Class 1:

EVE 161:
Microbial Phylogenomics

Class #4:
Phylogeny

UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

Where we are going and where we have been

- <u>Previous Class</u>:
  - 3. rRNA
- <u>Current Class:</u>
  - 4. Phylogeny
- <u>Next Class:</u>
  - 5. Tree of Life

Phylogeny

- Eisen Make Up Office Hours Tomorrow
- Storer 5331 10:30-11:30

- Cassie Ettinger Office Hours
- Today Storer 5331 10:30-11:30

Phylogeny

Phylogeny

First Phylogenetic Tree?

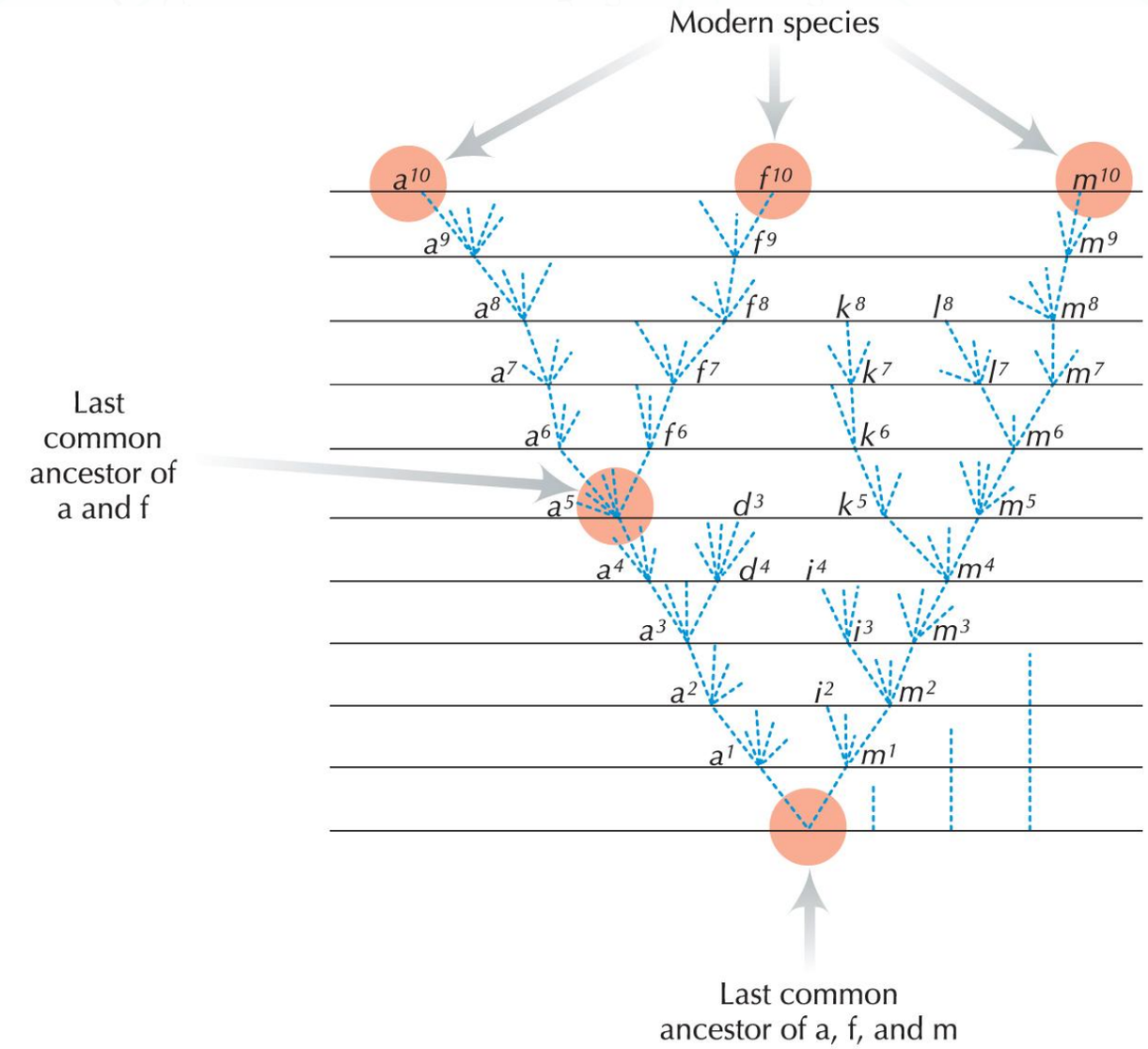
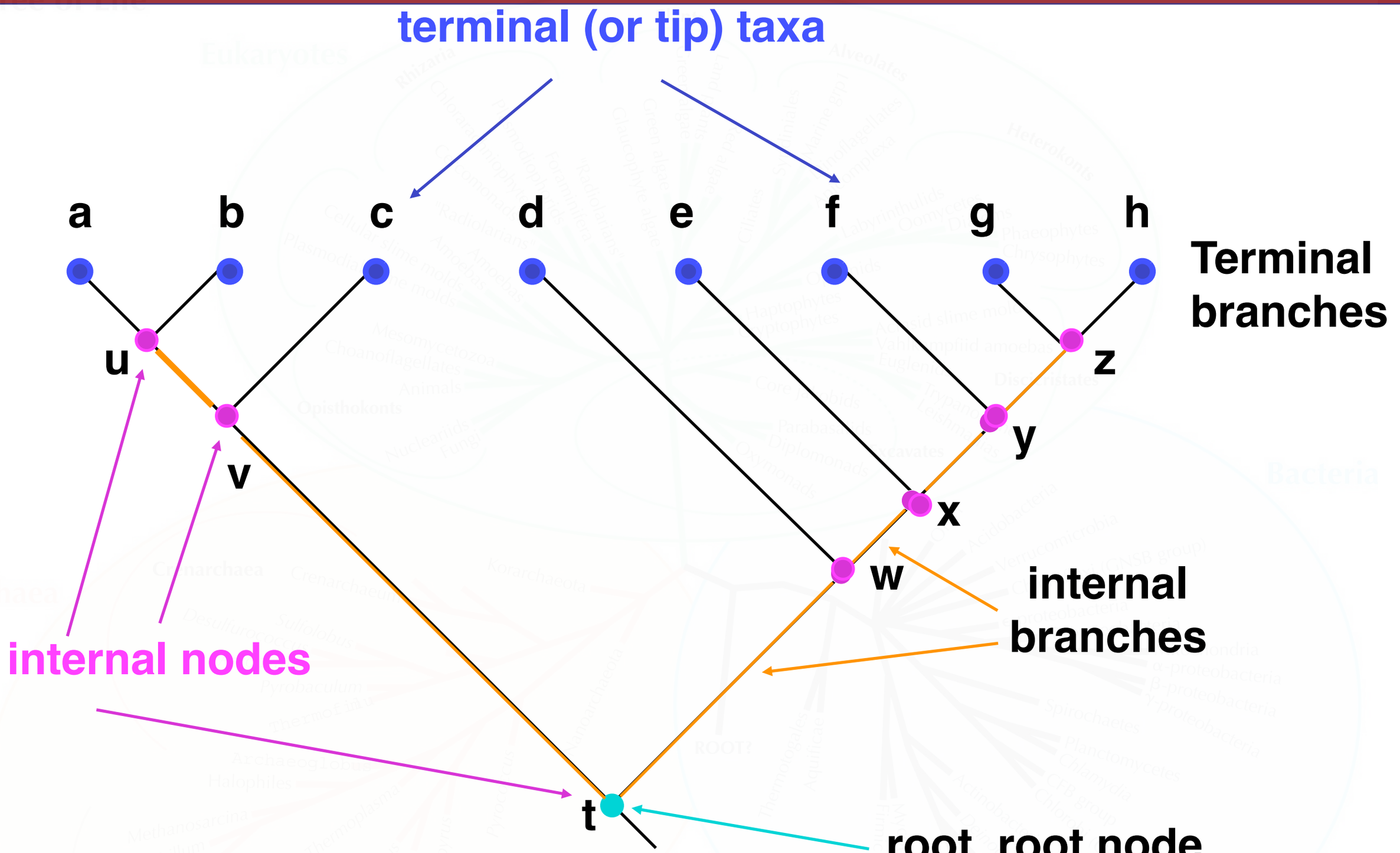


FIGURE 5.1. Vertical inheritance as represented in the only figure in *On the Origin of Species*. The evolution of species over time is represented by the branching of the tree with a single species giving rise to one or more descendent species. The descendent species inherit traits from the parental species in the form of "vertical" inheritance. Modern species are represented by a^{10}, f^{10}, and m^{10} with all other lineages having become extinct. a^{10} and f^{10} can trace their ancestry to a common ancestral species a^5. In turn, a^{10} and f^{10} can jointly trace their common ancestry with m^{10} to the species at the bottom of the figure.

5.1, reprinted from Darwin C., *On the Origin of Species*

Parts of a phylogenetic tree



Internal nodes represent hypothetical ancestral taxa

Branch Rotation

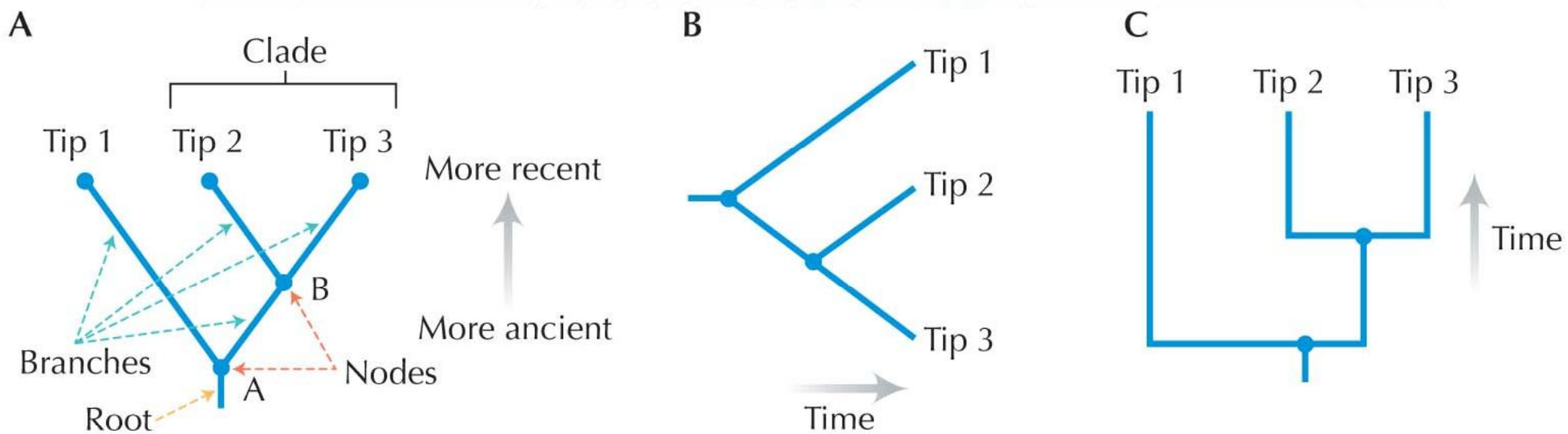


FIGURE 5.2. Model phylogenetic trees showing tips, nodes, and branches. (*A*) All the key elements of a phylogenetic tree. This tree shows the evolutionary history of three operational taxonomic units (OTUs). OTUs can represent species, individuals, genomes, genes, or other entities having an evolutionary history. (*Thick blue lines*) The branches represent the evolution of the OTUs over time. In this tree, evolutionary time is shown progressing from bottom to top, and thus this is known as a vertical tree. When considering evolutionary time from the past to the present, nodes (*blue circles*) represent the points at which one lineage separated into two. When considering evolutionary time from the present to the past, nodes represent the common ancestor of the organisms above the node. In this case, Tip 2 and Tip 3 share a common ancestor at node B. All descendants from node B, including Tip 2 and Tip 3, can be considered a clade or monophyletic group (see Fig. 5.3 for more detail on clades). The separation of taxa on the horizontal axis and the angles of the branches have no real meaning; it is done to be visually pleasing. (*B*) The tree in *A* has been rotated 90°. Such horizontal trees contain the same information as vertical trees. In this case, evolutionary time progresses from left to right, and the separation on the *y*-axis has no meaning. (*C*) A T-branching tree. It too has the same information as the trees in *A* and *B,* but the branches are drawn with a T-shaped junction instead of a V-shaped junction. In each of these trees, the "root" of the tree is the branch leading up to the common ancestor of all taxa shown in the tree.

Types of Trees

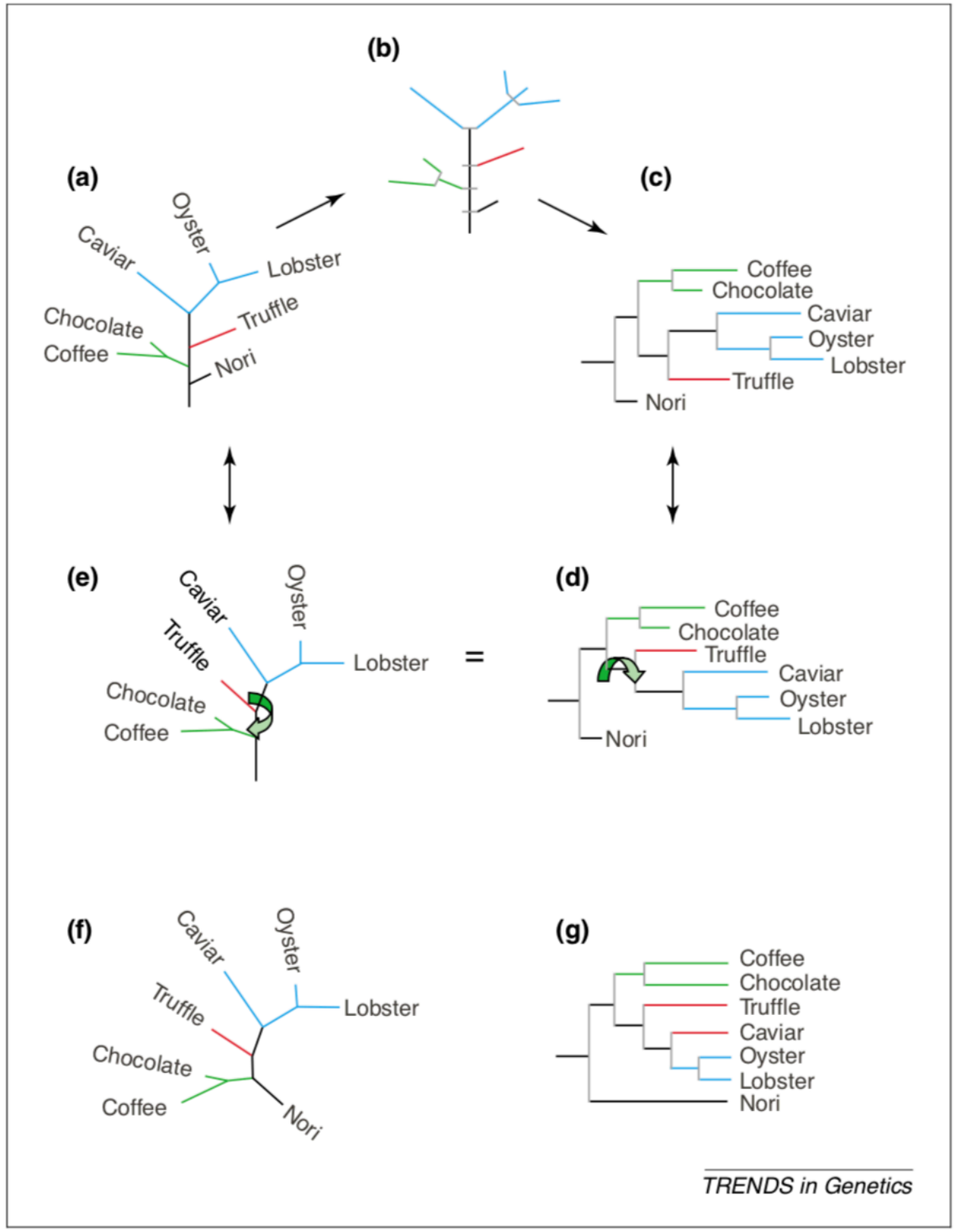


Fig. 2. Phylogenetic tree styles. All these trees have identical branching patterns. The only differences are (f), which is unrooted. (g) is a cladogram, so the branch lengths are right justified and not drawn to scale (i.e. they are not proportional to estimated evolutionary difference).

Phylogram (Tree with Lengths)

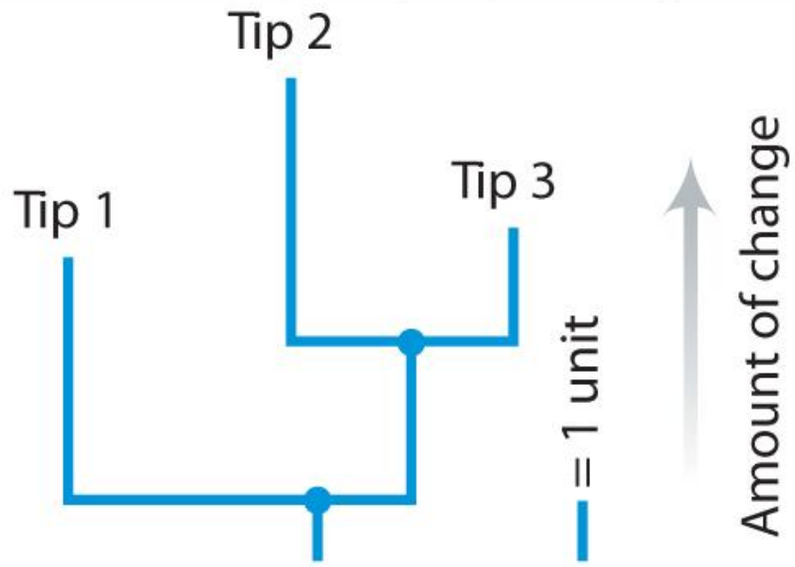


FIGURE 5.5. A phylogram (shown) is a phylogenetic tree that indicates the amount of evolution in addition to the branching order. The amount of evolution is represented by the branch lengths along the time axis (in this example, the vertical axis). In this tree, Tip 2 and Tip 3 share a common ancestor to the exclusion of Tip 1. However, during the time since they diverged from their common ancestor, Tip 2 has undergone more change. If Tip 2 and Tip 3 are modern organisms, this means that the rate of evolution in the lineage leading up to Tip 2 was greater than that leading up to Tip 3. Differences in rates of evolution are common and can be due to many factors such as different mutation rates, different population sizes, and different selective forces. Regardless of the cause, it is frequently very useful to incorporate such differences into evolutionary trees. In phylograms, a scale bar defines how much change is represented per unit length.

Phylogenetic Groups

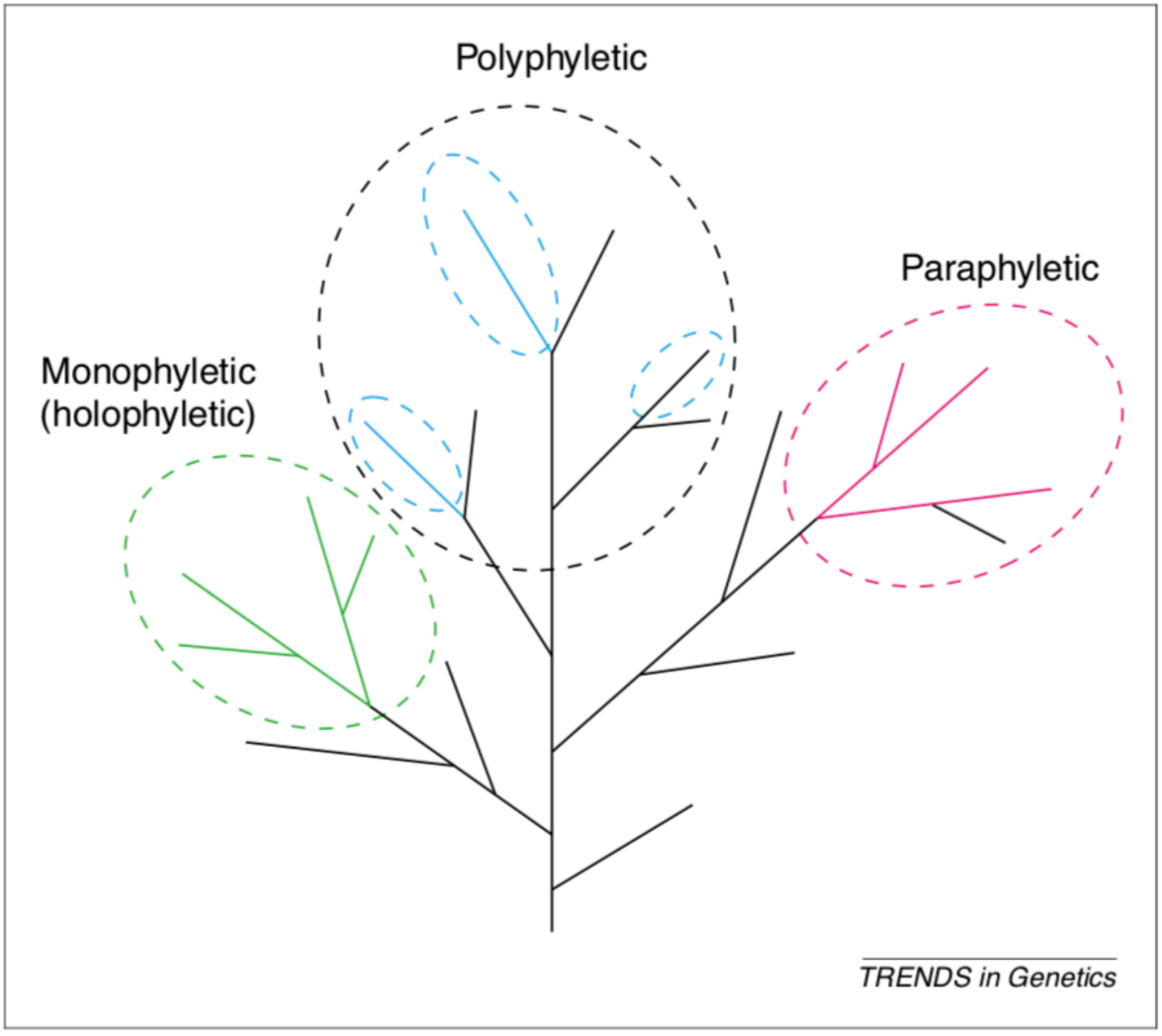


Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and 'polyphyletic'.

Phylogenetic Groups

Monophyletic **Paraphyletic** **Polyphyletic**

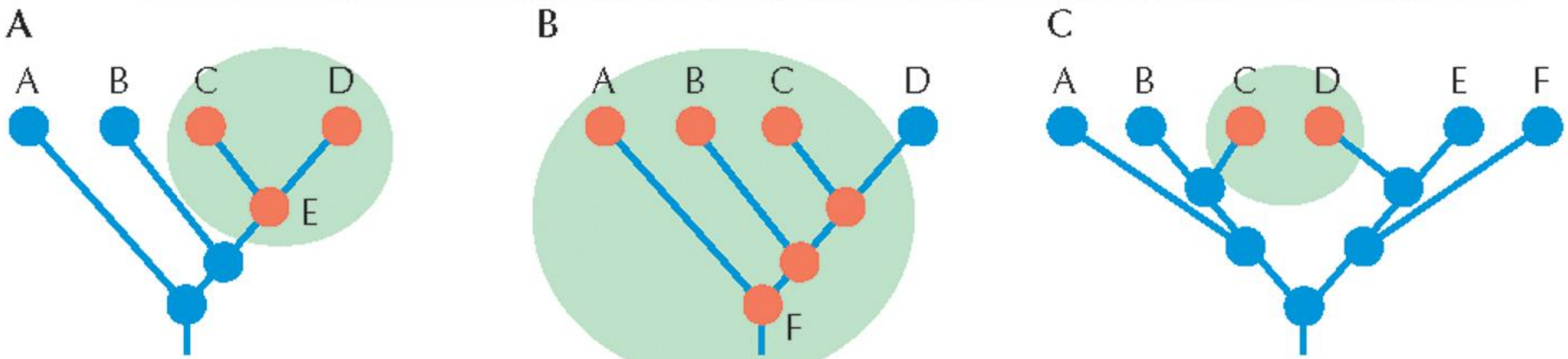


FIGURE 5.3. Different types of phylogenetic groups. In each panel, the phylogenetic group is indicated by a *green shaded circle*. (*A*) Monophyletic group. All species (C and D) in the group share a common ancestor (E) not shared by any of the other species. (*B*) Paraphyletic group. All species in the group share a common ancestor (F), but some species (D) have been excluded from the group. (*C*) Polyphyletic group. A grouping of lineages each more closely related to other species not in the group than they are to each other.

Branch rotation

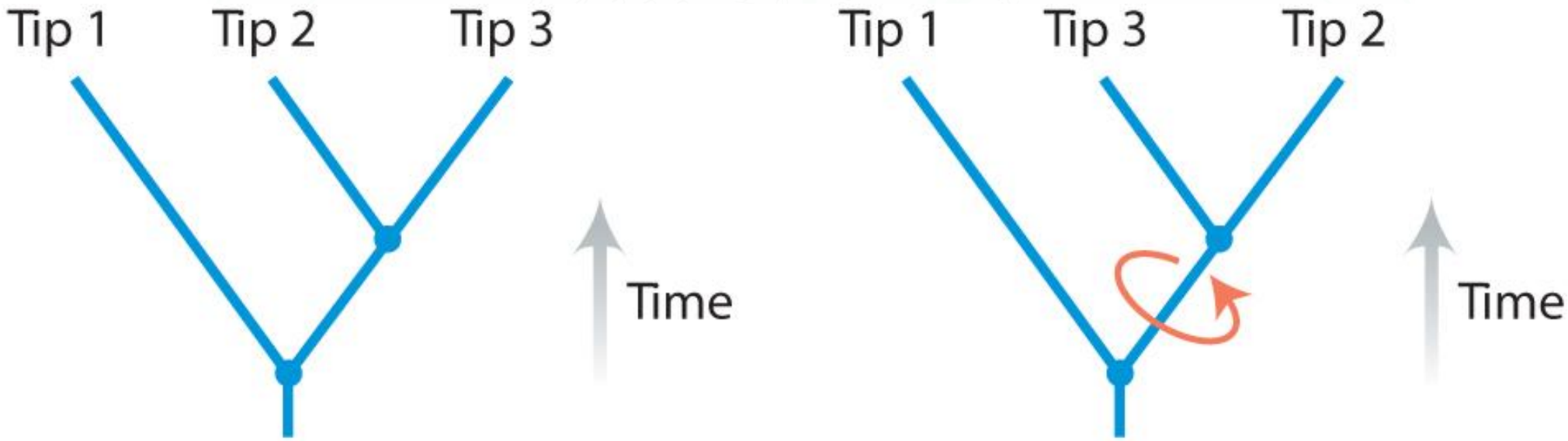


FIGURE 5.6. Branch rotation does not change the information in a phylogenetic tree. The two trees are the same except the branch leading up to the ancestor of Tip 2 and Tip 3 has been rotated such that Tip 3 is on the left and Tip 2 is on the right. The tree drawings on *left* and *right* are simply different forms of the same tree. It is useful to consider a phylogenetic tree like a mobile: The branches are free to rotate, but the branching patterns never change.

Characters

- A heritable feature of an organism is known as a **character (**also **character trait** or **trait**).

- The form that a character takes is known as its **state** (also known as **character state**).
  - Note: Presence/absence can be a **state**

- Example:
  - Character = heart
  - Character state = present/absent
  - Character state = # of chambers

Characters ancestry is critical to understand

- Characters that are inherited from a common ancestor are **homologous.**
- Species change over time
  - Known (generally) as **divergence,** or **divergent evolution**.
  - Species change over time due to the combined processes of mutation, recombination, drift, selection, etc

Homology vs. Orthology vs. Paralogy

Data matrices

TABLE 27.1. Character matrix showing character states for nine traits (A–I) in eight OTUs (1–8)

	Character Traits								
OTUs	A	B	C	D	E	F	G	H	I
1	α	20	1	α	+	+	2	i	–
2	β	10	1	α	+	–	2	ii	–
3	β	14	1	—	+	–	1	iii	–
4	α	23	1	—	–	–	1	ii	–
5	β	13	1	—	–	+	1	ii	–
6	α	23	1	α	–	+	2	i	–
7	β	14	1	β	–	+	2	ii	–
8	α	20	1	α	+	+	2	i	–

Alignment

		Alignment Position								
		1	2	3	4	5	6	7	8	9
OTUs	A	A	A	C	T	A	T	G	G	C
	B	A	C	T	A	T	G	G	C	A
	C	A	T	A	C	T	A	T	G	G
	D	A	T	A	C	T	A	T	G	G
	E	A	A	T	G	G	C	C	C	A
	F	A	A	C	T	T	T	G	G	C
	G	A	C	T	A	T	G	G	C	A
	H	A	A	C	T	A	T	G	G	C

Alignment with Gaps

		Alignment Position												
		1	2	3	4	5	6	7	8	9	10	11	12	13
OTUS	A	A	–	A	C	T	A	T	G	G	C	–	–	–
	B	A	–	–	C	T	A	T	G	G	C	A	–	–
	C	A	T	A	C	T	A	T	G	G	–	–	–	–
	D	A	T	A	C	T	A	T	G	G	–	–	–	–
	E	A	–	–	–	–	A	T	G	G	C	C	C	A
	F	A	–	A	C	T	T	T	G	G	C	–	–	–
	G	A	–	–	C	T	A	T	G	G	C	A	–	–
	H	A	–	A	C	T	A	T	G	G	C	–	–	–

Sequence Alignment

FIGURE 5.16. Example of a multiple sequence alignment, showing alignment of a portion of the RecA proteins from different bacterial species. Amino acids conserved across most or all of the species are highlighted in *red*. *Letters* are the abbreviations of different amino acids (see Fig. 2.23).

Alignment Scoring (BLAST)

Match 1

Query	AATTAATTAACC
Database	AATTAATTAACC
Score	111111111111 = Total 12

Match 2

Query	AATTAATTAACC
Database	AACCAATTAACC
Score	110011111111 = Total 10

TABLE 27.3. DNA substitution matrix and some word matches

Query Sequence	Database Sequence A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Blosum Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
		2	0	-2	0	-1	0	0	0	1	0	0	0	1	0	1	-1	1	1	-1	S
C	9		2	-1	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
S	-1	4		2	-2	-1	-1	0	0	-1	-1	-1	1	1	0	-1	0	0	2	1	P
T	-1	1	5		2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	A
P	-3	-1	-1	7		2	0	-1	-2	0	1	1	0	0	-1	0	-1	1	2	4	G
A	0	1	0	-1	4		3	-1	-1	0	0	1	-1	0	-1	0	-1	0	0	0	N
G	-3	0	-2	-2	0	6		2	-1	-1	-1	0	-1	0	0	0	0	2	1	3	D
N	-3	1	0	-2	-2	0	6		1	0	0	2	2	1	-1	0	0	2	2	4	E
D	-3	0	-1	-1	-2	-1	1	6		0	-2	0	1	1	-1	0	0	1	3	3	Q
E	-4	0	-1	-1	-1	-2	0	2	5		2	-1	0	1	0	-1	0	1	2	2	H
Q	-3	0	-1	-1	-1	-2	0	0	2	5		-1	-1	0	-1	1	0	1	3	-4	R
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8		1	-2	-1	1	1	2	3	1	K
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5		-2	-1	-1	0	1	2	4	M
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5		-1	1	0	0	1	3	I
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5		-1	0	-1	1	2	L
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4		0	1	2	4	V
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4		-1	-2	1	F
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		-1	2	Y
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		-1	W
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Guide Tree for Alignment

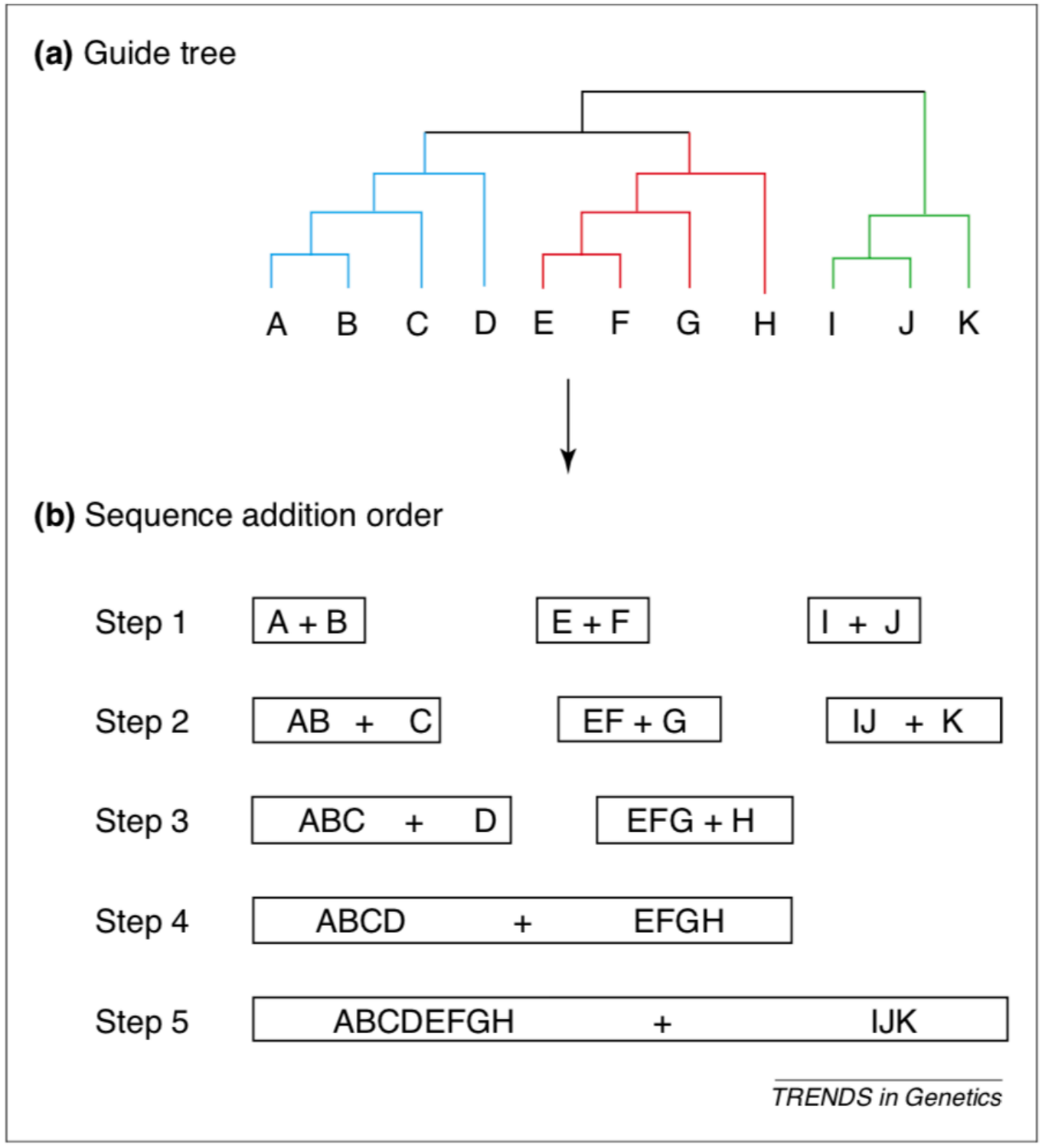


Fig. 4. Steps in progressive sequence alignment. (a) The first step is to calculate the guide tree. (b) This determines the order in which sequences are added to the growing alignment.

Refining Alignment

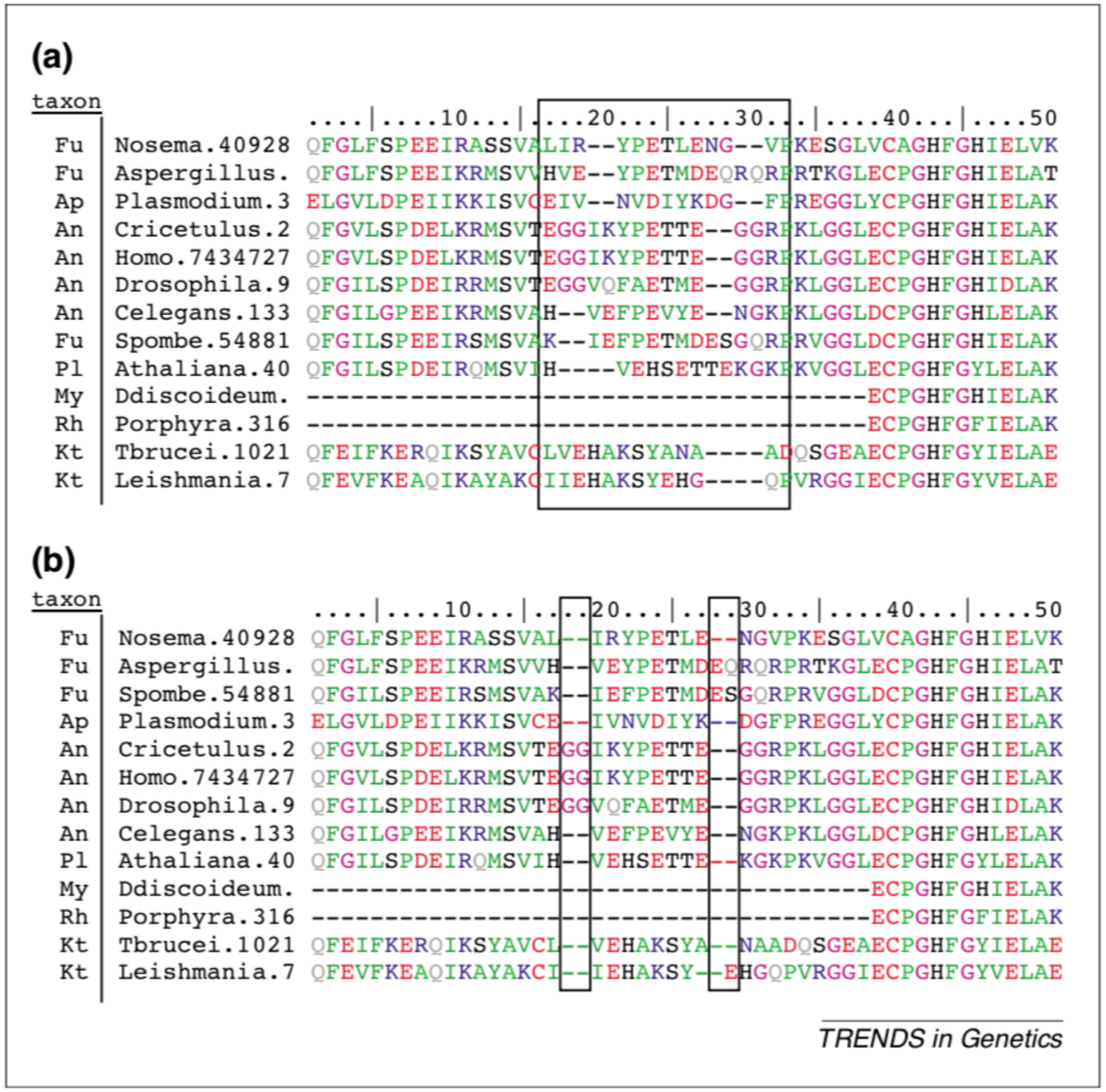
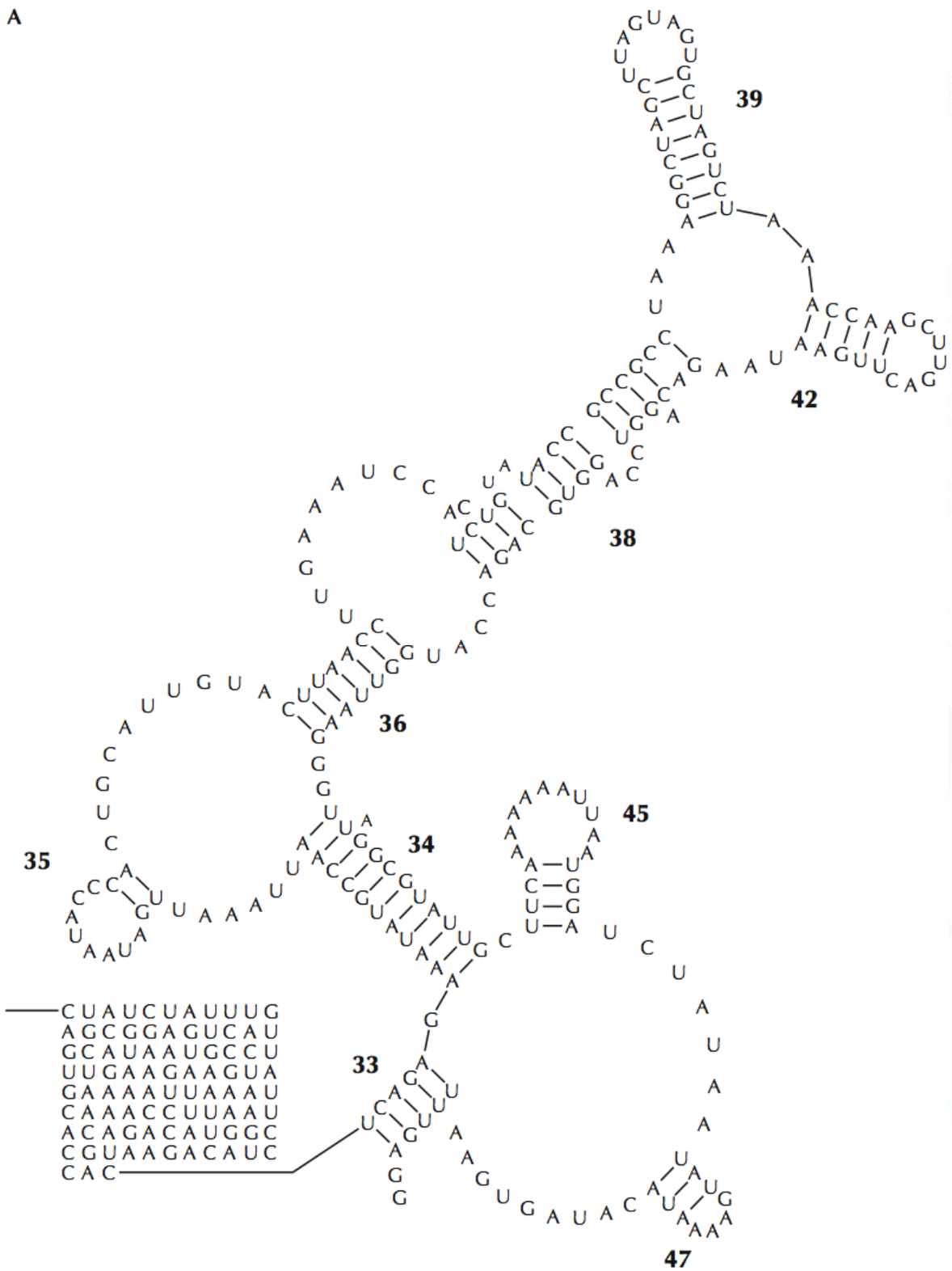
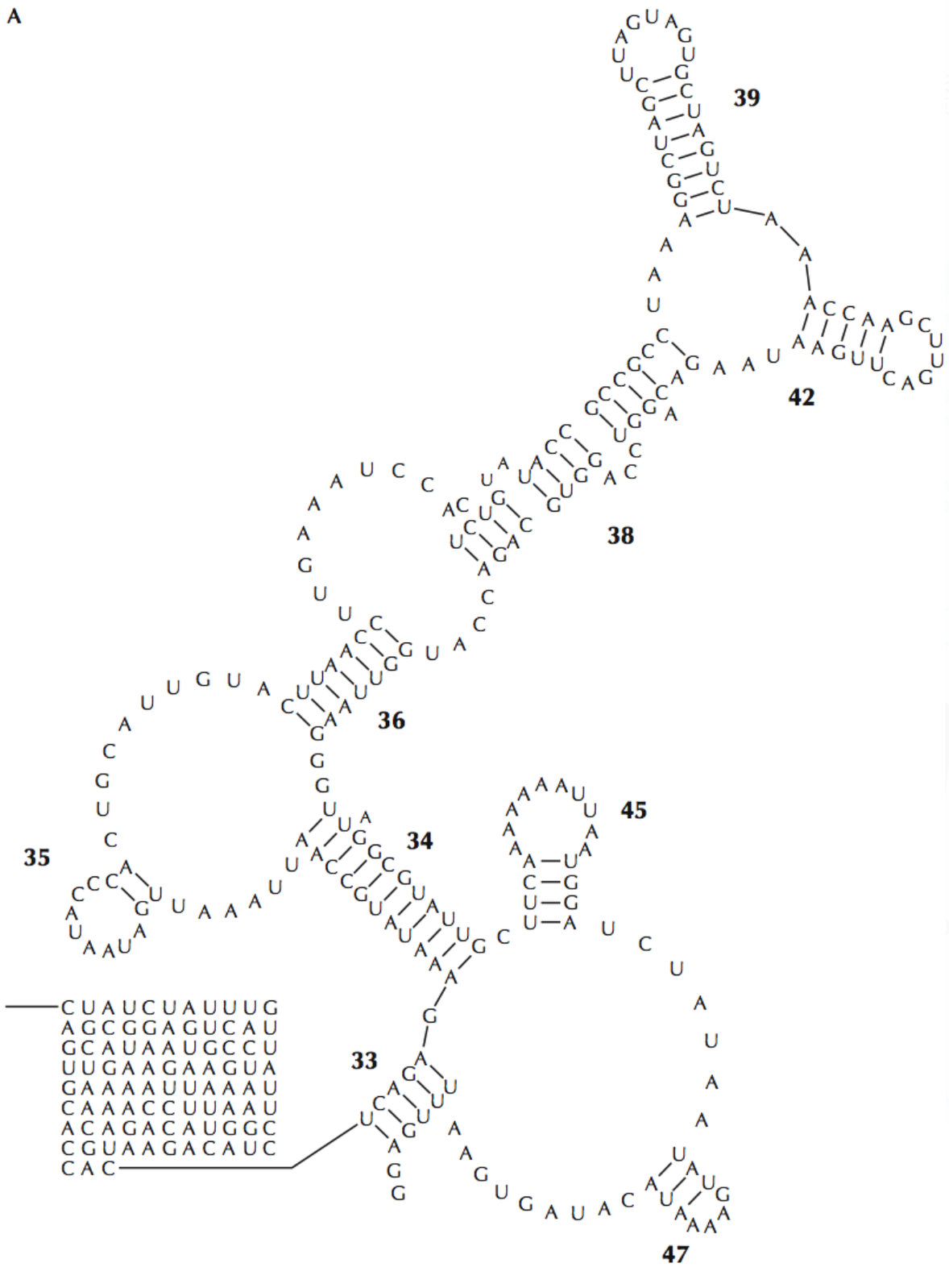


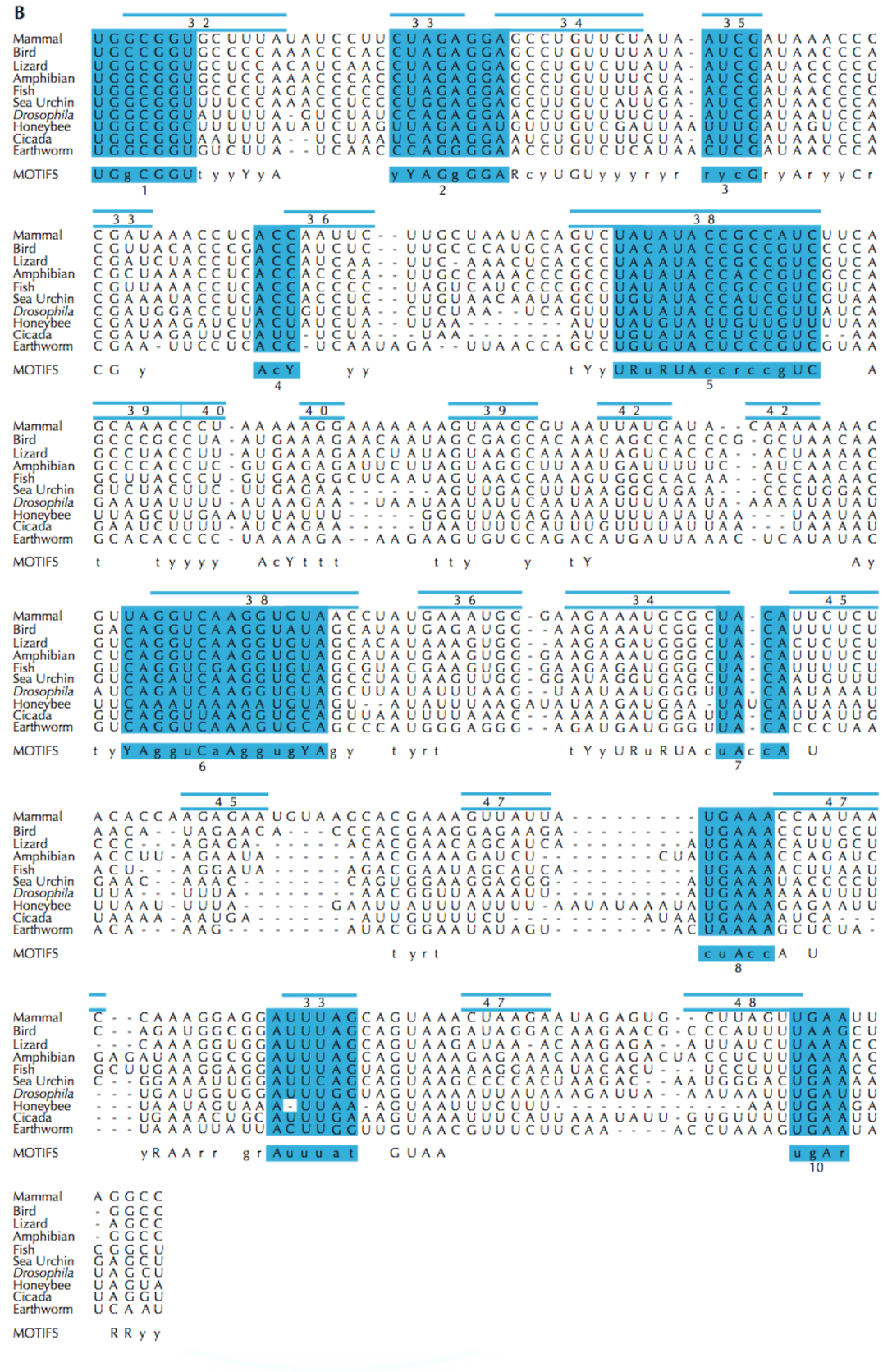
Fig. 5. Refining an alignment. (a) The raw output from a ClustalX alignment of rpb1 sequences, which predicts six insertion/deletion events (boxed), some of which are blatantly inconsistent with known taxonomy. (b) The refined alignment makes much better evolutionary sense, because it shows only two insertion events in well-defined taxonomic groups (animals and higher fungi). Taxon labels are Fu (fungi), An (animals), Pl (green plant), Ap (apicomplexan), Rh (rhodophyte), My (mycetozoan), Kt (kinetoplastids). In (b), the sequence from *Saccharomyces pombe* has been placed adjacent to the other fungi to make these relationships more obvious.

Structure Guided Alignment

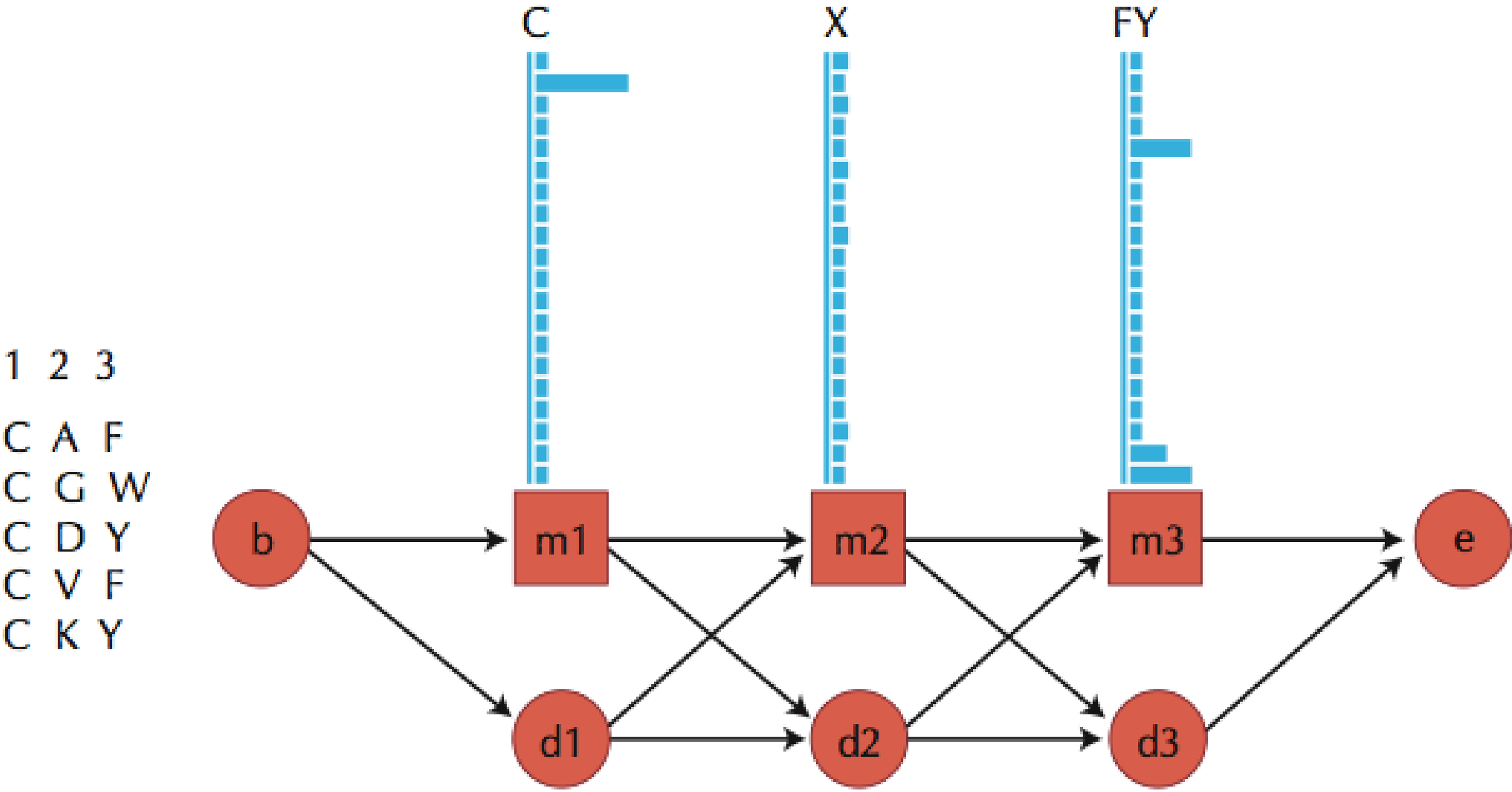


Structure Guided Alignment





HMM Alignment



Tree reconstruction methods

TABLE 27.4. Molecular phylogenetic methods

Method	Description
Parsimony	Possible trees are compared and each is given a score that is a reflection of the minimum number of character state changes (e.g., amino acid substitutions) that would be required over evolutionary time to fit the sequences into that tree. The optimal tree is considered to be the one requiring the fewest changes (the most parsimonious tree).
Distance	The optimal tree is generated by first calculating the estimated evolutionary distance between all pairs of sequences. These distances are then used to generate a tree in which the branch patterns and lengths best represent the distance matrix.
Maximum likelihood	This method is similar to parsimony methods in that possible trees are compared and given a score. The score is based on how likely the given sequences are to have evolved in a particular tree given a model of amino acid or nucleotide substitution probabilities. The optimal tree is considered to be the one that has the highest probability.
Bayesian	A variant of maximum likelihood in which the likelihood of a tree itself is calculated.

Based on Table 3 in Eisen J.A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**:163–167.

Distance Phylogenetics

- Calculate distances between taxa
- Use an algorithm to infer a tree from those distances
- Based on principle that divergence occurs after organisms separate

Parimony Phylogenetics

- Based on Principle of Parsimony / Occam's Razor
- Possible trees are given a score of the fewest number of sequence changes required to fit data into tree
- Score for all possible trees

Likelihood Phylogenetics

- Based on Bayes' Theorem
- Trying to calculate the Probability of the Data, Given a Hypothesis
- For each tree, calculate probability that the data (e.g., sequence alignment) could come from that tree
- Score for all possible trees

- Prob($H|D$) = Prob(H and D) ÷ Prob(D)
- Prob($H|D$) = **Prob($D|H$)** x Prob(H) ÷ Prob(D)

Bayesian Phylogenetics

- Based on Bayes' Theorem
- Trying to calculate the Probability of the Hypothesis, Given the Data
- For each tree, calculate probability that the tree could come from that data ((e.g., sequence alignment)
- Score for all possible trees

- Prob($H|D$) = Prob(H and D) ÷ Prob(D)
- **Prob($H|D$)** = $\underline{\text{Prob}(D|H)}$ x Prob(H) ÷ Prob(D)

Bayesian

- Based on Bayes' Theorem
- Prob(H|D) = Prob(H and D) ÷ Prob(D)

- **Prob(H|D)** = Prob(D|H) x Prob(H) ÷ Prob(D)

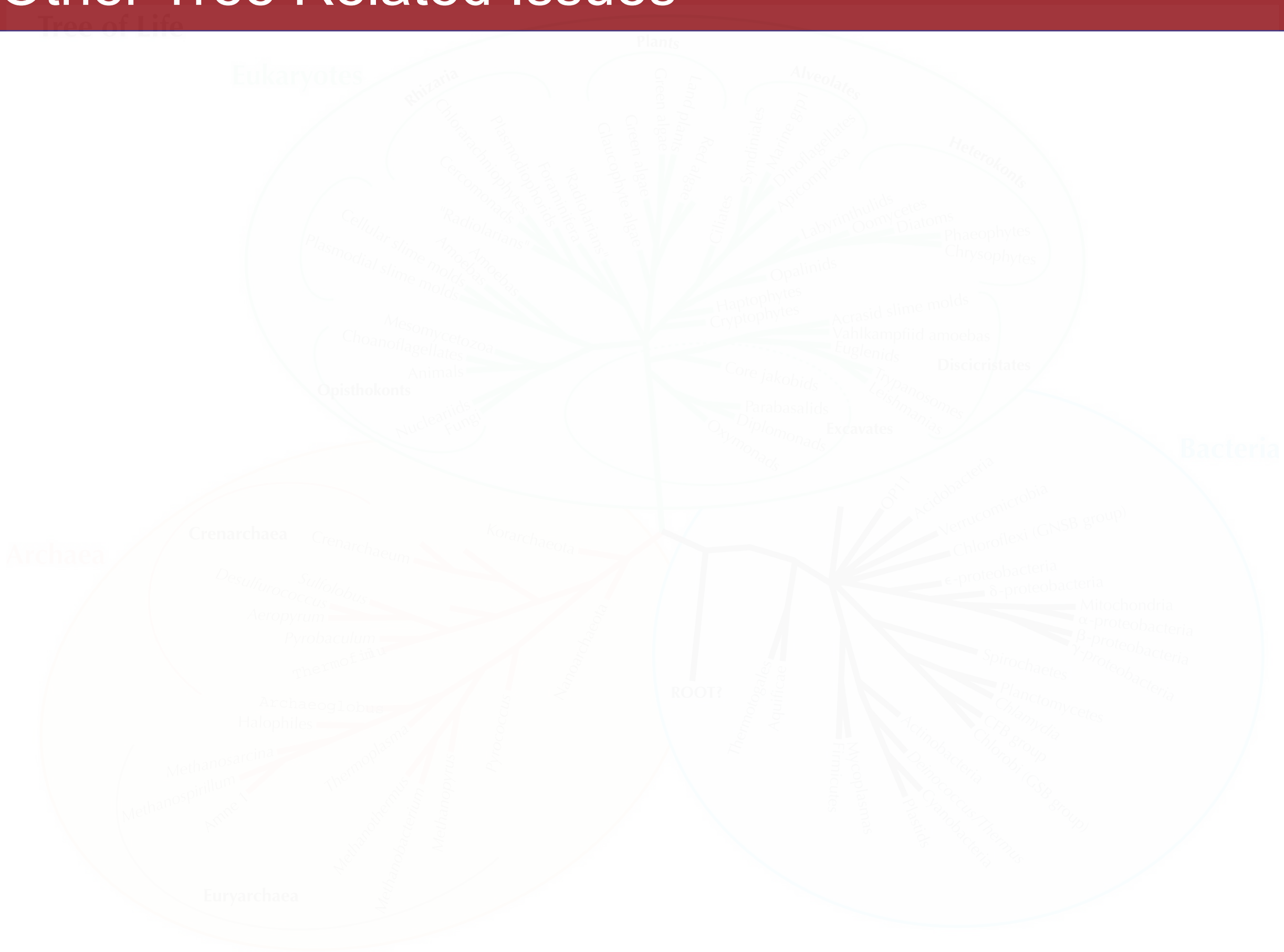
Many possible trees

TABLE 27.5. Number of possible branching patterns versus number of OTUs

Taxa	Rooted Trees[a]	Unrooted Trees[b]
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

[a] $N_r = (2n - 3) \times (2n - 5) \times (2n - 7) \times \cdots \times 3 \times 1 = (2n - 3)!/[2_{n-2} \times (n - 2)!]$.
[b] $N_u = (2n - 3) \times (2n - 5) \times (2n - 7) \times \cdots \times 3 \times 1 = (2n - 5)!/[2_{n-3} \times (n - 3)!]$.

Other Tree Related Issues



Bootstrapping

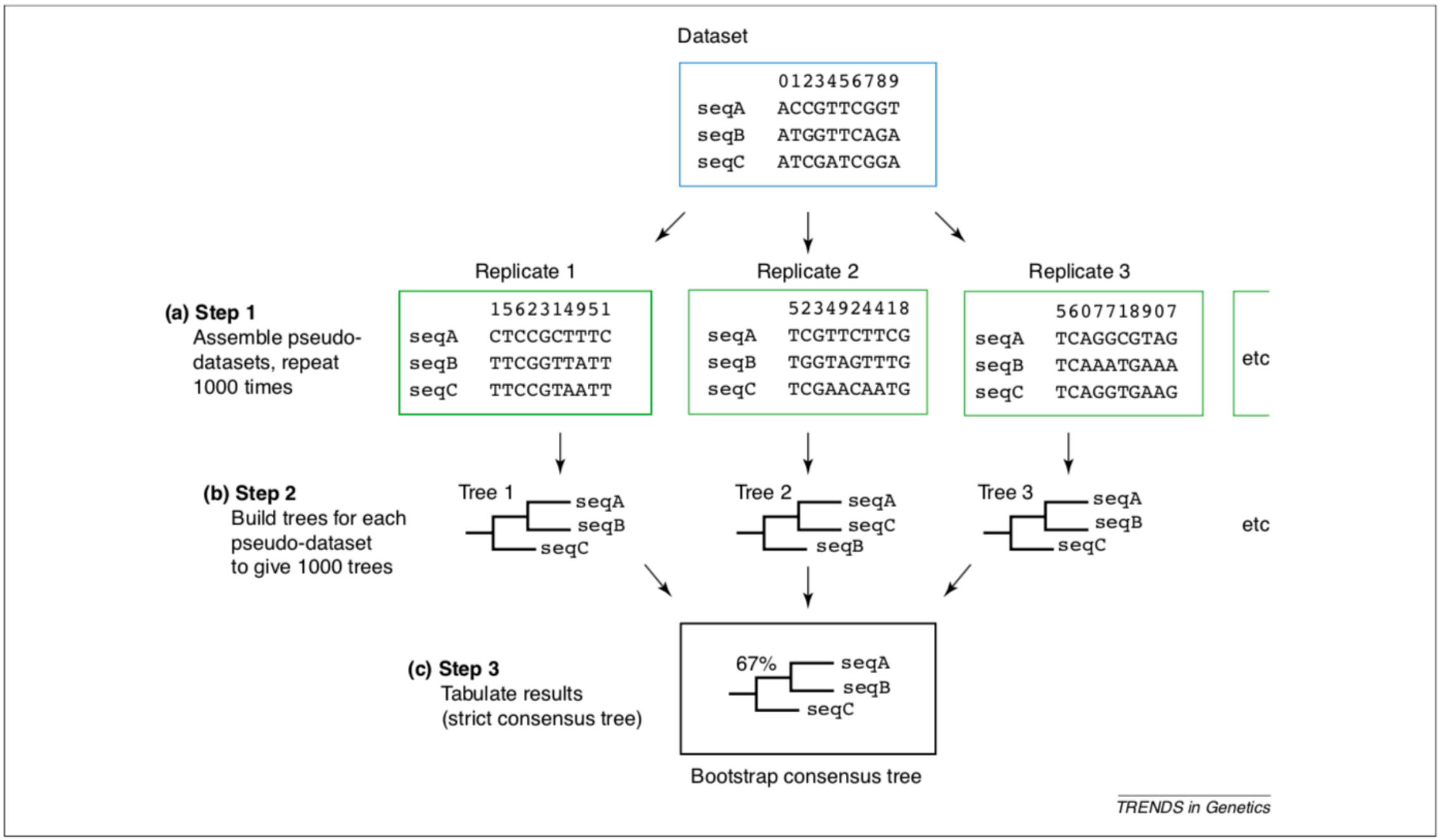


Fig. 6. Bootstrap analysis proceeds in three steps. The dataset is randomly sampled with replacement to create multiple pseudo-datasets of the same size as the original ((a), three are shown in this example). (b) Individual trees are constructed from each of the pseudo-datasets. (c) Each of the pseudo-dataset trees are scored for which nodes (groupings) appear and how often. In this case, a node uniting seqA plus seqB is found in two of the three replicate trees. This gives a bootstrap support for this grouping of 2/3 or 67%.

Bootstrapping

A Bootstrapping

Alignment of sequences

```
Species  1    ATGTTGGATGGTGAT
Species  2    ATGTTGGAAGGAGAA
Species  3    ATGTTAGGAGAAGAA
Species  4    ATGTCAGCAGCCGCC
```

Bootstrapping alignment #1

```
Species  1    ATGTTGGAGGGTAAT
Species  2    ATGTTGGAGGGAAAA
Species  3    ATGTTAGGGGAAAAA
Species  4    ATGTCAGCGGCCCCC
```

Bootstrapping alignment #2

```
Species  1    ATGGGGGATGGTGAT
Species  2    ATGGGGGAAGGAGAA
Species  3    ATGGGAGGAGAAGAA
Species  4    ATGGGAGCAGCCGCC
```

Jacknifing

B Jackknifing

Alignment of sequences

```
Species  1        ATGTTGGATGGTGAT
Species  2        ATGTTGGAAGGAGAA
Species  3        ATGTTAGGAGAAGAA
Species  4        ATGTCAGCAGCCGCC
```

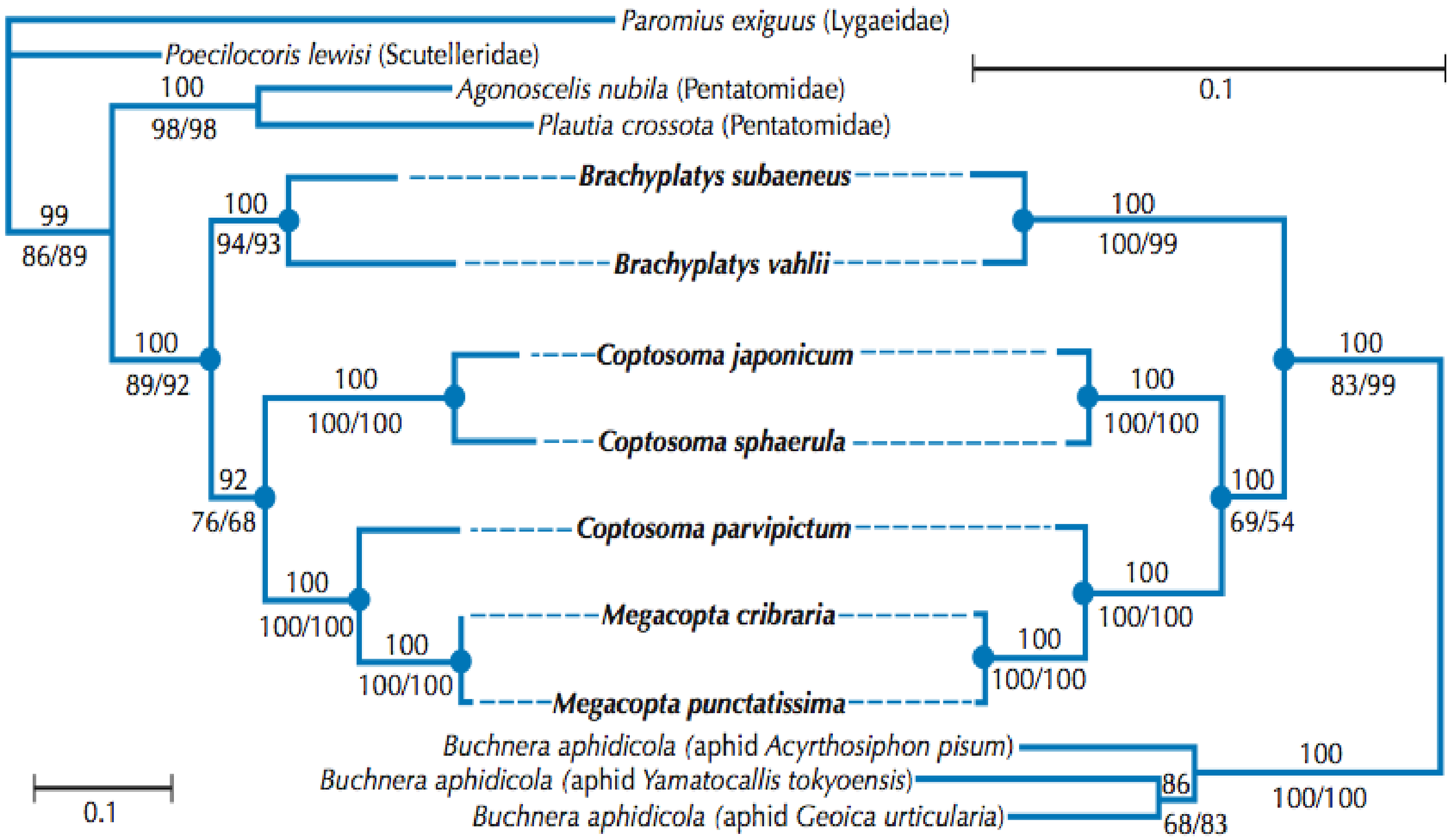
Jackknifed alignment 1 (10 columns kept)

```
Species  1        AT-T-GGA-GG-GA-
Species  2        AT-T-GGA-GG-GA-
Species  3        AT-T-AGG-GA-GA-
Species  4        AT-T-AGC-GC-GC-
```

Jackknifed alignment 2 (10 columns kept)

```
Species  1        --GTTGG--GGT-AT
Species  2        --GTTGG--GGA-AA
Species  3        --GTTAG--GAA-AA
Species  4        --GTCAG--GCC-CC
```

Congruence


Paromius exiguus (Lygaeidae)
Poecilocoris lewisi (Scutelleridae)
Agonoscelis nubila (Pentatomidae)
Plautia crossota (Pentatomidae)
0.1
Brachyplatys subaeneus
Brachyplatys vahlii
Coptosoma japonicum
Coptosoma sphaerula
Coptosoma parvipictum
Megacopta cribraria
Megacopta punctatissima
Buchnera aphidicola (aphid Acyrthosiphon pisum)
Buchnera aphidicola (aphid Yamatocallis tokyoensis)
Buchnera aphidicola (aphid Geoica urticularia)
100
98/98
99
86/89
100
94/93
100
89/92
100
100/99
100
83/99
100
100/100
100
100/100
92
76/68
100
69/54
100
100/100
100
100/100
100
100/100
100
100/100
86
68/83
100
100/100
0.1

Masking

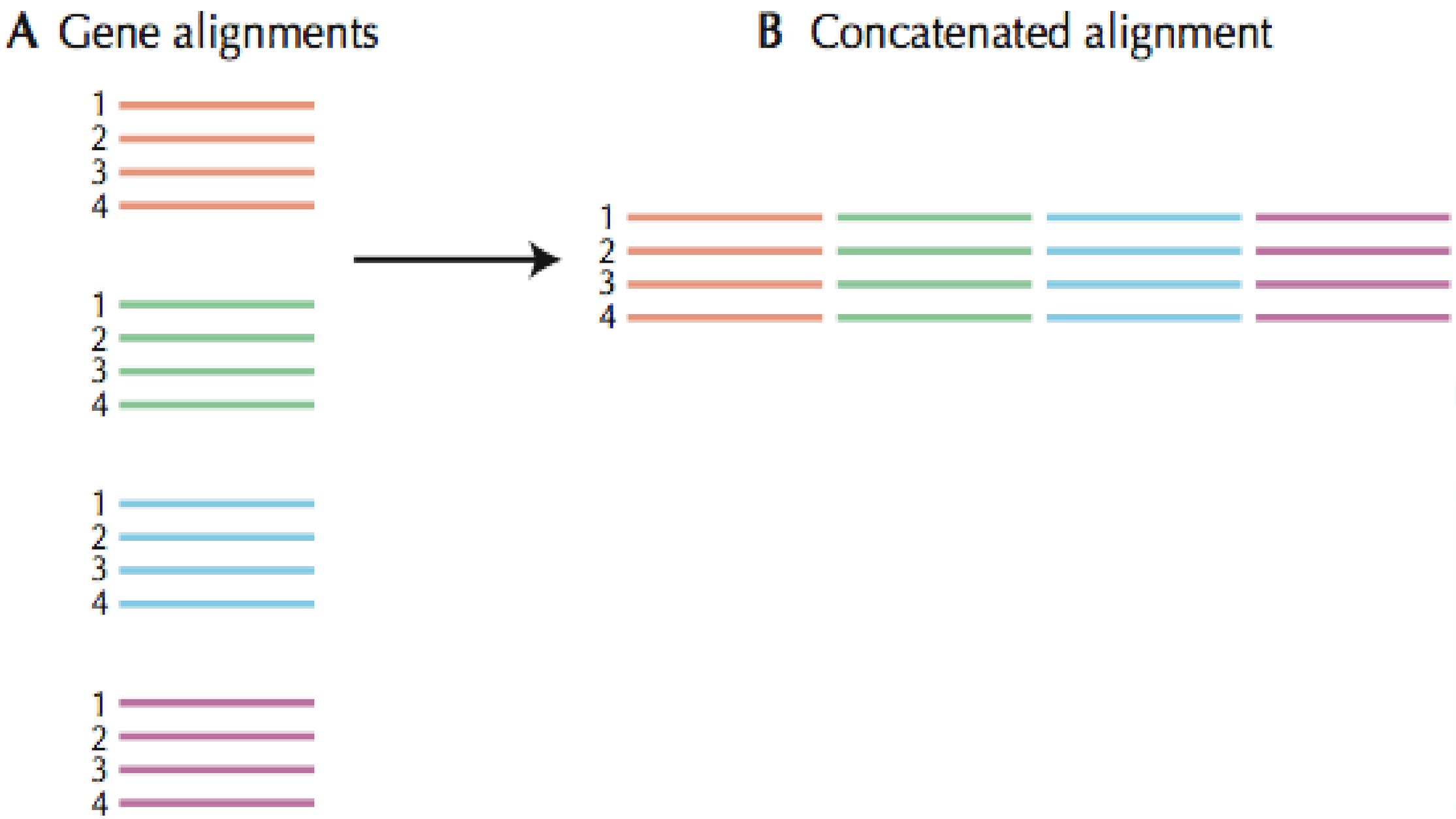
A

Protein	1	DEMGLGKK----R--VESTR
Protein	2	DEMGVGKRGH-----LESRK
Protein	3	DDMGLGRWK---R--VESTE
Protein	4	DDMGLGHKKKK-----VDSTK
Mask		11111111100000011111

B

Protein	1	DEMGLGKKVESTR
Protein	2	DEMGVGKRLESRK
Protein	3	DDMGLGRKVESTE
Protein	4	DDMGLGHKVDSTK

Concatenation



Homoplasy

A

Sequence 1 G A C G A G
Sequence 2 G C C G A C

B

Sequence 1 | Sequence 2

G A C G A G | G C C G A C

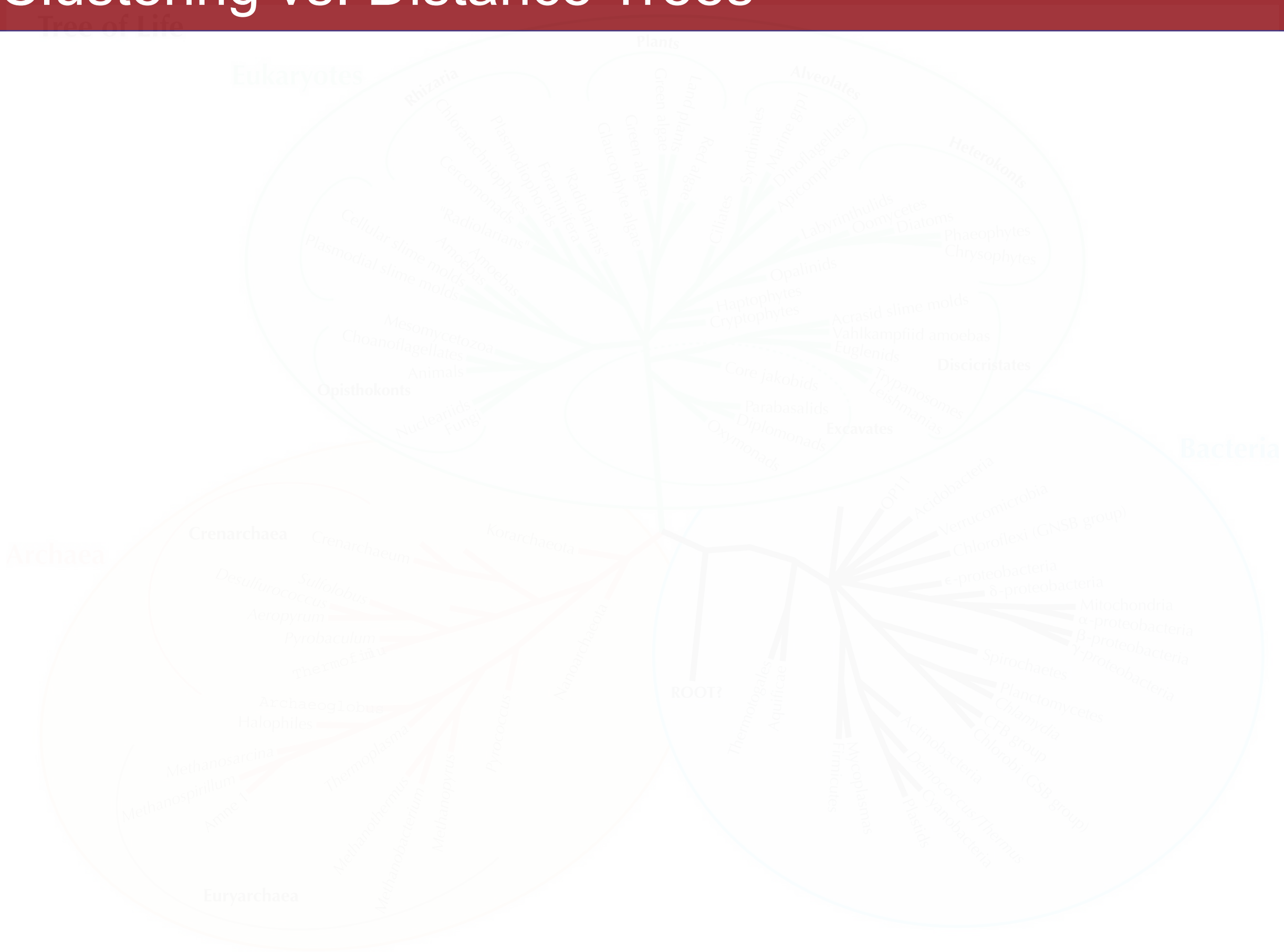
↑ G → A at position 5 | ↑ C → A at position 5

G A C G G G | G C C G C C

↑ G → A at position 2 | ↑ G → C at position 5; G → C at position 6

G G C G G G | G C C G G G } Ancestral sequences

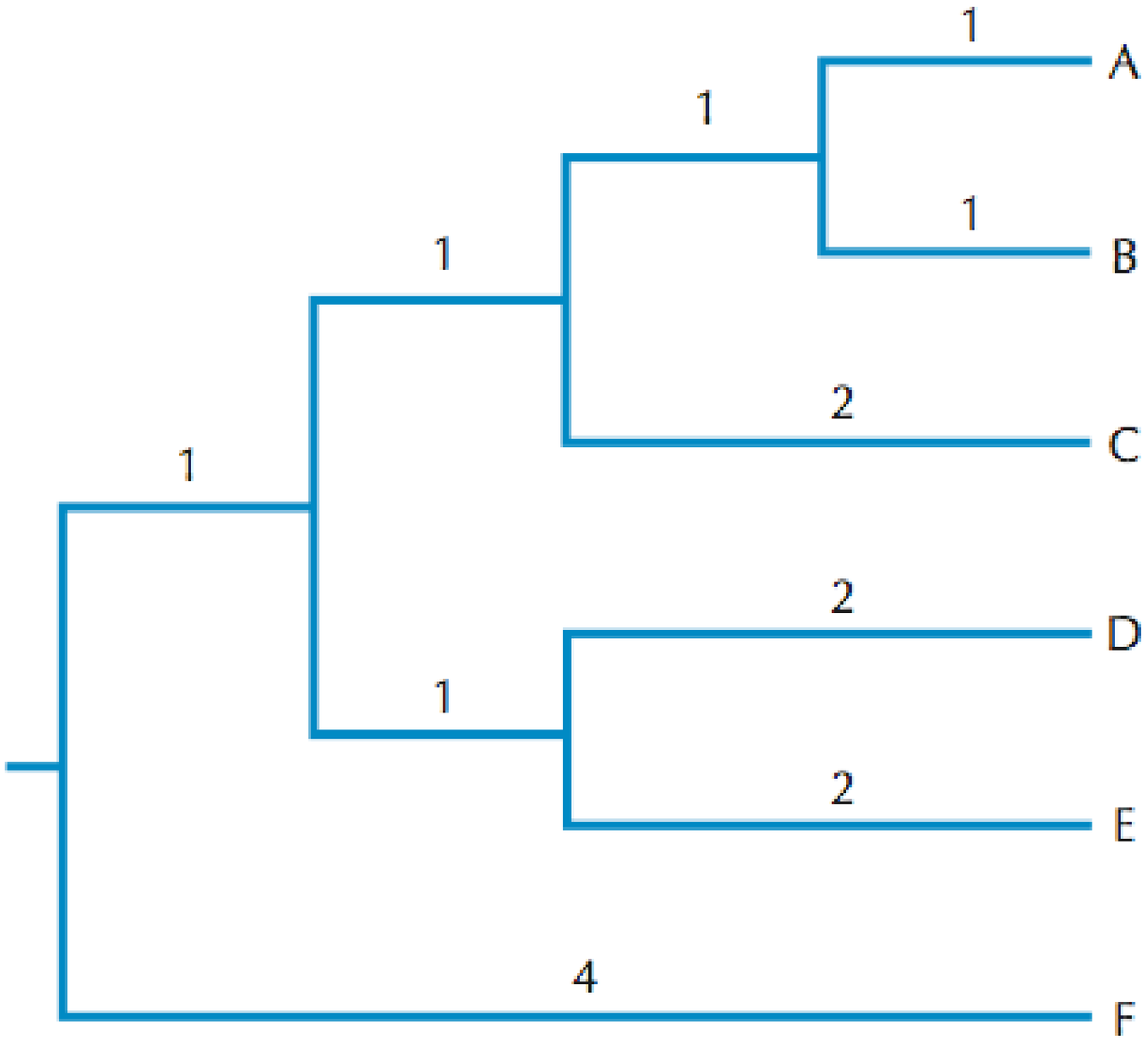
Clustering vs. Distance Trees



UPGMA

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic mean (UPGMA) algorithm

The True Tree



Distance Matrix For True Tree

TABLE 27.6. Distance matrix

OTUs	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

Collapse Diagonal

TABLE 27.7. Diagonal matrix—Step 1

OTUs	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

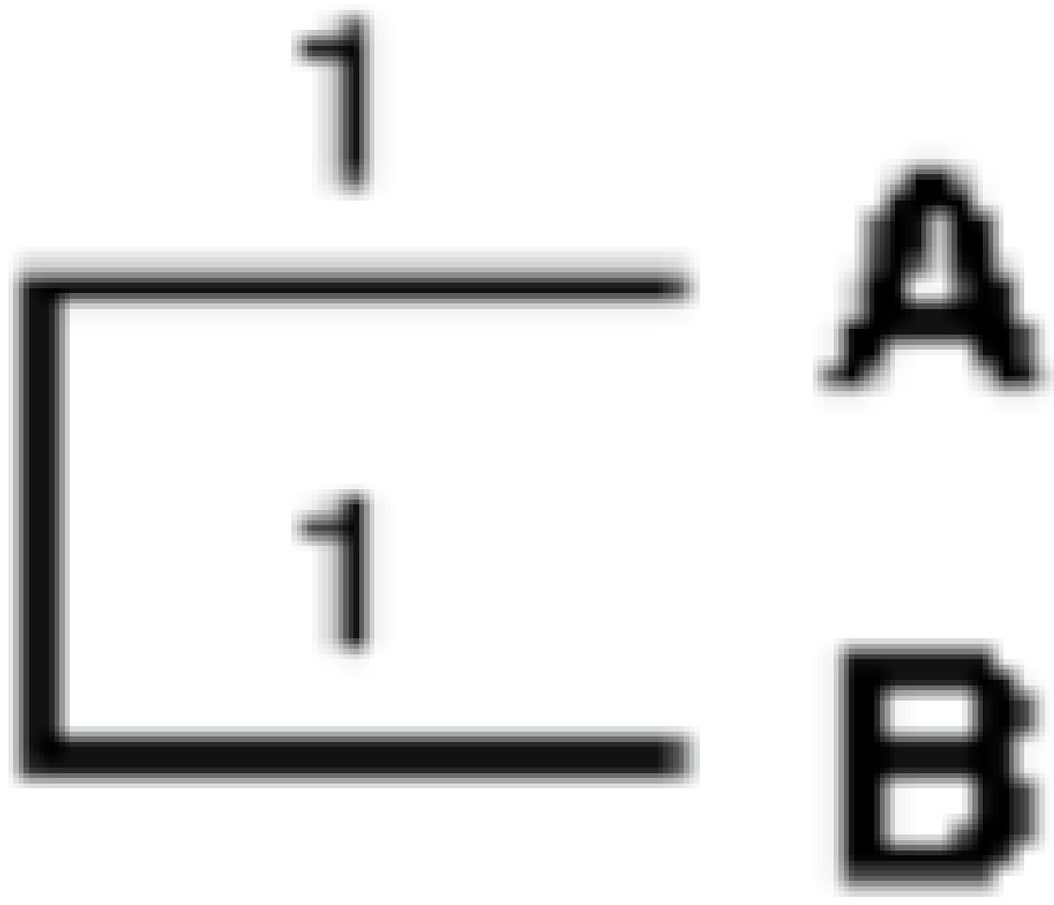
Identify Lowest D

OTUs	A	B	C	D	E
B	**2**				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Join Those Two Taxa

- Create branch with length = D
- 2
- A----------B

Make D from Node Equal



Create New Distance Matrix

- Merge Two OTUs joined in previous step (AB)

- $D_{x,AB} = 0.5 * (D_{x,A} + D_{x,B})$

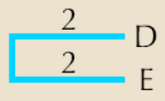
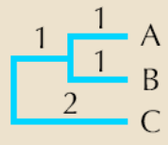
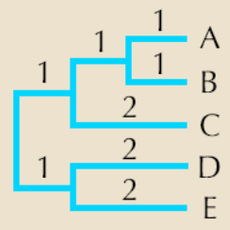
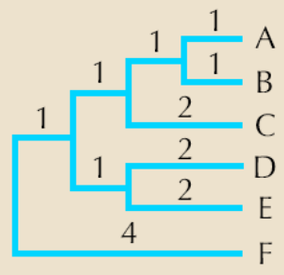
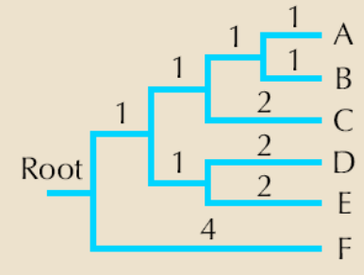
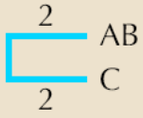
New Matrix

TABLE 27.8. Diagonal matrix—Step 2

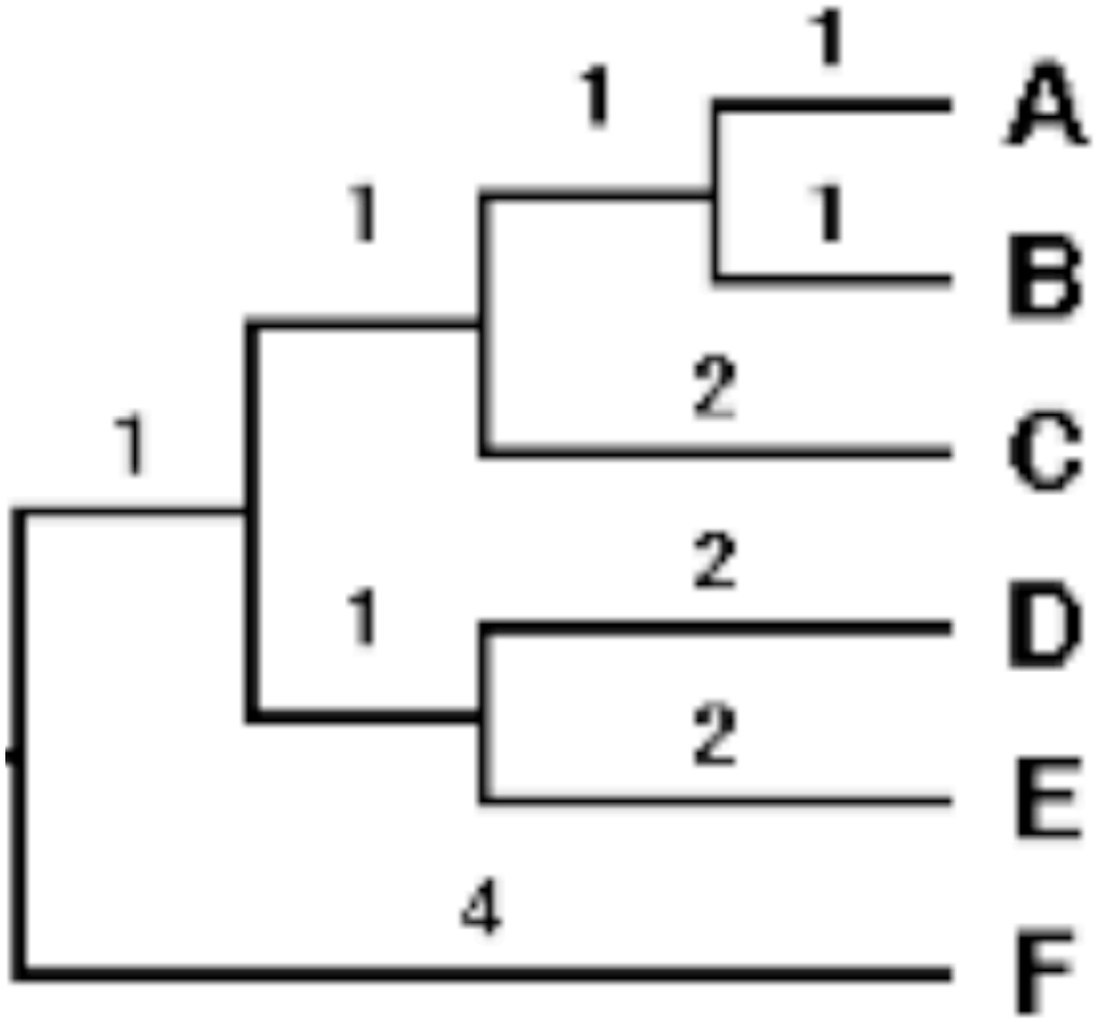
OTUs	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

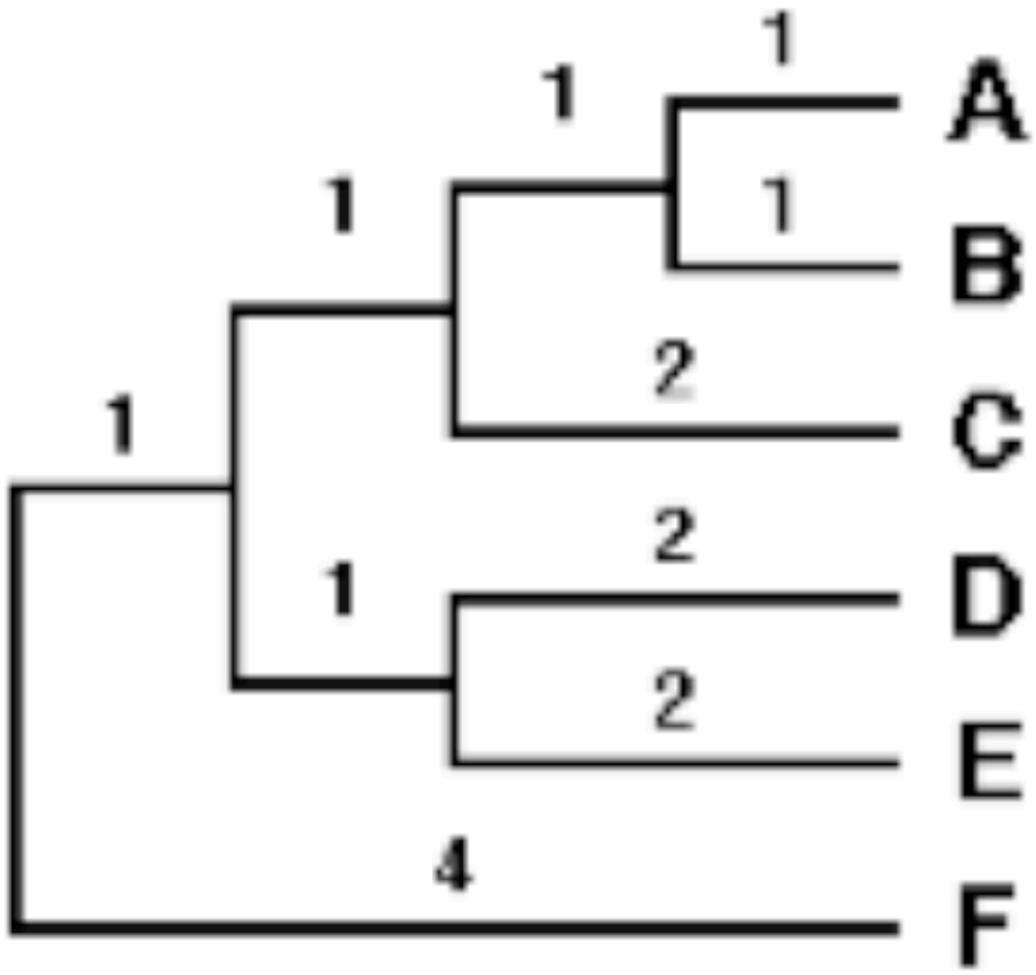
UPGMA

TABLE 27.9. Example of UPGMA tree construction

Step	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Distance matrix	OTUs | A B C D E B | 2 C | 4 4 D | 6 6 6 E | 6 6 6 4 F | 8 8 8 8 8	OTUs | AB C D E C | 4 D | 6 6 E | 6 6 4 F | 8 8 8 8	OTUs | AB C DE C | 4 DE | 6 6 DE | 8 8 8	OTUs | ABC DE DE | 6 F | 8 8	OTUs | ABCDE F | 8	No new matrix
Identify	A↔B = 2 smallest D	AB↔C = 4 D↔E= 4	AB↔DE = 6 C↔DE = 6	ABC↔DE	ABCDE↔F	
Taxa joined	A and B	D and E	AB and C	ABC and DE	ABCDE and F	
Subtree						
Comments on tree drawing	The distance between A and B is 2 units. A subtree is drawn with the branch point halfway between the two. Thus, each branch is 1 unit in length.	Branching done as in Step 1. Because the distance from AB to C is also 4, that pair could have been selected as well.	First a subtree is drawn with AB and C:  The the AB subtree is attached to the AB branch at a point equal to the length of the A and B branches.	The tree is first done as in Step 3 with the ABC and DE subtrees replacing the branches.	The tree is now complete but unrooted.	The tree can then be rooted using midpoint rooting which tries to balance all the tips to reach the same end point. Note this is the tree that we started with to build the distance matrix.

Compare to True Tree

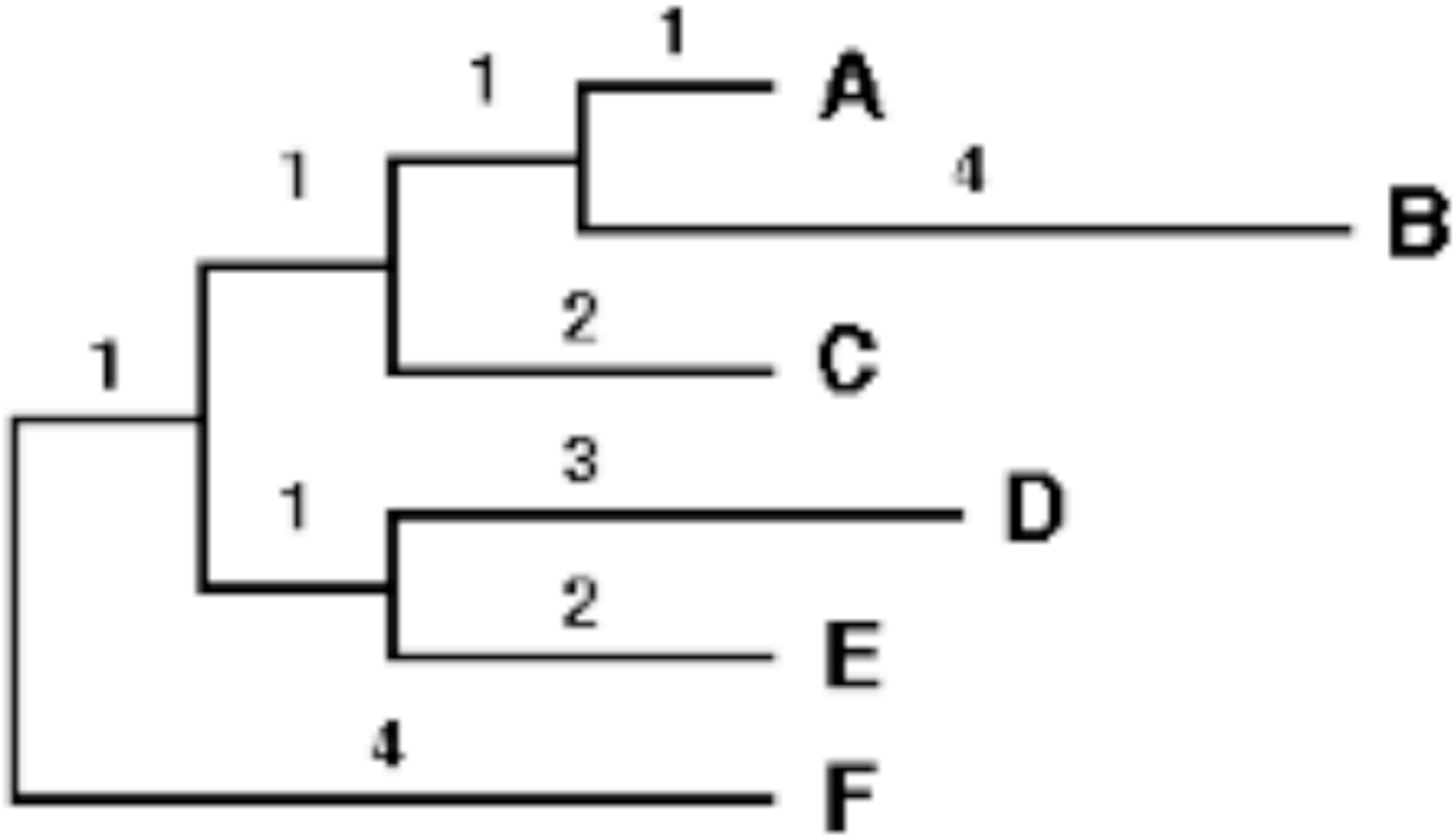




But …

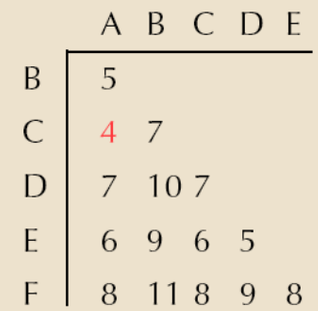
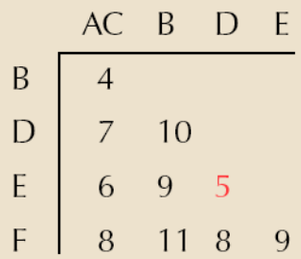
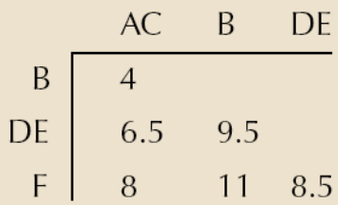
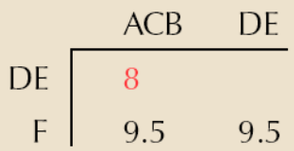
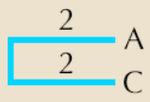
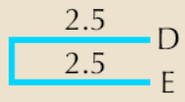
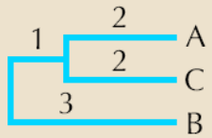
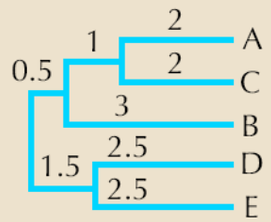
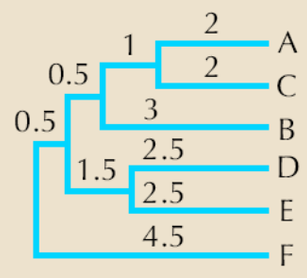
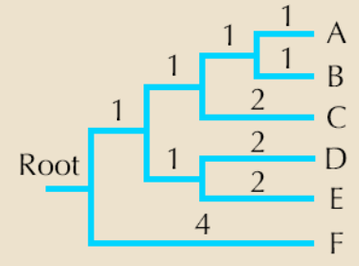
- What is evolutionary rates not equal

Unequal rates

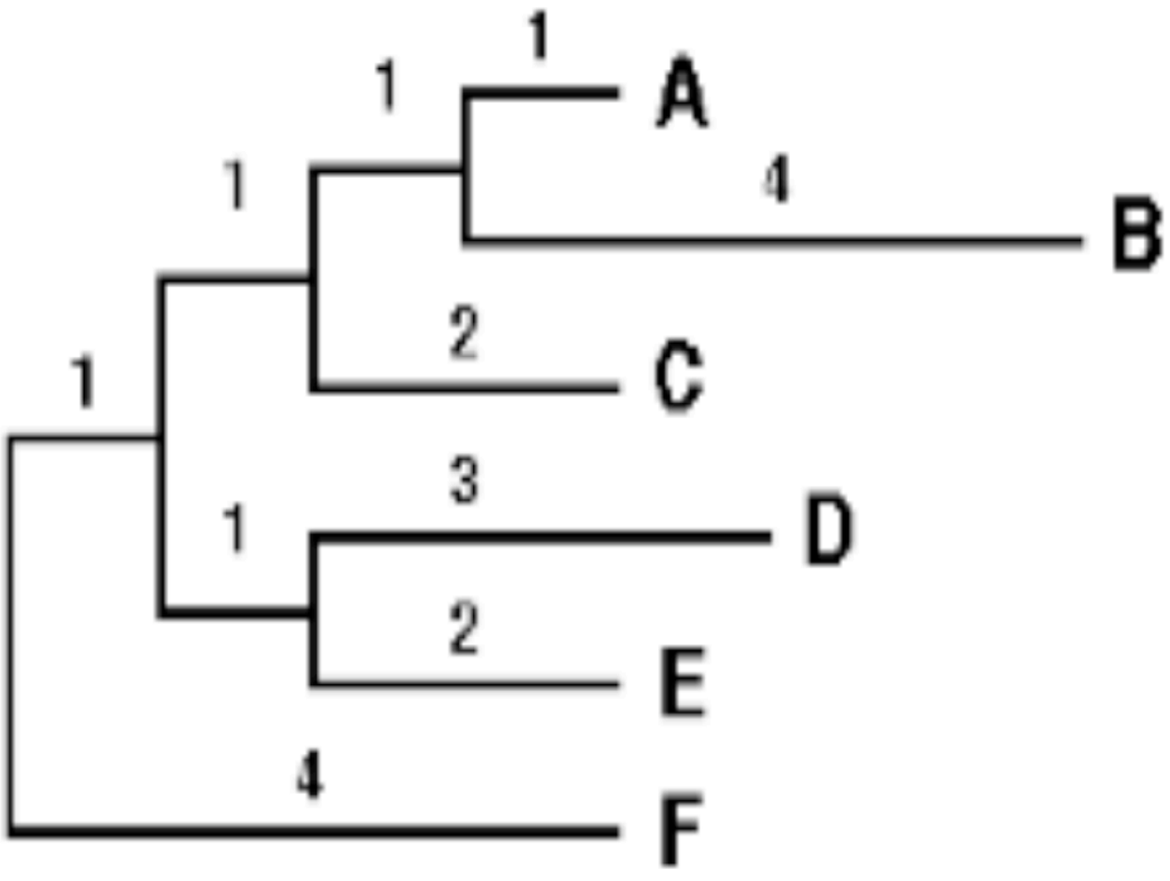


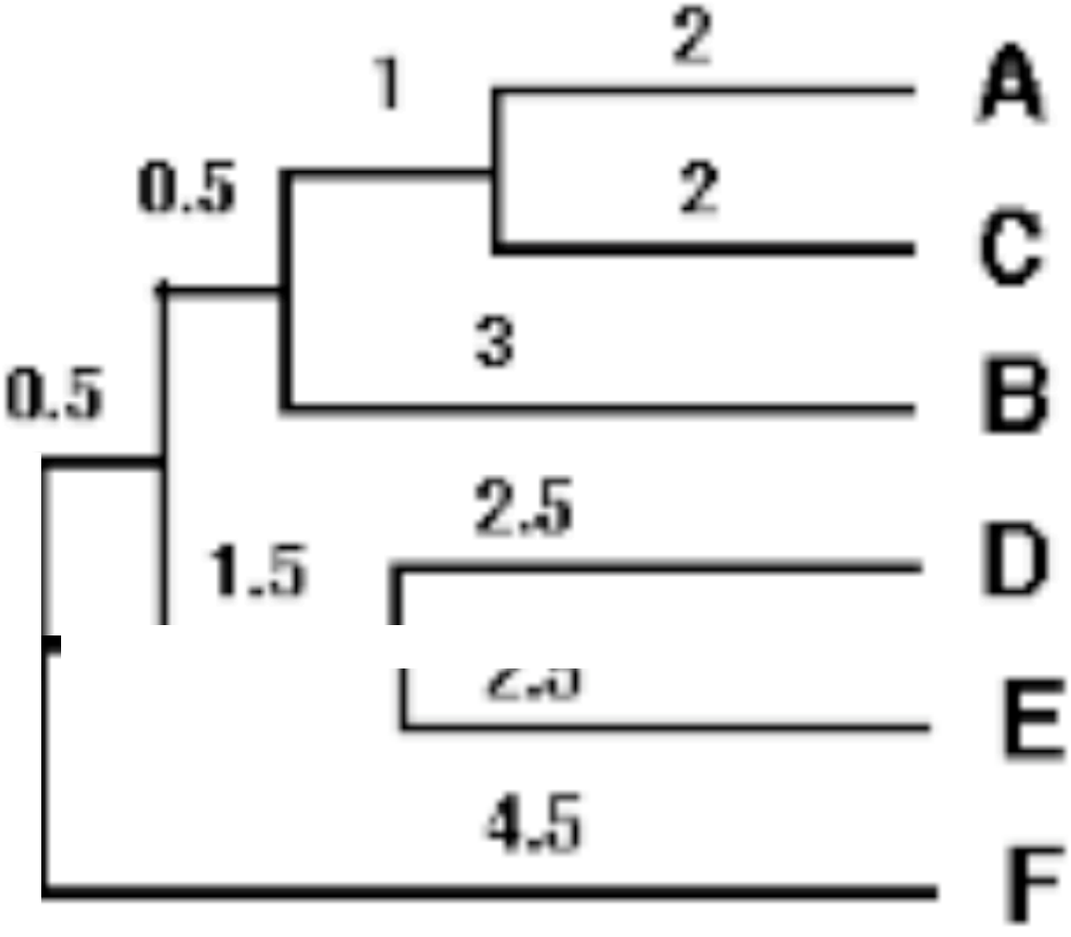
UPGMA with Unequal rates

TABLE 27.10. UPGMA tree construction errors

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Distance matrix						No new matrix
Identify smallest D	A↔C = 4	D↔E = 5	AC↔B = 4	ACB↔DF = 8	ABCDE↔F = 9	
Taxa joined	A and C	D and E	A and C with B	ABC with DE	ABCDE with F	
Subtree						
Comments					Note how this is not the same as the starting tree.	

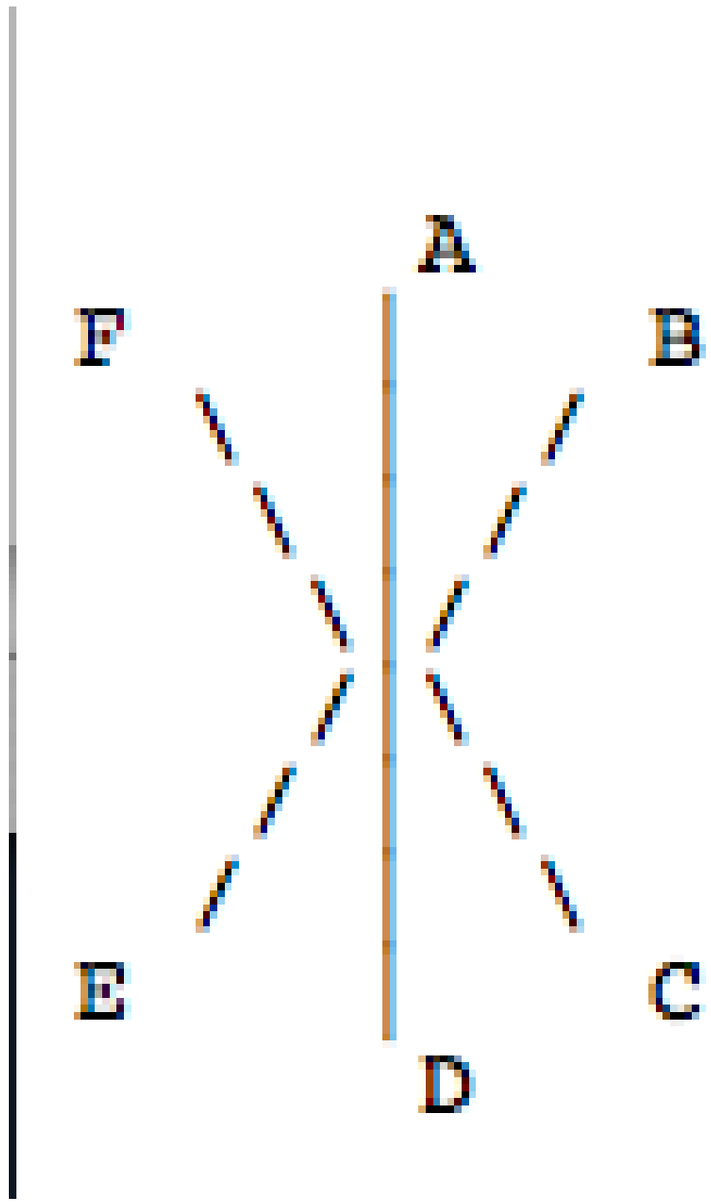
Compare to True Tree





Neighbor Joining

Start with Star Tree



Calculate S

- For each OTU, calculate a measure (S) as follows.
- S is the sum of the distances (D) between that OTU and every other OTU, divided by N-2 where N is the total number of OTUs.
- This is a measure of the distance an OTU is from all other OTUs

Calculate Pairwise Distances

- Calculate the distance D_{ij} between each OTU pair (e.g., I and J).

Identify Closest Pair

- Identify the pair of OTUs with the minimum value of $D_{ij} - S_i - S_j$

Joining Taxa

- As in UPGMA, join these two taxa at a node in a subtree.

Calculate Branches

- Calculate branch lengths. Unlike UPGMA, neighbor joining does not force the branch lengths from node X to I (D_{xi}) and to J (D_{xj}) to be equal, i.e., does not force the rate of change in those branches to be equal. Instead, these distances are calculated according to the following formulas
- $D_{xi} = 1/2\ D_{ij} + 1/2\ (S_i - S_j)$
- $D_{xj} = 1/2\ D_{ij} + 1/2\ (S_j - S_i)$

Calculate New Matrix

- Calculate a new distance matrix with I and J merged and replaced by the node (X) that joins them. Calculate the distances from this node to the other tips (K) by:

- $D_{xk} = (D_{ik} + D_{jk} - D_{ij})/2$

Distance Matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

S

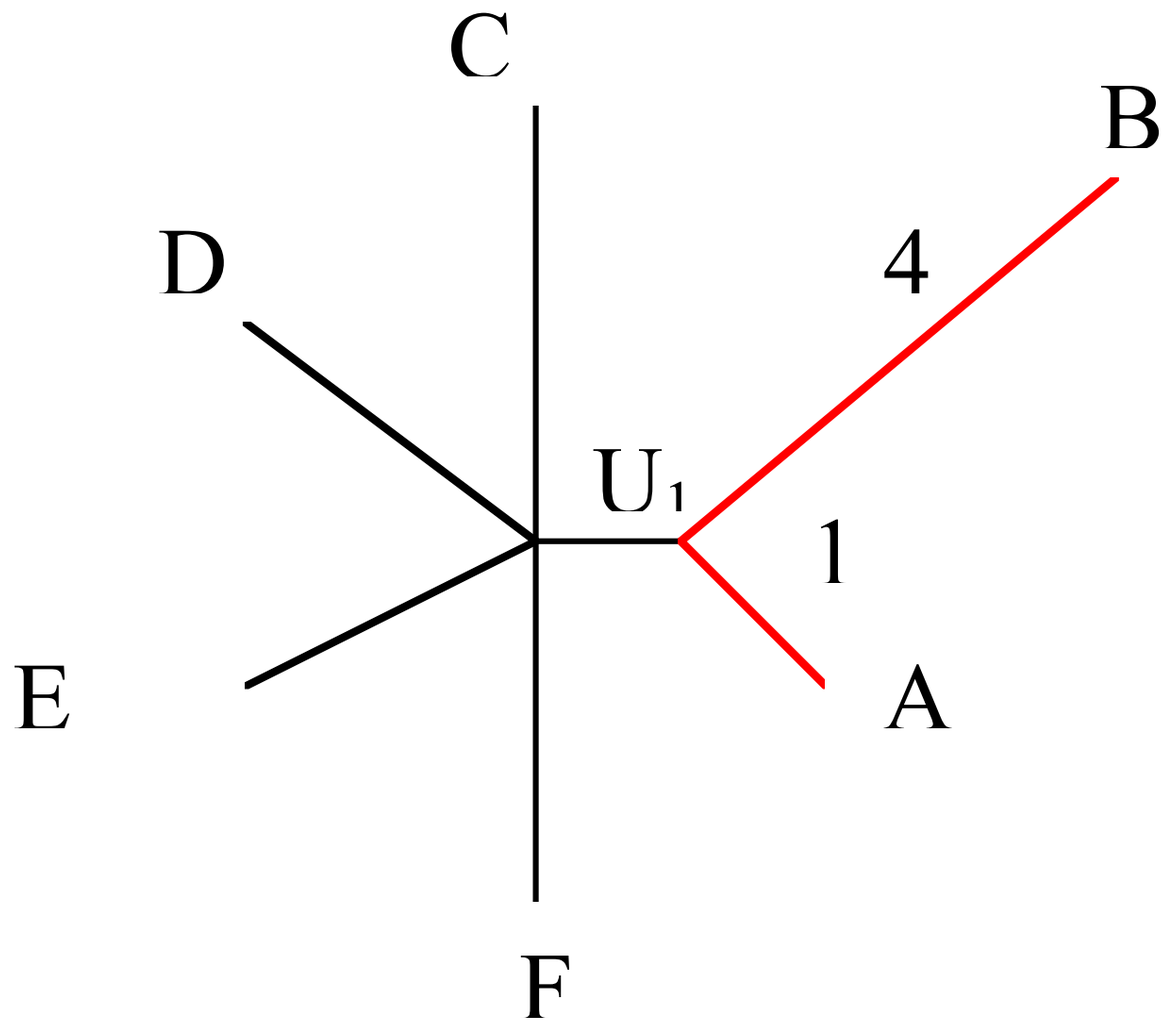
- $S_A = (5+4+7+6+8) / 4 =$ 7.5
- $S_B = (5+7+10+9+11) / 4 =$ 10.5
- $S_C = (4+7+7+6+8) / 4 =$ 8
- $S_D = (7+10+7+5+9) / 4 =$ 9.5
- $S_E = (6+9+6+5+8) / 4 =$ 8.5
- $S_F = (8+11+8+9+8) / 4 =$ 11

M

- $M_{ij} = D_{ij} - S_i - S_j$
- Smallest are
- $M_{AB} = 5 - 7.5 - 10.5 = -13$
- $M_{DE} = 5 - 9.5 - 8.5 = -13$
- Choose one of these (AB here)

New Tree

- Create a node (U) that joins pair with lowest M_{ij} such that
- $S_{IU} = D_{IJ}/2 + (S_I - S_J) / 2$
- Merge U with other taxa
- Join I and J according to S above and make all other taxa in form of a star


C
B
D
4
U_1
1
E
A
F

New Matrix

- $D_{XU} = D_{IX} + D_{JX} - D_{IJ}$ where I and J are those selected from above.

TABLE 27.11. Neighbor-joining example

Cycle 1

Distance matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Step 1

S calculations

S_x = (sum all D_x)/(N – 2), where N is the # of OTUs in the set.

$S_A = (5+4+7+6+8)/4 = 7.5$
$S_B = (5+7+10+9+11)/4 = 10.5$
$S_C = (4+7+7+6+8)/4 = 8$
$S_D = (7+10+7+5+9)/4 = 9.5$
$S_E = (6+9+6+5+8)/4 = 8.5$
$S_F = (8+11+8+9+8)/4 = 11$

Step 2

Calculate pair with smallest (M), where $M_{ij} = D_{ij} - S_i - S_j$.

Smallest are
$M_{AB} = 5 - 7.5 - 10.5 = -13$
$M_{DE} = 5 - 9.5 - 8.5 = -13$
Choose one of these (AB here).

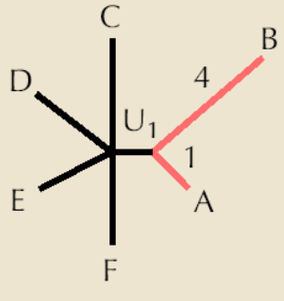
Step 3

Create a node (U) that joins pair with lowest M_{ij} such that $S_{IU} = D_{ij}/2 + (S_i - S_j)/2$.

U_1 joins A and B:
$S_{AU_1} = D_{AB}/2 + (S_A - S_B)/2 = 1$
$S_{BU_1} = D_{AB}/2 + (S_B - S_A)/2 = 4$

Step 4

Join i and j according to S above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.



Step 5

Calculate new distance matrix of all other taxa to U with $D_{xU} = D_{ix} + D_{jx} - D_{ij}$, where i and j are those selected from above.

Cycle 2

	U_1	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

Step 1

$S_{U_1} = (3+6+5+7)/3 = 7$
$S_C = (3+7+6=8)/3 = 8$
$S_D = (6+7+5+9)/3 = 9$
$S_E = (5+6+5+8)/3 = 8$
$S_F = (7+8+9+8)/3 = 10.6$

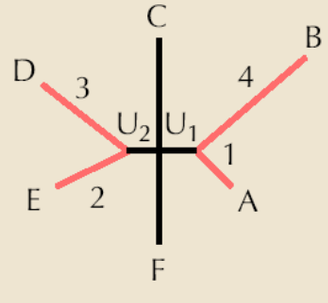
Step 2

Smallest is
$M_{CU_1} = 3 - 7 - 8 = -12$
$M_{DE} = 5 - 9 - 8 = -12$
Choose one of these (DE here).

Step 3

U_2 joins D and E:
$S_{DU_2} = D_{DE}/2 + (S_D - S_E)2 = 3$
$S_{EU_2} = D_{DE}/2 + (S_E - S_D)/2 = 2$

Step 4



Step 5

Cycle 3

	U_1	C	U_2
C	3		
U_2	3	4	
F	7	8	6

Step 1

$S_{U_1} = (3+3+7)/2 = 6.5$
$S_C = (3+4+8)/2 = 7.5$
$S_{U_2} = (3+4+6)/2 = 6.5$
$S_F = (7+8+6)/2 = 10.5$

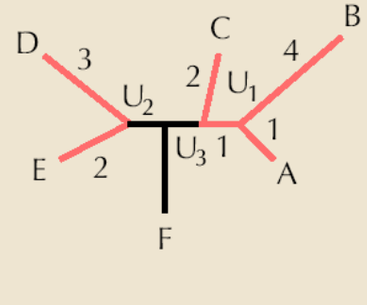
Step 2

Smallest is
$M_{CU_1} = 3 - 6.5 - 7.5 = -11$

Step 3

U_3 joins C and U_1:
$S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$
$S_{U_1U_3} = D_{CU_1}/2 + (S_{U1} - S_C)/2 = 1$

Step 4



Step 5

Cycle 4

	U_2	U_3
U_3	2	
F	6	6

Step 1

$S_{U_2} = (2+6)/1 = 8$
$S_{U_3} = (2+6)/1 = 8$
$S_F = (6+6)/1 = 12$

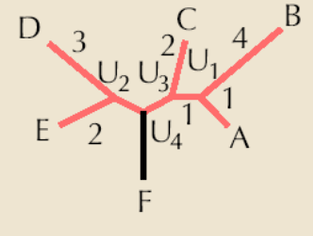
Step 2

Smallest is
$M_{U_2F} = 6 - 8 - 12 = -14$
$M_{U_3F} = 6 - 8 - 12 = -14$
$M_{U_2U_3} = 2 - 8 - 8 = -14$
Choose one of these ($M_{U_2U_3}$ here).

Step 3

U_4 joins U_2 and U_3:
$S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U2} - S_{U3})/2 = 1$
$S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U3} - S_{U2})/2 = 1$

Step 4



Step 5

Cycle 5

	U_4
F	5

Step 1

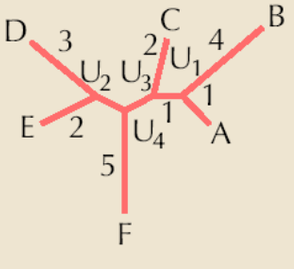
Because N – 2 = 0, we cannot do this calculation.

Step 2

Step 3

For last pair connect with branch = D.
Here $D_{U_4F} = 5$.

Step 4

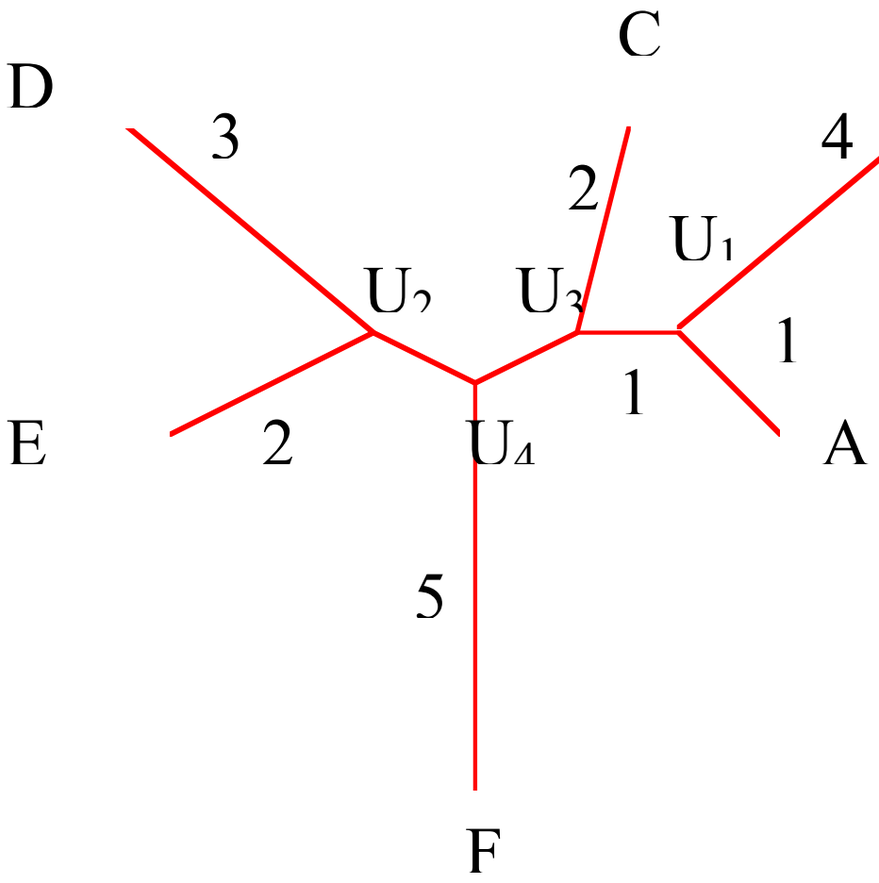


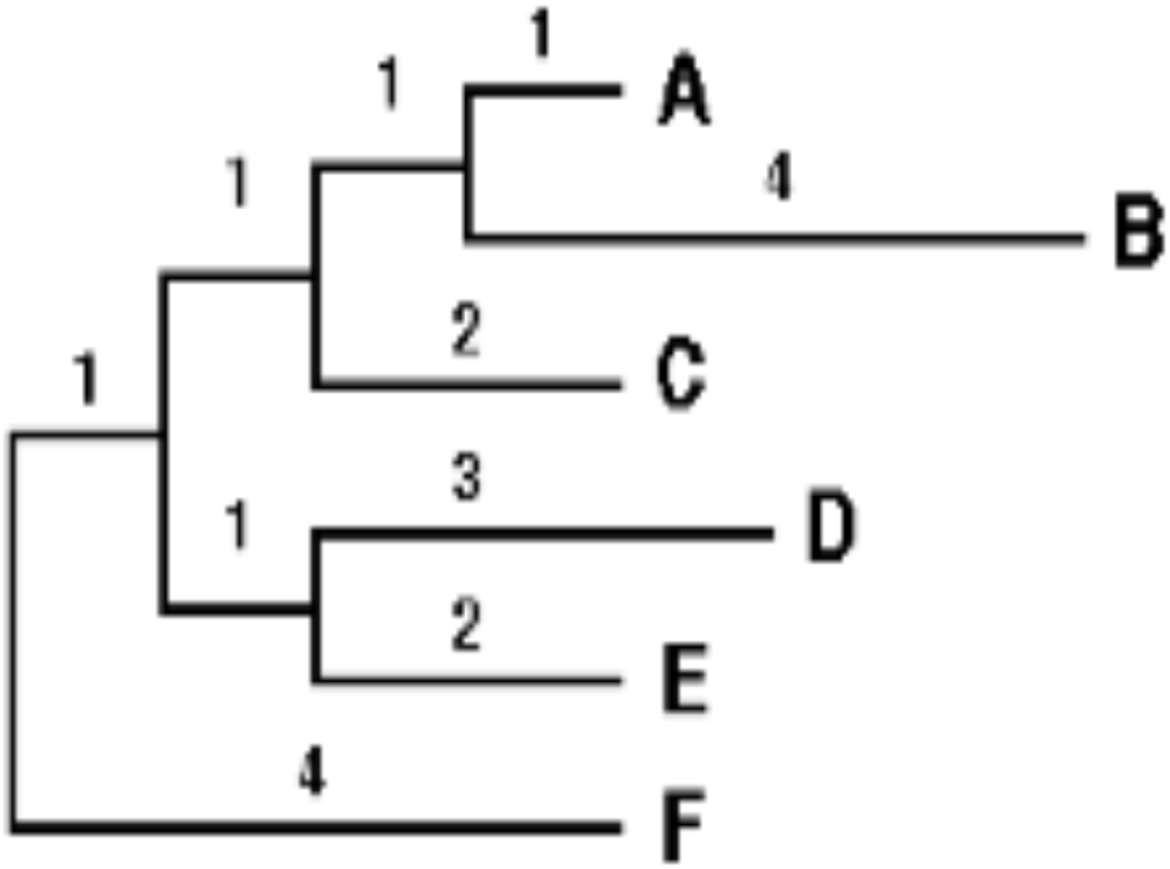
Step 5

Comments

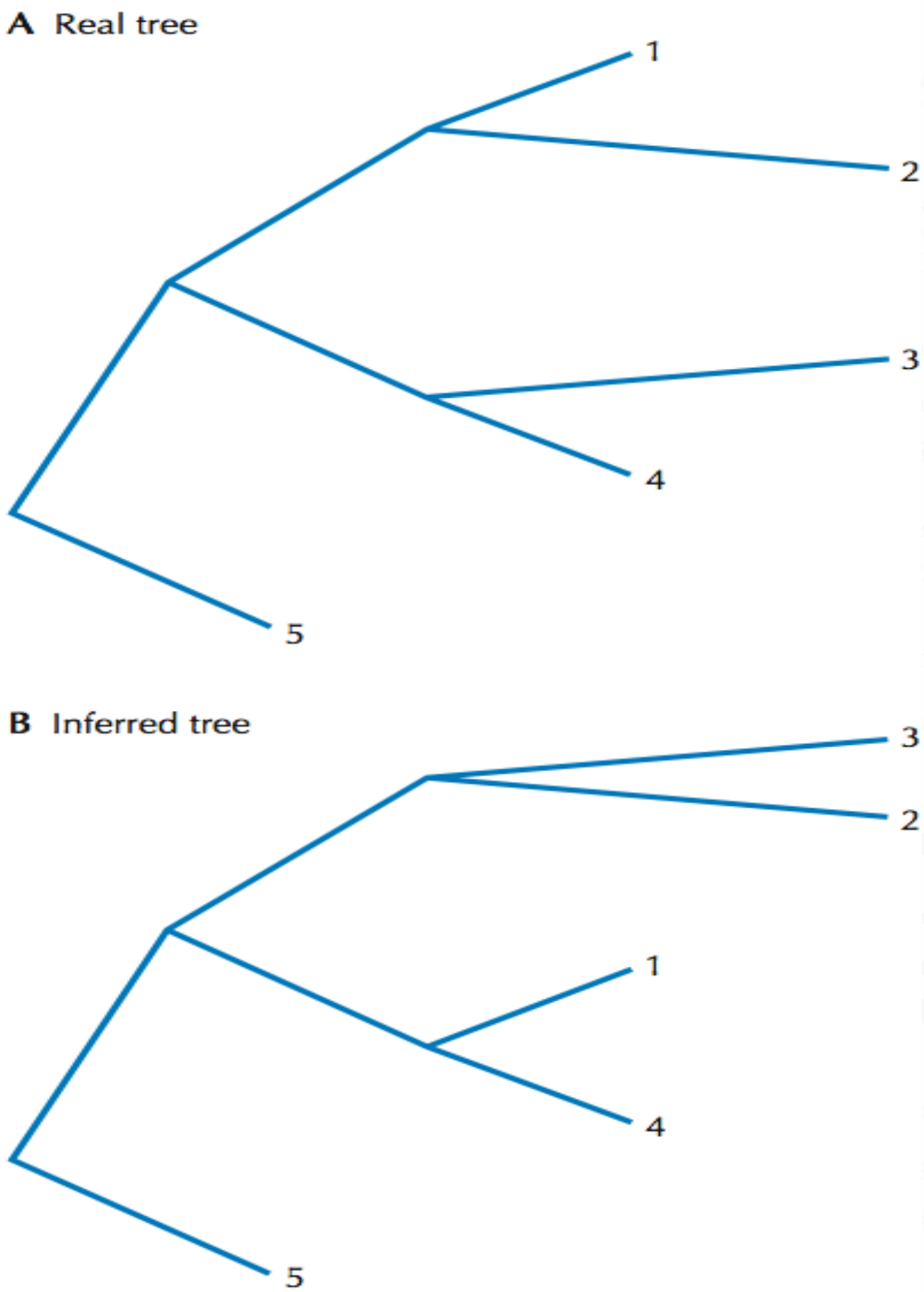
Note this is the same tree we started with (drawn in unrooted form here).

Compare to True Tree



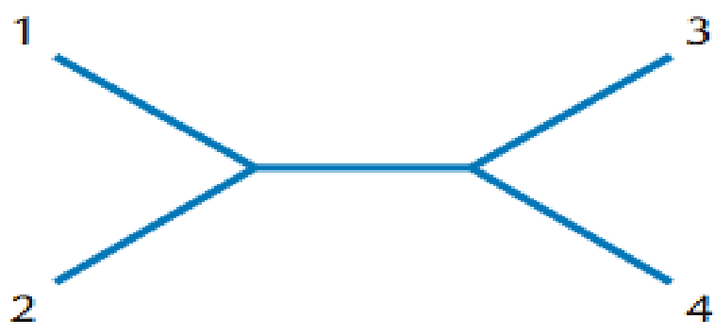


Long branch attraction

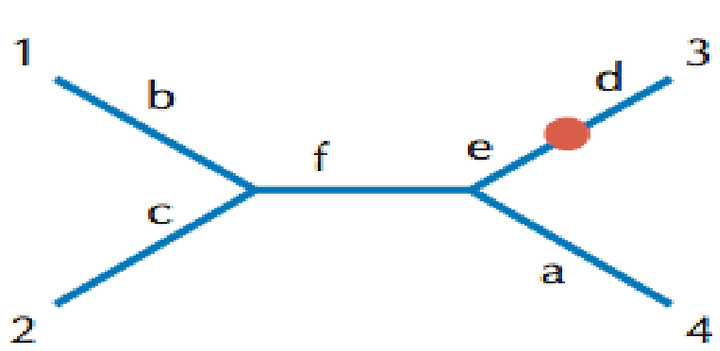


Rooting

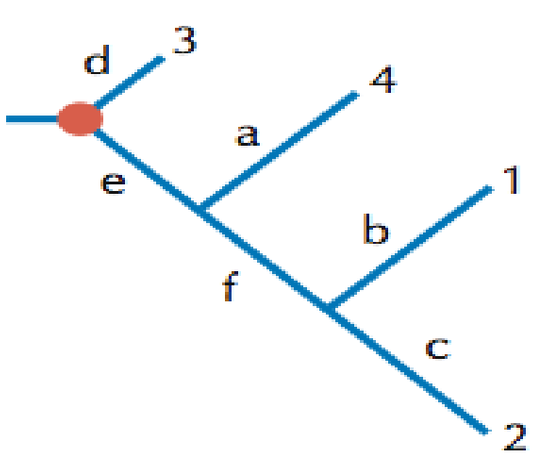
A



B



C



Rooting TOL Review