Wilson Lab – December 2019

Assembling Datasets and Performing Phylogenetic Analyses

Necessary Software. Most of this should be available for either the Mac and PC.

- Excel
- A text editor designed for programming, such as BBEdit (Mac) or Notepad++
- Mesquite by Maddison and Maddison LINK
- MUSCLE automated alignment tool LINK
- FigTree to view and modify your phylogenetic tree LINK

Assemble data in Excel spreadsheet

1. Create and organize sequences and associated metadata (organism name, specimen ID number, dates, locations, etc.) into the first tab a working spreadsheet.
    a. Individual taxa specimens are ordered in rows. These are your "operational taxonomic units" or OTUs.
    b. Organize specimen metadata and sequence data in columns.
    c. From here you will organize and re-organize by copy/pasting rows of data in an order that suits your fancy.
2. Use "Find and Replace" to remove offending characters.
    a. There can be no "-" in the columns that hold the metadata.
    b. Remove such characters as /()\?!,.;:'". Only use alpha-numeric and underscore characters. *This helps downstream with programs and analyses that don't read these characters.*
    c. A minus sign character "-" can only ever be used in a string of sequence data, either nucleotide or amino acid sequences. NEVER IN METADATA COLUMNS!
3. Organize datasets in new tabs.
    a. Copy information from main tab into a new tab.
    b. Organize/eliminate columns for FASTA files
    c. Organize information to be included in sequence name into the columns on the left.
    d. Follow by separator column with "_>_" for each sequence.
    e. Put column with sequence data after separator column. If sequence missing, place "-" in cell.
4. Save file.

Convert excel spreadsheet data into a FASTA file.

1. Copy/paste spreadsheet data into text editor (TextWrangler/BBEdit for Mac, Notepad++).
2. Use "Find and Replace" (your best friend!) to edit file.
    a. Search for tab characters ("\t"), replace with underscore ("_").
    b. Search for return characters ("\r"), replace with newline and greater than symbol ("\r>") to indicate the start of sequence name.
        i. You'll need to manually add ">" to the first line.
        ii. Double check the last lines of the file to ensure there are no extra ">" at the beginning of blank lines
    c. Search for your defined separator. In this example it's "_>_". Replace this with a return/new line character ("\r" or "\n").
    d. You should have a series of sequences that are FASTA formatted, looking something like this:

```
>SOME_RANDOM_SEQUENCE
ATTGCGCTAGCTTCGAACCGCTTAGGAGGAGCCTCGATCGCGTACTAACT
```

3. Save file using either a ".fas" or ".fasta" suffix.

Open file in sequence editor and perform automated alignment.

1. Launch Mesquite and Open the .fasta file you just made.
    a. When prompted select FASTA (DNA/RNA).
    b. You will then be prompted to save the file as a .nex file. Proceed as directed.
    c. If the file fails to open, pay attention to the warning signal that pops up. This will give you some inclination on how the .fasta file was poorly formatted or where offending characters might be located within the sequence.
2. Perform automated alignment, go to: Matrix>Align Multiple Sequences>Muscle Align…
    a. Proceed with either Separate or *No*
    b. Unless the "Path to Muscle:" was selected before, then direct Mesquite to the application using the Browse button.
    c. *OK*. Now wait a 30-90 seconds, depending on the complexity of your dataset, until the analysis is complete.
3. Trim ends, front and back, by selecting all the columns between the first set of informative characters to the end.
    a. Uninformative characters are when a single column is all one nucleotide, OR only one of the taxa has a different nucleotide from the rest. Blanks with a "-" don't count. Only evaluate cells with an ATCG. Columns (characters) with 2 or more cells with a nucleotide that is different from the rest are considered informative.
4. Scan matrix to proof any misaligned characters. This will take some skill in pattern recognition, and a certain amount of effort to effectively recognize areas that require correction. However, there will be instances where you cannot decide the correct course of action. In such instances, no action is better than trying to force a correction.
    a. To enhance the viewable characters in your matrix go to Display>Widths>Narrow Columns and/or Thin Rows. Birds Eye View will narrow the columns to maximize the number of colored characters you can view.
    b. Tools to move and adjust individual characters or blocks of characters are on the bottom right toolbar. Pick, choose and play to figure out how these work.
5. Export nucleotide matrix in Phylip format.
    a. File>Export
    b. Choose Phylip (DNA/RNA)
    c. Set "Maximum length of taxon names" to be >10. I tend to set it to 100.
    d. *Export*
    e. Save file as a .phy file.
6. Save your .nex file and all the modifications you've made.

Run a RAxML analysis on the CIPRES Web Portal.

1. Login to your account on the CIPRES Science Gateway at www.phylo.org/portal2/
    a. Create an account if you haven't done so.
2. Create a folder for your project. Here you will store all the data matrices to analyze as well as run the tasks for analysis.
3. Upload you dataset
    a. Select Data in your project folder
    b. Select the Upload Data button.
    c. Use the brows button to find the .phy file to upload. OPTIONAL: add whatever info you want into the boxes provided.
    d. Save
4. Analyze your data
    a. Select Tasks in your project folder
    b. Create New Task button

   c. Name your analysis in the description bar. Here it's good to put the name of the analysis run (in this case it's a RAxML) analysis, then the taxonomic group you are evaluating, and also the type of data (i.e. the gene used in the dataset).
   d. Select Input Data – This will take you back to the Data section of your project where you will select the data you just uploaded.
   e. Select Tool – This is where you will be presented with a dizzying array of analytical tools. For our purposes we're only interested in the RAxML analysis and MrBayes. Here select RAxML-HPC2 on XSEDE. (XSEDE is the supercomputer cluster. There's options to dive deeper into the explanations of all these terms if you want to at your leisure.)
   f. Select button next to Input Parameters – This will provide a large set of options for modifying the maximum likelihood analysis. The only ones you should worry about a this point are:
      i. Set Name for Output File – This should mirror the description of the analysis.
      ii. Outgroup – Put the taxa that you have declaired for your outgroup. They should be displayed EXACTLY AS WRITTEN in your data file. Each taxon should be separated by a single comma. NO SPACES.
      iii. Click on Advanced Parameters at the bottom.
      iv. Scroll down to Configure Bootstrapping.
      v. Bootstrap iterations – Add and extra zero to make it 1000.
      vi. Save Parameters
   g. Select Save and Run Task – To begin your ML analysis with bootstrapping.
5. You should receive an email notifying you when your analysis is done. Otherwise you can monitor the tasks by clicking the Refresh Tasks button on the top right of the page.
6. When the analysis is complete, select View Output button to the right of the task.
   a. You will want to download the RAxML_bipartitions.*your_filename* file for further analysis on your computer.
7. Troubleshooting – If there is no such file, then there was likely something wrong with the analysis.
   a. View the STDOUT file as this will provide some indication of what is wrong with either the analysis parameters or your input file.
   b. Go back to the previous steps to make sure you've adequately removed offending characters, etc. from your dataset. Then re-upload the analysis. This can be done easily by selecting the CLONE button. Here all you will need to do is replace the data file in which to run the analysis on.

View your phylogenetic tree in FigTree.

1. Open FigTree
2. When prompted you can either change the "label" to "bootstrap", or simply ignore.
   a. It's asking you want to call the numbers associated with each branch. These are the bootstrap bipartition data.
3. Add bootstrap data to your tree.
   a. Check the box next to Branch Labels. A bunch of numbers and data will appear on the branches of the tree. This isn't the data we want, we need to tell FigTree what we want.
   b. Click the arrow next to this to present the options for branch labels. Here you can modify size, font, color, etc. for the branch labels.
   c. In the dropdown menu select the last option, which will be either "label" or "bootstrap" or whatever you decided to call the bipartition data. This is your bootstrap % for the branches on your tree. Generally speeking 90-100% is great. 80-89% is good, 70-79% is OK, <69% is considered week and not worth reporting.
4. Play with FigTree to modify your tree as you see fit, but save file under a different name. You can always go back to the original file if you modify past the point of return.