**Bioinformatics Lab 7**

**RAxML, MrBayes, and BEAST2**

**Week 8**

**Goal**

To estimate fast ML and Bayesian phylogenies from large datasets using concatenated alignments. Alternatively, to estimate a Bayesian phylogeny from a smaller dataset using a partitioned mitochondrial alignment.

**Links to relevant tutorials / data**

* The RAxML-NG ReadMe:

<https://github.com/amkozlov/raxml-ng/blob/master/README.md>

* BEAST2 tutorial: <https://taming-the-beast.org/tutorials/Introduction-to-BEAST2/>
* The MrBayes website: <https://nbisweden.github.io/MrBayes/download.html> and GitHub page: <https://github.com/NBISweden/MrBayes>

**Operating system**

Mac, Linux, or PC (although RAxML-NG may not work well on a PC). BEAST is a standalone program that should work well on any system, but is more complex to set up an analysis.

**Background**

This lab involves installing and running **RAxML-NG, MrBayes, or BEAST2**.

RAxML-NG is a phylogeny inference program that uses maximum-likelihood (ML) as an optimality criterion. It is one of the fastest ML/Bayesian (versus distance or parsimony) methods out there, making it the preferred choice for very large datasets. It can be used for tree estimation, bootstrapping to assess support for a tree topology, evaluating the likelihood of alternative models of substitution, etc.

MrBayes is a program for Bayesian inference, so uses posterior distributions of model parameters as its optimality criteria. MrBayes provides a flexible framework that can conduct all sorts of analyses, but we will only use it for tree estimation.

BEAST2 is also a program for Bayesian tree inference, especially for partitioned alignments. It provides lots of flexibility in assignment of priors, especially for divergence time data (e.g. using molecular clock assumptions), but as such it is useful only for smaller datasets of just a handful of loci.

I am providing three msa datasets to choose from. The primate.phy is 13,472 base pairs from 12 species of primates and should run quickly on MrBayes or RAxML. The primate-mtDNA.nex file is 898 base pairs of a three mitochondrial genes.

**Steps**

1. ***Preparing the input files***
2. Download the data from Canvas (under Phylogenetics Module). These are alignments in Nexus format (.nex) or Phylip format (.phy), which are similar to the fasta format that we’ve been working with.
3. Choose phylip format for RAxML-NG, Nexus format for MrBayes, and the mtDNA alignment for BEAST2. I suggest using a different folder for each phylogenetics software that you’re running.
4. [Optional: upload your alignment to the Cipres Science Gateway (<https://www.phylo.org/>) and run it on the jModelTest to estimate the best model of sequence evolution to use for your alignment. You will need to make an account, but it’s free. Then upload your msa to a new data folder, and run an analysis, selecting jModelTest as the tool to use.]
5. ***RAxML-NG***
6. If you haven’t already, download RAxML-NG from <https://github.com/amkozlov/raxml-ng>. Unzip and move the unzipped folder containing the binary to an appropriate location. You can then call the binary in terminal using ./raxml-ng from within the containing folder, by specifying the full path from wherever you are (e.g., “/Applications/raxml-ng\_v1.1.0\_macos\_x86\_64/raxml-ng”), or by adding the location to your PATH.
7. Perform a “single tree inference on DNA alignment” (estimate a tree) using RAxML-NG, following the instructions in the Readme.

*Note: If using the Epinecrophylla alignment, this will take at least several hours to run and uses quite a bit of RAM. I recommend testing it quickly in class and then starting it sometime when you aren’t going to need your computer (e.g., at night before you go to bed) so you can leave it running. If the default single tree inference analysis is impractical on your machine, either find a different computer/cluster to run it on, or reduce the number of tree searches performed. By default, RAxML-NG conducts searches on 20 starting trees (10 random, 10 parsimony). You can reduce this using the –tree flag as follows:*

*--tree pars{N},rand{M}*

*Where N and M are the number of searches to run on parsimony and random trees, respectively. Reduce these from the default 10 as needed (although keep in mind you may end up with a poor tree if you do too few searches).*

1. View your tree! It can be viewed in the free GUI program FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>), as well as in R using various R packages. Note that FigTree requires a jdk (java development kit): <https://www.oracle.com/java/technologies/downloads/>. *Note: Your tree will look nicer once you re-root it on the outgroup. Find the lemur sample (or the branch leading to the lemur and tarsier) and click “Reroot” in FigTree.*
2. [Optional: If you completed step 5 and have time, you may want to complete a bootstrap analysis to evaluate support for your best tree (particularly if you are using your own dataset). To do this, try using the instructions in the Readme (steps 2-4). Alternatively, rather than re-estimate the tree, you can use the existing tree from step 6 using the –support flag.]
3. ***MrBayes***
4. If you haven’t already, install MrBayes. A pre-compiled version is available for Windows on the Installation page of the MrBayes website. If you are on a Mac, the instructions at <https://github.com/NBISweden/MrBayes> worked for me. Here is the MrBayes manual: <https://github.com/NBISweden/MrBayes/blob/develop/doc/manual/Manual_MrBayes_v3.2.pdf>
5. Navigate to your working folder (the one containing the Nexus alignment output in step 4). Start MrBayes. The default install location of MrBayes is generally “/usr/local/bin/”, so try typing:

/usr/local/bin/mb

1. Load the data:

execute primate.nexus

1. Set the evolutionary model to GTR with gamma-distributed rate variation and a proportion of invariable sites:

lset nst=6 rates=invgamma

1. Start the MCMC:

mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=1000

1. After the 20,000 generations specified above, you will be prompted to decide if you want to continue the analysis. Typically, you would want to continue it if the value of “standard deviation of split frequencies” printed to the screen is above 0.01. We will probably be able to reach that benchmark in a reasonable amount of time with this dataset, but if not, it is okay to stop the analysis after 20,000 generations for the purposes of this assignment.
2. Once the analysis is stopped, summarize the posterior distribution of parameter values including substitution model parameters and ESS (effective sample size) values. You typically want average ESS values >100 for a publishable analysis.

sump

1. [Optional: View traces of the posterior parameter distributions (these should end in “.p”) using the program Tracer (<http://tree.bio.ed.ac.uk/software/tracer/>). We will use this program more in the next week or two with the BEAST analysis, but could be fun to get a sneak preview with this first Bayesian analysis.]
2. Summarize the posterior distribution of trees (this is printed to a file ending in :

sumt

1. The file produced by the above step includes posterior probability for each node in the tree. As with the RAxML tree, it can be viewed in FigTree or R (see above). It may still be quite unresolved after a short MCMC!
2. ***BEAST2***
3. If you haven’t already, install BEAST2: <https://www.beast2.org/> This should be as easy as clicking on the appropriate file for your operating system and adding it to your Applications folder.
4. Follow the instructions in the Introduction to BEAST2 tutorial using the primate-mtDNA.nex alignment file: <https://taming-the-beast.org/tutorials/Introduction-to-BEAST2/>.

**Products**

* + One ML or one Bayesian tree from concatenated alignments, or one Bayesian tree from partitioned alignments, each with 12 tips representing some primate species, including an outgroup. The Bayesian trees might be unresolved (too short an analysis), but the ML tree should be decent.
  + Export a .pdf / .png / .jpg file of your tree to your Bioinformatics folder and push your data to GitHub.
  + Write a short Methods and Results text based on the program that you used (RAxML-NG, MrBayes, or BEAST2). Outline 1) what you did, 2) what you found, and 3) why you think it is interesting/relevant. This should be three paragraphs (one for each of the three topics):

1. In order to make the phylogenetic tree, I used the BEAST2 program. I imported the MSA into the BEAUti software to set the parameters for the phylogenetic analysis. This includes the model, which, in this case, a model called HYK was used. Clock model can be used to set an evolutionary estimated time for the tree, and priors can be used to set the model parameters. After setting all the parameters, the file was made into an XML file, which was run by BEAST in order to make a tree file. Then the output from BEAST can be utilized in TreeAnnotator, which sets the parameters for the tree (not the model used to obtain the data but the way the tree looks etc.). Then the output from TreeAnnotator can be opened in FigTree, which allows the actual tree to be changed (node labels/bars, tip length). Then it can be exported into a pdf file.

2. BEAST2 is a very heavily customizable program, which is amazing but probably comes with a steep learning curve. It is understandable why those working with phylogeny would use a program like BEAST2. It is computationally heavy as well, but it gives very useful results.

3. Comparing with my classmates that used RAxML, the trees look slightly different. This is because of the different phylogenetic models that were used. Even the confidence intervals look different because of this as well. This is why the program and model that is used is important when it comes to phylogenetic analysis.