Some of these slides have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

Paul has many great tools for teaching phylogenetics at his web site:

http://hydrodictyon.eeb.uconn.edu/people/plewis

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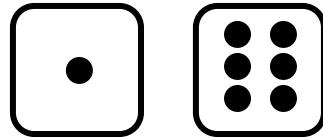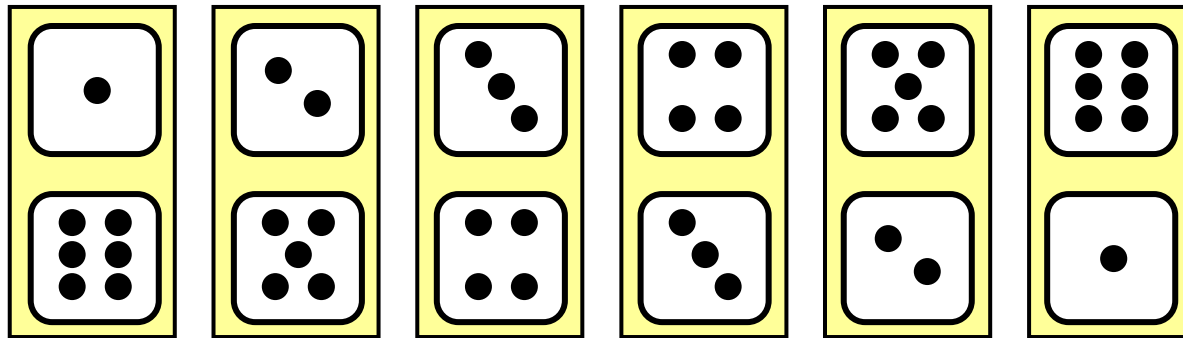
$$(1/6) \times (1/6) = 1/36$$

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)?"

3

Likelihood of a single sequence

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

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G$$

$$= \pi_A{}^{12} \pi_C{}^{7} \pi_G{}^{7} \pi_T{}^{6}$$

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

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.

How can we calculate the likelihood score

Under the JC (or K2P) model:

$$\ln L \; = \; 12 \ln \pi_A + 7 \ln \pi_C + 7 \ln \pi_G + 6 \ln \pi_T$$
$$= \; 12 \ln 0.25 + 7 \ln 0.25 + 7 \ln 0.25 + 6 \ln 0.25$$
$$= \; -44.361$$

How can we calculate the likelihood score

Under the F81 (or HKY or GTR) model:

$$\ln L \;=\; 12 \ln \pi_A + 7 \ln \pi_C + 7 \ln \pi_G + 6 \ln \pi_T$$

But what are the values for the parameters: $\pi_A, \pi_C, \pi_G, \pi_T$?

In many cases we refer to these parameters as "nuisance parameters." They must be specified in order to calculate the likelihood, but we are not interested in them by themselves.

ML parameter estimates

We can find the maximum likelihood estimates of the parameters to give us the ML score: the maximum likelihood obtainable under this model:

$$
\begin{aligned}
\ln L &= 12 \ln \pi_A + 7 \ln \pi_C + 7 \ln \pi_G + 6 \ln \pi_T \\
&= 12 \ln \widehat{\pi_A} + 7 \ln \widehat{\pi_C} + 7 \ln \widehat{\pi_G} + 6 \ln \widehat{\pi_T} \\
&= 12 \ln 0.375 + 7 \ln 0.21875 + 7 \ln 0.21875 + 6 \ln 0.1875 \\
&= -43.091
\end{aligned}
$$

But how did I get the numbers to fill for the parameters? How do we know that $\widehat{\pi_A} = 0.375$ and $\widehat{\pi_C} = 0.21875$

ML parameter estimates

We might guess that:

$$
\begin{aligned}
\widehat{\pi_A} &= \frac{12}{32} = 0.375 \\[2mm]
\widehat{\pi_C} &= \frac{7}{32} = 0.21875 \\[2mm]
\widehat{\pi_G} &= \frac{7}{32} = 0.21875 \\[2mm]
\widehat{\pi_T} &= \frac{6}{32} = 0.1875
\end{aligned}
$$

but how do we prove it?

ML parameter estimates

For simple problems we solve for the point in parameter space for which derivatives with respect to all parameters are 0 (we also have to consider boundary points).

We would have to do constrained optimization because

$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

and that is a pain.

ML parameter estimates

We can reparameterize:

$$
\begin{aligned}
r &= \pi_A + \pi_G \\
a &= \frac{\pi_A}{\pi_A + \pi_G} \\
c &= \frac{\pi_C}{\pi_C + \pi_T}
\end{aligned}
$$

and always recover the original parameters:

$$
\begin{aligned}
\pi_A &= ra \\
\pi_G &= r(1-a) \\
\pi_C &= (1-r)c \\
\pi_T &= (1-r)(1-c)
\end{aligned}
$$

ML parameter estimates

$$
\begin{aligned}
\ln L &= 12 \ln \pi_A + 7 \ln \pi_C + 7 \ln \pi_G + 6 \ln \pi_T \\
&= 12 \ln [ra] + 7 \ln [(1-r)c] + 7 \ln [r(1-a)] + 6 \ln [(1-r)(1-c)]
\end{aligned}
$$

Recall that:

$$
\frac{\partial \ln f(x)}{\partial x} = \frac{\frac{\partial f(x)}{\partial x}}{f(x)}
$$

$$
\begin{aligned}
\ln L &= 12 \ln [ra] + 7 \ln [(1-r)c] + 7 \ln [r(1-a)] + 6 \ln [(1-r)(1-c)] \\
\frac{\partial \ln L}{\partial a} &= \frac{12r}{ra} + \frac{7(-r)}{r(1-a)} \\
&= \frac{12}{a} - \frac{7}{(1-a)} \\
0 &= \frac{12}{\hat{a}} - \frac{7}{(1-\hat{a})} \\
\hat{a} &= \frac{12}{19}
\end{aligned}
$$

$$\ln L = 12 \ln [ra] + 7 \ln [(1-r)c] + 7 \ln [r(1-a)] + 6 \ln [(1-r)(1-c)]$$

$$\frac{\partial \ln L}{\partial c} = \frac{7(1-r)}{(1-r)c} + \frac{6 - (1-r)}{(1-r)(1-a)}$$

$$= \frac{7}{c} - \frac{6}{(1-c)}$$

$$0 = \frac{7}{\hat{c}} - \frac{6}{(1-\hat{c})}$$

$$\hat{c} = \frac{7}{13}$$

$$\ln L = 12\ln[ra] + 7\ln[(1-r)c] + 7\ln[r(1-a)] + 6\ln[(1-r)(1-c)]$$

$$\frac{\partial \ln L}{\partial r} = \frac{12a}{ra} + \frac{7(-c)}{(1-r)c} + \frac{7(1-a)}{r(1-a)} + \frac{6(-(1-c))}{(1-r)(1-c)}$$

$$= \frac{12}{r} - \frac{7}{(1-r)} + \frac{7}{r} - \frac{6}{1-r}$$

$$= \frac{19}{r} - \frac{13}{(1-r)}$$

$$0 = \frac{19}{\hat{r}} - \frac{13}{(1-\hat{r})}$$

$$\hat{r} = \frac{19}{32}$$

ML inference displays "scale invariance" so we can just transform the ML estimates into our original parameters:

$$\widehat{\pi_A} = \hat{r}\hat{a} \quad = \quad \left(\frac{19}{32}\right)\left(\frac{12}{19}\right) = \frac{12}{32}$$

$$\widehat{\pi_G} = \hat{r}(1-\hat{a}) \quad = \quad \left(\frac{19}{32}\right)\left(\frac{7}{19}\right) = \frac{7}{32}$$

$$\widehat{\pi_C} = (1-\hat{r})\hat{c} \quad = \quad \left(\frac{13}{32}\right)\left(\frac{6}{13}\right) = \frac{7}{32}$$

$$\widehat{\pi_T} = (1-\hat{r})(1-\hat{c}) \quad = \quad \left(\frac{13}{32}\right)\left(\frac{6}{13}\right) = \frac{6}{32}$$

Likelihoods on the simplest possible tree

GA⟶GG

$$L = L_1 L_2$$

$$= \Pr(G)\Pr(G \to G)\Pr(A)\Pr(A \to G)$$

$$= \Pr(G)\Pr(G \to G|\nu)\Pr(A)\Pr(A \to G|\nu)$$

$$= \left(\frac{1}{4}\right)\left(\frac{1}{4} + \frac{3}{4}\, e^{\frac{-4\nu}{3}}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4} - \frac{1}{4}\, e^{\frac{-4\nu}{3}}\right)$$

$$d = \frac{1}{4} - \frac{1}{4}\, e^{\frac{-4\nu}{3}}$$

$$\left(\frac{1}{4} + \frac{3}{4}\, e^{\frac{-4\nu}{3}}\right) = 1 - 3d$$

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

$$= \frac{(1 - 3d)d}{16}$$

$$\frac{\partial \ln L}{\partial d} = \frac{1 - 6d}{16}$$

$$0 = \frac{1 - 6\hat{d}}{16}$$

$$\hat{d} = \frac{1}{6}$$

$$\hat{\nu} = 0.82396$$

$$L = 0.005208$$

You may recall that the JC distance correction from lecture 8 looked like this:

$$\nu = \frac{-3}{4} \ln\left(1 - \frac{4p}{3}\right)$$

If you put in $p = 0.5$, because half the sites differ in our example then you the same branch length:
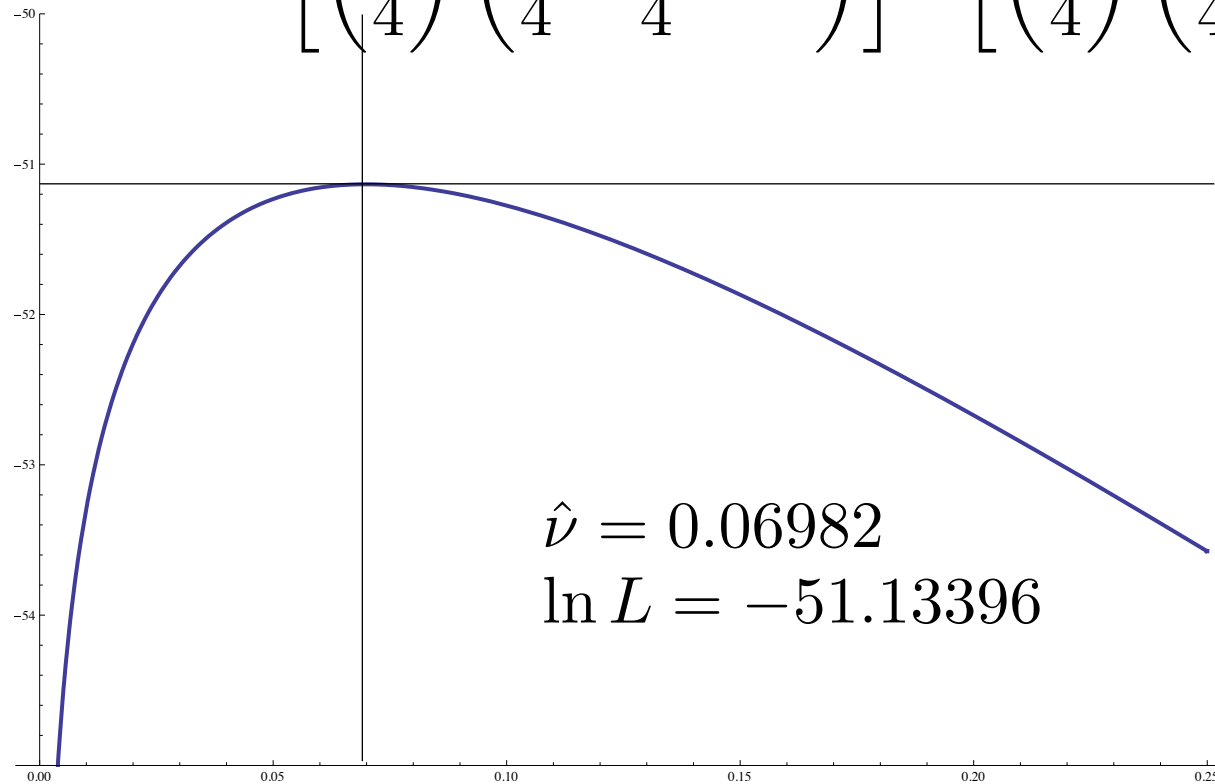
$$\nu = 0.82396$$

Our JC distance correction formula is actually an ML estimator of the branch length between a pair of taxa.

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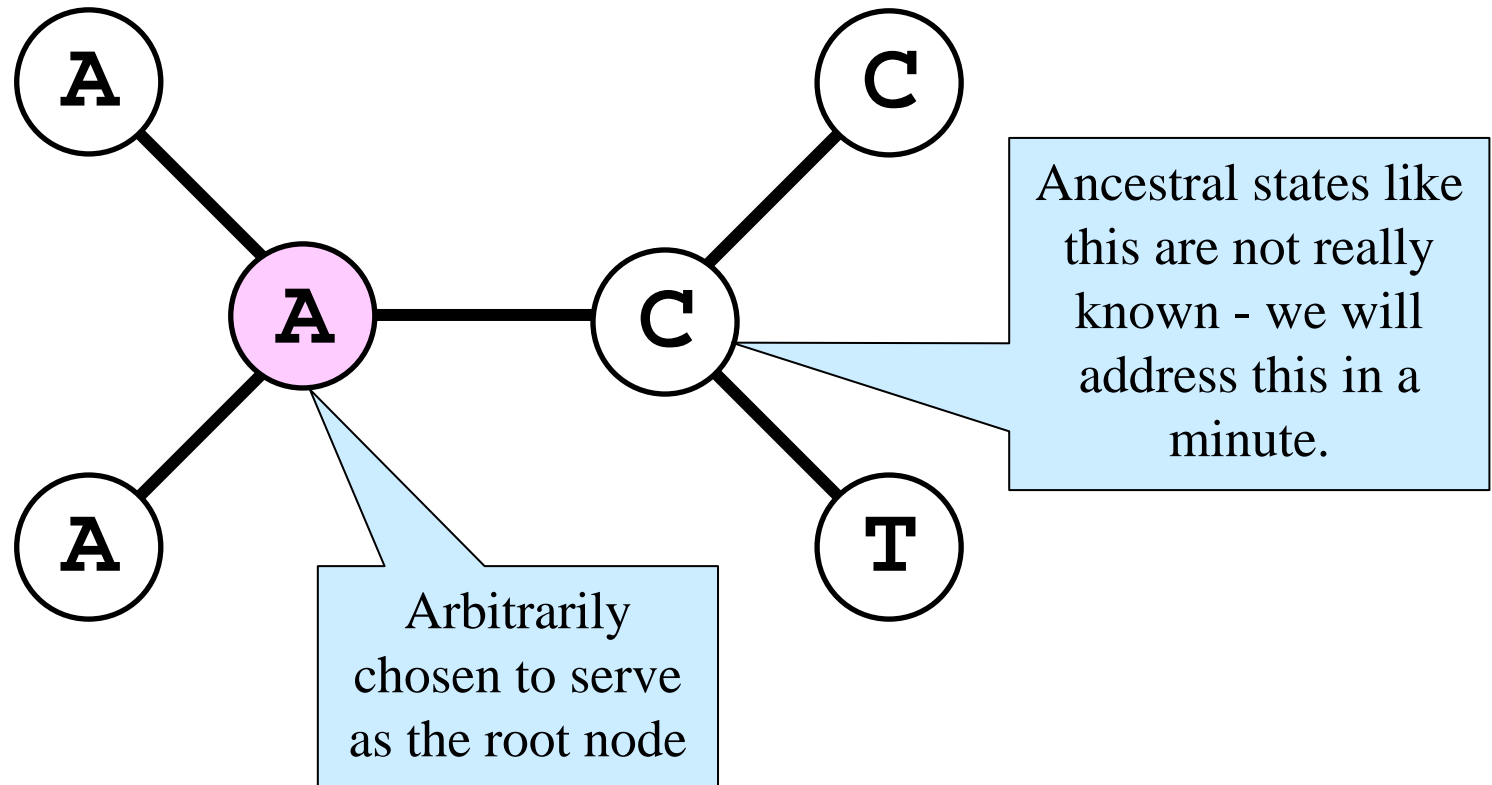
gorilla   GAAGTCCTTGAGAAATAAACTGCACACTGG

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)

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

Arbitrarily chosen to serve as the root node

9

Likelihood for site k

v_5 is the expected no. substitutions for just this segment of the tree

$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4v_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4v_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4v_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4v_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4v_5/3}\right]$$

π_A

$P_{AA}(v_1)$ $P_{AA}(v_2)$ $P_{AC}(v_3)$ $P_{CT}(v_4)$ $P_{CC}(v_5)$

10
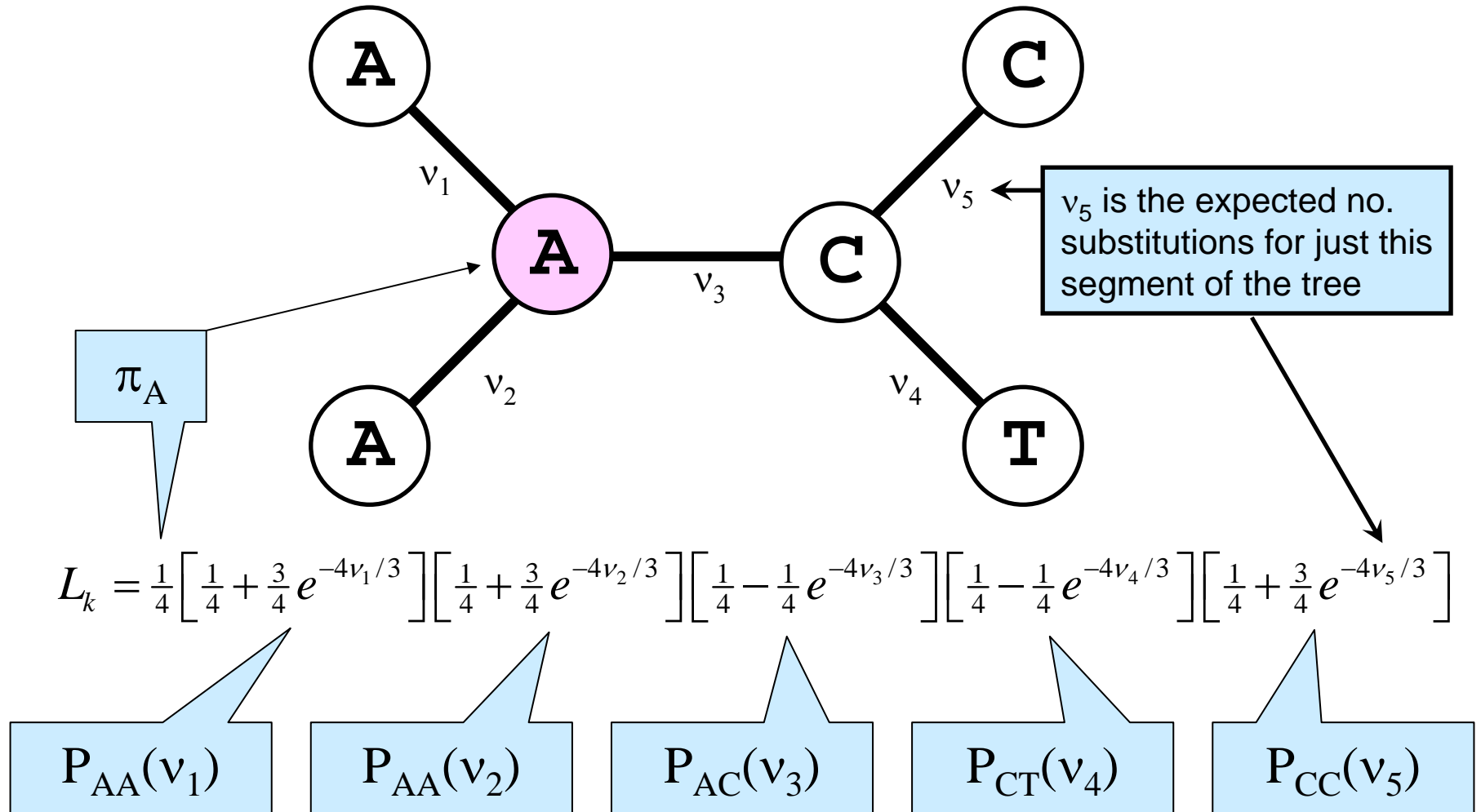
Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum

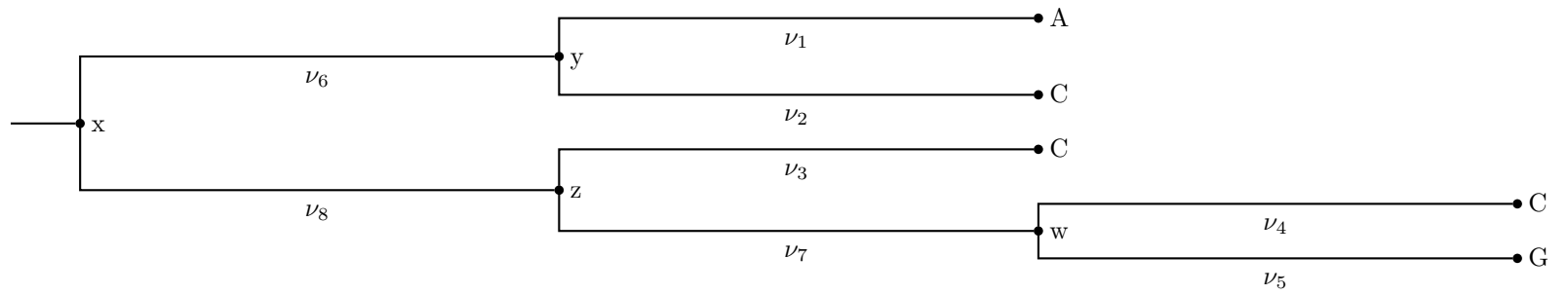11
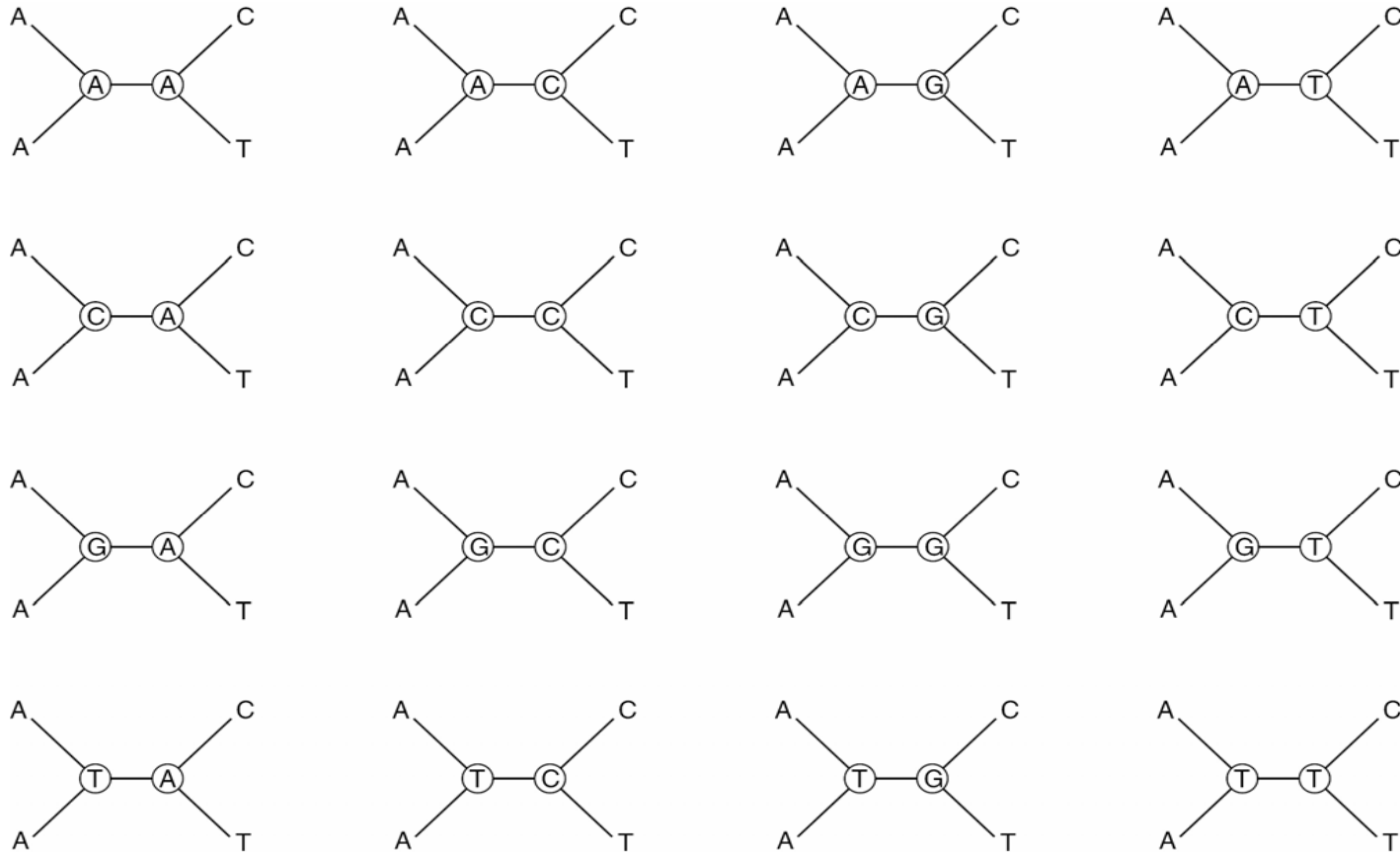
Pruning algorithm*
(same result, much less time)
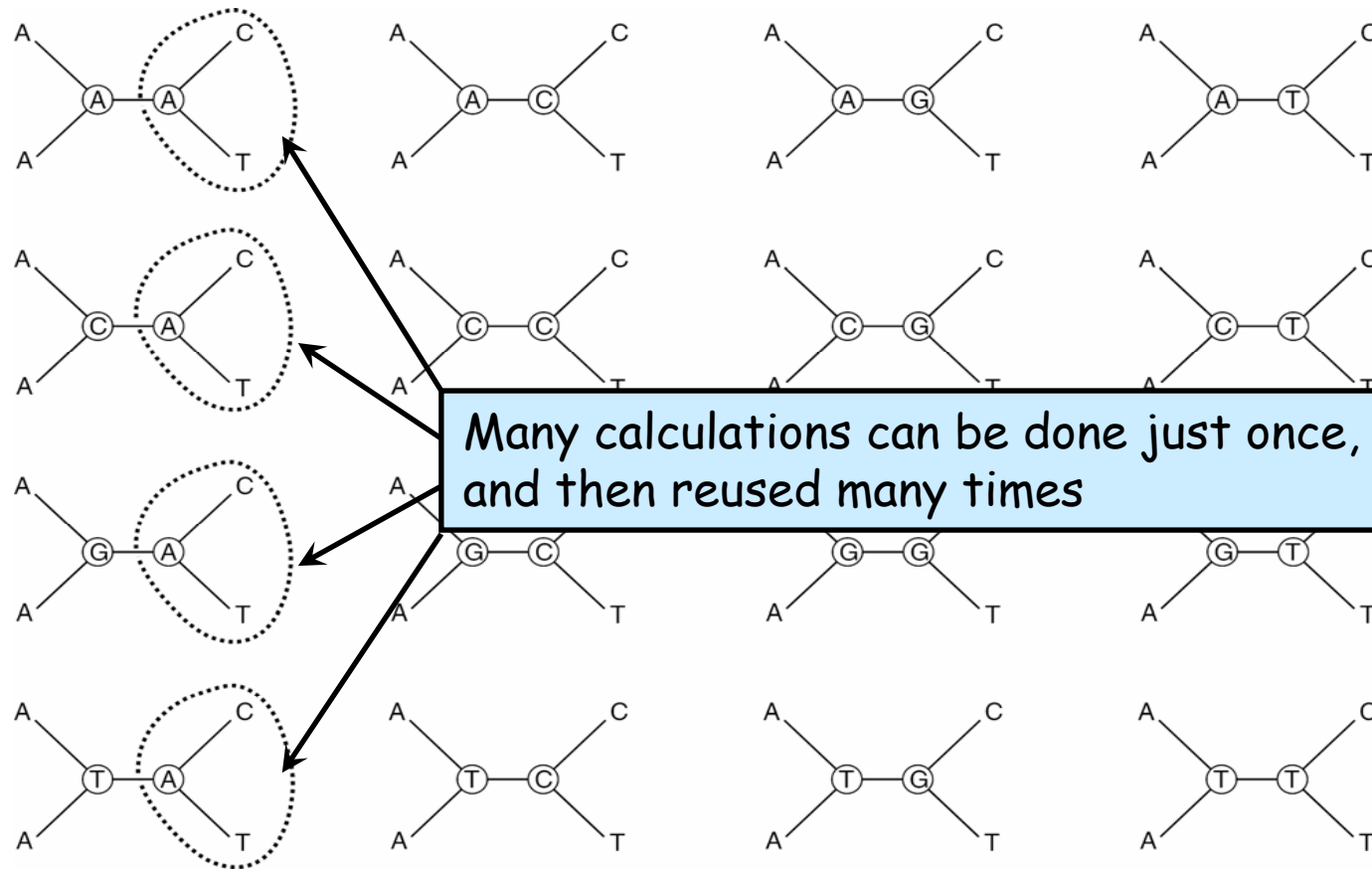
Many calculations can be done just once, and then reused many times

12

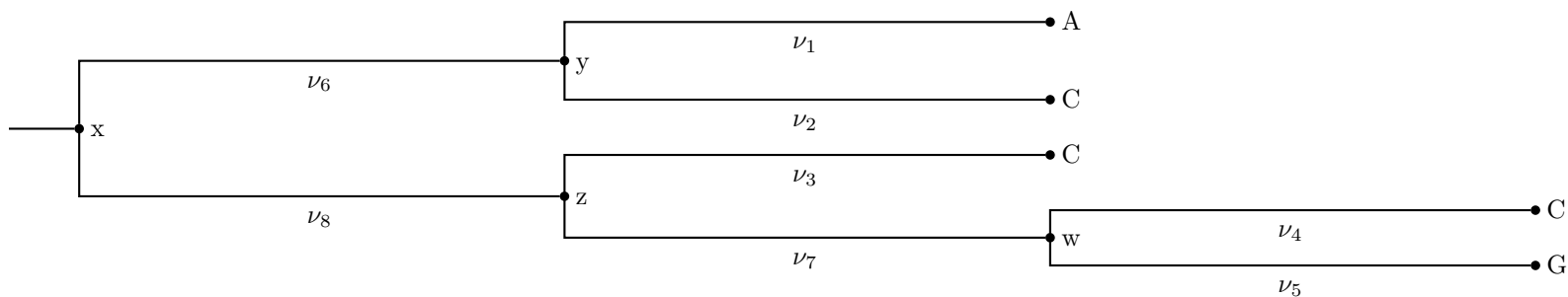| Taxon	Character
1	A
2	C
3	C
4	C
5	G
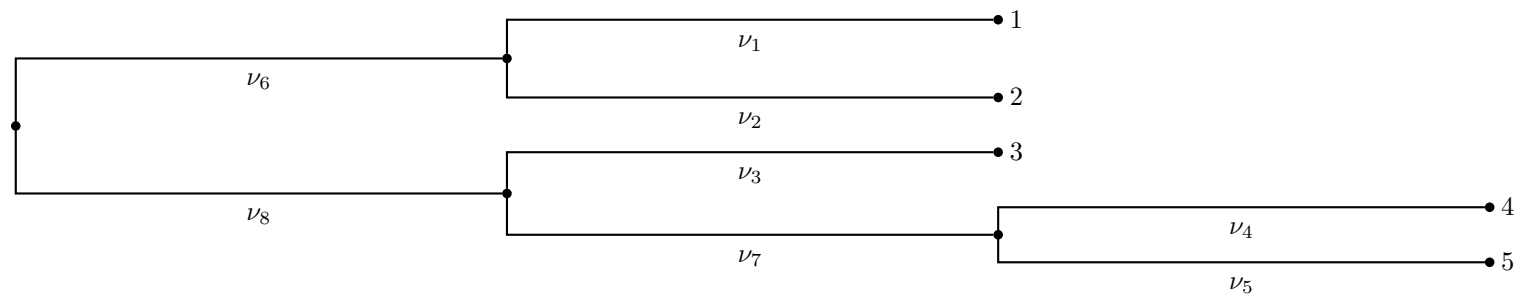
$$L = \sum_x \sum_y \sum_z \sum_w \Pr(x, y, z, w, A, C, C, C, G | \boldsymbol{\nu})$$
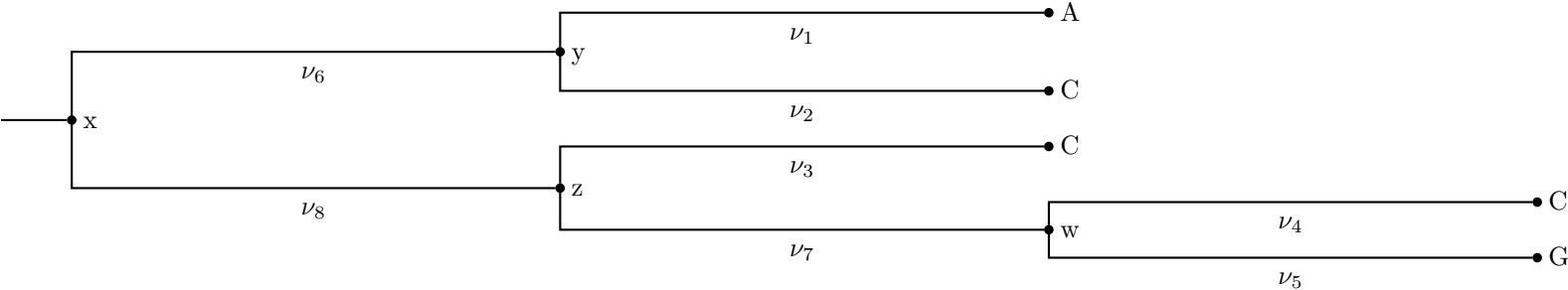
$$L = \sum_x \sum_y \sum_z \sum_w \Pr(x)\Pr(y|x,\nu_6)\Pr(A|y,\nu_1)\Pr(C|y,\nu_2)\cdots$$

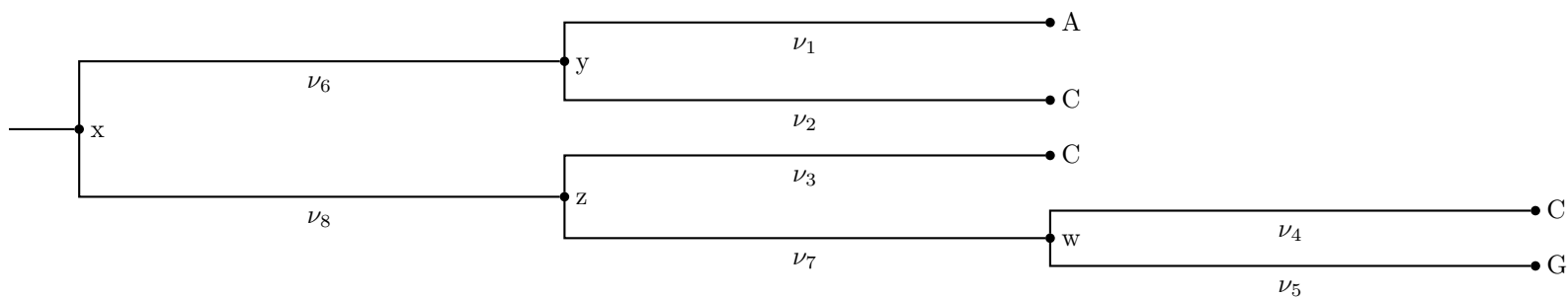$$\Pr(z|x,\nu_8)\Pr(C|z,\nu_3)\Pr(w|z,\nu_7)\Pr(C|w,\nu_4)\Pr(G|w,\nu_5)$$

$$L = \sum_x \sum_y \sum_z \Pr(x)\Pr(y|x,\nu_6)\Pr(A|y,\nu_1)\Pr(C|y,\nu_2)\cdots$$

$$\Pr(z|x,\nu_8)\Pr(C|z,\nu_3)\left(\textcolor{red}{\sum_w \Pr(w|z,\nu_7)\Pr(C|w,\nu_4)\Pr(G|w,\nu_5)}\right)$$

$$L = \sum_x \sum_y \Pr(x)\Pr(y|x,\nu_6)\Pr(A|y,\nu_1)\Pr(C|y,\nu_2)\cdots$$

$$\left(\textcolor{cyan}{\sum_z \Pr(z|x,\nu_8)\Pr(C|z,\nu_3)}\left(\textcolor{red}{\sum_w \Pr(w|z,\nu_7)\Pr(C|w,\nu_4)\Pr(G|w,\nu_5)}\right)\right)$$

$$L \;=\; \sum_x \Pr(x)\left(\sum_y \Pr(y|x,\nu_6)\Pr(A|y,\nu_1)\Pr(C|y,\nu_2)\right)\cdots$$

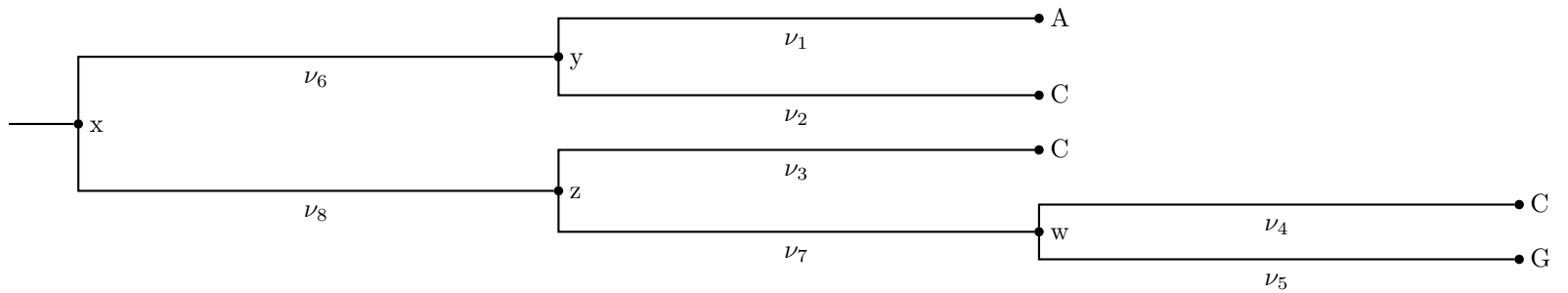$$\left(\sum_z \Pr(z|x,\nu_8)\Pr(C|z,\nu_3)\left(\sum_w \Pr(w|z,\nu_7)\Pr(C|w,\nu_4)\Pr(G|w,\nu_5)\right)\right)$$

Maximum likelihood is a lot of work

- Site likelihoods involve products of transition probabilities, summed over ancestral states
- Overall log-likelihood for a tree is sum of site log-likelihoods
- Overall log-likelihood must be maximized!
  - must find MLEs for all edge lengths and all model parameters
  - this involves computing the overall log-likelihood **many, many times** (try turning on **logiter** in PAUP to get a feel for how much work this involves)
- Maximized lnL can now be compared to maximized lnL from other trees

Is it worth it?

- Uses all information
  - Parsimony ignores constant and autapomorphic sites
  - Distance methods ignore information not captured in pairwise comparisons

- Model generality
  - Some models possible with distance methods, but some quantities cannot be estimated reliably (e.g. variation in rates across sites)
  - Many parsimony variants exist, but parsimony does not allow estimation of the step matrix entries, for example
  - Many complex models are only possible under likelihood or Bayesian methods (which have a likelihood foundation)