CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

https://users.cs.fiu.edu/~giri/teach/BioinfF18.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.

Slide by Pevsner

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.

Five kingdom system (Haeckel, 1879)

Slide by Pevsner

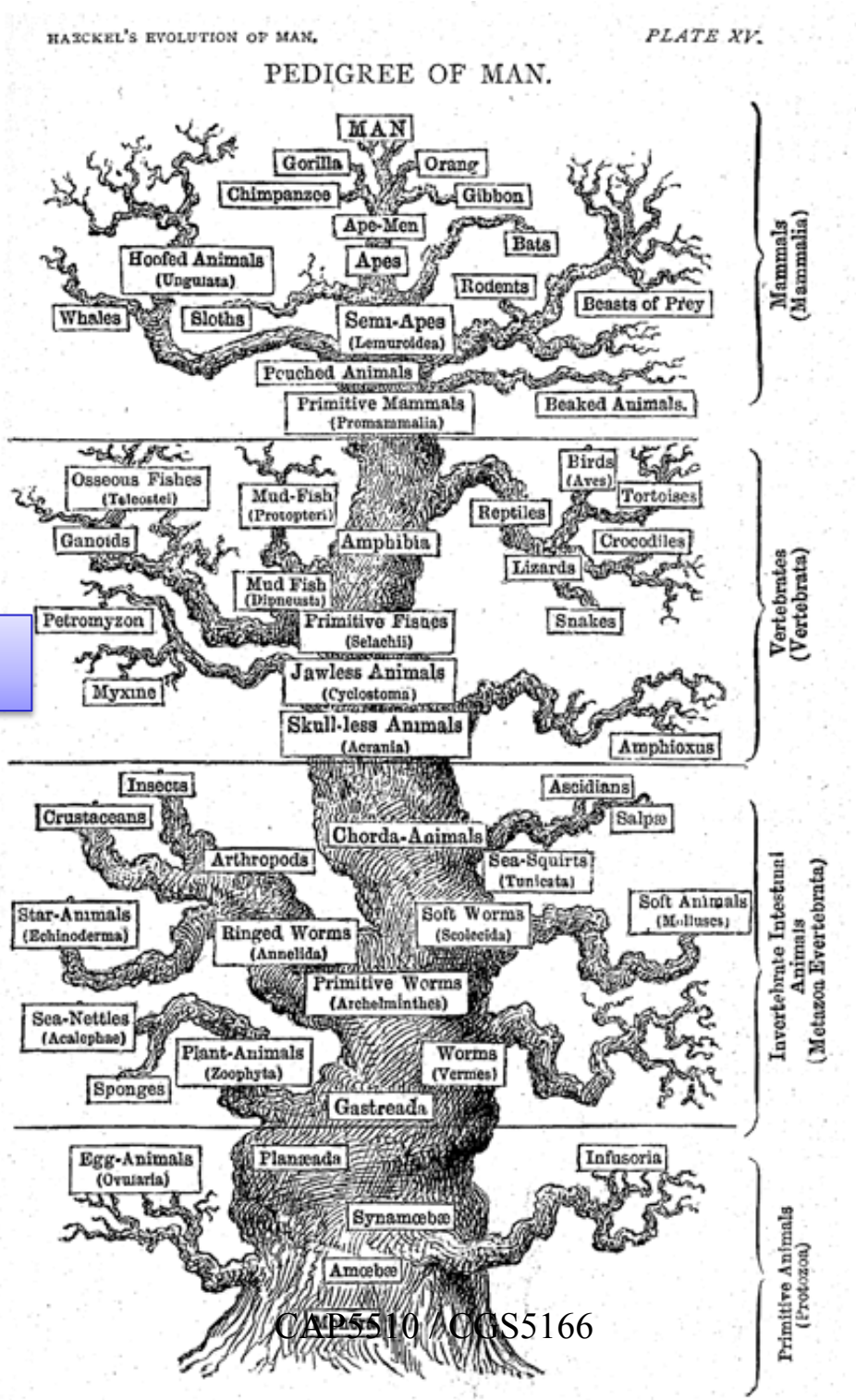
animals

plants

fungi

protists

monera



mammals

vertebrates

invertebrates

protozoa

Evolution & Phylogeny

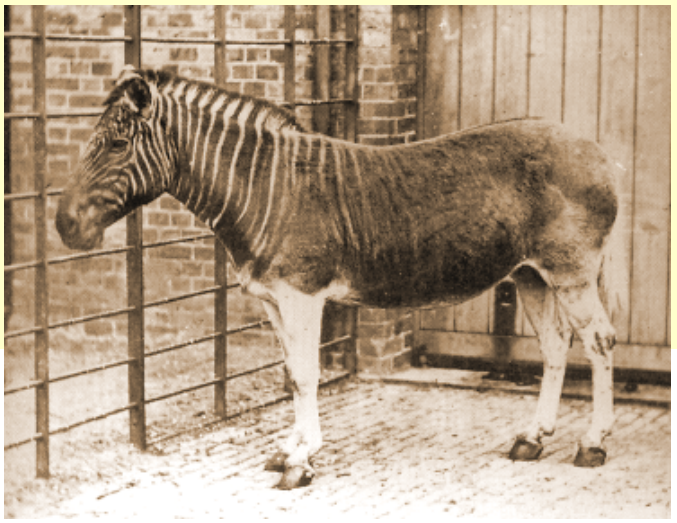
- At the molecular level, evolution is a process of mutation with selection.
- Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life.
- Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny relied on the comparison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

Slide by Pevsner

Questions for Phylogenetic Analysis

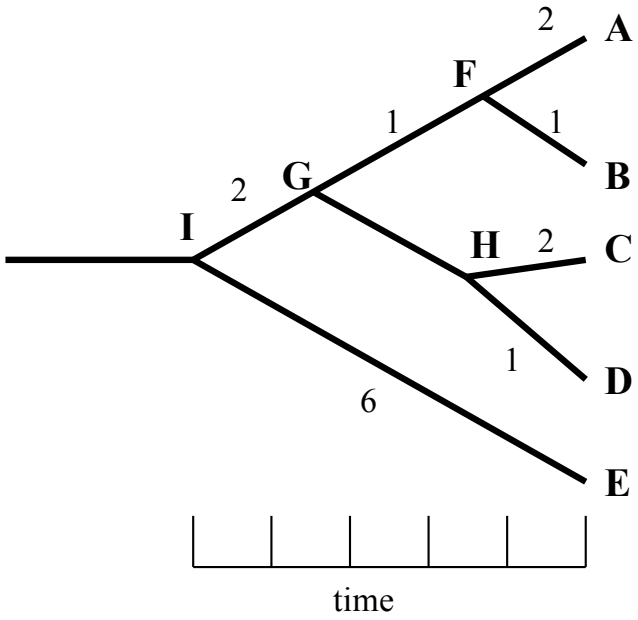
- How many genes are related to my favorite gene?
- How related are whales, dolphins & porpoises to cows?
- Where and when did HIV or other viruses originate?
- What is the history of life on earth?
- Was the extinct quagga more like a zebra or a horse?

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.

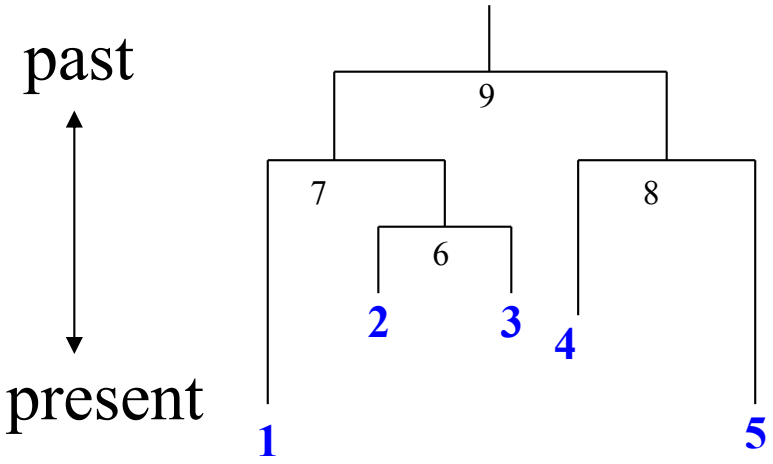


Slide by Pevsner

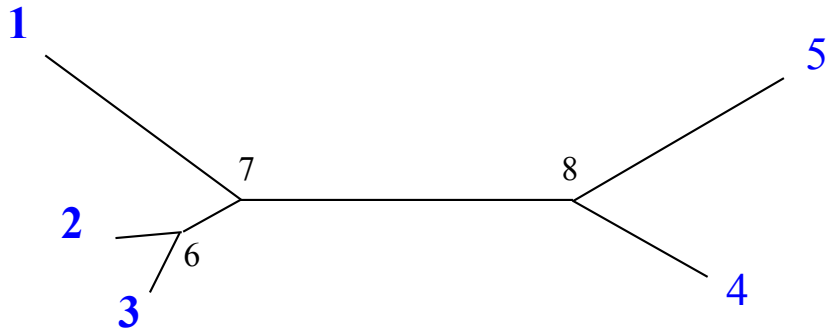
Tree Roots

- The root of a phylogenetic tree represents the common ancestor of the sequences. Some trees are unrooted, and thus do not specify the common ancestor.
- A tree can be rooted using an outgroup (that is, a taxon known to be distantly related from all other OTUs).

Tree nomenclature: roots

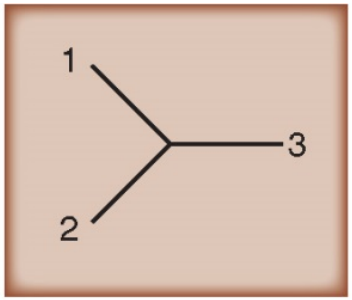


Rooted tree (specifies evolutionary path)

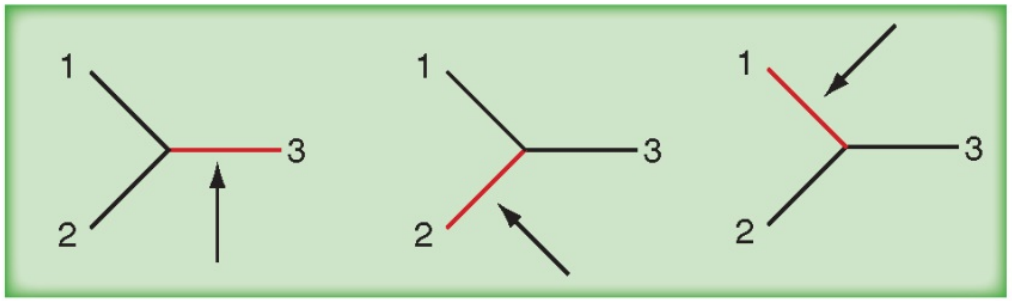


Unrooted tree

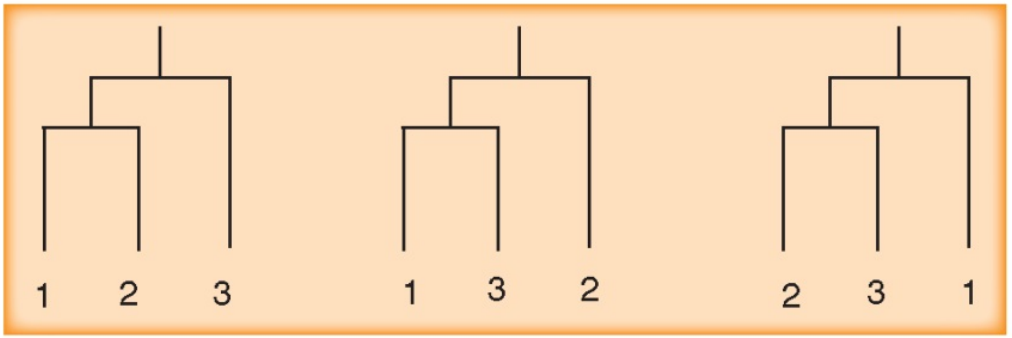
Numbers of rooted and unrooted trees: 3 OTUs



For three operational taxonomic units (OTUs) there is one possible unrooted tree.

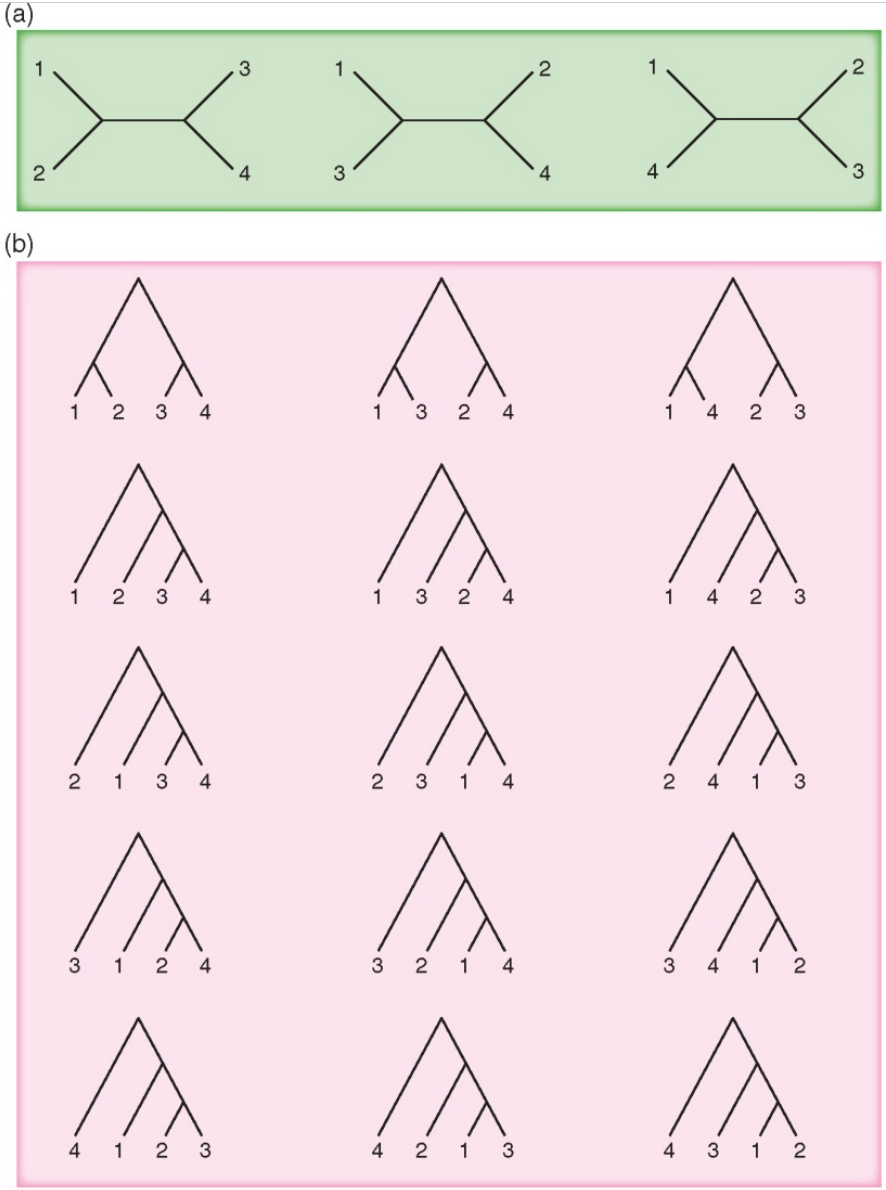


Any of the three edges can be selected to form a root.



Three rooted trees are possible.

Numbers of rooted and unrooted trees: 4 OTUs



For 4 OTUs there are three possible unrooted trees.

For 4 OTUs there are 15 possible rooted trees.

There is only one of these 15 trees that accurately describes the evolutionary process by which these four sequences evolved.

Tree Nomenclature

Slide by Pevsner

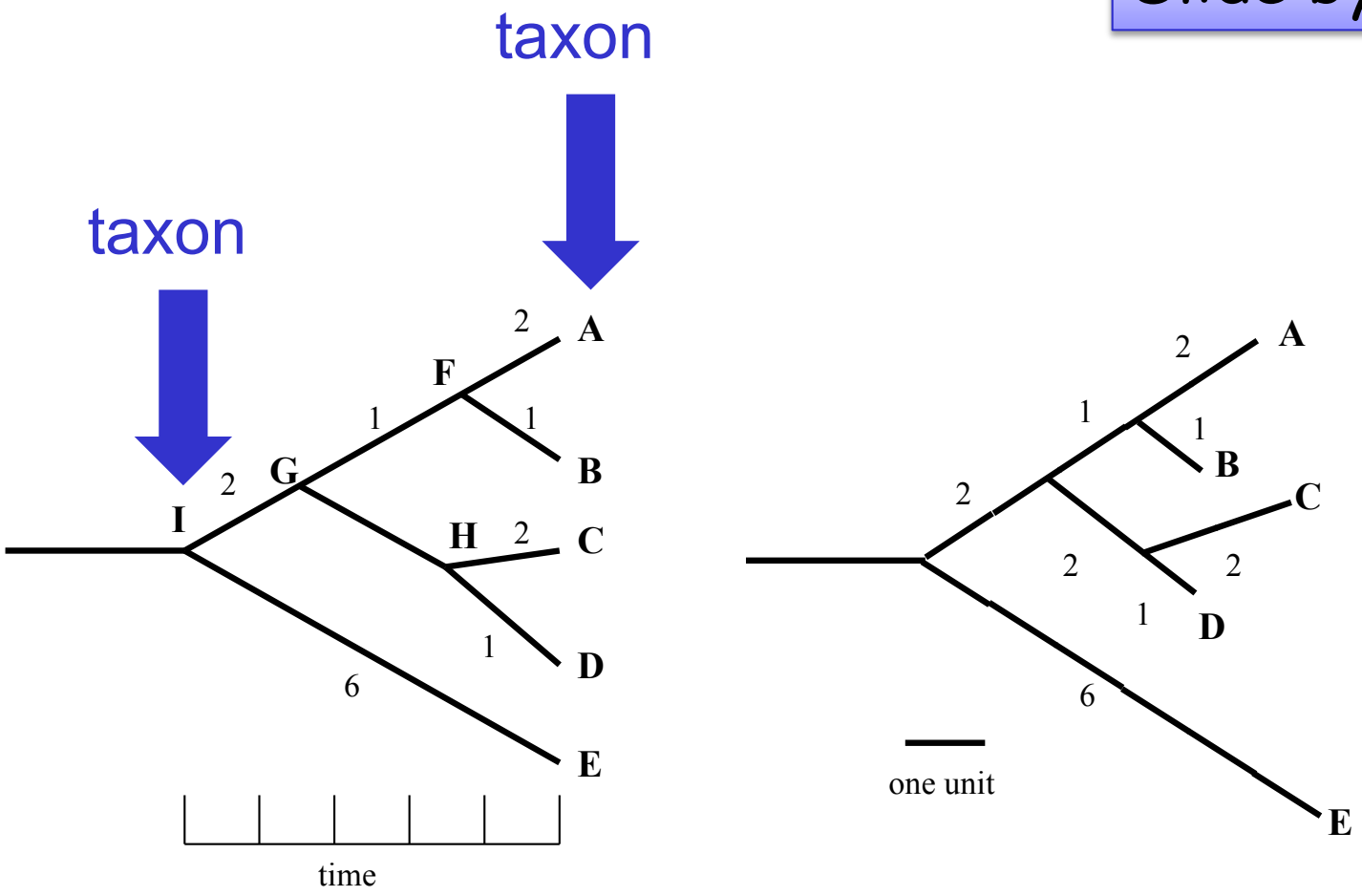


Fig. 7.8
Page 232

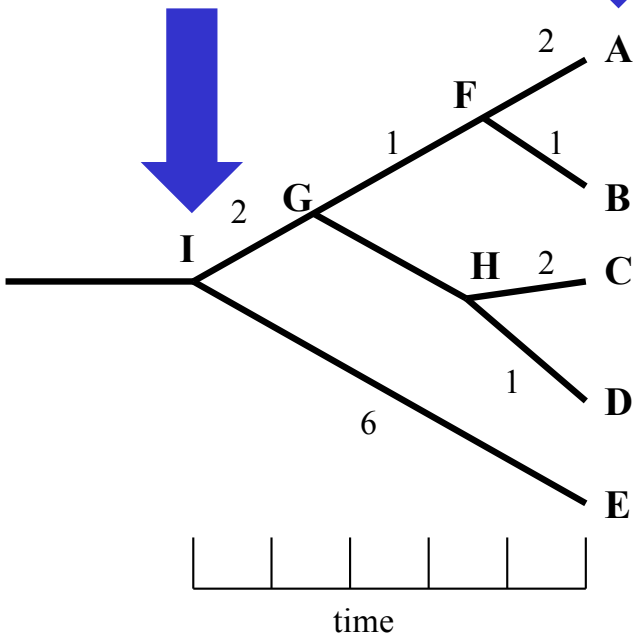
Tree nomenclature

Slide by Pevsner

operational taxonomic unit (OTU)

such as a protein sequence

taxon



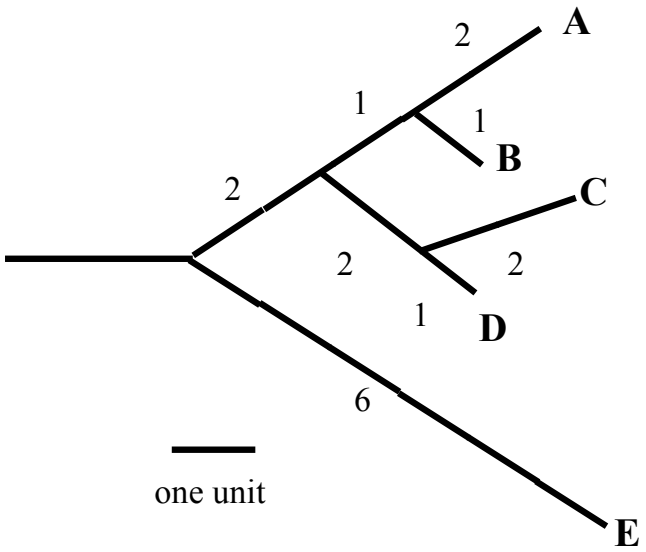


Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

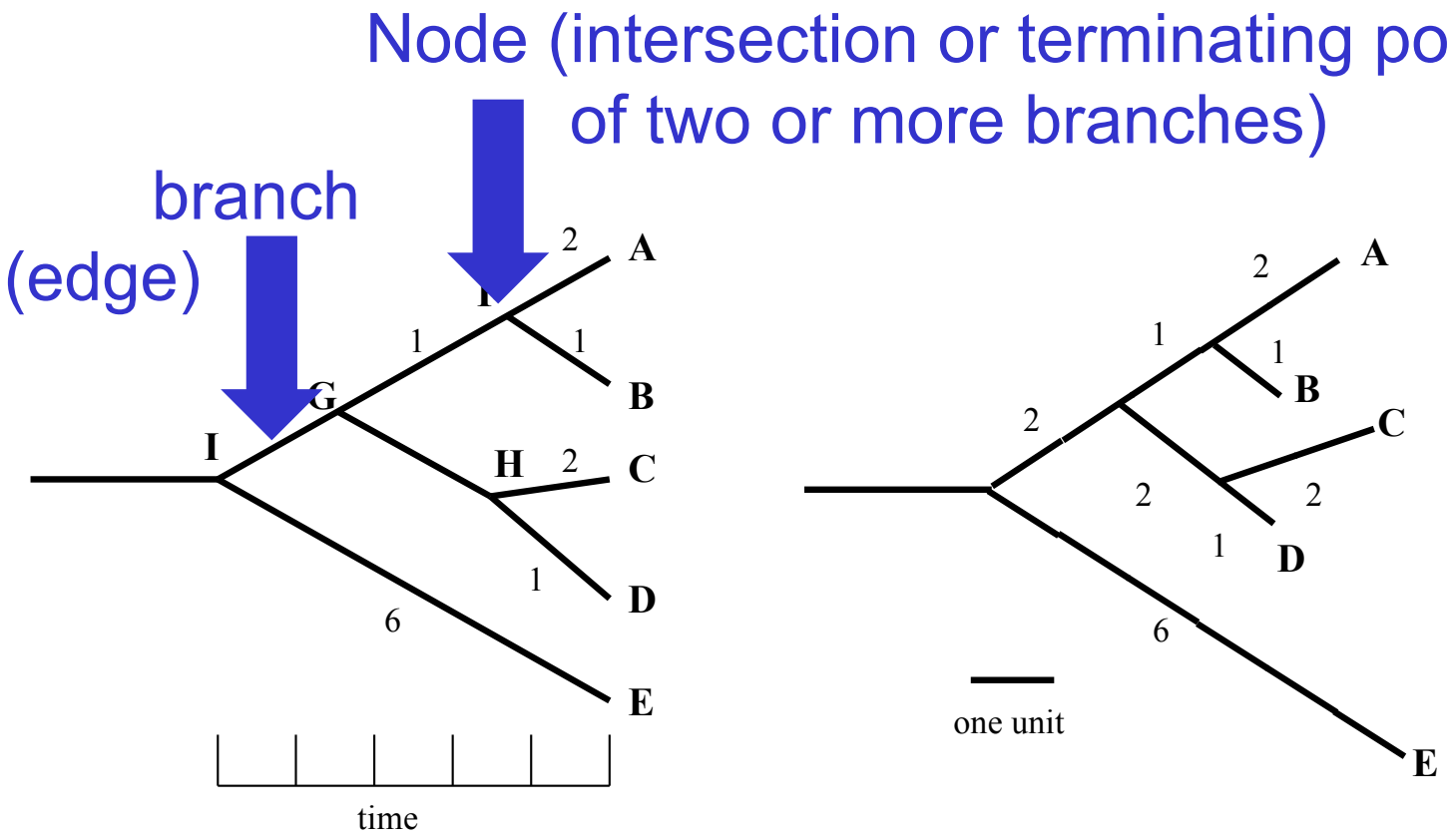
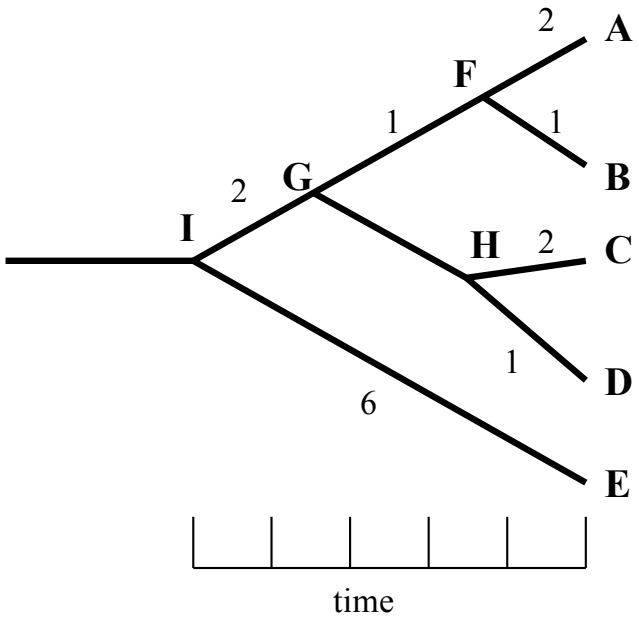


Fig. 7.8
Page 232

Tree nomenclature

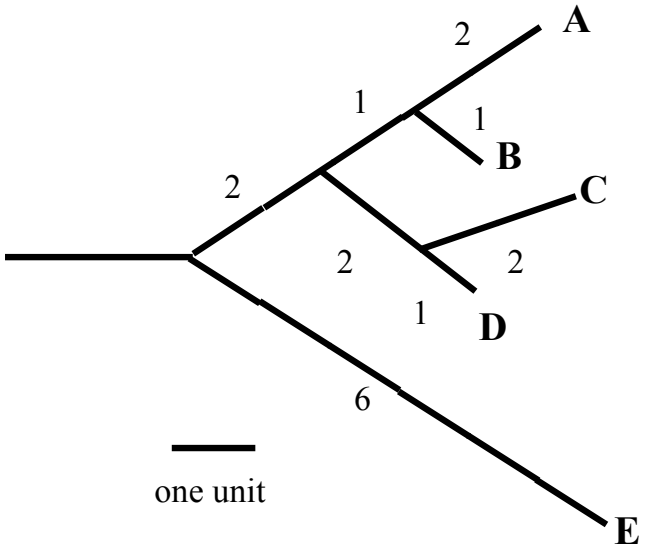
Slide by Pevsner

Branches are unscaled...



…OTUs are neatly aligned, and nodes reflect time

Branches are scaled...



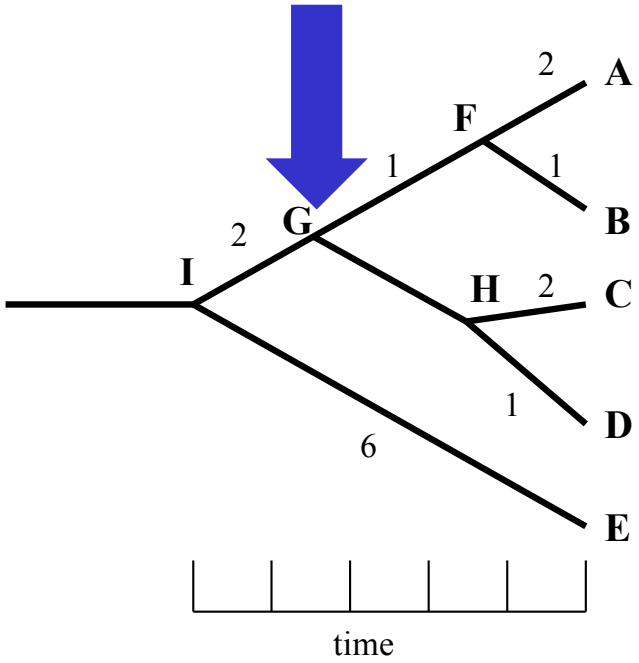
…branch lengths are proportional to number of amino acid changes

Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

bifurcating internal node



multifurcating internal node

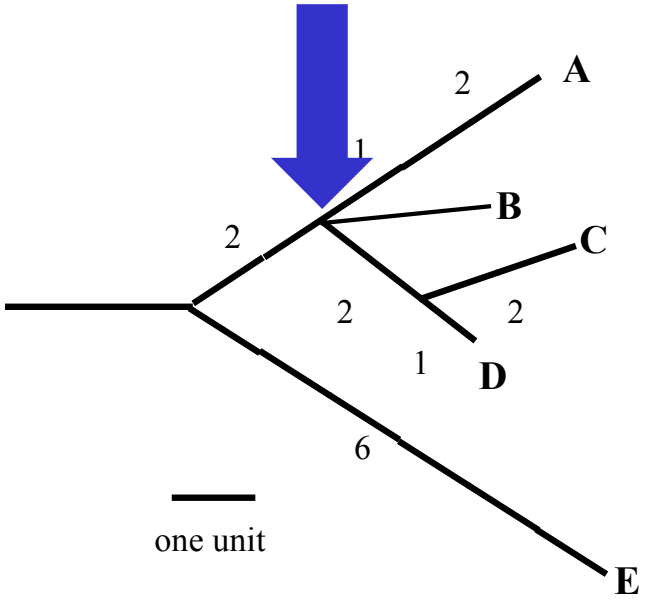
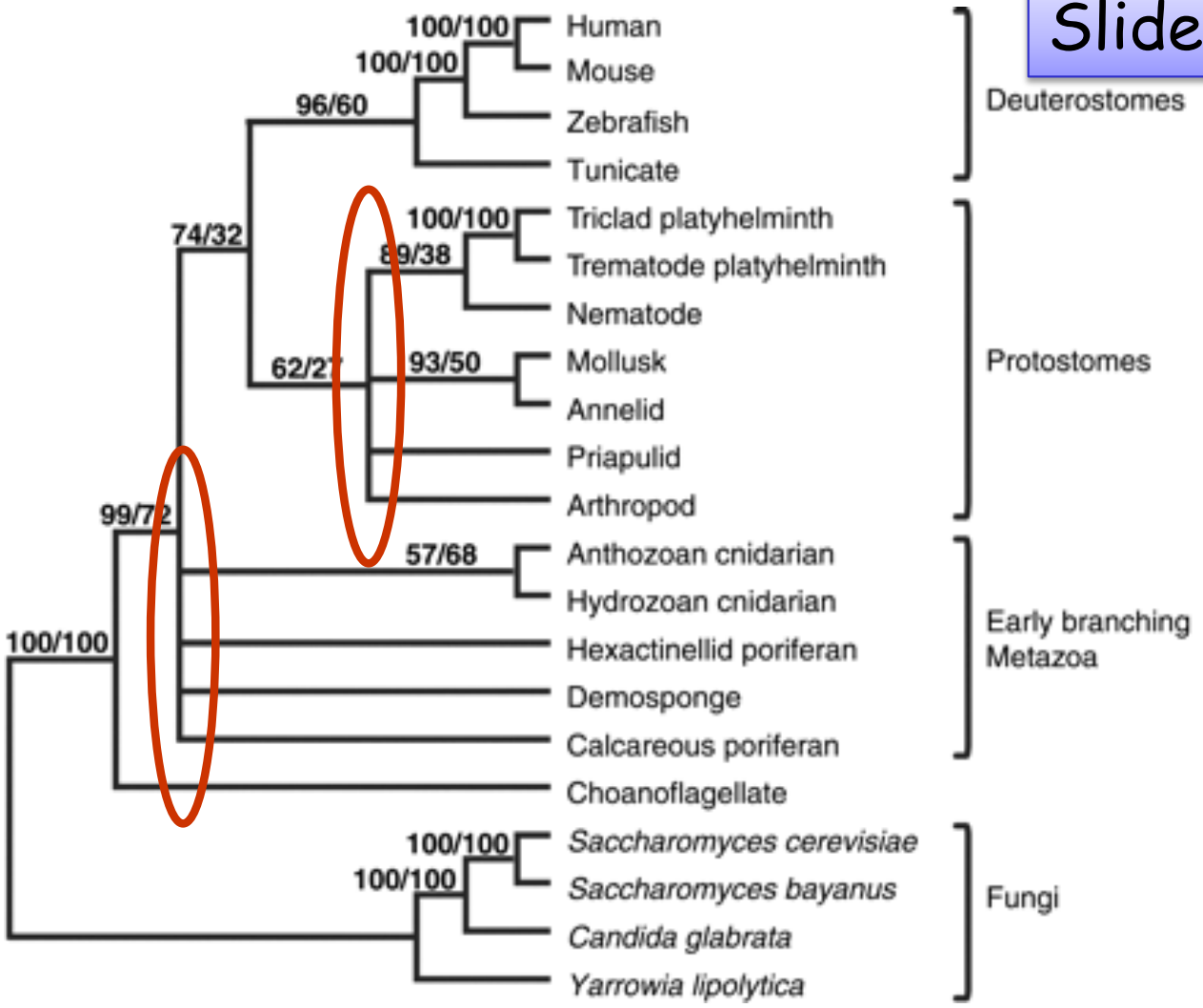


Fig. 7.9
Page 233

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

Slide by Pevsner

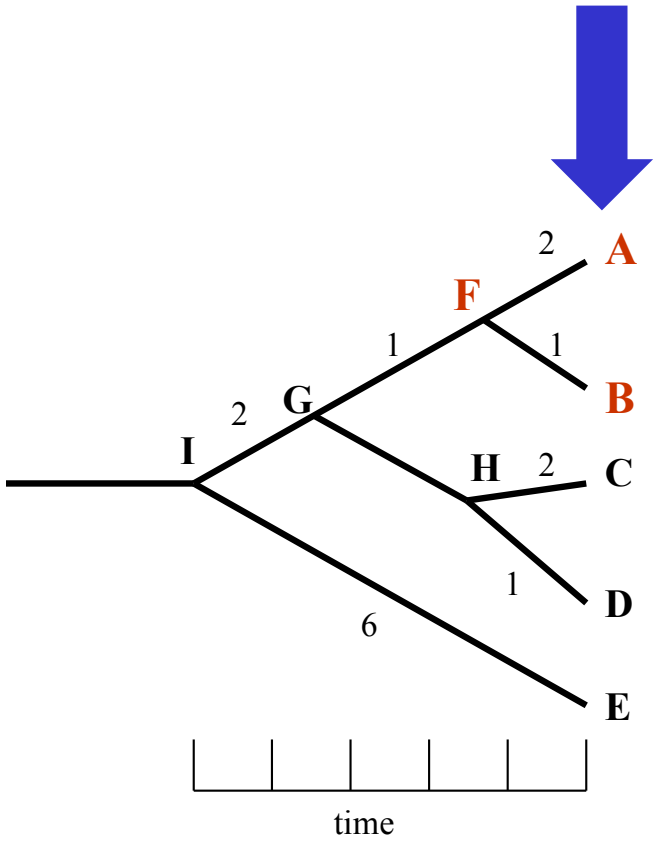


Rokas A. et al., Animal Evolution and the Molecular Signature of Radiations Compressed in Time, *Science* 310:1933 (2005), Fig. 1.

Tree nomenclature: clades

Slide by Pevsner

Clade ABF (monophyletic group)



Clade

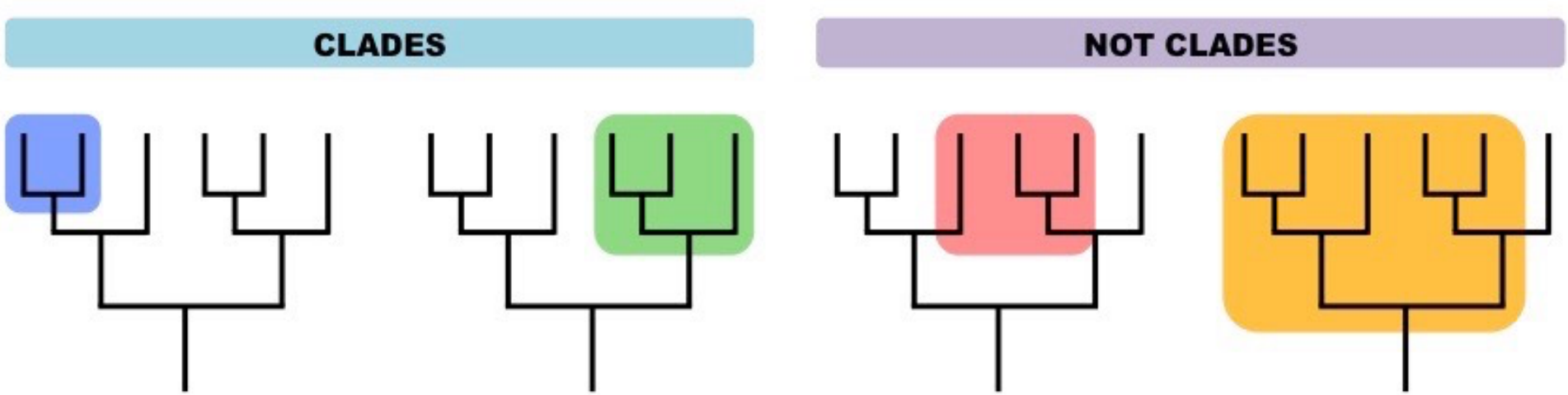
group of organisms believed to have evolved from a common ancestor

Monophyletic

a **group** of organisms that consists of all the descendants of a common ancestor

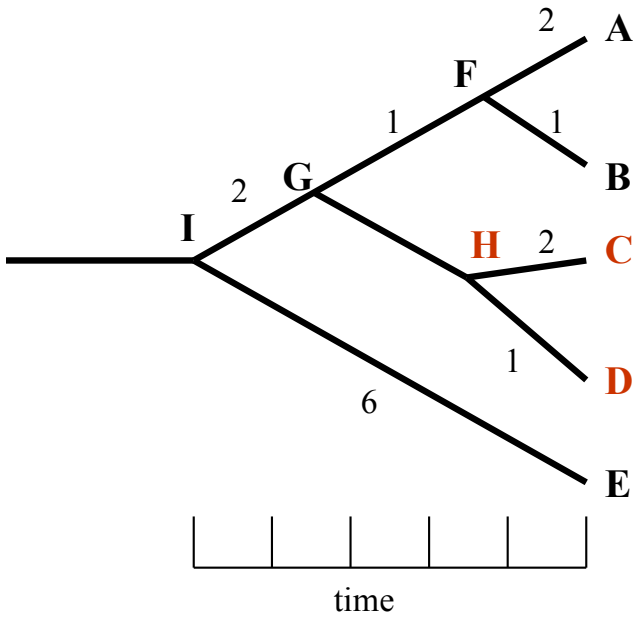
Fig. 7.8
Page 232

Tree nomenclature: clades



Tree nomenclature

Slide by Pevsner



Clade CDH

Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

Clade ABF/CDH/G

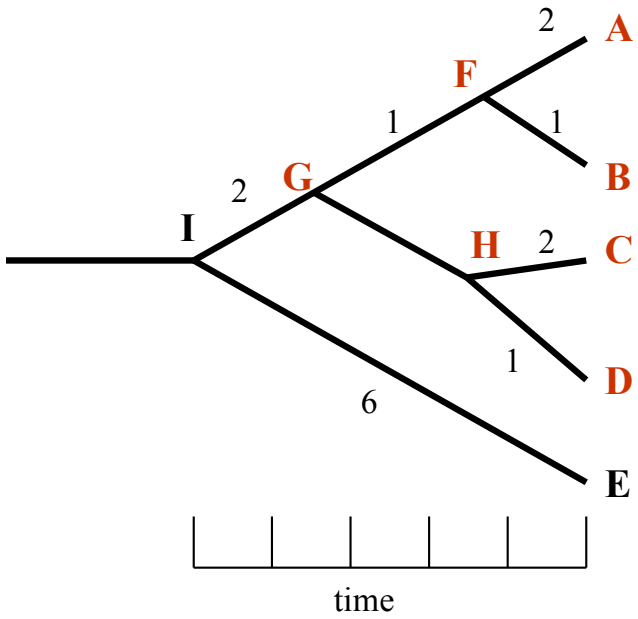
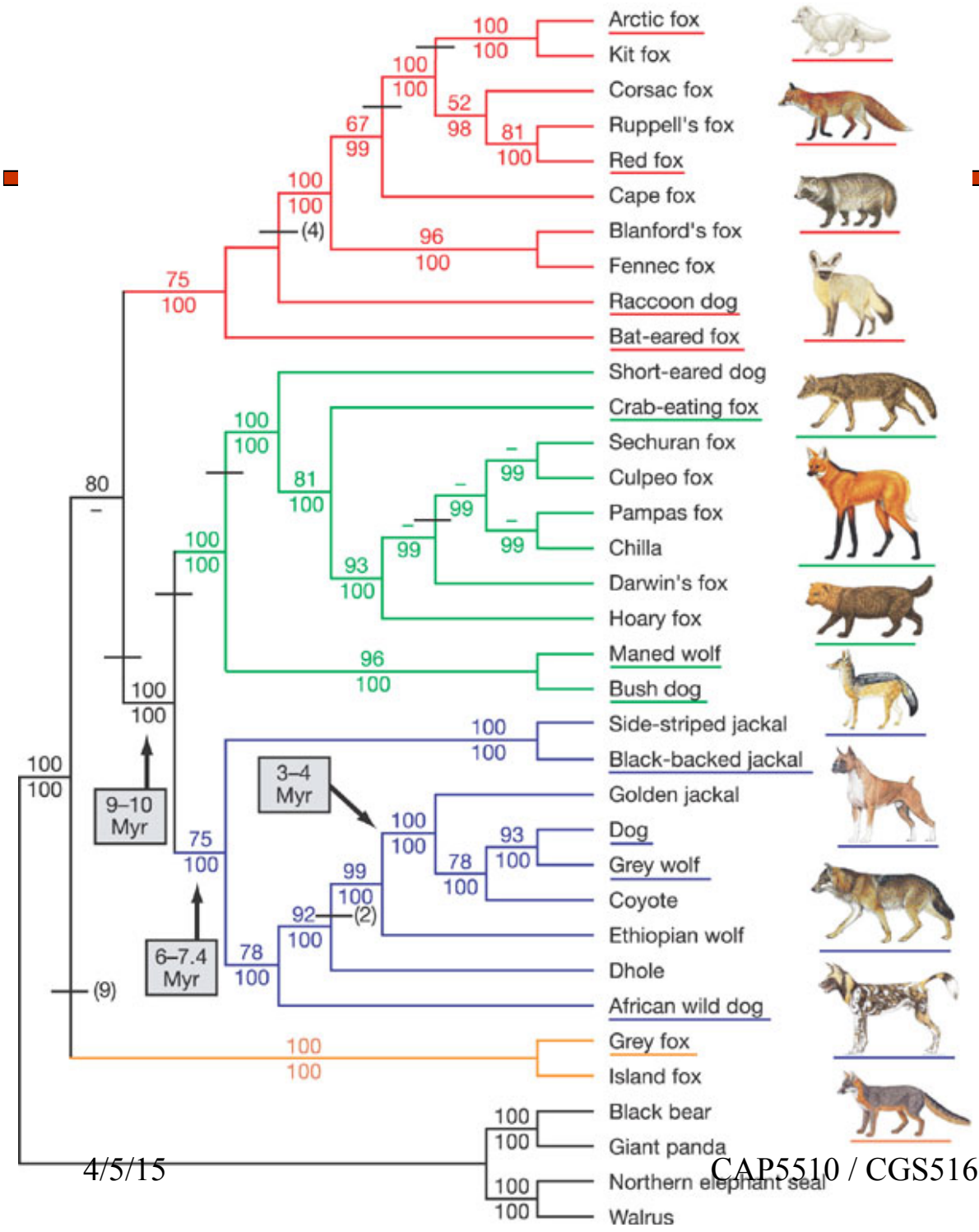


Fig. 7.8
Page 232

Examples of clades

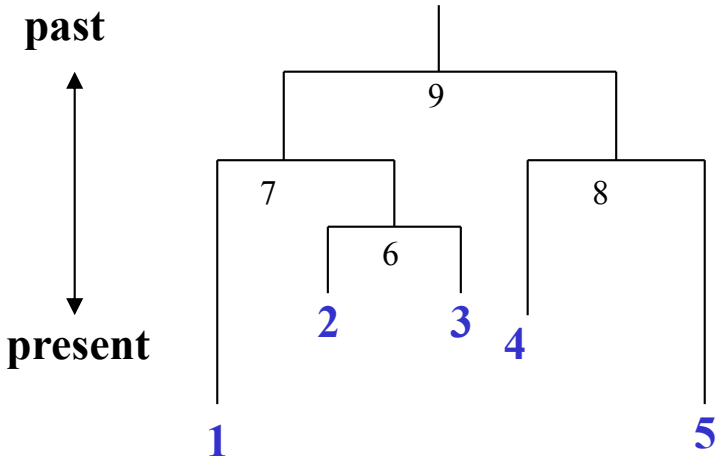
Slide by Pevsner

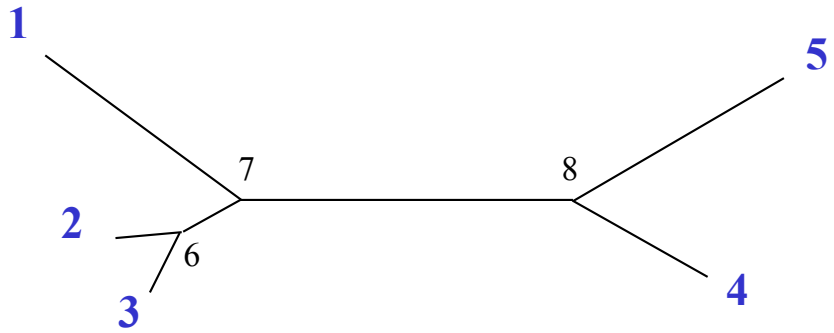


Lindblad-Toh et al., *Nature* 438: 803 (2005), fig. 10

Tree nomenclature: roots

Slide by Pevsner





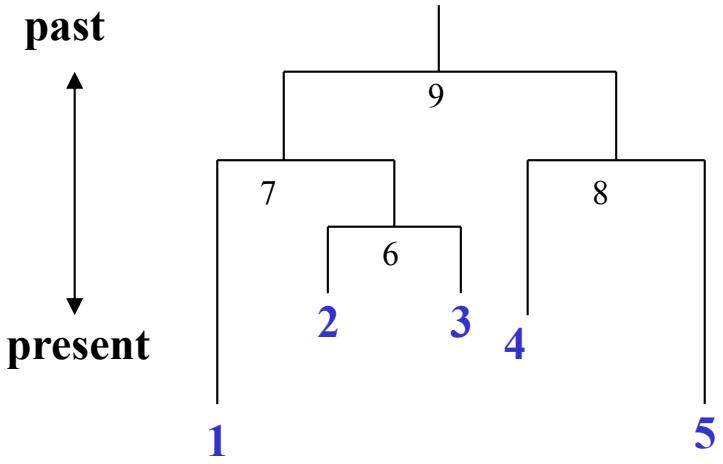
Rooted tree (specifies evolutionary path)

Unrooted tree

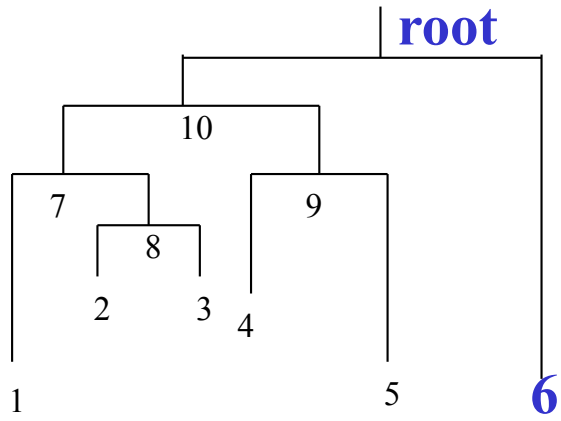
Fig. 7.10
Page 234

Tree nomenclature: outgroup rooting

Slide by Pevsner



Rooted tree



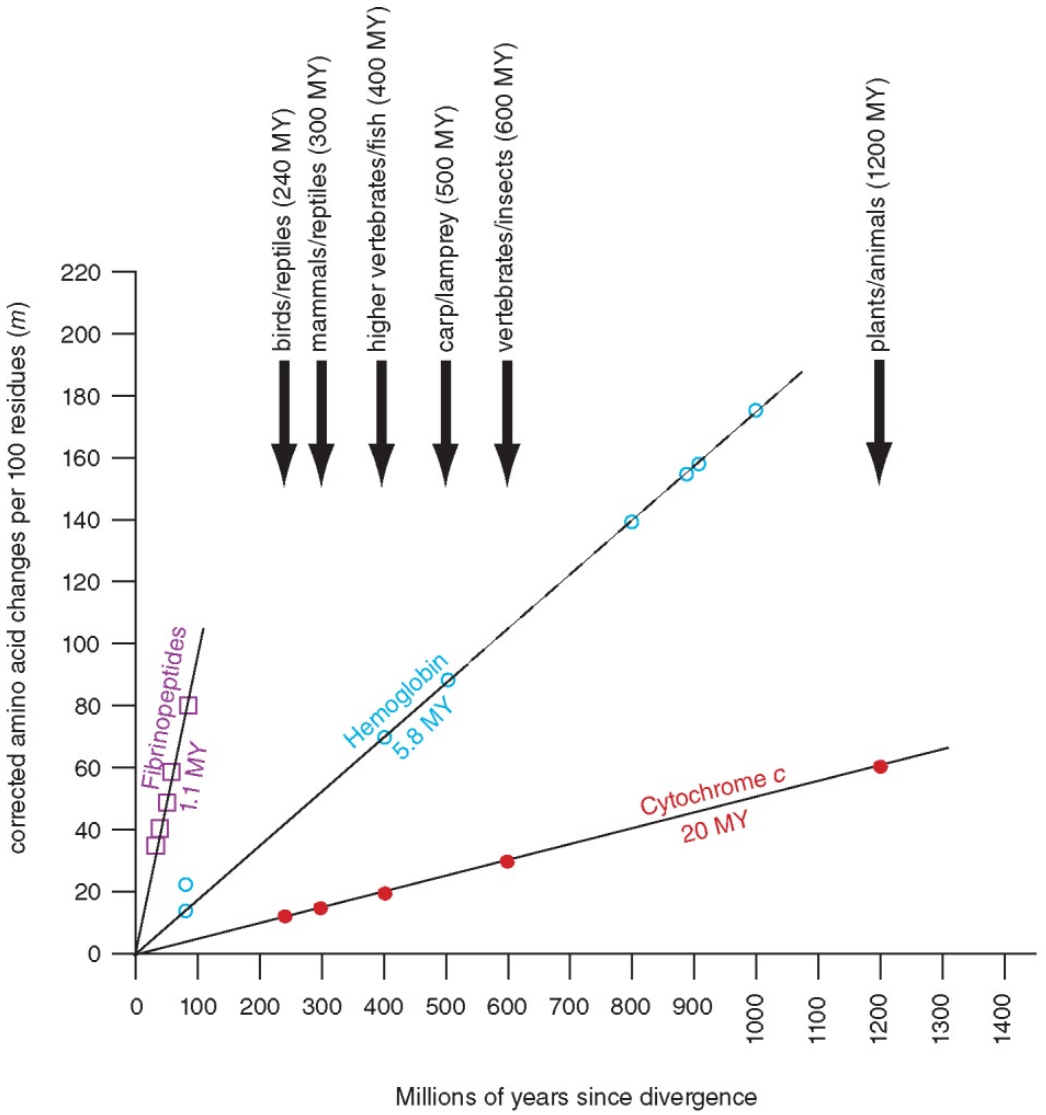
Outgroup
(used to place the root)

Fig. 7.10
Page 234

Molecular Clock Hypothesis

- The molecular clock is a figurative term for a technique that uses the **mutation rate** of biomolecules to deduce the time in prehistory when two or more life forms diverged.
- In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes c, and fibrinopeptides.
- Some proteins appeared to evolve slowly, while others evolved rapidly.
- Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock:
  - **For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages**.

Molecular Clock Hypothesis



Molecular Clock Hypothesis — Implications

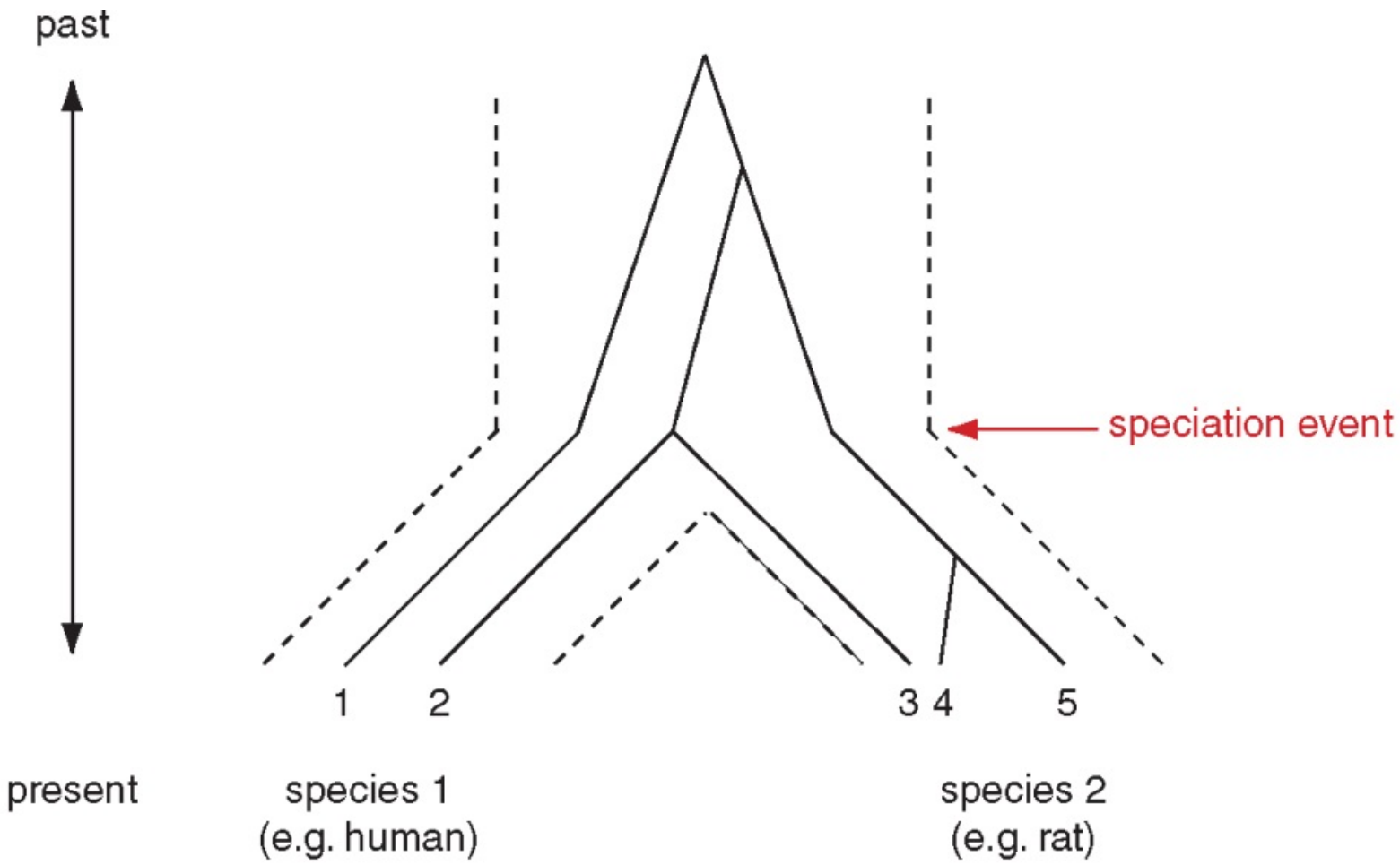
- If protein sequences evolve at constant rates, they can be used to estimate the times that sequences diverged. This is analogous to dating geological specimens by radioactive decay.

Species trees versus gene/protein trees

Molecular evolutionary studies can be complicated by the fact that both species and genes evolve. Speciation usually occurs when a species becomes reproductively isolated. In a species tree, each internal node represents a speciation event.

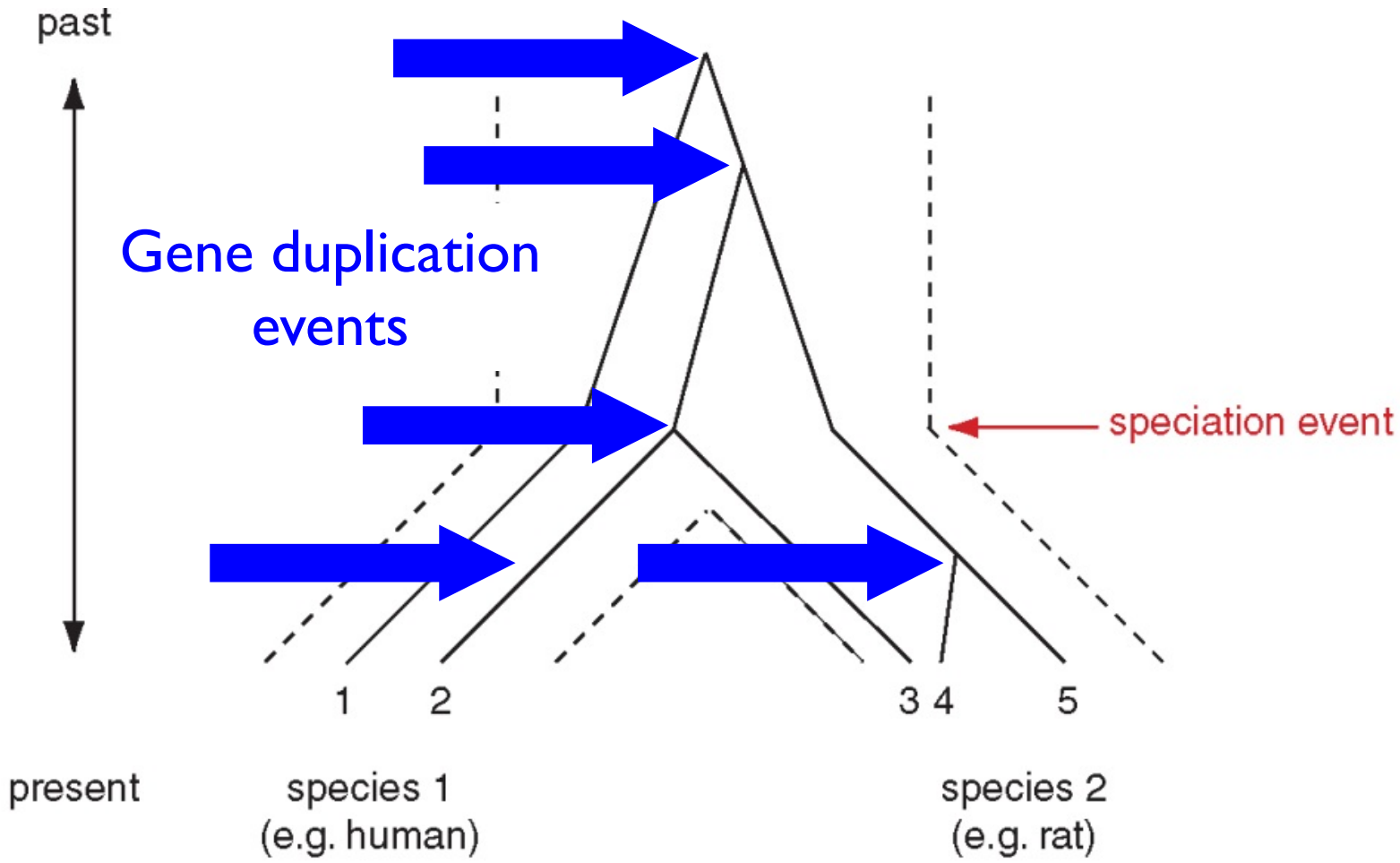
Genes (and proteins) may duplicate or otherwise evolve before or after any given speciation event. The topology of a gene (or protein) based tree may differ from the topology of a species tree.

Species trees versus gene/protein trees



B&FG 3e
Fig. 7.13
Page 267

Species trees versus gene/protein trees



B&FG 3e Fig. 7.13 Page 267

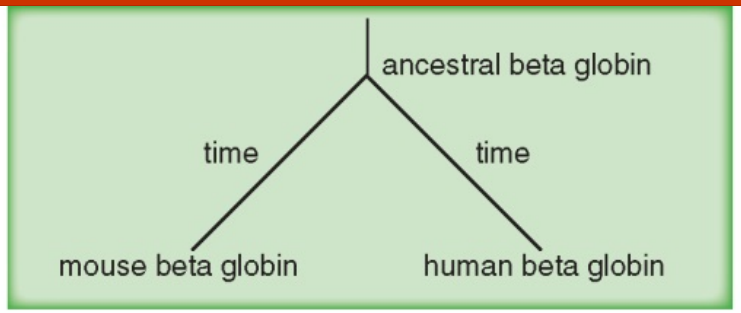
A gene (e.g. a globin) may duplicate *before* or *after* two species diverge!

Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

Some substitutions in a DNA sequence alignment can be directly observed: single nucleotide substitutions, sequential substitutions, coincidental substitutions. Additional mutational events can be inferred by analysis of ancestral sequences.

Two sequences (human and mouse) and their common ancestor: we can infer which DNA changes occurred over time



ancestral		M	V	H	L	S	P	V	E	K	S	A	V
human		M	V	H	L	T	P	E	E	K	S	A	V
mouse		M	V	H	L	T	D	A	E	K	S	A	V

ancestral	5′	ATG	GTG	CAT	CTG	AGT	CCT	GTT	CAG	AAG	TCT	GCT	GTT 3′
human	5′	ATG	GTG	CAT	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT 3′
mouse	5′	ATG	GTG	CAC	CTG	ACT	GAT	GCT	GAG	AAG	TCT	GCT	GTC 3′

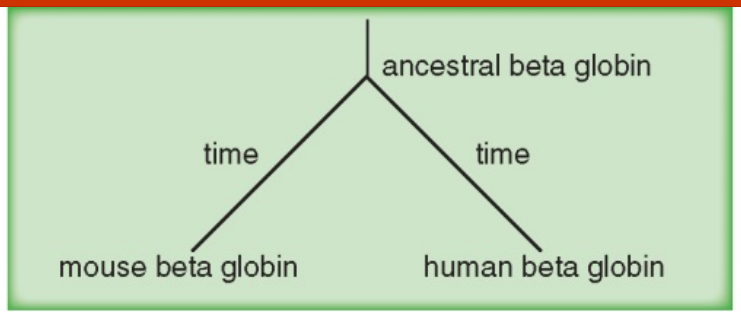
ancestral
human
mouse

protein

DNA

B&FG 3e
Fig. 7.15
Page 269

Two sequences (human and mouse) and their common ancestor: we can infer which DNA changes occurred over time



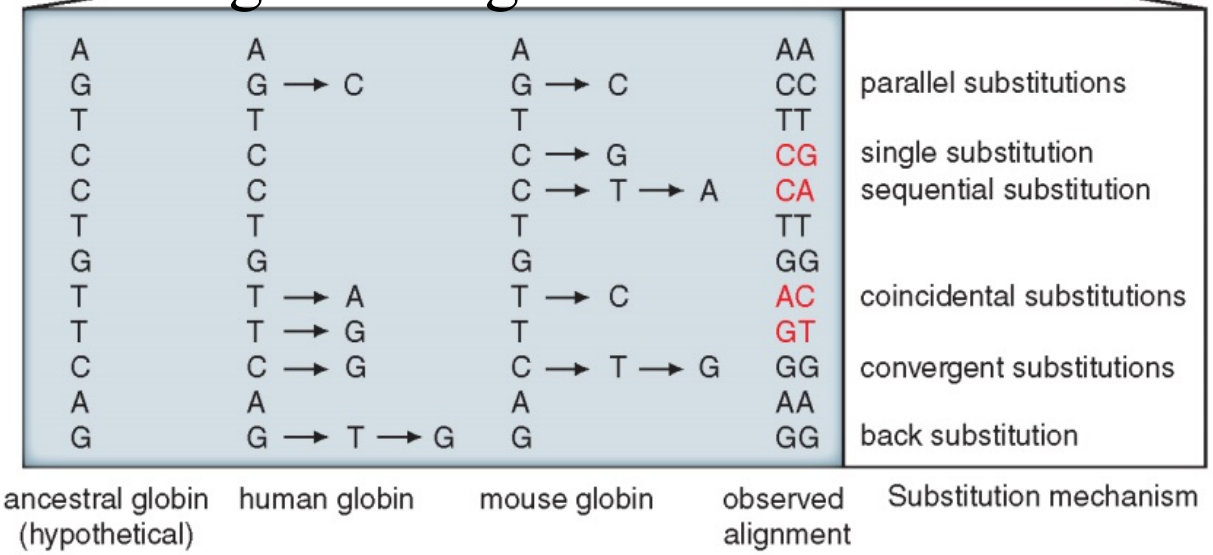

```
ancestral      M   V   H   L   S   P   V   E   K   S   A   V
human          M   V   H   L   T   P   E   E   K   S   A   V
mouse          M   V   H   L   T   D   A   E   K   S   A   V

ancestral  5′ ATG GTG CAT CTG AGT CCT GTT CAG AAG TCT GCT GTT 3′
human      5′ ATG GTG CAT CTG ACT CCT GAG GAG AAG TCT GCC GTT 3′
mouse      5′ ATG GTG CAC CTG ACT GAT GCT GAG AAG TCT GCT GTC 3′
```

ancestral globin

human globin

mouse globin



parallel substitutions

single sequential

coincidental

convergent

back substitution

B&FG 3e
Fig. 7.15
Page 269

Step matrices: number of steps required to change a character

(a)

	A	C	T	G
A	0	1	1	1
C	1	0	1	1
T	1	1	0	1
G	1	1	1	0

nucleotide step matrix

(b)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
C		0	2	3	1	1	2	2	3	2	3	2	2	3	1	1	2	2	1	1
D			0	1	2	1	1	2	2	2	3	1	2	2	2	2	2	1	3	1
E				0	3	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2
F					0	2	2	1	3	1	2	2	2	3	2	1	2	1	2	1
G						0	2	2	2	2	2	2	2	2	1	1	2	1	1	2
H							0	2	2	1	3	1	1	1	1	2	2	2	3	1
I								0	1	1	1	1	2	2	1	1	1	1	3	2
K									0	2	1	1	2	1	1	2	1	2	2	2
L										0	1	2	1	1	1	1	2	1	1	2
M											0	2	2	2	1	2	1	1	2	3
N												0	2	2	2	1	1	2	3	1
P													0	1	1	1	1	2	2	2
Q														0	1	2	2	2	2	2
R															0	1	1	2	1	2
S																0	1	2	1	1
T																	0	2	2	2
V																		0	2	2
W																			0	2
Y																				0

amino acid step matrix

B&FG 3e Fig. 7.16 Page 271

For amino acids, between 1 and 3 nucleotide changes are required to change one residue to another.

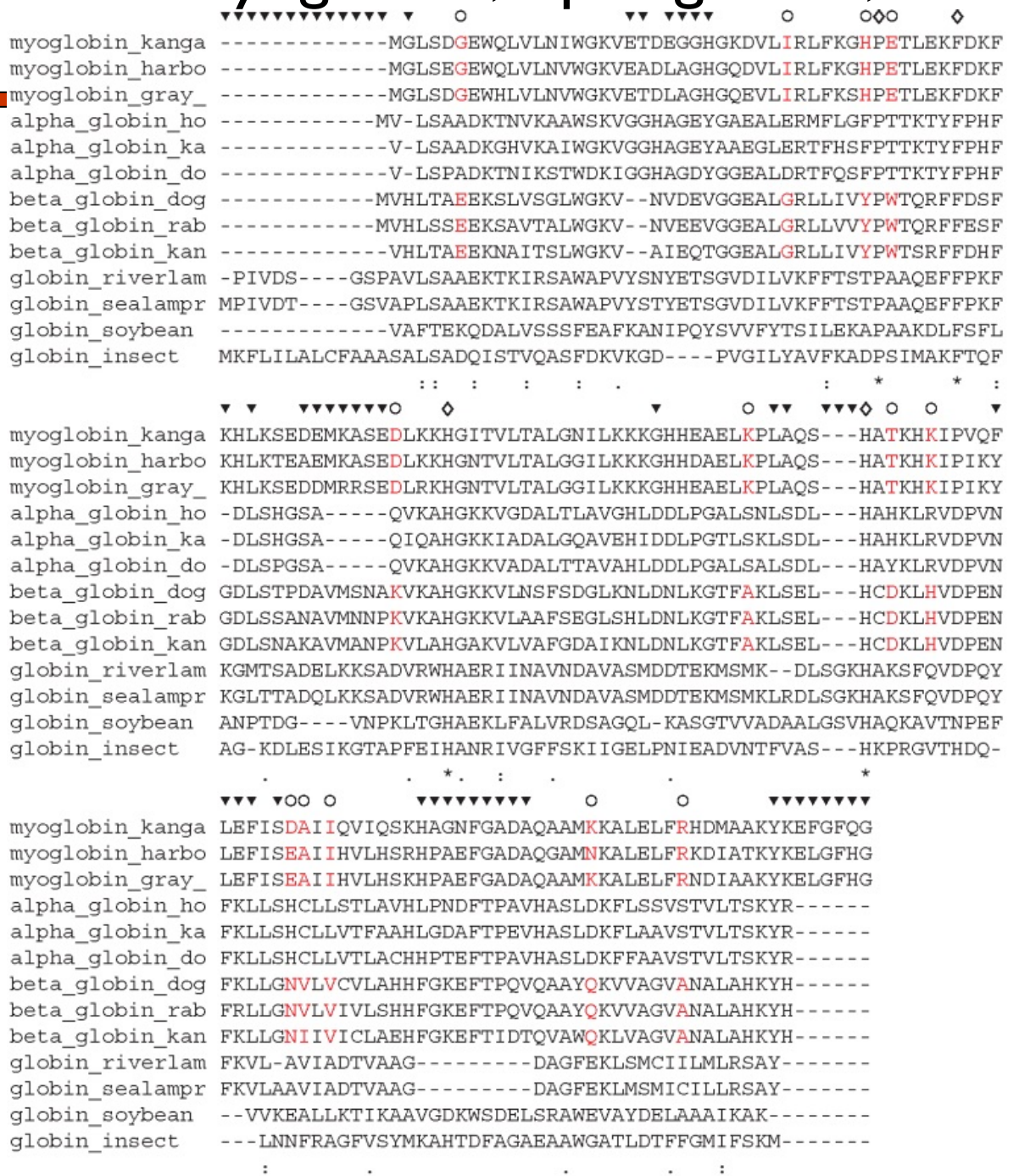
Stage 2: Multiple sequence alignment

The fundamental basis of a phylogenetic tree is a multiple sequence alignment.

(If there is a misalignment, or if a nonhomologous sequence is included in the alignment, it will still be possible to generate a tree.)

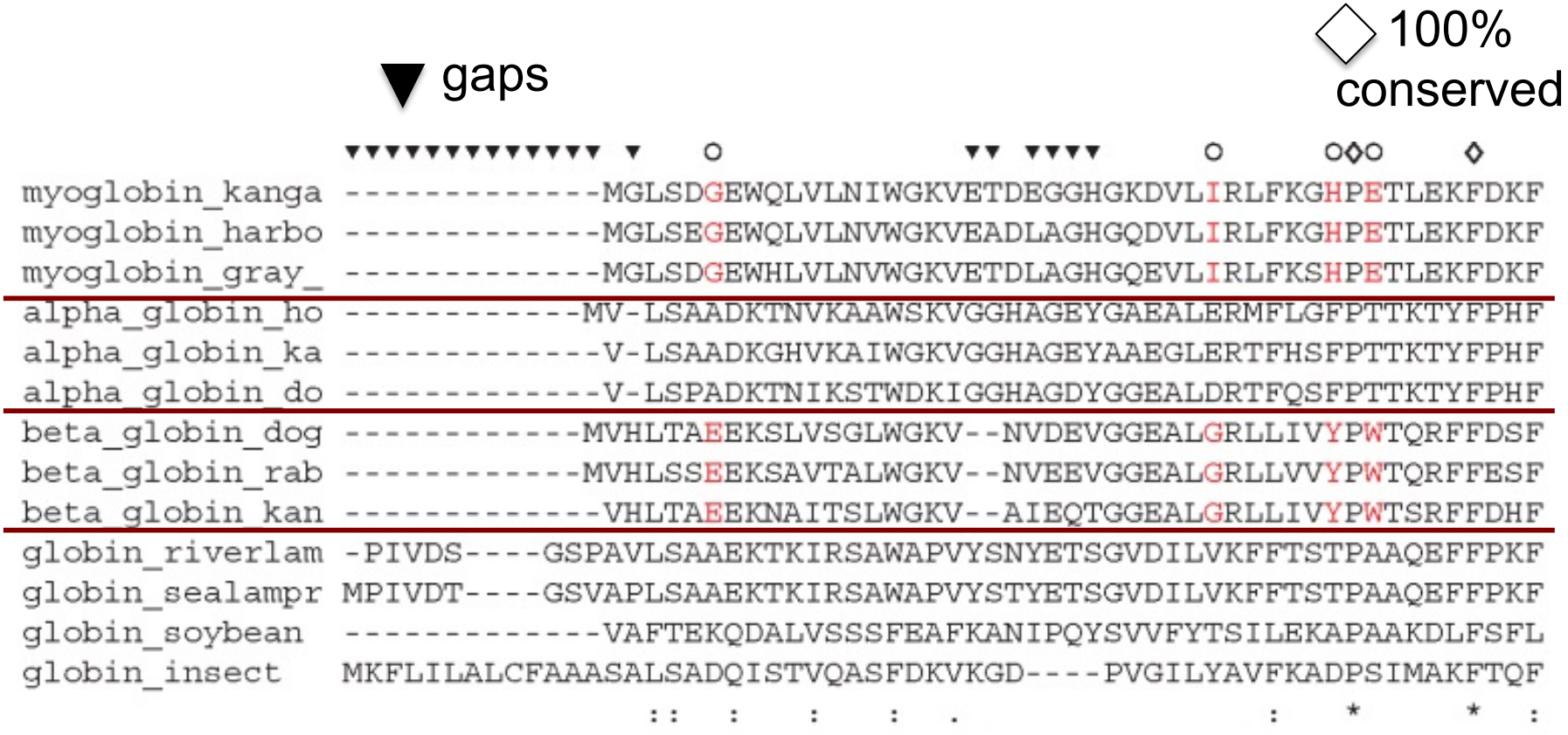
Consider the following alignment of 13 homologous globin proteins (see Fig. 3.2)

Multiple alignment of myoglobins, alpha globins, beta globins



B&FG 3e
Fig. 7.17
Page 272

Open circles: positions that distinguish myoglobins, alpha globins, beta globins



B&FG 3e
Fig. 7.17
Page 272

Stage 2: Multiple sequence alignment

[1] Confirm that all sequences are homologous

[2] Adjust gap creation and extension penalties as needed to optimize the alignment

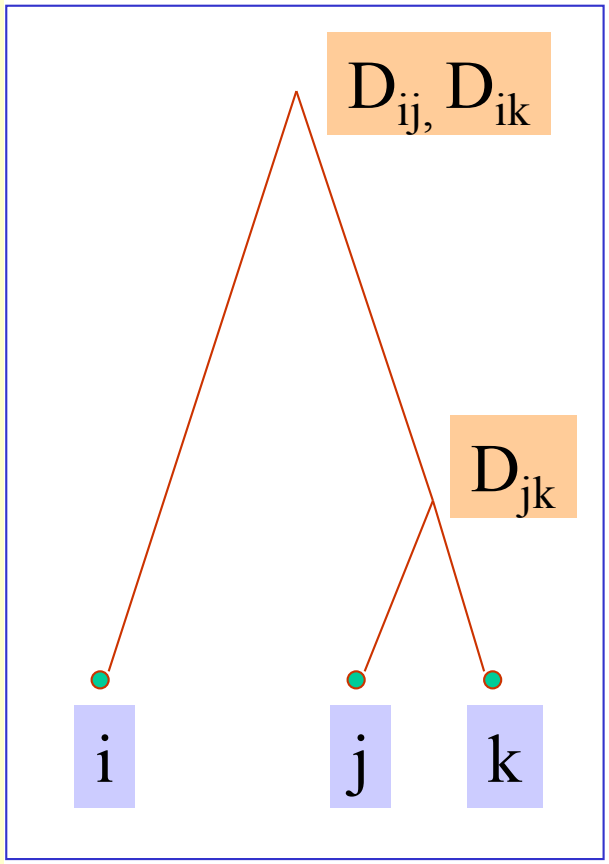
[3] Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for all taxa (delete columns having incomplete data).

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

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

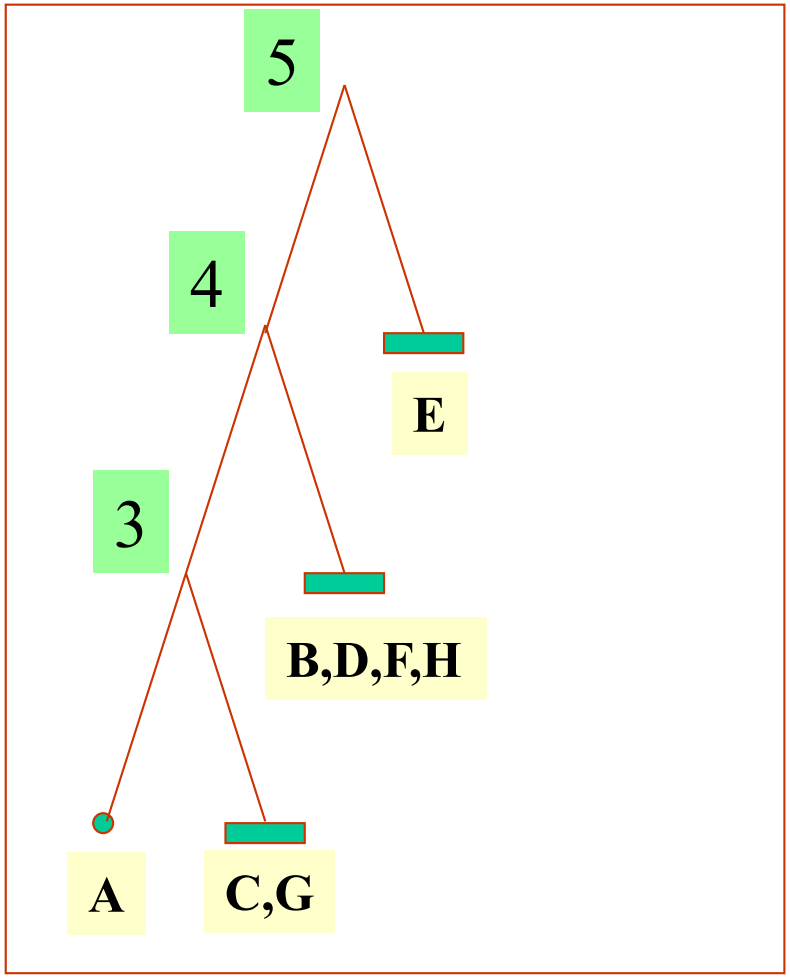
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

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

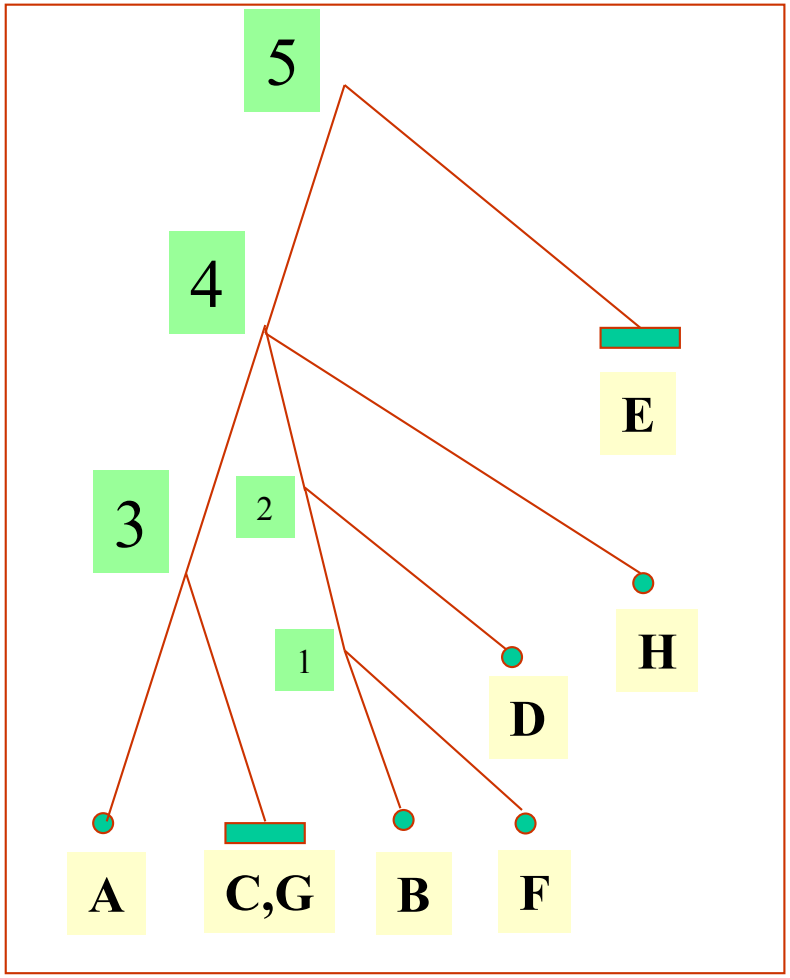
Ultrametric: Example

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



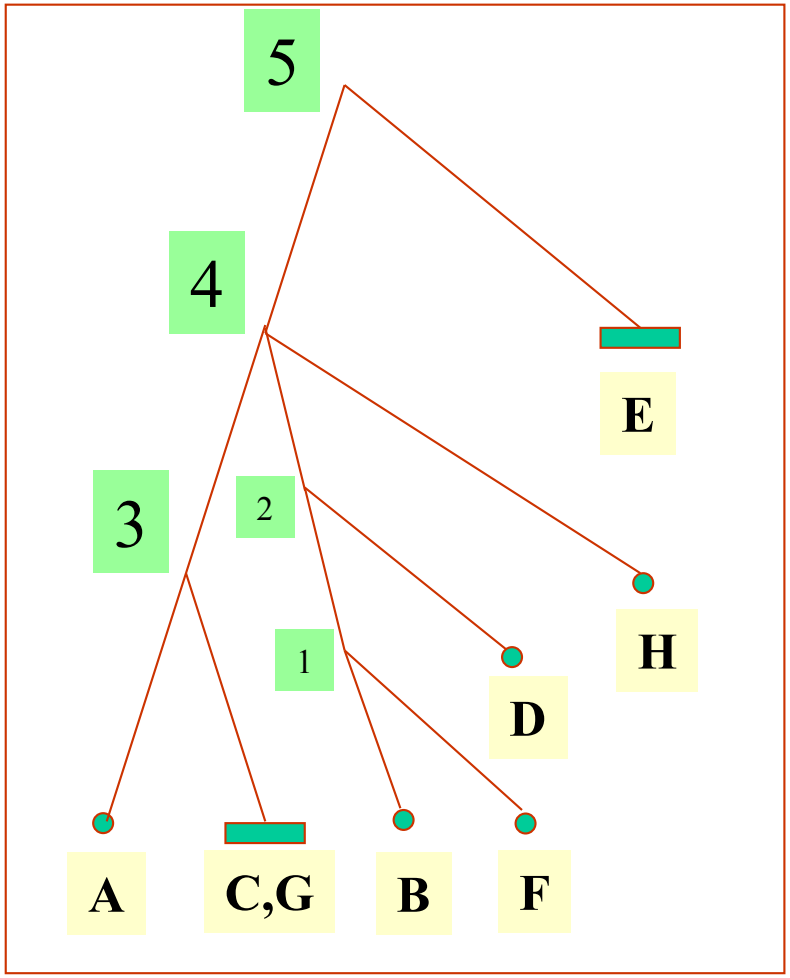
Ultrametric: Example

	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								



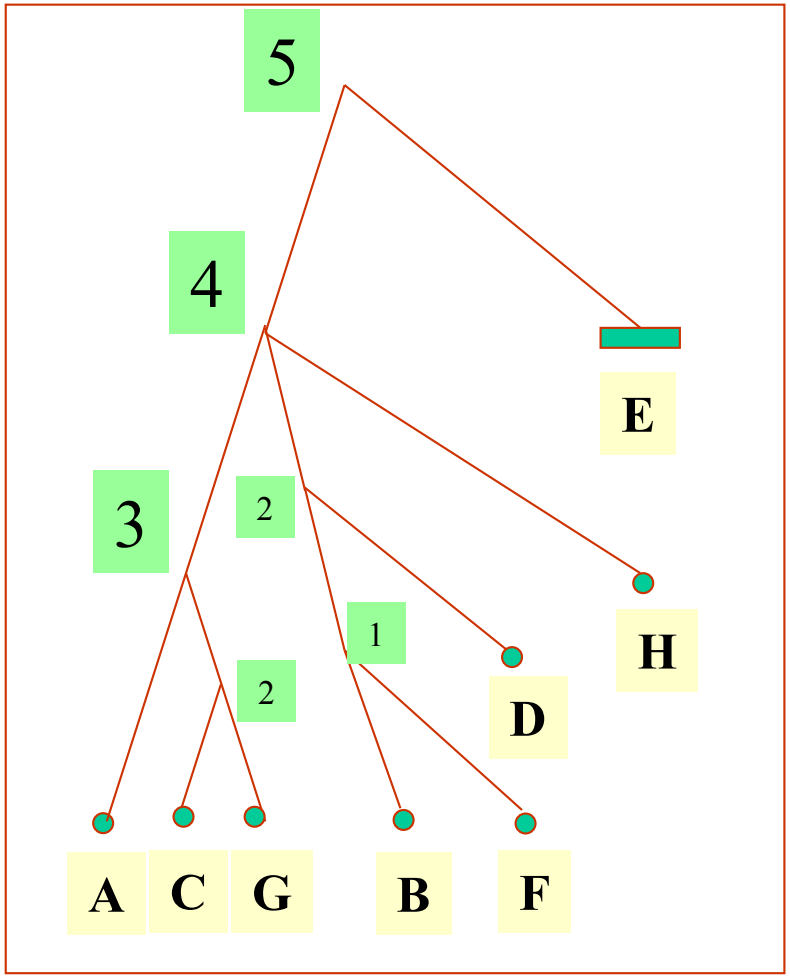
Ultrametric: Distances Computed

	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: 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

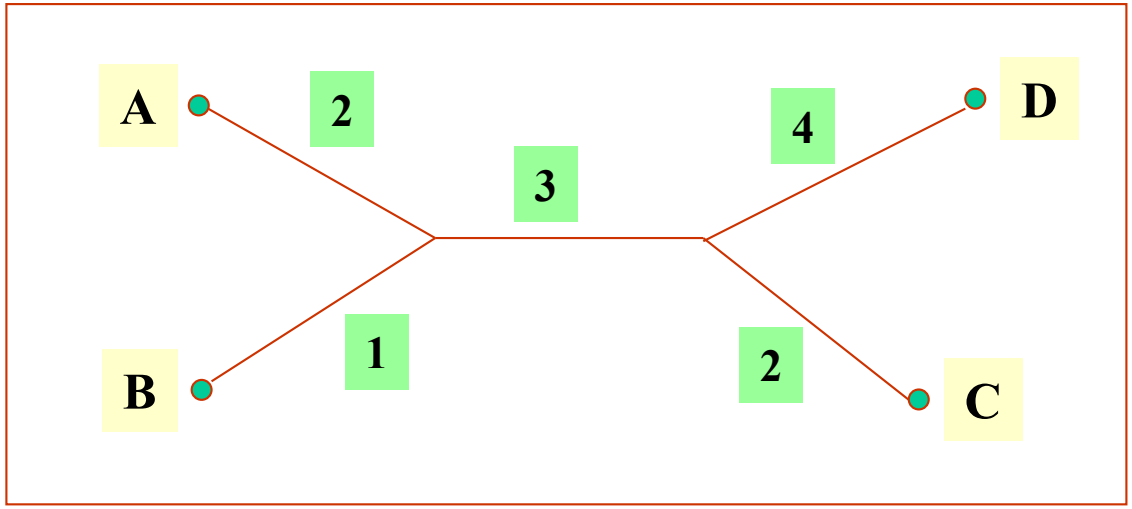
Ultrametric Data Sources

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

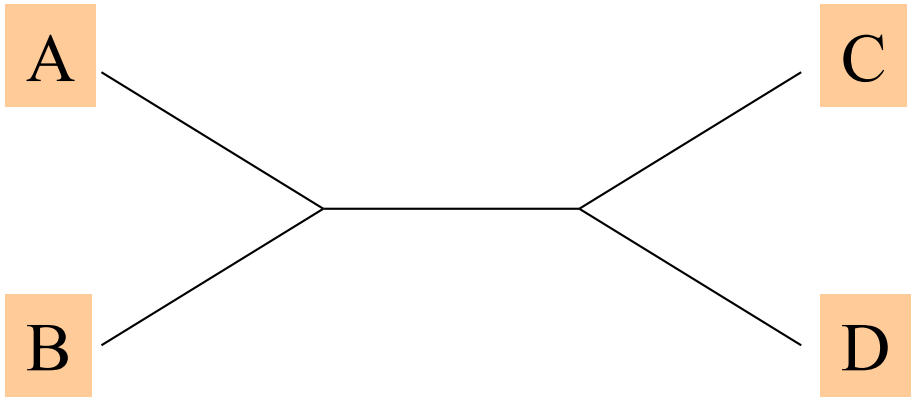
Additive-Distance Trees

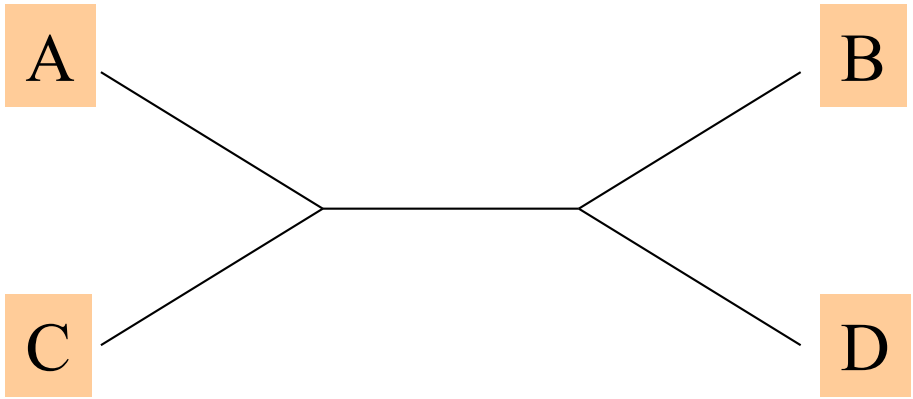
Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

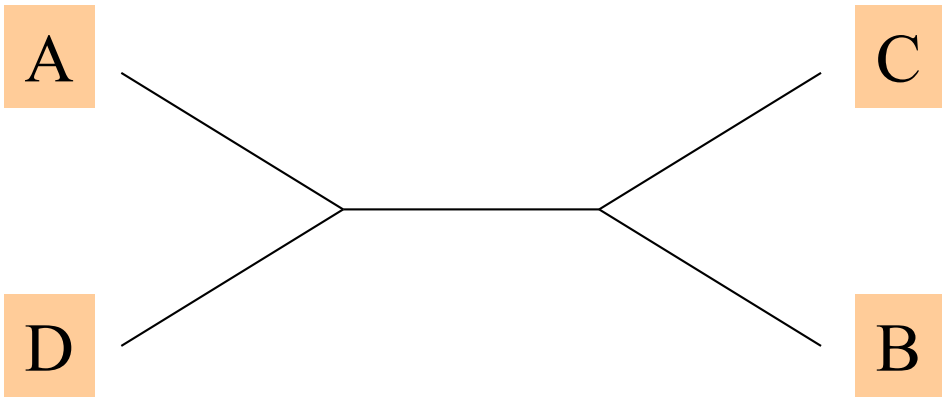
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

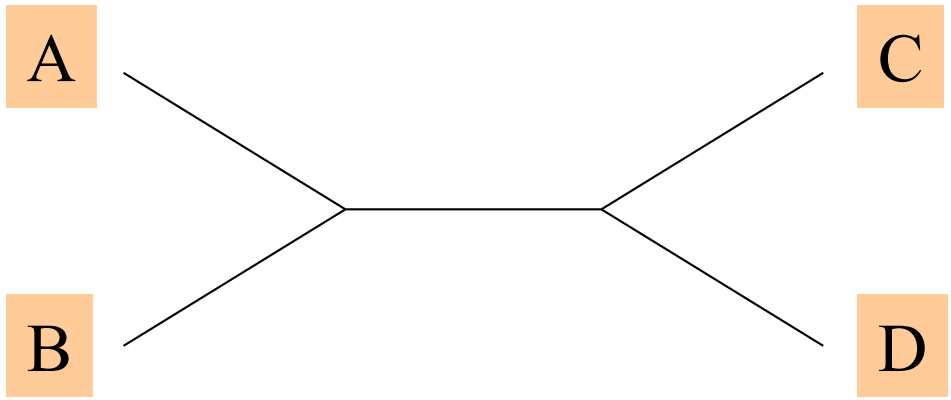






Four-Point Condition

- If the true tree is as shown below, then
  1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and
  2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

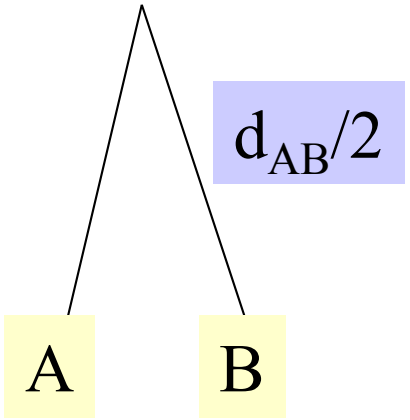


Unweighted pair-group method with arithmetic means (UPGMA)

	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
- Now apply UPGMA

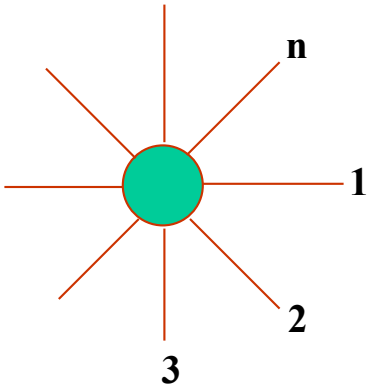
$$D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \left(\sum_{k=1}^{n} D_{kO} \Big/ n \right)$$

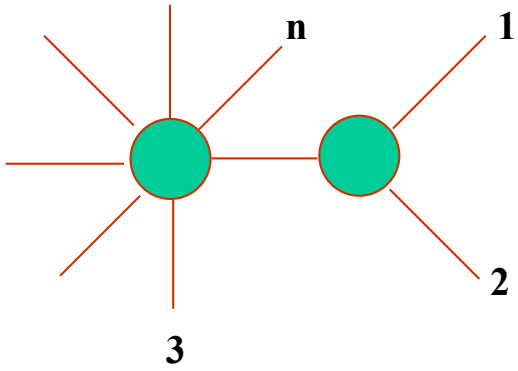
Saitou & Nei: Neighbor-Joining Method

- 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.

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

Neighbor-Joining

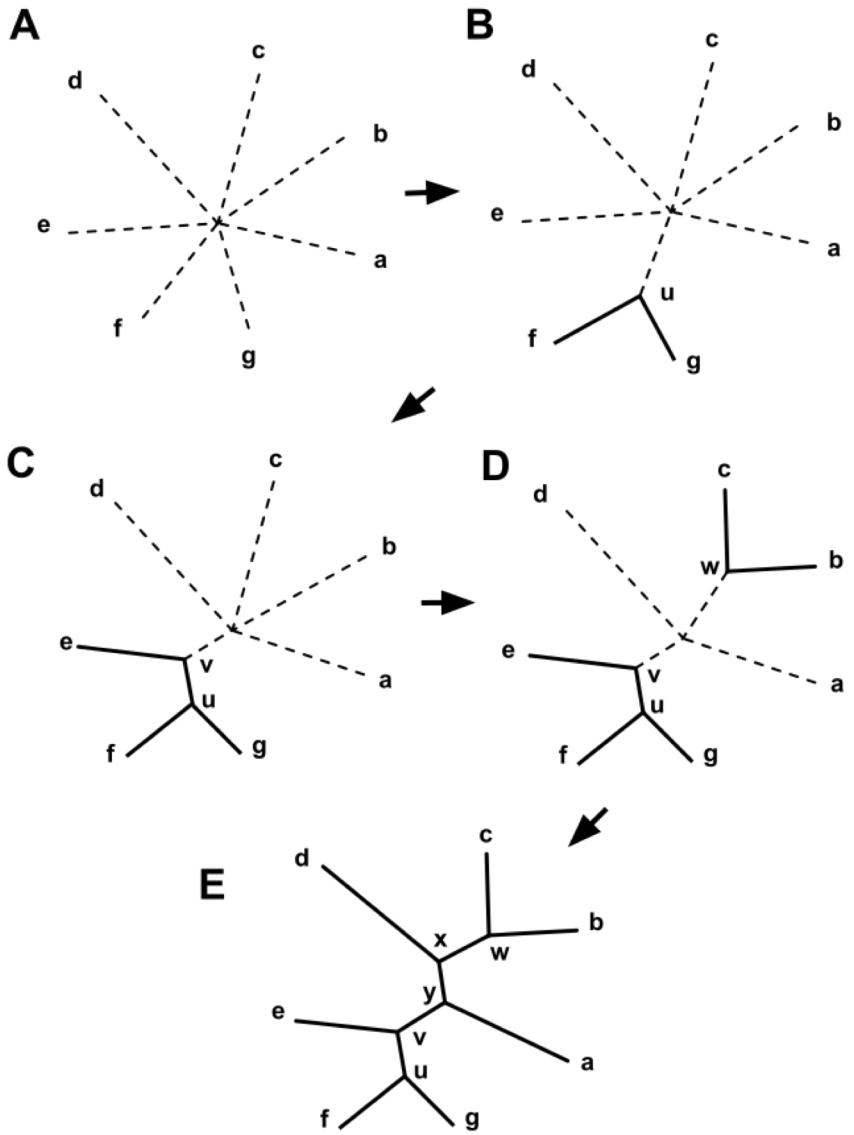




$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

http://en.wikipedia.org/wiki/Neighbor_joining

Neighbor-joining method



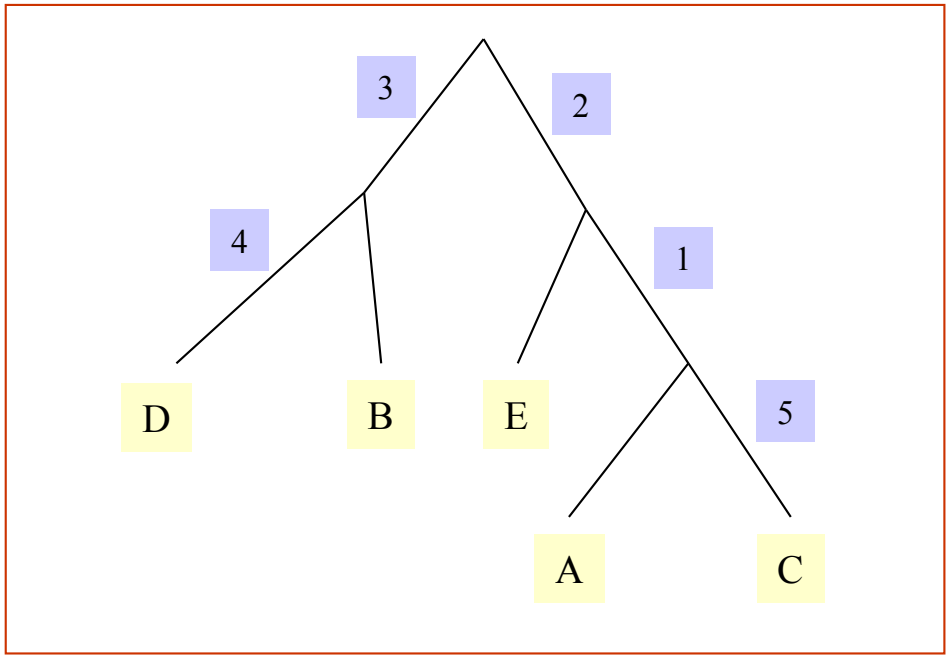
Constructing Evolutionary/Phylogenetic Trees

- 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 ⇔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

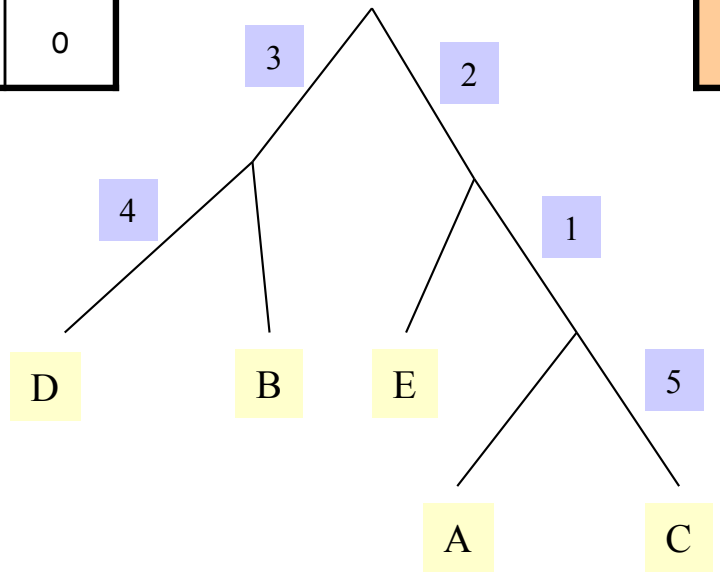


Example

- 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

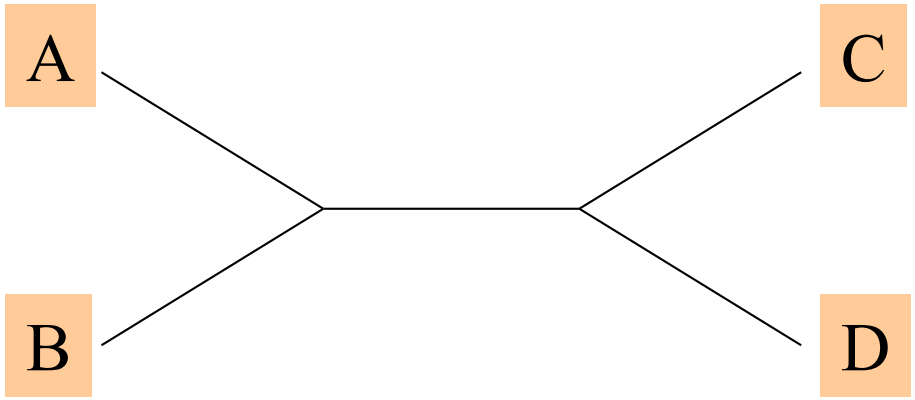
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

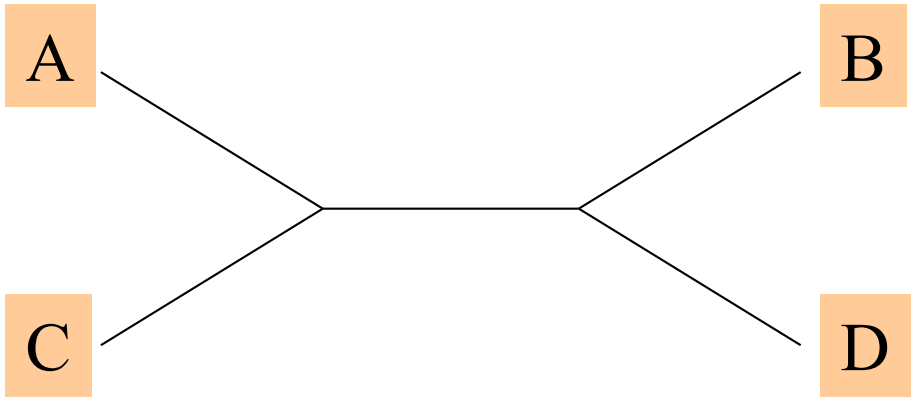
	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

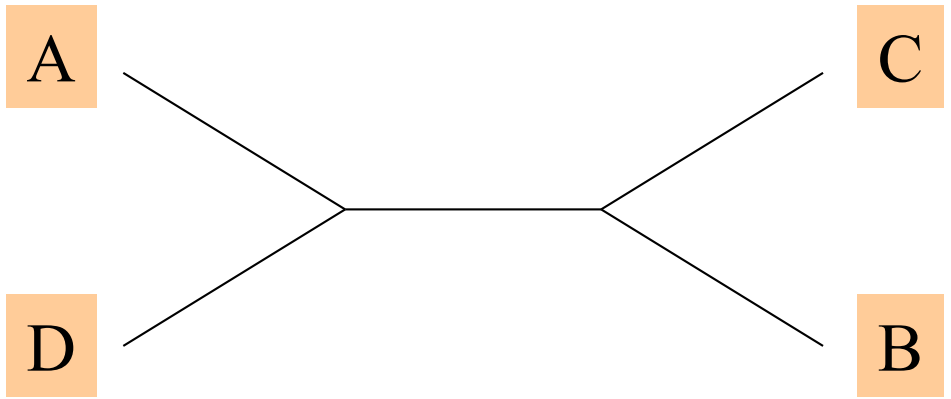
Maximum Parsimony Method: Example

	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







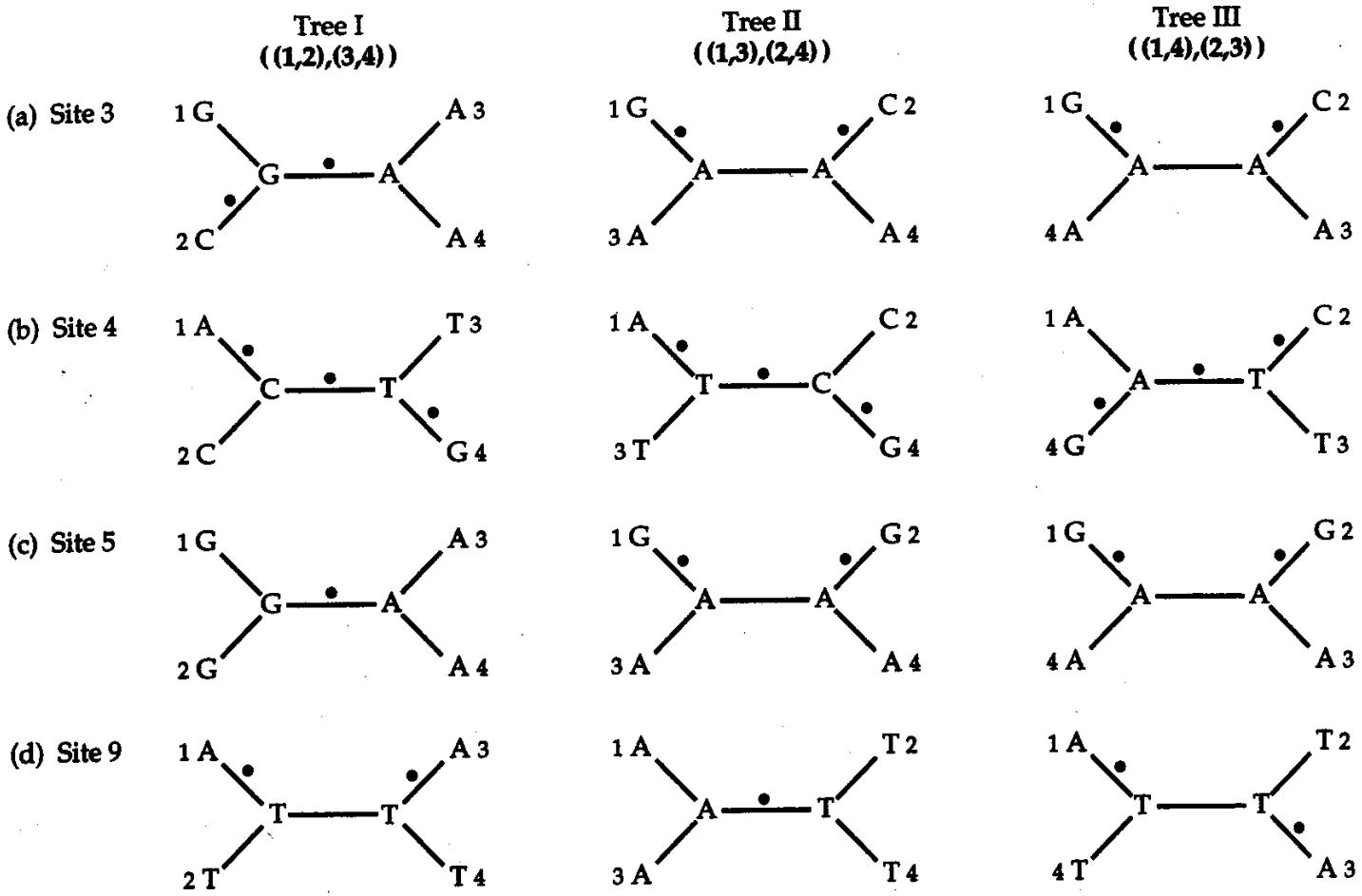


FIGURE 5.14 Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number 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

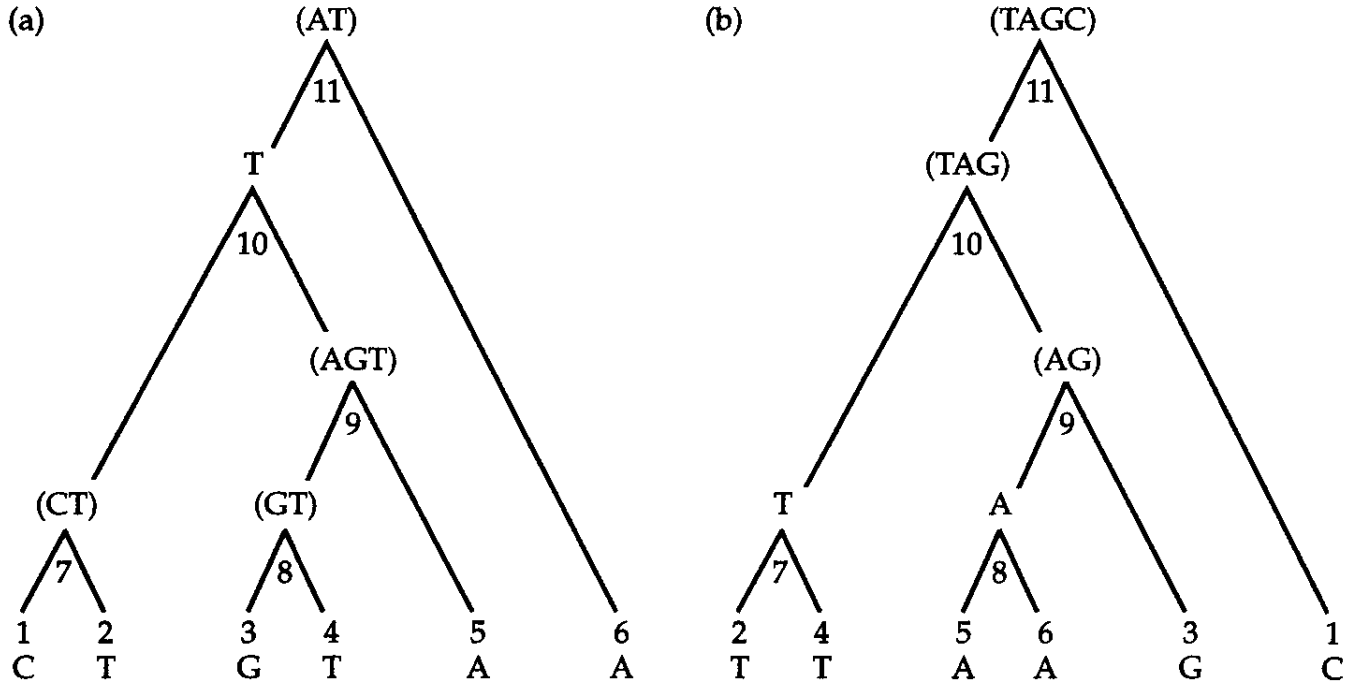


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal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

Searching for the Maximum Parsimony Tree: Exhaustive Search

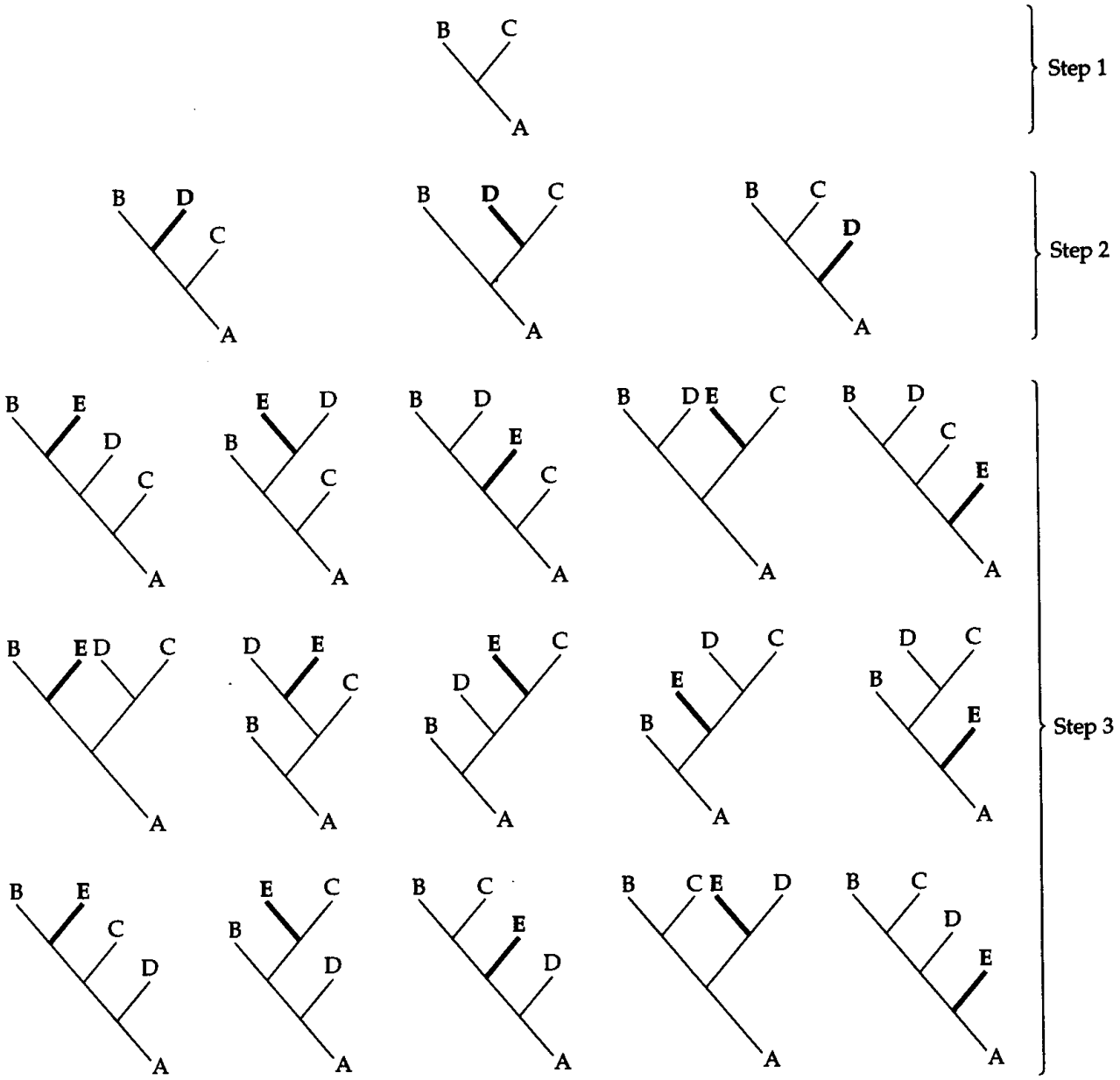
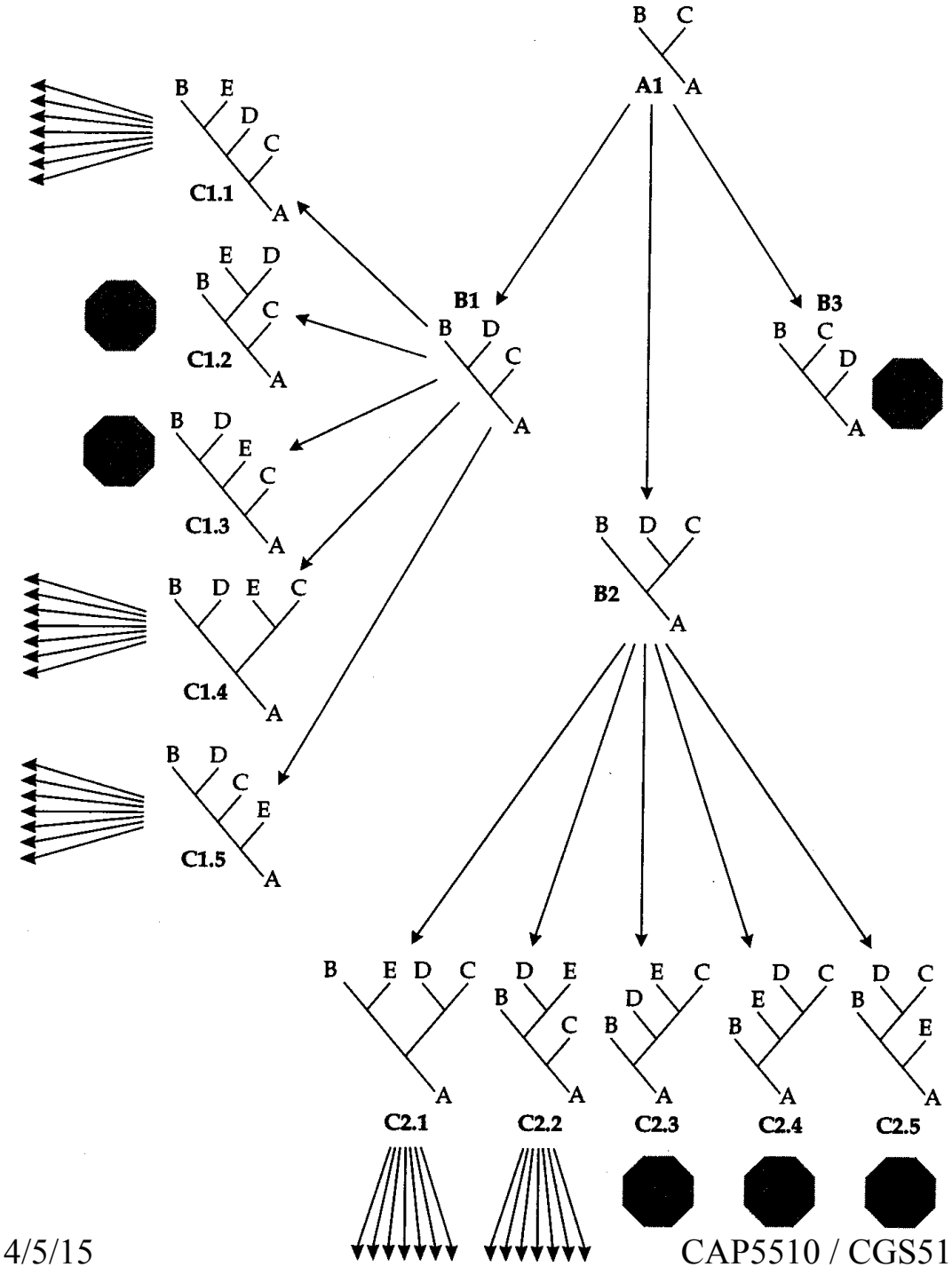


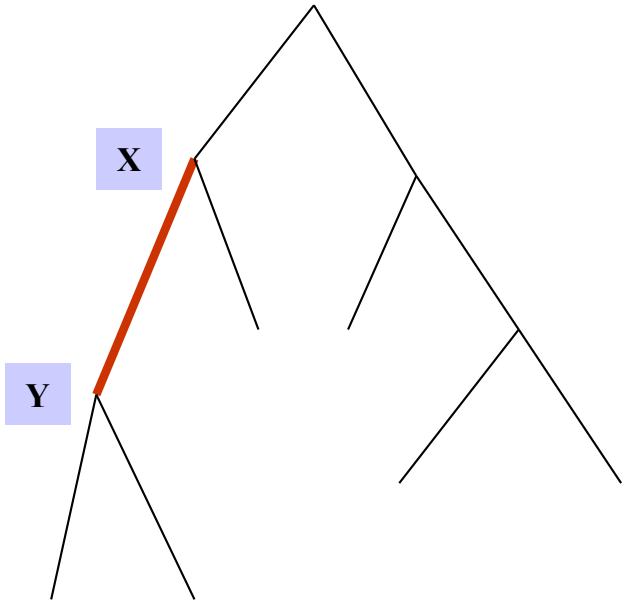
FIGURE 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four 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. Modifed 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

(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

(b)

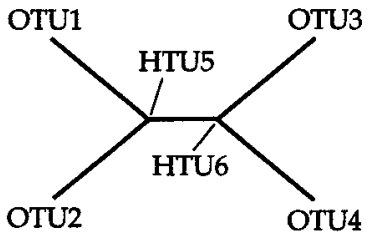
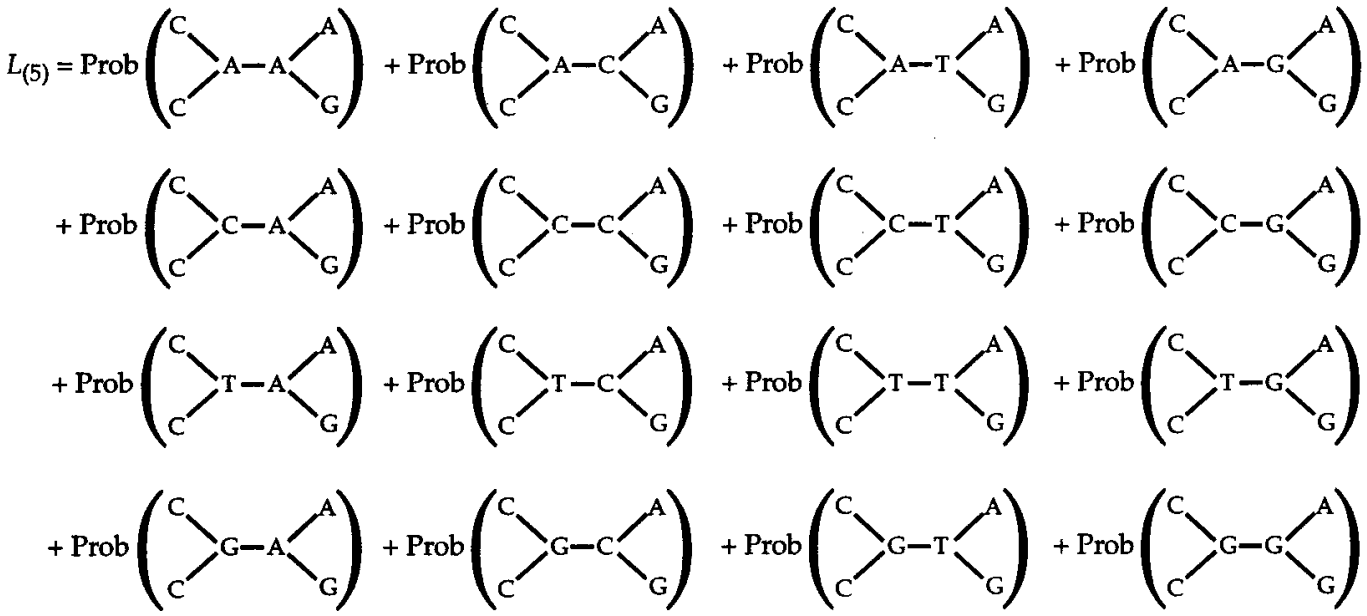


FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length *n*. (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums 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).

(c)



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

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

Computing Maximum Likelihood Tree