Integrative Biology 200B
"Ecology and Evolution"

University of California, Berkeley, Spring 2011
by NM Hallinan, updated by Nick Matzke

Lab 8: Molecular Evolution

There are many different features of genes and genomes that can be explored using phylogenetic methods. Today we're going to do a likelihood test for different rates of evolution in different parts of a DNA sequence. This is in general an important part of studying gene evolution. Knowing if different parts of genes evolve at different rates allows us to: use the appropriate model of sequence evolution when deducing gene phylogenies; detect the affects of natural selection on genes; and better understand the patterns and processes involved in the evolution of genes and genomes.

We will attempt to detect different rates of evolution between introns and exons within a gene. Our data set consists of five paralogous Protein Tyrosine Kinase genes from the *Caenorhabditis elegans* genome. As they are paralogous genes, their phylogeny is not well known and can only be inferred from the sequences themselves. Although we will only test for differences between these two broad regions in just a few genes, the same general principles can be applied to all likelihood tests for rate variation.

Finally, we will also use this test to illustrate the differences between Maximum Likelihood and Bayesian model testing. We will do the same test comparing the same two models using each method. We will explore the difference between joint estimation and marginal estimation of both the models and the nuisance parameters, and learn how to interpret the outcome from a Metropolis Hastings MCMC.

Note: The specific models employed here are used for DNA, but once you understand how a probabilistic substitution model plus a tree confers likelihood on a DNA alignment, you are prepared to understand how similar models can be devised for amino acids, morphological characters, etc.

Lab Prep

1. Download the appropriate RAxML executable from this page http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm . The executables are about a third of the way down the page. Most Mac people probably have Macs with Intel chips (iMAC, I think). You can try the pthreads version if you like -- threading allows RAxML to use multiple processors in parallel, which can speed up your jobs -- but it is probably simpler to not use those for now. Unpack it and put it in a folder named RAxML.

2. Download MrBayes from this page http://mrbayes.csit.fsu.edu/download.php . Unpack it, run it, and put it in a folder named MrBayes.

3. Download these two files from our website and put them in the RaxML folder: PTK_Nem_phyl and Parts

4. Download this file and put it in the MrBayes folder: PTK_Nem_small.nex

The Model

Nucleotide substitution models

There are multiple models for describing the probability of one nucleotide turning into another along one branch of a phylogeny. All rely on establishing a rate of change between every pair of nucleotides; these rates can be described in a ***transition matrix***. For example under the Jukes-Cantor model every nucleotide has an equal chance of changing into every other nucleotide; such a model can be described with this matrix:

		To			
		A	C	G	T
	A	–	α	α	α
	C	α	–	α	α
From	G	α	α	–	α
	T	α	α	α	–

The transition matrix and the branch lengths can be used to calculate the probability of going from one nucleotide to another along any branch. A zero branch length will result in no changes, so that every site has a 100% chance of being in the same state that it began. On the other hand, after a branch of infinite length the probability of any site having a particular nucleotide will be that nucleotide's ***equilibrium frequency***, regardless of what nucleotide it started out as. The equilibrium frequency depends only on the transition matrix, and not on the starting state. Under the Jukes-Cantor model every base has an equilibrium frequency of 0.25, so after an infinite amount of time every site an equal chance of being any base. For intermediate branch lengths between zero and infinity the probability of going from one state to another will depend on both the starting state and the transition matrix.

Another common model is *Kimura's two-parameter* in which there is a different rate of change for transitions (among pyrimidines, C and T, or among purines, A and G) and transversions (between pyrimidines and purines).

		To			
		A	C	G	T
	A	–	β	α	β
	C	β	–	β	α
From	G	α	β	–	β
	T	β	α	β	–

Question 1. Provide a matrix for a model in which there is one rate for transversions, one rate for transitions among pyrimidines and another for transitions among purines.

We are going to use the *General Time Reversible (GTR)* model. Time reversible means that the probability of going from state *x* to state *y* when going from node *A* to node *B*, is the same as going from state *y* to state *x* when going from node *B* to node *A*. When you use a reversible model, the root of the tree does not have to be defined. All the models we have looked at so far are reversible. You might assume that the GTR model has six rates and looks like this:

		To			
		A	C	G	T
	A	–	α	β	γ
	C	α	–	δ	ε
From	G	β	δ	–	σ
	T	γ	ε	σ	–

Under this model the rate of change between any two states is the same in either direction. This model actually has five free parameters, not six, because the entire matrix must be constrained to evolve at rate 1. However, you can make this model be even more general and still be time reversible by including the equilibrium frequencies of each base (π).

		To			
		A	C	G	T
	A	–	$\pi_C\alpha$	$\pi_G\beta$	$\pi_T\gamma$
	C	$\pi_A\alpha$	–	$\pi_G\delta$	$\pi_T\varepsilon$
From	G	$\pi_A\beta$	$\pi_C\delta$	–	$\pi_T\sigma$
	T	$\pi_A\gamma$	$\pi_C\varepsilon$	$\pi_G\sigma$	–

This adds three new parameters to the model for a total of eight. It is three and not four, because all the equilibrium frequencies have to add to one. In practice many programs do not actually fit the equilibrium frequencies, but instead use the empirical frequencies, that is to say the actual frequencies of the bases in the sequence. The empirical frequencies are usually very close to the ML frequencies. This is the case for *RAxML*, but not *MrBayes*. Nevertheless, we can still consider these free parameters.

Gamma-distributed Rates

Does every site in the sequence evolve at the same rate? Probably not. This feature of gene evolution is often approximated using the gamma distribution. The idea is that there is a probability that any site will evolve at a given rate and that probability is drawn from the gamma distribution. For each site, the likelihood is calculated for every possible rate; the likelihood under each rate is multiplied by the probability of that rate under the gamma distribution; and all those likelihoods are added together to calculate the total likelihood for the site.

This distribution has two parameters, α, which controls the shape of the distribution, and β, which controls the spread of the distribution. The β parameter is held constant for all these models and only the shape of the distribution is varied. Thus this adds one parameter to our model.