## The gamma distribution model of among-site rate variation.

1. The models we discussed in class all assume that the same branch lengths apply to every site. That is, they assume that the probability of substitution along any branch is the same for every site in the sequence. This assumption will be wrong if some sites are subject to stronger constraints than others. The ML branch lengths using such a model will be a "compromise" over fast- and slow-evolving sites, and they will be incorrect for most sites, which can lead to incorrect inference of phylogeny, ancestral sequences, etc.

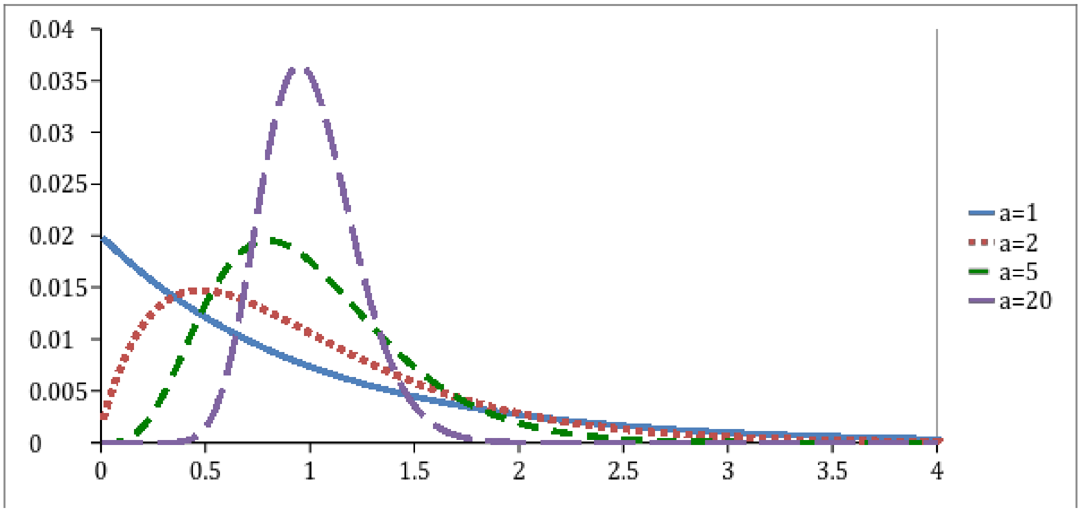
2. Thus we are motivated to develop models that incorporate among-site rate variation (**ASRV**). Because we don't know a priori which sites are fast and slow (or how fast and slow each one is), we usually use a "mixture model" that first calculates the likelihood at that site given several different branch length sets and then summing those up to yield the total likelihood. That is:

$$P(D_i|T,\ M) = \sum_j P(D_i|T,\ M,\ Bj)$$

   where $D_i$ is the data at one site in the sequence, T is the tree, M is the model, and $B_j$ is some set of branch lengths on the tree.

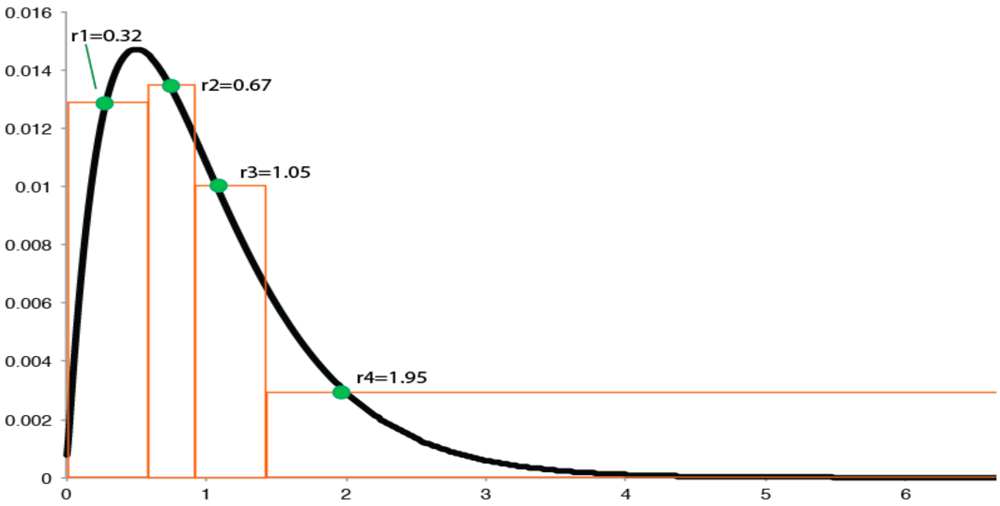
3. By far the most commonly used mixture model of ASRV is the so-called Gamma ASRV model. This approach uses a set of branch length multipliers, such that all of the branch lengths $B_j$ are simply the branch lengths in some common "base" set of branch lengths *B*, all multiplied by the same rate multiplier $r_j$. When r is >1, the entire tree is stretched by a scale factor equal to *r.* If r is <1, the entire tree is shrunk by that scale factor. By using several different values of *r,* we can model the situation in which some sites evolve rapidly and other sites evolve slowly. (Note that if a site is slow, it is slow across the entire tree, and if it is fast, it is fast across the entire tree. This comes from the fact that the multiplier *rj* is applied to all the branches on the tree.)

4. So where do the values of *r* come from? This is what the gamma distribution is used for. The gamma distribution is convenient, because it is a probability distribution that can take a large number of shapes (all monotonic) by changing the value of a single parameter, α, also called the "shape parameter." The illustration below shows some gamma distributions under different values of α, the mean is maintained at 1.



5. As the graph shows, if a is small, then the probability that any site will evolve slowly is high, but there is a small probability that the site will evolve very fast. (You can also read this as "most sites evolve very slowly, but a few sites evolve very rapidly."). If a is large, there is little rate variation, and sites almost certainly have rates very close to the base branch length.

6. So how do we put this into action? We need specific values of *r* to multiply the branch lengths by, calculate likelihoods, and then sum over them. The continuous distributions above make this hard, but it is easy to use a discrete approximation of any continuous gamma distribution. Suppose we want to approximate the distribution when α=2.0, using four discrete categories. We simply split the distribution up into four bins, each of which contains 25% of the total area under the curve of the distribution (and that total area is always 1, because it is a probability distribution). The *r* for that bin is the weighted mean of the distribution in that bin.



7. So now we have four branch length multipliers $r_j$, each with equal probability ($p_j$=0.25). We take each $r_j$, multiply all the branch length on the tree by that factor, and calculate the likelihood given those branch lengths. We weight each likelihood by the probability of that set of branch lengths under the model (which is 0.25), and sum the likelihoods to get the total likelihood. We sum because we don't know which is true, so it could be that r1 is true OR r2 is true OR r3 is true OR r4 is true. In that case, we add the probabilities.

$$P(D_i|T,\ M) = \sum_j P(T,\ M,\ B * r_j) * p_j$$

8. Different values of α will yield different values of $r_j$ , of course. A low value of α models a situation with a lot of rate variation: the $r_j$s will range from very small to very large, yielding some categories of the model with very short branch lengths and some categories with very long branch lengths. A large of value of α will yield $r_j$ values all close to one. Thus, the likelihoods calculated will be different when different values of alpha are used.

9. We can therefore find the best-fit estimate of alpha by calculating the likelihood of many values of alpha using a hill-climbing approach. The ML estimate of alpha is the one with the highest likelihood: that is, the one that yields the highest probability that the sequence data we observe would have evolved. Thus alpha becomes a free parameter of the model to be optimized by maximum likelihood. The likelihood of a tree (or of an ancestral state, or anything else) is then the probability of the data given the tree, the model, and the ML estimates of all the free parameters, including alpha.

10. When rate variation is present, this mixture model works much better than a model that assumes all sites evolve at the same rate. Although the same set of branch length multipliers is applied to all sites, the mixture allows sites that evolved very slowly to contribute a lot of likelihood to the sum when the branch lengths are short (that is, the term of the sum using the smaller values of $r_j$). A fast-evolving site will contribute most to the sum when the branch lengths are long – that is, when the larger values of $r_j$ are used.

11. We can categorize sites into rate bins after the fact by asking what fraction of the total likelihood at that site came from using each $r_{j.}$ The posterior probability that a site *i* is in the slowest bin (the one with the smallest multiplier $r_1$) is:

$$PP(r_{j=1,i}|T,\ M) = \frac{P(T, M, B{*}r_{j=1}){*}p_{j=1}}{\sum_j P(T, M, B{*}r_j){*}p_j}$$

Because our bins all have equal sizes, the $p_j$ terms all cancel out, so the PP that a site is “in” a bin is simply the fraction of the total likelihood at the site contributed when the branch lengths for that bin are assumed.

12. The use of 4 categories to discretize the distribution was arbitrary. We could use 5, 8, 16, or 100. In fact, studies have found that the number of categories used doesn’t matter much to the likelihood; you can get a decent approximation with 4, a good one with 8, and not much change as you go higher. Thus, people typically use a 4- or 8 category discretization. Note that no matter how many categories you use, the only parameter that determines the values of $r_j$ is $\alpha$, the shape parameter of the distribution. This is the only free parameter in the model: all the rate multipliers, however many there are, depend entirely on $\alpha$. This means that when a likelihood ratio test is used to determine whether the gamma-distributed ASRV model should be used, only one degree of freedom is required.