Using PHYLIP Software to Generate Neighbor-Joining or UPGMA Trees from Genetic or Morphological Distance Matrices

Introduction:
PHYLIP (created by Joe Felsenstein) is a very flexible program for conducting phylogenetic analyses from genetic or morphological data sets. It can be used for a number of different analyses. Our lab uses it frequently to generate Neighbor-Joining trees depicting relationships among samples from distance matrices.

Many programs will generate trees depicting relationships among individuals. However, it is often useful to create trees that depict relationships among sets of individuals that fall within predefined population samples, such that each tip of the tree represents one of the populations in the analysis. If one has 20 individuals in each of five samples, a tree of 100 individuals can be difficult to interpret. Often, it is useful to draw a tree of the relationships among the five samples based on the best estimate of how they are related, as inferred from the individual data.

We have used PHYLIP for generating trees depicting patterns of divergence among samples for:

1. Data on genetic divergence among samples based on genetic distance matrices generated using allele or haplotype frequency data
2. Morphological divergence among samples based on Procrustes distances between sample consensus configurations.

How does PHYLIP work?
PHYLIP can be downloaded from: http://evolution.genetics.washington.edu/phylip.html

It is not one program but a series of modular programs. When you download it, you will see a folder called **doc** (for the program documentation in html format [will open in your browser]), a folder called **exe** (for the executables or the individual programs themselves), and a folder called **src** (for the source code). The main instructions for the package can be found in the **main.html** file located in the doc folder. It is useful to review that file before starting. There are separate instruction files for every program in PHYLIP but one typically needs to read only the instructions for the programs that will be used.

To run an analysis, you have to know which program to use for what you want to do and then run it by finding it in the exe folder and clicking on it. The program will use an input file that you give it, ask your for directions on the options you want for the analysis, conduct the analysis (often within a few seconds), and then spit out the output file(s) into the exe folder. Often, a particular analysis will involve you running your data through several of the programs in a given sequence.

Creating a Neighbor-Joining tree
You will need to have a pair-wise distance matrix as a simple text file. This matrix depicts the difference between every pair of samples in your data set. What distance measure you use will depend on the type of data. For morphometric data, we typically use the Procrustes distance and each number in the distance matrix will be the Procrustes distance between each pair of consensus (sample average)

configurations in the study. The distance measure that you use matters; the analyses conducted are only as good as the data used. If the data are crap, the analysis will be crap.

It is often useful to have the data in an Excel file and then save it as a tab-delimited text file. The distance matrix can be a full matrix or an upper or a lower triangle distance matrix (one simply tells PHYLIP which, when prompted).

An example of an input file is listed below:

		13											
S01	0.000												
S04	0.000	0.000											
S05	0.000	0.000	0.000										
S06	0.037	0.037	0.037	0.000									
S07	0.000	0.000	0.000	0.037	0.000								
S08	0.259	0.259	0.259	0.291	0.259	0.000							
S09	0.385	0.385	0.385	0.385	0.385	0.545	0.000						
S11	0.385	0.385	0.385	0.385	0.385	0.566	0.330	0.000					
S12	0.536	0.536	0.536	0.536	0.536	0.684	0.362	0.303	0.000				
S13	0.578	0.578	0.578	0.578	0.578	0.743	0.486	0.275	0.227	0.000			
S14	0.352	0.352	0.352	0.384	0.352	0.272	0.674	0.649	0.778	0.798	0.000		
S15	0.278	0.278	0.278	0.308	0.278	0.258	0.602	0.602	0.753	0.794	0.211	0.000	
S16	0.299	0.299	0.299	0.299	0.299	0.473	0.180	0.303	0.392	0.526	0.604	0.516	0.000

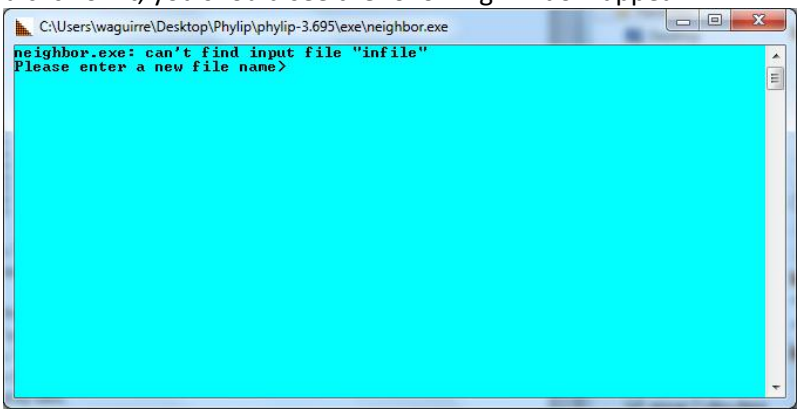
The size of the text is reduced so that it fits. The first row (13) indicates the number of samples. Hit tab and then write the number of samples in your data set.

The first column has the sample labels. In this case they are sites, S01 (Site 1), S04 (Site 4), etc. This column should be ten characters wide so if your specimens have less than ten characters in their names, fill in the rest of the characters with spaces. The columns that follow are the genetic distances between each pair of samples and the diagonal of 0s is included. The columns should be separated by tabs.

This is an example of a lower triangle distance matrix because only the data in the bottom left part of the matrix are included. The top right would have the same data mirroring the bottom left, so one can leave it out.

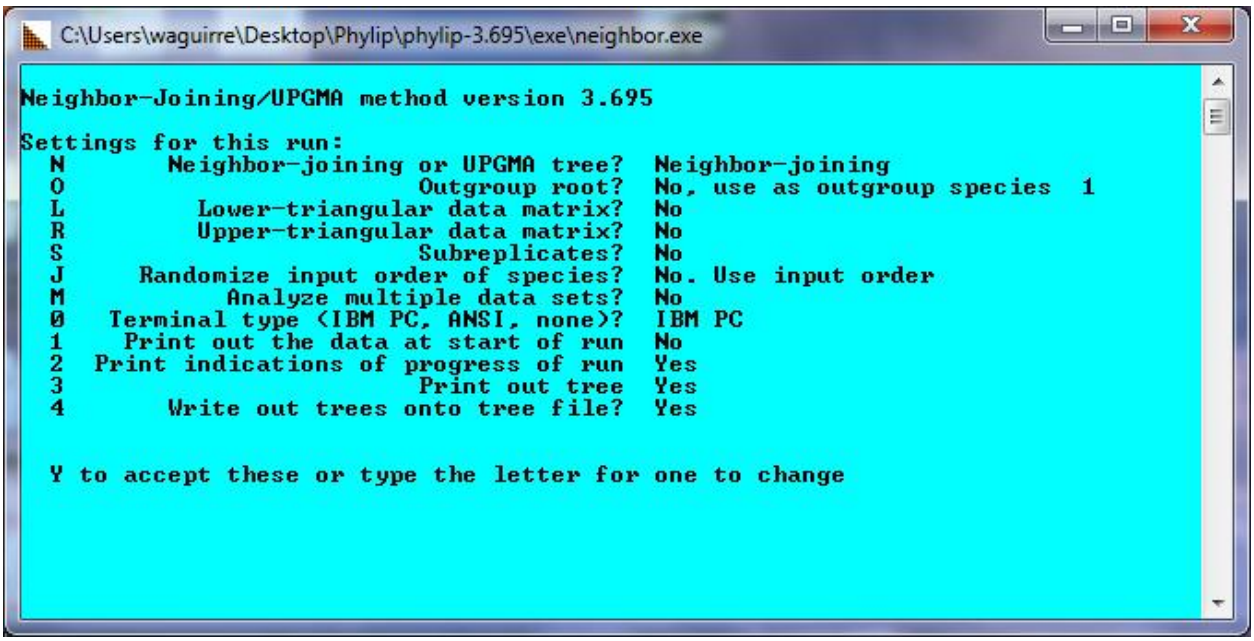
This genetic distance matrix would be the input file. Remember that it should be tab-delimited text file. Word files (.doc) have hidden characters in them and will not work.

Put this input file in the **exe folder**. Now click on the program **Neighbor**. Neighbor is a program that conducts a Neighbor-Joining cluster analysis of your data based on the distances among samples provided. When you click on it, you should see the following window appear:



Just type in the name of your input file. Remember to make sure that it is in the exe folder (where the Neighbor program is running from) and that you type in the file extension if it has one (e.g., .txt if the file is named infile.txt). If your file is named infile with no extension, the Neighbor program will skip this window.

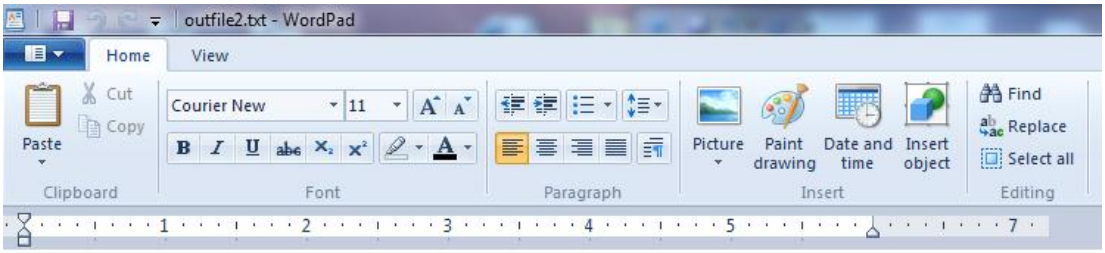
You should then get the following window of menu options. Here, the Neighbor program is basically asking you for modifications from the default conditions under which it runs. It allows you to run a UPGMA analysis instead of a Neighbor-Joining analysis, designate one of your groups as the outgroup (sample at the base of the tree), etc.



In our example, all can be left under the default conditions except for option L. Our data matrix will typically be a Lower-triangle matrix so this should be changed. Type “L” and click “enter”. The L option will change from “No” to “Yes”. Once this is done, you can type “Y” (accept the conditions) and click “enter” and the program will run the analysis.

For a typical data set, the analysis should run almost instantaneously and be done within a second or two. The program will create two output files, an “outfile” file and an “outtree” file, and these will appear in the exe folder that you have been working in. The outfile is the one you care about. Open it in a word processor. I use Wordpad (under Accessories in Windows).

The results of the analysis, including your tree, will appear in the file as text. The tree is what we care about. PHYLIP does have programs within it to draw nicer trees but we typically just copy the tree and then paste it into Paint (also under Accessories in Windows) and make a nicer tree by hand. See the next page for what the text version of the tree looks like. Remember that what matters are the horizontal distances. Vertical distances do not count so you can change the position of branches as long as the change does not alter the horizontal lengths of the branches. For example, I would probably modify the position of S07 in the tree below so that it appears on the top side of the tree closer to S05 and S06.




```
S14             0.35200   0.35200   0.35200   0.38400   0.35200   0.27200
                0.67400   0.64900   0.77800   0.79800   0.00000   0.21100
                0.60400
S15             0.27800   0.27800   0.27800   0.30800   0.27800   0.25800
                0.60200   0.60200   0.75300   0.79400   0.21100   0.00000
                0.51600
S16             0.29900   0.29900   0.29900   0.29900   0.29900   0.47300
                0.18000   0.30300   0.39200   0.52600   0.60400   0.51600
                0.00000
```

```
  +S04
  !
  !  +S05
  !  !
  !  !    +S06
  !  !    !
  !  !    !         +------S08
  !  !  +-8 +------6
  9-10  ! ! !       !    +-------S14
  !  !  ! ! !       +---5
  !  !  ! ! !           +---S15
  !  !  ! +-7
  !  !  !   !             +-------S09
  !  !  !   !           +-3
  !  +-11   !           ! +--S16
  !     !   +-----------4
  !     !               !     +---S11
  !     !               +-----2
  !     !                     !      +----S12
  !     !                     +------1
  !     !                            +-------S13
  !     !
  !     +S07
  !
  +S01
```

This is your tree!

```
remember: this is an unrooted tree!

Between        And            Length
-------        ---            ------
   9           S04            0.00000
   9             10           0.00000
  10           S05            0.00000
  10             11           0.00000
```

Created by Windsor Aguirre
Last modified Sep/3/15