PhyloMath Lecture 9 – March 30th 2004

C.A. Fyler

Last time we established the likelihood for one site (*k*) in a four taxon tree (Fig. 1) by calculating the likelihood at every possible ancestral combination (16 in total). The overall likelihood for that site being:

$L_k = \pi_G \Sigma_i \Sigma_j P_{Gi}(d_2) P_{iT}(d_1) P_{ij}(d_3) P_{jA}(d_4) P_{jA}(d_5)$

To calculate the likelihood of a tree incorporating multiple sites we would multiply the likelihoods of all sites $L = L_1 L_2 ... L_\kappa ... L_n$, **or** take the Log likelihoods of all the sites and add them to avoid ridiculously small values that occur when small numbers are multiplied ($\ln L = \ln L_1 + \ln L_2 + ... + \ln L_\kappa + \ln L_n$).

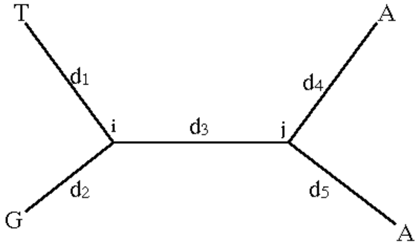


Figure 1

In a more realistic example we would not be given the branch lengths and would estimate them as well. The estimation process is like having 6 knobs[1] each allowing adjustment of a single parameter. There is a common meter measuring the overall likelihood, and the goal is to fiddle with all the knobs until you can push the meter no higher. After maximizing the likelihood of one parameter, the other parameters in your model will not generally be maximized anymore. So, you can imagine that finding the ML for your tree would take a lot of knob fiddling!

Felsenstein's Pruning Algorithm

Felsenstein (1981) was the first one to apply the pruning algorithm to likelihood. Notice that with 16 possible ancestral combinations we are calculating some of the same numbers over and over again (Paul's handout of the 16 possible trees for all possible ancestral state combinations demonstrates this point). We can economize on computation by using the computer's memory to store the results of certain computations. This is a trade off since we are sacrificing the computer's memory in exchange for speed. The method of pruning can be derived simply by moving summation signs to the inside of the equation.

Recall:

$L_k = \pi_G \Sigma_i \Sigma_j P_{Gi}(d_2) P_{iT}(d_1) P_{ij}(d_3) P_{jA}(d_4) P_{jA}(d_5)$

[1] 6 knobs in this case because we decided on a K80 model. 5 branch lengths + κ.

Moving the summation signs in gives us:

$L_\kappa = \pi_G \Sigma_i P_{Gi}(d_2) P_{iT}(d_1) \Sigma_j P_{ij}(d_3) P_{jA}(d_4) P_{jA}(d_5)$

You may notice that the terms for the tips of the tree in this equation (G,T) (A,A) are in exact correspondence to the structure of the tree in Figure 2.

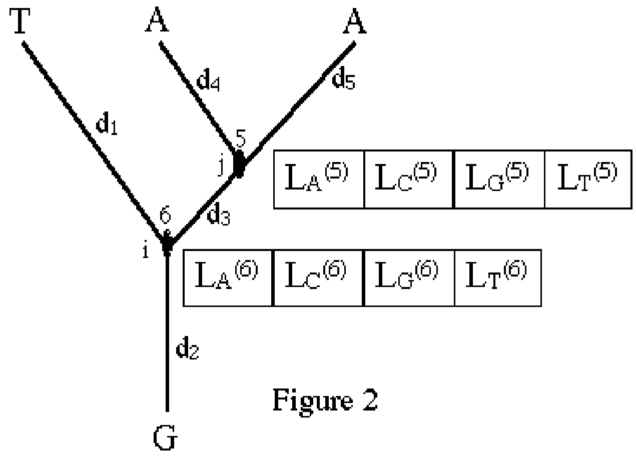


Figure 2

Now we can calculate the conditional probabilities of the sub-trees:

To get ourselves aquatinted: $L_G^{(5)} = P_{GA}(d_4) P_{GA}(d_5)$
$L_T^{(5)} = P_{TA}(d_4) P_{TA}(d_5)$
$L_T^{(6)} = P_{TT}(d_1) \Sigma_j P_{Tj}(d_3) L_j^{(5)}$

Therefore, the general formula for node 5 is: $\mathbf{L_j^{(5)} = P_{jA}(d_4) P_{jA}(d_5)}$
And, the general formula for node 6 is: $\mathbf{L_i^{(6)} = P_{iT}(d_1) \Sigma_j P_{ij}(d_3) L_j^{(5)}}$

The key to the pruning algorithm is that once the general formulas are computed, they need not be continually recomputed. Conditional Likelihood Arrays[2] save time but can stack up memory. For example, 500 patterns + 200 taxa will take up as much as 3MB to just store temporary results compared to the 100 kb that it will take to store the actual data. 30x more memory to store intermediate formulas!

From the above example, the overall Likelihood of our site will be: $\mathbf{L_k = \pi_G \Sigma_i P_{Gi}(d_2) L_i^{(6)}}$

We can substitute the general formulas for nodes 5 and 6 into the overall likelihood equation to confirm that it is the same as our original equation for the likelihood of a site.

$$L_k = \pi_G \sum_i P_{Gi}(d_2) \left[P_{iT}(d_1) \sum_j P_{ij}(d_3) L_j^{(5)} \right]$$

First substitute in $L_i^{(6)} = P_{iT}(d_1) \Sigma_j P_{ij}(d_3) L_j^{(5)}$

[2] Called conditional because it is conditional on a node having a particular state.

Then substitute in $L_j^{(5)} = P_{jA}(d_4)\, P_{jA}(d_5)$:

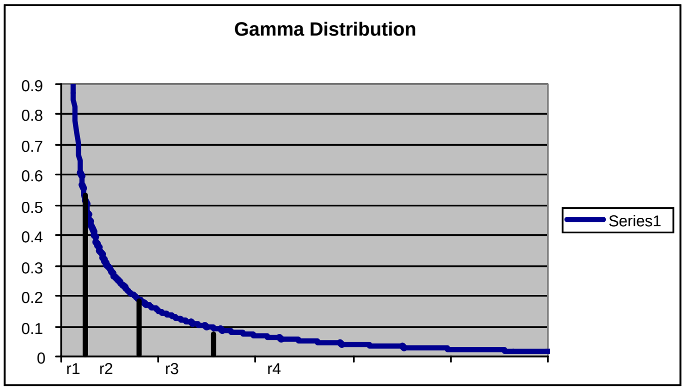
$$L_k = \pi_G \sum_i P_{Gi}(d_2)\left[P_{iT}(d_1) \sum_j P_{ij}(d_3)\left(P_{jA}(d_4) P_{jA}(d_5)\right)\right] \Leftarrow \text{this is the same equation we started with!!}$$

CALCULATING LIKELIHOOD WITH RATE HETEROGENEITY

To incorporate rate heterogeneity first specify the shape parameter (α) of the gamma distribution and then determine the mean rate of DNA evolution for subsets of the distribution. Using more divisions (and therefore more rates) will allow for more of your data to be explained properly but there is a trade off between the number of rates and computation time. In this case we divided the gamma distribution into four equal areas.

$$L_k = \Pr(D_k \mid \theta)$$

where θ = all parameters in the model, and D_k = the data for site k



As with ancestral states, we do not know which of the four representative rates applies to this site, so the likelihood for site *k* is a sum over all the rates.

$$L_k = \Pr(D_k \text{ and } r_1 \mid \theta) + \Pr(D_k \text{ and } r_2 \mid \theta) + \Pr(D_k \text{ and } r_3 \mid \theta) + \Pr(D_k \text{ and } r_4 \mid \theta) \Leftarrow \text{Joint Probability}$$

$$= \Pr(r_1)\Pr(D_k \mid r_1, \theta) + \Pr(r_2)\Pr(D_k \mid r_2, \theta) + \Pr(r_3)\Pr(D_k \mid r_3, \theta) + \Pr(r_4)\Pr(D_k \mid r_4, \theta) \Leftarrow \text{Conditional probability}$$

Note added by Paul:

Each of the four terms above (e.g. Pr(D_k and r_1 | θ)) involves a sum over the 16 possible ancestral state combinations (or at least involves the equivalent of such a sum; the actual work done is less because of the use of the pruning algorithm).

Remember that we divided up the gamma distribution evenly, so

Pr(r_1) = Pr(r_2) = Pr(r_3) = Pr(r_4) = 1/4 (or 1/the number of categories)

To compute $\Pr(D_k \mid \theta)$, we need to know the transition probabilities. Here are the transition probabilties for the JC69 model:

$$P_{ij}(\alpha t)=\begin{cases}\frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & i=j\\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & i\neq j\end{cases}$$

In order to compute $\Pr(D_k \mid r_1,\theta)$, we need only substitute transition probabilities that are conditional on r_1:

$$P_{ij}(r_1\alpha t)=\begin{cases}\frac{1}{4}+\frac{3}{4}e^{-4r_1\alpha t} & i=j\\ \frac{1}{4}-\frac{1}{4}e^{-4r_1\alpha t} & i\neq j\end{cases}$$

The same substitution is then done for the other relative rates.

Invariant sites (pinvar) It is often realistic to assume that there is some proportion of the sites that are invariant (have zero rate of change). The invariant sites model rate is either 0 or *r*, therefore:

$$P(\mathrm{D}_{\kappa}|r_{1,}\theta)=\Pr(rate=0)\Pr(D_{\kappa}|rate=0,\theta)+\Pr(rate=r)\Pr(D_{\kappa}|rate=r,\theta)$$

The quantity $\Pr(\text{rate}=0)$ is just the parameter pinvar, whereas $\Pr(\text{rate}=r)$ is 1 – pinvar. The transition probabilities become very simple when the rate is assumed to be 0. For the JC69 model

$$P_{ij}(0\alpha t)=\begin{cases}\frac{1}{4}+\frac{3}{4}e^{-4(0)\alpha t}=1 & i=j\\ \frac{1}{4}-\frac{1}{4}e^{-4(0)\alpha t}=0 & i\neq j\end{cases}$$

The relative rate r can be determined from knowledge of pinvar and from the fact that relative rates are normalized so that their expected value is 1:

$$E(\text{rate})=(0)\Pr(\text{rate}=0)+(r)\Pr(\text{rate}=r)=1$$
$$r=\frac{1}{\Pr(\text{rate}=r)}=\frac{1}{1-\text{pinvar}}$$

Thus, the JC69 transition probabilities for the case in which the rate = *r* would be:

$$P_{ij}(r\alpha t)=\begin{cases}\frac{1}{4}+\frac{3}{4}e^{-4\alpha t/(1-\text{pinvar})} & i=j\\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t/(1-\text{pinvar})} & i\neq j\end{cases}$$

MOLECULAR CLOCK ASSUMPTIONS

A molecular clock assumes that on average the rate of molecular evolution is invariable throughout long periods of evolutionary time across multiple lineages. Enforcing a molecular clock allows you to assume that the α part of the equation is the same for each branch length.

For example, with out a molecular clock using a JC Model:

$$d_1 = 3\alpha_1 t_1$$
$$d_2 = 3\alpha_2 t_2$$

If $d_1 > d_2$, then we don't know if it is due to a difference in t or a difference in α. If a molecular clock is enforced then the rates are the same for every part of the tree so:

$$d_1 = 3\alpha t_1$$
$$d_2 = 3\alpha t_2$$

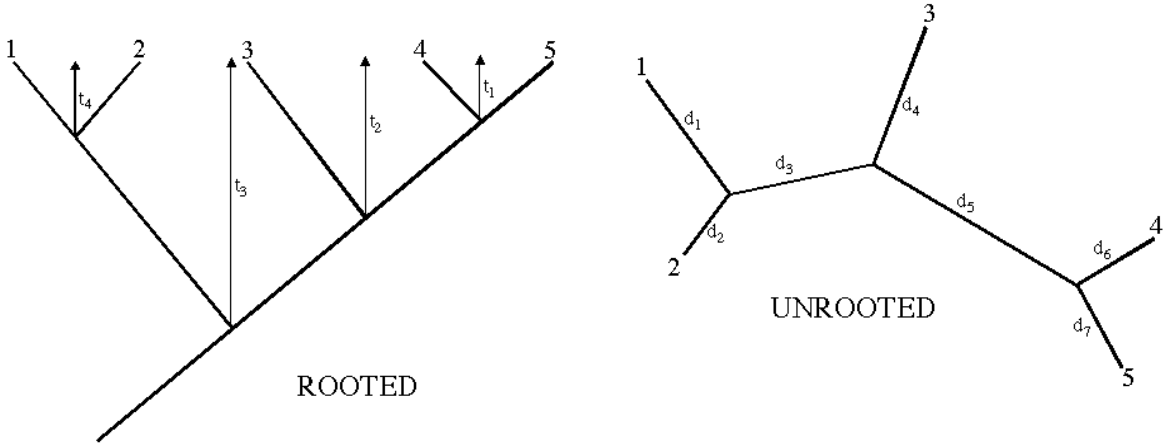


Figure 3

The rooted tree in figure 3 assumes a molecular clock. In the rooted tree all the tips are equidistant from the root and the number of substitutions are the same for equivalent amounts of time regardless of the lineage. The rooted MC tree has four branch lengths instead of 7 and we can calculate the likelihood using t_1-t_4 instead of the internal branches.

ROOTED TREE	UNROOTED TREE
MC Assumptions	NO MC Assumptions
No. nodes = 2n-1 (9 in this case)	No. nodes = 2n-2 (8 in this case)
No. interior nodes = n-1	No. interior nodes = n-2
No. branch lengths (= no. interior nodes) = n-1 (4 in this case)	No. branch lengths = 2n-3 (7 in this case)

Do the data make the molecular clock assumption implausible?
L_o = likelihood under the assumption of a molecular clock (the constrained, simpler hypothesis is always the null hypothesis).
L_1 = likelihood under the alternate "no clock" hypothesis (unconstrained hypothesis).

If the Likelihood ratio is about 1 then both models (with and without the MC) are explaining the data equally well. If $L_1 / L_o >> 1$ then you do not believe the null hypothesis as much.

L_1 will always be greater or equal to L_o (because we could always constrain the extra parameters of the no clock model to match any particular configuration in the constrained model) but we want to find out if the no clock model is a significantly better explanation of our data given the difference in parameterization. To do this, perform a Likelihood ratio test.

Likelihood ratio test (LRT) = 2(lnL$_1$ – lnL$_o$)

Since the distribution of LRT is nearly identical to a χ^2 distribution, we can use the chi-square distribution with the degrees of freedom equal to the *difference* in the number of parameters between the two models. The number of substitution model parameters (e.g. κ,α) are the same between the two models, so the only difference is in the number of branch length parameters:

d.f. = (no. branch lengths in unconstrained model) – (no. branch lengths in clock model)
=(2n-3) – (n-1)
= (2n-n) –3 +1
= n-2

BAYESIAN ANALYSES

Given a hypothesis θ (such as a possible tree) and some data *D*, the probability of the hypothesis given the data (the **posterior probability** of hypothesis θ) is:

$$\Pr(\theta|D) = \frac{\Pr(\theta \,\&\, D)}{\Pr(D)}$$

The joint prob. in the numerator can be written as a conditional probability:

$$\Pr(\theta \,\&\, D) = \Pr(\theta)\Pr(D|\theta)$$

The Pr(θ) is called the **prior probability** of hypothesis θ, and Pr(D| θ) is the likelihood of hypothesis θ (note: do not say “likelihood of the data”, because the data are constant; likelihoods are functions of hypotheses). Substituting back into the first equation:

$$\Pr(\theta|D) = \frac{\Pr(\theta)\Pr(D|\theta)}{\Pr(D)}$$

The equation above that relates the posterior probability to the product of the prior probability and likelihood is known as **Bayes’ Theorem**, or Bayes’ Rule. The denominator is a constant (it depends only on the data, which is itself constant for any given analysis). It is sometimes referred to as the **marginal probability of the data**, where *marginal* refers to the fact that it is the total probability considering all possible values of θ:

$$\Pr(D) = \Sigma_\theta \Pr(\theta)\Pr(D|\theta)$$

The quantity on the left is the posterior probabliity of θ. The sum of the posterior probabilities of all possible hypotheses θ must be 1:

$$\sum_{\theta} \Pr(\theta \mid D) = \sum_{\theta} \frac{\Pr(\theta)\Pr(D \mid \theta)}{\Pr(D)} = 1$$

It is more common to use Bayes' Rule in the context of continuous hypotheses. For example, you can think of every possible value of the transition/transversion rate ratio k (from 0 to infinity) as a separate hypothesis, in which case there are an infinite number of hypotheses to consider. In this case, probabilities become probability densities and sums become integrals, but Bayes' Rule works exactly the same way:

$$f(\kappa \mid D) = \frac{f(\kappa) f(D \mid \kappa)}{f(D)}$$

SUMMARY-

$$f(\theta \mid D) \propto f(\theta) f(D \mid \theta)$$

WHERE

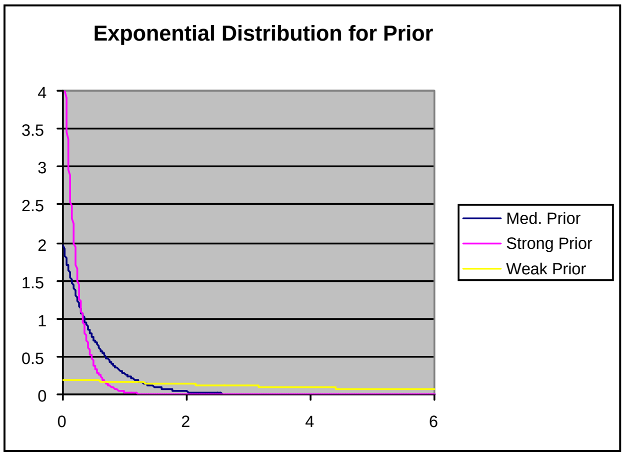
$f(\theta|D) =$ posterior probability density, we make all of our inferences from this

$f(\theta) =$ prior density, what we presume to know independent of our data

$f(D|\theta) =$ likelihood, contains the information contained in the data

If we have a "strong" prior then the data will have less influence on the posterior. Strong priors are said to be *informative*. If the prior is "weak", then the data will have more influence on the posterior. Weak priors are said to be uninformative or flat, if they specify the same prior probability density for every value, or vague if they give preference to some values over others, but not by much.

The following chart shows three exponential priors. One is described as a strong prior. Note how little density is given to values above 2 in this prior compared to values less than, say, ½. Now look at the prior describe as weak and compare the density for 2 vs. the density at ½. This prior puts almost as much weight on 2 as it does on ½.


Exponential Distribution for Prior
Med. Prior
Strong Prior
Weak Prior