Lab 02:

Introduction to Nexus and Newick formats;
Introduction to FigTree and Mesquite

Updated by Will Freyman

1 Before you begin

1.1 Lab grading

These labs are graded based on participation. The way I'll keep track of your participation is by asking you to keep copies of various files or answer questions. At the end of the exercise you'll email me the files and answers to your questions.

1.2 Software needed

Please download and install the following software:

1. Mesquite: `http://mesquiteproject.wikispaces.com/`
2. FigTree: `http://tree.bio.ed.ac.uk/software/figtree/`
3. A plain text editor such as:
   - Sublime Text: `http://www.sublimetext.com/`
   - TextWrangler: `http://www.barebones.com/products/textwrangler/`
   - Vim: `http://www.vim.org/`

1.3 Files needed

Please download the following files:

1. Amblygnathus.nex `http://ib.berkeley.edu/courses/ib200/labs/02/Amblygnathus.nex`

2 Introduction

Today we will learn about the **Nexus** and **Newick** file formats, which are widely used by phylogenetic programs. By having common file formats the results of one program can be viewed or analyzed in another (although this does not always work as well as one would hope). You can manipulate these file formats directly in a text editor, but often it is better to use a program like **Mesquite** or **FigTree** and save the results to be used in another program.

3 Making Trees Pretty in FigTree

FigTree is developed by Andrew Rambaut's research lab, who are well known for developing the the Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software package. FigTree is a highly useful way to quickly view phylogenetic trees and to produce publication-ready figures. Here we'll just cover some quick basics.

Open the `Amblygnathus.nex` file in FigTree. `Amblygnathus.nex` is a Nexus file, and it contains multiple phylogenetic trees in the same file. Click the `Current Tree:` tab (left hand side of window) to scroll through the trees. Try experimenting with the options under the `Layout` tab. Use the `Selection Mode` tab (at the top of window) to select a `Clade`. Color one of the clades on the trees. Under the `Trees` tab (right side of window) click `Order nodes` to *ladderize* the tree. Using the `Node Labels` tab (right side of window) add the node ages to the tree. Explore other options to modify the appearance of the tree. Under the `File` menu, export the ladderized tree with the colored clade and node labels (and any other changes you feel like making) as a pdf. You'll email me the pdf file when you complete the lab.

Now let's convert the Nexus file into a Newick file. Remember, the Nexus file contains multiple trees, so to export the single tree you've been editing follow these steps: under the `Edit` menu select `Copy`, and then under the `File` menu select `New`. Then under the `Edit` menu select `Paste`. Now select `Export Trees` under the `File` menu.

4 Newick and Nexus File Formats

In a plain text editor, open the Newick file you just saved. Newick format is a commonly used way of representing tree topologies as text. Put simply, monophyletic clades are surrounded by parentheses and sister clades are separated by commas. For example, a simple tree could be written as (((A,B),C),(D,E)). Newick format also contains information about branch lengths (after colons) and node names (after closed parentheses).

Now lets take a look at the Nexus file `Amblygnathus.nex` by opening it in a text editor. Every Nexus file starts out with `#NEXUS` and then is followed by a brief description of the file surrounded by brackets. This is followed by the actual data of the Nexus file, which are organized into several blocks. Each block starts with a line `BEGIN BLOCK NAME;` and finishes with an `END;`. The lines between, which hold the actual data, are often indented. Each program creates and uses different blocks with information and commands that are particular to it, but there are several block types that are almost universally used and contain the most fundamental information for phylogenetic analysis.

The data block (ie. `BEGIN DATA`) contains your data matrix. Some Nexus files use a taxa block (ie. `BEGIN TAXA`) and a character block (ie. `BEGIN CHARACTERS`) instead of the data block. Most programs, but not all, are flexible and can use either format. The data block basically contains all the same information as the other two blocks.

The data block must contain a `DIMENSIONS` and a `FORMAT` line, which describe the data in the matrix. Datatype determines the basic class of data: discrete; continuous; protein; or DNA. After that there are several commands, which describe what symbols are allowed and what they represent. For example, does a `-` mean unknown or missing. The characters block may also contain other information such as `charlabels` (names for the different characters) and `statelabels` (names for the different character states for each character) or `charstatelabels` (both types of info combined in one command). Next is the actual data matrix `MATRIX`. Each line of the matrix starts with the name of the taxon represented and is followed by a series of symbols representing the character states for the various characters in the matrix. Each taxon line must have exactly the same number of characters in the same order.

You will also see `BEGIN TREES`, which is the start of the trees block that contains your phylogenetic tree information. A Nexus file will only contain a tree block, when it is necessary to import a tree into a program. Tree blocks often start with a `TRANSLATE` command, which is required for a number of programs, but not all. It is a list of consecutive numbers followed by the names of the taxa that those numbers will represent. This is followed by a `TREE` line (in Amblygnathus.nex there are 21 different trees), which describes the tree in Newick format using the numbers assigned in the translate command for the names of the taxa.

5 Intro to Mesquite

Mesquite was developed by Wayne and David Maddison (twin brothers who are both phylogeneticists!) as a tool for interpreting phylogenetic information. The strengths of Mesquite are creating and editing data matrices, examining the distribution of features on a phylogeny, and testing hypotheses about character evolution. This program will be a great way to explore the data you collect for your final project, especially if you are not quite yet comfortable with R. However, Mesquite is NOT the program you should use for any of your tree-building. You need to implement maximum parsimony, maximum likelihood, or Bayesian inference in another program such as PAUP* or MrBayes to build your trees.

Mesquite is modular, which means that the program is set up as a bunch of modules that all do different functions, such as draw a picture or do a parsimony analysis. Some modules use other modules and the modules are used in combination to perform an analysis. Thus Mesquite is very flexible and capable of doing analyses that it was never intended to do. Unfortunately it also means that Mesquite can be difficult to use, because it is not always clear where to find the appropriate command in its menus.

6 Editing a Data Matrix in Mesquite

Open Mesquite and select `File>Open` to open the `Amblygnathus.nex` file. A *Project* window will open showing the *Character Matrix.* Here you can edit and add characters and character states. There are several tools along the left side that allow you to manipulate the matrix. When you hold the cursor over each of the buttons, a description of what it does appears at the bottom of the window.

To the far left is a panel showing some of the other viewing options. Click the `Taxa>List & Manage Taxa`. The *Taxa Block* can be used to edit information about taxa. Change some of the taxon names.

Now go back to the *Character Matrix* by clicking the tab at the top of the window, or the button on the far left. Now let's add some data to the matrix. Select the edit tool again and click on the various cells in the matrix. Since this is a categorical data matrix, all the characters must be consecutive integers starting with 0. Use the tools on the left of the screen to add a new character. Just make up the data.

A matrix filled entirely with 0s and 1s is often not the easiest thing to interpret; we would be much better served if the characters and character states were more descriptive. At the bottom left are five small buttons next to a blue i that look like little windows. Select the one second from the right, the `Show State Names Editor Window`. For example maybe character 1 is *common sense* and its states could be 0, *absent*, and 1, *present.* Now go back to the *Character Matrix* by clicking the tab at the top of the window. As you can see all the 0s and 1s have been replaced with actual words. One thing to note is that when adding or editing character states in the matrix you still need to type in 0 and 1, even if the states now have more interesting names.

Now save the Nexus file with your new character. When you finish the lab you'll email me a copy of the modified Nexus file.

7 Ancestral Character Reconstruction Using Parsimony

In the far left panel click `Trees>View Trees`. A new tab will appear with a tree of 29 taxa in it. On the top left side there is a menu button that allows you to scroll through the 21 different trees stored in this file. Scroll through these trees to see some of the differences in topology. In the left-hand column of the window are several tools for modifying these trees. More options are in the `Display` menu. Explore these different functions.

Ive already mentioned that Mesquite is not a program for building trees or for creating alignments of sequence data. The whole purpose of Mesquite is to analyze how characters change on trees. There are two very broad categories of data: discrete and continuous. A discrete character might be *Mandible indentation* and the states could be *present* or *absent*. A continuous character might be something like body size. Lets look at some discrete characters.

You are still probably in the *Tree Window*. From the *Tree Window* go to `Analysis>Trace Character History`. Select `Parsimony Ancestral States` and hit OK. (We will discuss the different options here later on in the semester.) Neat! Now each of the morphological characters in the data matrix is mapped onto the tree with the ancestral states colored along the branches. You can go through the different characters by clicking on the arrows in your *Trace Character* box in your *Tree Window*. Scroll on over to Character #5 and look at the branch connecting *darlingtoni* and *bicolor*. There should be several colors along this branch. Any ideas as to why that is? If you have no idea, discuss with your fellow lab mates or ask me. This is called an ambiguous reconstruction. FYI: Character #5 of the Amblygnathus data set is the number of spines on the internal sac of the male aedeagus (ie. penis). Because beetles are awesome like that and have spines on their genitals.

What is an ambiguous reconstruction? Write the answer to this and include it in the email you send to me at the end of the lab.

In Mesquite it is also possible to reconstruct ancestral states for discrete characters under different parsimony models (for example unordered states). You can also reconstruct ancestral states for discrete characters using likelihood. But well wait to do this another time.

Please email me the following:

1. Pdf file of the pretty tree you made in FigTree.
2. The Nexus file you modified by adding a new character in Mesquite.
3. Your answer to *What is an ambiguous reconstruction?*