**Introduction to Bayesian phylogenetics and model testing**

November 12th, 2019

All the programs and data files we will need today can be found in the Github folder for Day3, as well as flash drive in class. If you need the files, please ask.

**MrBayes Tutorial**

We will be working with MrBayes, which is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. MrBayes uses a Markov Chain Monte Carlo approach to estimate and sample the posterior distribution of model parameters. RAxML and other ML methods only produce a single tree but using MCMC approaches we will generate a posterior distribution of trees which can help incorporate phylogenetic uncertainty. In the end we will generate a consensus tree with support values at the nodes. More information about the program and manuals can be found here: <http://mrbayes.sourceforge.net/>

Create a folder to run the MrBayes analysis. Make sure to include the mb executable file and the **nexus** file of interest. To start off we will use a smaller subset of data in the “**ruhf\_32\_by\_5000.nex”** file containing 32 species and 5,000 bp of data. Also select one of the 79 coding genes that you have previously used.

Opening the terminal. If you downloaded MrBayes via Homebrew, just type “**mb**”, if you downloaded the program through GitHub, you will need to navigate to the folder with the MB executable, and launch MrBayes by typing:

**./mb**

At the MrBayes prompt, launch nexus file with. We will walk through the software with a small data set, but it will be up to you to run it with your chosen gene of interest:

**execute ruhf\_32\_by\_5000.nex**

To check what models are available for you to specify type in:

**help lset**

Nst is the number of substitution parameters. A value of 1 suggest that all possible nucleotide transitions have a single rate (JC). A value of 2 suggests one rate for transitions and a second for transversions (HKY). What model is specified with 6 substitution values? \_\_\_\_\_\_\_\_\_

If you look at the rates category, you will also see options such as equal, gamma, propinv, and invgamma. If we want just a gamma distribution but no specification for invariant sites, what would we choose?

For now, we are just going to stick with a GTR model of molecular evolution with a gamma distribution, to set the model type:

**lset nst=6 rates=gamma**

Before do any analyses we want to check the priors. To see how those set type:

**help prset**

Default settings for the substitution rates for the GTR model (Revmatpr) and state frequencies (Statefreqpr) are set as a flat Dirichlet model (values=1.0). This is appropriate if we want to estimate these parameters from the data assuming no prior knowledge about their values. However, we can fix those rates, such as if we were running a JC or SYM model that require nucleotide frequencies to be equal (this command would be preset statefreqpr=fixed(equal)). Default settings can be specified as prset statefreqpr=Dirichlet(1,1,1,1).

To look at all priors and settings, type:

**showmodel**

How many parameters are being estimated during our analysis? \_\_\_\_\_\_\_

For an initial analysis we do a mcmc run with 50,000 generations sampling every 100 generations. With this sampling, how many trees we will have in the posterior distribution? \_\_\_

**mcmc ngen=50000 samplefreq=100**

On screen you will see the status of the run. The way it is set, we are running 2 separate runs each form a different random starting tree with 4 chains each. The cold chain is in square brackets [] while the three heated changes are in parentheses (). The cold chain will always have the smallest likelihood score (best score). Ideally you will want to run this until the deviation of split frequencies is less than 0.01. After your initial runs are done, the program will ask you if you want to continue. What was the split frequency? \_\_\_\_\_\_

Before summarizing any results, we will need to perform a burnin. The burnin removes the earliest samples prior to thorough exploration of tree space. The most common approach is to remove the first 25% of the generations. Summarize all estimated parameters after a 25% burnin, remember you will have to tell the program how many generations constitute 25% of the total posterior distribution. If you only ran 50,000 generations sampling every 100, you have a total of 500 trees to choose from. If you ran more generations to reach convergence, how many trees do you? \_\_\_\_\_\_\_\_\_\_\_\_

**sump burnin=125**

This produces a table that summarizes the parameters. ESS values (Effective Sample Size; or the number of effectively independent draws from the posterior distribution) of 200 are thought to be well sampled, while values of 10,000 are considered a waste of computational resources. Based on the table, did your analysis sample the parameters well? \_\_\_\_\_\_\_

Summarize the trees from posterior distribution using the same burnin as for the sump command:

**sumt burnin=125**

This produces a cladogram in text view format, as well as a file that can be opened in Figtree (nex.con.tre) with support values given at the nodes. A well supported node in a Bayesian framework is usually 0.95. Play around with Figtree to place the support values on the tree, as well as selecting the appropriate root.

**Tracer**

Let’s explore the test data a bit more in Tracer. Open the program and drag & drop your files ending in run1.p and run2.p to the boxes under the words “Trace Files”. If you select one at a time you can look at the ESS values and distribution of all estimated parameters. If you highlight both at the same time, we can compare the two different chains that were ran together to see how close they were to each other. If you click the Trace button along the top, do you see a “fuzzy caterpillar” in each parameter? Are you ESS values large enough to have confidence in?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**BEAST tutorial**

We are going to do a walkthrough of how to set up, run, and visualize a BEAST analysis. BEAST is a Bayesian platform for phylogenetic analysis of molecular sequences. However, BEAST goes beyond many other programs by estimating rooted, time-calibrated phylogenies, as well was a framework for testing evolutionary hypotheses conditioning on a single tree topology. The learning curve for BEAST can be quite steep and frustrating. We will be limited in what we can do today, but I am always happy to chat more if you have additional questions (jbl256@cornell.edu). More examples of modules/tutorials can be found here: <http://www.beast2.org/tutorials/index.html>

Before we can run Beast, we will need to make a configuration file using the standalone program BEAUTi. Start this program by double clicking on the executable file in the programs folder. We will use the 32-tip nexus file.

Click File -> Import alignment and select ruhf\_32\_by\_5000.nex. To look at the file to make sure it was uploaded correctly, double click on the row in the table.

We will skip over the **Tip** tab. This setting allows you to specify dates for fast evolving viruses or sub-fossil ancient DNA material, which we do not have in our current analysis. In most cases, this option is not necessary**.**

In the **Site Model** tab you can set the model of molecular evolution for the sequence data. In previous analyses we have used the GTR+G model, so we will do that here as well. Make sure to click estimate for substation rate and type 4 for the Gamma Category count. Click the GTR model and make sure “Frequencies” are set to empirical (meaning these will be calculated from the alignment).

The next tab is the **Clock** tab where we can set model for the molecular clock. As opposed to RAxML and MrBayes, Beast uses a molecular clock so that the produced trees have a timescale. The simplest model is “Strict clock”, which assumes that all branches on the tree have the same rate of evolution. Given our sampling of plants, this is not likely true, however for now we will leave this set as strict.

Next tab is the **Priors**. The tree prior is first, coalescent models are better suited for populations while the others are better suited for species level data. We will select the Yule model which is the simplest speciation model where each lineage is assumed to have speciated at a fixed rate. The only parameter for this is birth rate, which we will use a uniform prior. The rest of the priors are from the parameters we told the program to estimate. We will use a gamma prior for transition rates. Unlike other tree building programs we do not need to specify an outgroup in Beast. Part of the analysis is to find the most probable root of the tree, but if we wanted to specify “In group” samples, we would do that in the priors tab by selecting the “+ Add Prior” button, you will do this a bit later on, but for the first run don’t use this.

The last tab is **MCMC**, where we set the number of generations, burnin, and initializations attempts. For now, set chain length to 5,000,000. The default here is 10 million, which should normally be sufficient for small data sets to reach appropriate ESS values. Hopefully with this setting this run will take less than 10 minutes per million generations. Larger data sets, such as the single gene data sets, may need 50-100 (or more) million generations to reach appropriate convergence.

Now we are ready to save the XML file. Click file -> Save. At any time you can reopen the XML file in BEAUTi by clicking open and select the file.

We are now ready to start the analysis. Open the Beast program by double clicking the icon. Select the XML file you just saved. Click Run. You will see the progress of the analysis. This data set should run in a matter of minutes.

At the end of the run BEAST creates a table listing each operator, how many times it was used, how much time they took, and other details. This may be useful for optimization of runs, but for the moment we will skip this.

We cannot look at each individual tree produced in the BEAST run, but we can look at the overall distribution of trees to see if our run needs more generations. Will look at the output from the BEAST run using the program Tracer. Click File -> Import Trace File. Click on the Trace tab at the top. For the posterior, we should see a “fuzzy caterpillar” and not a directional change in the trace. This suggests that the run fully mixed. We can also look at the ESS values on the left, with values above 200 suggestive of a converged run. Keep in mind that the authors of BEAST suggest that ESS values should not be used for categorical/discrete parameters, nor for posterior, likelihood or prior density samples. However, low ESS values for these other cases may reflect underlying problems. If our ESS values and traces look good, we can move on to building the MCC tree.

To create a consensus tree summarizing the posterior distribution of trees, we will use TreeAnnotator. Double click the icon. Set the burnin percentage to 0.25, and the target tree type to Maximum Clade Credibility. Your input file is the trees file created from the Beast run. If you want to give a name to the output file, do so now. Then click Run. This will create a tree file we can open in FigTree. Node support values from the posterior probability can be found under node labels -> display -> probability.

**Node calibration**

If we wanted to use fossil calibration points for our tree to get an accurate estimate of age, we would first need to set a taxonomic group by clicking “+ Add Prior” and selecting the species/tips we wanted in the group. We would also need to enforce monophyly, so select the check box. Clicking on the black triangle next to the prior brings up the mean and standard deviation to fill in with our best estimates of the age of the group.

We need to fix the node for the angiosperms. This includes the following taxa: *Chloranthus spicatus, Trochodendron aralioides, Drimys garanadensis, Piper cenocladum, Calycanthus floridus, Acorus americanus, Kingia australis, Certophyllum demersum, Illicium oligandrum, Nymphaea alba,* and *Amborella trichopoda*. In the Priors tab create a new prior. Give it a name. Include the 11 taxa above, and click monophyletic. To start off we will use a Normal distribution. The constrain age for the angiosperms will be 130 MY with a confidence interval spanning 110-150 MY.