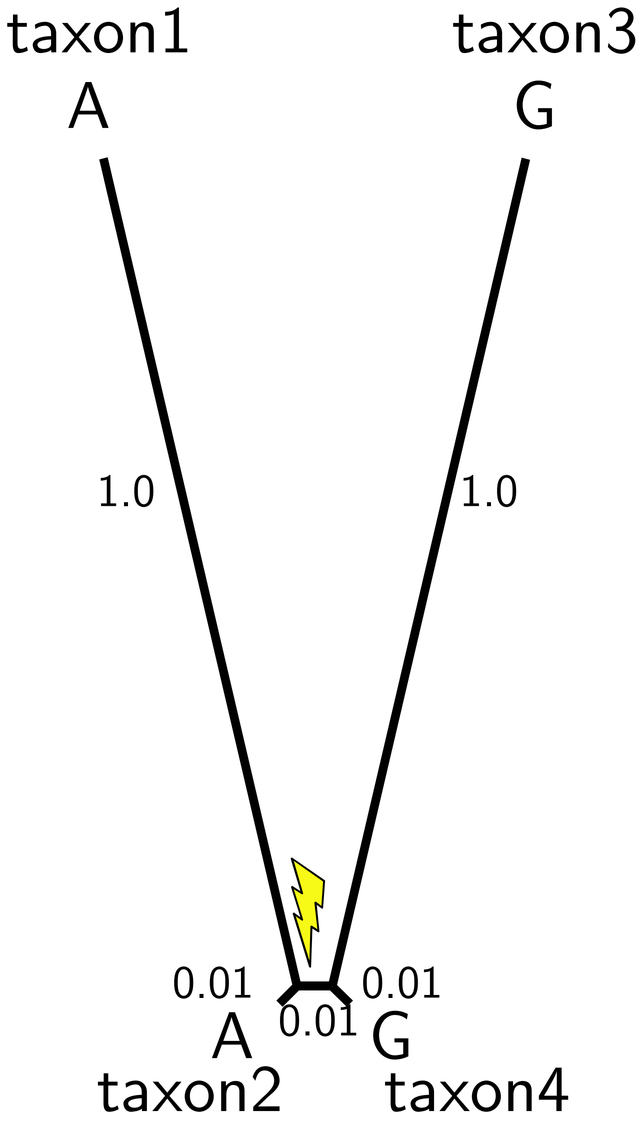
Thanks to Paul Lewis and Joe Felsenstein for the use of slides

Review

- Hennigian logic reconstructs the tree if we know **polarity** of characters and there is **no homoplasy**
- UPGMA infers a tree from a distance matrix:
  - groups based on **similarity**
  - fails to give the correct tree if rates of character evolution vary much
- Modern distance-based approaches:
  - find trees and branch lengths: patristic distances $\approx$ distances from character data.
  - do **not** use all of the information in the data.
- Parsimony:
  - prefer the tree that **requires** the fewest character state changes. Minimize the number of times you invoke homoplasy to explain the data.
  - can work well if if homoplasy is not rare
  - fails if homoplasy very common **or is concentrated on certain parts of the tree**

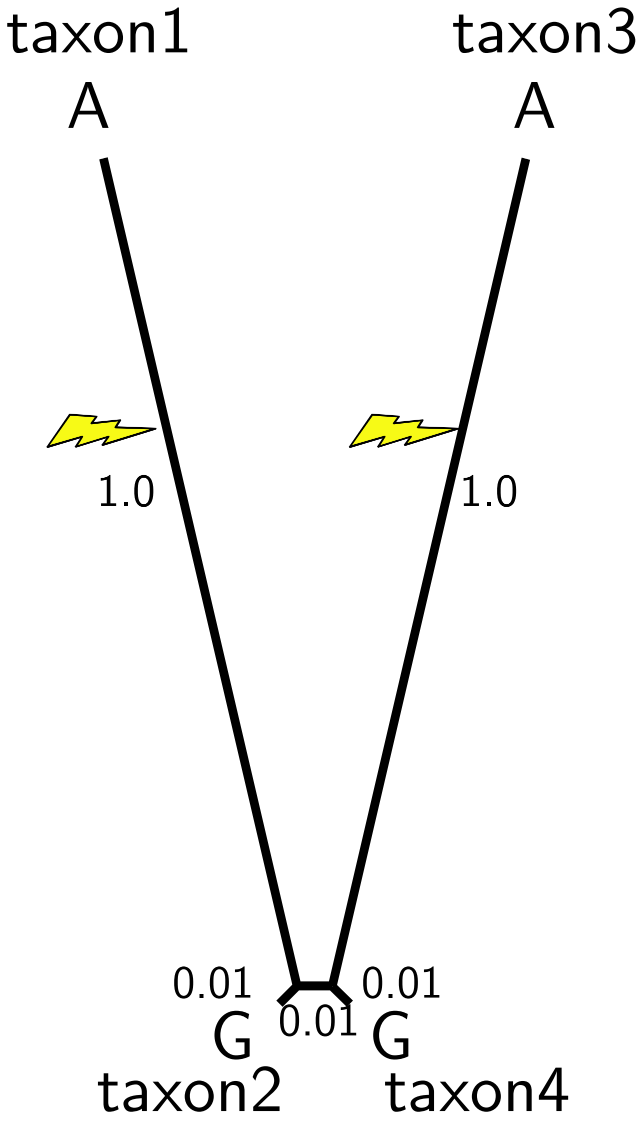
Long branch attraction



Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

Long branch attraction



Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).

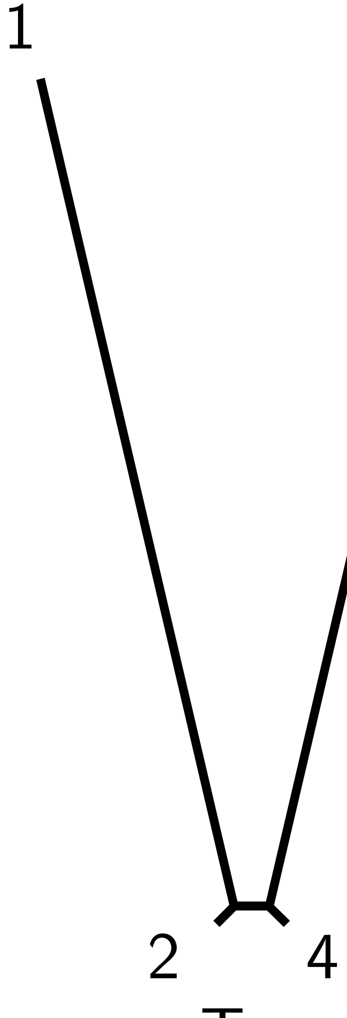
Long branch attraction data

Under such a tree misleading characters are more common that characters that favor the true tree.

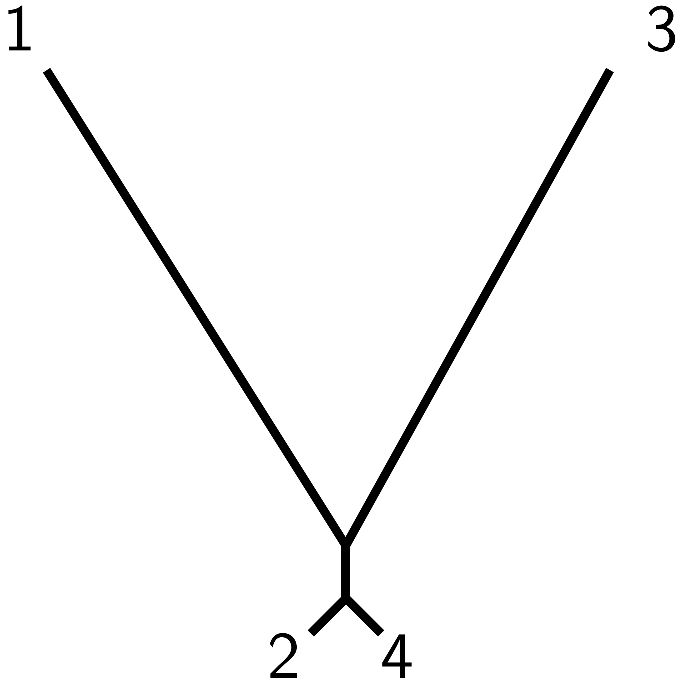
	Rare					Common			
taxon1	A	A	C	C		A	A	C	C
taxon2	A	A	C	C		G	C	T	G
taxon3	G	C	T	G		A	A	C	C
taxon4	G	C	T	G		G	C	T	G

Long branch attraction

Parsimony is almost guaranteed to get this tree wrong.



True



Inferred

Likelihood

X is the data.

T is the tree.

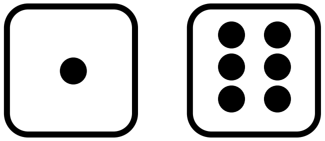
ν is a vector of branch lengths.

$\Pr(X|T,\nu)$ is the *likelihood*; this is sometimes denoted $L(T,\nu)$.

Maximum likelihood: find the T and ν that gives the highest likelihood.

Combining probabilities

- Multiply probabilities if the component events must happen simultaneously (i.e. whereever you would naturally use the word AND when describing the problem)

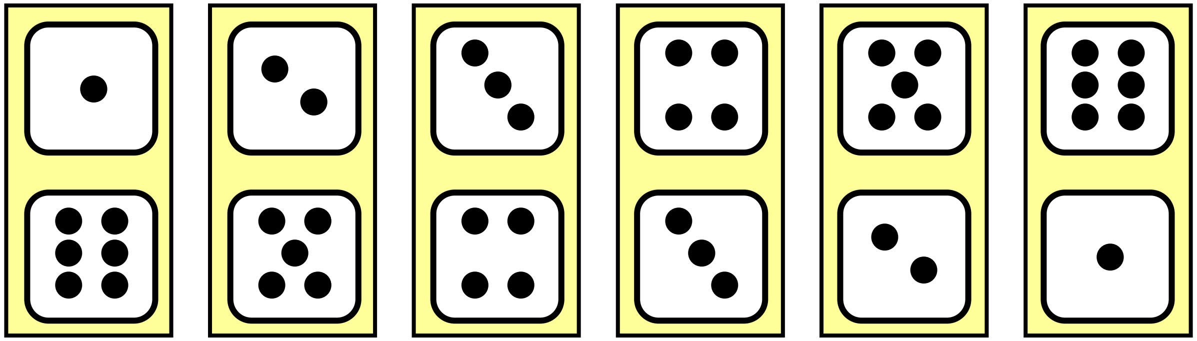


$(1/6) \times (1/6) = 1/36$

What is the probability of rolling two dice and having the first show 1 dot AND the second show 6 dots?

Combining probabilities

- Add probabilities if the component events are mutually exclusive (i.e. whereever you would naturally use the word OR)



(1/36) + (1/36) + (1/36) + (1/36) + (1/36) + (1/36) = 1/6

What is the probability of rolling 7 using two dice? This is the same as asking "What is the probability of rolling (1 and 6) OR (2 and 5) OR (3 and 4) OR (4 and 3) OR (5 and 2) OR (6 and 1)?"

Likelihood of a single sequence

First 32 nucleotides of the $\psi\eta$-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$
\begin{aligned}
L &= \pi_G\pi_A\pi_A\pi_G\pi_T\pi_C\pi_C\pi_T\pi_T\pi_G\pi_A\pi_G\pi_A\pi_A\pi_A\pi_T\pi_A\pi_A\pi_A\pi_C\pi_T\pi_G\pi_C\pi_A\pi_C\pi_A\pi_C\pi_A\pi_C\pi_T\pi_G\pi_G \\
&= \pi_A^{12}\pi_C^{7}\pi_G^{7}\pi_T^{6}
\end{aligned}
$$

$$\ln L = 12\ln\left(\pi_A\right) + 7\ln\left(\pi_C\right) + 7\ln\left(\pi_G\right) + 6\ln\left(\pi_T\right)$$

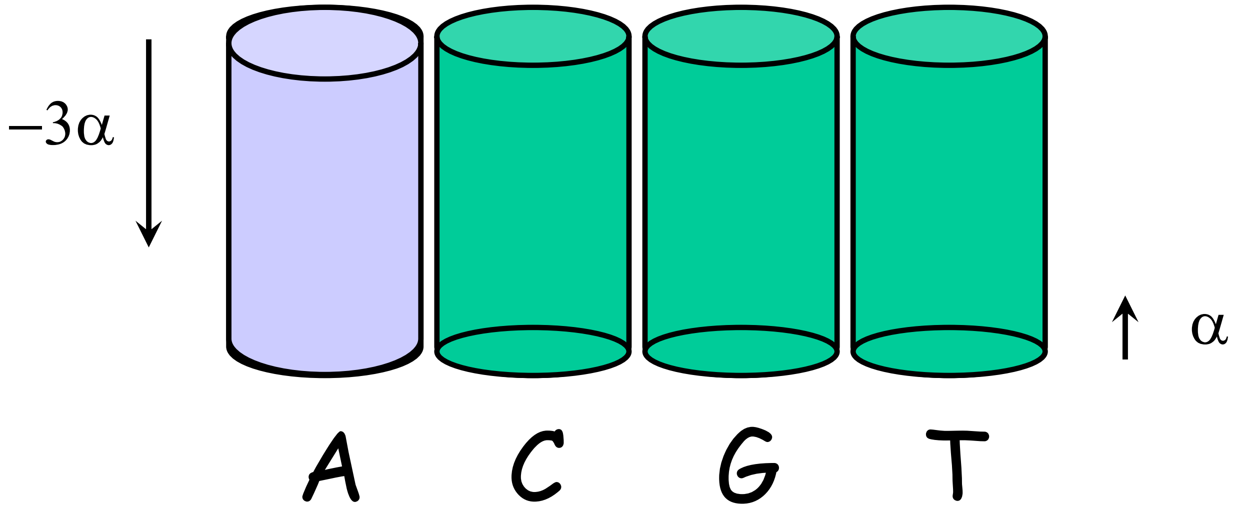
We can already see by eye-balling this that the F81 model (which allows unequal base frequencies) will fit better than the JC69 model (which assumes equal base frequencies) because there are about twice as many As as there are Cs, Gs and Ts.

Likelihoods on the simplest possible tree

$$\texttt{GA}\longrightarrow\texttt{GG}$$

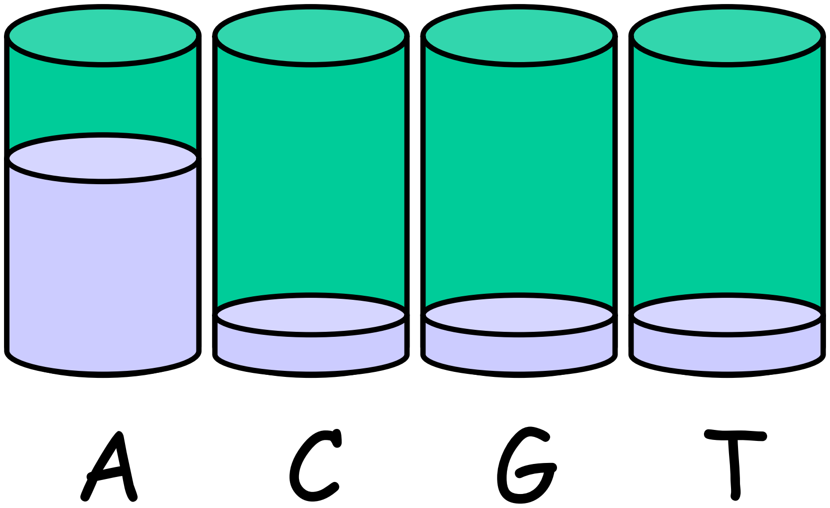
$$\begin{aligned}
L &= L_1 L_2 \\
&= \Pr(G)\Pr(G \rightarrow G)\Pr(A)\Pr(A \rightarrow G) \\
&= \Pr(G)\Pr(G \rightarrow G|\nu)\Pr(A)\Pr(A \rightarrow G|\nu)
\end{aligned}$$

Water analogy (time 0)



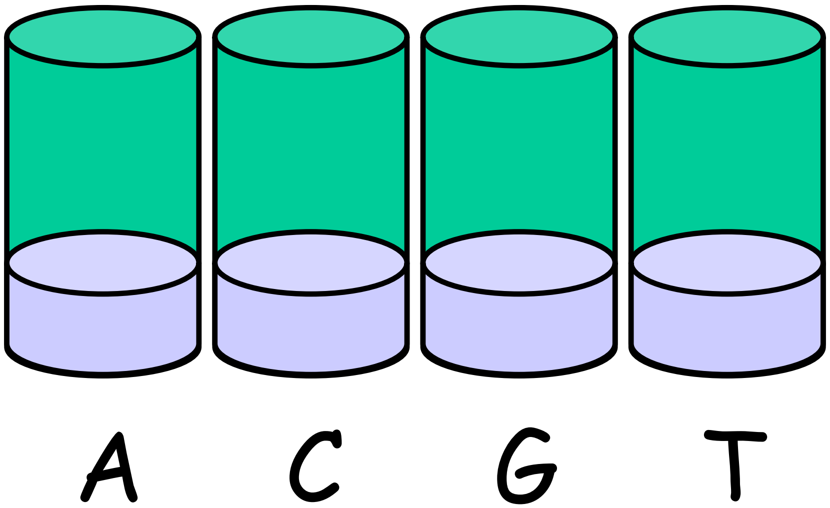
- Start with container A completely full and others empty
- Imagine that all containers are connected by tubes that allow same rate of flow between any two
- Initially, A will be losing water at 3 times the rate that C (or G or T) gains water

Water analogy (after some time)



A’s level is not dropping as fast now because it is now also *receiving* water from C, G and T

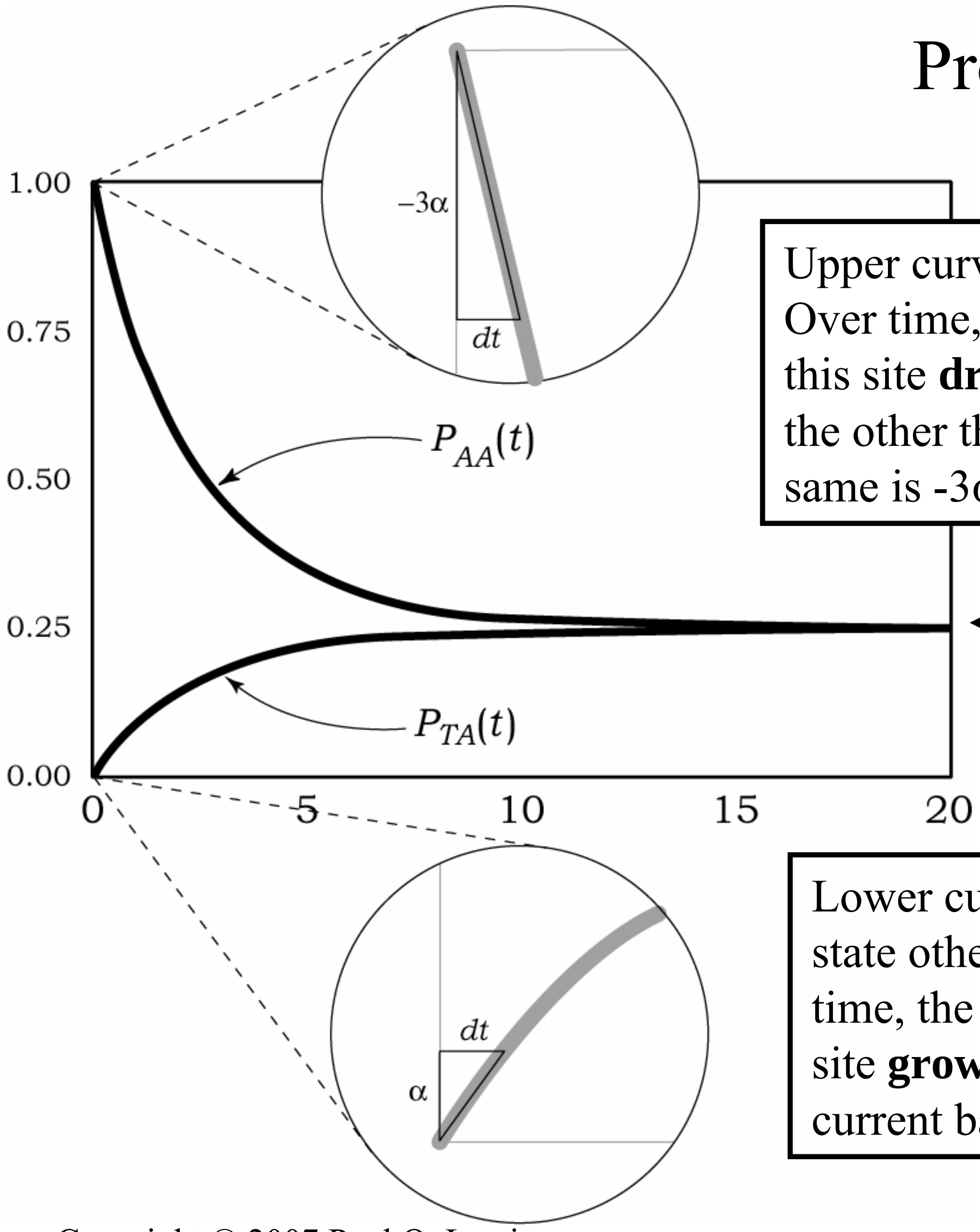
Water analogy (after a very long time)



Eventually, all containers are one fourth full and there is zero *net* volume change – **stationarity** (equilibrium) has been achieved

(Thanks to Kent Holsinger for this analogy)

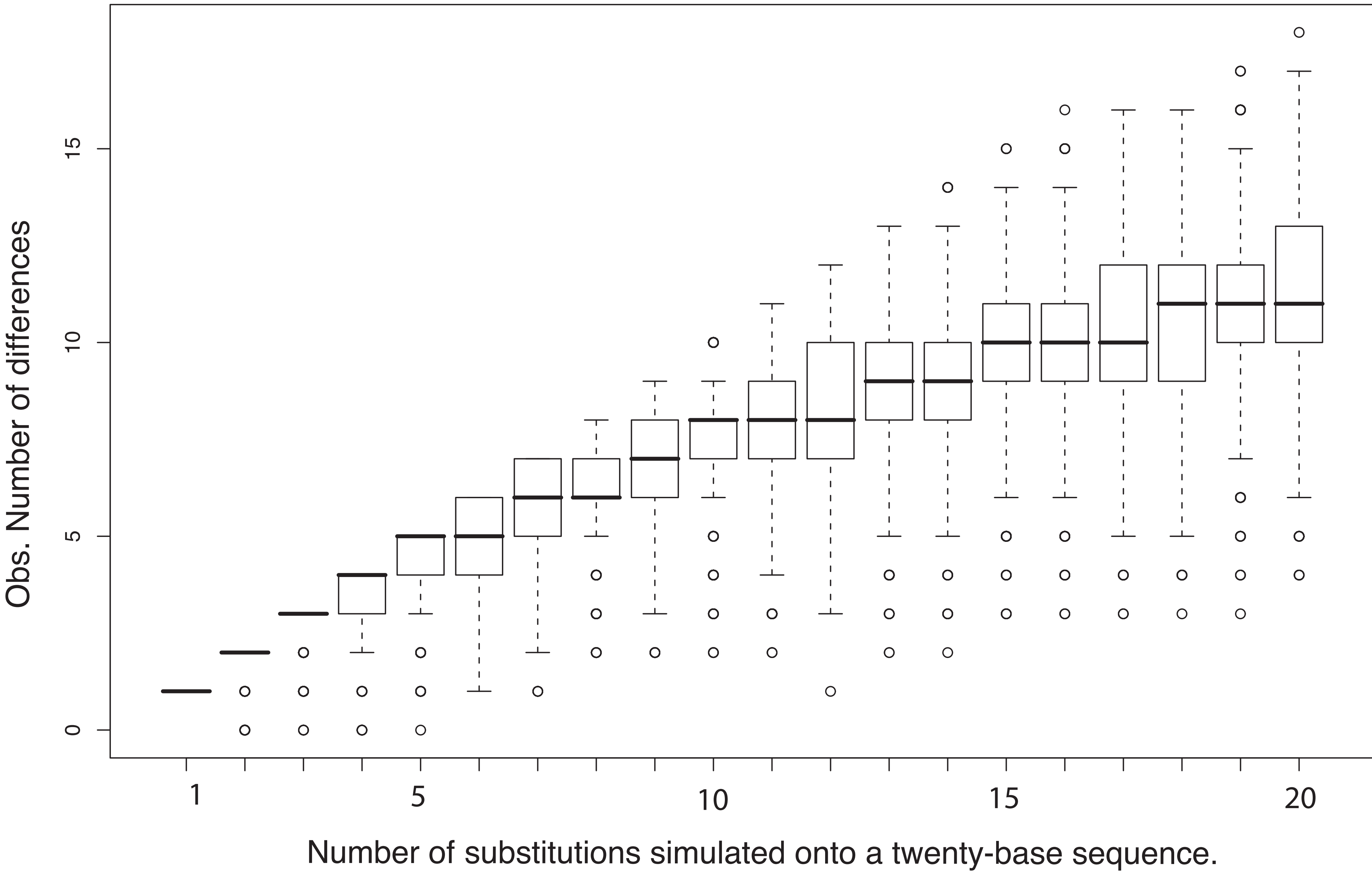
Probability of "A present" as a function of time



Upper curve assumes we started with A at time 0. Over time, the probability of still seeing an A at this site **drops** because rate of changing to one of the other three bases is 3α (so rate of staying the same is -3α).

The equilibrium relative frequency of A is 0.25

Lower curve assumes we started with some state other than A (T is used here). Over time, the probability of seeing an A at this site **grows** because the rate at which the current base will change into an A is α.



Obs. Number of differences
0
5
10
15
1
5
10
15
20
Number of substitutions simulated onto a twenty-base sequence.

Jukes-Cantor model

$$\begin{aligned}
\Pr(G \rightarrow G|\nu) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-4\nu}{3}} \\
\Pr(A \rightarrow G|\nu) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-4\nu}{3}}
\end{aligned}$$

Likelihoods on the simplest possible tree

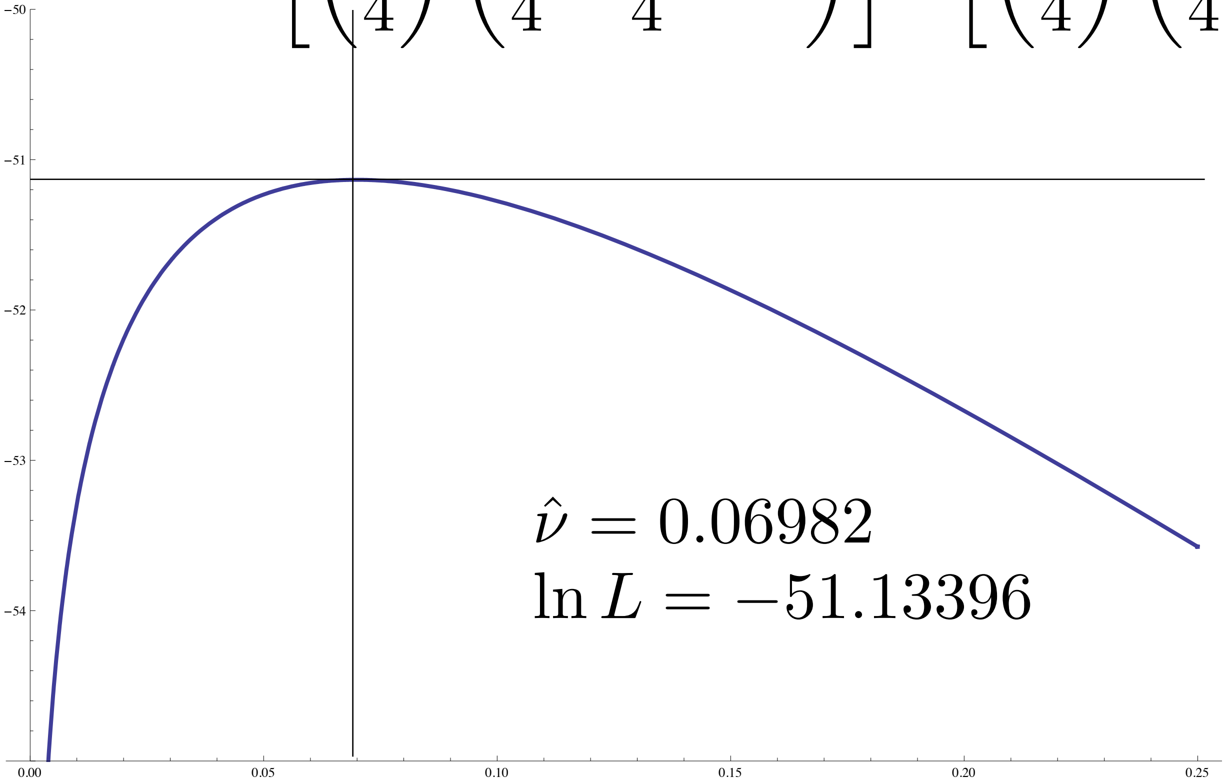
$$\mathtt{GA}\longrightarrow\mathtt{GG}$$

$$\begin{aligned}
L &= L_1 L_2 \\
&= \Pr(G)\Pr(G \rightarrow G)\Pr(A)\Pr(A \rightarrow G) \\
&= \Pr(G)\Pr(G \rightarrow G|\nu)\Pr(A)\Pr(A \rightarrow G|\nu) \\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}+\frac{3}{4}\,e^{\frac{-4\nu}{3}}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}-\frac{1}{4}\,e^{\frac{-4\nu}{3}}\right)
\end{aligned}$$

The first 30 nucleotides of the $\psi\eta$-globin gene

```
gorilla    GAAGTCCTTGAGAAATAAACTGCACACTGG
orangutan  GGACTCCTTGAGAAATAAACTGCACACTGG
```

$$L = \left[\left(\frac{1}{4}\right)\left(\frac{1}{4} + \frac{3}{4}\,e^{\frac{-4\nu}{3}}\right)\right]^{28}\left[\left(\frac{1}{4}\right)\left(\frac{1}{4} - \frac{1}{4}\,e^{\frac{-4\nu}{3}}\right)\right]^{2}$$

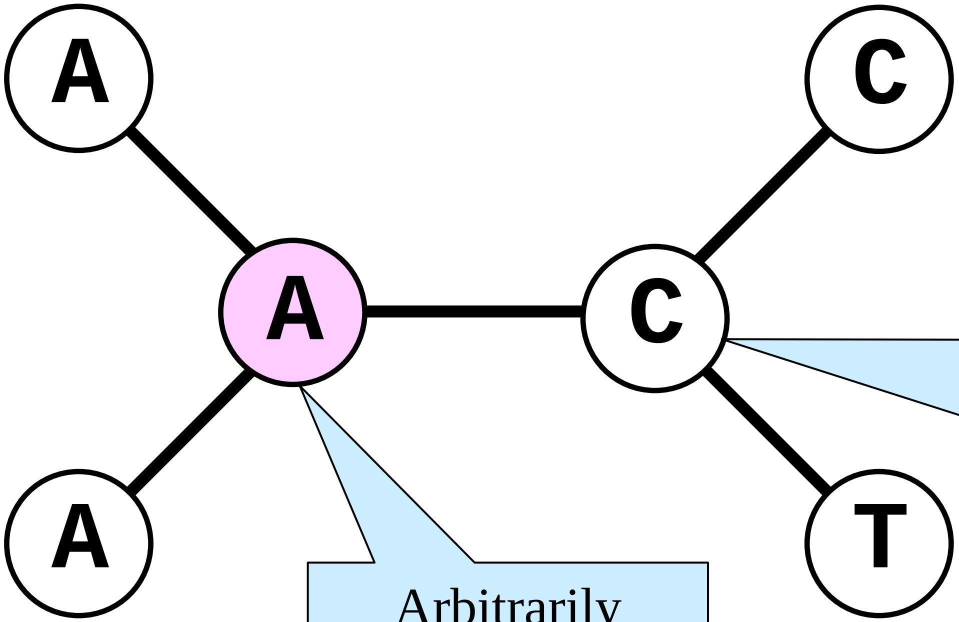


$\hat{\nu} = 0.06982$

$\ln L = -51.13396$

Likelihood of a tree

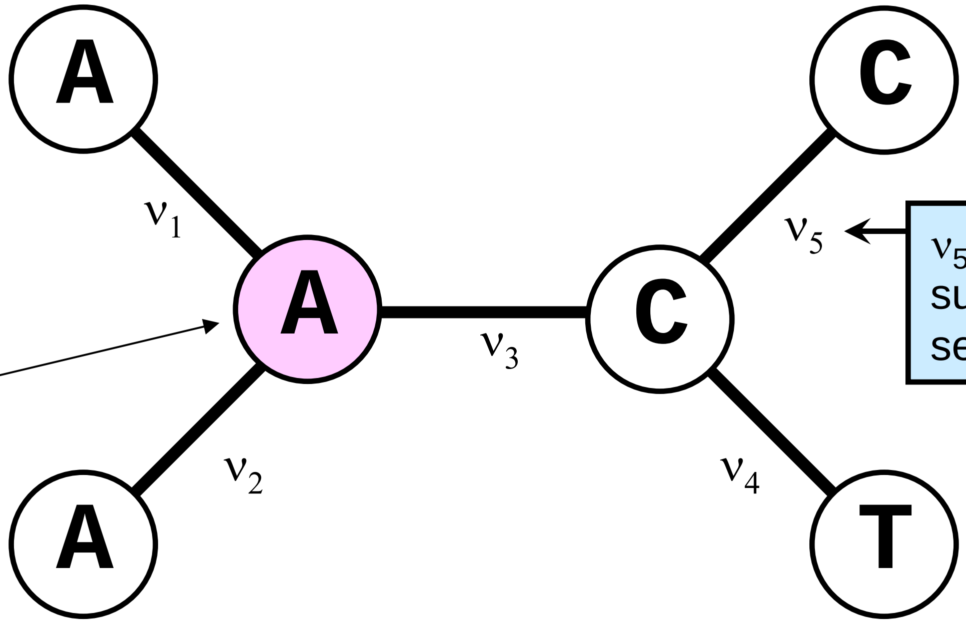
(data for only one site shown)



Arbitrarily chosen to serve as the root node

Ancestral states like this are not really known - we will address this in a minute.

Likelihood for site k



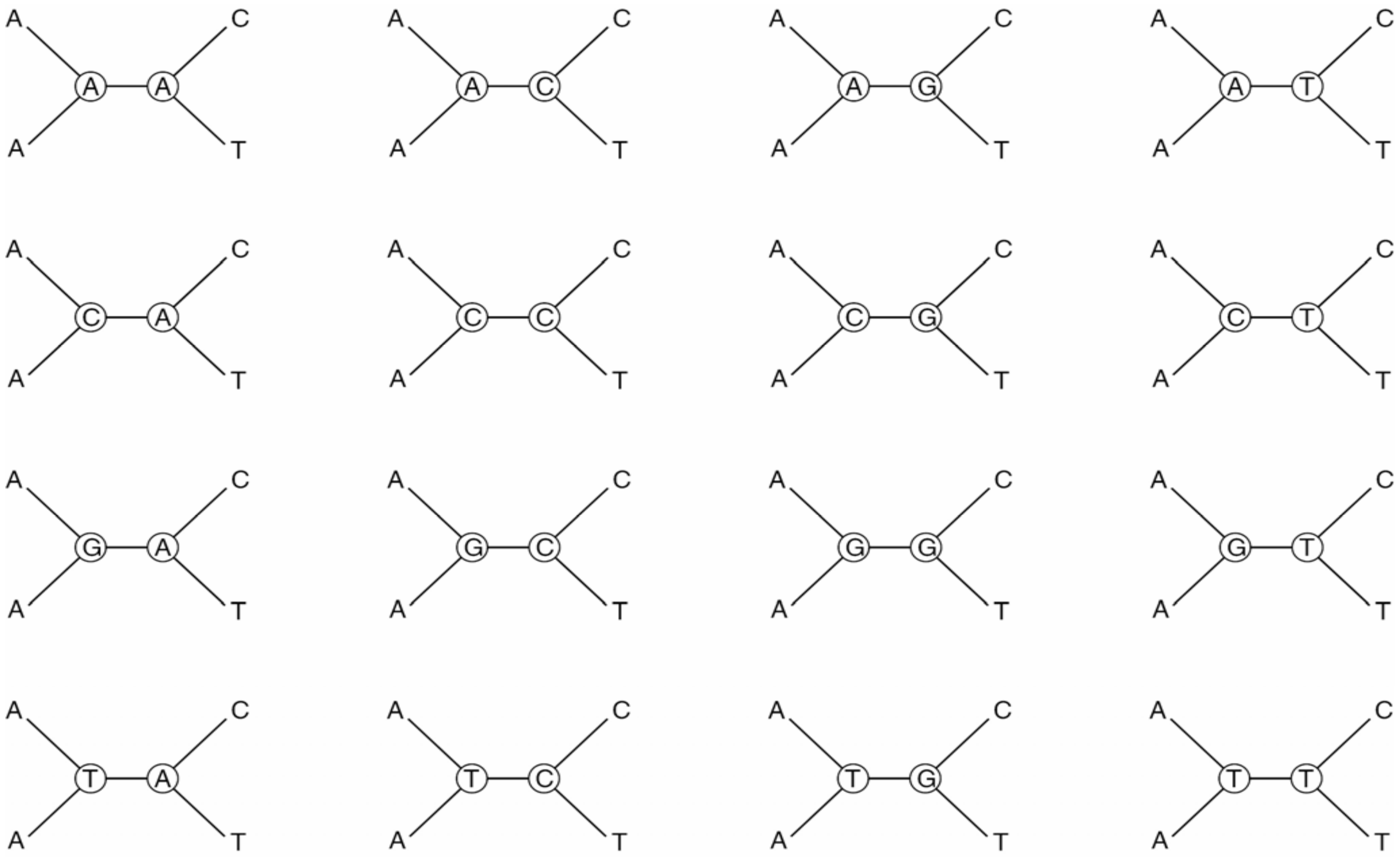
ν_5 is the expected no. substitutions for just this segment of the tree

π_A

$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_5/3}\right]$$

$P_{AA}(\nu_1)$ $P_{AA}(\nu_2)$ $P_{AC}(\nu_3)$ $P_{CT}(\nu_4)$ $P_{CC}(\nu_5)$

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum



Likelihood and Bayesian procedures

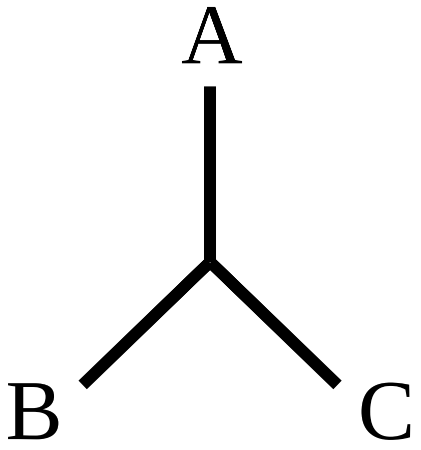
1. very computationally intensive,
2. Use all of the information in the data,
3. Let us estimate the forces of character evolution while estimating trees,
4. Uses models to detect concerted patterns of homoplasy (this is how likelihood based procedures avoid long-branch attraction).

Tree Searching

Parsimony and ML give us ways to deciding whether one tree is fits our data better than another tree, but . . .

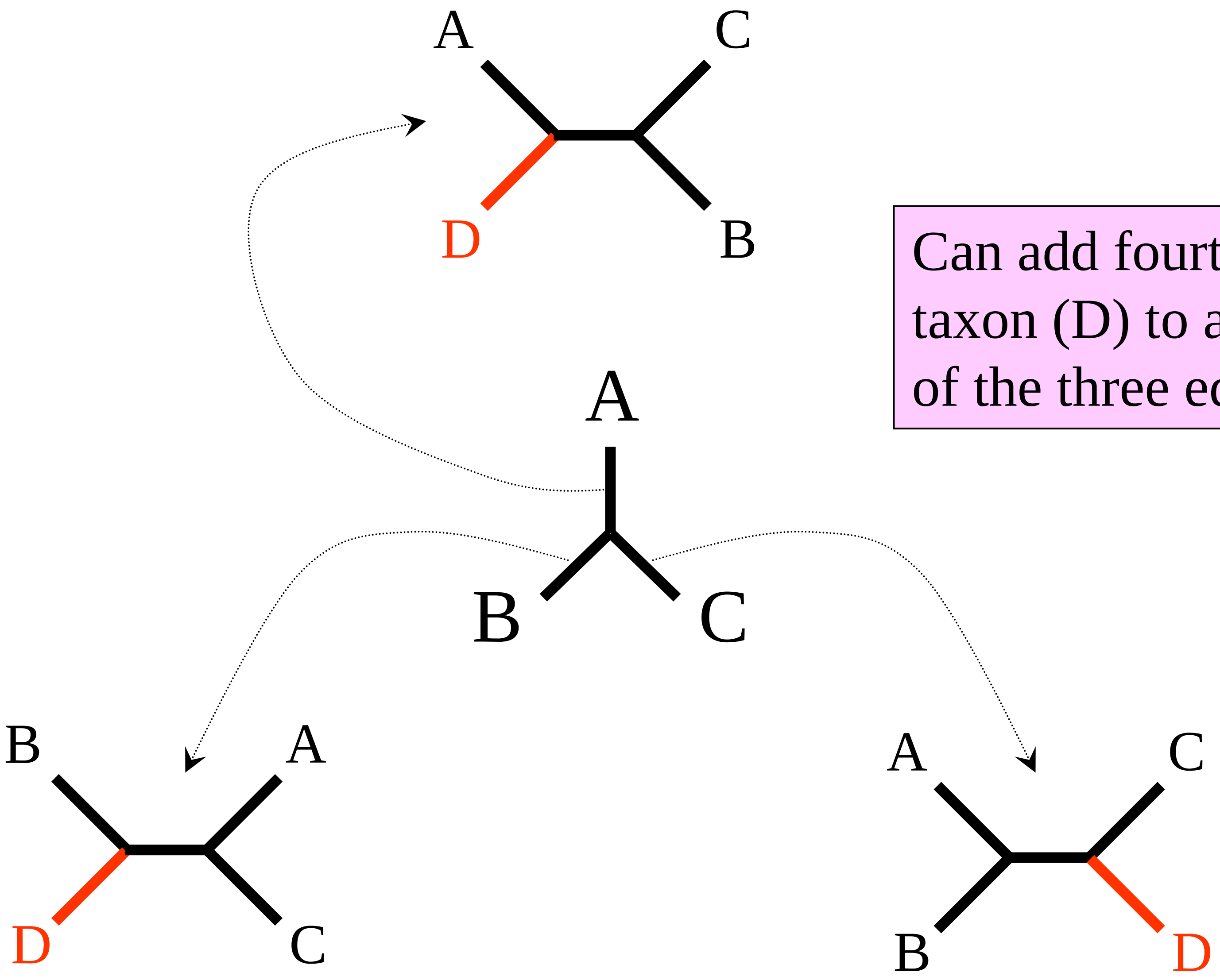
How do we find the best tree?
(or one that is good enough)

Exhaustive Enumeration



With the first three taxa, create the trivial unrooted tree

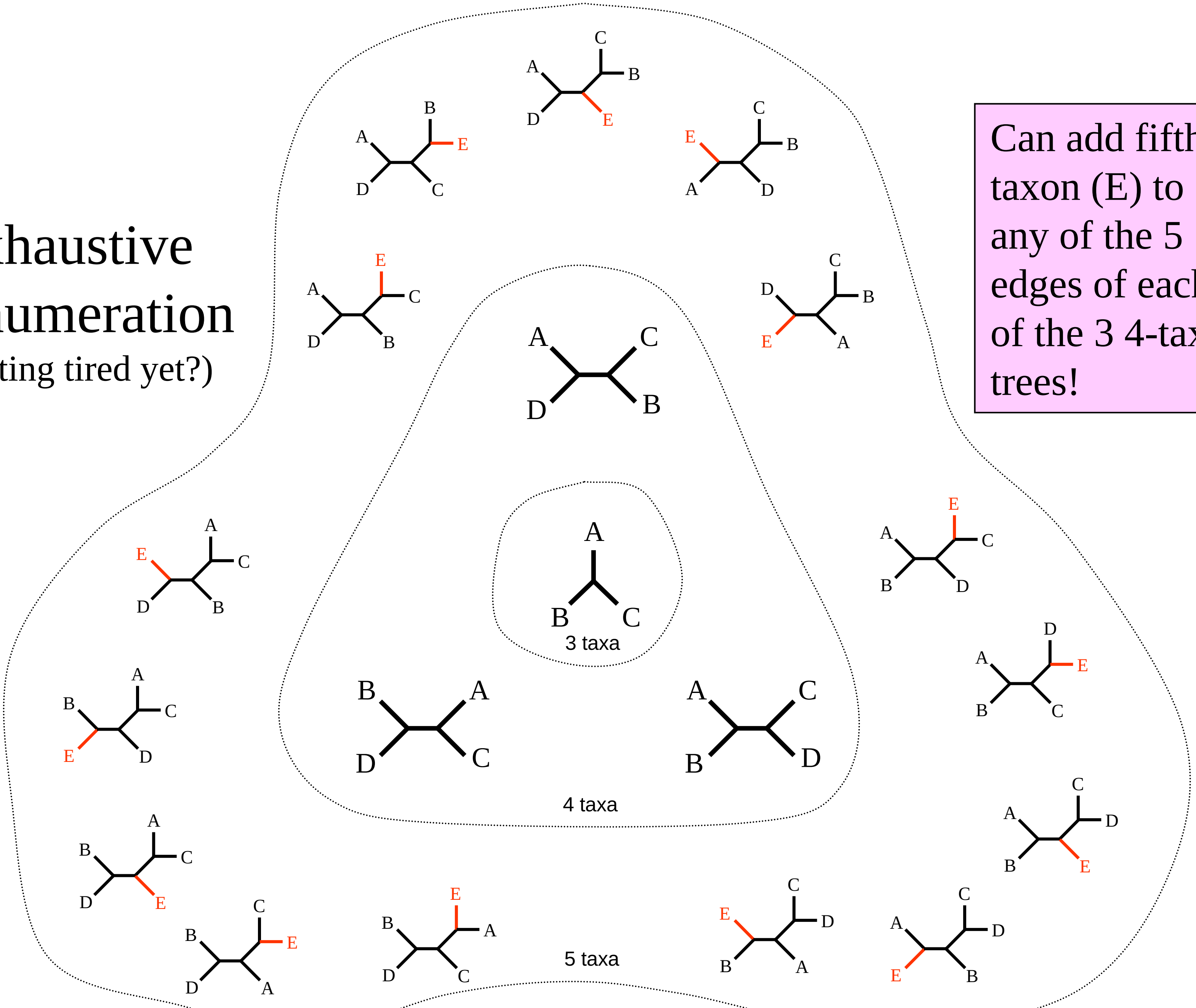
Exhaustive Enumeration...



Can add fourth taxon (D) to any of the three edges

Exhaustive Enumeration

(getting tired yet?)



Can add fifth taxon (E) to any of the 5 edges of each of the 3 4-taxon trees!

Tips	Number of unrooted (binary) trees	
4	3	
5	15	
6	105	
7	945	
8	10,395	
9	135,135	
10	2,027,025	
11	34,459,425	
12	654,729,075	
13	13,749,310,575	
14	316,234,143,225	
15	7,905,853,580,625	
16	213,458,046,676,875	
17	6,190,283,353,629,375	
18	191,898,783,962,510,625	
19	6,332,659,870,762,850,625	
20	22,164,309,5476,699,771,875	
21	8,200,794,532,637,891,559,375	
22	319,830,986,772,877,770,815,625	
23	13,113,070,457,687,988,603,440,625	> 21 moles of trees
24	563,862,029,680,583,509,947,946,875	

For N taxa:

$$\begin{aligned}
\text{\# unrooted, binary trees} &= \prod_{i=3}^{N-1}(2i-3) \\
&= \prod_{i=4}^{N}(2i-5) \\
\text{\# rooted, binary trees} &= \prod_{i=3}^{N}(2i-3) \\
&= (2N-3)(\text{\# unrooted, binary trees})
\end{aligned}$$