

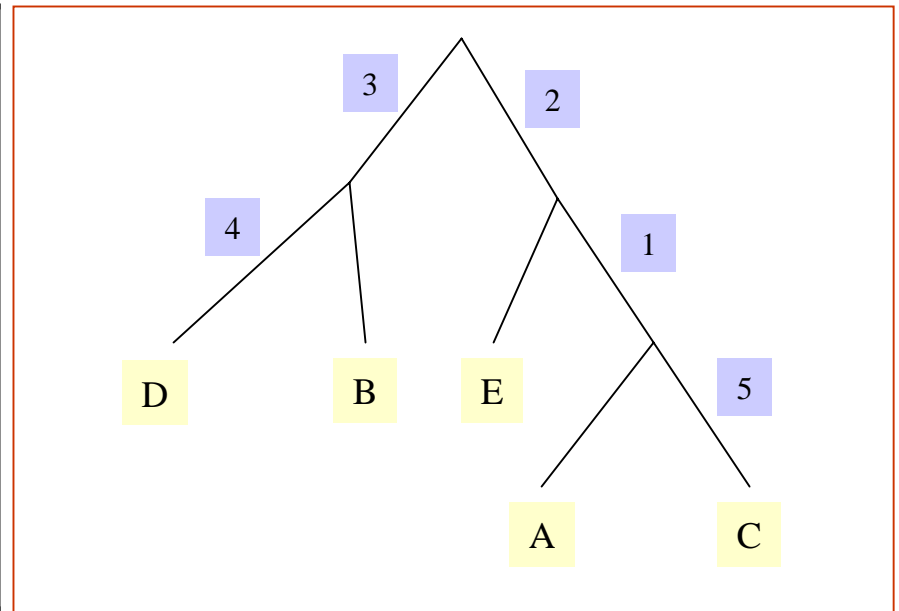
# Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

# Character-based Methods

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

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



# Example

- **Perfect phylogeny** does not always exist.

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

# Maximum Parsimony

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

# Examples of Character Data

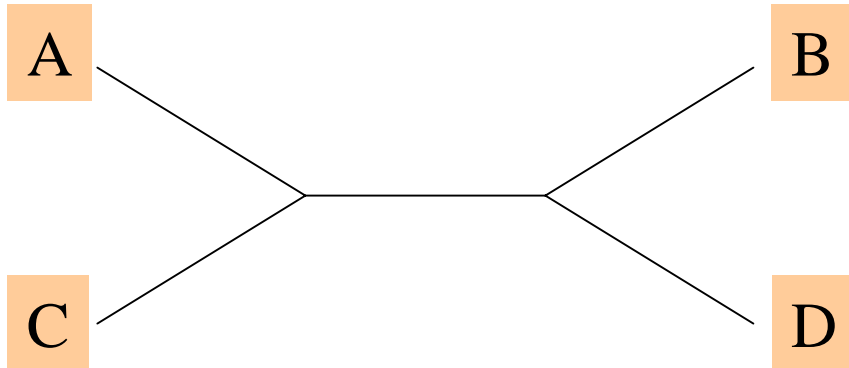
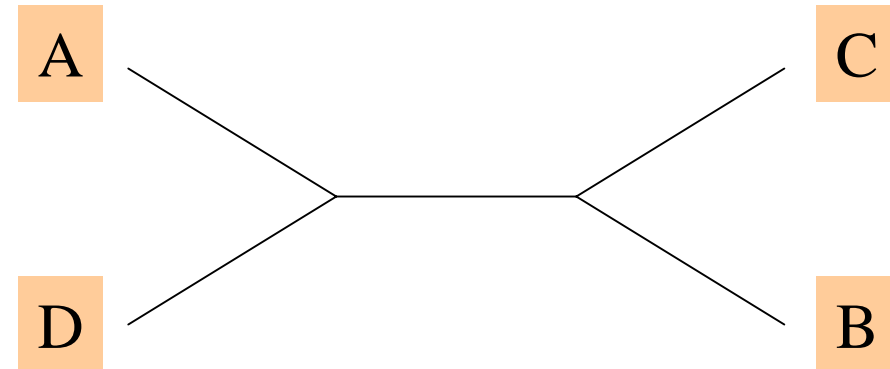
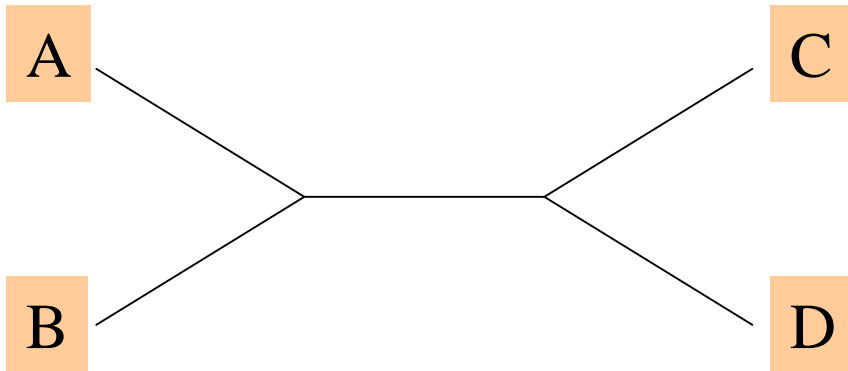
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

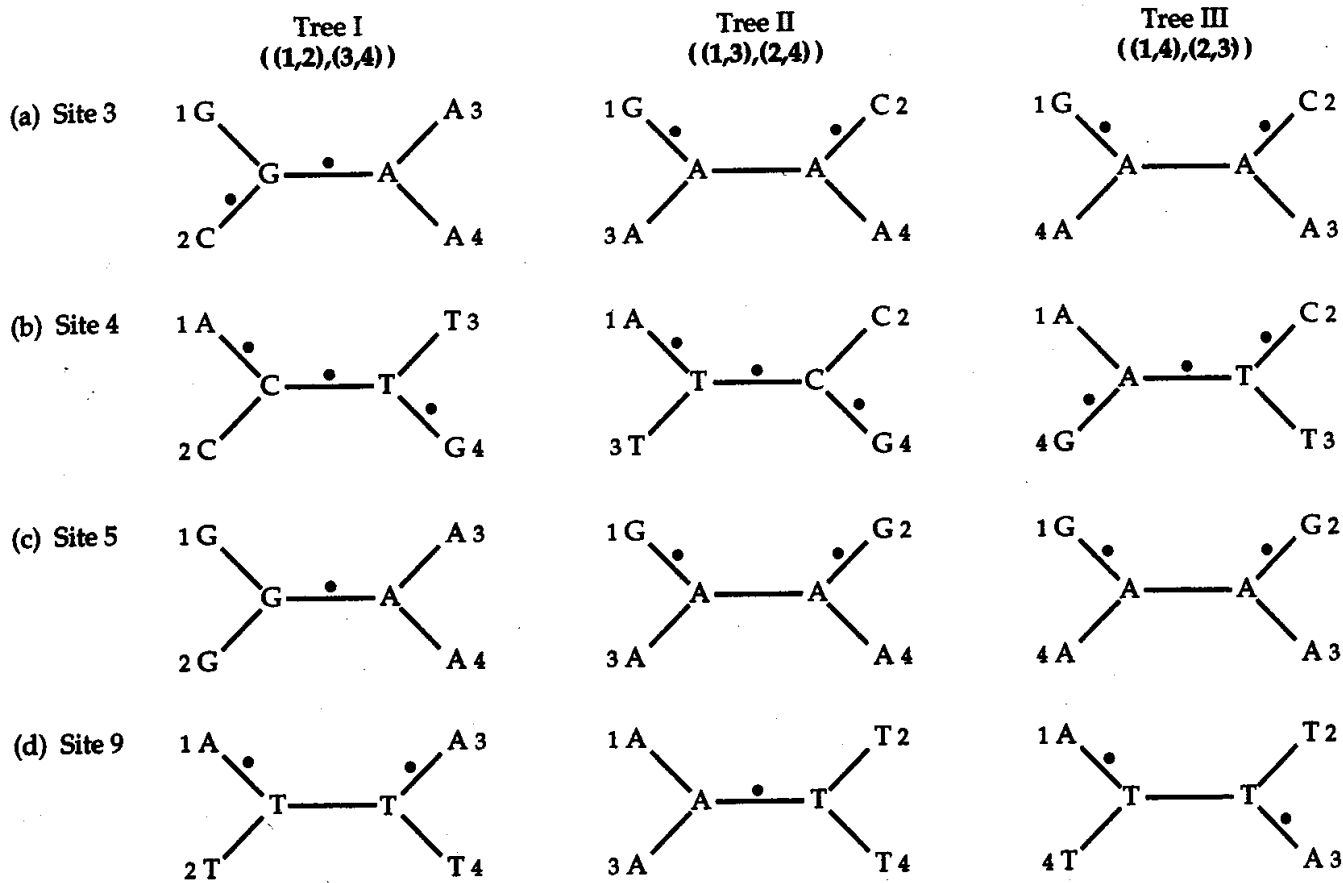
	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

# Maximum Parsimony Method: Example

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

# Unrooted Trees on 4 Taxa



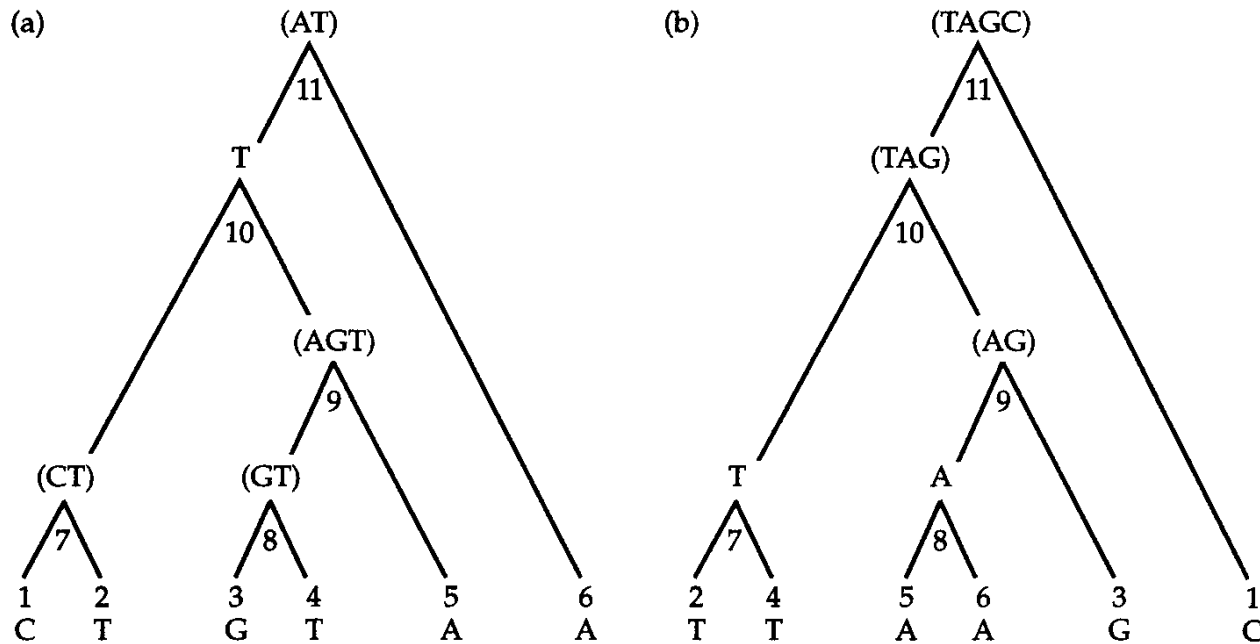


**FIGURE 5.14** Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number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

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

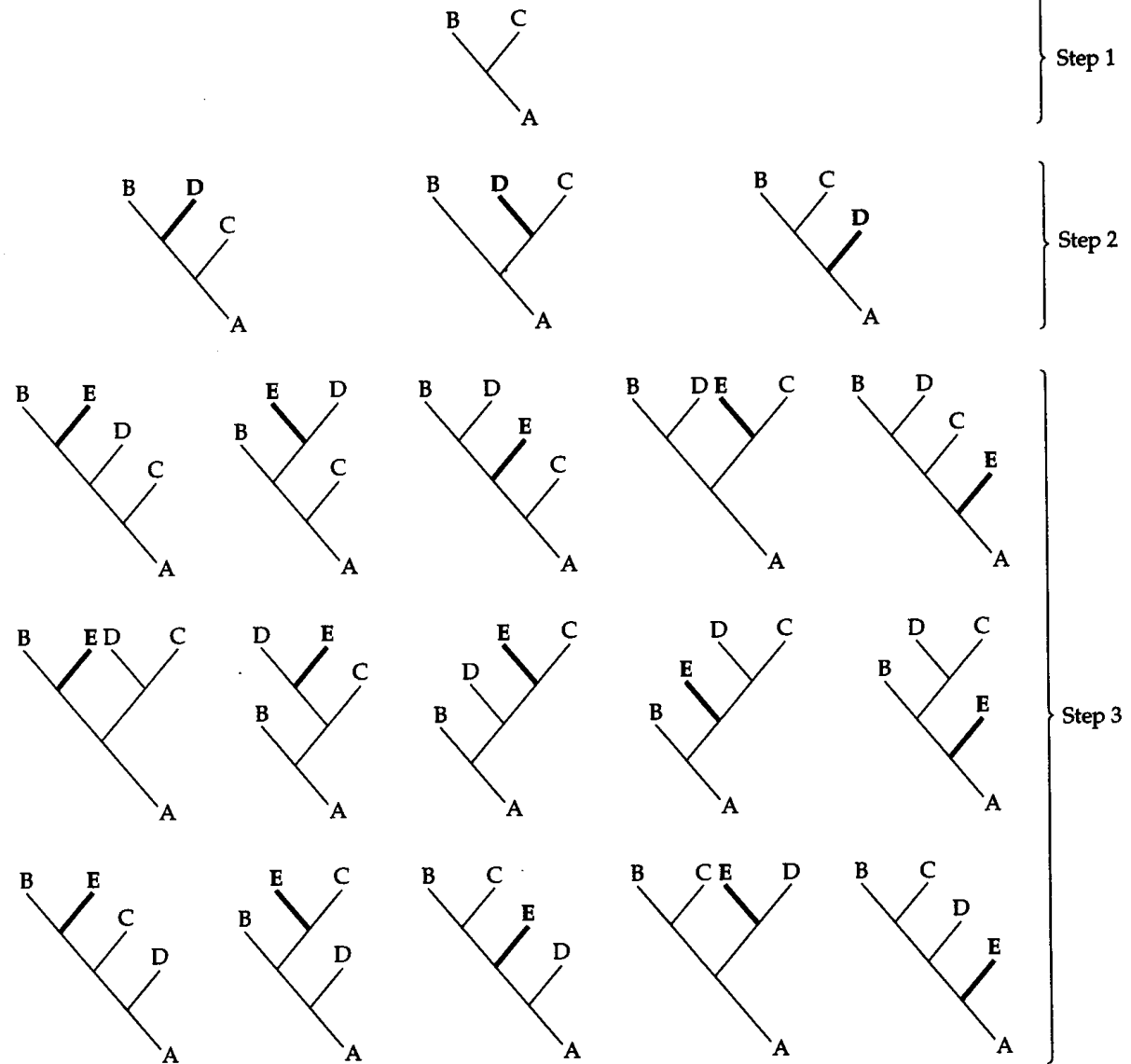


# Inferring nucleotides on internal nodes

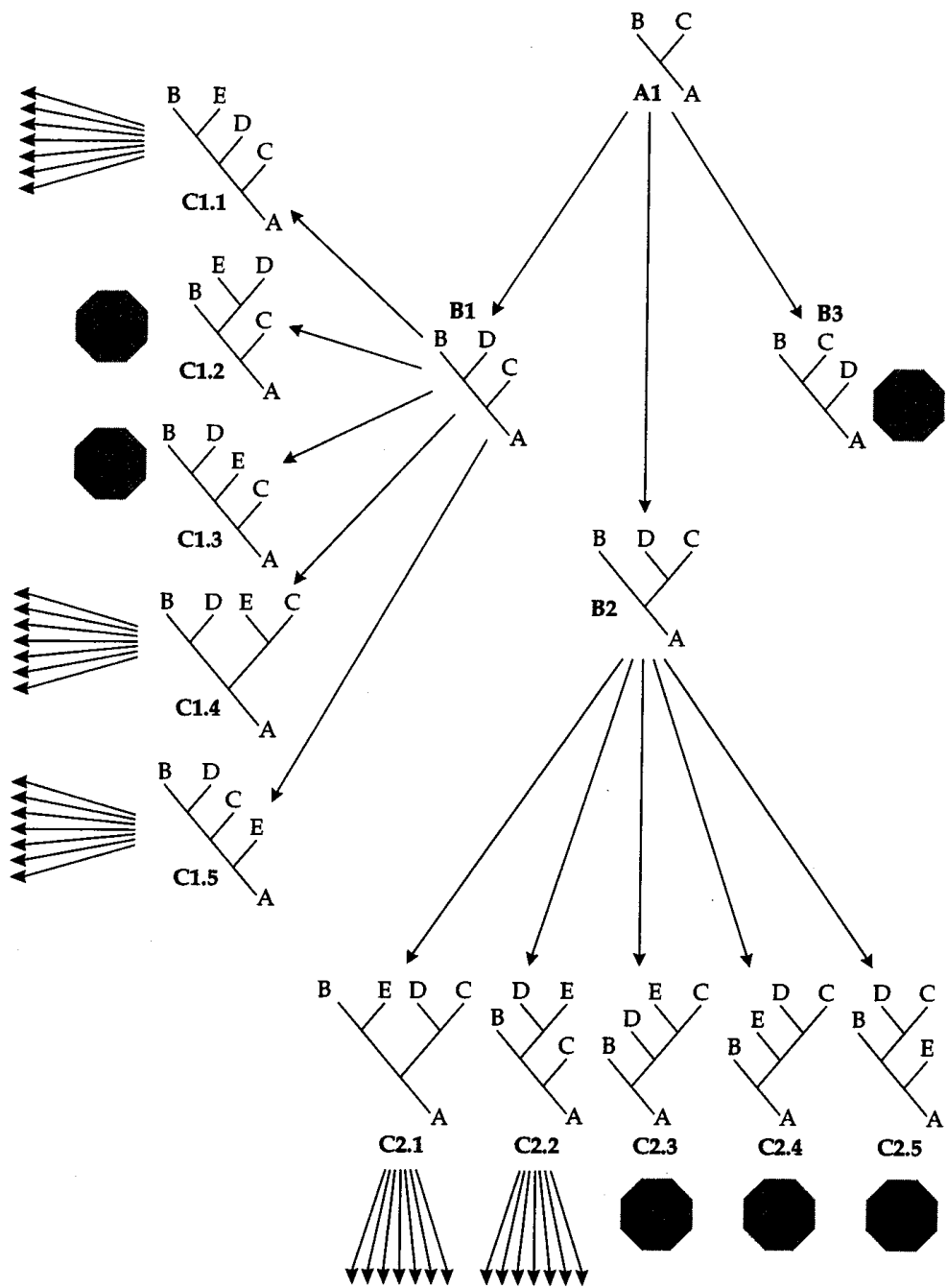


**FIGURE 5.15** Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an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

# Searching for the Maximum Parsimony Tree: Exhaustive Search



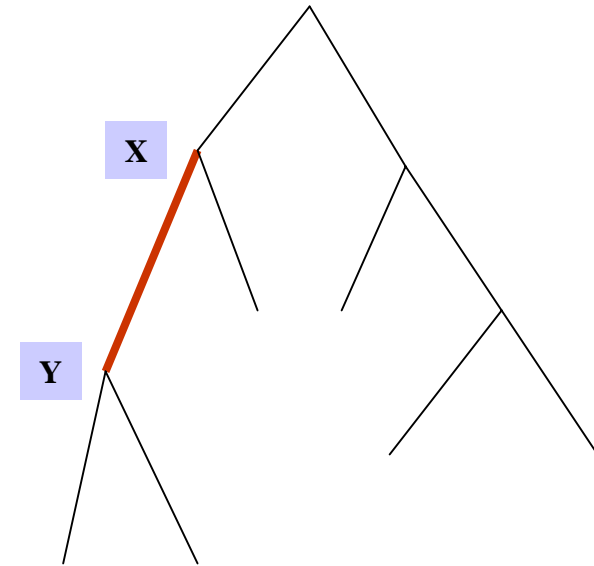
**FIGURE 5.16** Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for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



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

# Probabilistic Models of Evolution

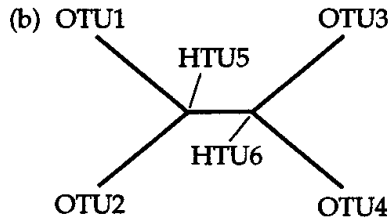
- Assuming a **model of substitution**,
  - $\Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$ ,
- Using this formula it is possible to compute the likelihood that data  $D$  is generated by a given phylogenetic tree  $T$  under a model of substitution. Now find the tree with the maximum likelihood.



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

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

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



(c)

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

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

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

# Computing Maximum Likelihood Tree

# Models of Nucleotide Substitution

**Jukes-Cantor**

$1-3\alpha$	$\alpha$	$\alpha$	$\alpha$
$\alpha$	$1-3\alpha$	$\alpha$	$\alpha$
$\alpha$	$\alpha$	$1-3\alpha$	$\alpha$
$\alpha$	$\alpha$	$\alpha$	$1-3\alpha$

**Kimura 3ST**

$1-\alpha-\beta-\gamma$	$\alpha$	$\beta$	$\gamma$
$\alpha$	$1-\alpha-\beta-\gamma$	$\gamma$	$\beta$
$\beta$	$\gamma$	$1-\alpha-\beta-\gamma$	$\alpha$
$\gamma$	$\beta$	$\alpha$	$1-\alpha-\beta-\gamma$

**PAM Matrix for Amino Acids**

# PHYLIP's Character- based Methods

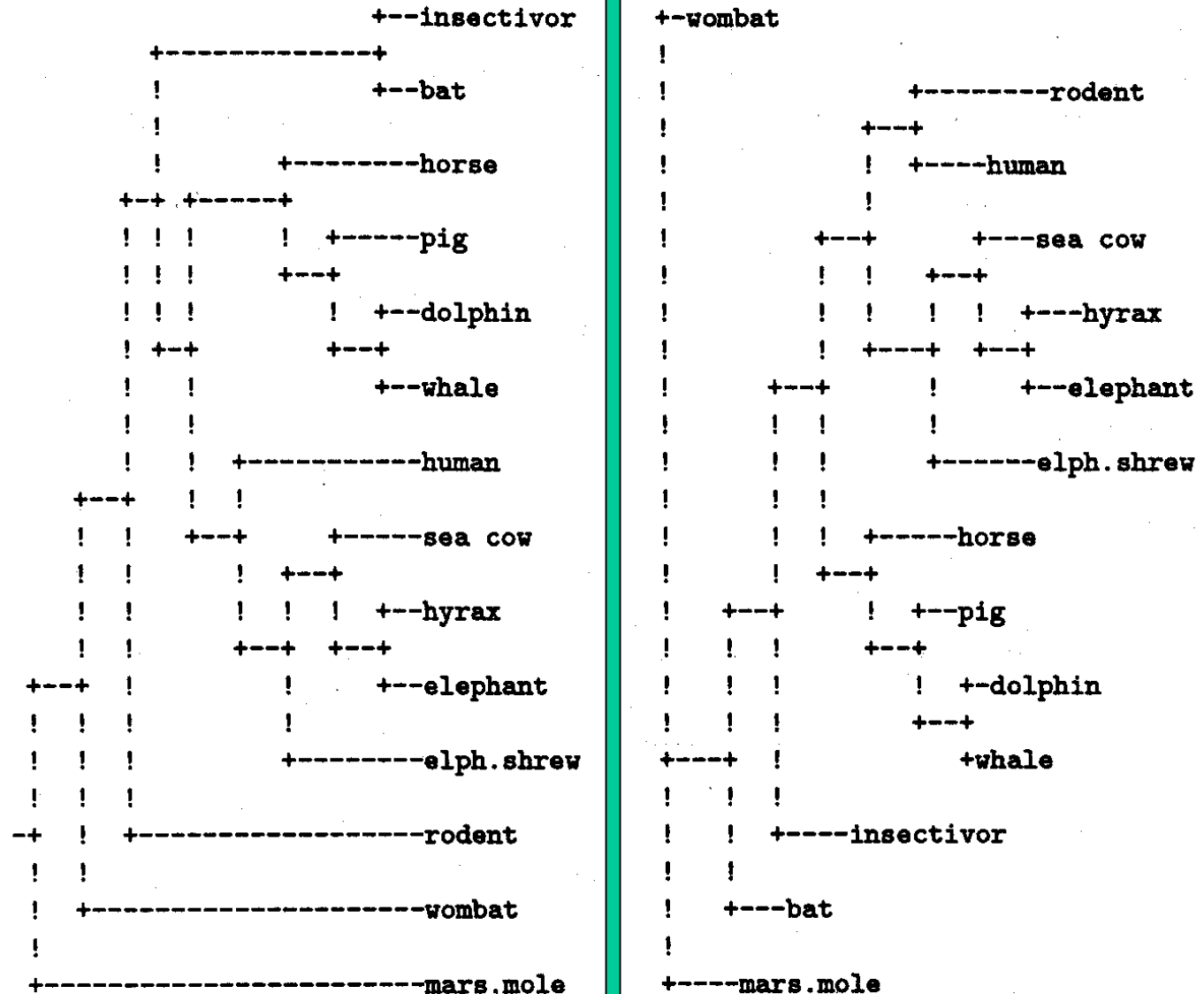


FIGURE 14.3. Parsimony and Maximum Likelihood trees

# PHYLIP's Distance- based Methods

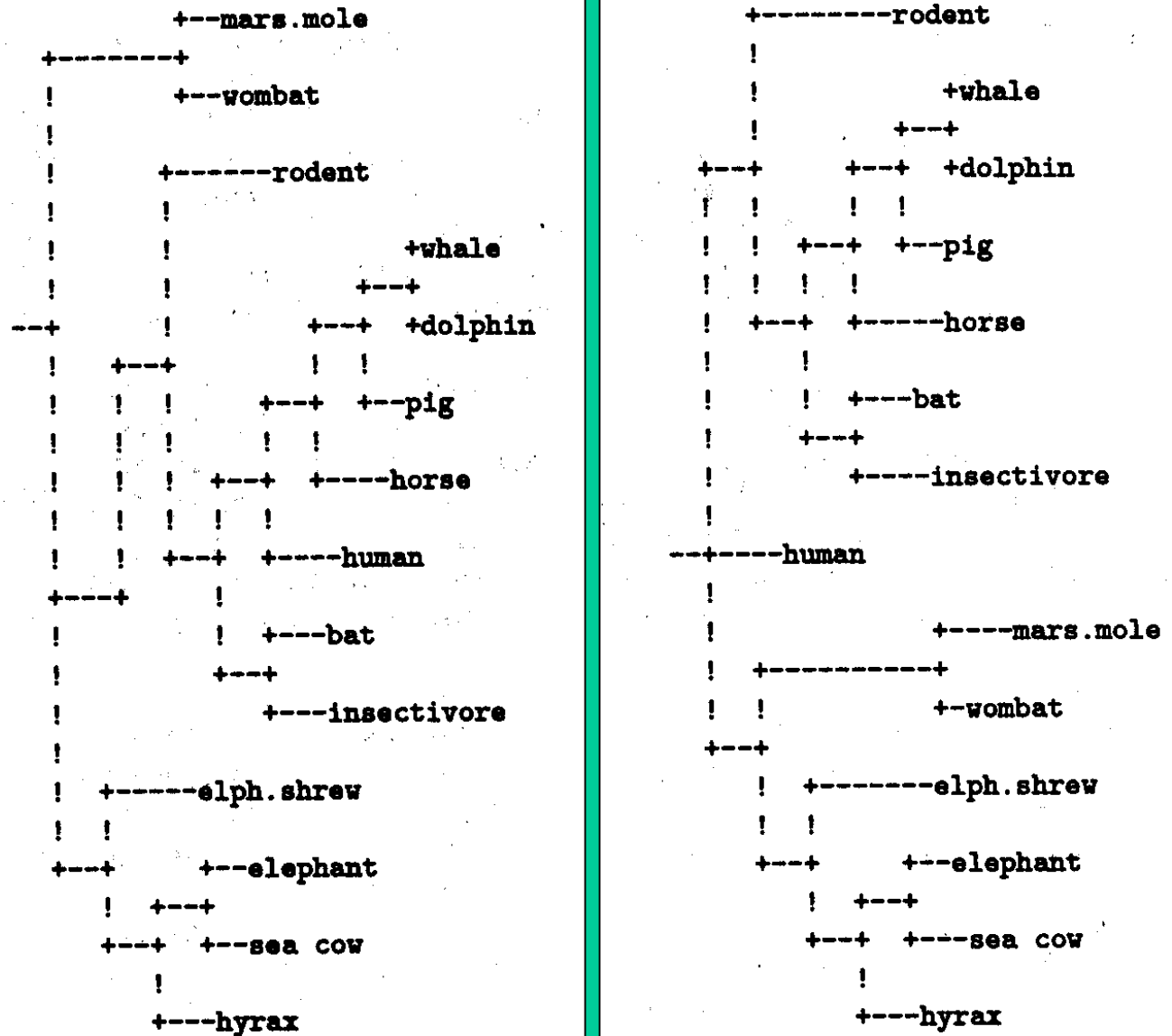


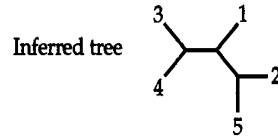
FIGURE 14.2. UPGMA and neighbor-joining trees



# Bootstrap: Estimating confidence level of a phylogenetic hypothesis

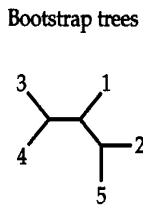
(a) Sample

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	G	A	G	G	G	A	G	G	A	C	C	C	G	A	T	C	A	A	A	A	A
2	G	C	G	T	G	G	G	G	A	A	C	C	G	G	A	G	A	A	A	A	A
3	C	A	G	A	G	A	G	A	A	A	A	C	A	G	A	G	T	A	A	A	C
4	C	A	A	A	G	A	G	C	A	A	C	G	A	G	T	T	A	A	A	A	C
5	G	C	G	G	A	C	A	G	A	A	A	A	G	A	T	T	A	A	A	A	T



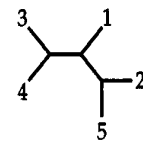
Pseudosample 1

	1	1	1	1	2	6	6	6	8	8	10	13	13	13	13	15	16	17	17	19	
1	G	G	G	G	A	A	A	A	A	G	G	C	G	G	G	G	T	C	A	A	A
2	G	G	G	G	C	G	G	G	G	A	G	G	G	G	A	G	A	A	A	A	A
3	C	C	C	C	A	A	A	A	A	A	A	A	A	A	A	A	G	G	G	G	T
4	C	C	C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
5	G	G	G	G	C	C	C	C	G	A	G	G	G	G	T	A	A	A	A	A	A



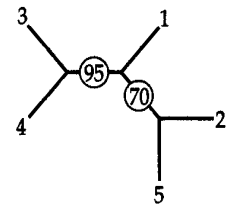
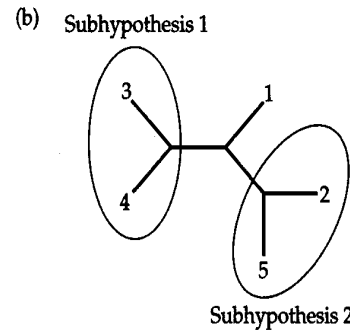
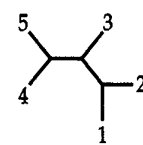
Pseudosample 2

	2	2	2	2	5	7	8	8	9	10	11	12	12	14	14	17	17	18	20	20	
1	A	A	A	A	A	G	G	G	G	A	C	C	C	C	A	A	A	A	A	A	A
2	C	C	C	C	G	G	G	G	A	A	C	C	C	G	G	A	A	A	A	A	A
3	A	A	A	A	A	G	A	A	A	A	C	A	A	A	A	A	A	A	A	A	A
4	A	A	A	A	G	G	A	A	C	A	A	C	C	A	A	A	A	A	A	A	C
5	C	C	C	C	A	A	G	G	A	A	A	A	A	A	A	A	A	A	A	A	A



Pseudosample n

	3	3	3	5	5	6	7	7	9	9	11	11	11	11	11	12	12	18	18	18	
1	G	G	G	G	G	A	G	G	A	A	C	C	C	C	C	C	C	C	A	A	A
2	G	G	G	G	G	G	G	G	A	A	C	C	C	C	C	C	C	C	A	A	A
3	G	G	G	G	G	A	G	G	A	A	C	C	C	C	C	A	A	A	A	A	A
4	A	A	A	G	G	A	G	G	C	C	A	A	A	A	A	A	C	C	A	A	A
5	G	G	G	A	A	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A



**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

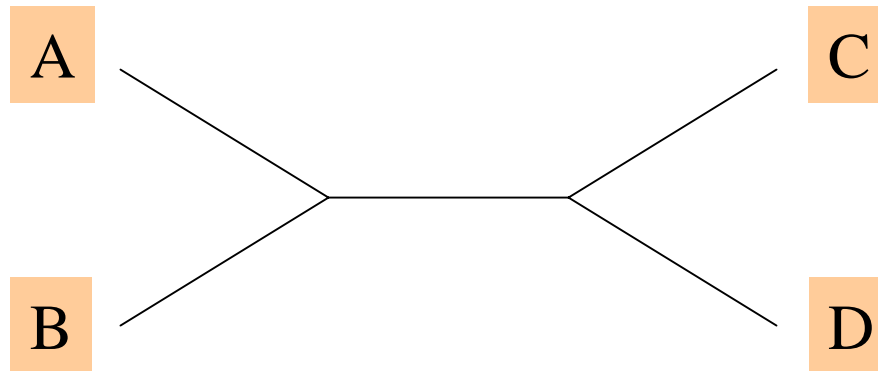
- Evolutionary divergences are strictly bifurcating, i.e., observed data can be represented by a tree. [Exceptions: transfer of genetic material between organisms]
- Sampling of individuals from a group is enough to determine the relationships
- Each individual evolves independently.

# 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.

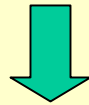
# Additive Metrics: Four-Point Condition

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



# 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<sup>rd</sup> 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