**Week 8**

**Tree Building Session**

In this lab we will be re-BLASTing the sequences from last week, using the Ribosomal Database Project to create a sequence alignment, and using a virtual machine to run bioinformatics software that will build and visualize phylogenetic trees from that alignment.

**Getting into the Virtual Machine:**

Login to the computer with kerberos

click on user icon in upper left corner

click on class software

click on FRS 2 Virtual Machine

wait - be patient...ok

Don’t touch anything until the window turns blue and prompted to put in password: frs2

click login

click away the two pop-up windows

click on little weird mouse talking thing upper left hand corner

click on terminal emulator

STOP DO NOT PROCEED WITHOUT INSTRUCTOR APPROVAL

**Meanwhile:**

Start by re-BLASTing your .fasta file from last week. You’ll want the BLAST results handy as you attempt to build the phylogenetic tree.

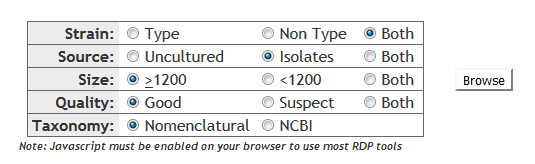
As a reminder, please make sure under “Choose Search Set” that:

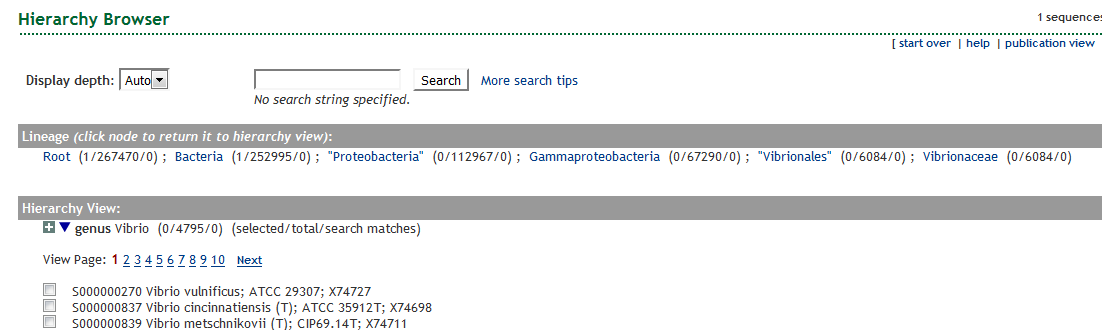
“Nucleotide collection (nr/nt)” is chosen for database

“Uncultured/environmental sample sequences” is checked

You may have to uncheck “Align two or more sequences” to access these options

* Navigate to RDP and login (using the account you made last week):
* <http://rdp.cme.msu.edu/myrdppub>
* Find your uploaded sequence and click the “+” to add it to your cart
* Click on “BROWSERS” (on top of the webpage)
* Click on “Isolates” to select only isolates for further analysis. Then click “Browse” ([Fig. 10](https://peerj.com/articles/960/#fig-10)).



* Search genus/species of interest (whatever came up as the closest genus match in BLAST). Type in the full name of interest (genus and species epithet) in the search box. 

*Now select all the species you think could represent your sample (these are going to become branches of your tree). Select by clicking on the checkbox or “+” to select all the listed names. Select those that are most similar in name to your top 3-4 BLAST results. Start by searching for the species, if there are less than 20-30 sequence matches for that species then you can select them all. If there are many more, you should manually select ~20 representatives from that species. Do this for each of the species that matched at the highest level of* ***identity*** *in your BLAST results. Note: click “refresh” whenever a series of selections is made.* This adds each species to your SEQCART. A SeqCart of 10-50 species (depending on your genus) will work well for tree building.

* Next you need to choose an outgroup to root your tree. (from the genus closest matched in your BLAST result, go back by one classification hierarchy *e.g. genus--> family* and choose an outgroup that you like) --This outgroup should be selected as in the previous step above when choosing the species that closely represent your sample. **Do not forget to click “refresh” as well after selecting a few species for the outgroup.** Not the browser refresh, the refresh next to the search box on the page.

*As discussed at the start of class, the best outgroup is as close as possible to the group of interest without falling within the group of interest. Best is to choose a type strain of a species from a genus within the same family as your group of interest.*

* View your selection by clicking on the SEQCART tab on top of the webpage (This will show you how many comparative sequences you have selected.)
* Click on “download”, leave the download options as the defaults (fasta, aligned, uncorrected), and then click on the appropriate download button (the appropriate download button is the bottom one for Model RDPX-Bacteria), **DO NOT click the download button on the top right of the page**). Save the file to your **desktop** and then rename it something informative. Afterwards, upload the file to the appropriate folder in the Google Drive.

Repeat this process for each isolate

**Step 2: Clean up your new FASTA file.**

The RDP alignment will have taxon names that most of the downstream software tools will not tolerate because they include special text characters. So, we have written a little Perl script (cleanup.pl) that will remove those special characters and replace them with underscores.

This script has been placed in the Virtual Machine on these computers. The first step is to get the file into the Virtual Machine as well. Open a web browser \*WITHIN THE VIRTUAL MACHINE\* by clicking the weird mouse thing to find the web browser icon in the dropdown. Navigate to your Google Drive and download the file into the virtual machine.

To run cleanup.pl:

* Go to the Terminal Emulator window that you already opened
* Navigate to the Desktop by using the cd (current directory commands)

username$ cd Desktop/

* Call the perl script by entering the following command.

username$ cleanup.pl -i **myrdp\_downloa.d\_𝓧𝓧\_seqs.fa** -o **FRS##clean**.fa

(𝓧𝓧 => look on Desktop to get number, ## => your sample)

*Your input file is denoted by –i, your output file is denoted by –o*

*Put clean somewhere in your output file name.* The new file will appear on the Desktop within the virtual machine if done correctly.

This will give you a new and improved file that FastTree can read!

Repeat Step for each isolate

**Step 3: Build a Tree with FastTree**

* Use the following command in the terminal to run FastTree

username$ FastTree -nt FRS##clean.fa > FRS##clean.tre

*Note: FRS##clean.fa should be your cleaned FASTA file from the previous step and may have a different name if you renamed it.*

*FRS##clean.tre will be your new output file and you should rename it*

*(nt = nucleotide tree) input file, “>” means write to the output file.*

*You better name it .tre or dendroscope won’t recognize it.*

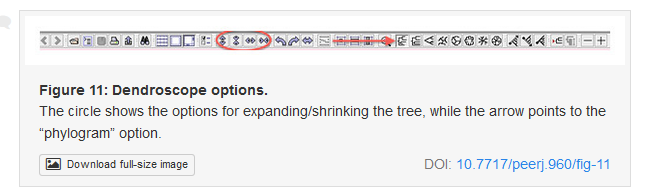
*Repeat Step for each isolate*

Then you need to move the .tre file back into Google Drive. \*\*Be sure you use the Google Drive that is open in the Virtual Machine\*\*!

**Step 4: View your Tree in Dendroscope**

* Find Dendroscope by typing “Dendroscope” in the search bar under the strange mouse icon
* Open from File → open → your file on the desktop
* Select “Interpret as node labels” when asked (if this does not come up, you did something wrong.)

*Once the tree is visible, the first step is to re-root the tree to the outgroup. Expand the tree by clicking the expansion button (labeled in*[*Fig. 11*](https://peerj.com/articles/960/#fig-11)*), then scroll through the tree to locate the outgroup. Click on the beginning of the taxon name, to select it, and re-root the tree by going to edit and selecting “re-root”. (If you don’t get a tree, you probably made a mistake in what file you inputted into the terminal. Start from calling the perl script again and make sure you clean up the RDP file. Don’t give up, you can do this!)*



*We recommend viewing the tree as a phylogram, which can be accomplished by clicking on the phylogram button (labeled in*[*Fig. 11*](https://peerj.com/articles/960/#fig-11)*). From this tree it should be possible to determine the phylogenetic placement of the candidate sequence, and in some cases to give it a name with more certainty than a simple BLAST search.*

Repeat Step for each isolate

Save Dendroscope files to the FRS 2 Folder.

**Summary of Unix/Linux commands and terms**

$ **ls** lists files and directories (folders). If left as just “ls” this command will list the files and directories in your current location. If a “path” is added afterwards (e.g., ls /usr) this command will list the files and directories in that location.

$ **cd** use to change directories

$ **cd ..** use to move up one directory

$ **cd directory\_name** use to move to that directory

$ **cd** ∼ use to move to the home directory of the current user

$ **less file\_name** view a file, type q to exit

A few quick definitions:

*command line*—the command line is where you type commands in a terminal window

*script*—a computer program. Usually computer programs are called scripts when they perform relatively simple functions that are limited in scope. Scripts are typically only run from the command line

*directory*—a folder

*compile*—turning a human-readable file into a computer-executable program