

# **BSC3402L (6): Experimental Biology — Comparative Genomics**

**Laboratory Section: Thursdays from 4:00 to 6:00 PM.**

## **Phylogenetic Analysis 3**

**Week Seven, Thursday, March 1, 2007**

**Author and Instructor: Fredrik Ronquist**

Lab covers: (1) Bayesian model selection using Bayes factors; (2) Inferring site rates; (3) Phylogenetic inference of protein data; (4) Inferring ancestral states; and (5) Inferring ancestral changes.

Fredrik Ronquist  
School of Computational Science  
Florida State University  
Tallahassee, FL 32306-4120  
ronquist@scs.fsu.edu  
850-645-1325

## 1. Bayesian model selection using Bayes factors

As we discussed in the lecture this morning, Bayes factors provide an easy way of comparing models against each other. We will start this lab by using Bayes factors to do some model selection for the `primates.nex` data set. Specifically, we will ask: (1) whether the Jukes-Cantor model describes the evolution of this sequence better than the GTR model; (2) whether there is significant rate variation in this sequence; and (3) whether the evolution of this sequence is clock-like.

First we will compare the Jukes-Cantor model with the GTR model. Recall that the stationary state frequencies are equal in the Jukes-Cantor model whereas they are allowed to vary in the GTR model. Furthermore, all rates are equal in the Jukes-Cantor model whereas each reversible rate pair is allowed to be different in the GTR model. To compare these two models, we need to run a MCMC analysis to convergence on each one of them. Then we estimate the overall likelihood of each model using the harmonic mean of the chain likelihoods after burn in. Finally, we compute the Bayes factor, which is the ratio of the two model likelihoods, and assess the significance against a table of critical values (see the lecture this morning).

Before we start, create a directory called `phylo3` and copy `primates.nex` there. Then navigate to the new directory and start MrBayes there by typing **mb**. Finally read the data set by typing **execute primates.nex**. Now we need to set the substitution model to Jukes-Cantor. A good first step is to look at the current model. Do this by typing **showmodel**. You will get a table towards the end of the output. It should look like this:

```
Active parameters:

Parameters
-----
Statefreq      1
Topology       2
Brlens         3
-----

1 -- Parameter = Statefreq
   Prior      = Dirichlet
2 -- Parameter = Topology
   Prior      = All topologies equally probable a priori
3 -- Parameter = Brlens
   Prior      = Branch lengths are Unconstrained:Exponential(10.0)
```

There are three free parameters in this model, of which the stationary state frequencies is one. In the Jukes Cantor model, however, the stationary state frequencies should be fixed to be equal; they should not be free to vary under a Dirichlet prior. To make sure they are fixed, change the prior for the state frequencies first by typing **prset statefreqpr=fixed(equal)** and then check the model again with **showmodel**. The output should now look like this:

```
Model settings:

Datatype = DNA
Nucmodel = 4by4
Nst      = 1
Covarion = No
# States = 4
          State frequencies are fixed to be equal
Rates    = Equal
```

Active parameters:

```
Parameters
-----
Statefreq      1
Topology       2
Brlens         3
-----

1 -- Parameter = Statefreq
   Prior       = Fixed
2 -- Parameter = Topology
   Prior       = All topologies equally probable a priori
3 -- Parameter = Brlens
   Prior       = Branch lengths are Unconstrained:Exponential(10.0)
```

Note that the stationary state frequency parameter is now fixed to be equal.

Now we can run a MCMC analysis under this model. Let us first try to run the analysis for 20,000 generations by typing **mcmc ngen=20000**. You should get adequate convergence in about 1 minute, as indicated by the standard deviation of split frequencies. Answer **no** when MrBayes asks you whether you want to continue the run. To calculate the harmonic mean of the likelihood values, we need to summarize the samples using an adequate burn in. Since MrBayes by default discards 25 % of the samples when it calculates its convergence diagnostic, the standard deviation of split frequencies, we will do the same. After 20,000 generations, sampled every 100<sup>th</sup> generation, we will have 200 samples, 25 % of which is 50. Therefore type **sump burnin=50**. You should now get a lot of information about the parameter samples. Try to find a small table that looks like this:

```
Estimated marginal likelihoods for runs sampled in files
"primates.nex.run1.p" and "primates.nex.run2.p":
(Use the harmonic mean for Bayes factor comparisons of models)
```

Run	Arithmetic mean	Harmonic mean
1	-6431.57	-6442.38
2	-6431.53	-6442.30
TOTAL	-6431.55	-6442.34

This table contains the values we need. In my case, I got the overall log model likelihood estimate of -6442.34; your value should be close to this. Write this value down for future reference.

Now we need to change the model to a GTR model. First we need to allow all pair-wise substitution rates to be different by typing **lset nst=6**. Then we must allow stationary state frequencies to be different using **prset statefreqpr=dirichlet(1,1,1,1)** (this setting puts a flat Dirichlet distribution prior on the stationary state frequencies, that is, we consider all possible stationary state frequency values equally likely prior to analysis). When you type **showmodel** after having made these changes, your output should look like this:

```
Model settings:

Datatype = DNA
Nucmodel = 4by4
```

```

Nst      = 6
          Substitution rates, expressed as proportions
          of the rate sum, have a Dirichlet prior
          (1.00,1.00,1.00,1.00,1.00,1.00)
Covarion = No
# States = 4
          State frequencies have a Dirichlet prior
          (1.00,1.00,1.00,1.00)
Rates    = Equal

```

Active parameters:

```

Parameters
-----
Revmat      1
Statefreq   2
Topology     3
Brlens      4
-----

1 -- Parameter = Revmat
   Prior      = Dirichlet(1.00,1.00,1.00,1.00,1.00,1.00)
2 -- Parameter = Statefreq
   Prior      = Dirichlet
3 -- Parameter = Topology
   Prior      = All topologies equally probable a priori
4 -- Parameter = Brlens
   Prior      = Branch lengths are Unconstrained:Exponential(10.0)

```

Note that we now have a `revmat` parameter in addition to the `statefreq` parameter. The former contains the six pair-wise rates of nucleotide substitutions. To compute the model likelihood of the GTR model, we use the same strategy as above. First run 20,000 MCMC generations using **mcmc** (we do not have to specify the number of generations again because this setting is persistent and remains in effect from the last time we changed it). Answer yes every time MrBayes asks you whether you want to replace an output file (this will overwrite the previous results but the harmonic mean that we wrote down above is the only thing we need from these runs). You should not have any problem getting convergence within 20,000 generations so simply answer **no** when MrBayes asks you whether you want to continue the chain. Finally obtain the harmonic mean under the GTR model by typing **sump burnin=50**. The relevant table should now appear something like this:

```

Estimated marginal likelihoods for runs sampled in files
"primates.nex.run1.p" and "primates.nex.run2.p":
(Use the harmonic mean for Bayes factor comparisons of models)

Run   Arithmetic mean   Harmonic mean
-----
1     -5945.04           -5954.70
2     -5945.05           -5955.40
-----
TOTAL -5945.04             -5955.11
-----

```

The log of the harmonic mean of the model likelihood is now estimated to be -5955.11 (you should get something similar). This is almost 500 log likelihood units better than the previous value (-6442.34). In other words, the GTR model is strongly preferred (recall from this morning's lecture that a difference of 5 log likelihood units or more is considered strong evidence in favor of the better model).

So far, we have assumed that the rates are the same at all sites in the molecule. We can continue our Bayes factor comparisons by invoking the gamma model of rate variation across sites. First write down the log model likelihood of the GTR model under equal rates. Then assume gamma rates using **lset rates=gamma** and repeat your analysis. Is there significant rate variation in the evolution of the sequences? Choose the best assumption about rates (if necessary, go back to equal rates using `lset rates=equal`, but only if that is the preferred model). Finally, test the clock model by issuing **prset brlenspr=clock:uniform**. This associates the branch lengths in the tree with the simplest form of clock tree prior, the clock uniform prior, instead of allowing branch lengths to vary freely to yield a non-clock tree. Use exactly the same procedure as before and compare the clock model against the standard non-clock model. If the non-clock model is significantly better, the conclusion is that there is significant rate variation across lineages in our data set. Do you observe this?

## 2. Inferring site rates

Say that you are interested in examining the distribution of rates over sites in the `primates.nex` data set. In MrBayes, you can control the parameters that are reported during a MCMC run with the `report` command. Type **help report** to get a list of the available options. To infer site rates, you need to type **report siterates=yes**. Now, start a new MCMC run and stop it when you have a good sample from the posterior. As before, set 25 % of the samples as the burn-in when using **sump**. Your site rates will be presented in the parameter table at the end of the `sump` output. If you are interested, you can try to plot these rates as a diagram using the KSpread and KChart applications. Are there any regions with particularly slow or high rates? Do you see the expected rate differences among codon positions (first, second, third)?

## 3. Phylogenies for protein data

Inferring phylogenies from protein data is very similar to inferring phylogenies from nucleotide data except that the models you can use for protein evolution are different. Perhaps the most useful option in MrBayes is the possibility of integrating over a set of fixed rate matrices for protein data. To test this type of analysis, we need a protein dataset, which is provided for you on the course site on Blackboard as **avian\_ovomuroids.nex**. Copy this data set into your working folder, `phy103`. The data file "avian\_ovomuroids.nex" is set up to do a mixed-model protein analysis because it contains a MrBayes block at the end with the single command **prset aamodelpr=mixed**. Check that this is true by examining the nexus file in a text editor. Then read the file in MrBayes by starting the program again if you have stopped it (typing **mb**) and then executing the protein dataset by typing **execute avian\_ovomuroids.nex**. Perform an analysis of this dataset using the same approach you used previously for the `primates.nex` dataset. If you have difficulties getting a good sample of the posterior distribution within a reasonable amount of time, quit the analysis prematurely and continue with the next task.

## 4. Inferring ancestral states

For the next exercise, where we will be inferring ancestral states, we will add one binary character to the "primates.nex dataset". This character will be used for the geographic distribution of the species in the matrix. Remember to keep a copy of the original dataset if you want to go back to the unaltered file later. For

instance, you can do this in the terminal window by typing **cp primates.nex primates2.nex**, which will create a copy of the file `primates.nex` called `primates2.nex`. Enter the new character by opening the copy of the datafile (`primates2.nex`) and entering a character before the first nucleotide. Type '0' for the primates that are African and a '1' for the ones that are not. You also need to tell MrBayes that the data are now mixed. Do that by modifying the format line to "**format datatype=mixed(standard:1,DNA:2-);**". This tells MrBayes that the first character is a 'standard' character with state labels from 0 to 9, and that the remaining characters (from 2 to the end, the end being represented by a dot) are DNA characters. Also add one to the number of characters on the **dimensions** line. Now start MrBayes (**mb**) and execute the data file (**execute primates2.nex**). MrBayes will automatically divide the data into two partitions, one for the geographic character and one for the DNA. You can see this clearly if you type **showmodel** and examine the output. You might want to assume that the two partitions evolve under different rates; do that by typing **prset ratepr=variable**. In the default prior, the rates are assumed to be the same for all partitions (`prset ratepr=equal`) so we need to change the prior to variable to allow the rates to be different among partitions.

To infer ancestral states in MrBayes, you need to set up a constraint first for the group that you want to have the ancestral states for. The strength of the Bayesian MCMC analysis is that it integrates over phylogenies but if you want the states for a particular ancestor you want to sample only those trees that have that ancestor. Go back to your `primates.nex` dataset and look at your best tree. Choose one group for which you want to infer ancestral states and define constraints for that group using the command "constraints". Type **help constraints** to get information about how you do this. For instance, if you want to infer states for the ancestor of humans and chimps, set up this constraint by typing **constraint A -1 = Homo\_sapiens Pan** and then **prset topologypr=constraints(A)**. We also need to tell MrBayes to report the ancestral states. If we're smart, we will tell the program to report the ancestral states for only the first data partition (the geographic character) and not for the DNA partition. Do that by typing **report applyto=(1) ancstates=yes**. Finally, you need to set the DNA model to your favorite one by using 'prset' and 'lset', if necessary. To make sure that the changes only apply to the DNA partition (partition 2), use "**prset applyto=(2) ...**" and "**lset applyto=(2) ...**". Use **showmodel** to check your settings when you are done.

When these preparations are done you can run your analysis the same way as you have done previously. The **sump** command will report the probabilities of ancestral states, in our case the probabilities of the common ancestor of humans and chimps being African and non-African, respectively.

## 5. Inferring ancestral character changes (and states) with parsimony

If you are interested in looking at character changes in ancestors, I recommend that you use PAUP and parsimony. PAUP does not recognize mixed datasets so we will have to do this exercise with the original "primates.nex" file or some other file with only a single kind of data in it. Start PAUP by typing **paup** and then execute your data file. Then perform your favorite type of analysis and choose the settings such that it results in only one or a few trees. Now we can bring up that tree or those trees by typing **showtrees** as we have done previously. To get the changes for the ancestral branches, use the `describetrees` command, which gives you a number of options for printing information about the tree. For the ancestral changes, type

**describetrees / apolist=yes.** Apolist is an abbreviation of apomorphy list; apomorphy is used to denote character changes mapped to branches in a phylogenetic tree. Note the numbering system used to associate the changes with branches in the tree. If you are interested in ancestral states, type **describetrees / mprsets=yes** instead, and you will get the most parsimonious states (the most parsimonious reconstruction sets, or mpr sets) for all ancestors.

## **6. Assignment**

For your assignment, find an interesting question about a topic of your choice and describe how you can address that question using tools we have covered in the course. Maximum one page.