Day 2: Maximum Likelihood and Bayesian Approaches to Phylogenetics

Jesus Martinez-Gomez & Eugenio Valderrama

October 20th, 2020

Abstract

In today's worksheet we will analyze the same data using a Maximum Likelihood and Bayesian approaches using the programs RAxML and MrBayes. We will do these two ways, the first will be on the Linux servers you have access too. The second way will be using the amazing resource called CIPRES Science Gateway.

Table of Contents

Model Based Phylogenetics on the Linux server

Remote Login

First we will need to log into the server. Refer to Day1 worksheet on how to do this. Remebasher your user name is your netid, the server is the IP address of the computer and your password is your first name, capital first letter plus 6410 (except for Han, let me know if you forgot).

As a short reminder:

- Windows user will need to open Putty or Mobaxterm
- Mac user can simply open the terminal app and ssh

Update project folder with git

First everyone needs to change directory to the project file. When you first login to the server you will be put in the "username" folder. This is analogous to when you turn on your computer you go to your Desktop. A quick tip: typing 'cd' without any argument will take you to the "username" folder immediately; once there 'cd' to the class folder.

```
cd
cd PLBIO6401_EcologyEvolution_Module_2020
```

The next thing we are going to do is update our class folder. Eugenio and I added new material for Day2 as well as made a slight adjustment to the Day 1 lecture. We do this using git. Access this page with your *browser by clicking here*

```
git pull https://github.com/Jesusthebotanist/PLBIO6401_EcologyEvolution_Module_2020.git
ls
```

Now that we have updated our file, we'll find a new folder 'Day2_MLBayes'. This folder contains

1. A subfolder 'markdownFiles' - similar to Day1 this is the file I used to generate this document, if you are interested.
2. A copy of this pdf.

Lets make two directories for our Likelihood and Bayesian analysis.

```
mkdir Day2_MLBayes/ML
mkdir Day2_MLBayes/MLbootstrap
mkdir Day2_MLBayes/Bayes
```

Maximum Likelihood with RAxML

Introduction and Set up

RAxML is one of the most commonly used Maximum likelihood inference programs. While it is the program we will use, the past recent years have seen a nubasher of new Likelihood programs including IQ-Tree, GARLi, FAST ML, and others. Certain programs are more common in certain fields. For example IQ-Tree seem more common in the virus research and, I believe, is the inference method used to generate the *COVID19 phylogeny here*. If you ever use phylogenetics in your own research, make sure you look into these various programs to find the one suitable for you.

We will used the aligned fasta file we generated in Day 1. Our first step is to make a copy of this file into both our ML folder and our MLbootstrap folder. Then 'cd' into ML.

```
cp Day1_LinuxAlignmentParsimony/My_first_Parsimony_Analysis/ruhf_32_by_5000_aligned.fas \
Day2_MLBayes/ML

cp Day1_LinuxAlignmentParsimony/My_first_Parsimony_Analysis/ruhf_32_by_5000_aligned.fas \
Day2_MLBayes/MLbootstrap

cd Day2_MLBayes/ML/
```

Again we have RAxML pre-installed, to run it simply type 'raxmlHPC'. Unlike TNT which opens up a scripter where you type commands, many Linux programs have 'arguments' denoted by a dash '-' which can be be include in the terminal line when you run the program. The dash '-' is sometimes called a flag. The argument are different for different programs but one commonly shared one is the -h flag which pull up a help menu. Let try this out with RAxML

```
raxmlHPC -h
```

As you can see RAxML has quite a nubasher of different options. Often times there is a pdf manual which is easier to read these options. You can find the *RAxML manual here*. We do not need to specify all argument to run a program. In our analysis we will specify the following.

- -s The input sequence file name
- -m The substitution model
- -o The outgroups
- -p A random parsimony seed, this is for reproducibility purposes
- -n The output file name

We are essentially using the RAxML equivalent of the 'default' parameters. Importantly our substitution model is the GTR+gamma (i.e., GTRGAMMA), do you recall from lecture what this means? You may have notices at various points we've used backslash "", Linux interprets these as a new line. We use this

to make the code more human readable. You should be able to copy and paste the following.

```
raxmlHPC \
-s ruhf_32_by_5000_aligned.fas \
-m GTRGAMMA \
-o Pedinomonas_minor \
-p $RANDOM \
-n ruhf_32_by_5000.phy
```

This might take a bit of time to run. Once it is finished type 'ls', you'll notice a nubasher of output files were generated. Always read the manual in order to understand the output of a program. In this case we'll give you that information, the phylogeny will be in the file, 'RAxML_bestTree.ruhf_32_by_5000.phy'. Download this using Cyberduck and open it using FigTree. Do you see a phylogeny? Is Pedinomonas_minor the outgroup?

Running RAxML with Bootstrap

As we discussed in lecture an important part of phylogenetics is to gain a sense of how confident we are in our inference. In other words, how strongly support are certain topologies. Bootstrap is a common statistical method to assess support for many types of analysis including parsimony and likelihood phylogenetic approaches. In the following section we run the same analysis as above but this time using bootstrap. To do this we add a few extra arguments:

- -f this specifies the type of bootstrap approach. We will use 'a' which is rapid Bootstrap
- -x Random seed for rapid Bootstrap
- -N the number of bootstraps

```
cd ..
cd MLbootstrap

raxmlHPC \
-f a \
-s ruhf_32_by_5000_aligned.fas \
-m GTRGAMMA \
-o Pedinomonas_minor \
-p $RANDOM \
-N 100 \
-n ruhf_32_by_5000.phy \
-x $RANDOM
```

This will take a ~100x longer as it is running the analysis 100 times over, sampling with replacement. Again, use 'ls' to see the output files. The file with the phylogeny and bootstrap values is, "RAxML_bipartitions.ruhf_32_by_5000.phy".

download this with Cyberduck and view in Figtree. It will prompt you to name the information on the phylogeney, name it "Bootstrap". Click the drop down arrow in 'Node Labels' on the left hand tool bar. Then click on the drop down arrow for 'Display:' you should see 'Bootstrap' click on that. The Bootstrap value should appear on the nodes!

MrBayes

Now we will perform a Bayesian analysis. Unlike likelihood where you have to run a bootstrap analysis, the Bayesian approach has a natural way of accounting for confidence in what is called the posterior. In short, instead of estimating a single phylogeny, a Bayesian approach will infer an entire set of phylogenies with slightly different topologies and branch lengths. These represent the *posterior distribution* and they represent how certain we are in our analysis.

Set up

Similar to RAxML we will first copy our alignment over. Unfortunately, MrBayes does not read .fasta file it can only read .Nexus file format. There are a nubasher of ways to convert .fasta to .Nexu, our preferred method is to use Aliview. This is a program used to visualize alignments but can convert alignments into different formates, very useful!

1. Download **Aliview**
2. Using Cyberduck download 'ruhf_32_by_5000_aligned.fas' locally. You should have two copies in two different folde, they are exactly the same so pick either.
3. Open 'ruhf_32_by_5000_aligned.fas' Aliview. You can click around to check out the alignment you made!
4. Go to "File" then "Save As Nexus (Illegal character replaced by _ (e.g. for MrBayes)", this will automatically change the prefix to .nexus instead of .fas.
5. Open this file in your text editor and compare it to the .fas, they both contain the same information just different formating. A homework question below ask about these difference, make some notes.
6. Upload the to the Day2_MLBayes/Bayes folder.

Phylogenetic Inference with MrBayes

MrBayes is more similar to TNT than RAxML, in the sense that when you run MrBayes it'll open up it's own scripter which you type commands into. As opposed to specifying argument using flags like in RAxML.

First 'cd' to the Bayes directory, where you've uploaded the .nexus file.Notice I use '../..' this means go up two prior folder instead of one. You can include '..' in a relative path.

```
cd ../..
bash
```

Read in the data

```
execute ruhf_32_by_5000_aligned.nexus
```

Unlike the -h flag in RAxML here we use type the following to learn about about the specific programs.

```
help lset
```

We will specify a GTR+gamma model, the same model we used for RAxML. To do this we specify 'nst' which is the nubasher of substitution parameters. 'nset' = 6 corresponds to GTR which has six parameters. 'rates' is the used to specify gamma.

```
lset nst=6 rates=gamma
```

Similar to maximum likelihood we will specify an outgroup.

```
outgroup Pedinomonas_minor
```

The advent of a Bayesian analysis is you can specify priors. Priors are typically represented as distributions (e.g., normal, exponential, gamma etc...) which parameter values are drawn from. These distribution represent additional information not included in the matrix (i.e., alignment). In phylogenetic models people typically use the default priors. Prior come more into play when it comes to comparative methods. We will use default priors. To view priors type the following.

```
help prset
```

One last check is we can show the entire model using the following.

```
showmodel
```

We are ready to run the analysis. Remebasher Bayesian analysis us a MCMC to estimate the posterior distribution. We need to specify how long to run the MCMC (i.e., how many generation) and how often to save a value (i.e., samplefreq). While, ideally we'd want to save every sample, if these analysis run for a long time they can take up a ton of computer space! Lets run the MCMC for 50000 generation and save every 100 generations, it shouldn't take super long.

```
mcmc ngen=50000 samplefreq=100
```

At the end of the analysis Mrbayes automatically detect if the MCMC analysis

has ran for long enough, for this dataset 50000 generation is not enough but we will cut it short. It'll ask if you want to continue, for our purpose type no.

MCMC Processing We have generated a posterior distribution of all the model parameter and of the phylogeny, which is also a model parameter, but MrBayes saves them in two differen files. The next step is to remove the so called, 'burn in'. The first couple MCMC generations in an MCMC analysis contain parameter cobashination with a low posterior value. As such, we typically remove 25% of the MCMC generation as a rule of thubash. MCMC are tricky and more of an art than a science. The general rule is to run as many independent chains for as long as you can to get an accurate estimate of the posterior distribution. In practice researcher typically run 3 chains (we <3 the nubasher 3) for as long as it take for these chains to reach convergence, more on this later.

In Mrbayes rather than specifying a % of samples to remove, you need to tell it an exact nubasher, so we have to do a bit of mental math. We ran an analysis for 50000 generation but we only save every 100 samples, so our posterior distribution contains 5000 samples (50000/100). We want to remove 25% of 500 which is 125 samples (500 x .25). First we remove this from our parameter file.

```
sump burnin=125
```

This will generate some information including a table. Look at the 'avg ESS' column. Effective Sample Size (ESS) is a measure of whether a particular parameter converged or not. Typically ESS>200 is indicative of a parameter value converging. We will look at this more in depth later in Tracer.

Second we will remove the burnin form our posterior distribution of phylogenies; they are saved in a separate file from the rest of the parameters.

```
sumt burnin=125
```

Similar to TNT, this will generate a cartoon diagram of the phylogeny in your window. Lets quit bash for now.

```
quit
```

You'll notice a nubasher of new files have been generated. Again, it is important to alway read a program manual to know exactly what you are looking at. We'll be visualizing them in the next step.

Visualizing MCMC Trace and Phylogeny Assessing MCMC convergence is tricky and some programs, like Mrbayes have an internal method of doing so. However most people do analysis convergence in a seperate program called Tracer or using the R package coda. We will use Tracer.

1. Download **Tracer**. For Mac user download the .dmg, for Windows download the .zip.

2. Using Cyberduck download the 'ruhf_32_by_5000_aligned.nexus.con.tre' and the two .p files, 'ruhf_32_by_5000_aligned.nexus.run1.p' and 'ruhf_32_by_5000_aligned.nexus.run2.p'

In Tracer open both '.p' files by dragging them into the left most panel. Once loaded you'll see information pop up on the right hand panel. This is a representation of the MCMC estimate for the model parameters. Click on, "Trace" in the upper section of the right hand panel. Notice how in the first few generations the trace increase. This is the portion of the MCMC we want to burnin aka remove, because they are sampled from low posterior probability space. Look at option of the parameters on the left hand panel click on r(A<->C). This represent the parameter estimate for the transition probability from adenine to cytosine. Now click, "Marginal Density" on upper section of the rigth hand panel, these show density plots of these parameter estimates. For example, my density plot ranges has a medium around .11 that means the rate from A to C and C to A is ~.11. The 95% confidence interval is how certain I am in my estimate. For me this looks fairly wide. However, because we didn't let this converge, my results will differ from yours. As such, be weary of any biological interpretation on analysis that have not converged. Again, an indicator of convergence is all the ESS>200.

Lastly, lets open, 'ruhf_32_by_5000_aligned.nexus.con.tre' up our phylogeny in Fig Tree.

Model Based Phylogenetics on CIPRES

Alright you all have learned how to use Linux to run these programs. Linux give you the ultimately flexibility in running a program how you want. However, there are a few downsides:

- You'll probably need to learn a lot more Linux. This isn't a bad thing and luckily at Cornell there is a free, extremely helpful workshop ran by the Cornell BioHPC exactly for this called **Linux for Biologist**, we highly recommend it.
- You need to have access to server. Some labs have them, and sometime they have individuals who manage (e.g., install programs). But if they lack a manager you might have to be that person!
- If you join a lab that doesn't have a server but you still want to do things with Linux we recommend you look into XSEDE. XSEDE is a NSF funded computational resource. Essentially, NSF pays Universities with super computers to give them computing resources, that NSF in turn give to researchers with XSEDE accounts. XSEDE is free but you have to submit a proposal to get an computing hours. Get in touch with *Susan Mehringer* here at Cornell to learn more.

Luckily, there is an amazing resource for phylogenetics call CIPRES. CIPRES is a website that allows you to run phylogenetic and alignment programs on a super computer for FREE and its easy! CIPRES is part of XSEDE, but unlike XSEDE which you access through Linux, CIPRES has its own GUI. There are a few downsides, not all programs are on it (sorry TNT) and you do have a limited, but generous, nubasher of computing hours, but you can ask for more (so I"m told). Practically speaking, if any of you do a phylogenetic analysis we encourage you start with CIPRES as it will likely server your purpose. We'll be repeating our likelihood analysis and Bayesian analysis on CIPRES.

CIPRES

Setting up an running an analysis is fairly straightforward but you need to make an account if you don't already have one.

1. Make an account on CIPRES.
2. Log in (you may not have too)

RAxML on CIPRES

Uploading Data

First step is to upload data to CIPRES .

1. Download a local copied of the **aligned** ruhf_32_by_5000_aligned.fas using Cyberduck
2. On CIPRES go to "Create New Folder"
3. Under "Label" give this project a name, "Day2" and click "Save"
4. On the left hand side your folder should appear with two subfolder, "Data" and "Tasks". Click on "Data" folder
5. Click "Upload/Enter Data"
6. Upload your file by clicking the "Browse" button. You could also upload you file manually if you wish. To do the latter you'll need to open your file in a text editor copy and paste into the website.
7. Click "Save"

Running RAxML on CIPRES

1. Click on "Tasks" folder
2. Click on the "Create new task" page click "Select Data", it should give you an option to click on the aligned file you just uploaded. Then click the green button, "Select Data"

3. It'll automatically transfer you to the "Select Tool" sheet, (if not click on it). Here you can select the different programs available. There are a couple version of RAxML, we'll be using "RAxML-HPC v.8 on XSEDE"
4. Click "Set Parameters", here we will specify all the parameters we had to type in using flags. Notice, there is a short explanation of what the parameter is followed by a flag "-", this is the corresponding argument in RAxML. We'll be running the exact same bootstrap analysis we ran on our servers, below are the instruction for specifying the flags. **Note** Make sure to click on "Advance Parameter" to show more.

- In CIPRES you need to specify the computing time our analysis is small so we will use the default ".25" (*Note:* You will need to change this value in the homework)
- -n change to ruhf_32_by_5000_CIPRES.phy
- -o Type in "Pedinomonas_minor"
- -p Make sure this box is clicked
- -m Under "Nucleic Acid Option" header make click the GTRGAMMA bubble
- -f a under the "Configure the Analysis" header click the drop down arrow so that it reads, "Rapid bootstrap analysis / search for best-scoring ML tree"
- -N again under this header, click the box next to "Specify the nubasher alternative runs on distinct starting tree?". Change 10 to 100.
- -x under "Configure Bootstrapping", click the bubble next to "Rapid Bootstraping"

5. Click "Save Parameters" at the bottom, a window will pop up click "Okay"
6. Go back to "Task Summary" (if you weren't automatically transfered here). Write a "Description" for job, "ruhf_32_by_5000_CIPRES".
7. Click, "Save and Run Task"

You've submitted your job! Depending on how busy the server is this might take a little bit of time. However, small job are prioritized and will jump to the beginning of he queue. You'll receive an email when the job is finished. If this is taking to long start on the MrBayes section below.

Analysis Output

1. When your job is done, navigate back to "Tasks" on CIPRES. You'll see a list of the task you submitted (only one).
2. For that task you should see an option "View Status" or if you waited for a while it may say, "View Output"
3. You will be taken to a page with details of your job and see the "Output" option. Next to it click on "View (2)".
4. This is a list of all the RAxML output in addition to some files that the Linux server generated with information on the run. Click "Select All" and

the "Download Selected" at the bottom. This will down load a compressed (.zip) file.

5. Open "RAxML_bipartitions.ruhf_32_by_5000_CIPRES.phy" in FigTree and compare it to "RAxML_bipartitions.ruhf_32_by_5000.phy". Is the topology the same? Is the bootstrap support the same? Why? These will be homework questions!

MrBayes on CIPRES

MrBayes is similar to RAxML where we can use the GUI to input our parameter. However we will take a slightly different approach. Instead of manually clicking all the boxes, we can include our parameters in the .Nexus file. We'll have to manually edit the file and add what is called a "Mrbayes block". This will contain all the information CIPRES need to run MrBayes and it save us a few steps.

1. Open "ruhf_32_by_5000_aligned.nexus" in a text editor. You should have a local copy because this is what you generated with Aliview
2. Scroll down to the end of the file and in a new line type the following

```
BEGIN MRBAYES;
set autoclose=yes nowarn=yes;
lset nst=6 rates=gamma;
outgroup Pedinomonas_minor;
mcmc ngen=50000 samplefreq=100;
sumt burnin=125;
sump burnin=125;
END;
```

These are essentially all the instruction you typed in the MrBayes tutorial above with a few important notice:

- We do not need the 'execute' command as we independently upload the input file and CIPRES knows where it is.
- The first command "set auto close..." is new and necessary to avoid warnings warning.

3. Save this .nexus file and change the name to, "ruhf_32_by_5000_aligned_MrBayesReady"
4. Upload the file to CIPRES and start a new task. See RAxML instruction for this
5. Once you've reached the "Select Tool" section select, "MrBayes on XSEDE".
6. In "Set Parameters", you'll see that similar to RAxML there are a ton of option to choose from, but since we have included a MrBayes block we only need to check a few.
   - The one that reads, "My Data Contains a MrBayes data Block...". This tells CIPRES we have already specified all the information
   - The one that read "Run BEAGLE", this will make the analysis run faster!

7. Submit the job

Analysis Output

Once your job is finished download the entire output folder. Look at the model parameters in Tracer and compare it to you MrBayes run.

Homework

Answer the following questions. Send your answer to the questions in a word document or text file. For question 2 and question 3, choose a gene from the "Ruhfel_unaligned_fasta" folder and send the phylogenies (four phylogenies total), please use the same gene. We recommend you use the same gene from homework 1 it should already be aligned.

Important note:

- In question 2 and 3, you will need to specify a different species as an out group, than the one in the examples above "Pedinomonas_minor", to root the phylogeny. Pick any species in your dataset as the root, does not matter. You'll need to open the file in a text editor find a species in your dataset.
- Some of the genes in the folder are quite large and may take some time to run on the server. In order for the run to complete you will need to have a stable Internet connection and your computer cannot got to sleep. This is very inconvenient and there are way around this which we have not taught you, so we recommend you choose on of the following genes: petN, petL, psal, psbl, psbM, PsbT, rpl32, petG, rpl36. However, if you do run into this issue or are an adventure shoot us an email and we can teach you how to do it.

Questions:

1. Describe what a bootstrap analysis is. Why does it take longer to run a bootstrap analysis?
2. For question 2 and 3, what species did you use to root your phylogeny? If you are unsure read the Important note above.
3. Infer a RAxML phylogeny with 2 bootstrap replicates on the server and 100 bootstrap replicate CIPRES. Specify 168 hours instead of .5, (see the section, "Running RAxML on CIPRES" for a refresher on how to do this.) Compare the two phylogenies in Fig Tree. Is the topology the same, explain why or why not? Is the bootstrap support the same, explain why or why not?
4. Infer a MrBayes Phylogeny with 500 MCMC generations on the server and on 50000 Specify 168 hours instead of .5, (see the section, "Running

RAxML on CIPRES" for a refresher on how to do this.) Compare the two phylogenies in Fig Tree. Is the topology the same, explain why or why not? Is the posterior probability the same, explain why or why not?

About This Document

This document was written in markdown and knitted into a PDF with Pandoc, using the following code.

```
pandoc Day2_MLBayes_oct.md \
-f gfm \
-t latex \
--toc \
-V toc-title:"Table of Contents" \
-V linkcolor:blue \
-M title="Day 2: Maximum Likelihood and Bayesian Approaches \
to Phylogenetics" \
-M author="Jesus Martinez-Gomez & Eugenio Valderrama" \
-M date="October 20th, 2020" \
-M abstract="In today's worksheet we will analyze the \
same data using a Maximum Likelihood and Bayesian approaches \
using the programs RAxML and MrBayes. We will do these two \
ways, the first will be on the Linux servers you have \
access too. The second way will be using the amazing \
resource called CIPRES Science Gateway." \
--extract-media ./images \
--highlight-style tango \
-o Day2_MLBayes_Oct2020.pdf
```