**A fungus that attacks a fungus of fungus-growing ants: a story in three phylogenies**

For this assignment, you can work together and use any resources (including any software) that you choose. I provide details on how to perform the work using certain tools, but it is up to you. This assignment, along with the first exercise on alignment, is due October 7th. Please put everything in a single word file and email it to me.

**Background**

Data from:

Augustin et al. 2013. Yet More ‘‘Weeds’’ in the Garden: Fungal Novelties from Nests of Leaf-Cutting Ants. Plos One 8(12):e82265.

*Escovopsis* spp. are fungi which attack and consume fungi reared by fungus-growing ants. Until recently, while it was known that Escovopsis spp. were morphologically and ecologically diverse, with different morphotypes having different host ranges, few species had been formally named and described. As part of this work, the *Escovopsis* genus has been split into *Escovopsis* and *Escovopsioides.* See Augustin et al. (2013) for more information.

As part of a collaborative effort, colleagues and members of my lab have sequenced the genome of an ‘*Escovopsis’* strain isolated from a Atta cephalotes fungus-growing ant colony. The strain was isolated and sequenced prior to new Escovopsis and Escovopsioides species being described. We now need to know what we should call our ‘Escovopsis’ strain.

I have downloaded the sequences\* used to generate the phylogeny in figure 9 in Augustin et al. (2013). I have also included the ITS DNA sequence for the genome sequenced strain (though, you will notice it is currently a bit short). Your goal is to generate a phylogeny to determine what genus the genome sequenced strain most likely belongs to and whether the genome sequenced strain is closely related to any named *Escovopsis* or *Escovopsioides* species.

\* Random Helpful Hint: If you ever have a list of NCBI sequences you want to download, you can do it as a batch download here: <http://www.ncbi.nlm.nih.gov/sites/batchentrez>.

**Before We Get Started**

You or someone that you are working with may want to download:

Mega

Jmodeltest

Mr. Bayes

FigTree

Web addresses for downloading sites for these programs can be found on the phylogenetics resource page that I handed out in class.

Much of this assignment is using MEGA, which is available via the web for both Macs and PCs. Unfortunately, the Mac version is a bit unstable, but it works fine. For the Mac version, I recommend trying to avoid minimizing the window, which is a hard habit to break. **Save often in either platform.**

You do not need to use MEGA or the suggested programs. There are many options out there, and you are welcome to use programs that you are either familiar with or that you have been meaning to learn anyway.

This assignment is a fairly simple data set where the analyses will run quickly and you are not trying to keep track of a 100s of sequences and many decisions on settings for later publication. For more complex analyses, ALWAYS keep good notes, preferably in a lab notebook.

**Step 1: Assemble Sequences**

I have done this for you by downloading sequences from genbank. I have placed all the unaligned sequences in a file (AllEscovopsis\_ITS.text) so that you can see them. This file is in the fasta format. You should notice the sequence names follow each “>”. Sequence names should be kept simple, with no spaces or no alphanumeric characters. Any spaces you see are actually “\_”. Some programs will have a problem if you start the sequence name with a number, so best to avoid that. Always start your sequence name with a good identifier because some programs (e.g., Phylip) can truncate sequence names over a certain length. Here, I have kept track of the genbank accession number in the first string of characters. You can find these numbers in the Augustin et al. phylogeny.

**Step 2: Align Sequences**

With this fasta file, there are several alignment options. You can use ClustalX (as you did in the last exercise), clustal omega online, or muscle online (see list of phylogenetic resources) or you can import your sequences into a comprehensive program like MEGA and do the alignment through MEGA. If this was a protein coding gene, you could use translatorX, which will use the amino acid coding information to inform the alignment of the DNA sequence. This is not a protein coding gene.

To align your sequences in MEGA, Go to Align/Edit or Build Alignment. Then Data/open/retrieve sequences from file, you can then open your fasta file. If you do not see it, change the file extension to .fas (from .txt)

Once in the Alignment Explorer, Select muscle (the muscle arm icon). Align by DNA. Follow the instructions. The alignment is pretty fast for this few sequences. Look at the data.

Based on your previous exercise, adjust gap penalities to see if it strengthens your confidence in the alignment. Make sure to write down any details if you used anything other than the defaults, because you will need to include this information when you write your “methods” section.

There are large gaps on the ends of many sequences because some sequences are much shorter than others. Save this session as something like escovopsis\_align1.mas.

We will do a bit of sequence trimming since the sequences vary in length.

To do this, from the start of the alignment, go to the right. Find where you have information from more than two sequences. Go to Edit/select sites. In the gray bar above your first sequence, select the first site you want to delete. Go to the very beginning of the sequence file. Press shift then select the gray box for the first column. You should have highlighted an entire box. Click the X in the toolbar to delete this sequence block. Save this alignment session as Escovopsis\_ITS\_trim.mas. Go to Data/export data and export your file as a .meg file of the same name. Close the alignment viewer.

Note, an online program like Clustal Omega will let you export in the clustal format. You can then import this file, instead of the unaligned fasta file, into a program like MEGA. To do this**,** open MEGA. Go to File/convert file format to Mega. Select your output file from the alignment program and the format (e.g., clustal). Select import. Save this file as a .meg file. Go to file/Exit editor. Now, go to File/open a file or session. Open your .meg file. Open the data explorer (it says T A - - ). Look at your data. You can’t edit the data here, but you can check to look to make sure all the sequences are there, etc.. You should also be able to convert to other files formats here, though there are some line break issues with this in some platforms at least (see below when we get to model testing).

**Step 3: A Quick Look at the Data**

In the main screen of Mega, open your .meg aligned and trimmed sequences file.

From here, it’s always good to get a quick look at the data. To do this, construct a quick neighbor-joining (NJ) tree. For now, use the default settings. Go to Phylogeny/construct neighbor joining tree. In the yellow box that says tests of phylogeny, select bootstrap. Do 100 bootstrap replicates. A tree viewer will appear with two trees. The first tree is the original tree from the NJ search. The numbers on the branches represent bootstrap proportions. The second tree is a consensus tree with bootstrap proportions. Play with the buttons on the left. Here, you can select the root, swap around branches, etc..

This original look is important because, for example, you may see that some taxon is on a super long branch. Is this because it did not align properly or is this biologically meaningful? You would need to go back to check your alignment to find out. For large data sets, this is the easiest way to confirm you did not accidentally include the same sequence twice or you did not forget a taxon of interest. You should carefully check your data before proceeding.

**Step 4: Parsimony Tree**

We will compare a parsimony tree, a likelihood tree, and a Bayesian tree. In general, for most datasets that you work with, parsimony will take a long time but this is small.

Go to Phylogeny/ construct and test maximum parsimony tree(s). Everything yellow can be changed. Do 100 bootstrap replicates. Under gaps/missing, select partial deletion and make the coverage cutoff 75%. This will allow us to use some of the sites where we have data from some but not all of the taxa. Run the analysis.

The most parsimonius tree will appear in the tree viewer with bootstrap support values on the branches. Save this tree as a newick file (file/export current tree/newick).

**Step 5: Making a Tree Pretty**

MEGA can do some basic manipulation of your tree, but FigTree, a freely available program, has a lot of advantages when preparing trees for publication. Open your newick file in a text editor and look at it. This is a common format to save tree information. It looks something like this:

((A,B),C)

This means that A and B are a clade, and C is outside of that clade. Within your newick format, there is also additional information about bootstrap values for each branch.

Now, open this file up in FigTree. When it asks you about the values you are importing, call them bootstrap values. Under branch labels, make sure to select bootstrap (or whatever you called them) in the drop down menu. The best thing to do now is play around with it. Under tip labels, for example, you can change the font size. Change the root so that the non-Escovopsis and Escovopsiodes sequences are outgroups. Make your tree is legible and highlight the genome sequenced strain placement some way. To change taxa labels, highlight the name, and then select annotate. You can change the name to the genus and species. Make sure these are italicized in your final tree.

This tree graphic can be saved as a PDF (to be put into your homework, for example) or as a .eps file to edit in Adobe illustrator. If saving as a .eps file under options, save the text as a font rather than as a graphic. This will make it much easier to edit the taxa names later.

**Step 6: Model of Evolution**

For Maximum Likelihood and Bayesian analyses, you will need a model of evolution.

We will compare models of evolution in Jmodeltest, which can compare 56 models. MEGA can also conduct such analyses, but it supports fewer models. It’s also important to learn how to move between different tools.

We need our alignment in a nexus file before using Jmodeltest. To do this, we’d like to export our data from MEGA as a nexus file format. This is where one problem with phylogenetics hits us, and we might as well face it now. While MEGA has a tool to export as a nexus file format, at least on a Mac, this file has line breaks that prevent Jmodeltest from importing the sequences. Here is one way to get around this problem.

1. Go to Align and open a previous alignment session. Open your .mas file.
2. Export your data as a fasta file.
3. Open your fasta file in a text editor and copy it.
4. Go to <http://sequenceconversion.bugaco.com/converter/biology/sequences/index.php>
5. Select fasta to nexus. Paste in your fasta formatted sequences, run, copy your nexus format into a text file and save.

Look at the nexus file format. It is elegantly simple. At the top, you indicate how long the sequence block is and how many taxa you have. Then, you have the data matrix.

Now, open Jmodeltest. Load you nexus file. Run the analysis. Determine what is the most appropriate model of evolution under the AICc criteria. For more information, see the Jmodeltest manual.

**Step 7: Likelihood**

For a dataset this size, this likelihood analysis can be run in MEGA. Under phylogeny, select maximum likelihood. Select a model of evolution (if the model of evolution you determined in jmodeltest is not there, then use the Mega tool to find a model that will be available in Mega). Do 500 bootstrap reps. It’s pretty fast. As above, same the tree and make it pretty in figtree.

For a larger dataset, I would strongly recommend using GARLI, which is much faster computationally than many alternatives. For GARLI, you will run multiple searches to determine if you get the same topology each time. These searches can be run without bootstrap searches to determine support values. Then, you can run the bootstrap analysis. The GARLI manual recommends a number of parameters that can be altered to make the bootstrap analysis go faster. READ IT!

**Step 8: Bayesian**

For a Bayesian analysis, we will use Mr. Bayes. Mr. Bayes, like jmodeltest, uses a nexus file format. To potentially make your life easier in the future, we are going to amend a Mr. Bayes block to the end of the nexus file that you already have. This will contain all the information that Mr. Bayes needs to run. You will then execute this file.

Paste the following below the data matrix in your nexus file.

BEGIN MRBAYES;

set autoclose=yes nowarn=yes;

lset nst=6 rates=gamma;

mcmc ngen=1000000 samplefreq=500 printfreq=500 nchains=4 savebrlens=yes burninfrac=0.25;

sump burninfrac=0.25;

sumt burninfrac=0.25 conformat=simple;

END;

Let’s talk through what this means.

* The first line prevents it from asking you questions when you aren’t standing there.
* Lset is setting the model of evolution. In the above block, we are using a GTR+G model. The GTR model has 6 different substitution rates (nst) and gamma allows for rate variation across sites. For more information on defining models in Mr. Bayes, see the Mr. Bayes manual. You will need to change this line for your model of evolution.
* The next line means that we will run 2 separate runs (the default in Mr. Bayes) for 1 million generations each. Every 500 generations we will store that tree and print it to a file. Each run will consist of 4 interacting chains searching through tree space. We want to save the branch lengths on the trees so that we can construct a tree with branch lengths later. When we samples parameters, we will ignore the first 25% of trees as burnin.
* When the search is complete, we will generate two files. The first will summarize the parameters on all the trees after burnin (sump). You should check to see if this burn-in is appropriate (see below)
* The second file will summarize the trees after burnin (sumt).

To run Mr. Bayes on a Mac, open up the terminal (under applications/utilities) and type mb. If this does not work, try /usr/local/bin/mb or sudo /usr/local/bin/mb.

To run Mr. Bayes on a PC, I believe you should have an executable that you can double click on in the Mr. Bayes folder. There should not be a need to work through the terminal.

In either platform, to execute your file, type:

Execute yourpath/filename.nex

This might look something like

Execute Desktop/Escovopsis/Escovopsis\_ITS.nex

Let the program run. This should not take no more than an hour, and you can do other things. However, if your computer is old or you have other things to do, email me and I will send you the files. I just want you to go through the process of getting it set up and running.

The consensus tree generated from all the save trees after burnin is saved as a newick file called something like: Escovopsis\_ITS.nex.con.tre. Use figtree to make this tree pretty. Make sure to show the posterior probabilities on the branches rather than the branch lengths.

***More advanced Mr. Bayes, making sure you have reached stationarity.*** Remember for a Mr. Bayes search that we start at some random place in tree space. Thus, the first trees are not a particularly good estimate of evolution within these lineages. Above, we removed these first bad trees from the analysis by rather arbitrarily setting the burnin as the first 25% of trees. This is often more than sufficient, but you can check this more accurately. Read the Mr. Bayes manual for more information and consider using Tracer, http://tree.bio.ed.ac.uk/software/tracer/.

**A Few Notes on More Advanced Analyses**

Here, we had a small number of taxa and sequence data from only a single gene. There are many benefits of including data from multiple genes. However, you should never assume that the same model of evolution applies to all your data. I would recommend determining the model of evolution for each gene’s data separately and then partitioning your data by gene. Both GARLI (likelihood) and Mr. Bayes (Bayesian) can handle data partitions.

**Please Turn in the Following…**

1) Write a Methods and Results Section for your work. Include three figures (one for each tree). Make sure you include a caption for each figure, following conventions in published work on what information should be included in the caption. Make sure to include details of what programs you used, what model of evolution you used, how many generations you ran for your Bayesian analysis, etc.

2) In looking at your sequence alignment, how confident are you in each of your characters being homologous? Is this a problem? If so, how would you suggest someone using these data for a paper proceed?