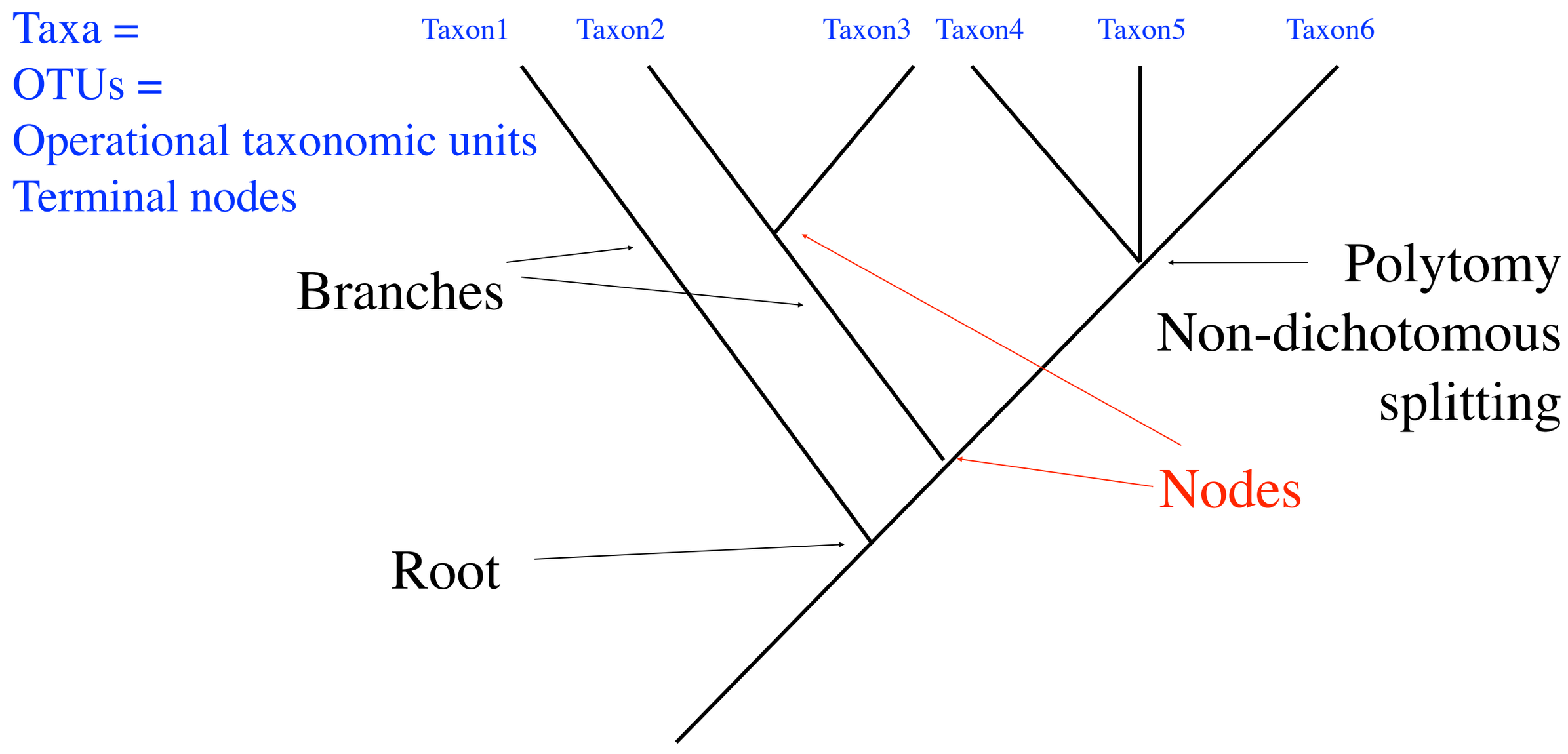
Week 4

Trees and Phylogenetics

Anatomy of a phylogenetic tree



Rooted and unrooted trees

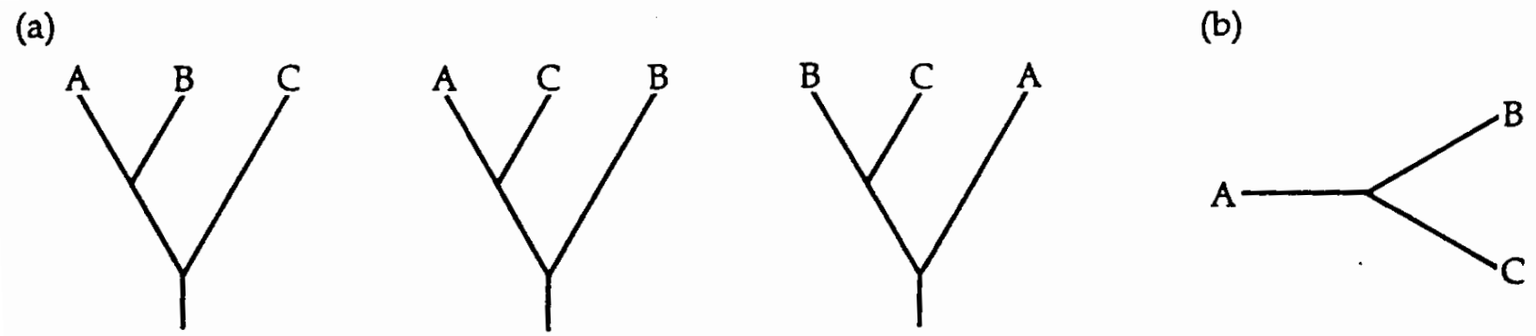


Figure 3. From three OTUs it is possible to construct three different rooted trees (a), but only one unrooted tree (b).

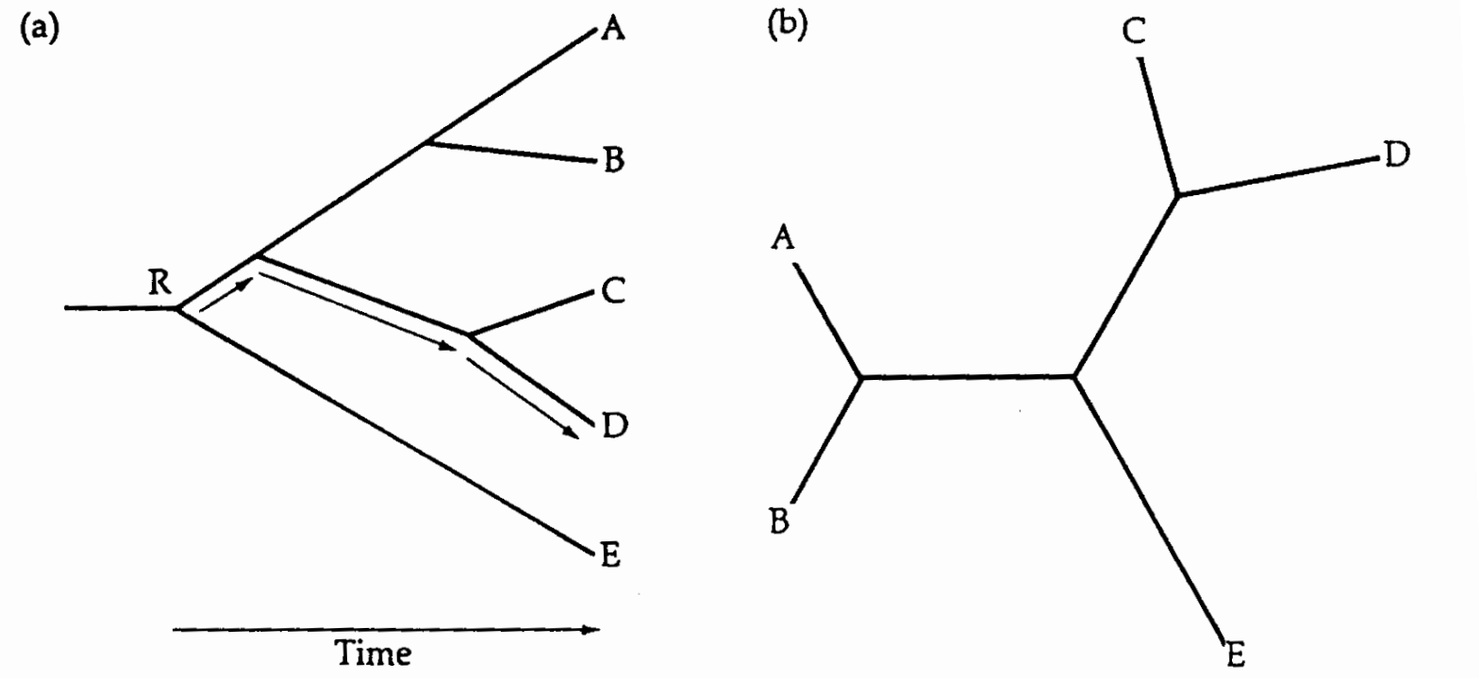


Figure 2. (a) Rooted and (b) unrooted phylogenetic trees. Arrows indicate the unique path leading from the root (R) to OTU D.

Tree shape vs. tree topology

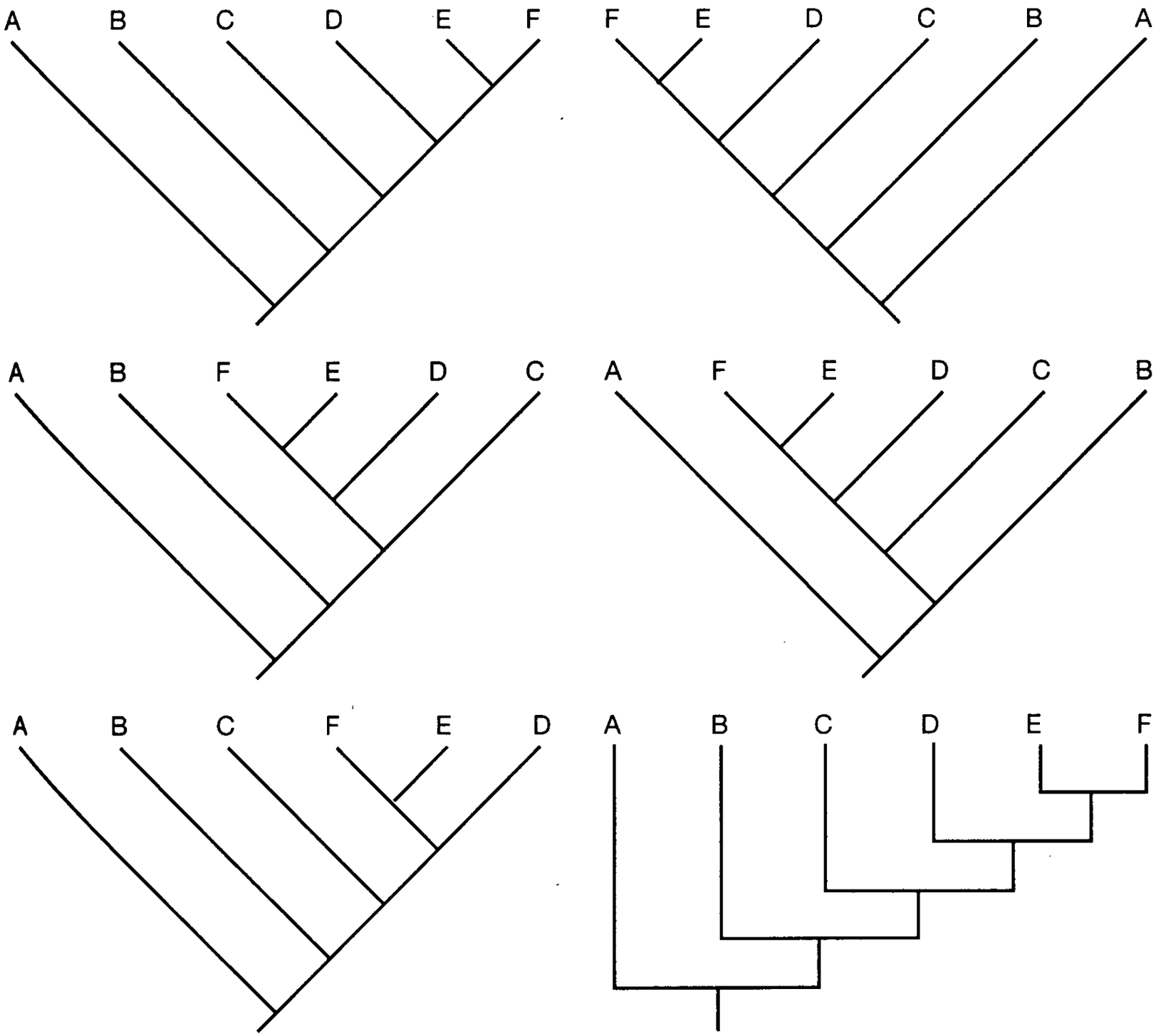


***Figure 10.2* Alternative ways of drawing the same tree**
These graphs show six ways to illustrate the same evolutionary relationships. These trees all happen to be oriented vertically, so that the basal taxon is at the bottom and derived groups are toward the top, but they could also be tipped 90° and presented horizontally. From Mayden and Wiley (1992).

Phylogenetic reconstruction: many possible trees from several taxa

The numbers of phylogenetic trees expands rapidly with the number of taxa (OTUs)

Note that rooted trees have the equivalent of one additional taxon, defined by the root

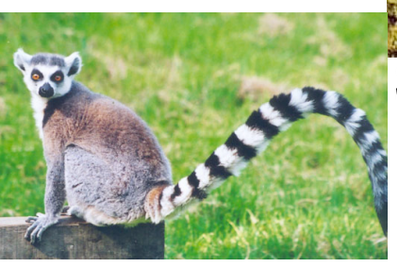
$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad (5.1)$$

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!} \quad (5.2)$$

Table 1. Numbers of possible rooted and unrooted trees for 1–10 OTUs.

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

From Felsenstein (1978).



Lemurs

Spider monkey

New World monkeys

macaque

Old World monkeys

Gibbon

Orang

Gorilla

Chimp

Bonobo

Australo-pithecus

Homo

~6 MYA

~7 MYA Hominins

12-15 MYA Hominids

15-18 MYA Hominoids

25 MYA Catarhini

40 MYA

63 MYA

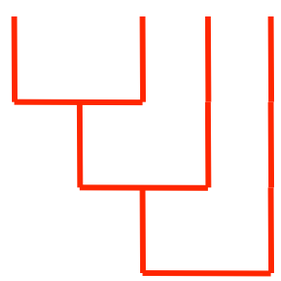
Primate Phylogeny

distance (phenetic) method of reconstructing a phylogenetic tree

- Collect as many characters as possible
- Construct a pair-wise distance between each taxon in your study
  - Euclidian (mophological)
  - Sequence divergence (molecular)
- Cluster taxa based on greatest overall similarity (smallest distance)
- Assumes homology outweighs analogy

Distance matrix

	Taxon1	Taxon2	Taxon3	Taxon4
Taxon1	0.0			
Taxon2	0.1	0.0		
Taxon3	0.2	0.3	0.0	
Taxon4	0.5	0.5	0.5	0.0



Distance phylogram

The distance method of UPGMA (unweighted pair group method with averages)

Pairs of taxa with the smallest distance score are clustered, and the remaining values. In the table are averaged to create a smaller table with one fewer columns e.g., human and chimp are paired, and the new distance of Gorilla to the average of human and chimp (HC) is 1.54 (= average of 1.51 &1.57).

Table 2. Mean (below diagonal) and standard error (above diagonal) of the number of nucleotide substitutions per 100 sites between OTUs.[a]

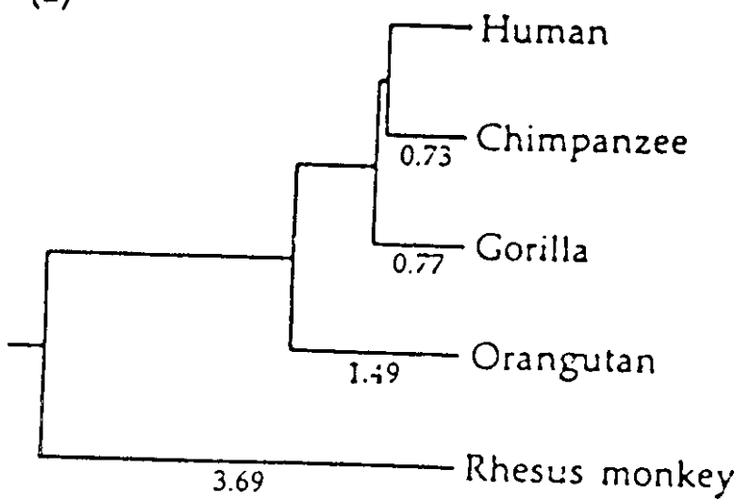
OTU	OTU				
	Human	Chimpanzee	Gorilla	Orangutan	Rhesus monkey
Human		0.17	0.18	0.25	0.41
Chimpanzee	1.45		0.18	0.25	0.42
Gorilla	1.51	1.57		0.26	0.41
Orangutan	2.98	2.94	3.04		0.40
Rhesus monkey	7.51	7.55	7.39	7.10	

From Li et al. (1987b).

[a] The sequence data used are 5.3 kb of noncoding DNA, which is made up of two separate regions: (1) the η-globin locus (2.2 kb) described by Koop et al. (1986b) and (2) 3.1 kb of the η-δ globin intergenic region sequenced by Maeda et al. (1983, 1988).

	OTU		
OTU	(HC)	G	O
G	1.54		
O	2.96	3.04	
R	7.53	7.39	7.10

(a)



UPGMA

	Human	Chimp	Gorilla	Orangu tan	Rhesus
Human	-				
Chimp	1.45	-			
Gorilla	1.51	1.57	-		
Orangu tan	2.98	2.94	3.04	-	
Rhesus	7.51	7.55	7.39	7.10	-

1) Find Smallest Cell in distance matrix

2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

distances between new node n and existing node i

$$d(n,i) = \frac{(d(a,i) + d(b,i))}{2}$$

UPGMA

	Human	Chimp	Gorilla	Orangu tan	Rhesus
Human	-				
Chimp	1.45	-			
Gorilla	1.51	1.57	-		
Orangu tan	2.98	2.94	3.04	-	
Rhesus	7.51	7.55	7.39	7.10	-

1) Find Smallest Cell in distance matrix

2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

distances between new node n and existing node i

$$d(n,i) = \frac{(d(a,i) + d(b,i))}{2}$$

UPGMA

	(H,C)	Gorilla	Orangutan	Rhesus
(H,C)	-			
Gorilla	1.54	-		
Orangutan	2.96	3.04	-	
Rhesus	7.53	7.39	7.10	-

1) Find Smallest Cell in distance matrix

2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

distances between new node n and existing node i

$$d(n,i) = \frac{(d(a,i) + d(b,i))}{2}$$

UPGMA

	((H,C),G)	Orangutan	Rhesus
((H,C),G)	-		
Orangutan	3.0	-	
Rhesus	7.46	7.10	-

1) Find Smallest Cell in distance matrix

2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

distances between new node n and existing node i

$$d(n,i) = \frac{(d(a,i) + d(b,i))}{2}$$

UPGMA

	(((H,C),G),O)	Rhesus
(((H,C),G),O)	-	
Rhesus	7.28	-

1) Find Smallest Cell in distance matrix

2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

distances between new node n and existing node i

$$d(n,i) = \frac{(d(a,i) + d(b,i))}{2}$$

UPGMA

	(((H,C),G),O)	Rhesus
(((H,C),G),O)	-	
Rhesus	7.28	-

1) Find Smallest Cell in distance matrix

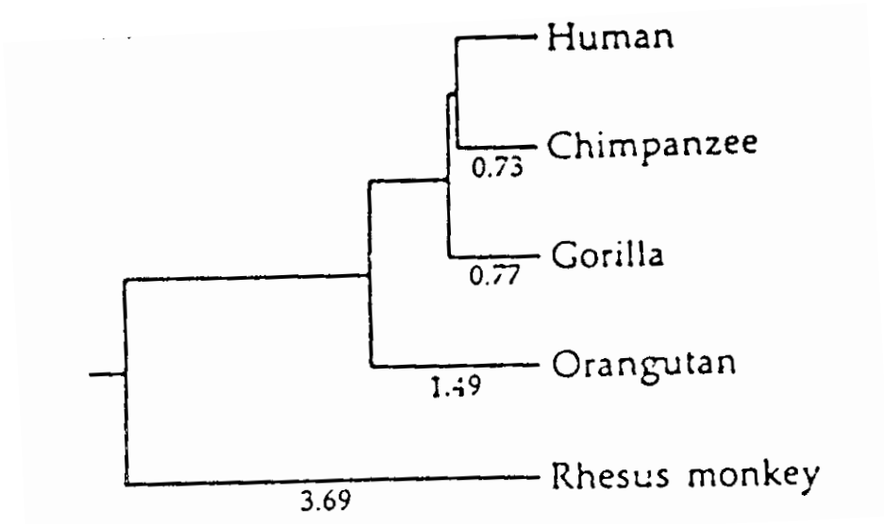
2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

So final tree:

(((H,C),G),O)



Neighbor Joining

	Human	Chimp	Gorilla	Orangu tan	Rhesus
Human	-				
Chimp	1.45	-			
Gorilla	1.51	1.57	-		
Orangu tan	2.98	2.94	3.04	-	
Rhesus	7.51	7.55	7.39	7.10	-

1) Find Smallest Cell in **transformed** distance matrix

2) Collapse those taxa into node

3) Recompute distance matrix

4) Repeat until one node left

distances between new node n and existing node i

$$d(n,i) = \frac{(d(a,i) + d(b,i) - d(a,b)}{2}$$

Neighbor Joining

	Human	Chimp	Gorilla	Orangu tan	Rhesus
Human	-				
Chimp	1.45	-			
Gorilla	1.51	1.57	-		
Orangu tan	2.98	2.94	3.04	-	
Rhesus	7.51	7.55	7.39	7.10	-

1) Find Smallest Cell in **transformed** distance matrix

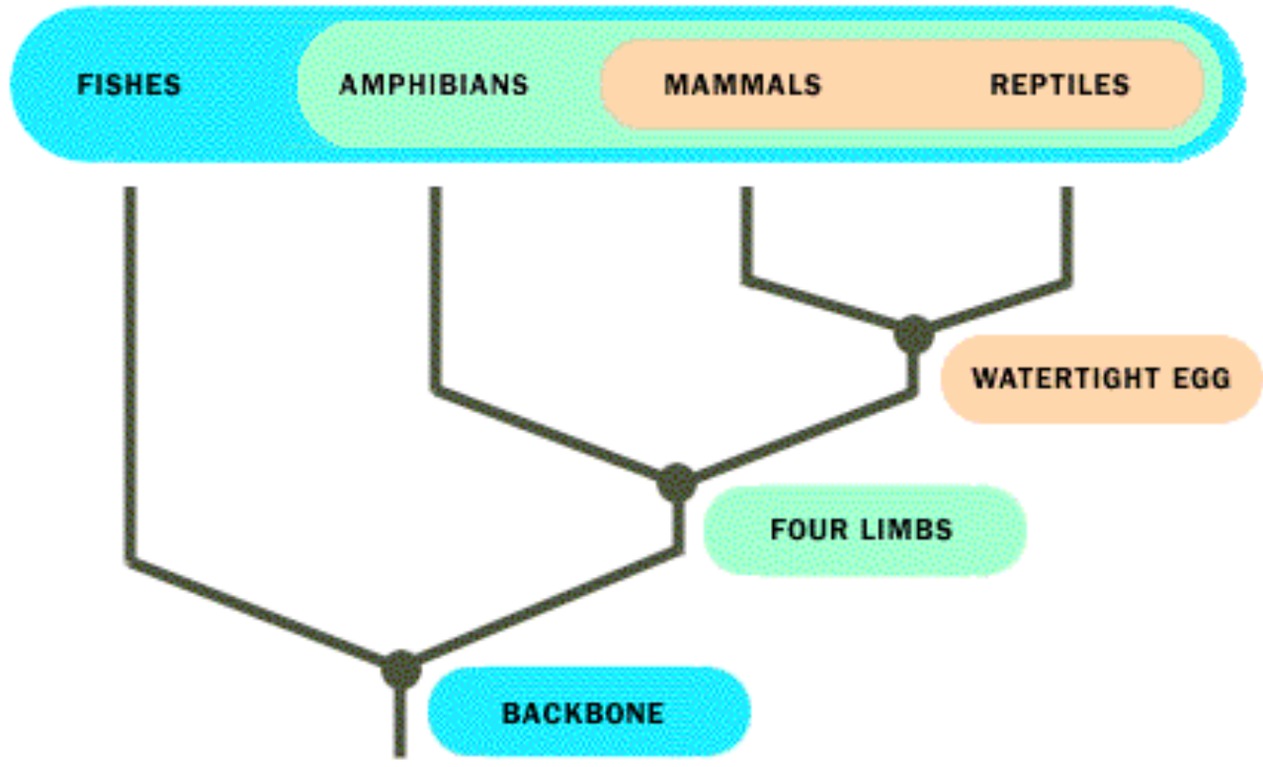
Major advantage is don't assume rates same across branches

$$Q(i,j) = (n-2) \times d(i,j) - \sum_{l=1}^{k} d(i,l) - \sum_{l=1}^{k} d(j,l)$$

Calculate UPGMA trees in jupyter notebook

Cladistic method of reconstructing a phylogenetic tree

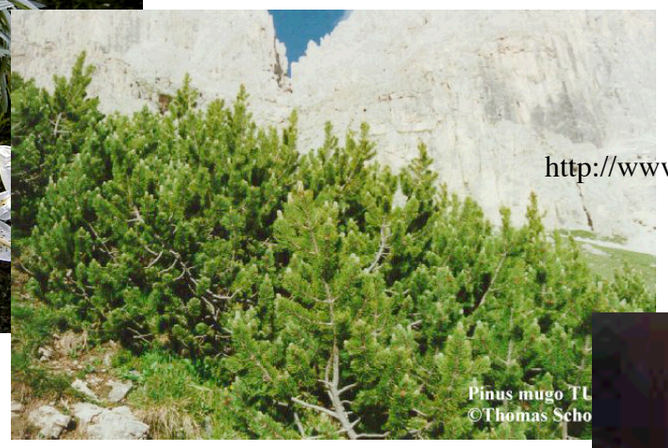
- Collect characters and character states
- Build a character state matrix
- Cluster organisms according to patterns of shared-derived characters
- Use parsimony to determine best tree



http://www.amnh.org/Exhibition/Fossil_Halls/cladistics.html

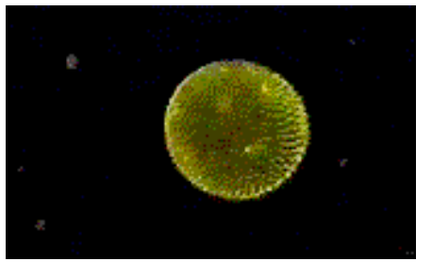
Synapomorphy = Shared, derived trait

http://www.plant-pictures.com/



http://www.home.aone.net.au/byzantium/ferns/about.html

http://home.clara.net/adhale/bryos/phframe.htm



http://www.microscopy-uk.org.uk/mag/art97b/volvoxms.html

Dicots

Monocots

Gymnosperms

ferns

Flowers

Seeds

Bryophytes (mosses)

Vascular Conducting system

Green Algae

Chlorophyll

Characters and Character states

- Character ~ trait
- Character state = alternative forms of a trait
- Eye color = character
- Blue, brown = character states

Character	A	B	C	D	E
Species 1	1	0	0	0	0
Species 2	1	1	1	0	0
Species 3	1	0	1	1	1
Species 4	1	0	0	1	1

Uninformative characters

conflicting character

Informative characters

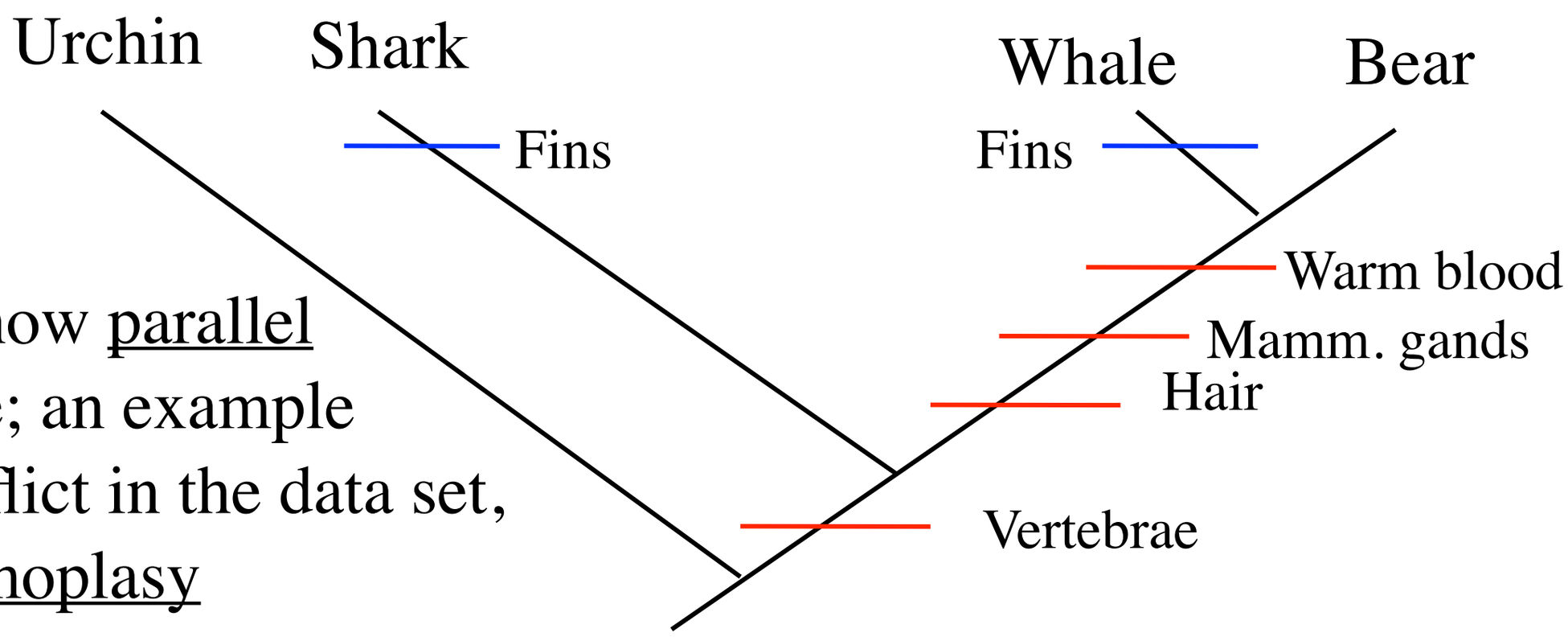
Uninformative = invariant, or in a single taxon

Conflicting = cluster taxa in a way that conflicts with clustering of other characters

Informative = both states shared by ≥ 2 taxa

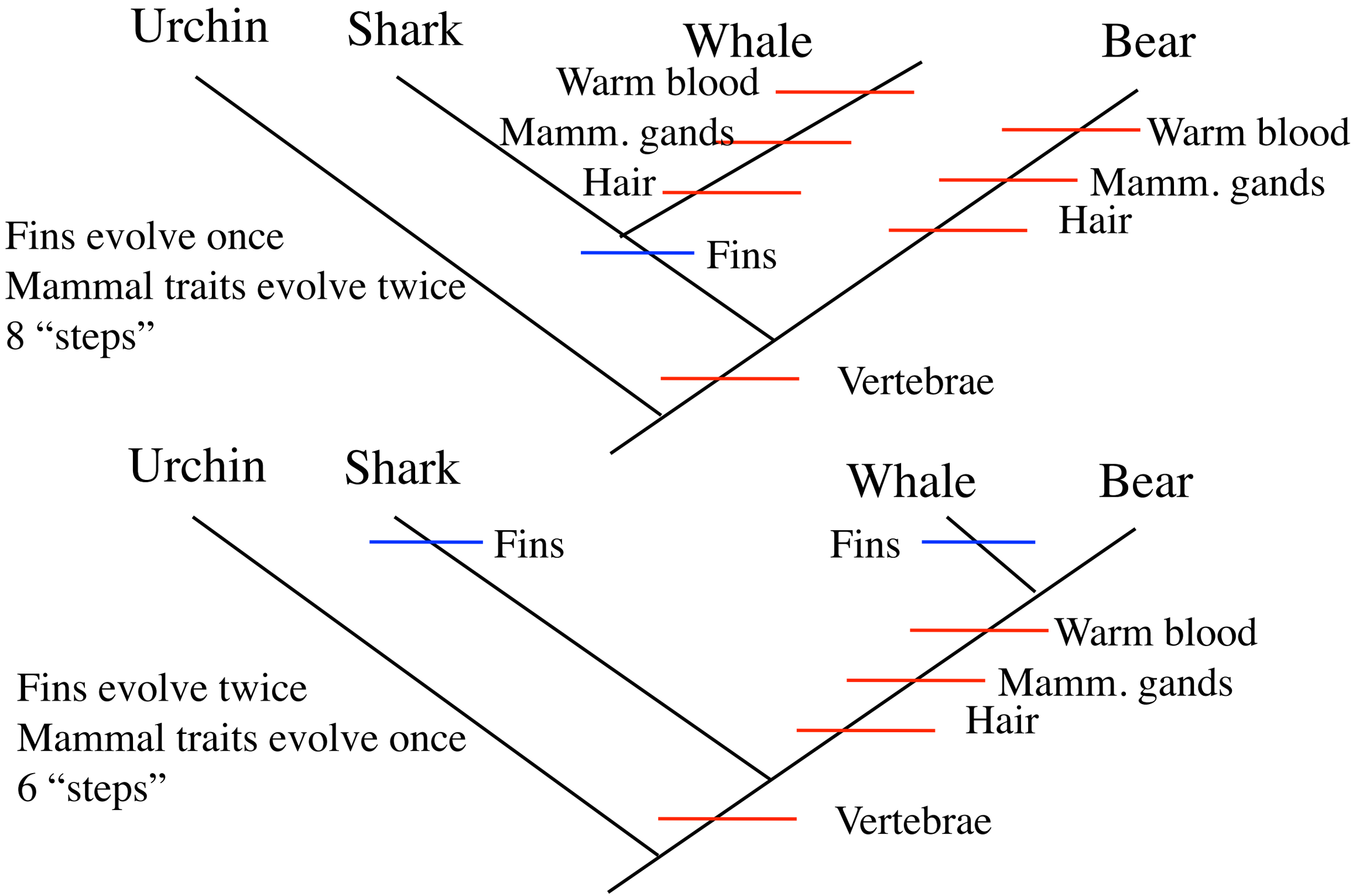
Parsimony example

	Vertebrae	Hair	Mammary glands	Warm blood	Fins
Urchin	0	0	0	0	0
Shark	1	0	0	0	1
Whale	1	1	1	1	1
Bear	1	1	1	1	0



Fins show <u>parallel</u> change; an example of conflict in the data set, or <u>Homoplasy</u>

Different Trees yield different interpreted evolutionary histories



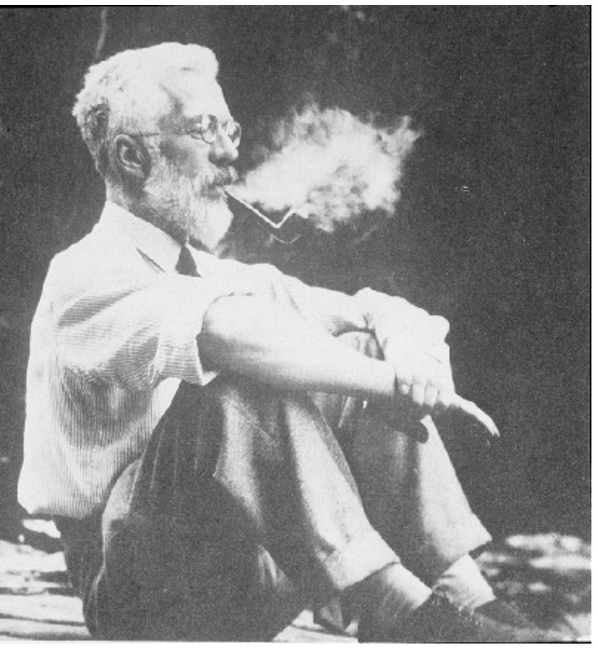
Building Trees with Parsimony

- Parsimony involves evaluating all possible tree topologies and giving each a Length score based on the total number of evolutionary changes that are needed to explain the observed data.

- The best tree is the one that requires the fewest hypothesized character state changes for all characters to account for all changes derived from a common ancestor.

Maximum Likelihood

And its application to phylogenies



R.A. Fisher- “Think Gray Matter!”

Joe Felsenstein

ML is a simple and powerful statistical method.

In words, want to find parameter values that maximize probability of observing data

$$L(\theta) = p(X|\theta)$$

Maximum Likelihood

Flipping Coins

Imagine you want to know if a coin is "fair"
you flip a coin $n = 10$ times, get 8 heads

Example of binomial random variable:

$$Prob\{X = x\} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

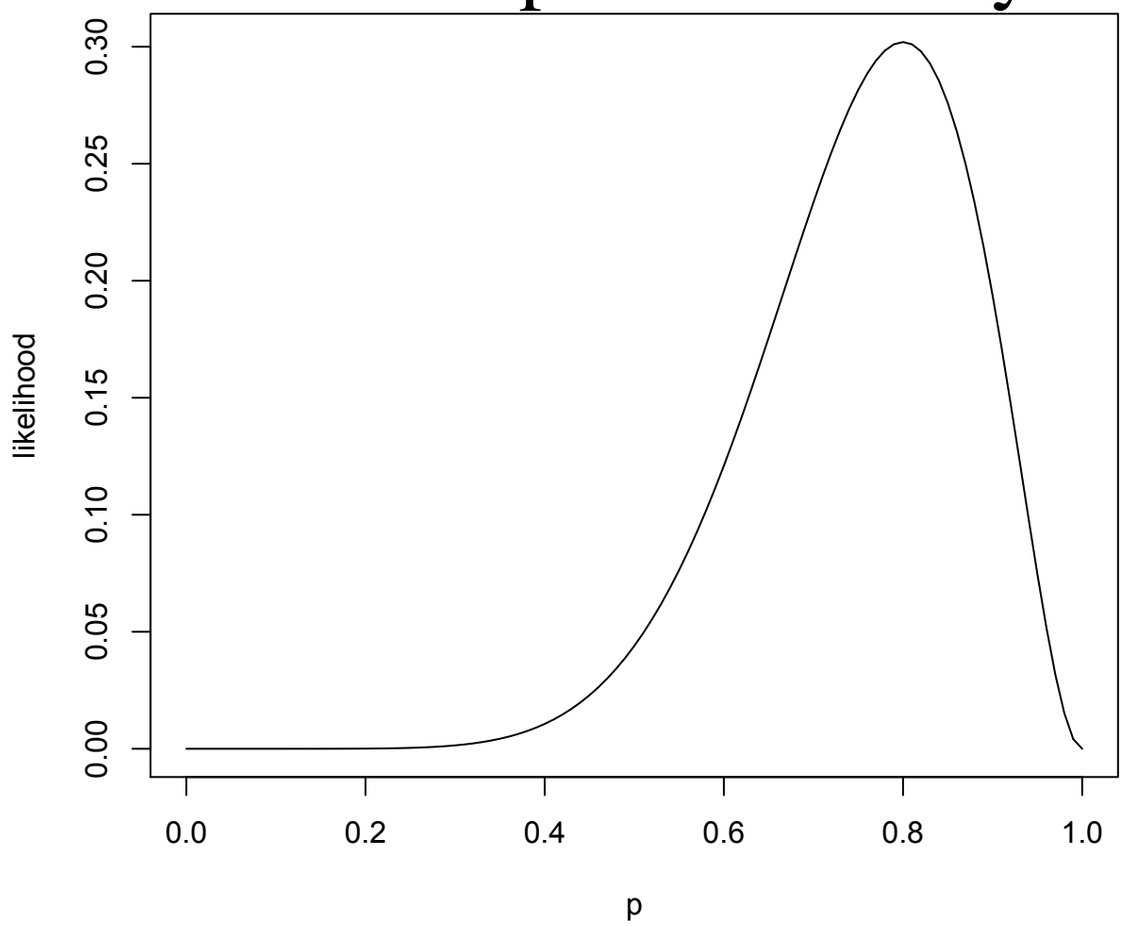
Probability of 8 heads:

$$Prob\{X = 8\} = \frac{10!}{8!(10-8)!} p^8 (1-p)^{10-8}$$

Maximum Likelihood

Flipping Coins

Let's estimate parameter P by ML



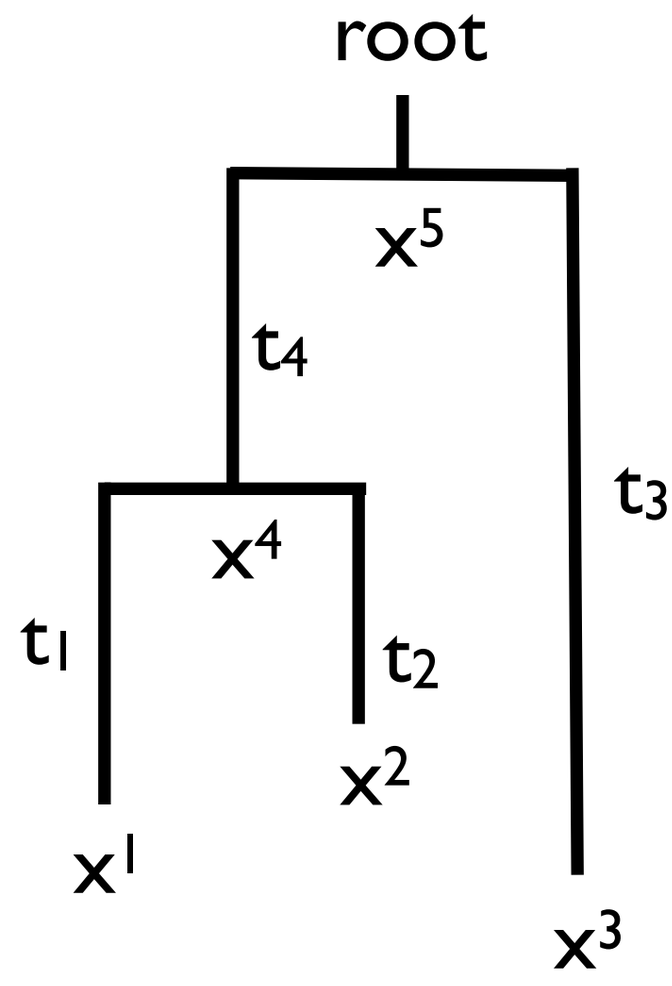
$$Prob\{X = 8\} = \frac{10!}{8!(10-8)!} p^8 (1-p)^{10-8}$$

Probability of 8 heads-- MLE of p = 0.8!

Maximum Likelihood

And its application to phylogenies

$$Prob(\text{data}|\text{tree})$$

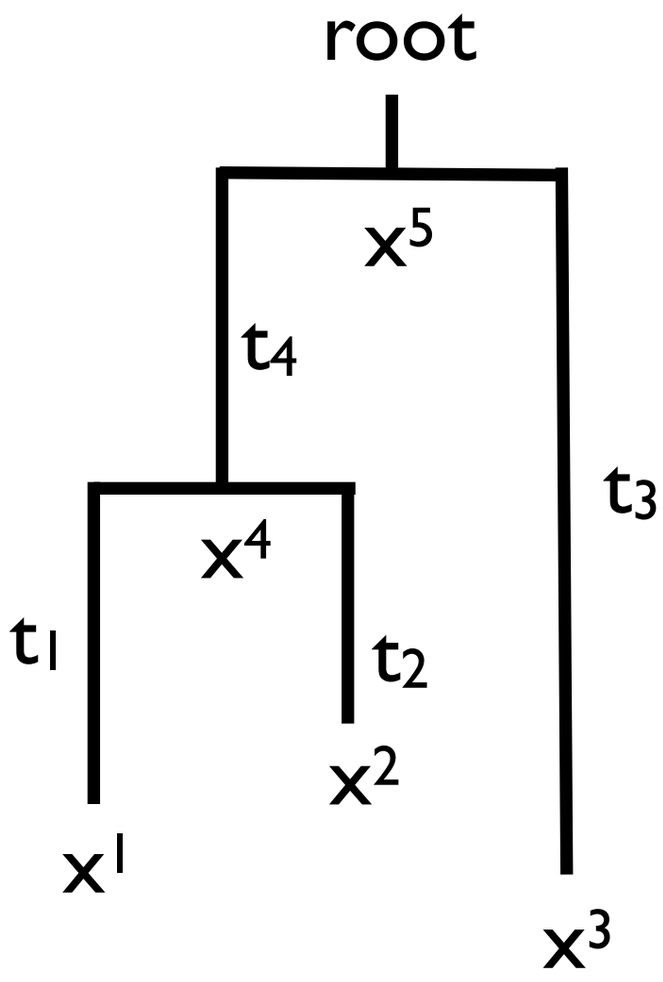


x^i are character states (i.e. ATGC)

t_i are branch lengths in units of time

Maximum Likelihood

And its application to phylogenies



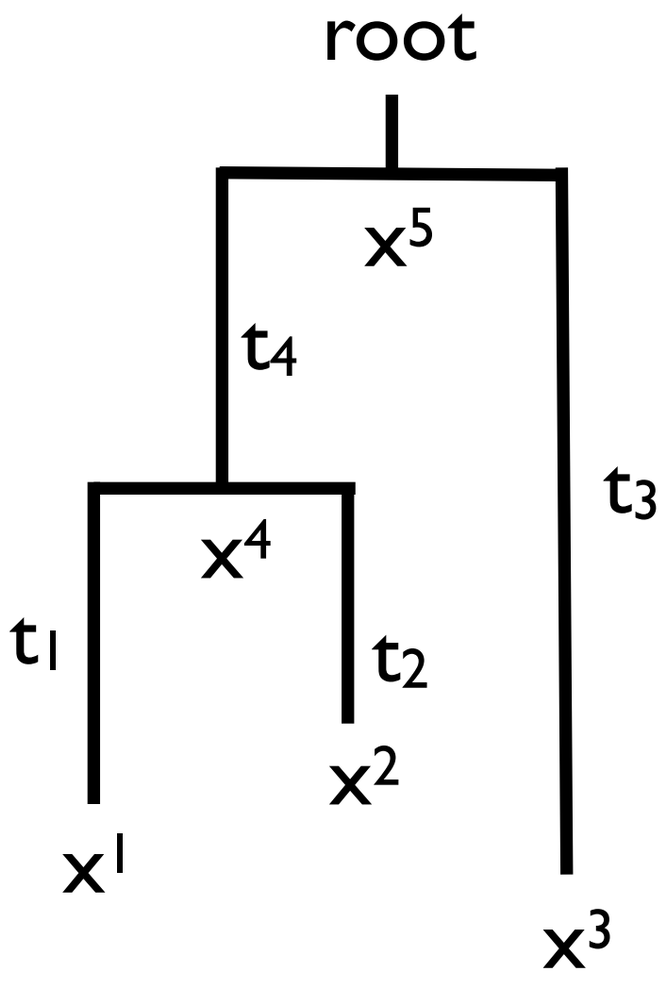
Decompose tree into, topology and set of branch lengths, $t^{\cdot}$

Also assume we can write down prob of going from one state to another, i.e. $Prob(x^i|x^j,t)$

$$
\begin{aligned}
Prob(\text{data}|\text{tree}) &= P(x^1,...,x^5|T,t^{\cdot}) \\
&= P(x^1|x^4,t_1)P(x^2|x^4,t_2)P(x^3|x^5,t_3)P(x^4|x^5,t_4)P(x^5)
\end{aligned}
$$

Maximum Likelihood

And its application to phylogenies



Can now search for tree with topology T and branch lengths $t.$ which maximizes

$P(x^{\cdot}|T, t.)$ ⟵ Lik. Function

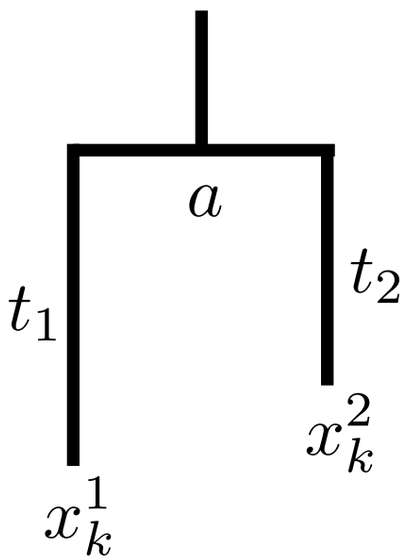
Two parts:

1) Search over topologies (big issue)

2) Optimization of branch lengths given topology-- easier

Maximum Likelihood

Likelihood to two taxa tree



Consider a single site, *k*, prob tree is

$$P(x_k^1, x_k^2, a|T, t_1, t_2) = q_a P(x_k^1|a, t_1) P(x_k^2|a, t_2)$$

Don't actually know ancestral state, so lets sum over all possibilities!

$$P(x_k^1, x_k^2, a|T, t_1, t_2) = \sum_a q_a P(x_k^1|a, t_1) P(x_k^2|a, t_2)$$

This is huge step!

If we have *n* sites to look at, assume independence so full likelihood is

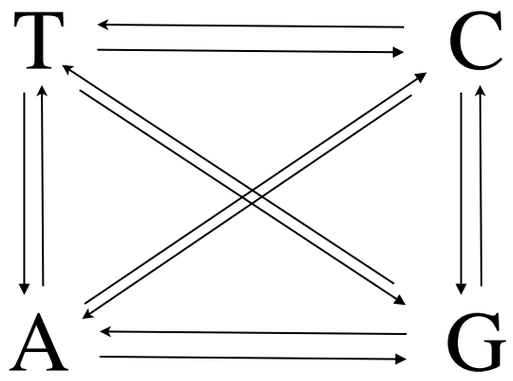
$$P(x^1, x^2, |T, t_1, t_2) = \prod_{k=1}^{n} P(x_k^1, x_k^2|T, t_1, t_2)$$

Maximum Likelihood

Probabilistic models of DNA evolution

To calculate Likelihood of tree, need some model of substitution

$$Prob(x^i|x^j,t)$$



Jukes-Cantor Model (1969)

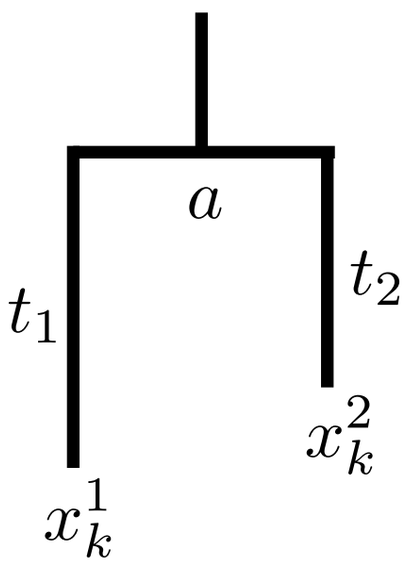
$$P = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}$$

If we assume continuous time the prob of a sub. or no sub respectively is

$$\begin{aligned} P(x^i|x^j,t) &= \frac{1}{4}(1-e^{-4\alpha t}), i \neq j \\ P(x^i|x^i,t) &= \frac{1}{4}(1+3e^{-4\alpha t}) \end{aligned}$$

Maximum Likelihood

Two sequences -- Jukes-Cantor Model



Seq1 CCGC
Seq2 CGGG

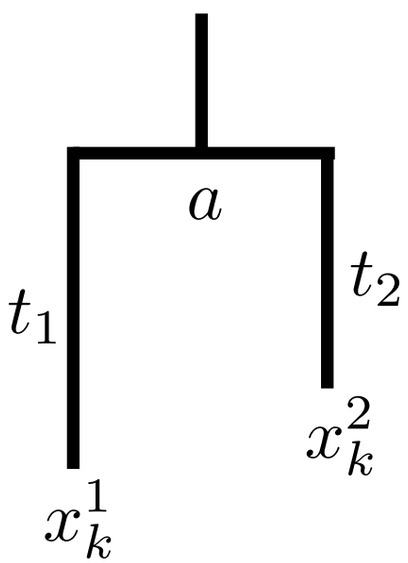
$$P(x_k^1, x_k^2, a|T, t_1, t_2) = \sum_a q_a P(x_k^1|a, t_1) P(x_k^2|a, t_2)$$

$$\begin{aligned} P(C, C|T, t_1, t_2) &= q_C P(C|C, t_1) P(C|C, t_2) \\ &\quad + q_A P(C|A, t_1) P(C|A, t_2) \\ &\quad + q_T P(C|T, t_1) P(C|T, t_2) \\ &\quad + q_G P(C|G, t_1) P(C|G, t_2) \\ &= \frac{1}{16}(1 + 3e^{-4\alpha(t_1+t_2)}) \end{aligned}$$

$$P(C, G|T, t_1, t_2) = \frac{1}{16}(1 - e^{-4\alpha(t_1+t_2)})$$

Maximum Likelihood

Two sequences -- Jukes-Cantor Model



Seq1 CCGC
Seq2 CGGG

If we have *n* sites to look at, assume independence so full likelihood is

$$P(x^1, x^2, |T, t_1, t_2) = \prod_{k=1}^{n} P(x_k^1, x_k^2 | T, t_1, t_2)$$

$$\begin{aligned} P(x^1, x^2 | T, t_1, t_2) \quad = \quad & P(C, C | T, t_1, t_2) \\ & \times P(C, G | T, t_1, t_2) \\ & \times P(G, G | T, t_1, t_2) \\ & \times P(C, G | T, t_1, t_2). \end{aligned}$$

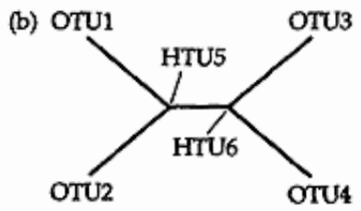
Maximum Likelihood

Four Taxon 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)



(c)

$L_{(5)}$ = Prob(C, C, A–A, G) + Prob(C, C, A–C, G) + Prob(C, C, A–T, G) + Prob(C, C, A–G, G)
+ Prob(C, C, C–A, G) + Prob(C, C, C–C, G) + Prob(C, C, C–T, G) + Prob(C, C, C–G, G)
+ Prob(C, C, T–A, G) + Prob(C, C, T–C, G) + Prob(C, C, T–T, G) + Prob(C, C, T–G, G)
+ Prob(C, C, G–A, G) + Prob(C, C, G–C, G) + Prob(C, C, G–T, G) + Prob(C, C, G–G, G)

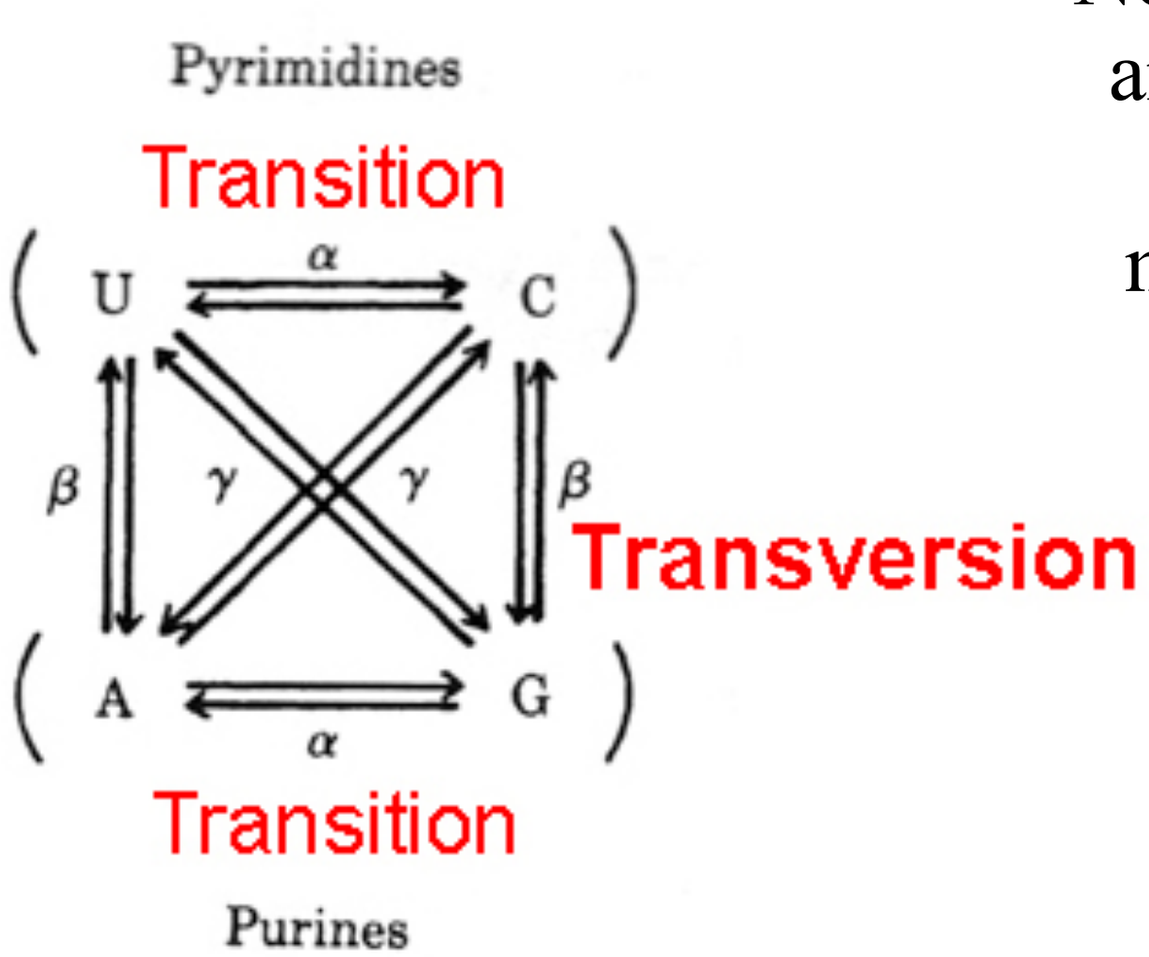
FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length *n*. (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums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

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

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

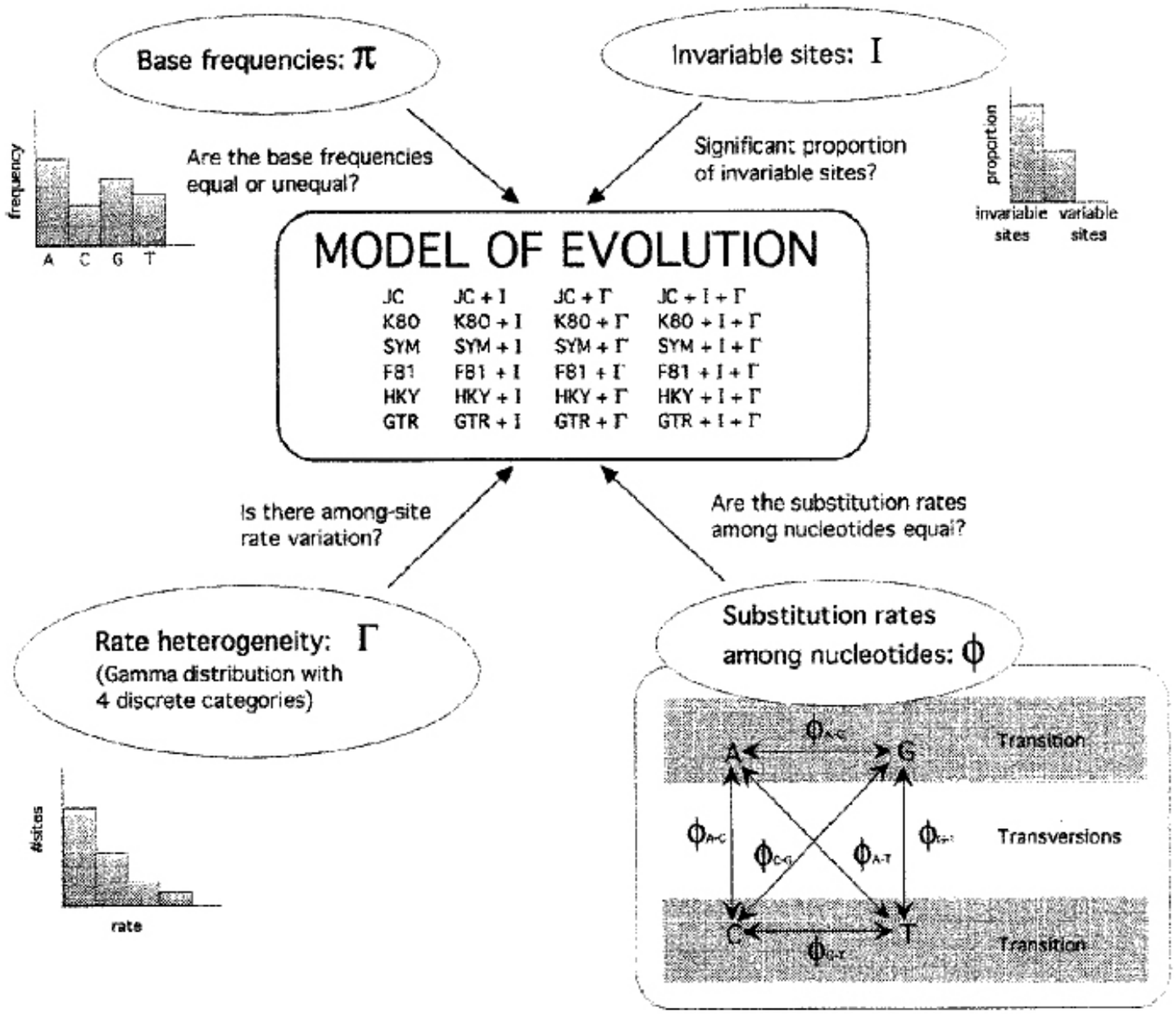
Models of Nucleotide Substitution

Correcting for the frequency of multiple hits.



Not all nucleotide substitutions are equally likely, because of how the DNA replication machinery detects mistakes.

Other "inequalities" to correct for when trying to estimate the frequency of multiple hits.

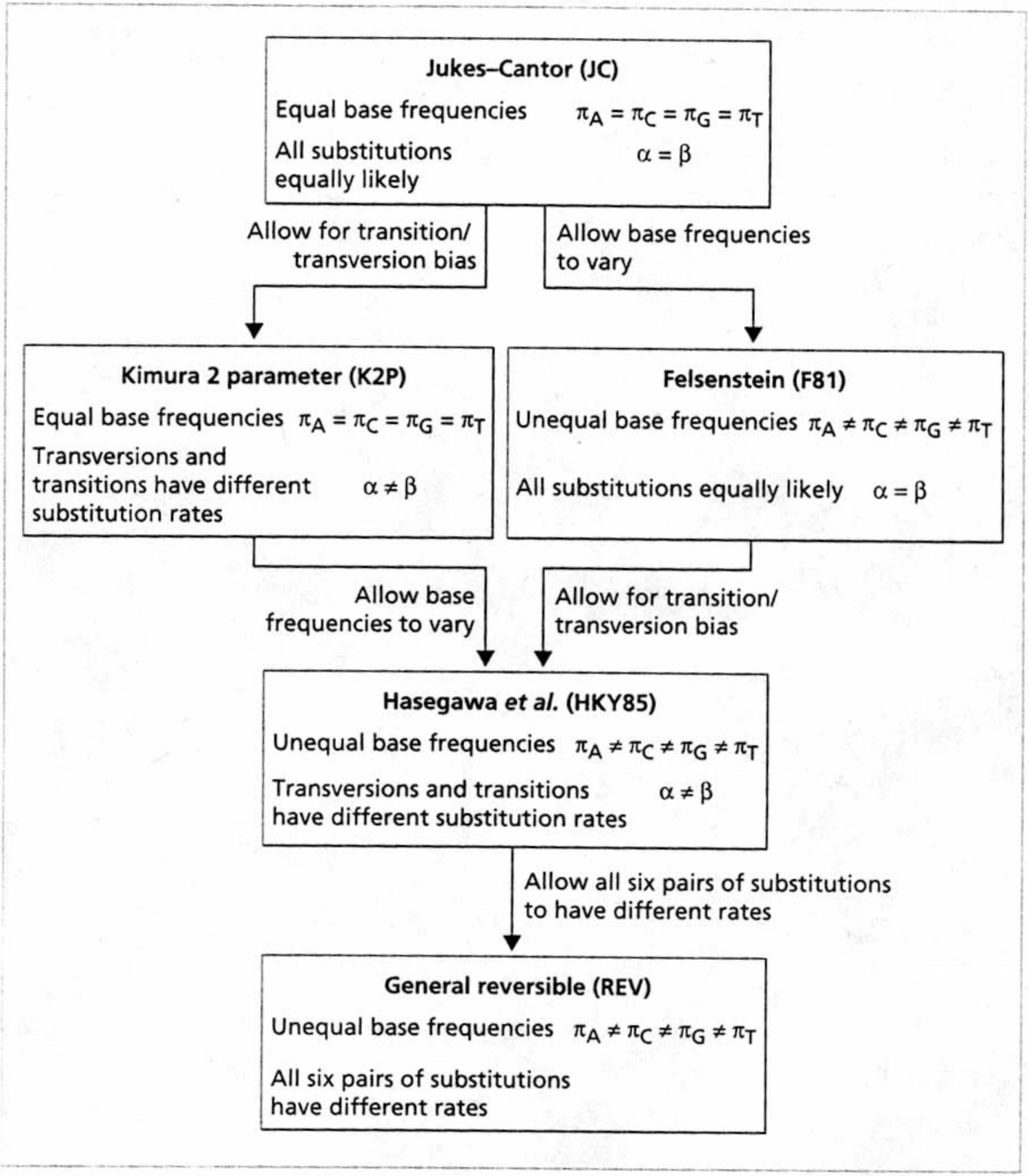


Model	Base frequencies	Substitution rates	Number of free parameters
JC	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}G} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{C\text{-}T} = \phi_{G\text{-}T}$	0
K80	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{G\text{-}T} \neq \phi_{A\text{-}G} = \phi_{C\text{-}T}$	1
SYM	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} \neq \phi_{A\text{-}G} \neq \phi_{A\text{-}T} \neq \phi_{C\text{-}G} \neq \phi_{C\text{-}T} \neq \phi_{G\text{-}T}$	5
F81	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}G} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{C\text{-}T} = \phi_{G\text{-}T}$	3
HKY	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{G\text{-}T} \neq \phi_{A\text{-}G} = \phi_{C\text{-}T}$	4
GTR	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} \neq \phi_{A\text{-}G} \neq \phi_{A\text{-}T} \neq \phi_{C\text{-}G} \neq \phi_{C\text{-}T} \neq \phi_{G\text{-}T}$	8

FIGURE 1. A comparison of models of nucleotide substitution. Model selection methods selected the best-fit model for the data set at hand among 24 possible models. See Table 1 footnote for explanation of acronyms for methods.

From David Posada and Keith Crandall, "Selecting the Best-Fit Model of Nucleotide Substitution," *Systematic Biology* 50 (2001), 580-601.

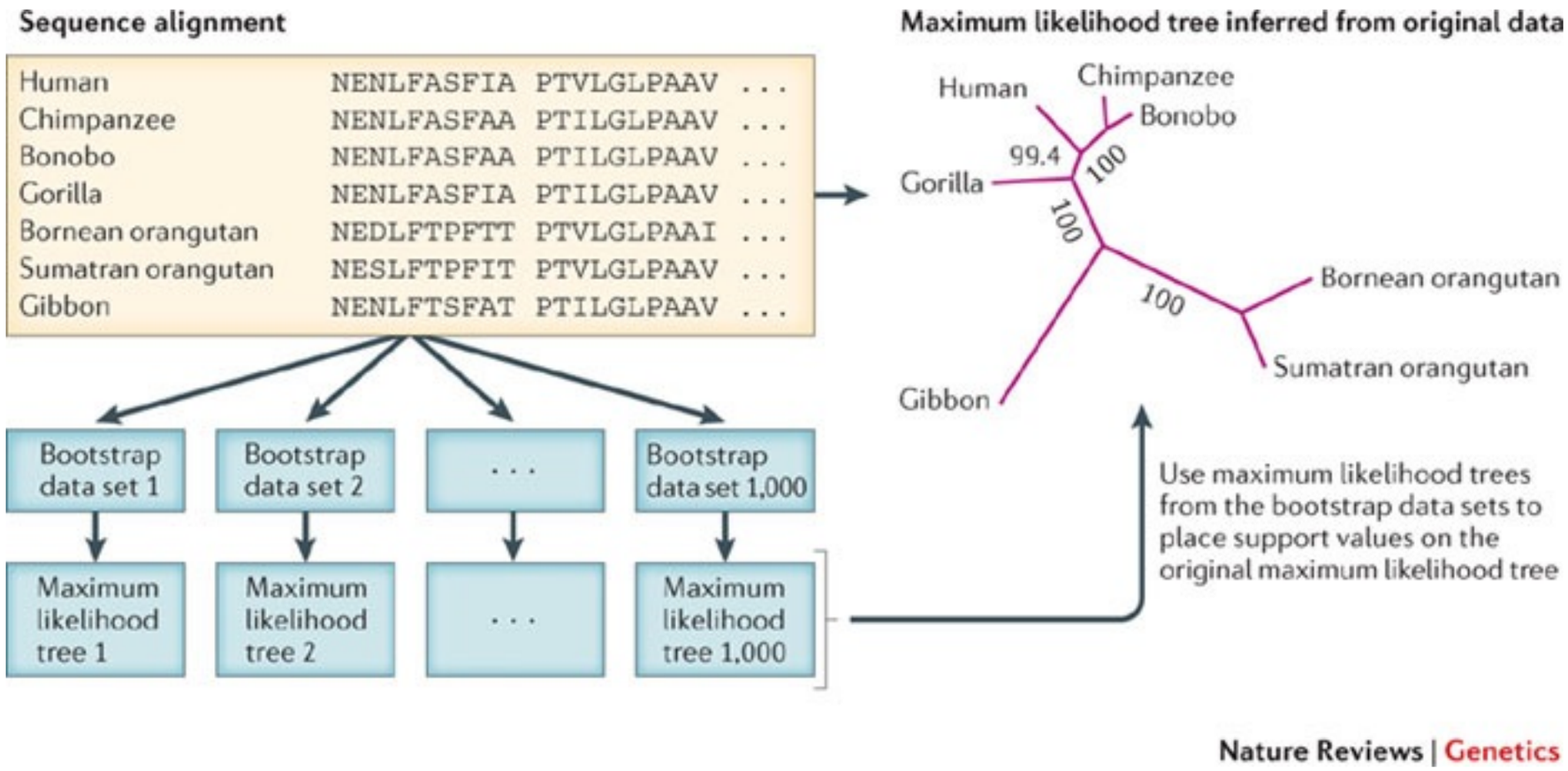
Assumptions of different models of molecular evolution.



Page and Holmes (1998)

Fig. 5.14 Interrelationships among five models for estimating the number of nucleotide substitutions among a pair of DNA sequences. The JC, K2P, F81 and HKY85 models can all be generated by constraining various parameters of the REV model.

Bootstrap for Confidence!



Randomly resample data matrix, make new tree, count how many times a node is supported