

Biology 545 – Lab 3 – Partitioned Analyses

As we'll discuss in Lecture, an increasing number of studies are using different models for various data partitions. It's only been possible to search tree space with partitioned models for a few years, so it's too soon to tell if it's actually a good thing to do. Certainly, partitioned (and mixed – more later) models can fit data better than unpartitioned (i.e., single partition) models, but there is the obvious danger of over-parameterization that is usually not addressed.

In this lab, we'll use two Bayesian approaches to partitioned analyses; estimation with the partitions assigned by codon positions, and mixed-model estimation, which doesn't assign sites to partitions *a priori*.

I. Analyses with *a priori* partitions using MrBayes.

For this exercise, we'll use the same data we've been using, which codes for Cytochrome b. Since it's protein-coding, we'll use the codon-structure as partition, as we did in the parsimony weighting exercise.

We'll have to modify the file somewhat.

First, delete the assumptions block.

Replace it with the following mrbayes block.

```
begin mrbayes;
  charset 1stpos = 1-727\3;
  charset 2ndpos = 2-727\3;
  charset 3rdpos = 3-727\3;
  partition codon = 3: 1stpos, 2ndpos, 3rdpos;
  set partition = codon;
end;
```

This does essentially the same thing in MrBayes that the assumptions block you just deleted did in PAUP; it erects and invokes character partitions and assigns sites to them based on codon positions.

Now, copy this mrbayes block into the file after the one above.

```
begin mrbayes;
  lset nst=2 rates=invgamma;
  prset applyto=(all) ratepr=variable;
  mcmc ngen=500000 samplefreq=100 nchains=1;
end;
```

The first line here identifies a HKY+I+G model. The parameters will be linked across all three partitions. We could easily decide not to do so, and allow unique base frequencies, transition ratios, proportions of invariable sites, and gamma shape parameters (along with branch lengths) for each of the partitions.

The second line allows each partition to have a unique average relative rate.

The third line actually runs the mcmc analyses, which should only take a half hour or so. This is actually a pretty approximate analysis for this exercise, and we wouldn't want to publish an analysis with chains of only half a million generations.

Once the analysis is finished, load the parameter file into excel. Calculate average for each of the rate multipliers the partitions. Note that three values are constrained to have a mean of 1. Do they make sense biologically?

II. Mixture model analyses with BayesPhylogenies.

Now we're going to run a similar analysis, but instead of each site being assigned to partitions a priori, we're going to let each site have some probability of belonging to each of three classes. This is a much more complex approach, but it lets us test the assumption that we can model the molecular evolution of this data set by assigning sites to codon positions.

Again, we'll have to modify the data set somewhat. To save you the hassle of converting formats, I've already done it and you can go to the course website and copy it (just as you did in the first lab).

Here, we'll use 3 GTR models, but we'll assume equal-rates.

I've already set the file up to run automatically when you open it. This run will take a few hours, so you should let it go overnight.

```
./BayesPhylogenies filename.nex
```

Once your run finishes, you want you to open the parameter file and see if the model detects codon structure. You'll be able to tell by looking at the pattern weights; if they're all $\sim 1/3$, the patterns in the data largely follow codon structure.

III. Just hand in your analysis of rate multipliers from MrBayes and pattern weights for BayesPhylogenies.