**Maximum-likelihood**

November 7th, 2019

All the programs and data files we will need today are on Github. Please pull that, I also have files on a flash drive if needed

**Model selection with jModeltest**

For most phylogenetics analyses, selecting the appropriate model of molecular evolution is the first step. Today we are going to use one of the most frequently used programs, jModeltest, to do this. You will need to select one of the 79 coding genes used previously from the Ruhfel et al. paper, and for model testing you should use the fasta format. To start the program, double click the “jModelTest.jar” file. The GUI shows up, which we will use.

**Click File -> Load DNA Alignment**

You will navigate to the folder where your data file resides and select the appropriate file. Once the file loads you should see several lines of output with the file you selected, the number of sequences, and number of sites.

**Click Analysis -> Compute Likelihood scores**

Here we have a lot of options. The default values are 11 substitution schemes, resulting in 88 models total. To save some time, change this value to 3. This will still test 12 models, the most common and easy to specify in MrBayes. We will also incorporate a gamma distribution with 4 categories, but make sure to uncheck the “invariant sites”. The last thing to change for now is the tree for likelihood calculations. The default is ML optimized, which is the best, but for timing, select “BIONJ”. If everything looks good to you, hit:

**“Compute Likelihoods”**

You can watch the progress and the current models being tested. After all models are tested:

**Analysis -> Do AIC calculation**

**Analysis -> Do BIC calculation**

We are doing both to see if they specify the exact same model.

**Results -> Show results table**

On the tabs along the top, click AIC, and sort the table based on AIC score with the smallest on top. Do the same with BIC. Does AIC and BIC select the same model? If not, which one is more complex?

What are the delta AIC scores for some common models?

GTR+G

HKY+G

HKY

JC

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

Now that we have our best fit model, we will start inferring the tree. Depending on what your best model was, RAxML may not support it. Do not fret, we will use that model for our Bayesian inference next week.

Before running anything in RAxML, call the help menu to see what the options are

If using RAxML-NG thru Git

**./raxml-ng -h**

If using RAxML after downloading via Conda

**raxmlHPC -h**

From here on out I’ll use the standard RAxML calls (raxmlHPC). If you want to use RAxML-NG look at the help menu and the will be slightly different

The basic execution of an alignment has the following options

**raxmlHPC -s alignment file -n prefix -m substitutionModel -p RandomSeedforParsimony**

For the Ruhf\_32\_by\_5000 data set, to get a basic tree you would do the following:

**raxmlHPC -s ruhf\_32\_by\_5000.phy -n 32\_initial -m GTRGAMMA -p $RANDOM**

This will produce several files. If you look at the RAxML\_info file, you will see the specifics of your data set and the run you just did. If you add “.tre” to the end of the RAxML\_bestTree file, you will see what your best tree looks like.

We did not specify an outgroup in the previous run. We will rerun the analysis with Pedinomonas\_minor.

**raxmlHPC -s ruhf\_32\_by\_5000.phy -m GTRGAMMA -o Pedinomonas\_minor -p $RANDOM -n 32\_outgroup**

Does using an outgroup (or a different outgroup) change your overall topology?

We are getting a topology, but we don’t know how well-supported it is. Let’s run 100 bootstrap replicates (for many studies you would try to run 1000) to see which clades are well-supported. For the sake of time, we will use rapid bootstrapping (-x) which run much faster than regular bootstrap (-b)

**raxmlHPC -f a -s ruhf\_32\_by\_5000.phy -m GTRGAMMA -o Pedinomonas\_minor -p $RANDOM -x $RANDOM -N 100 -n 32\_bootstrap**

After performing the bootstraps, RAxML start the search to find the ML tree. The RAxML\_bestTree still shows the ML tree, but the RAxML\_bipartitions will show the bootstrap support when opening in FigTree. In FigTree, make sure to show “Node labels” and select the appropriate option based on what you called the support values when you opened FigTree.

Now that you have a parsimony tree and a ML tree for the 32-tip data set, let’s compare them to see what looks different. Open up the “Comparative\_phylogenetics.R” script in R and set your working directory. We will us the first block of code to compare trees. Make sure your parsimony and ML trees are in the same directory. Run the first comparison. From the R output, how many nodes does each tree have? How many clades are unique in the parsimony tree vs unique in the ML tree? Are the clades that are different between the two have high bootstrap support?

\*in the script: pdf("Parsiomony\_vs\_ML.pdf",20,20), the numbers after PDF specify the width and height of the PDF file. For your single gene, you may need to make the height value larger so you can clearly see the tree and where the differences occur.

Comparative phylogenetics

Now you have a tree, or in this case several trees, can we start to ask some evolutionary based questions? In the R script there are two different types of analyses you can run with your given ML tree: 1) ancestral state reconstruction with continuous data or 2) ancestral state reconstruction with discrete data. There is a third option, state dependent diversification, but we will save that for the Bayesian module on Tuesday. In most cases we would already have some phenotypic data for our species of interest, but for now we will simulate the data. Pick either the continuous or discrete method here and run on the 32-tip tree. From your plotted tree, how is your trait evolving?

We have made our way through all the different pieces today: model testing, ML inference, tree comparison, and some basic comparative methods, all on the small data set. You will need to go through all these steps with the same gene that you selected for the parsimony analyses on Tuesday. After our Bayesian lab next Tuesday there will be homework questions involved with comparing the same gene between the different types of analyses. I’ll have that assignment ready for next class, but as you are working through these exercises, make sure to save the following information:

Using AIC, what is the preferred model of molecular evolution: \_\_\_\_\_

If it is not GTR+G, what is the deltaic score between your model and GTR+GAMMA: \_\_\_\_\_\_

Using 100 bootstrap replicates for your gene, are most of the nodes towards the MRCA (most recent common ancestor) well supported?

Comparing the parsimony and ML tree, how many nodes are shared between the two trees? Are the nodes that are unique to one data set or the other well supported (bootstrap support) in the ML tree?