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

I. Preparing the input files

1. 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.

2. 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.

3. [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.]

II. RAxML-NG

4. 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.

5. 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).

6. 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.*

7. [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.]

III. MrBayes

8. 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

9. 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
```

10. Load the data:

```
execute primate.nexus
```

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

```
lset nst=6 rates=invgamma
```

12. Start the MCMC:

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

13. 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.

14. 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
    ```

15. [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.]

16. Summarize the posterior distribution of trees (this is printed to a file ending in :

    ```
    sumt
    ```

17. 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!

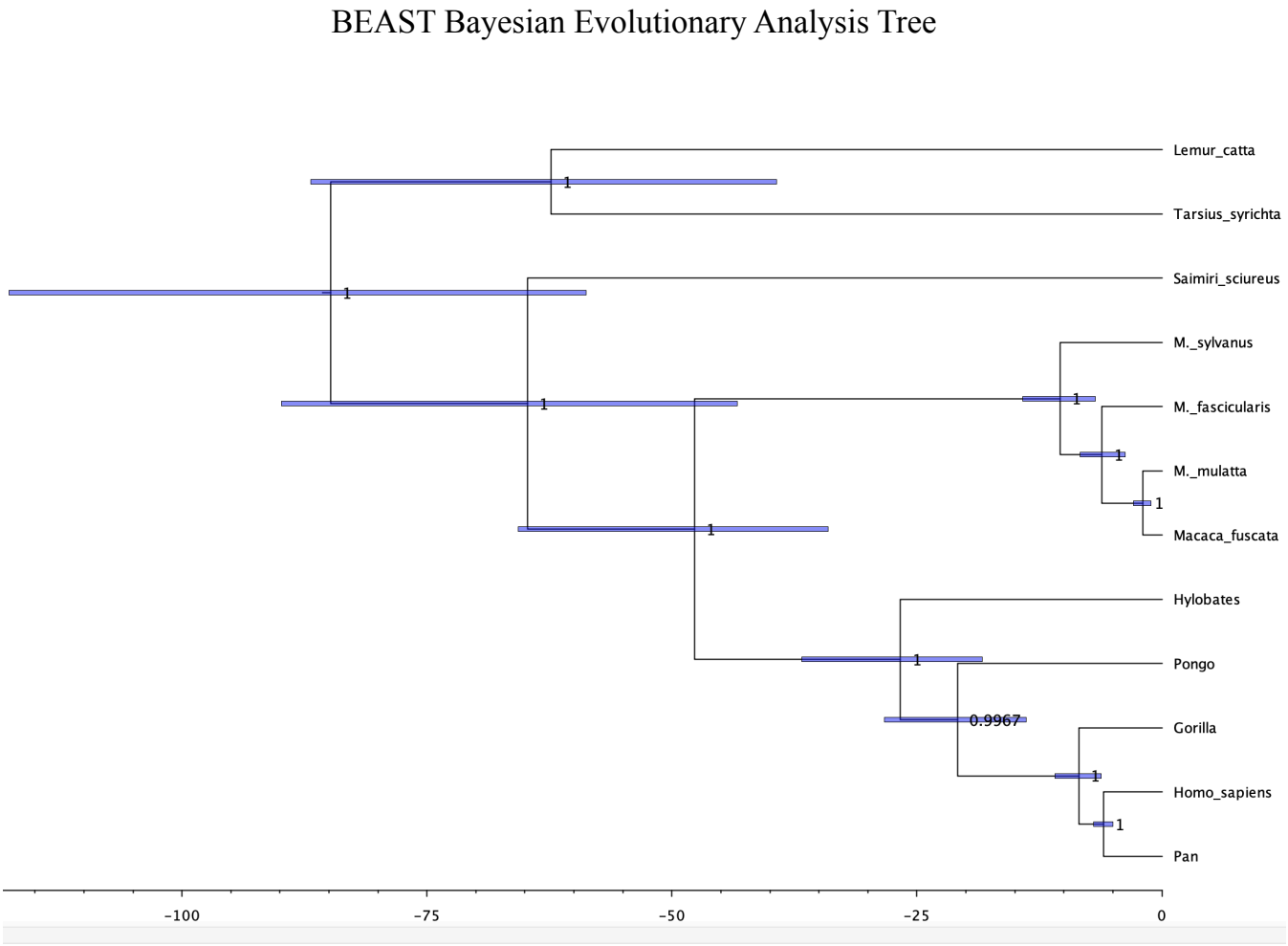
IV. BEAST2

18. 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.

19. 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. Given the primate-mtDNA.nex file, which consisted of 898 base pairs of three mitochondrial genes that were sequence aligned with the metadata of 12 primate mitochondrial genomes. First, I had to set up all the components of the BEAST2 analysis in BEAUti2 by preparing a configuration file in XML format. Even though it's possible to create such files from scratch in an editor it's not easy, therefore I used the user-friendly program (BEAUti) designed to produce the valid configuration for BEAST. Some aspects that were prepared in BEAUti includes but is not limited to: linking all four data partitions (noncoding, 1st, 2nd, and 3rd codon positions) to a singular evolutionary branch rate-distribution, setting t=0 (present day samples), setting up a standard clock model, using the simple Calibrated Yule Model as the tree prior along with other prior parameters, separating mean tree height from substitution rate parameters, adding a calibration node using prior fossil knowledge, setting the MCMC options to control the length of the chain (1,000,000) and the frequency of stored samples (200), and setting the file name to finally generate the XML file. With the configured file, I ran it in BEAST2 and with the data that was received I placed it into the Tracer application and the TreeAnnotator for data analysis. With the file created from the tree summarizing program TreeAnnotator, I uploaded the maximum clade credibility tree file into FigTree to be able to finally visualize the sought-after phylogenic tree.

2. The two programs that had represented the data in their own specific way includes; Tracer and FigTree. Tracer essentially showed the summary of the BEAST2 run of the 1,000,000 MCMC chain length primate data. More specifically, it provides summary statistics of log files that includes: the posterior, prior, the likelihood, tree likelihoods and other continuous parameters. It is also possible to analyze the data and specify a burn-in in order to have the Markov Chain time to reach its equilibrium distribution which was actually done in TreeAnnotator (10%). Basically, To summarize the use of Tracer it's the culmination of the data that has been received from the MCMC chain ran in BEAST and the tree product that is visualized in FigTree is the most likely posterior distribution estimate. Tracer does also allow you to compare different mutation rates corresponding to the different partitions in the alignment where the 3rd codon position had an exponentially higher mutation rate than the other groups(rest were similar). As for the phylogenetic tree, the maximum clade credibility tree type found the tree with the highest product of the posterior probability of all its nodes. The visualized tree in FigTree contains a reverse axis (time), node bars representing 95% highest posterior density (HPD), node labels to give the posterior probability and of course the evolutionary relationship between taxonomy tips. The node labels indicated that the posterior probability of nodes in the posterior set of trees was very high with values close to or equal to one. The 95% HPD intervals give a valid estimation of the divergence of species with respect to a given node, for example the ape clade first separated between 20-40 million years ago. Humans are most closely related to chimpanzees having split about 5 million years ago. The divergence between old-world and new world monkeys has a wide time range of between 40-90 million years ago.

3. This phylogenetic tree is really interesting and informative because it's essentially the representation of 1,000,000 datasets as one all inclusive and easy to interpret chart. If you really

analyze the information present in the tree you can find interesting information about the divergence of specific species and their presumed relationship with each other; all from a small subset of the entire genome (mitochondrial DNA). The data also shows the most probable times that specific claves may have diverged and if you were to acquire more data you could potentially piece together relationships between more distally related species.