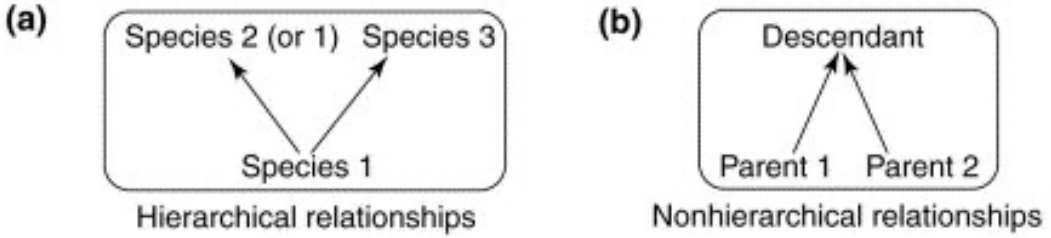
# Bioinformatics and Phylogenetics Fall 2016

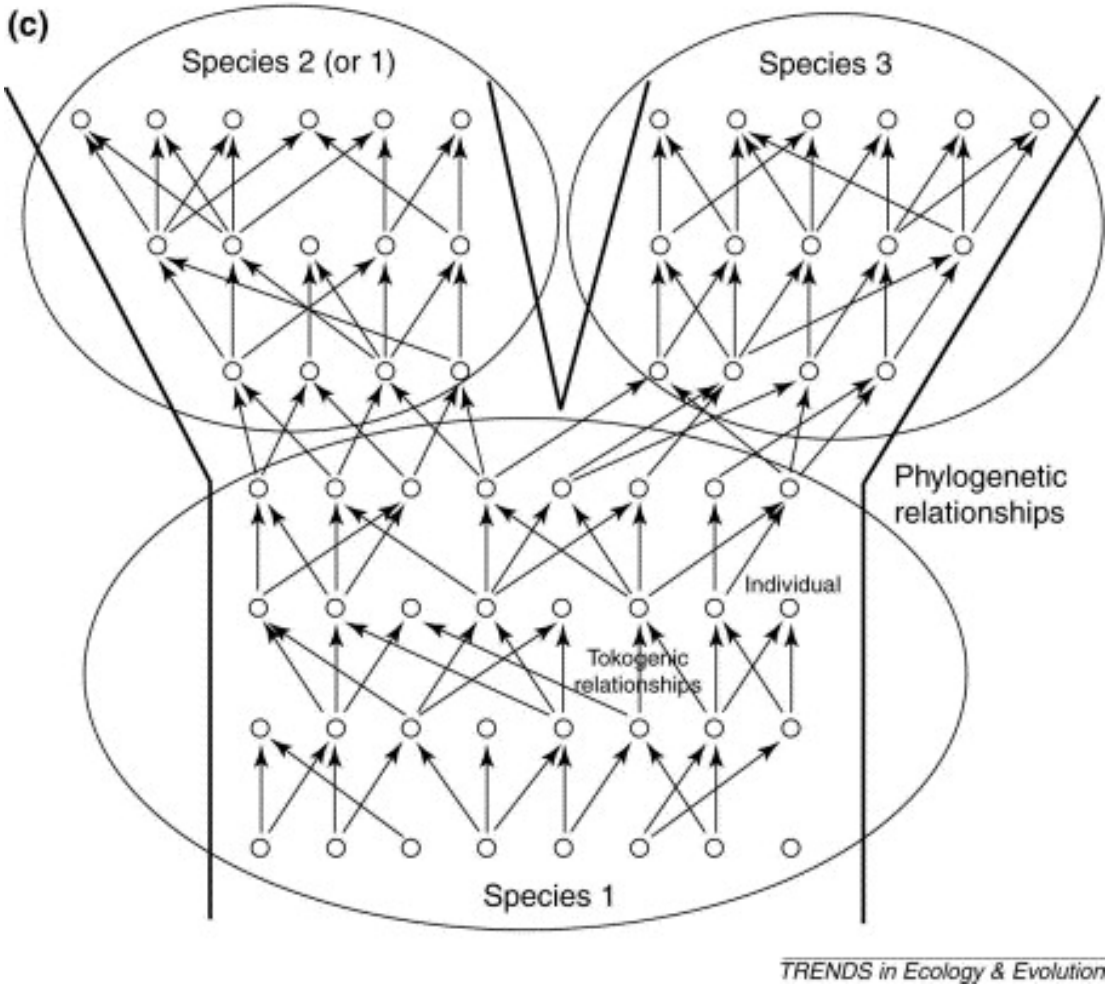
## Phylogenetic Methods Review

# Phylogenetic Trees

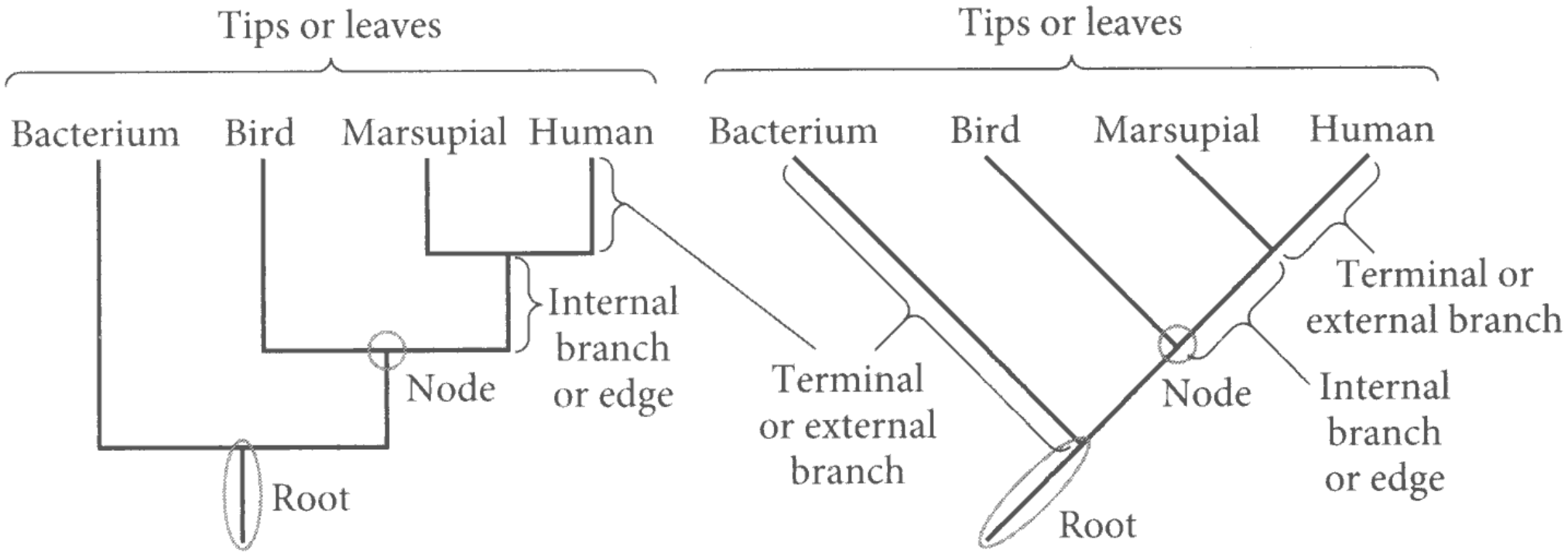



**Cladogenesis**:
Parent species splits into two

**Anagenesis**:
Transformation within a species



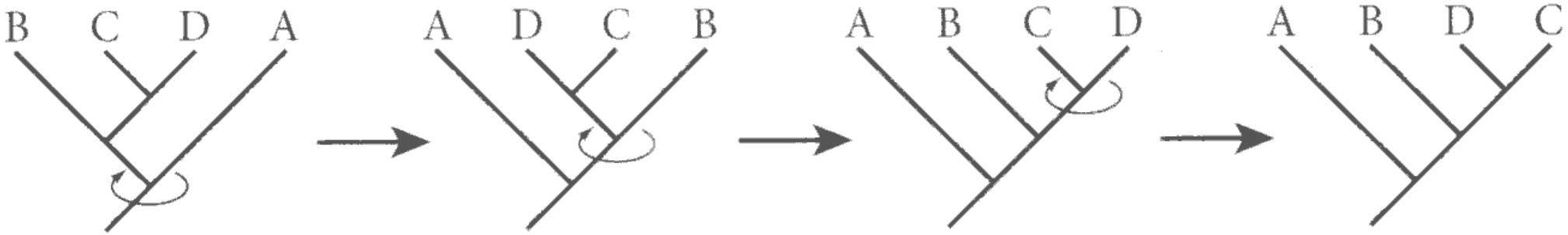
# Phylogenetic Trees



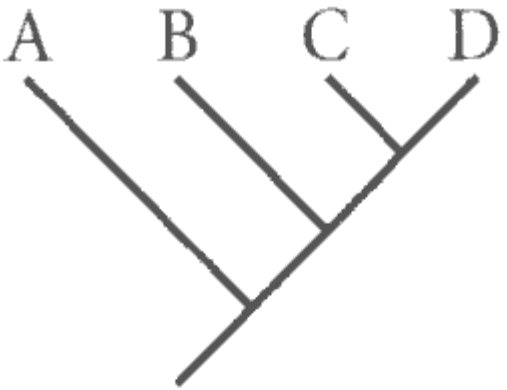
# Phylogenetic Trees

Rearranging phylogenetic trees:

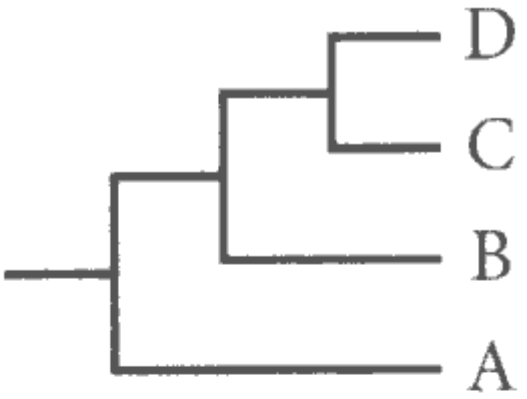
**Are the relationships the same or different?**



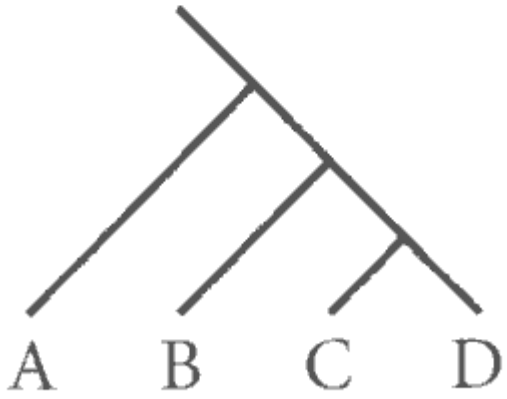
# Phylogenetic Trees



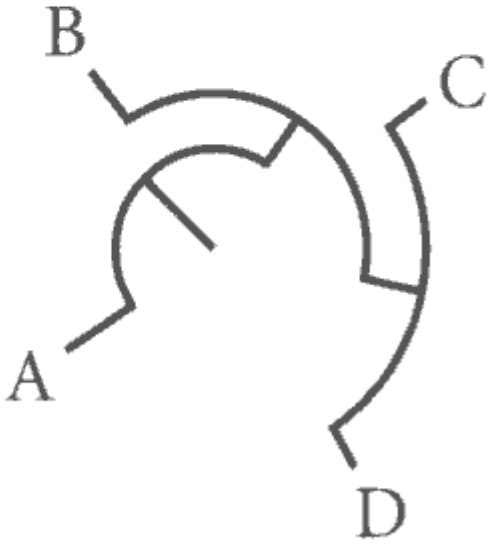

Diagonal-up




Rectangular-right




Diagonal-down

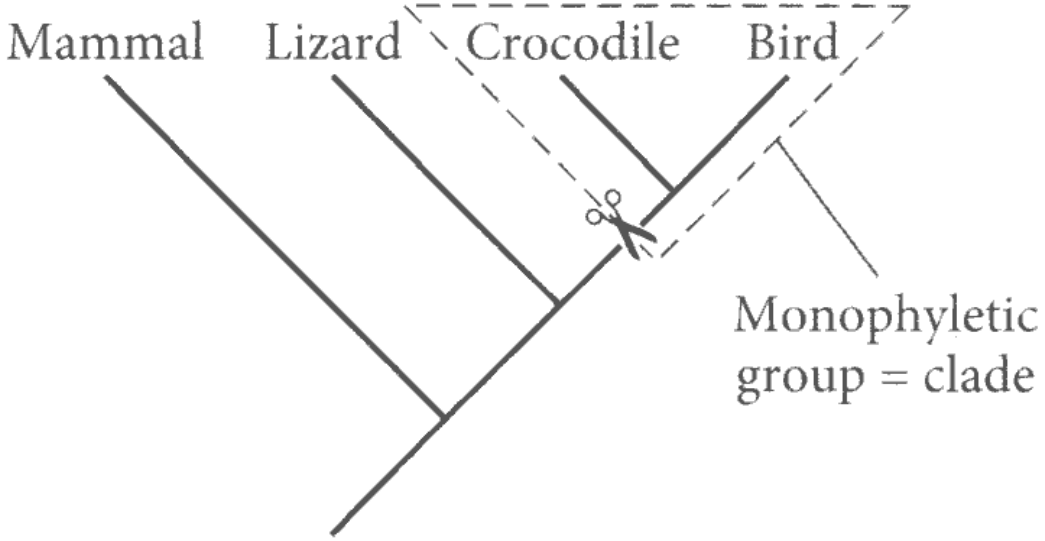


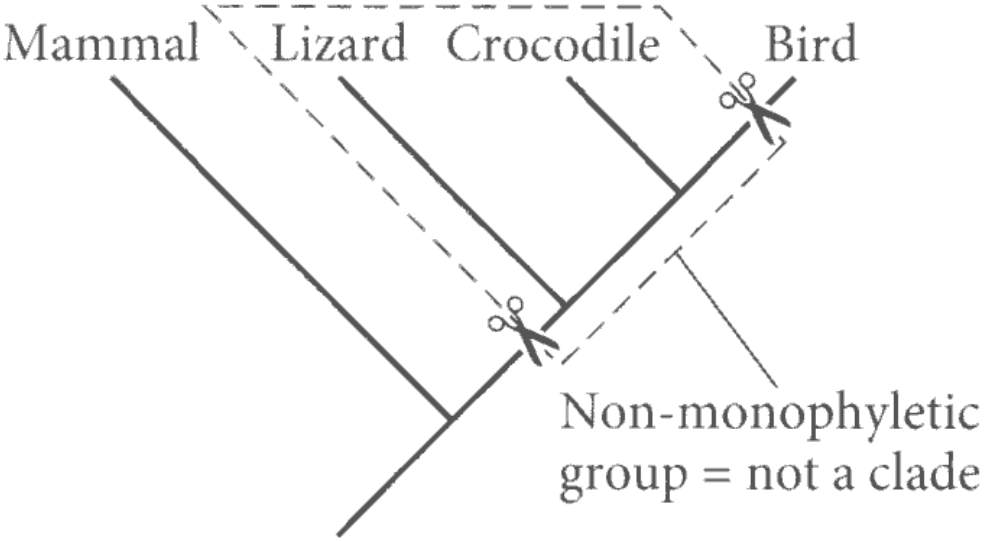

Circle

# Phylogenetic Trees

Monophyly, paraphyly, and polyphyly…

Reptiles are not a “natural” (monophyletic)!

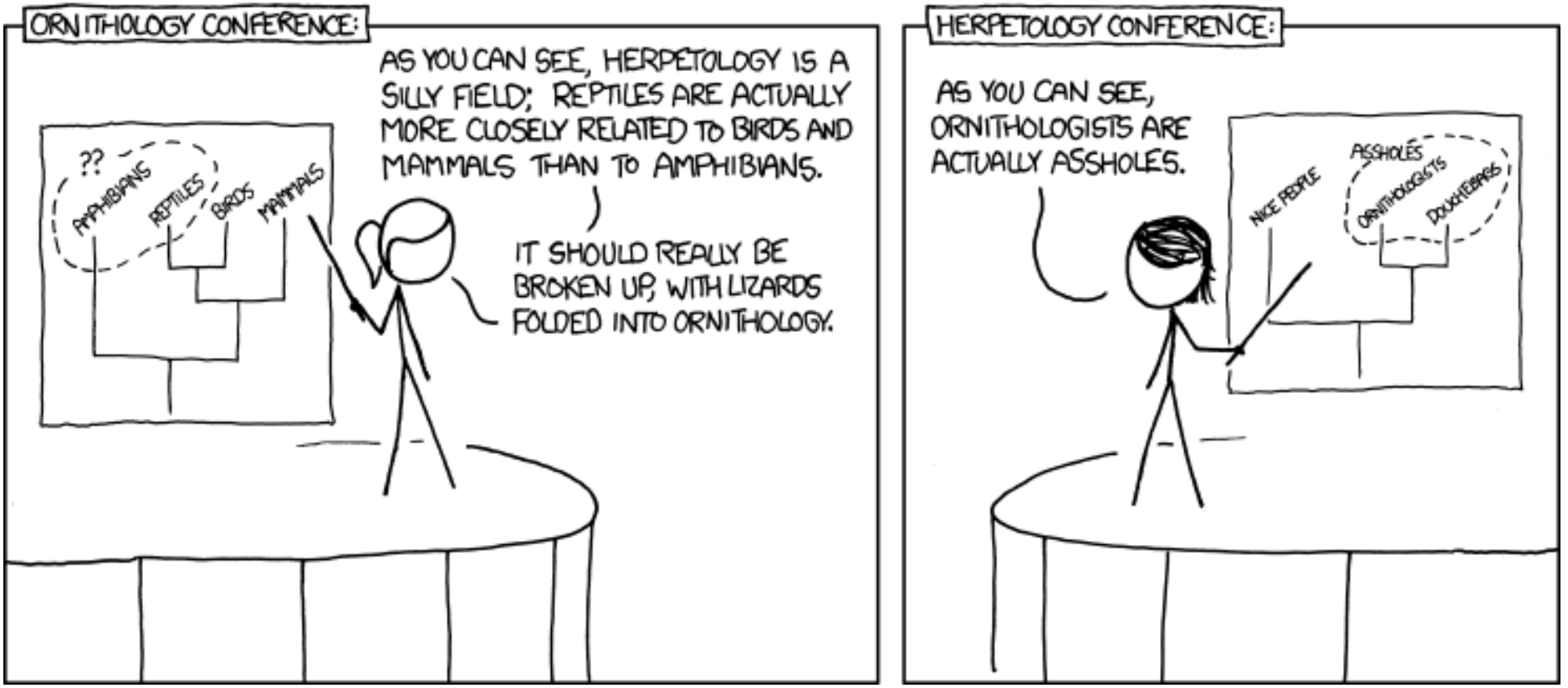


# Phylogenetic Trees

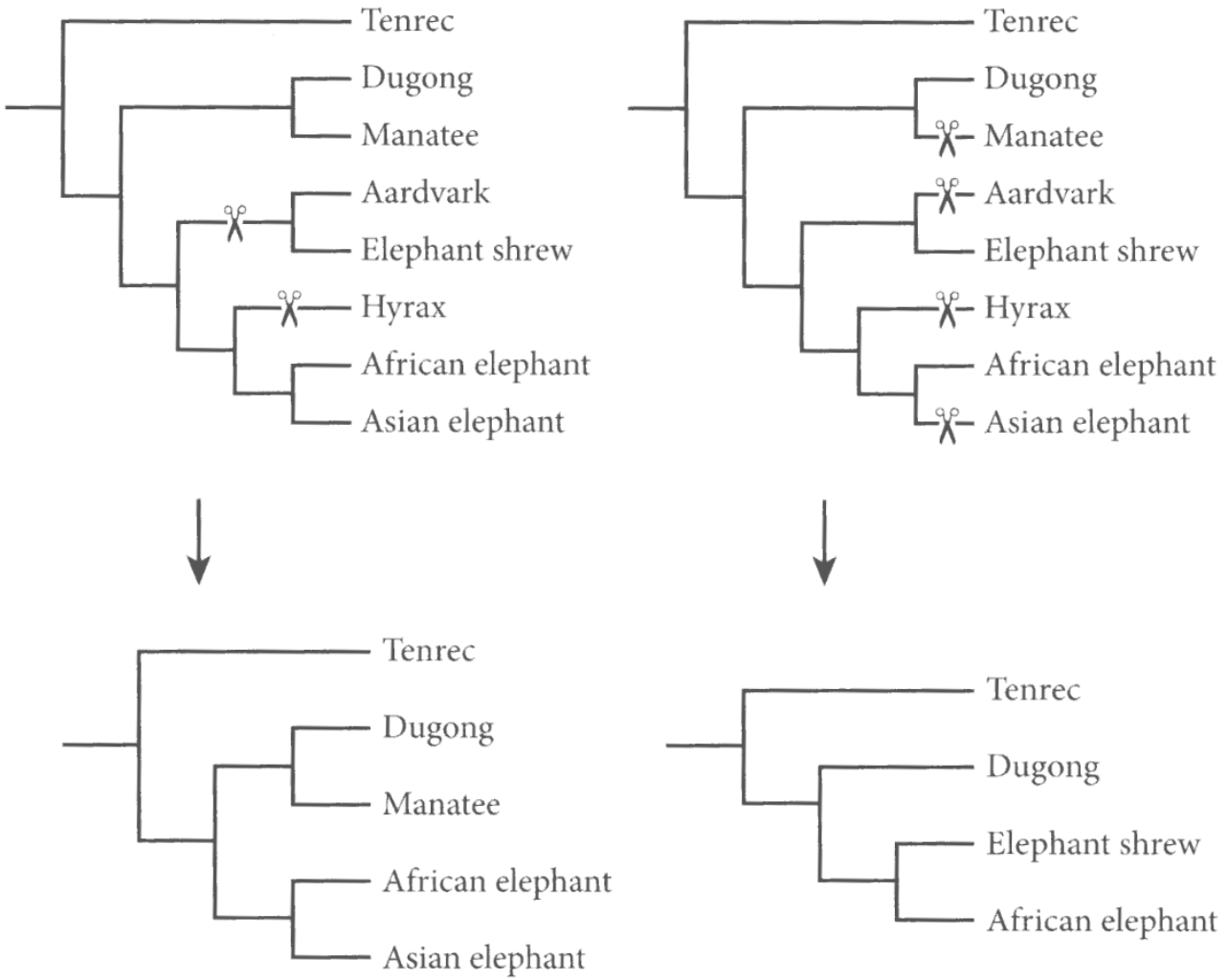
Monophyly, paraphyly, and polyphyly…

Reptiles are not a “natural” (monophyletic)!



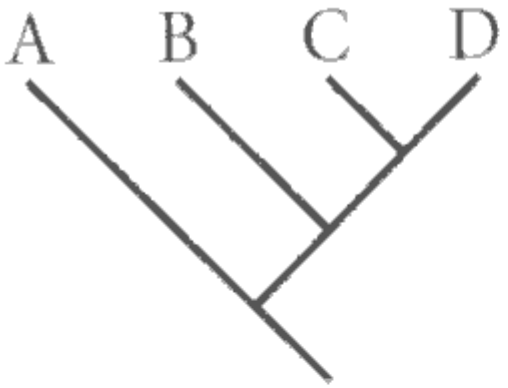
# Phylogenetic Trees

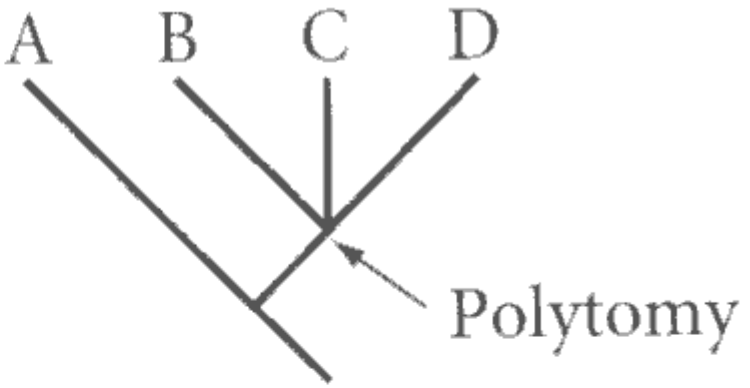
Pruning does not change the relationships among taxa

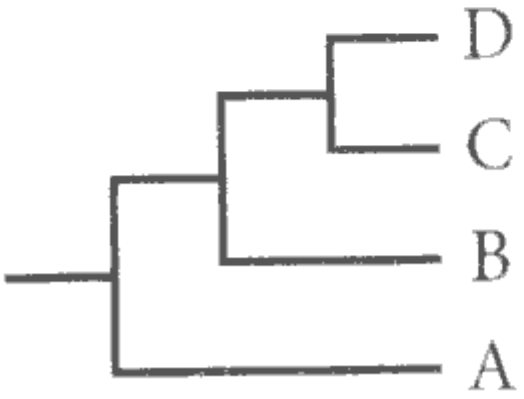


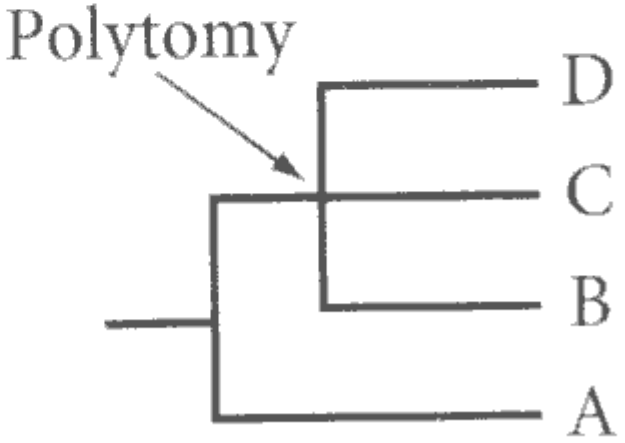
# Phylogenetic Trees

Expressing uncertainty of relationships



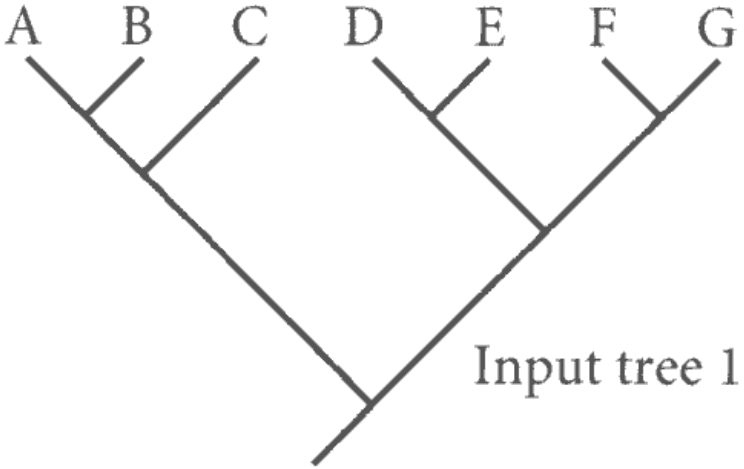



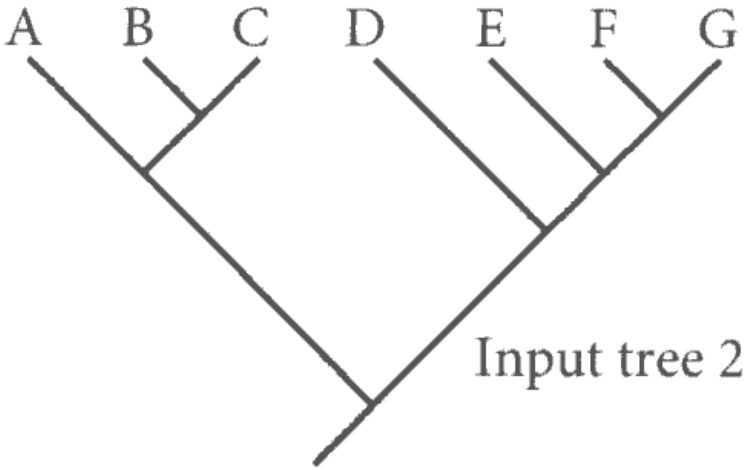



# Phylogenetic Trees

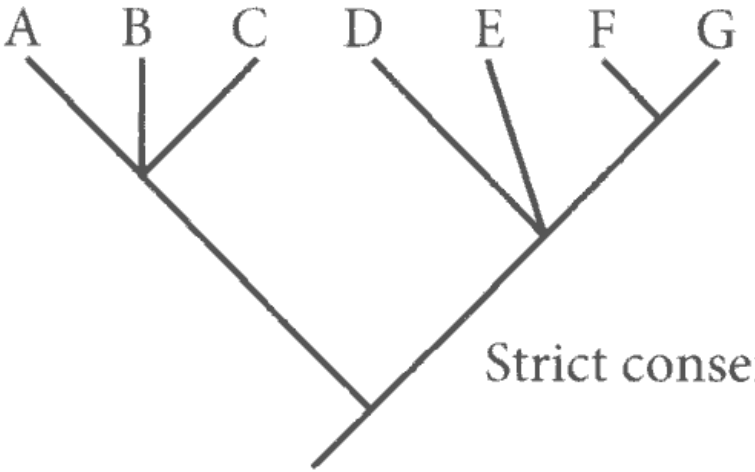
Summarizing competing phylogenetic hypotheses




Input tree 1




Input tree 2

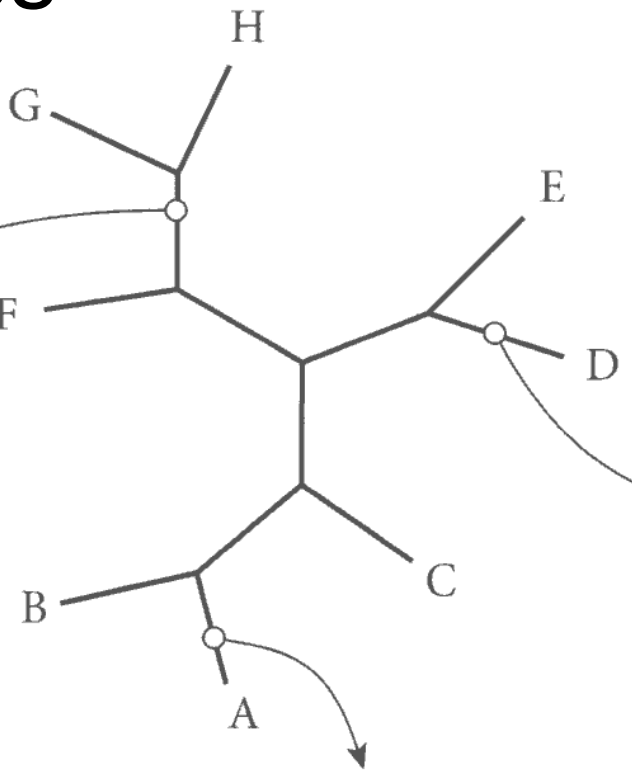


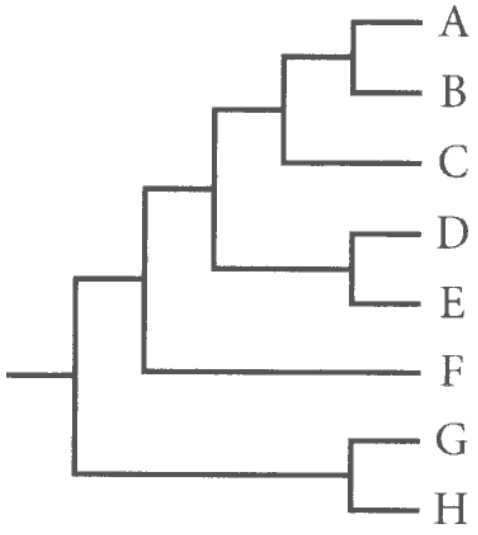

Strict consensus tree

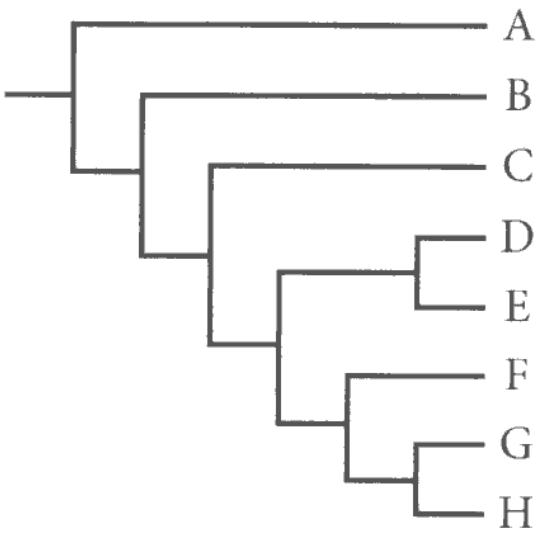
# Phylogenetic Trees

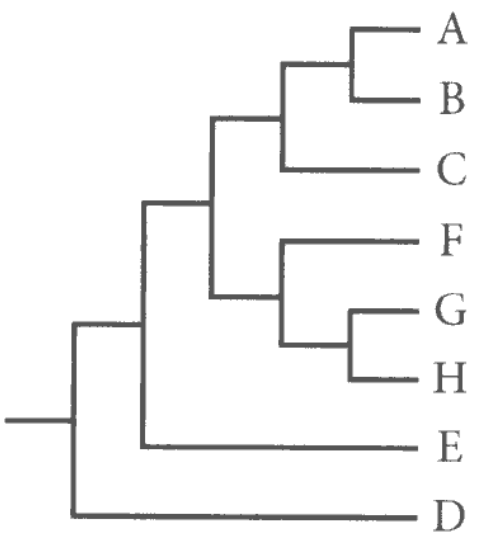
Rooting phylogenetic trees using an outgroup

**Are the relationships the same?**



# Character evolution

The same structure can be present in different species due to:

Homology (character evolved once in the last common ancestor)

or

Homoplasy (convergence or reversal/parallelism)

# Inferring Phylogenetic Trees

| No. | Character | States |
|---|---|---|
| 1 | Complexity of the mucus-coated surfaces in the nose (maxilloturbinals) | Minimally branched (olfactory surfaces in nasal passage) (0); highly branching (olfactory surfaces excluded from the nasal passage) (1) |
| 2 | Bony spur by the auditory bulla (paroccipital process) | Straight and projecting (0); cupped around auditory bulla (1) |
| 3 | Number of lower incisors | 2 (0); 3 (1) |
| 4 | Upper molar 1 | Present (0); absent (1) |
| 5 | Baculum (bone within the penis) | Present (0); absent (1) |
| 6 | Tail | Elongated (0); short (1) |
| 7 | Hallux (5th digit, or dewclaw, on hind leg) | Prominent (0); reduced or absent (1) |
| 8 | Claws | Nonretractable (0); retractable (1) |
| 9 | Prostate gland | Small and simple (0); large and bilobed (1) |
| 10 | Kidney structure | Simple (0); conglomerate (1) |
| 11 | External ear (pinna) | Present (0); absent (1) |
| 12 | Testis position | Scrotal (0); abdominal (1) |

| | Character state scoring | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Taxon | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Creodont | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ? | ? | ? | ? |
| Cat | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Hyena | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Civet | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Dog | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Raccoon | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bear | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Otter | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Seal | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Walrus | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Sea lion | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

# Inferring Phylogenetic Trees

| No. | Character | States |
|---|---|---|
| 1 | Complexity of the mucus-coated surfaces in the nose (maxilloturbinals) | Minimally branched (olfactory surfaces in nasal passage) (0); highly branching (olfactory surfaces excluded from the nasal passage) (1) |
| 2 | Bony spur by the auditory bulla (paroccipital process) | Straight and projecting (0); cupped around auditory bulla (1) |
| 3 | Number of lower incisors | 2 (0); 3 (1) |
| 4 | Upper molar 1 | Present (0); absent (1) |
| 5 | Baculum (bone within the penis) | Present (0); absent (1) |
| 6 | Tail | Elongated (0); short (1) |
| 7 | Hallux (5th digit, or dewclaw, on hind leg) | Prominent (0); reduced or absent (1) |
| 8 | Claws | Nonretractable (0); retractable (1) |
| 9 | Prostate gland | Small and simple (0); large and bilobed (1) |
| 10 | Kidney structure | Simple (0); conglomerate (1) |
| 11 | External ear (pinna) | Present (0); absent (1) |
| 12 | Testis position | Scrotal (0); abdominal (1) |

| | Character state scoring | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Taxon | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Creodont | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ? | ? | ? | ? |
| Cat | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Hyena | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Civet | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Dog | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Raccoon | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bear | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Otter | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Seal | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Walrus | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Sea lion | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

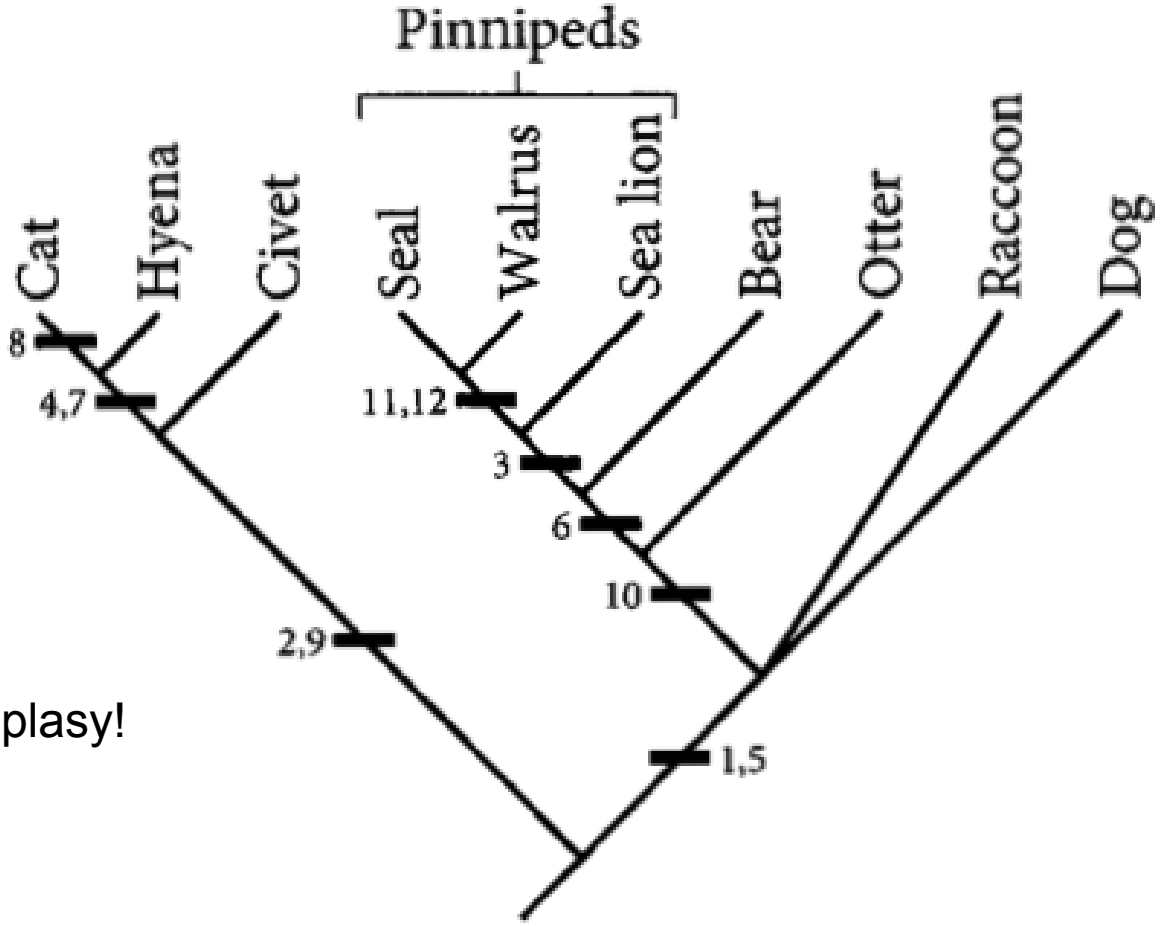
# Inferring Phylogenetic Trees

| | Character state scoring | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Taxon | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Creodont | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ? | ? | ? | ? |
| Cat | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Hyena | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Civet | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Dog | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Raccoon | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bear | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Otter | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Seal | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Walrus | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Sea lion | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

If characters display homoplasy it becomes difficult to reconcile tree with character matrix.

We need an optimality criterion:

Find the tree that minimizes homoplasy! (Maximum Parsimony)

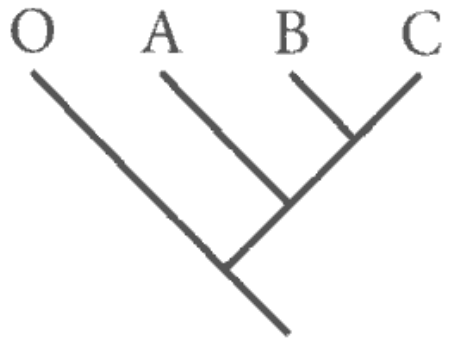


# Parsimony: basic principle

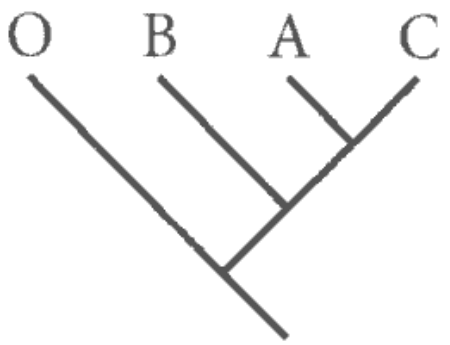
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| O | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| B | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

Parsimony aims at reducing the amount of homoplasy or number of steps needed to explain the evolution of characters (ie, find the “shortest” tree)

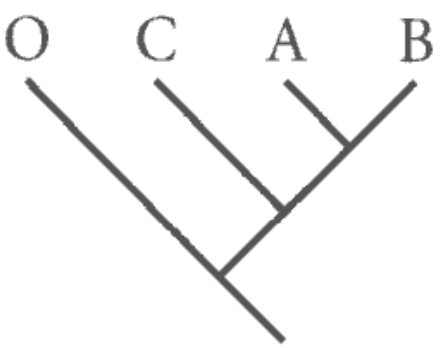
3 possible roted trees for 3 ingroup taxa




Tree 1




Tree 2

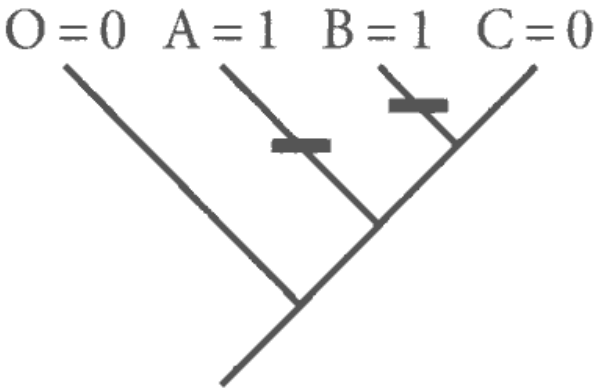


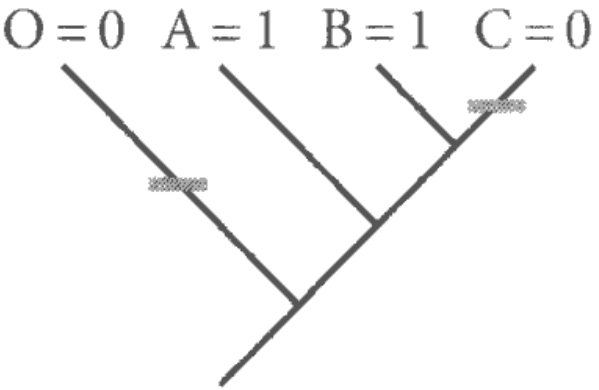

Tree 3

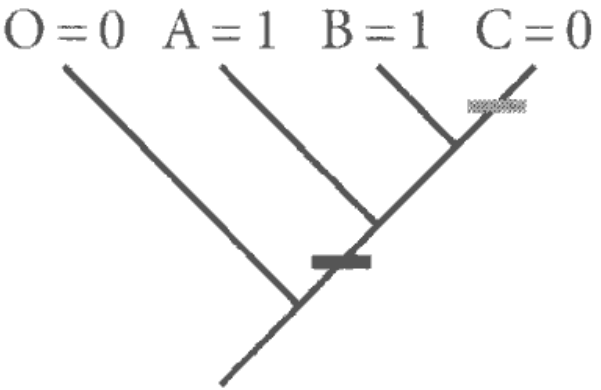
# Parsimony: basic principle

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| O | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| B | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

3 possible reconstructions of character 2 on tree 1



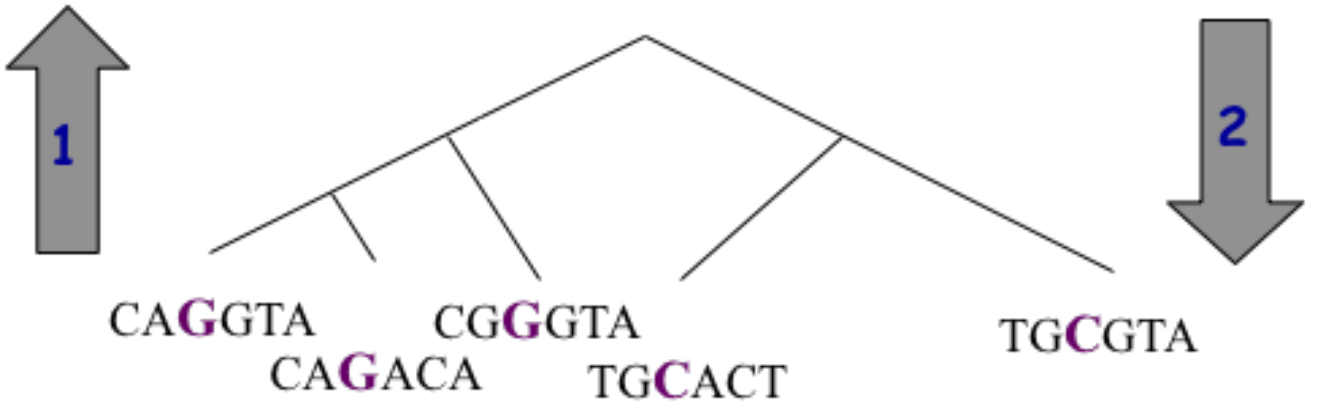



# Parsimony: Fitch algorithm

Execute independently for each character:

1. Bottom-up phase: Determine set of possible states for each internal node

2. Top-down phase: Pick states for each internal node



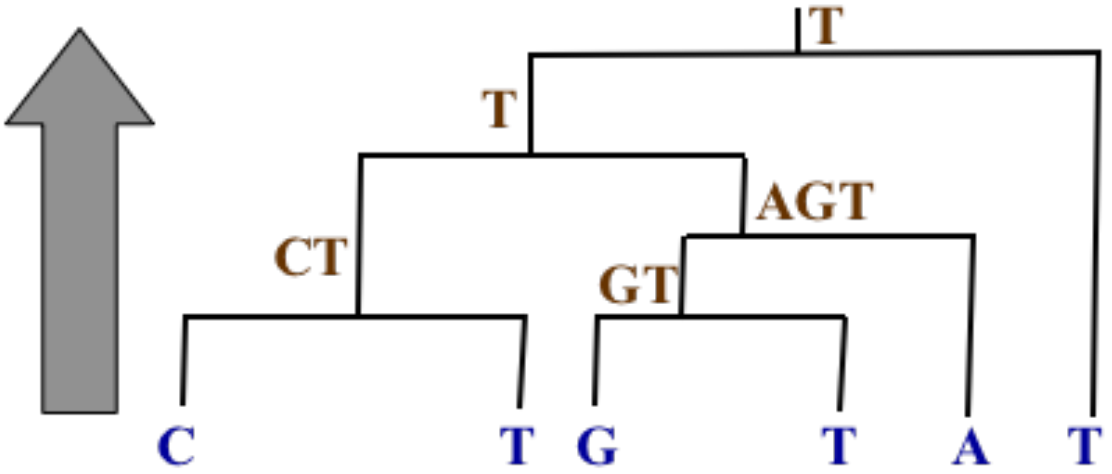
# Parsimony: Fitch algorithm

**Bottom-Up Phase**

Determine set of possible states for each internal node

- Initialization: $R_i = \{s_i\}$
- Do a post-order (from leaves to root) traversal of tree
  - Determine $R_i$ of internal node $i$ with children $j, k$:

$$R_i = \begin{cases} R_j \cap R_k & \text{if } R_j \cap R_k \neq \phi \\ R_j \cup R_k & \text{otherwise} \end{cases}$$




Parsimony-score = # union operations

score = 3

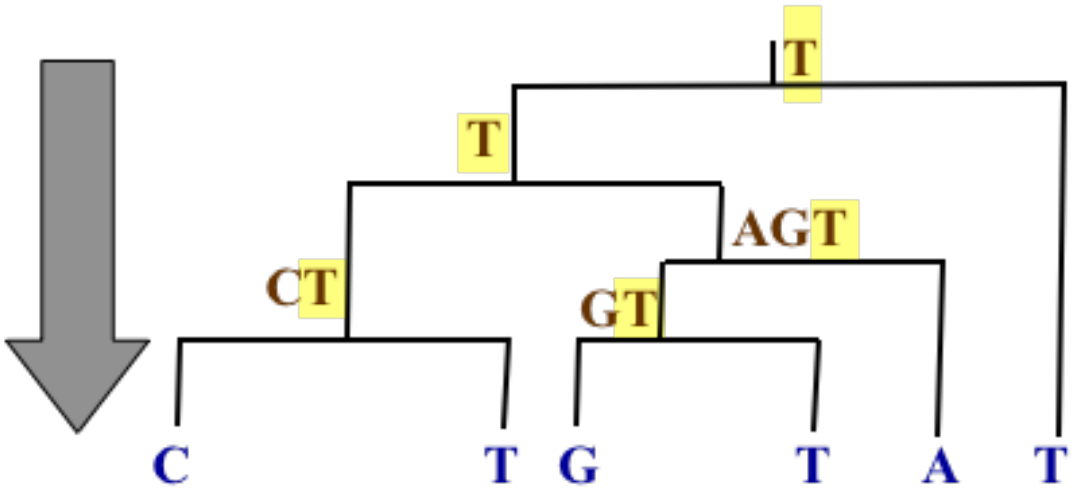
# Parsimony: Fitch algorithm

**Top-Down Phase**

## Pick states for each internal node

- Pick arbitrary state in $R_{root}$ for the root
- Do pre-order (from root to leaves) traversal of tree
  - Determine $s_j$ of internal node $j$ with parent $i$:

$$s_j = \begin{cases} s_i \;\; if \;\; s_i \in R_j \\ arbitrary \;\; state \in R_j \;\; otherwise \end{cases}$$




score = 3

# Parsimony: Fitch algorithm

Going up, for each character….

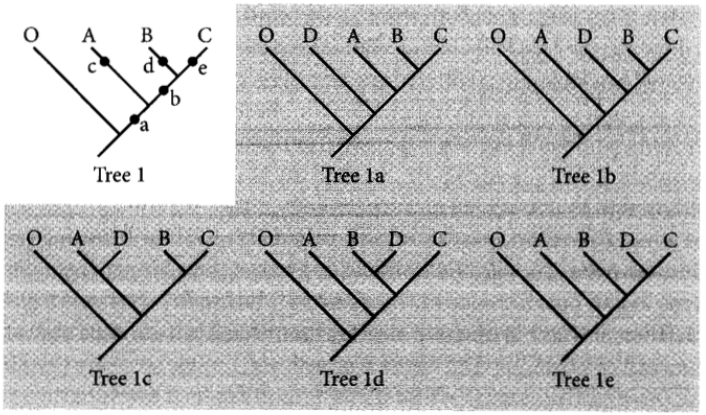
- If there is a common state, keep only it
- Otherwise, keep all the states and pay* a point

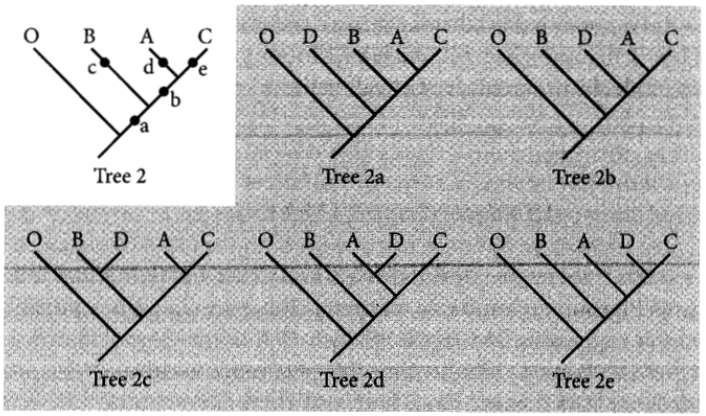
Going down, for each character….

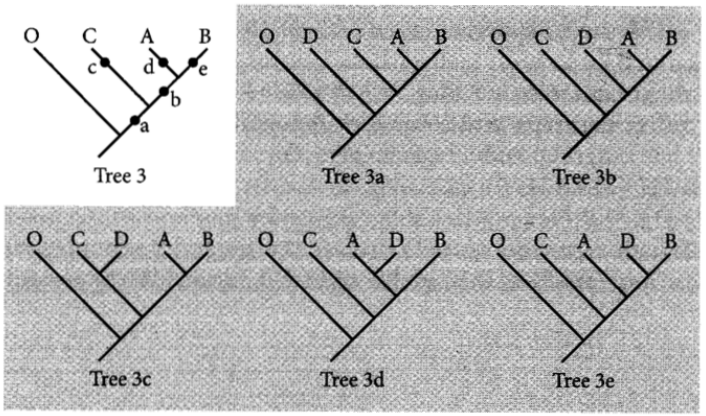
- Keep the common state between parent and child, if there is one
- If not, pick one from the child arbitrarily and pay* a point

*The penalty is paid either going up or going down, but not twice

# Finding optimal trees



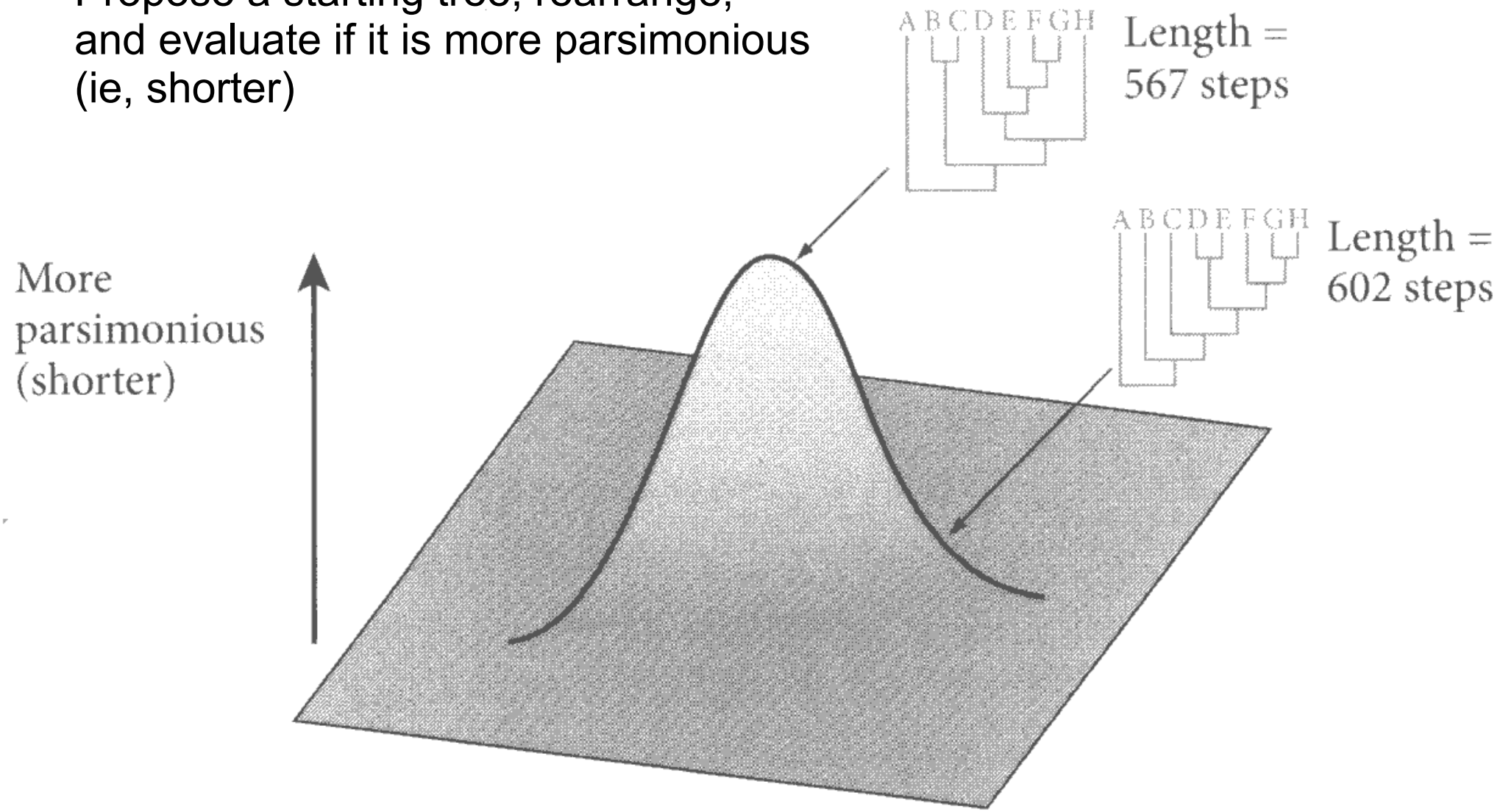



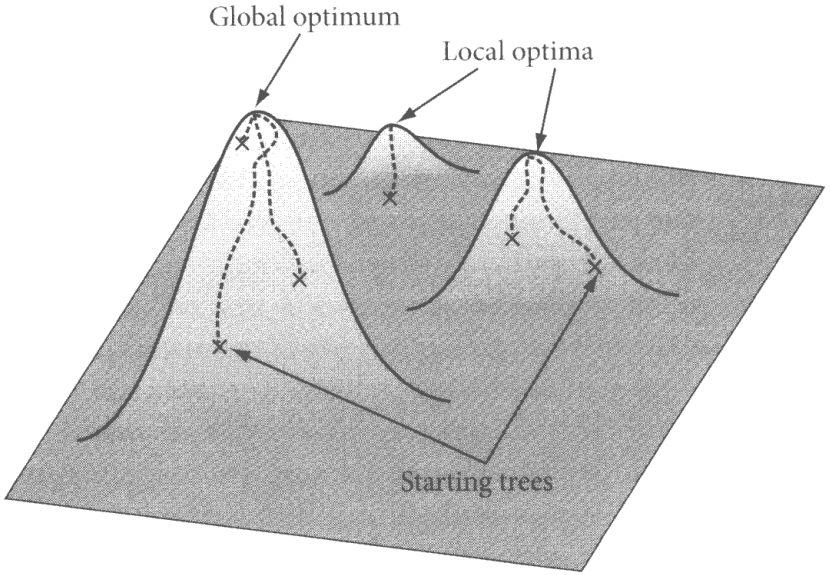

Tree space becomes very big quickly and
We cannot evaluate every possible tree!

# Heuristic tree searches (hill-climbing algorithm)

Propose a starting tree, rearrange, and evaluate if it is more parsimonious (ie, shorter)



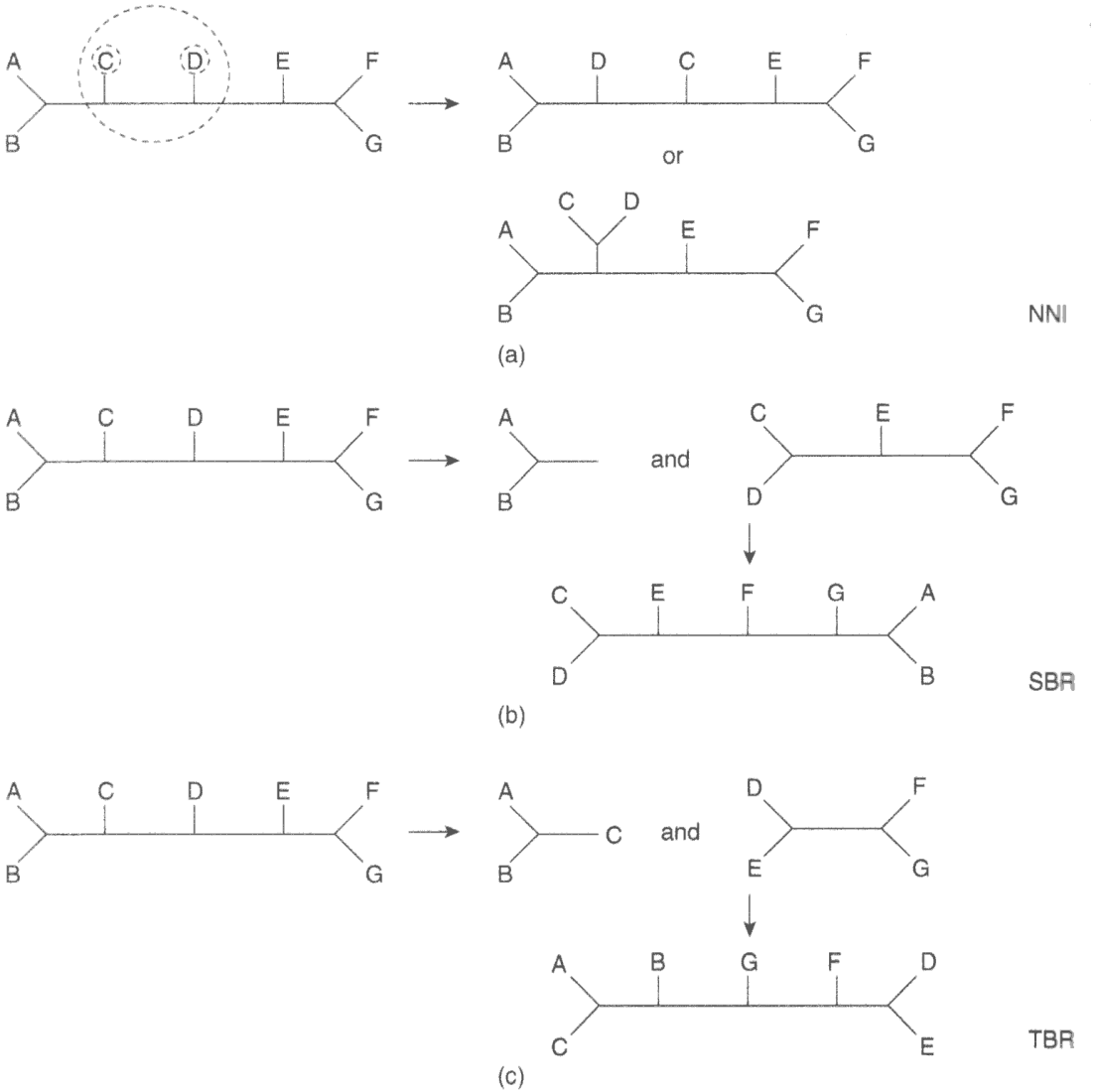
# Heuristic tree searches (hill-climbing algorithm)



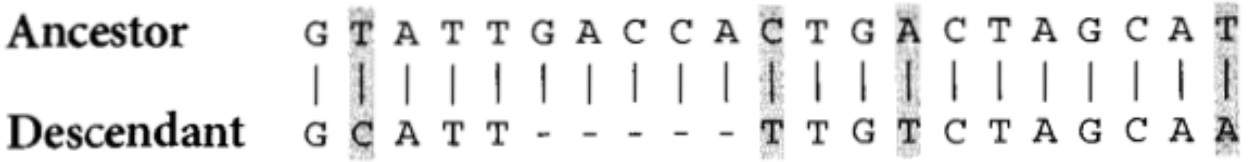

How can we escape local optima and find the global optimum?

Accept sub-optimal trees Sometimes...

Choose several random starting points...



# Parsimony: Molecular Data


Deletion


Insertion

We do not necessarily know *a priori* about ancestor/descendant relationships. So insertions and deletions are generally referred to as **indels**.

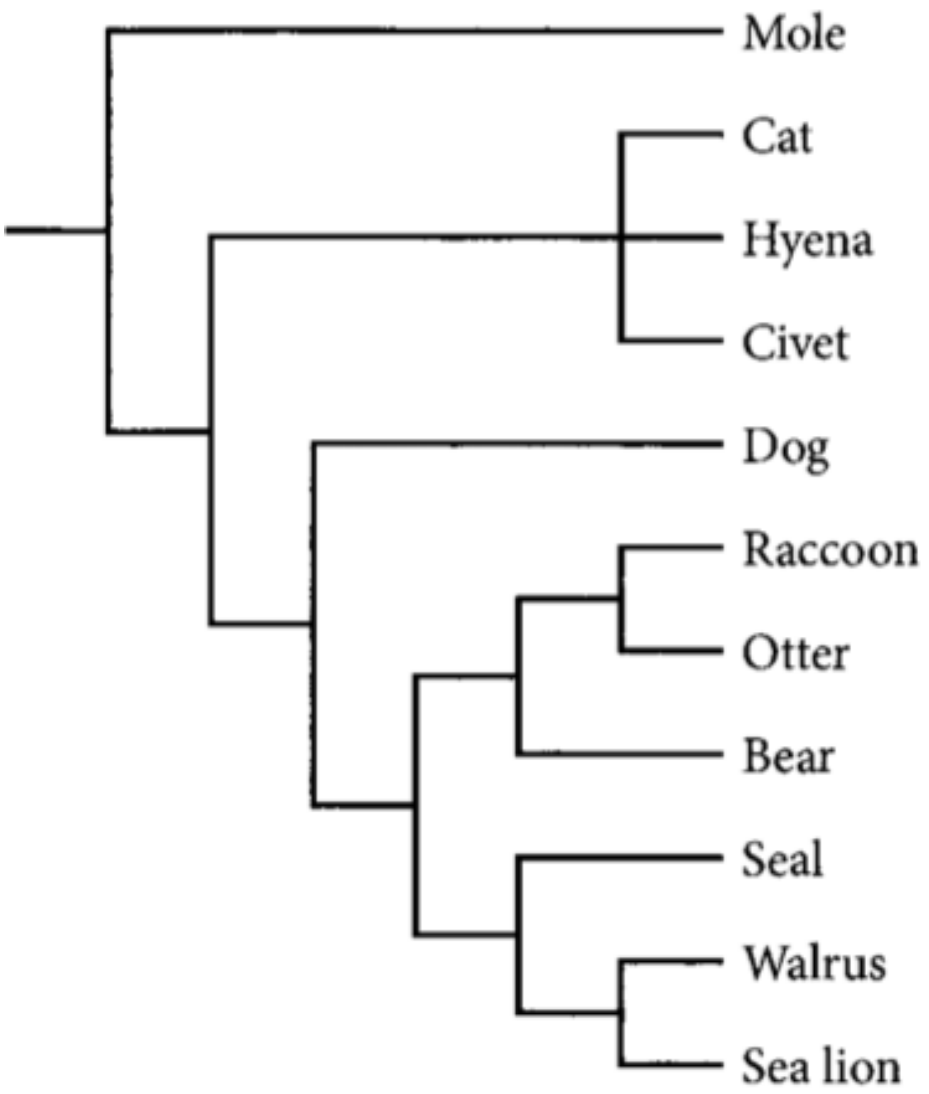
How many events does an indel represent?

# Parsimony: Molecular Data

| | Positions in DNA sequence | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Taxon | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Mole | G | T | T | A | A | - | C | T | T | C | T | C | A | C | T |
| Cat | G | T | T | G | A | - | C | C | T | C | T | T | A | C | T |
| Hyena | G | T | T | G | A | - | C | C | T | C | T | C | A | C | T |
| Civet | G | T | T | G | A | - | C | C | T | C | T | C | A | C | T |
| Dog | G | T | T | A | A | G | C | A | T | C | T | G | C | C | T |
| Raccoon | G | T | T | A | A | G | G | G | T | C | T | G | C | C | T |
| Bear | C | T | T | A | A | G | T | G | T | C | T | G | C | C | T |
| Otter | G | T | T | A | A | G | G | G | T | C | T | G | C | C | T |
| Seal | G | T | A | A | A | G | C | G | T | C | T | G | C | C | T |
| Walrus | G | T | A | A | A | G | C | G | T | C | T | G | C | C | T |
| Sea lion | G | T | A | A | A | G | C | G | T | C | T | G | C | C | T |

# Parsimony: Molecular Data

| | Positions in DNA sequence | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Taxon | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Mole | G | T | T | A | A | - | C | T | T | C | T | C | A | C | T |
| Cat | G | T | T | G | A | - | C | C | T | C | T | T | A | C | T |
| Hyena | G | T | T | G | A | - | C | C | T | C | T | C | A | C | T |
| Civet | G | T | T | G | A | - | C | C | T | C | T | C | A | C | T |
| Dog | G | T | T | A | A | G | C | A | T | C | T | G | C | C | T |
| Raccoon | G | T | T | A | A | G | G | G | T | C | T | G | C | C | T |
| Bear | C | T | T | A | A | G | T | G | T | C | T | G | C | C | T |
| Otter | G | T | T | A | A | G | G | G | T | C | T | G | C | C | T |
| Seal | G | T | A | A | A | G | C | G | T | C | T | G | C | C | T |
| Walrus | G | T | A | A | A | G | C | G | T | C | T | G | C | C | T |
| Sea lion | G | T | A | A | A | G | C | G | T | C | T | G | C | C | T |



# Parsimony: Models of Evolution

Step Matrix with character state changes receiving same weight

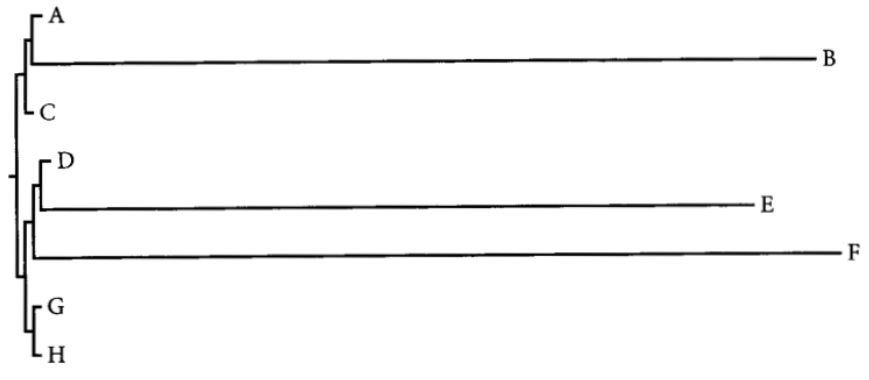
| | | To: | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| From: | A | 0 | 1 | 1 | 1 |
| | C | 1 | 0 | 1 | 1 |
| | G | 1 | 1 | 0 | 1 |
| | T | 1 | 1 | 1 | 0 |

Step Matrix upweighting transversions compared to transitions

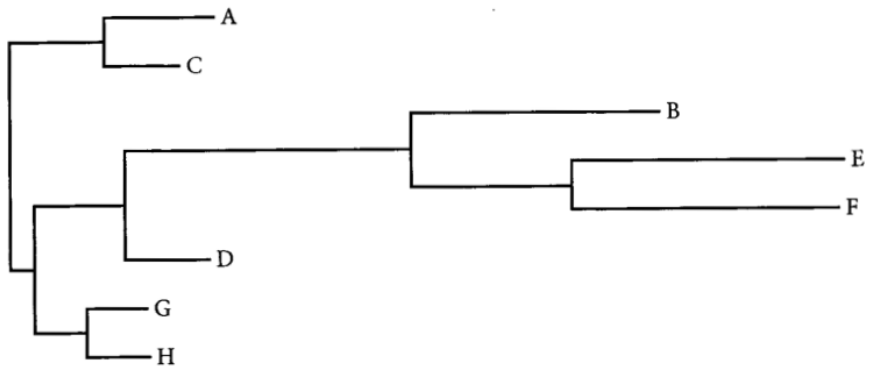
| | | To: | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| From: | A | 0 | 2 | 1 | 2 |
| | C | 2 | 0 | 2 | 1 |
| | G | 1 | 2 | 0 | 2 |
| | T | 2 | 1 | 2 | 0 |

# Problems of Parsimony

Parsimony does not take branch lengths into account. High rates of evolution, especially in different parts of the tree can lead to inference of wrong tree




True tree

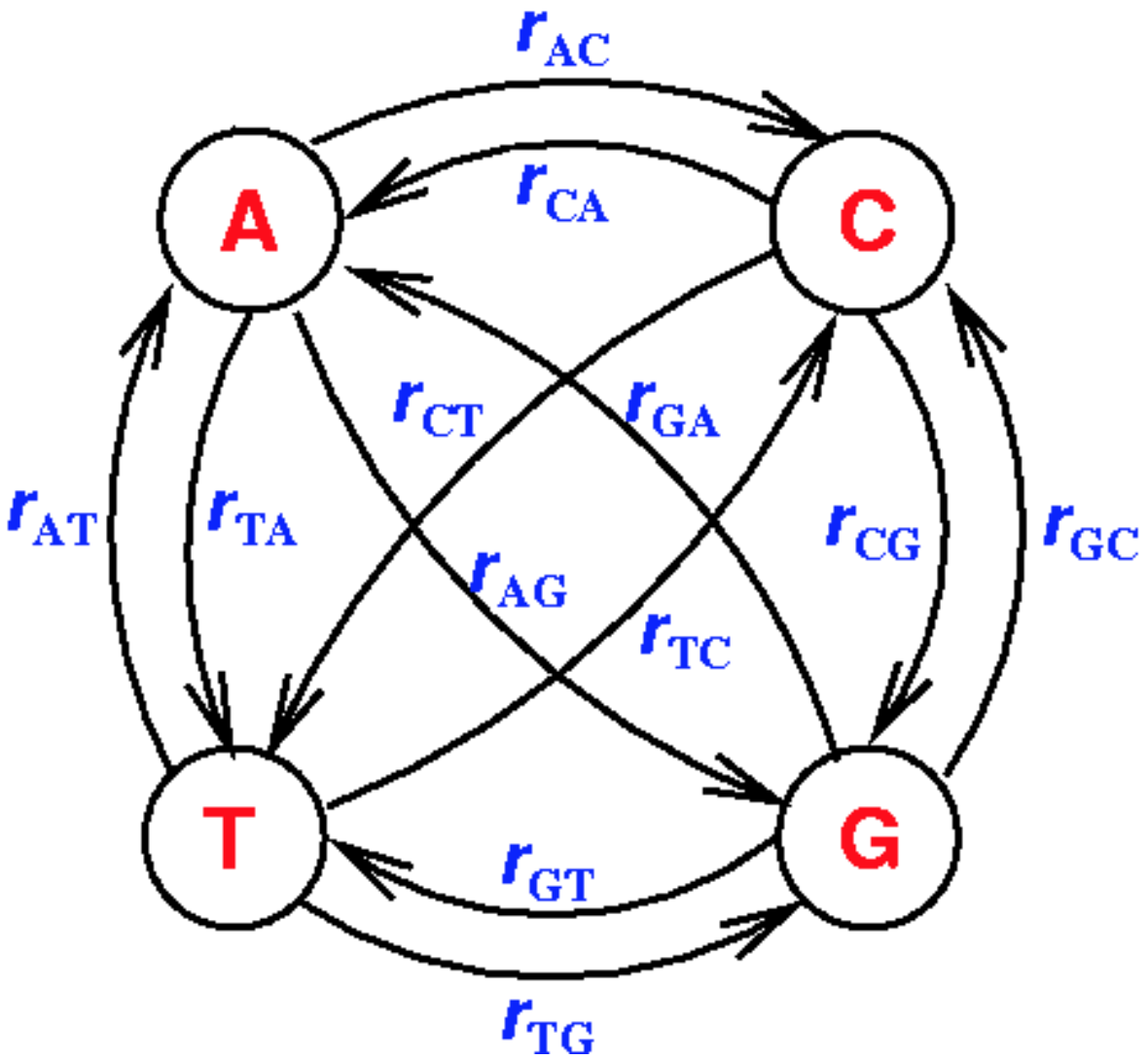



Unrooted tree estimated by parsimony

Character weighing may alleviate the issue but there is no formal way of testing weighting schemes in the parsimony framework.

# Models of Molecular Evolution

Substitution model specifies way in which characters evolve




Markov models:

Process in which probability of an event happening in some time window is dependent only on the state at that time and independent on how it came to be in that state (eg, coin toss).

# Models of Molecular Evolution

Jukes-Cantor model is the simplest substitution model

Core assumptions:
1) equal base frequencies
2) equal rates of substitution
3) rates of substitution are the same across the sequence

| | | To: | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| From: | A | $1/4 + 3/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ |
| | C | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 + 3/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ |
| | G | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 + 3/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ |
| | T | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 - 1/4e^{-4/3\mu t}$ | $1/4 + 3/4e^{-4/3\mu t}$ |

**FIGURE 8.5 Substitution probability matrix under the JC model of DNA sequence evolution.** The mutation rate, in substitutions per unit time, is denoted $\mu$. The time interval over which evolution is allowed to happen is denoted $t$.

# Models of Molecular Evolution

F81 model:

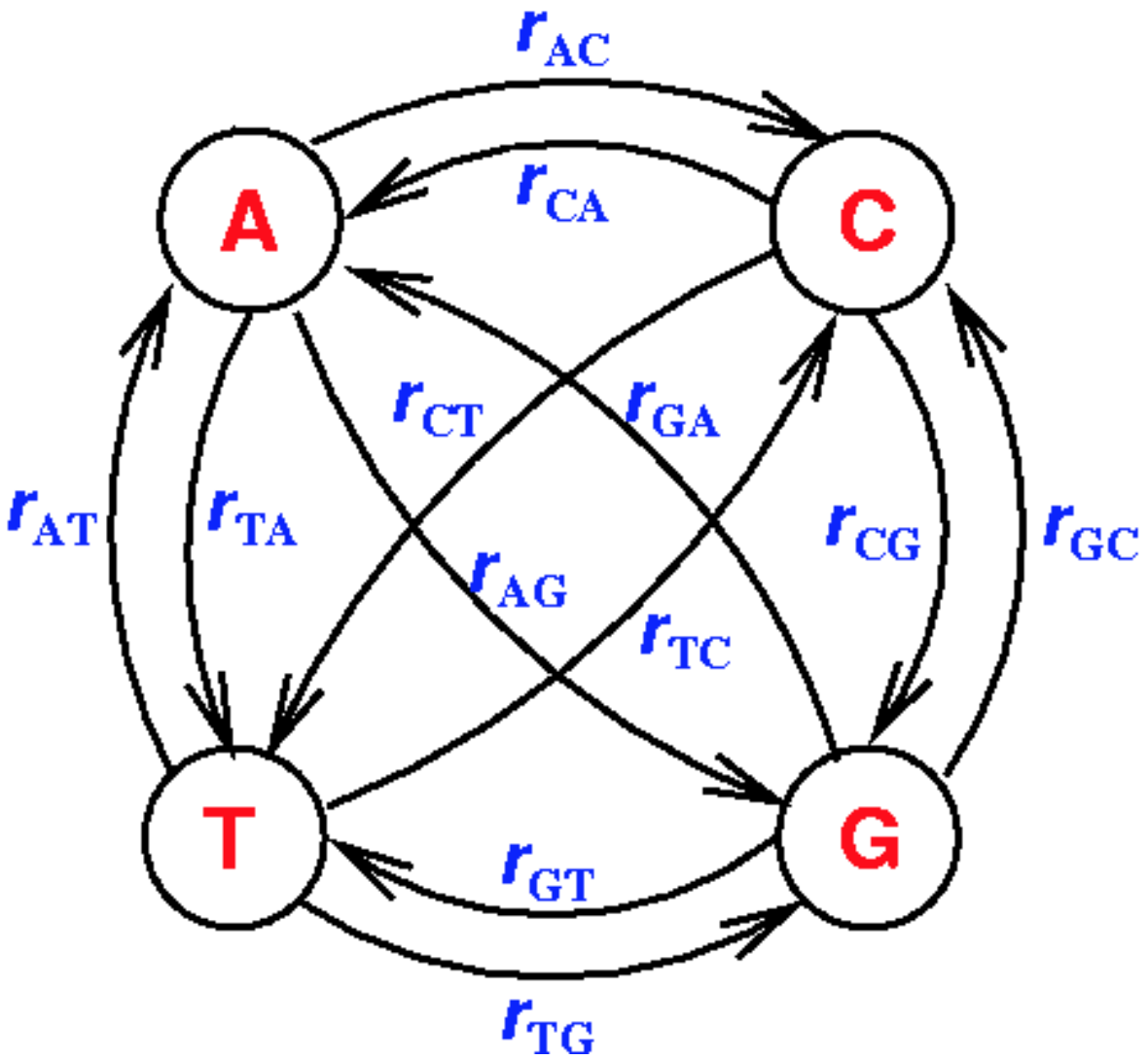
m is a modified version of μt that takes the effect of unequal base frequencies into account. Rare bases will persist for less time than common bases that will often be substituted by themselves.

| | | To: | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| From: | A | $\pi_A + (1 - \pi_A)e^{-mt}$ | $\pi_C(1 - e^{-mt})$ | $\pi_G(1 - e^{-mt})$ | $\pi_T(1 - e^{-mt})$ |
| | C | $\pi_A(1 - e^{-mt})$ | $\pi_C + (1 - \pi_C)e^{-mt}$ | $\pi_G(1 - e^{-mt})$ | $\pi_T(1 - e^{-mt})$ |
| | G | $\pi_A(1 - e^{-mt})$ | $\pi_C(1 - e^{-mt})$ | $\pi_G + (1 - \pi_G)e^{-mt}$ | $\pi_T(1 - e^{-mt})$ |
| | T | $\pi_A(1 - e^{-mt})$ | $\pi_C(1 - e^{-mt})$ | $\pi_G(1 - e^{-mt})$ | $\pi_T + (1 - \pi_T)e^{-mt}$ |

**FIGURE 8.8 Substitution probability matrix under the F81 model of DNA sequence evolution.** Base frequency notation is the same as Figure 8.6. The effective mutation rate, after correcting for base compositional inequality (see text), is denoted *m*. The time interval over which evolution is allowed to happen is denoted *t*.

# Models of Molecular Evolution

Substitution model specifies way in which characters evolve




Markov models:

Process in which probability of an event happening in some time window is dependent only on the state at that time and independent on how it came to be in that state (eg, coin toss).

# Maximum Likelihood

Probability of the data given the hypothesis

| Toss | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Likelihood |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Result |  | | | | | | | | | | |
| Prob. if fair | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.001 |
| Prob. if biased | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.056 |

**FIGURE 8.15 Likelihood of ten sequential heads for a fair or a biased coin.** The likelihood is the product of the probabilities of each individual toss, e.g. $0.5^{10}$, under the fair coin hypothesis.

# Maximum Likelihood

Only unknown is this tree is branch length.

How long is it under JC model?

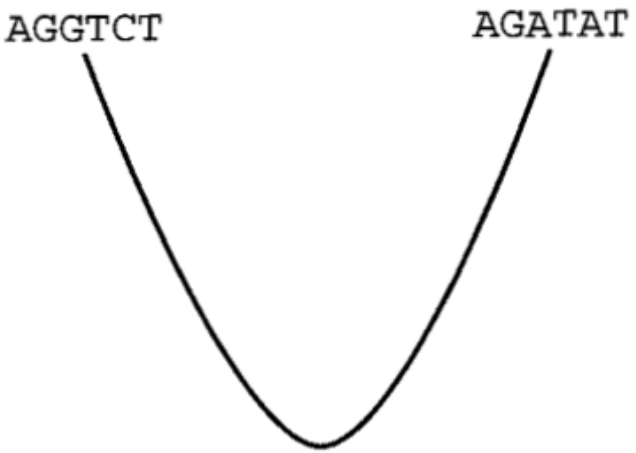



**FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

# Maximum Likelihood

Only unknown in this tree is branch length.

How long is it under JC model?

Cannot be infinitely long since 4 matching bases (¾ of bases [0.75 distance) should be mismatches after infinite time)

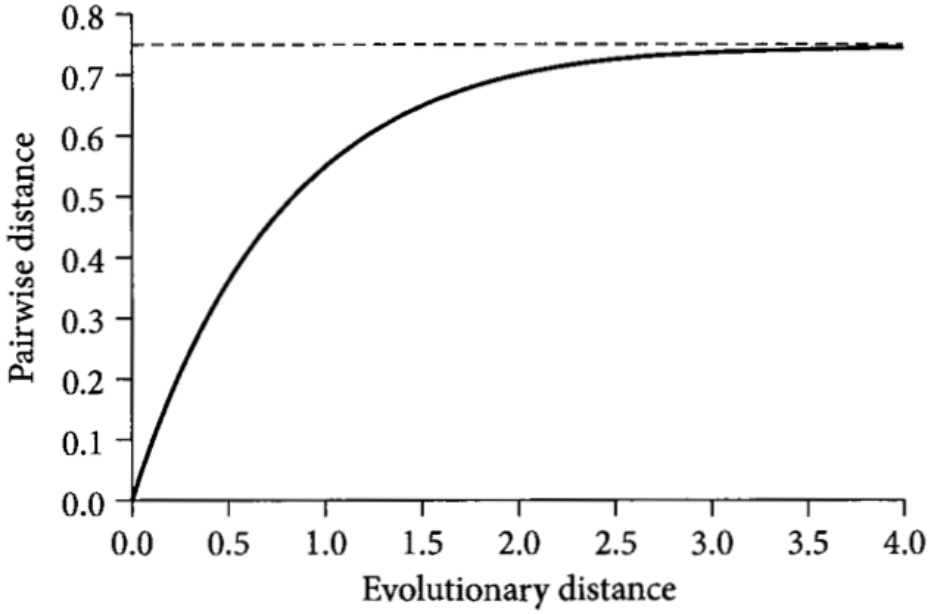



**FIGURE 8.12 Relationship of evolutionary and pairwise distances under the JC model.** The dashed line indicates the maximum expected pairwise distance, 0.75.

# Maximum Likelihood

Only unknown is this tree is branch length.

How long is it under JC model?

Since two bases are mismatches it cannot be zero length

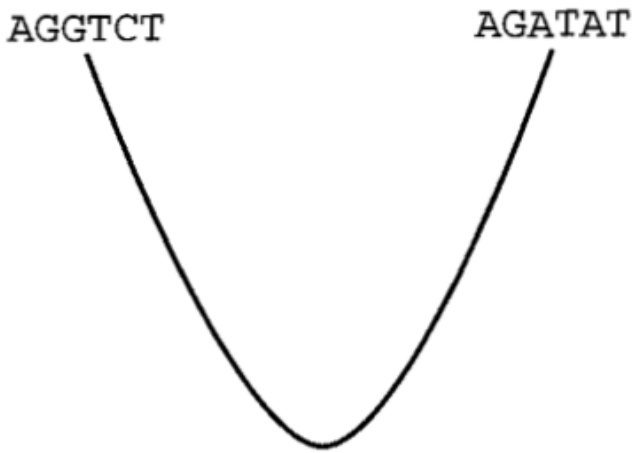



**FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

# Maximum Likelihood

The rate of evolution (μ) and time (t) will determine the branches length. μ and t are difficult to disentangle but all we need to worry about is their product μt.

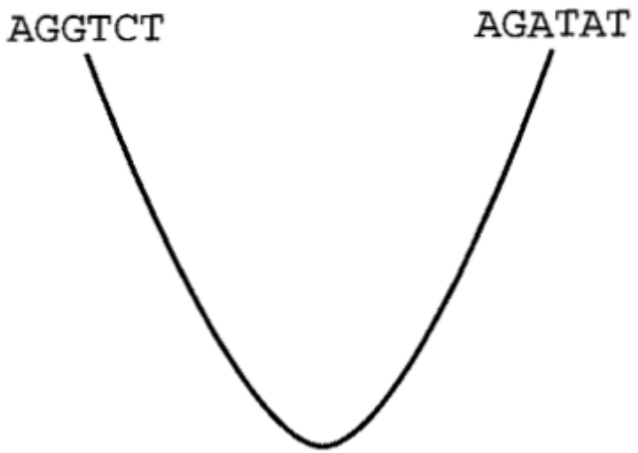



**FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

# Maximum Likelihood

The rate of evolution (μ) and time (t) will determine the branches length. μ and t are difficult to disentangle but all we need to worry about is their product μt.

The probability of the first position in one taxon to be A is ¼ (equal base frequencies)
Given that, the probability of observing A in both taxa is:
$¼ (¼ + ¾ e^{-4/3\mu t})$

We can substitute any value for μt into the equation to obtain the probability that this site evolved under that branch length. This is the **site likelihood**.

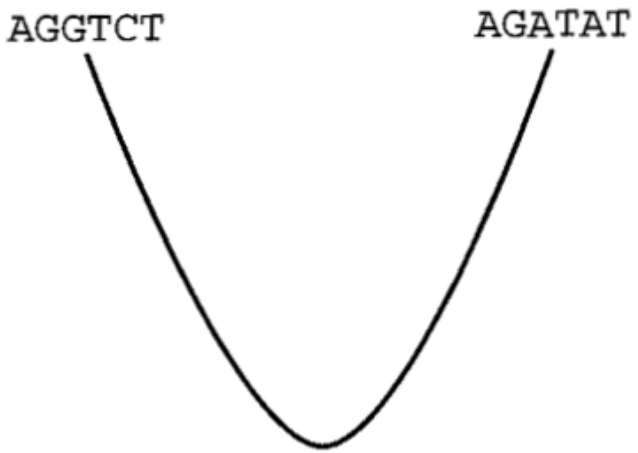



**FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

# Maximum Likelihood

The probability (site likelihood) of a mismatch is
$¼ (¼ - ¼ e^{-4/3\mu t})$

The likelihood of a tree knowing the entire characters matrix is the product of all site likelihoods.

$[¼ (¼ + ¾ e^{-4/3\mu t})]^2 * [¼ (¼ + ¾ e^{-4/3\mu t})]^4$

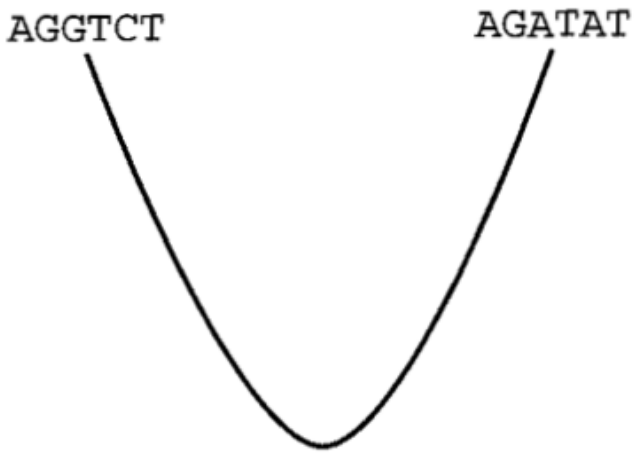



**FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

# Maximum Likelihood

The likelihood of a tree knowing the entire characters matrix is the product of all site likelihoods.

Branch length of 0.44 has highest likelihood ($0.595 \times 10^{-6}$)

Log likelihood is -14.33 (avoids computer issues with small numbers. Highest likelihood corresponds to least negative log likelihood.

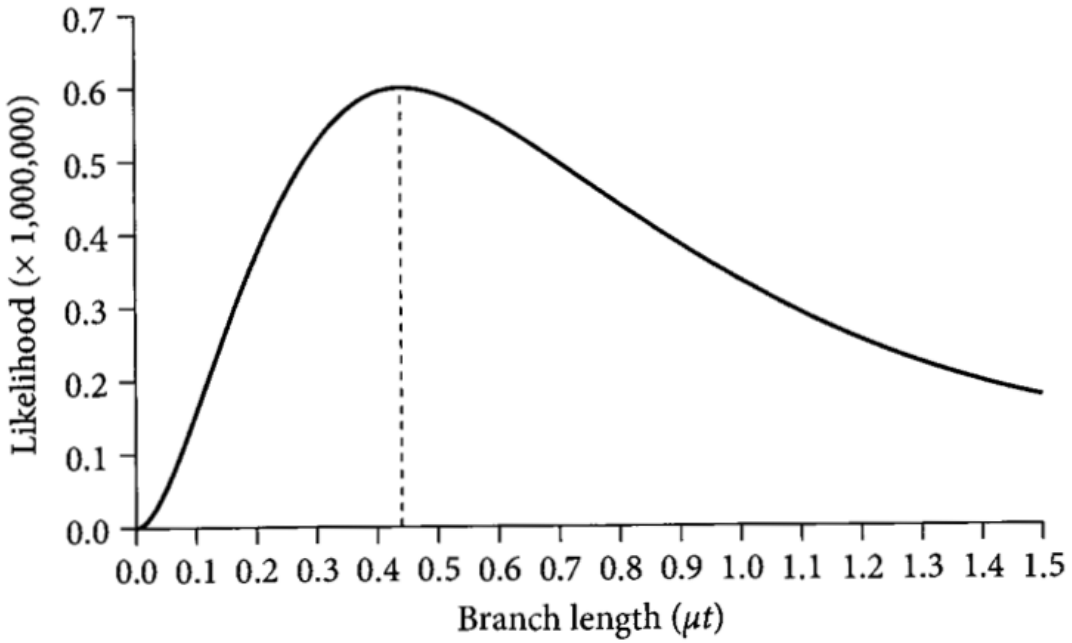



**FIGURE 8.17 Likelihood values for different branch lengths given the data shown in Figure 8.16.**

# Maximum Likelihood

The probability (site likelihood) of a mismatch is
$\frac{1}{4}(\frac{1}{4} - \frac{1}{4} e^{-4/3\mu t})$ – probability of a match: $\frac{1}{4}(\frac{1}{4} - \frac{3}{4} e^{-4/3\mu t})$

The likelihood of a tree knowing the entire characters matrix is the product of all site likelihoods.

$[\frac{1}{4}(\frac{1}{4} + \frac{3}{4} e^{-4/3\mu t})]^2 * [\frac{1}{4}(\frac{1}{4} + \frac{3}{4} e^{-4/3\mu t})]^4$

Internal node *a* has an unknown sequence. To get site likelihoods we sum over all possible nucleotides for each site and then multiply the site likelihoods.

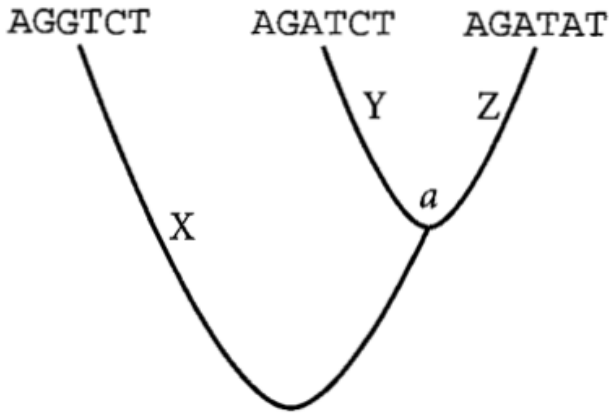
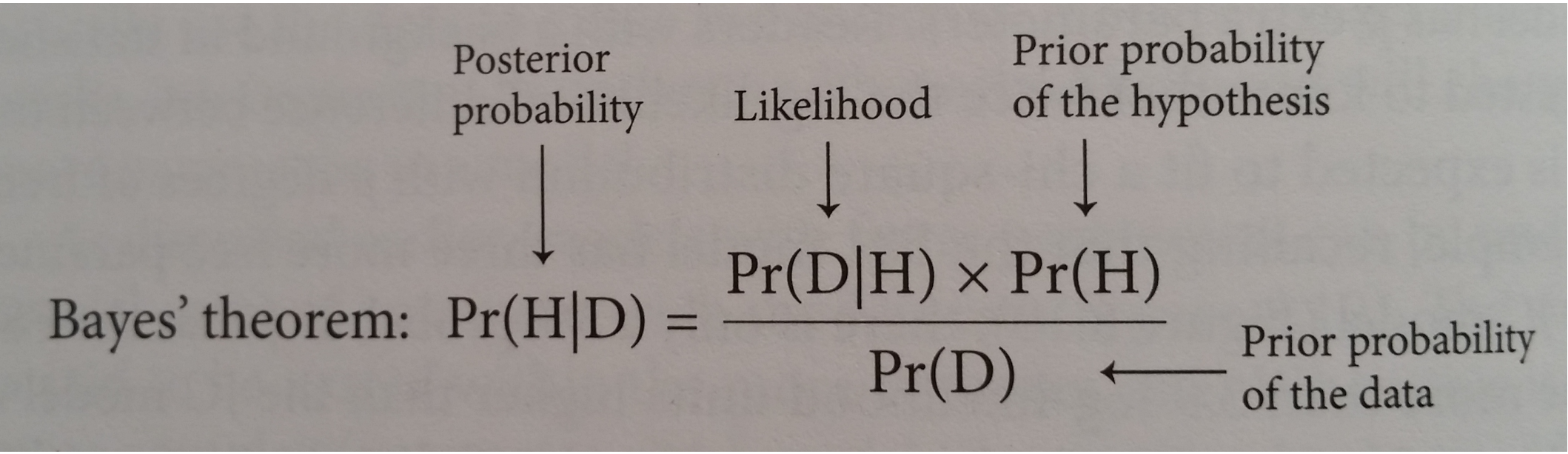



**FIGURE 8.18 Three-taxon tree with a six base-pair sequence at each tip.** The only items of uncertainty are the three branch lengths (X, Y, and Z) and the sequence at internal node *a*. We use the maximum likelihood criterion to estimate the value of the branch lengths, while summing over all possible sequences at node *a*.

# Bayesian Phylogenetics

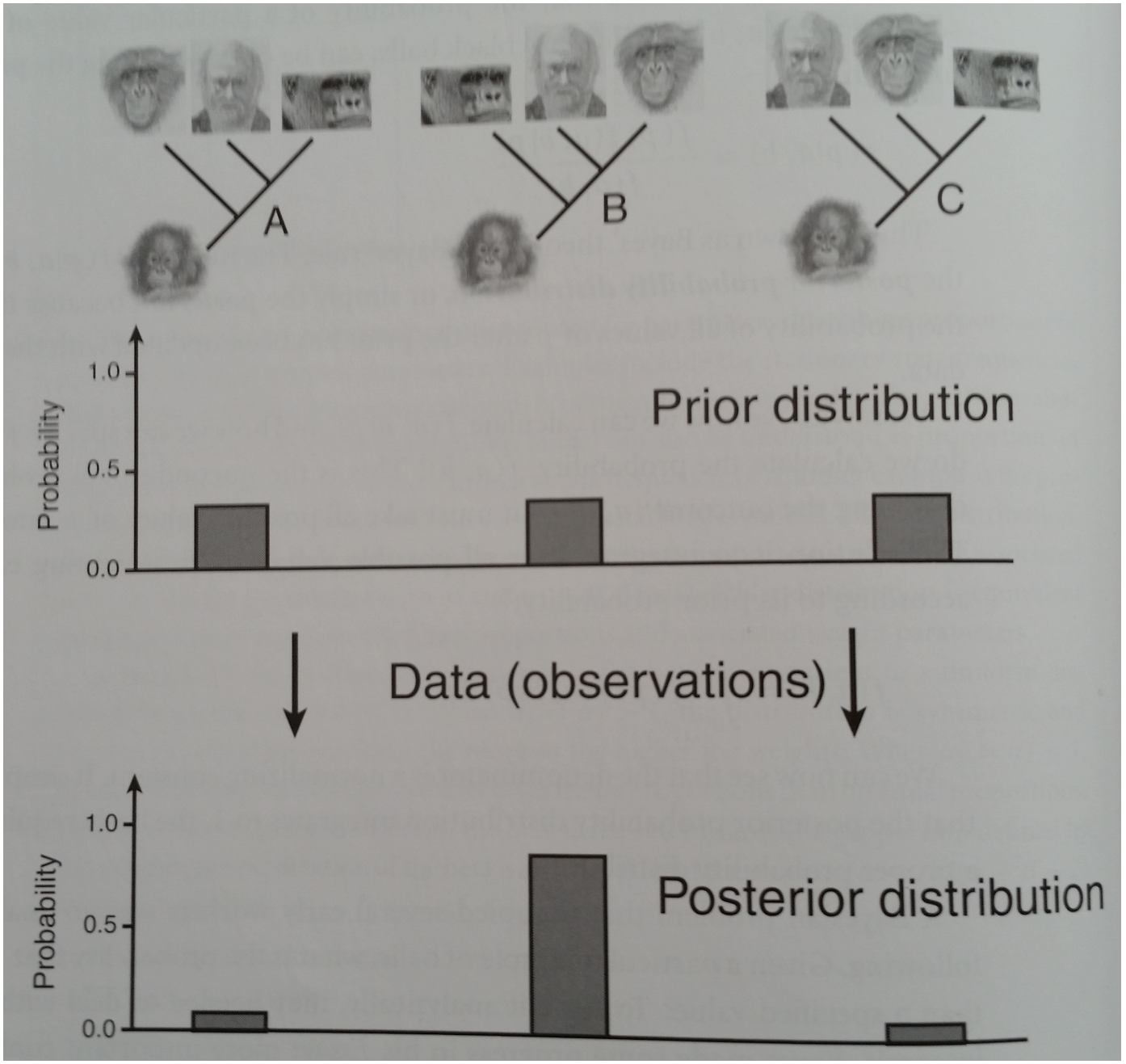
Probability of the hypothesis given the data, conditional on the prior probability of a hypothesis.

Recall the likelihood is the probability of the data given the hypothesis.



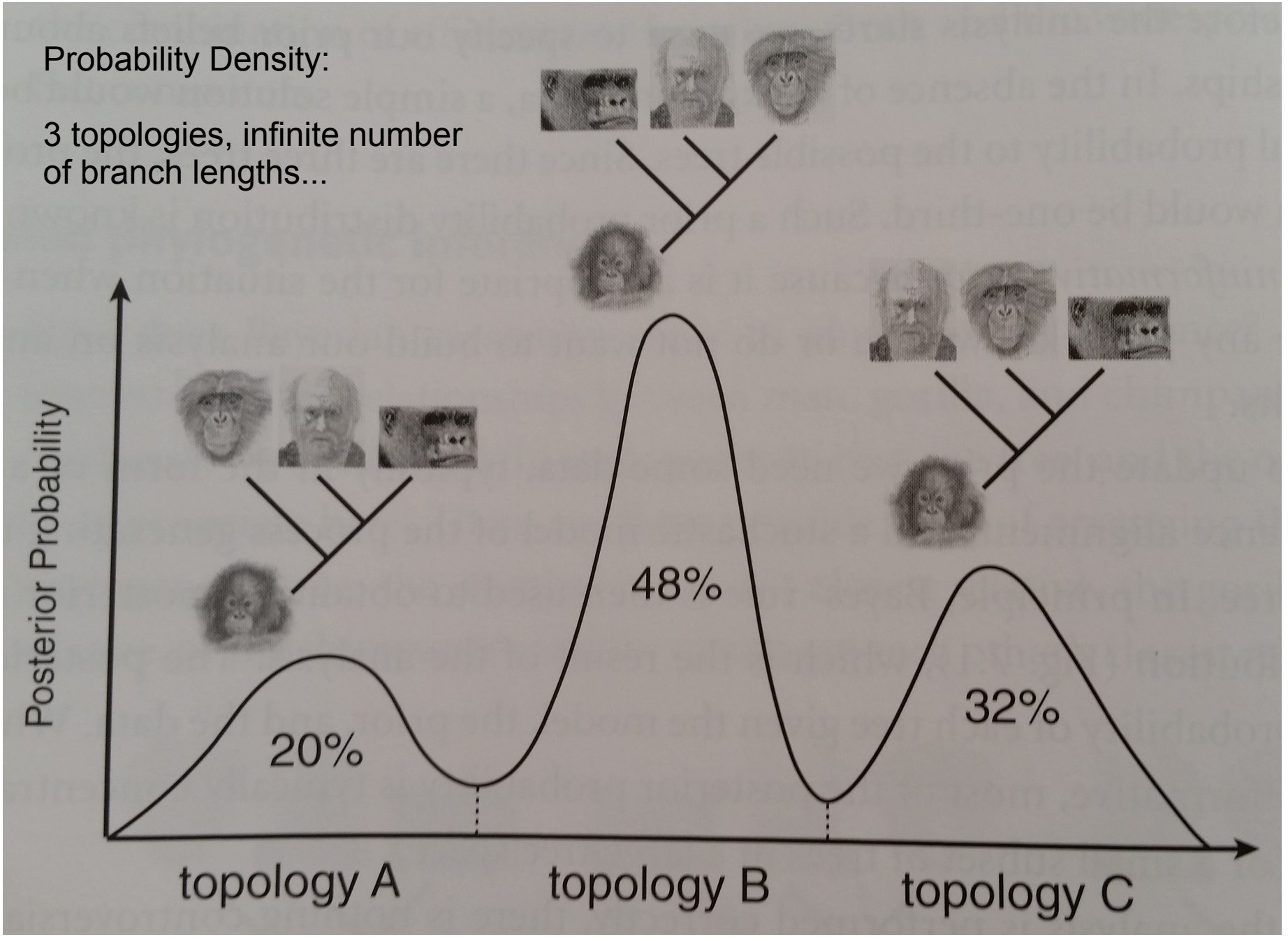

$$\text{Prob}(H \mid D) = \frac{\text{Prob}(H)\ \text{Prob}(D \mid H)}{\sum_H \text{Prob}(H)\ \text{Prob}(D \mid H)}$$

# Bayesian Phylogenetics

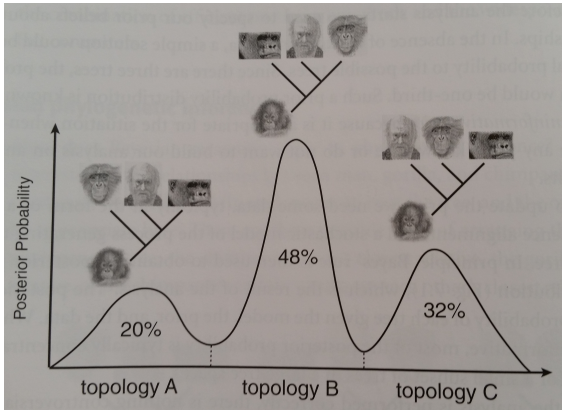



Uninformative, flat, or uniform prior

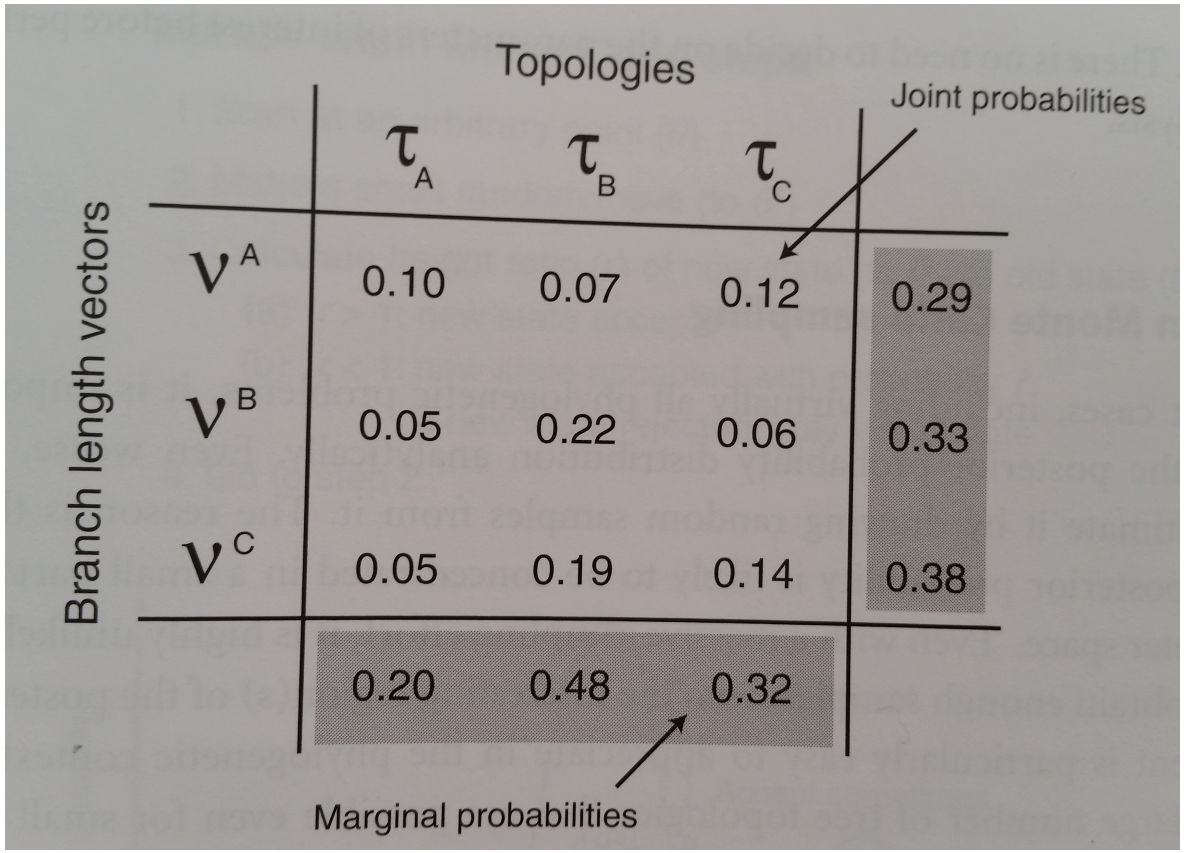
# Bayesian Phylogenetics



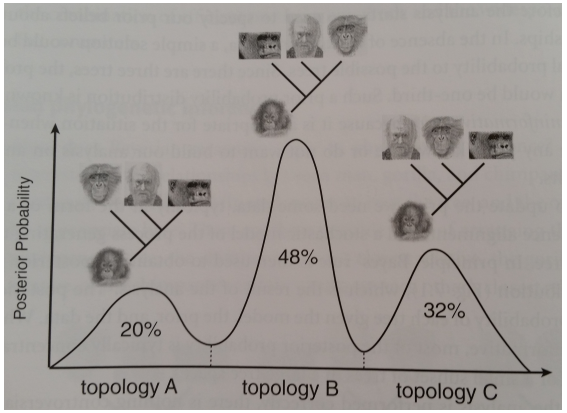
# Bayesian Phylogenetics




Integrate over branch lengths for each topology to derive the (marginal) probability of the topology



# Bayesian Phylogenetics




Integrate over branch lengths for each topology to derive the (marginal) probability of the topology

However, the branch lengths vector is literally infinitely long!

Topologies

Joint probabilities

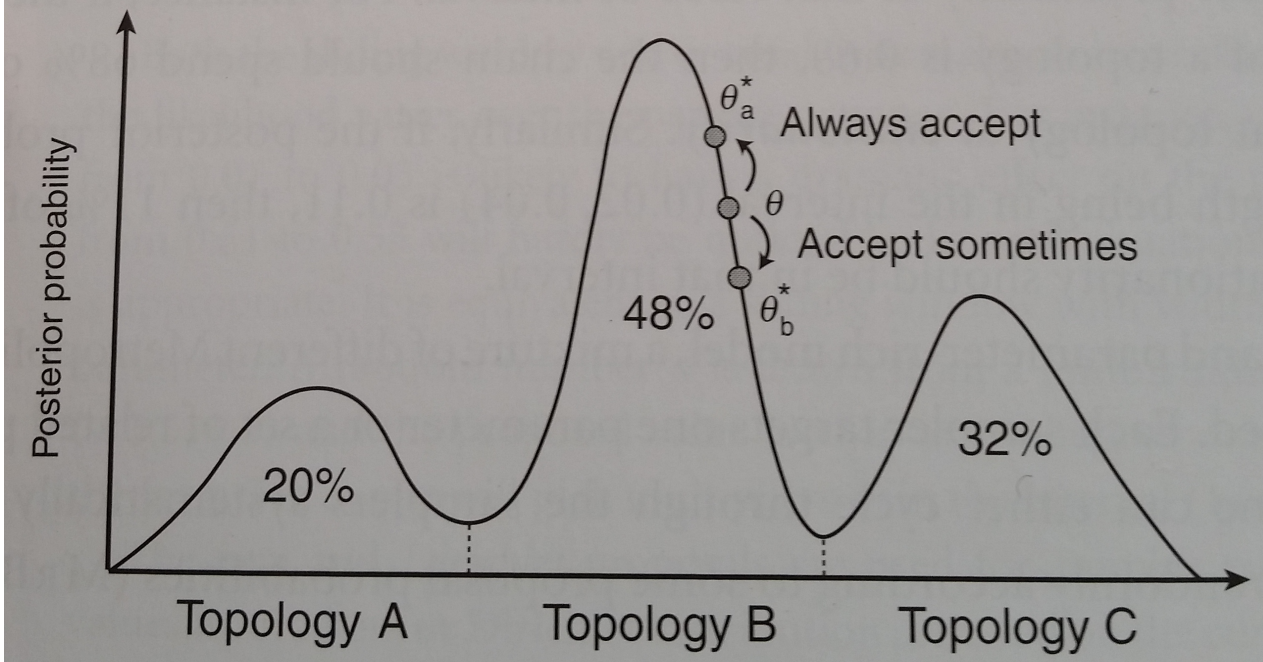
| Branch length vectors | $\tau_A$ | $\tau_B$ | $\tau_C$ | |
|---|---|---|---|---|
| $\nu^A$ | 0.10 | 0.07 | 0.12 | 0.29 |
| $\nu^B$ | 0.05 | 0.22 | 0.06 | 0.33 |
| $\nu^C$ | 0.05 | 0.19 | 0.14 | 0.38 |
| | 0.20 | 0.48 | 0.32 | |

Marginal probabilities

# Bayesian Phylogenetics

**Markov chain Monte Carlo steps**

1. Start at an arbitrary point ($\theta$)
2. Make a small random move (to $\theta^*$)
3. Calculate height ratio ($r$) of new state (to $\theta^*$) to old state ($\theta$)
   (a) $r > 1$: new state accepted
   (b) $r < 1$: new state accepted with probability $r$
   if new state rejected, stay in old state
4. Go to step 2




Better fitting tree topologies are will be more frequent in the posterior sample than ill-fitting ones...

MCMC guaranteed to find posterior if run long enough

MCMC robot with 1 chain

# How do we know how good our best tree is?

Bayesian phylogenetics gives us a probability distribution of topologies.

What about our confidence in the maximum likelihood or parsimony tree?

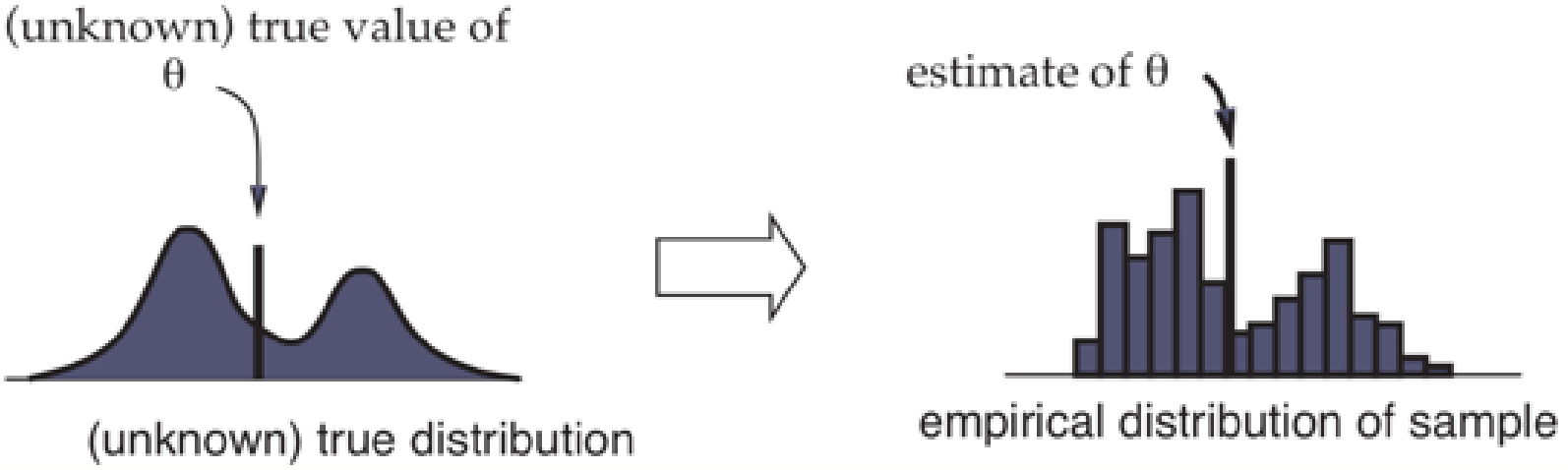
# Evaluating a tree: the bootstrap

Data in hand are probably somewhat like the underlying universe of possible data sets

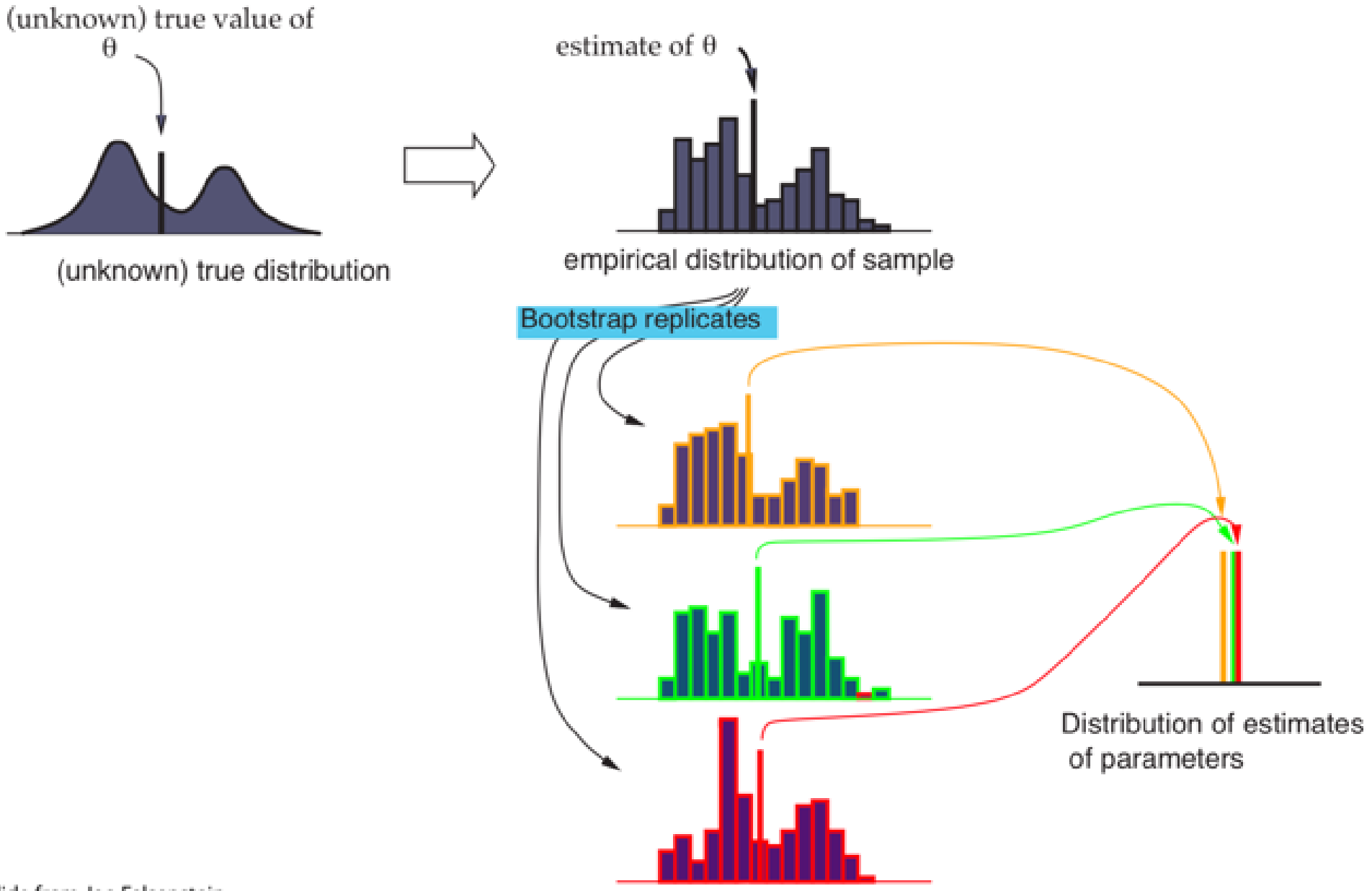
We can simulate other possible data sets by drawing randomly from the data already collected

This seems to get information from nothing, lifting yourself off the ground by your bootstraps...

# The bootstrap

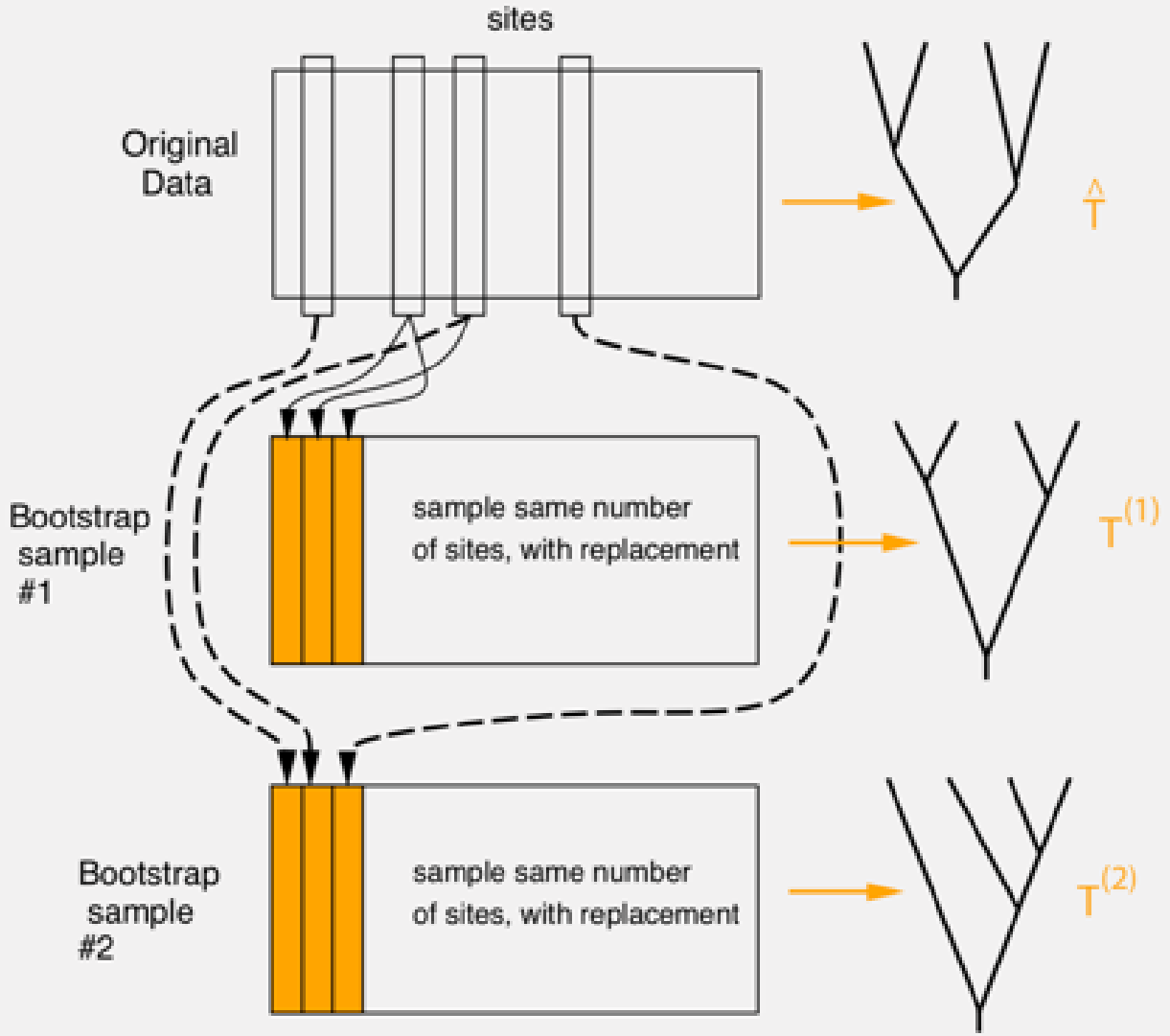


# The bootstrap




Slide from Joe Felsenstein

# The bootstrap for phylogenies




Slide from Joe Felsenstein

## The simplest phylogenetic test would compare two trees

Null: If we had no sampling error (infinite data) $T_1$ and $T_2$ would explain the data equally well.

Test Statistic:

$$\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$$

Expectation under null:

$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

**KH Test**

H0: The two trees are equally supported

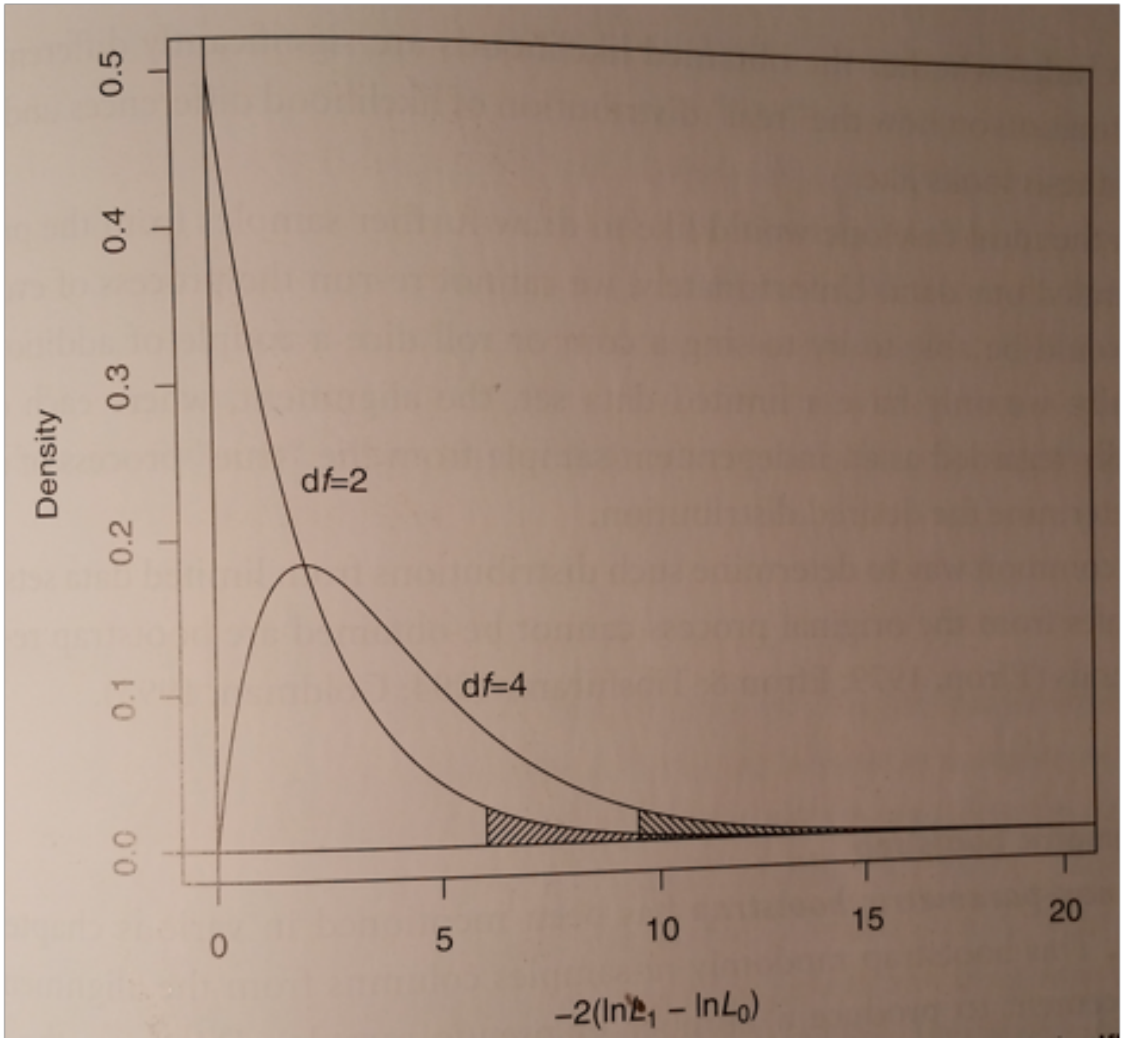
HA: The two trees are not equally supported

**SH Test**

H0: All trees (including the ML tree) are equally good explanations of the data

HA: Some or all trees are not equally good explanations of the data

# Comparing phylogenetic trees using log likelihood ratios




We could use log likelihood ratios to test for fit of substitution model since the models are nested and the test statistic is approx. chi2 distributed.

However, trees are not nested in each other!

## The simplest phylogenetic test would compare two trees

Null: If we had no sampling error (infinite data) $T_1$ and $T_2$ would explain the data equally well.

Test Statistic:

$$\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$$

Expectation under null:

$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

## KH Test

1. Examine the difference in $\ln L$ for each site: $\delta(T_1, T_2 \mid X_i)$ for site $i$.
2. Note that the total difference is simply a sum:

$$\delta(T_1, T_2 \mid X) = \sum_{i=1}^{M} \delta(T_1, T_2 \mid X_i)$$

3. The variance of $\delta(T_1, T_2 \mid X)$ will be a function of the variance in "site" $\delta(T_1, T_2 \mid X_i)$ values.

## RELL bootstrap

Often, the MLE of numerical parameters (including branch lengths) do not change much when we bootstrap.

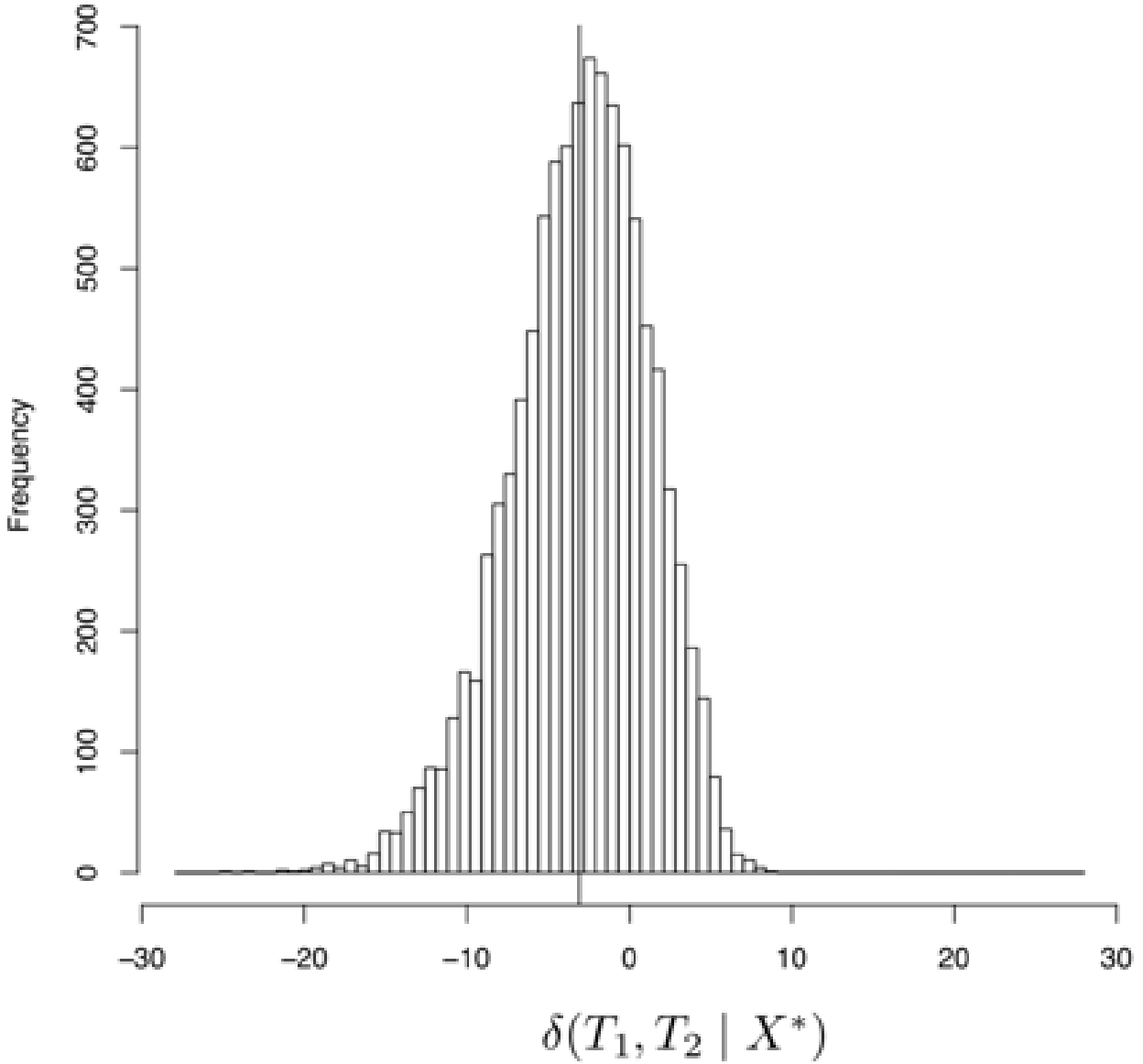
So, we can simply resample the site $\ln L$ values and sum them (rather than reoptimizing parameters).

This is called the RELL bootstrap (Kishino et al., 1990, and Felsenstein). It is not a "safe" replacement for normal bootstrapping (especially on large trees; Stamatakis et al., 2008) when you want to estimate clade support.

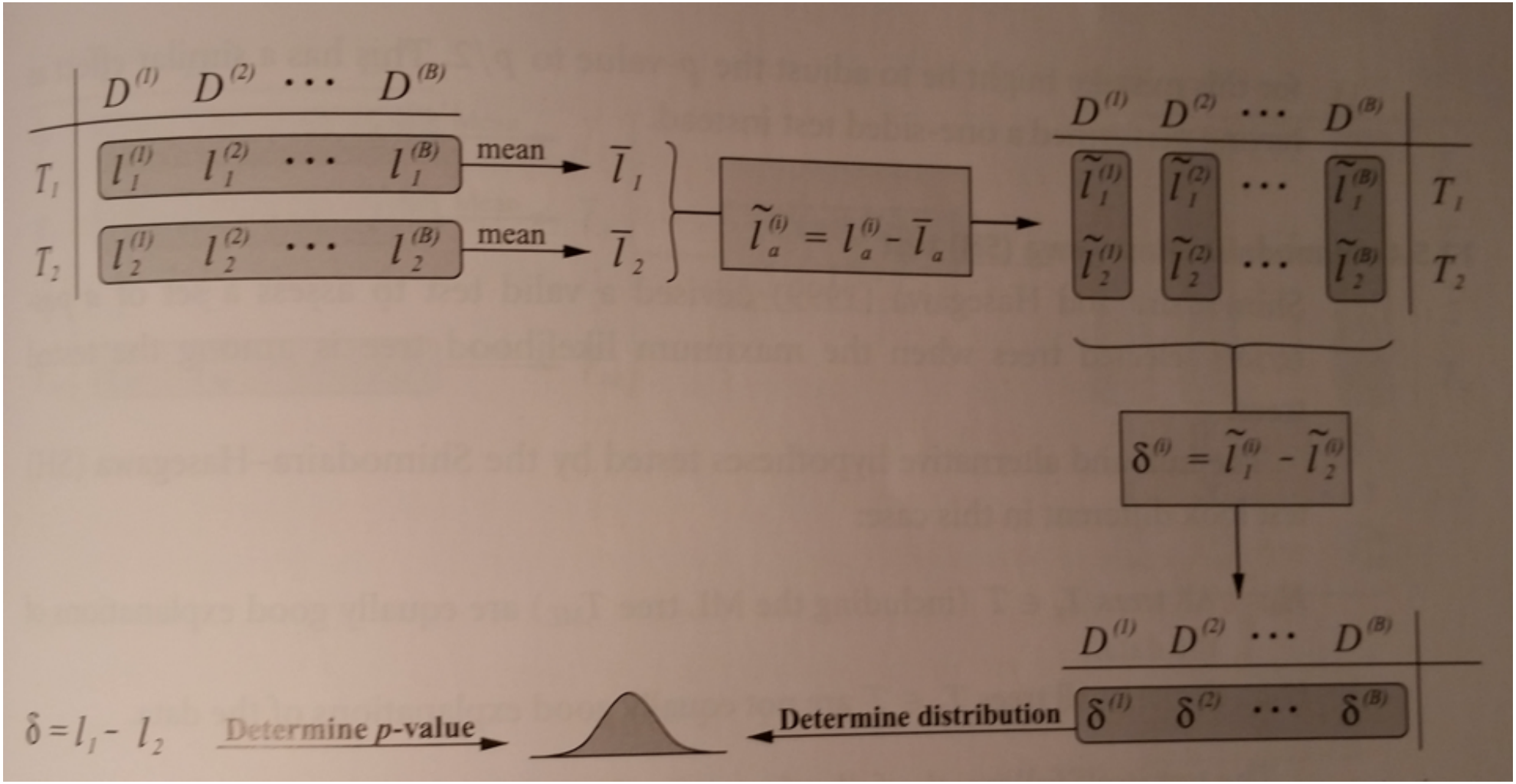
But it should be good enough for helping us learn about the standard error of the $\ln L$.

And it is really fast.

The (RELL) bootstrapped sample of statistics.
Is this the null distribution for our $\delta$ test statistic?



# KH Test Summary



**What if start out with only one hypothesized tree, and we want to compare it to the ML tree?**

---

The KH Test is **NOT** appropriate in this context (see Goldman et al., 2000, for discussion of this point)

**Multiple Comparisons**: lots of trees increases the variance of $\delta(\hat{T}, T_1 \mid X)$

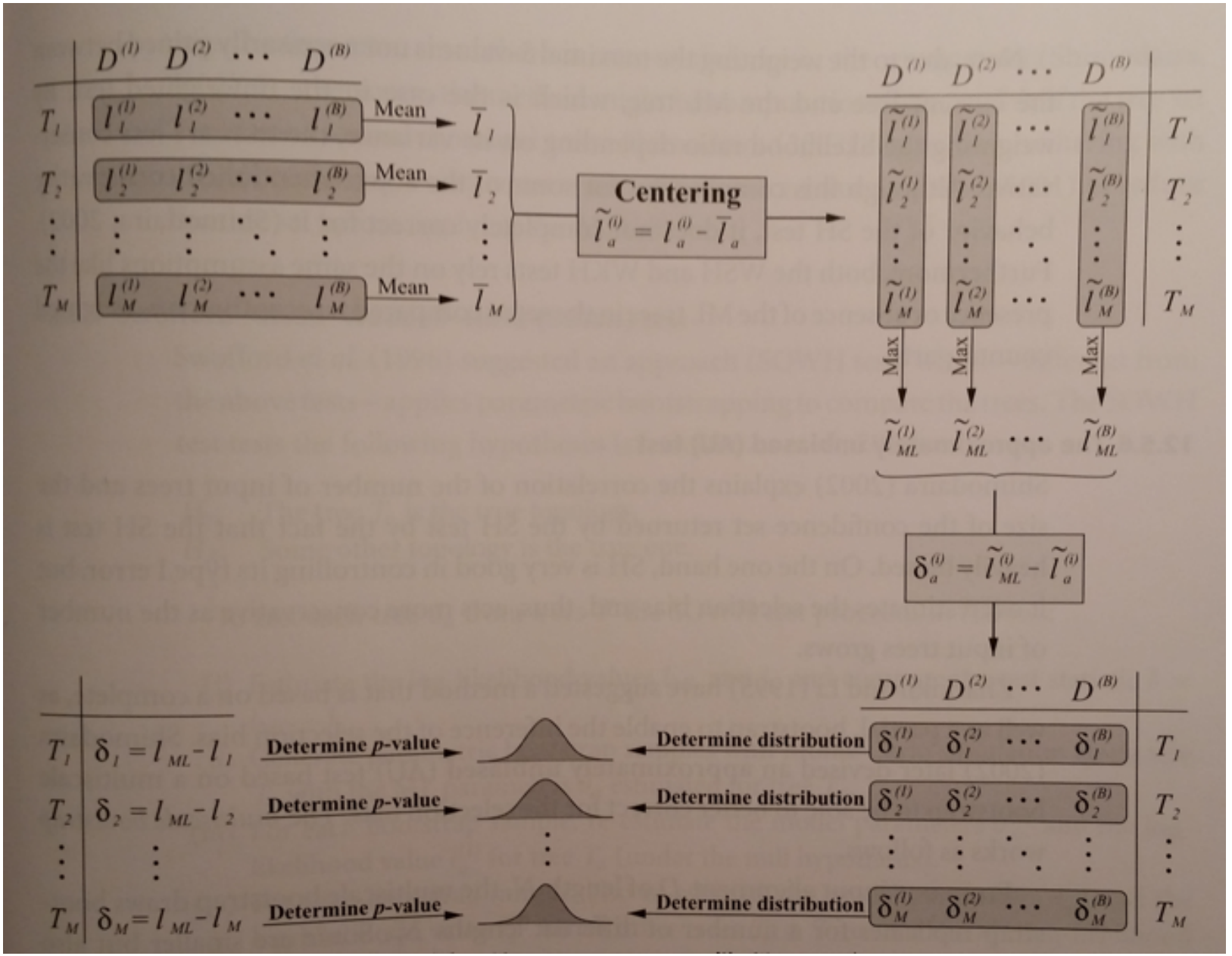
**Selection bias**: Picking the ML tree to serve as one of the hypotheses invalidates the centering procedure of the KH test.

## Using the ML tree in your test introduces selection bias

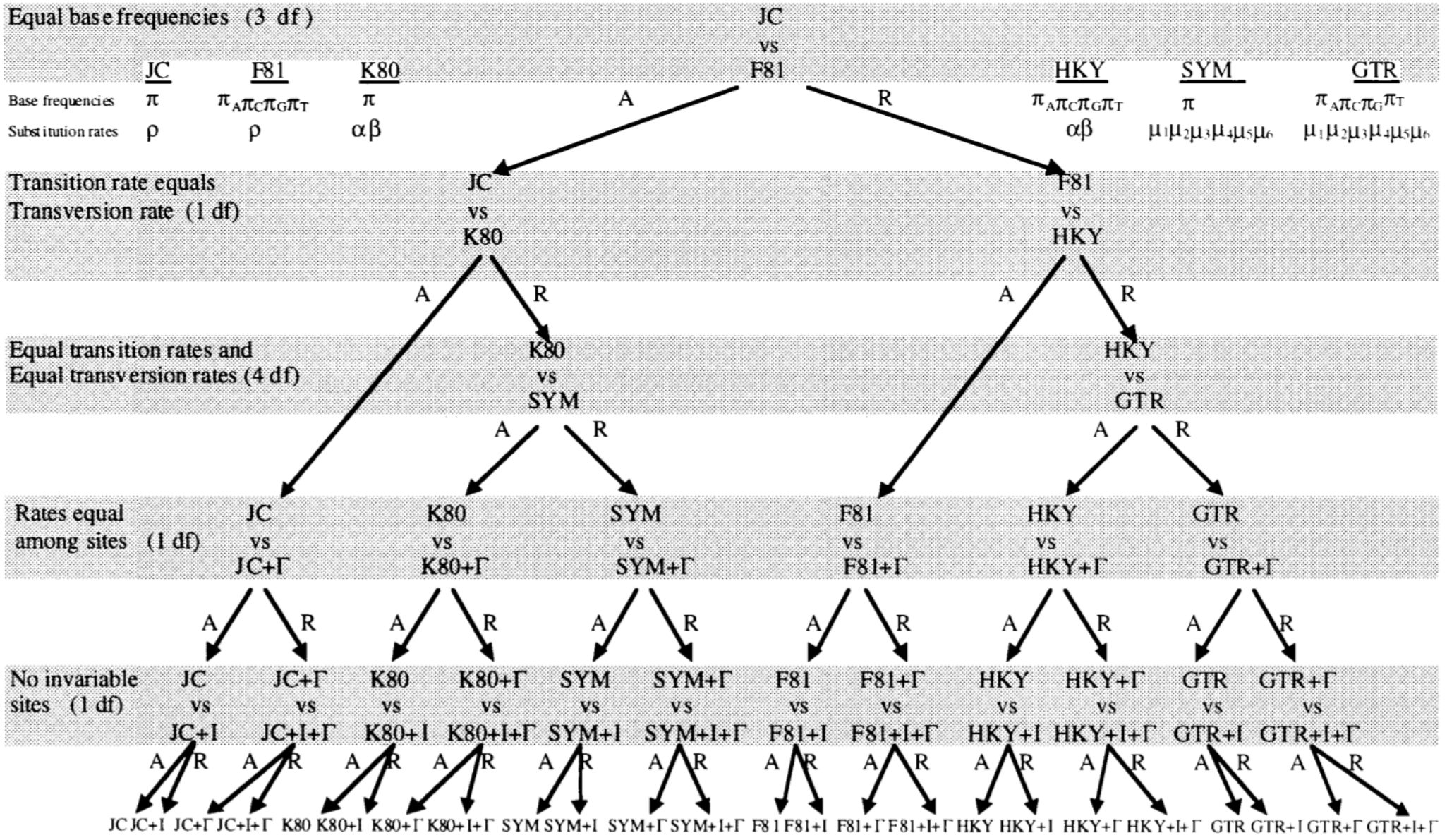
Even when the $H_0$ is true, we do not expect $2\left[\ln L(\hat{T}) - \ln L(T_1)\right] = 0$

Imagine a competition in which a large number of equally skilled people compete, and you compare the score of one competitor against the highest scorer.

# SH Test Summary



# We haven't talked about selecting the best substitution model yet!



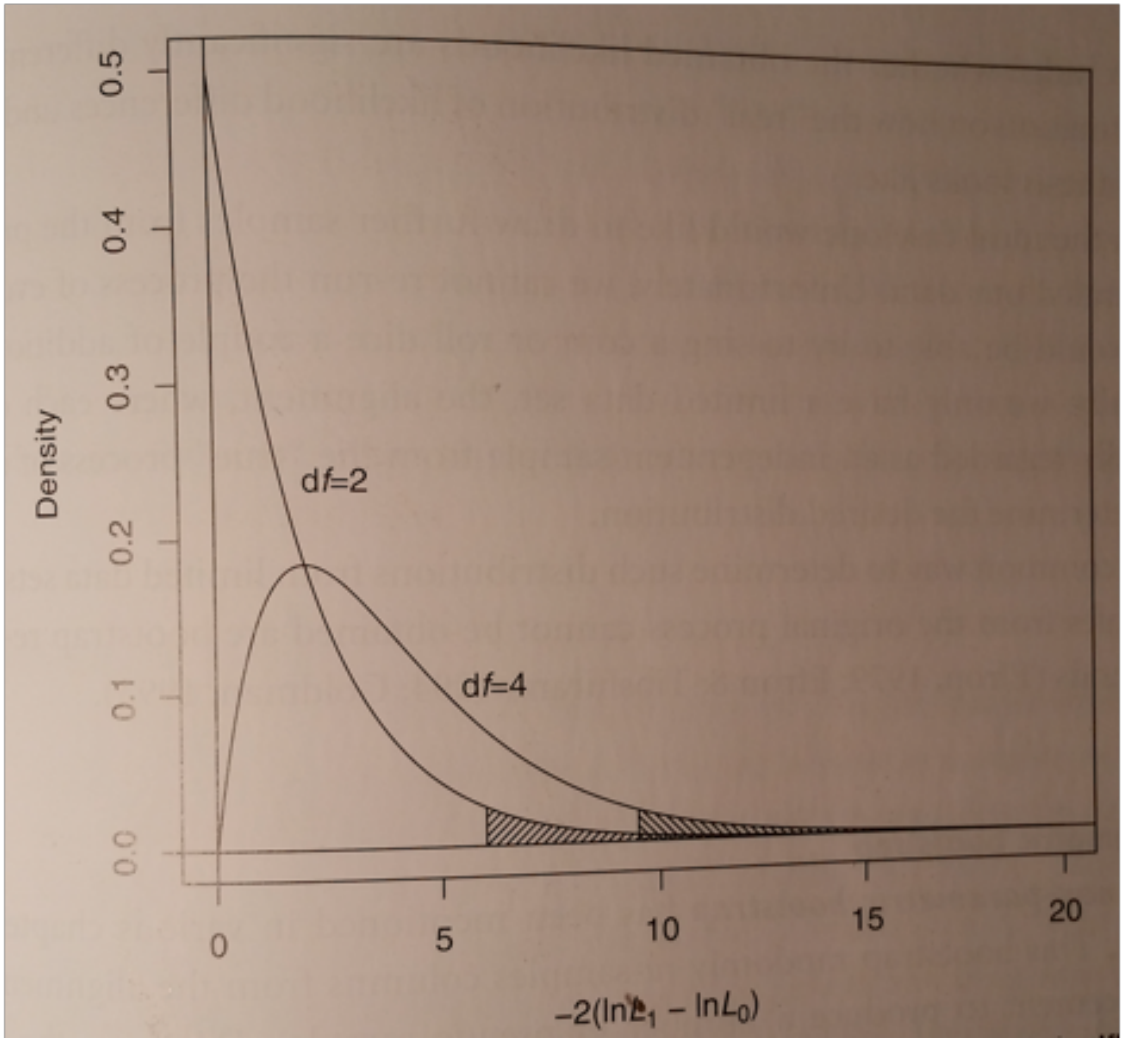
# Testing Model Fit

**Likelihood Ratio for comparing trees versus models of evolution. When comparing trees (T1 and T2) the trees vary. When comparing substitution models (second equation) the tree is held constant but the model changes.**

Test Statistic:

$$\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$$

$$\wedge = \frac{\max\ [L_0\ (Null\ \ Model \mid Data)]}{\max\ [L_1\ (Alternative\ \ Model \mid Data)]}$$




Nested model comparison using the Likelihood ratio test statistic are chi2 distributed. Using the number of additional model parameters in the alternate models as degrees of Freedom, we can find the p-value.

# Testing Model Fit

**Akaike Information Criterion (AIC)**

Suppose data generated by some unknown process **f**. Two candidate models represent **f**: **g1** and **g2**.

If we knew **f**, then we could find the information lost from using **g1**; similarly, the information lost from using **g2** to represent **f** could be found. We would then choose the candidate model that minimized the information loss.

We do not know **f**. We can estimate how much more (or less) information is lost by g1 than by g2; ie, which model fits the data better (Akaike 1974)

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

k is the number of estimated parameters; ln(L) the log likelihood

**Smaller AIC indicates better fit of the model to the data**