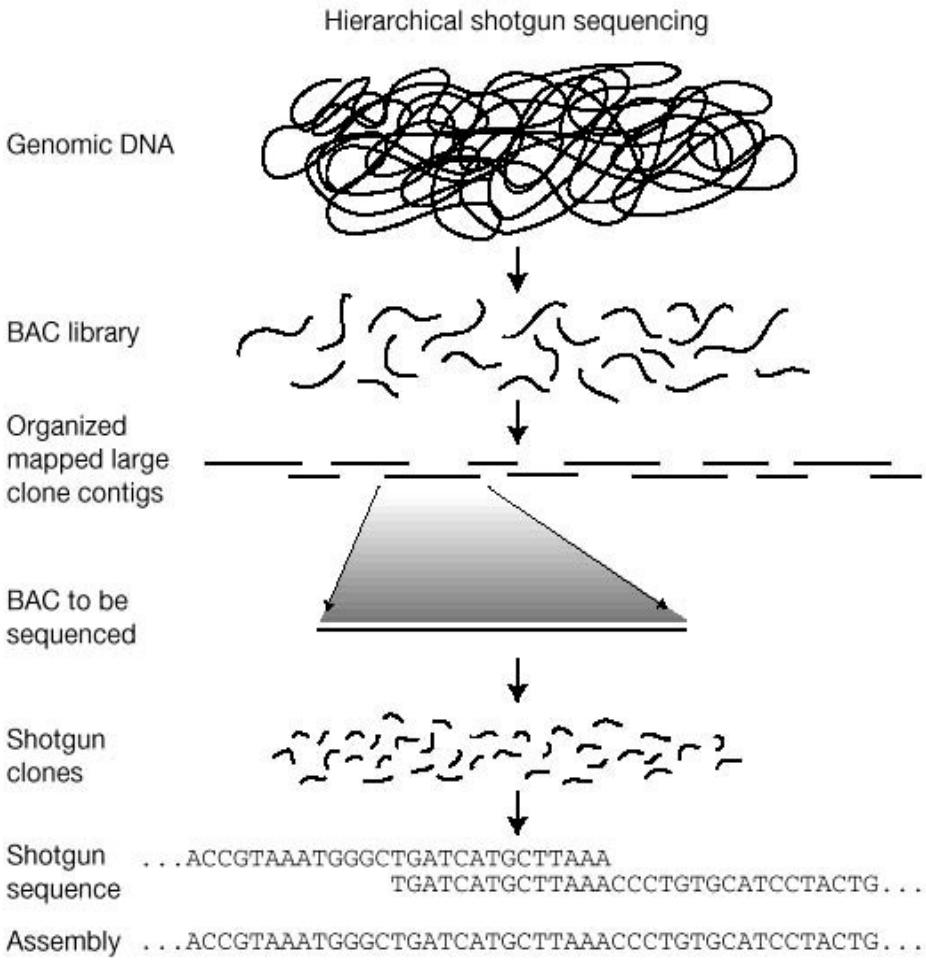
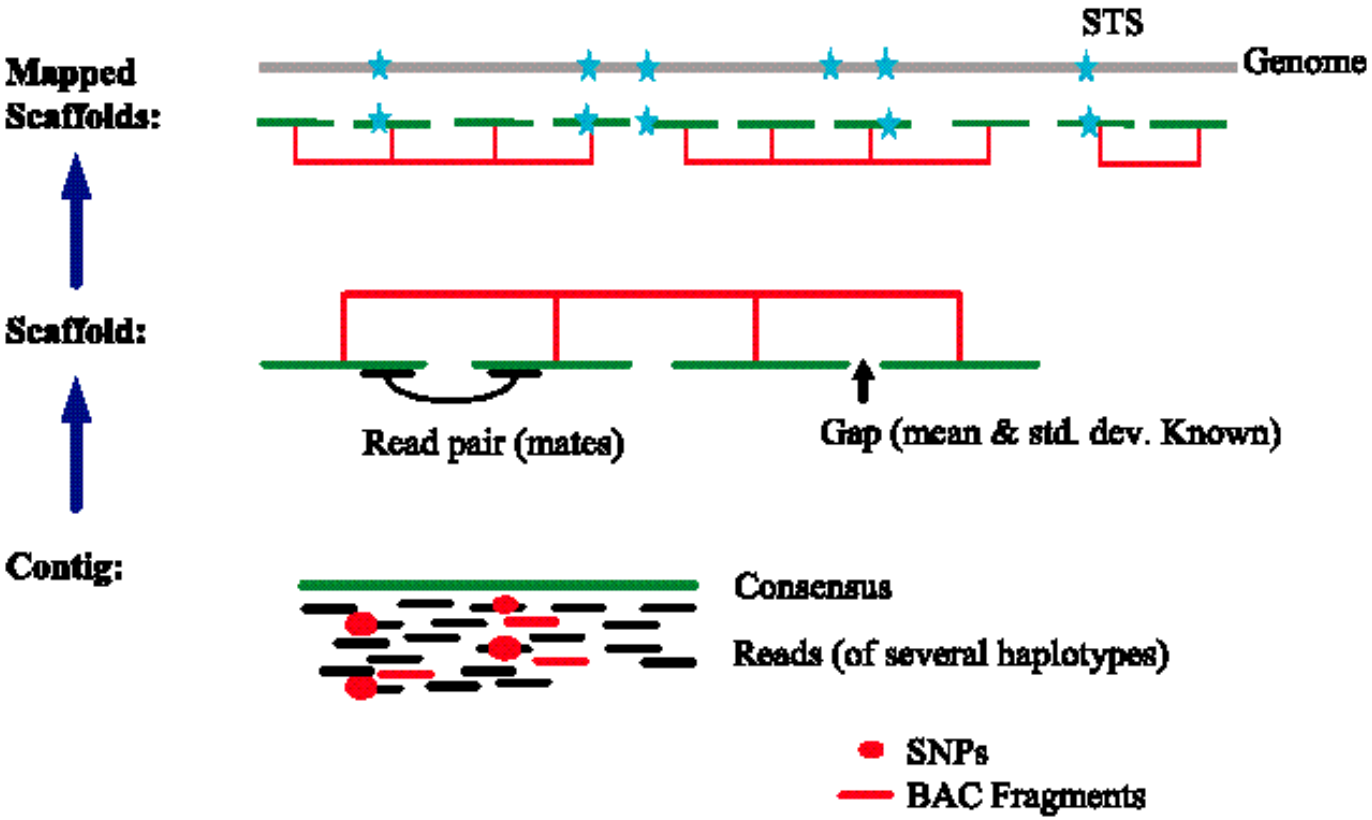
Sequencing

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Human Genome Project

Play the Sequencing Video:

- Download Windows file from

http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe

- Then run it on your PC.

Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-
  - **--ACCGT--**
  - **----CGTGC**
  - **TTAC-----**
  - **-TACCGT—**
  - **TTACCGTGC**

Assembly: Complications

- Errors in input sequence fragments (~3%)
  - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
  - Fragment contained in a repeat
  - Repeat copies not exact copies
  - Inherently ambiguous assemblies possible
  - Inverted repeats
- Inadequate Coverage

Assembly: Complications

w = AGTATTGGCAATC

z = AATCGATG

u = ATGCAAACCT

x = CCTTTTGG

y = TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
--------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

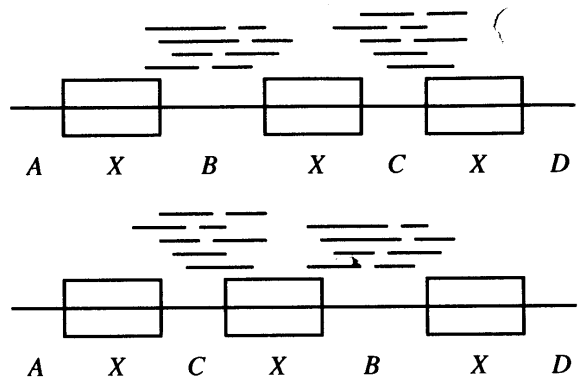


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

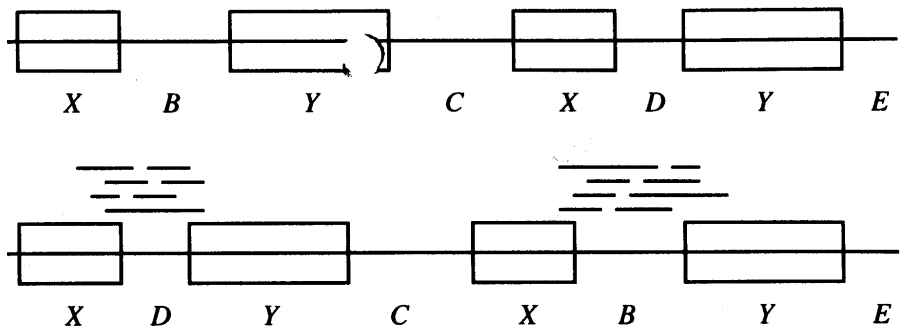


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

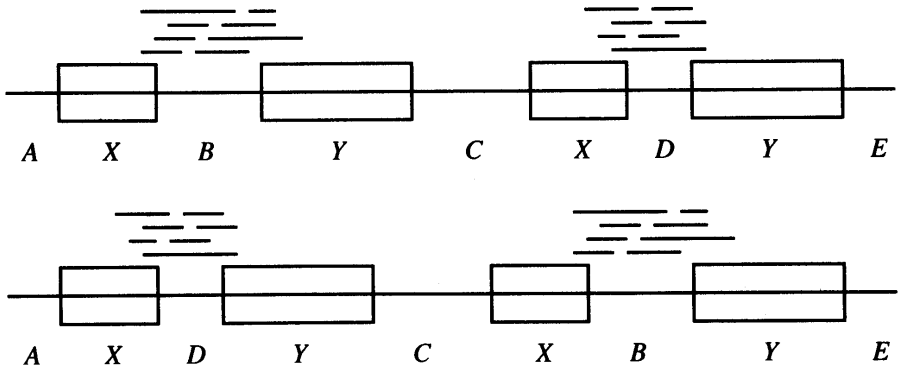


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

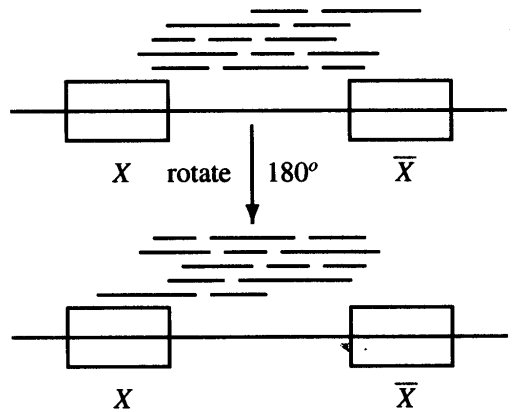


FIGURE 4.10

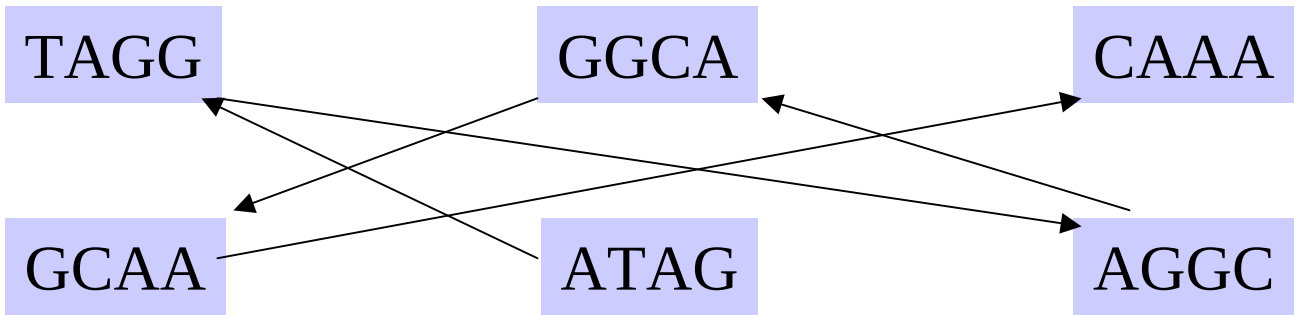
Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.

Miscellaneous

- Contig: A continuously covered region in the assembly.
- Other sequencing methods:
  - Sequencing by Hybridization (SBH)
  - Dual end sequencing
  - Chromosome Walking (see page 5-6 of Pevzner's text).

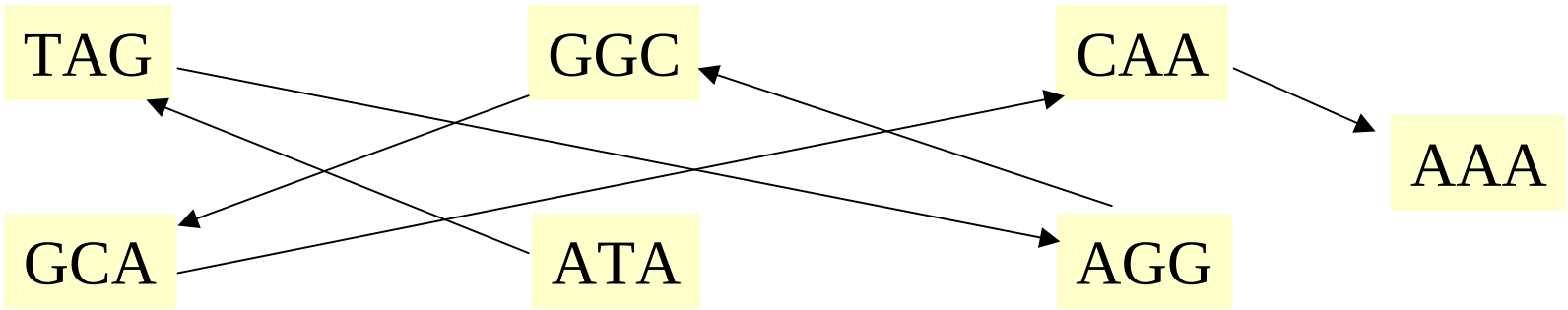
SBH

- Suppose that the only length 4 fragments that hybridize to S are: **TAGG, GGCA, CAAA, GCAA, ATAG, AGGC**. Then what is S, if it is of length ~9?



Hamiltonian Path Problem

SBH



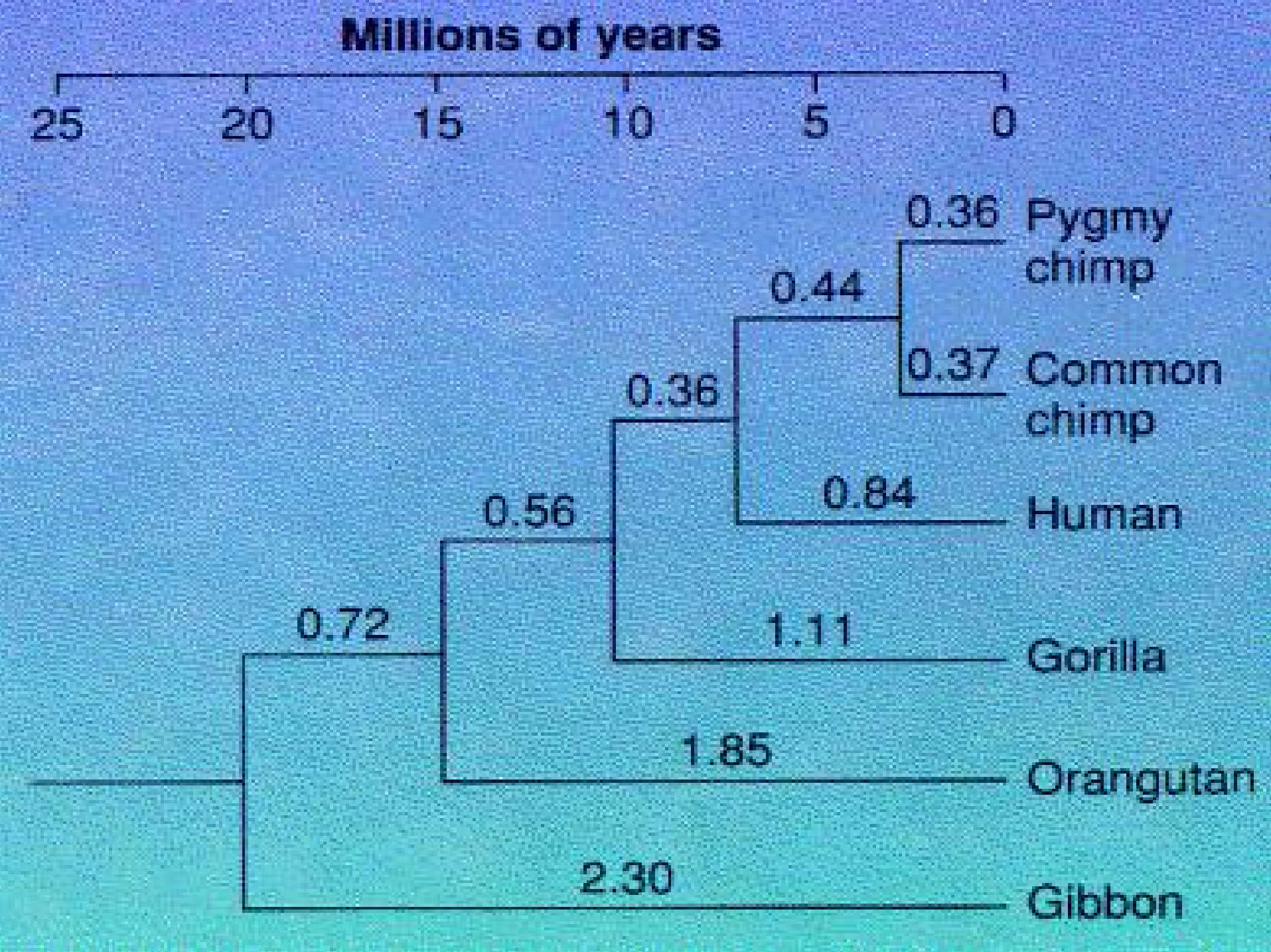
Eulerian Path Problem

Assembly Software

- Parallel EST alignment engine (http://corba.ebi.ac.uk/EST") with a CORBA interface to alignment database. Can perform ad hoc assemblies. Can act as foundation for CORBA-based EST assembly and editing package. [Parsons, EBI]
- Software using multiple alternative sequence assembly "engines" writing to a common format file [Staden, Cambridge] (http://www.mrc-lmb.cam.ac.uk/pubseq/index.html).
- Phrap,(http://bozeman.genome.washington.edu/phrap.docs/phrap.html)
- Assembler (TIGR) for EST and Microbial whole-genome assembly (http://www.tigr.org/softlab/)
- FAK2 and FAKtory (http://www.cs.arizona.edu/people/gene/) [Myers]
- GCG (http://www.gcg.com)
- Falcon [Gryan, Harvard] fast (rascal.med.harvard.edu/gryan/falcon/)
- SPACE, SPASS [Lawrence Berkeley Labs] (http://www-hgc.lbl.gov/inf/space.html)
- CAP 2 [Huang] (http://www.tigem.it/ASSEMBLY/capdoc.html)

Theory of Evolution

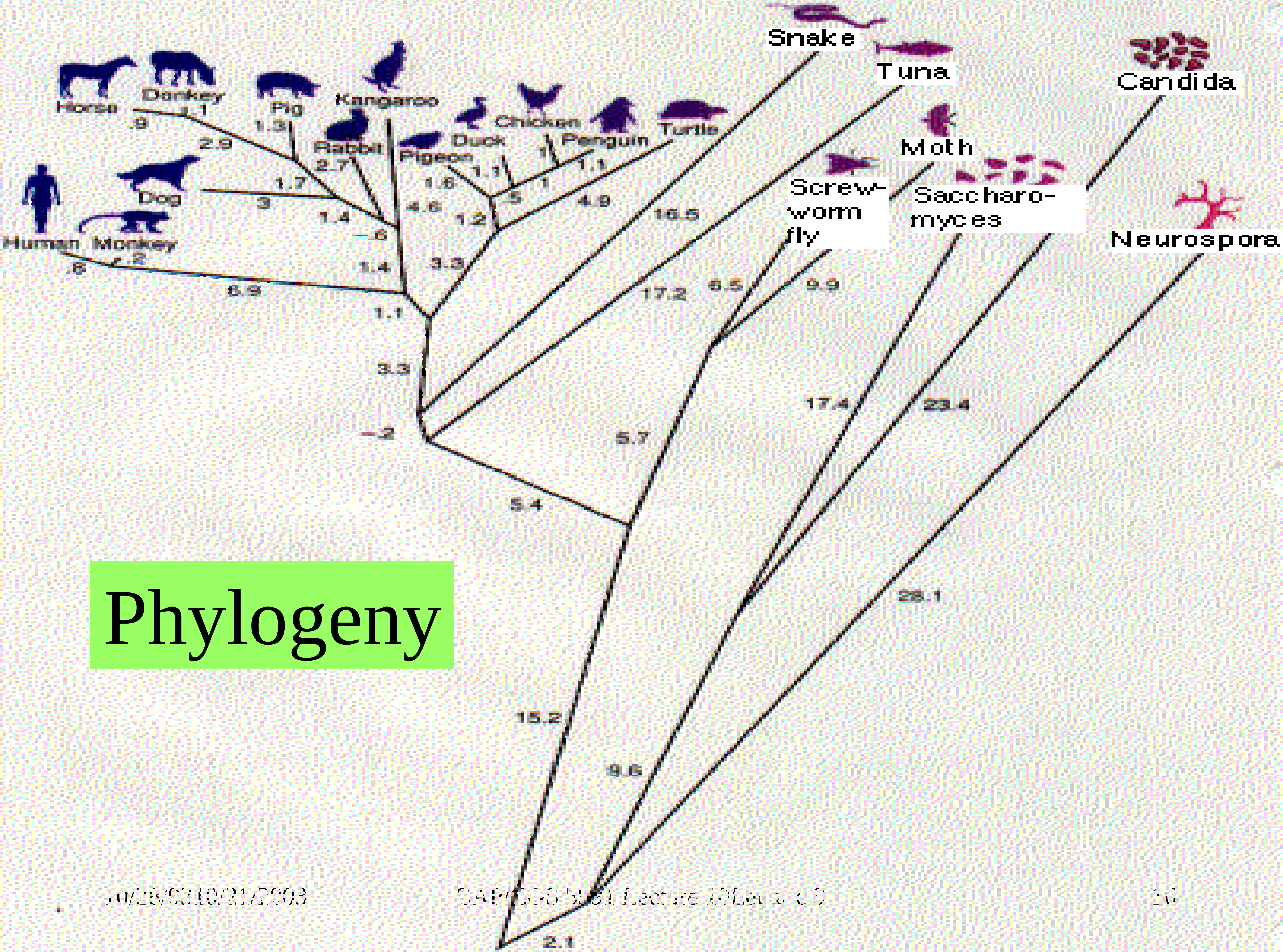
- Charles Darwin
  - **1858-59:** *Origin of Species*
  - 5 year voyage of H.M.S. Beagle (1831-36)
  - Populations have variations.
  - Natural Selection & Survival of the fittest: *nature selects best adapted varieties to survive and to reproduce.*
  - Speciation arises by splitting of one population into subpopulations.
  - Gregor Mendel and his work (1856-63) on inheritance.


Millions of years
25
20
15
10
5
0
0.36
Pygmy chimp
0.44
0.37
Common chimp
0.36
0.84
Human
0.56
0.72
1.11
Gorilla
1.85
Orangutan
2.30
Gibbon

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.

Phylogeny

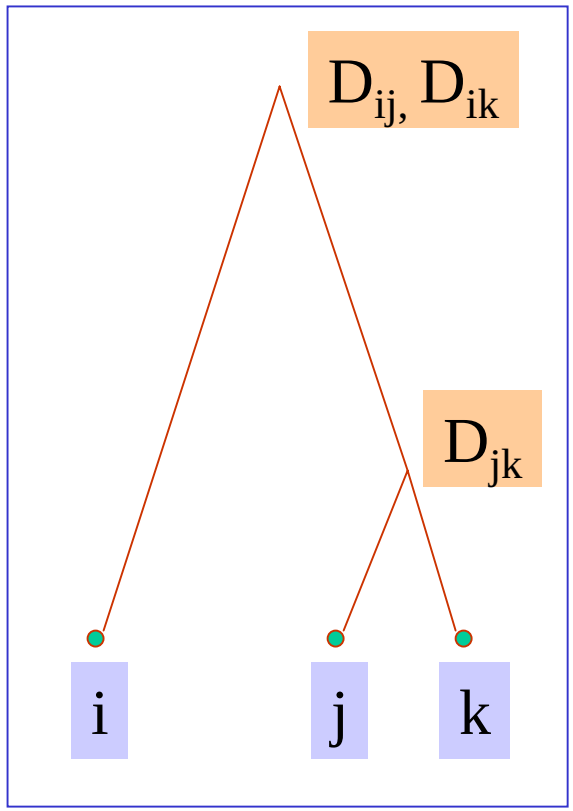

Snake
Tuna
Candida
Horse
Donkey
Pig
Kangaroo
Chicken
Duck
Penguin
Turtle
Rabbit
Pigeon
Moth
Dog
Human
Monkey
Screw-worm fly
Saccharo-myces
Neurospora

Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric [Nodes are labeled; distance = label of LCA]
    - Additive: [Edges have weights; distance = length of path]
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based [Edges have labels; set of changes = set of labels on path]
    - 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

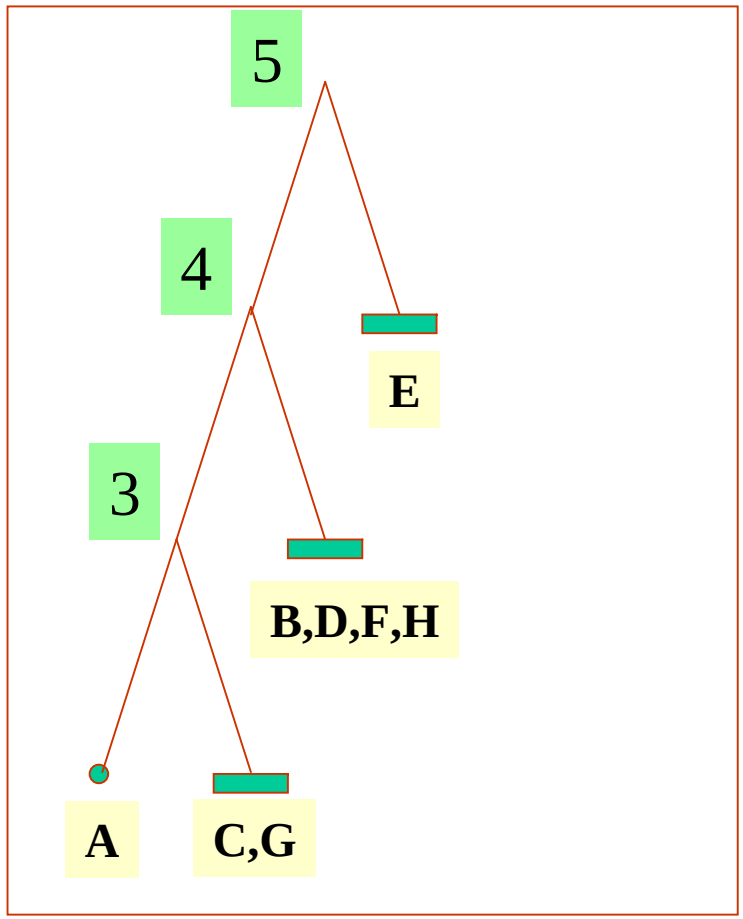
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

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

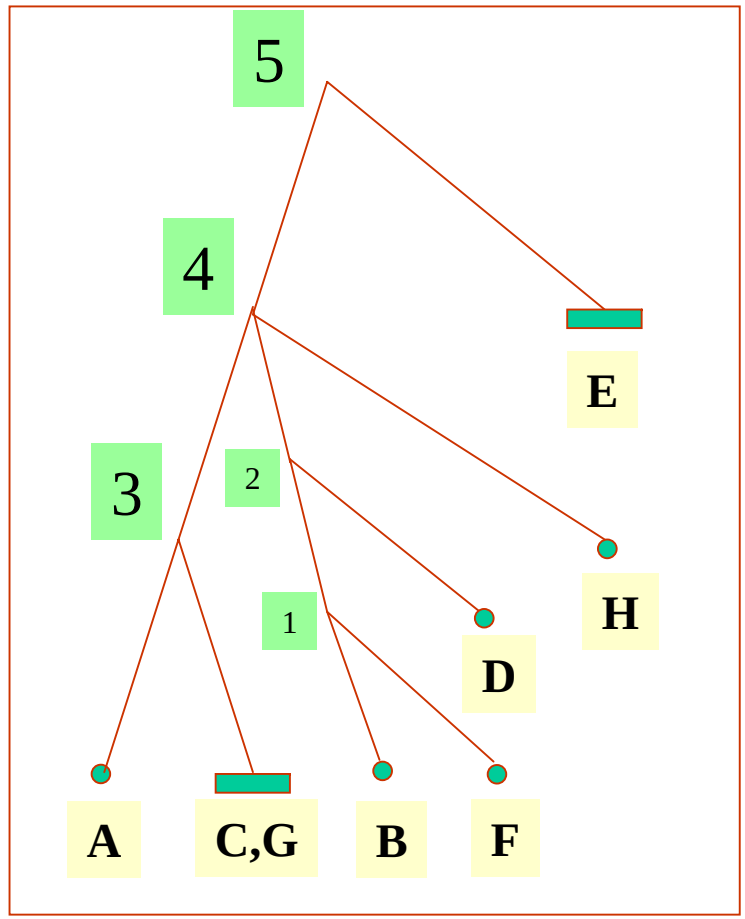
Ultrametric: Example

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



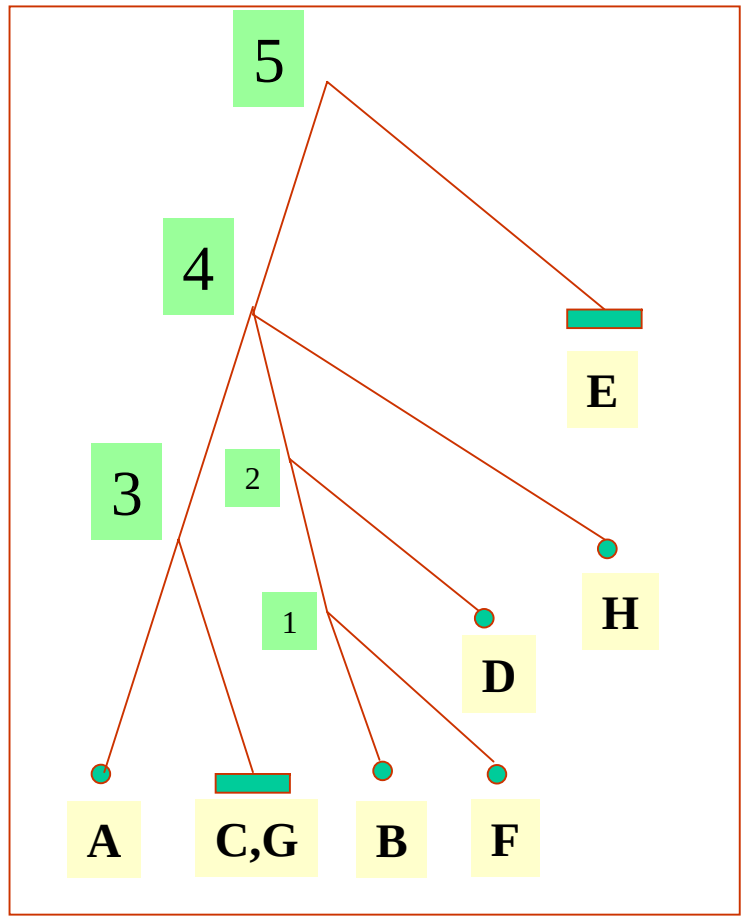
Ultrametric: Example

	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

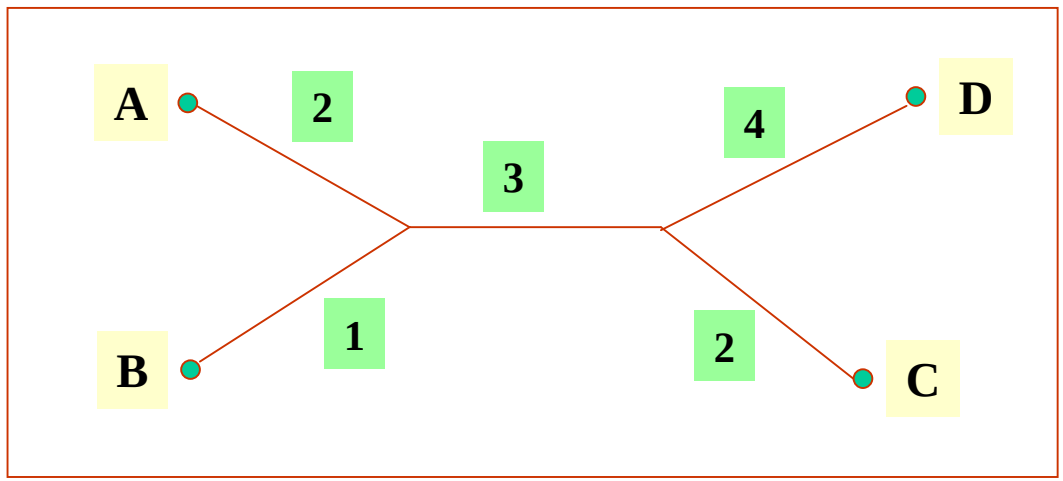
	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



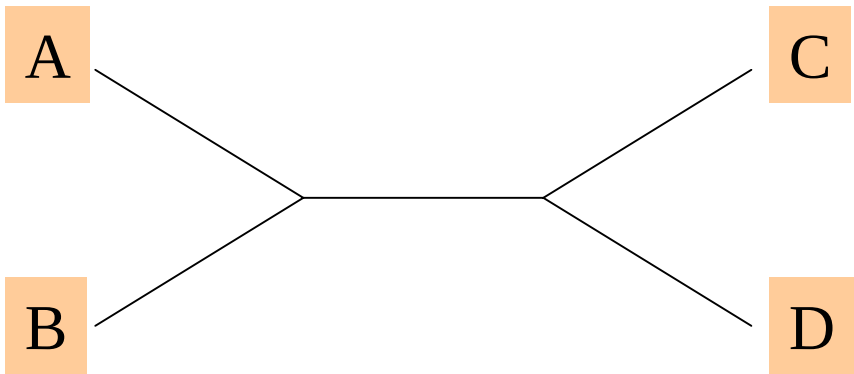
Additive-Distance Trees

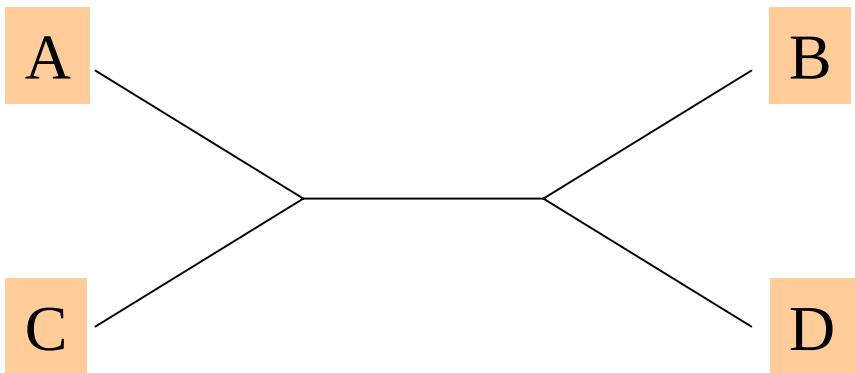
Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

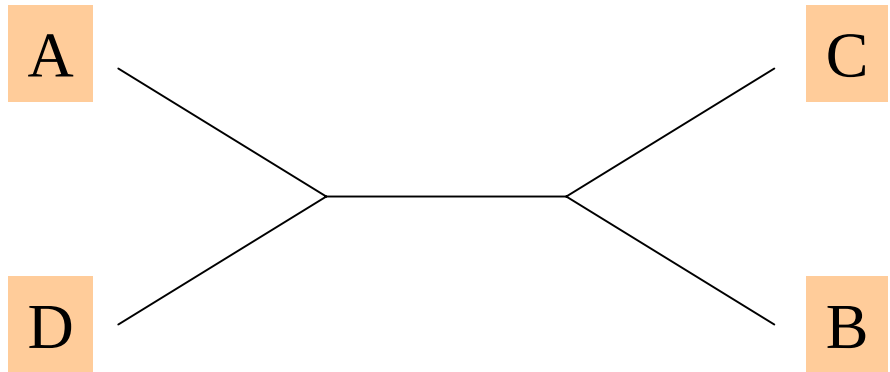
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

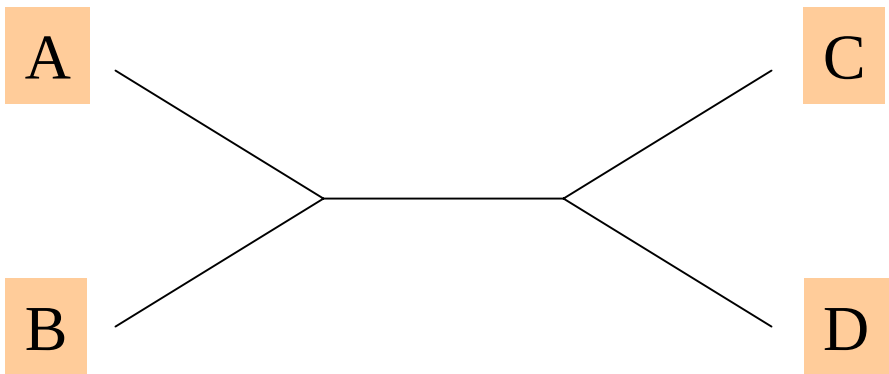






Four-Point Condition

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

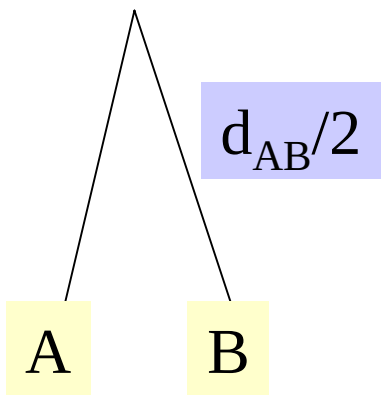


Unweighted pair-group method with arithmetic means (UPGMA)

	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

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

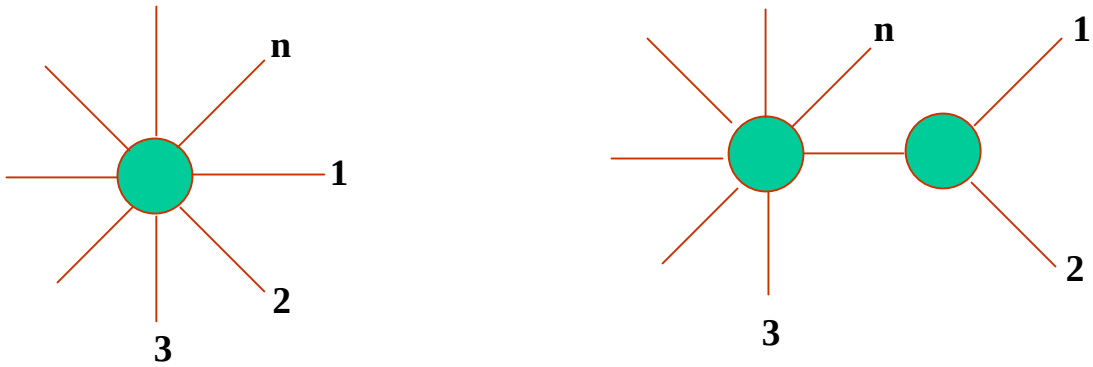
- Now apply UPGMA

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.

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

Neighbor-Joining



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

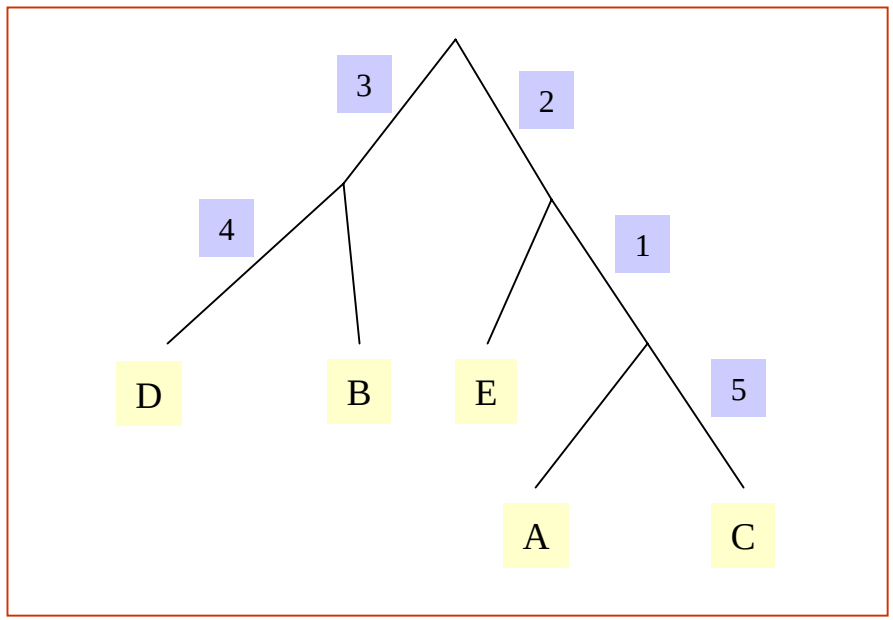
Constructing Evolutionary/Phylogenetic Trees

- 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	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

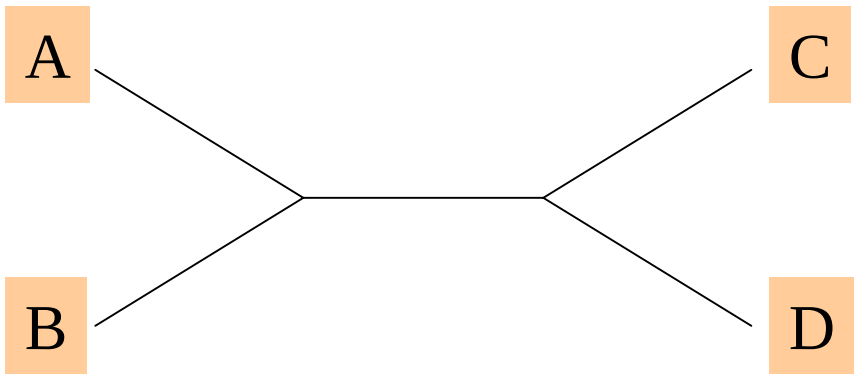
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

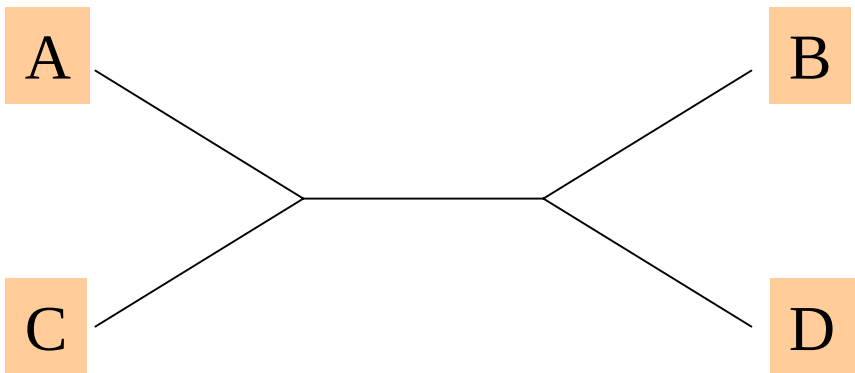
	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

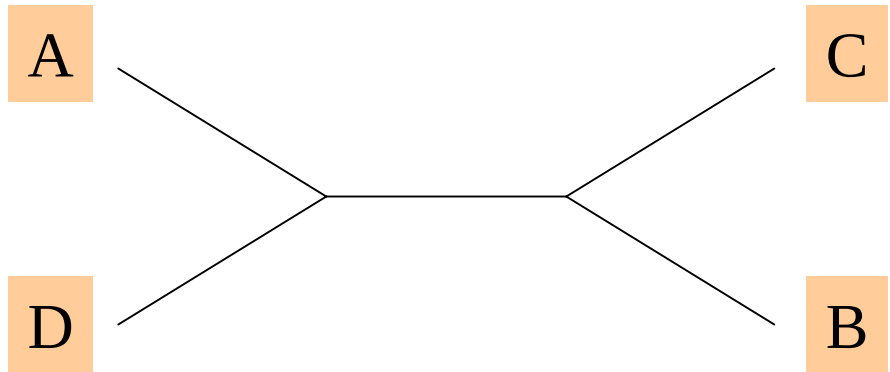
Maximum Parsimony Method: Example

	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







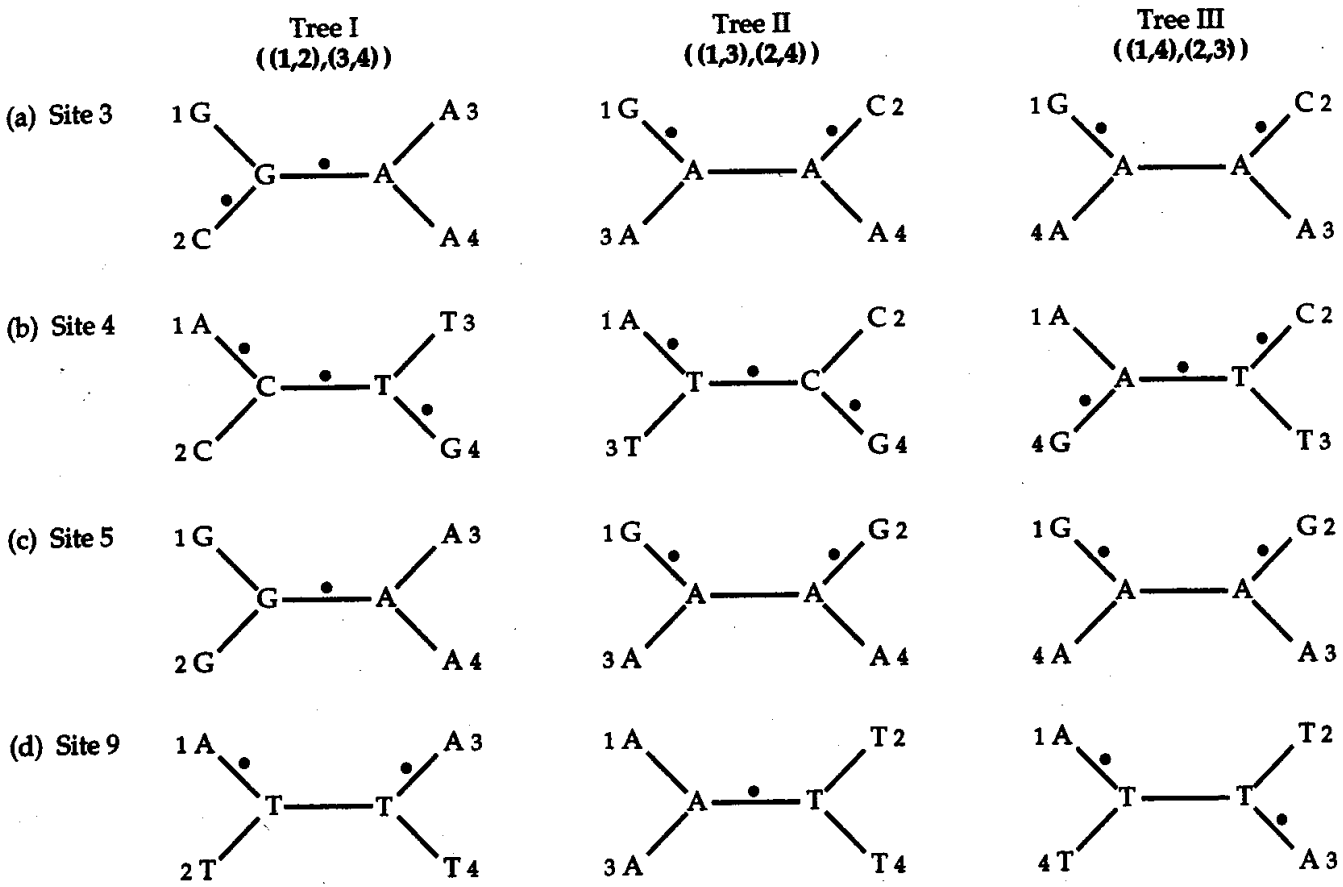


FIGURE 5.14 **Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number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

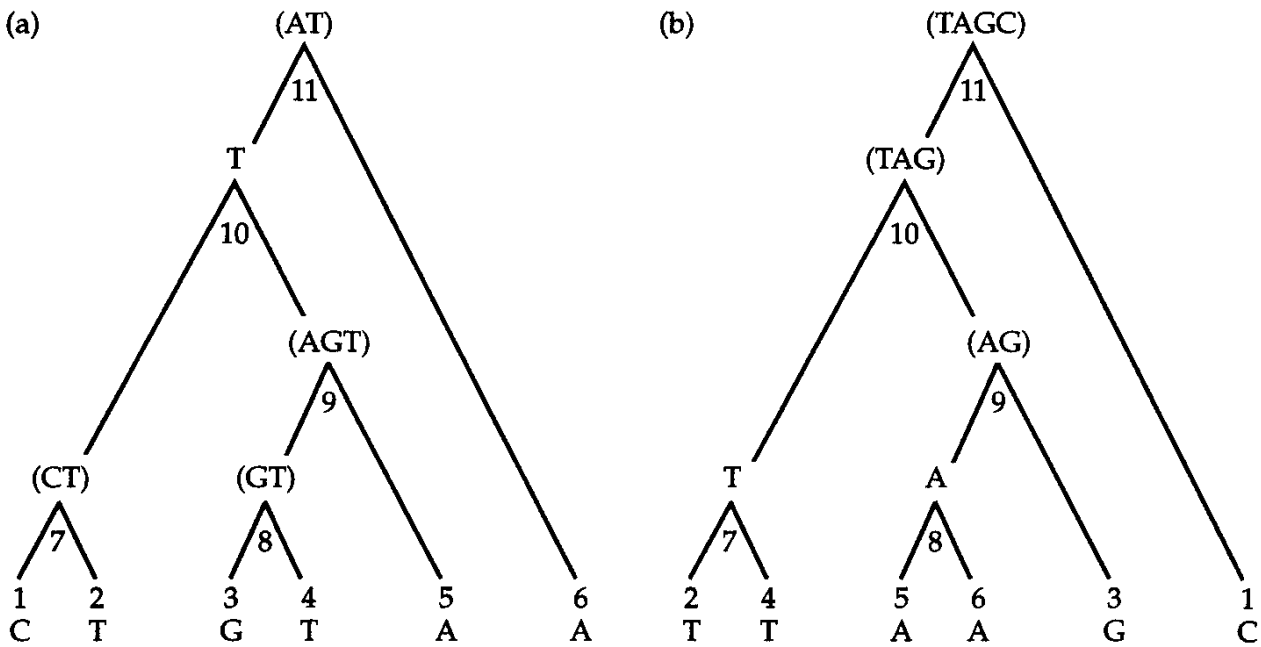


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal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

Searching for the Maximum Parsimony Tree: Exhaustive Search

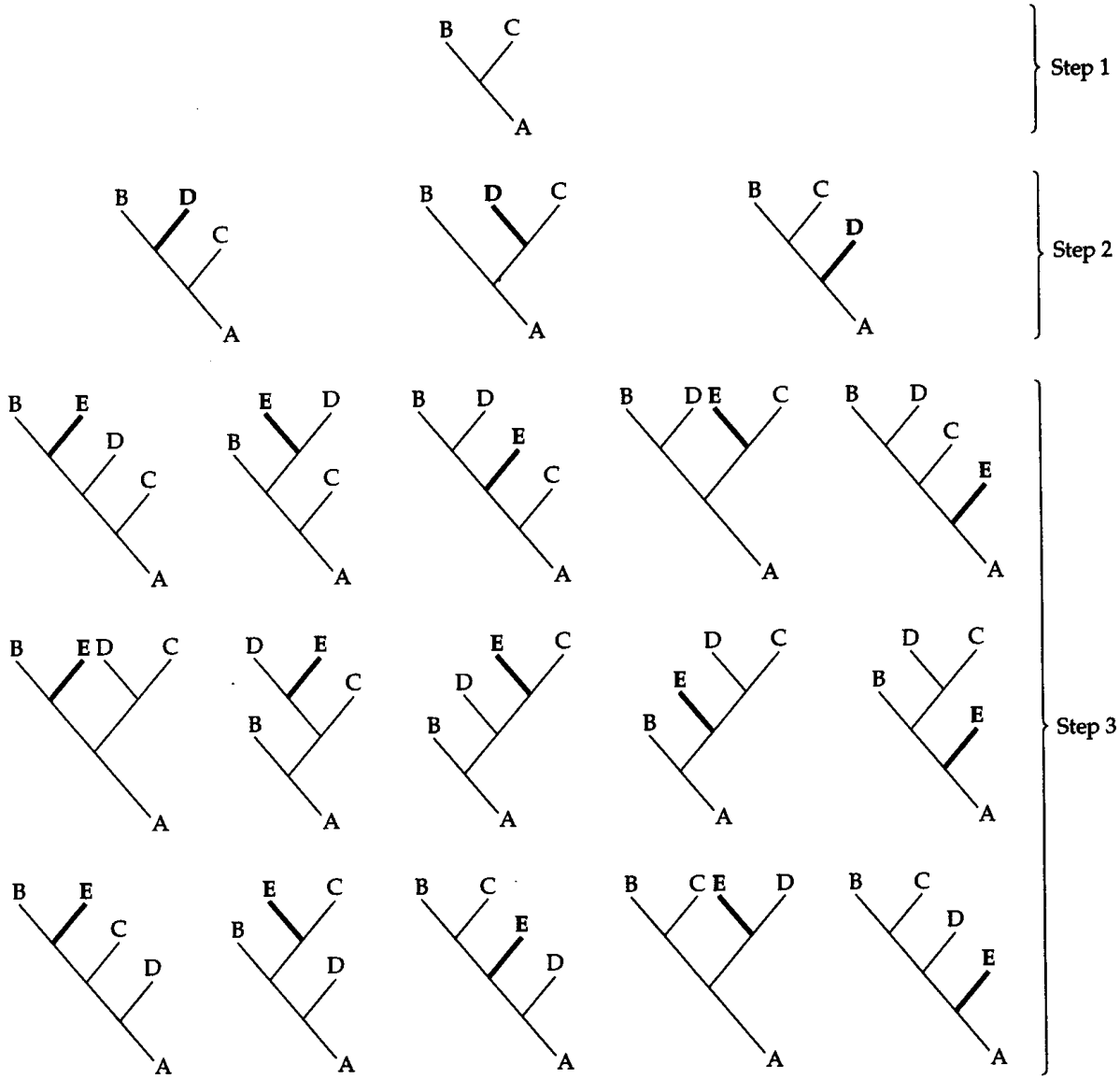
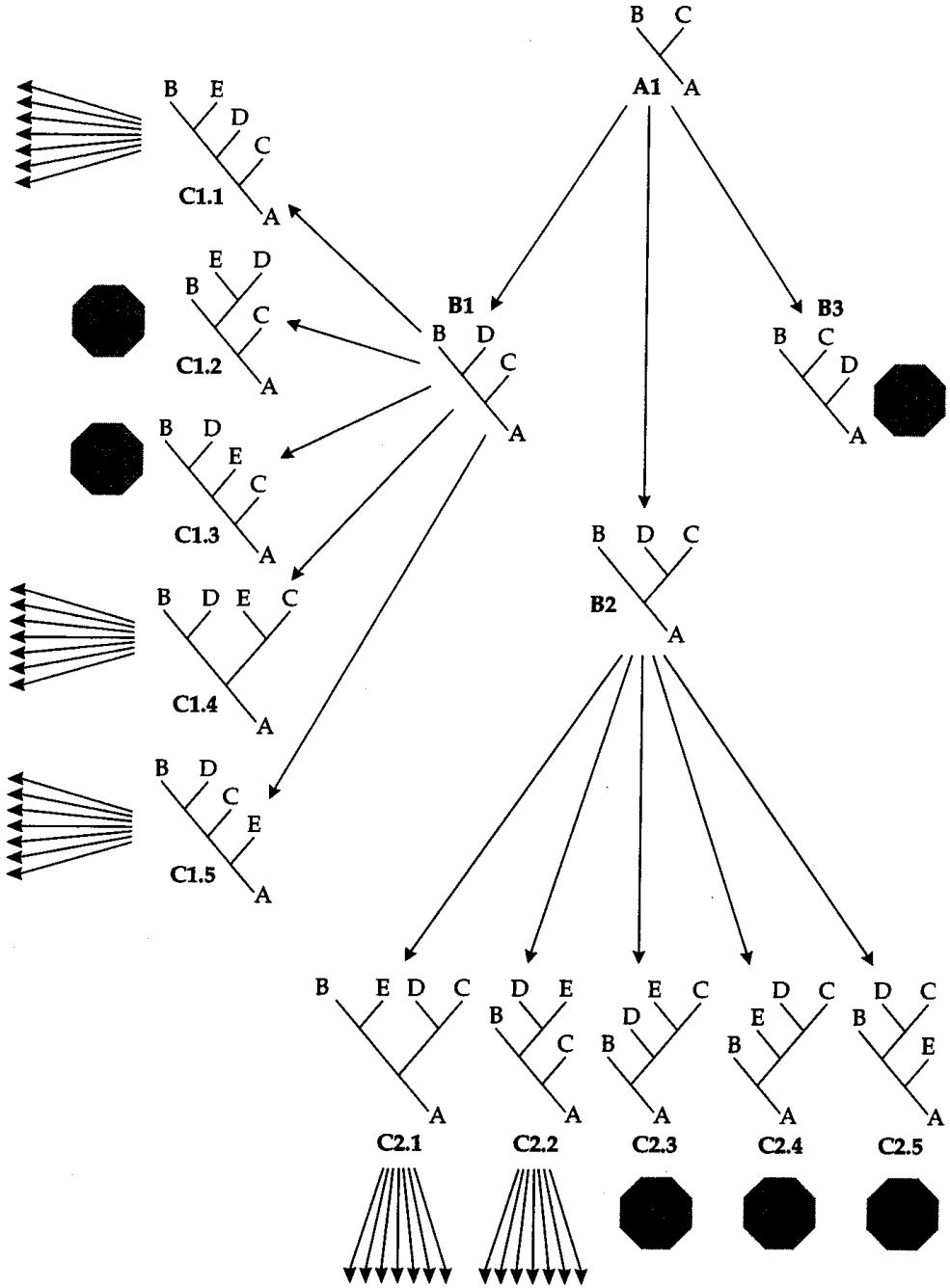


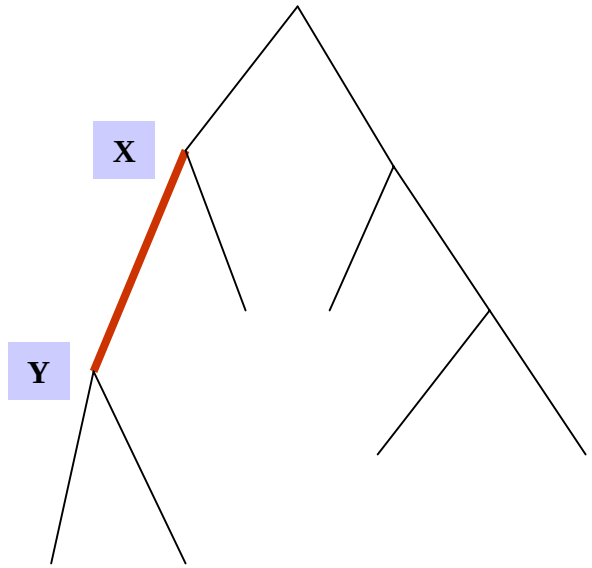
FIGURE 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four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

Computing Maximum Likelihood Tree

(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)

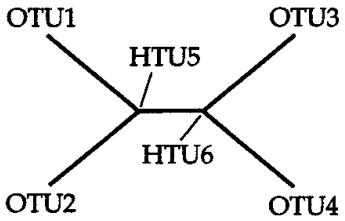
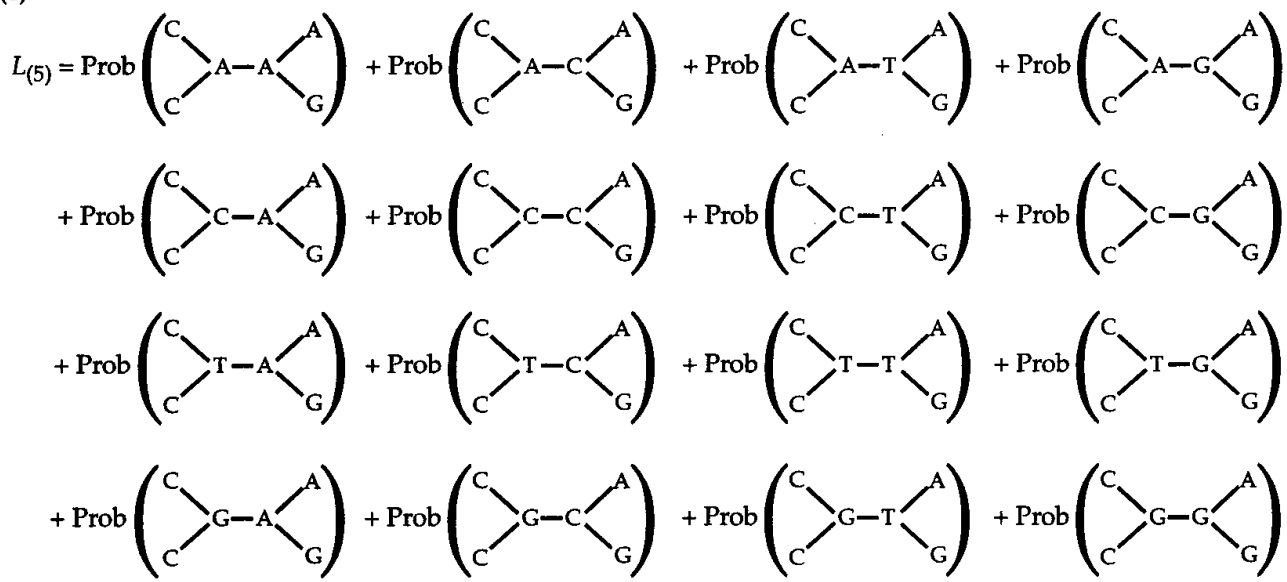


FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length *n*. (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums 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) $$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \ldots \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$$

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

Models of Nucleotide Substitution

Jukes-Cantor

1-3α	α	α	α
α	1-3α	α	α
α	α	1-3α	α
α	α	α	1-3α

Kimura 3ST

1-α-β-γ	α	β	γ
α	1- α- β-γ	γ	β
β	γ	1- α- β-γ	α
γ	β	α	1- α- β-γ

PAM Matrix for Amino Acids

PHYLIP's Character-based Methods

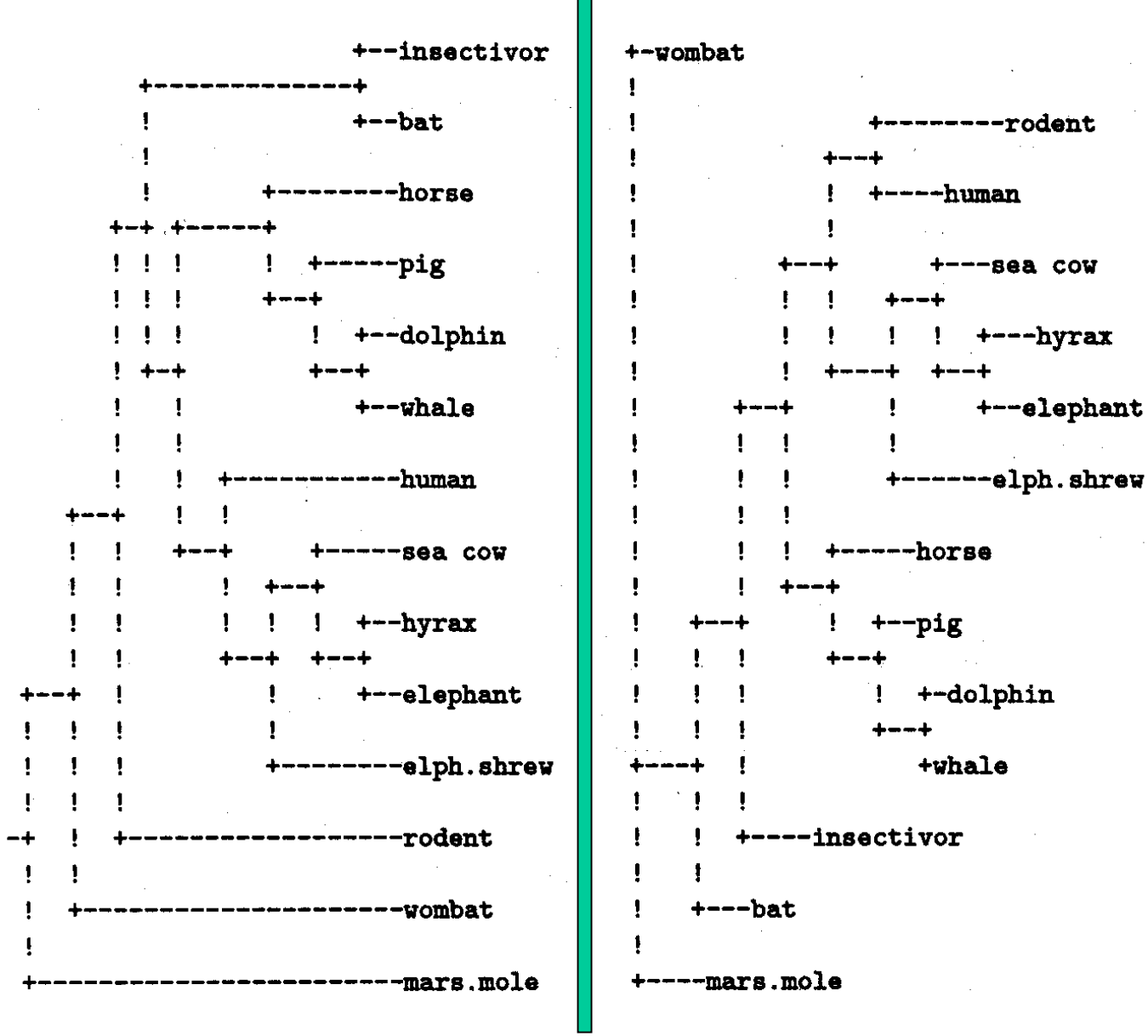


FIGURE 14.3. Parsimony and Maximum Likelihood trees

PHYLIP's Distance-based Methods

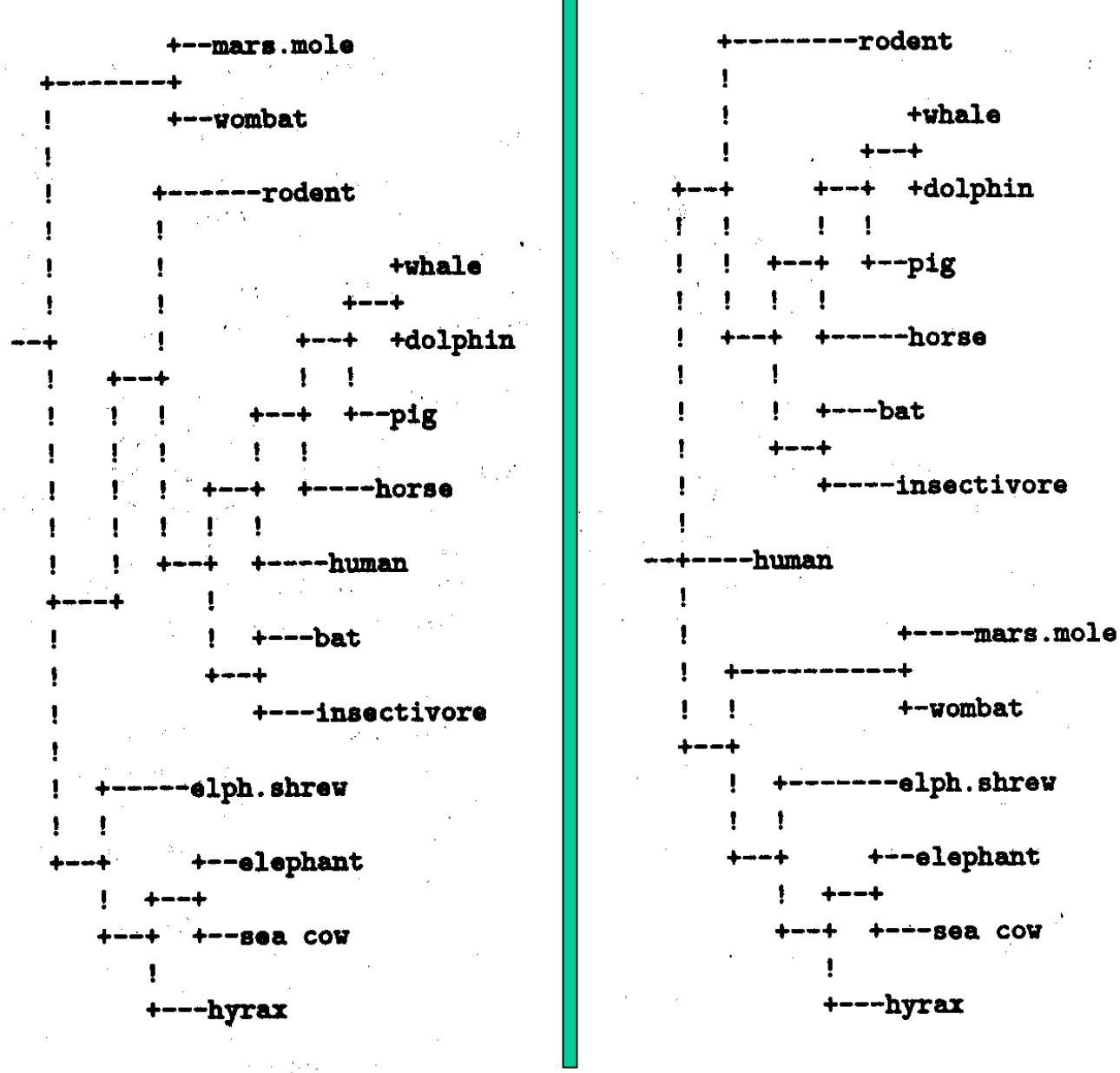


FIGURE 14.2. UPGMA and neighbor-joining trees

Bootstrap: Estimating confidence level of a phylogenetic hypothesis

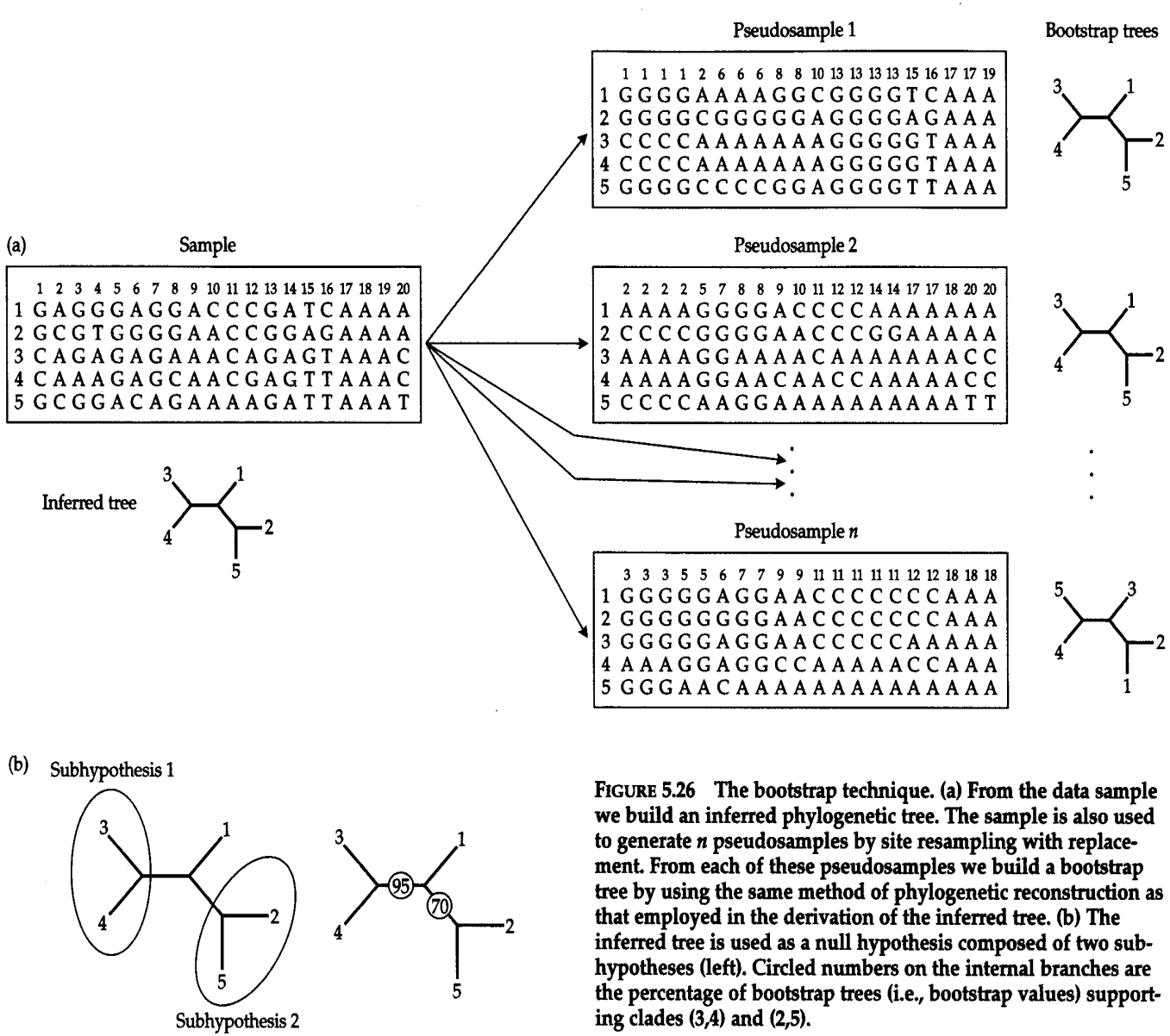


FIGURE 5.26 The bootstrap technique. (a) From the data sample we build an inferred phylogenetic tree. The sample is also used to generate *n* pseudosamples by site resampling with replacement. From each of these pseudosamples we build a bootstrap tree by using the same method of phylogenetic reconstruction as that employed in the derivation of the inferred tree. (b) The inferred tree is used as a null hypothesis composed of two subhypotheses (left). Circled numbers on the internal branches are the percentage of bootstrap trees (i.e., bootstrap values) supporting clades (3,4) and (2,5).

Tree Evaluation Methods

- Bootstrap
- Skewness Test (Randomized Trees)
- Permutation Test (Randomized Characters)
- Parametric bootstrap
- Likelihood Ratio Tests

Computing Evolutionary Relationships: Basic Assumptions

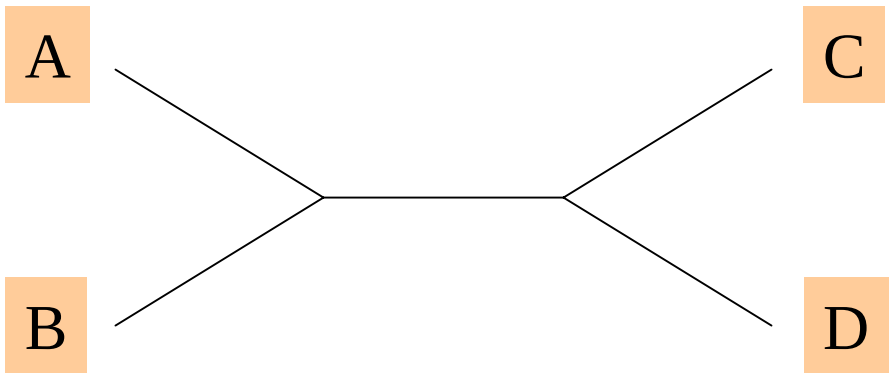
- Evolutionary divergences are strictly bifurcating, i.e., observed data can be represented by a tree. [Exceptions: transfer of genetic material between organisms]
- Sampling of individuals from a group is enough to determine the relationships
- Each individual evolves independently.

Comparisons

- UPGMA (fast)
  - assumes rate constancy; rarely used.
- Additive Methods (fast)
  - If four-point condition is satisfied, works well
  - Quality depends on quality of distance data
  - Does not consider multiple substitutions at site
  - Long sequences & small distances, small errors.

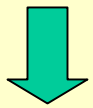
Additive Metrics: Four-Point Condition

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



Ultrametric vs Additive Metrics

- Check whether A is an additive distance matrix.



- Check whether U is an ultrametric matrix

COMMENT: This is just a reduction. It does not mean that one can build a tree for the same data!

Comparisons

- Maximum Parsimony
  - Character-based methods trusted more.
  - MP assumes “simplest explanation is always the best”!
  - No explicit assumptions, except that a tree that requires fewer substitutions is better
  - Faster evolution on long branches may give rise to homoplasies, and MP may go wrong.
  - Performance depends on the number of informative sites, and is usually not so good with finding clades.
  - If more than one tree, builds consensus trees.

Comparisons

- Maximum Likelihood
  - Uses all sites, unlike MP method.
  - Makes assumptions on the rate and pattern of substitution.
  - Relatively insensitive to violations of assumptions
  - Not very robust (if some sequences very divergent).
  - SLOW! Computationally intensive. Optimum usually cannot be found, since the search space is too large.
  - Fast heuristics exist.
  - Simulations show that it's better than MP and ME.

Phylogenetic Software

- PHYLIP, WEBPHYLIP, PhyloBLAST (large set of programs, command-line)
- PAUP (point-&-click) (MP-based)
- PUZZLE, TREE-PUZZLE, PAML, MOLPHY (ML-based)
- MrBAYES (Bayesian Methods)
- MACCLADE (MP?)
- LAMARC

Recipes to minimize errors in phylogenetic analysis

- Use large amounts of data. Randomize order, if needed.
- Exclude unreliable data (for e.g., when alignment is not known for sure)
- Exclude fast-evolving sequences or sites (3^{rd} codon positions), or only use it for close relationships.
- Most methods will incorrectly group sequences with similar base composition (Additive methods are robust in the presence of such sequences)
- Check validity of “independence” assumptions (e.g. changes on either side of a hairpin structure)
- Use all methods to look at data.

Alignments

- Inputs for phylogenetic analysis usually is a multiple sequence alignment.
- Programs such as CLUSTALW, produce good alignments, but not good trees.
- Aligning according to secondary or tertiary trees are better for phylogenetic analysis.
- Which alignment method is better for which phylogenetic analysis method? OPEN!

Other Heuristics

- Branch Swapping to modify existing trees
- Quartet Puzzling: rapid tree searching