Integrative Biology 200A *University of California, Berkeley*
"PRINCIPLES OF PHYLOGENETICS" *Spring 2006*

Lab 1: Introduction to PHYLIP

Introduction

Today we will be learning about some of the features of the PHYLIP (PHYLogeny Inference Package) software package. PHYLIP was developed by Joe Felsenstein, works on most operating systems, and is available for free online.

Methods that are available in the package include parsimony, distance matrix, and likelihood, as well as bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites, distance matrices, and binary (0/1) discrete-state characters. More information about PHYLIP and copies of the program are available at: http://evolution.genetics.washington.edu/phylip.html.

PHYLIP consists of about 30 programs that perform different algorithms on various types of data. Today, we'll be focusing on MIX, CLIQUE, CONSENSE, and NEIGHBOR.

Creating the Input Files for MIX and CLIQUE

We need to use a specific format to enter data into PHYLIP. Today we'll concentrate on entering some binary discrete state characters. The data sets we'll be using are the one from you Wagner tree and compatibility homework exercises. Make sure to use both data sets. Here are instructions on how to build your input file:

-First, open Text Edit on your computer.

-On the first line, put in 5 blank spaces. Then, add the number of taxa in the data set. In the case of the Wagner tree example, this will be 5.

-Insert another four spaces, and then put in the number of characters in the data set. For the Wagner tree example, this will be 6.

-On the next line, enter your first taxon. The taxon name must be 10 characters long, and can include punctuation and space. After the taxon name put the codings for the six characters. Your finished entry should look something like: Ancestor 000000.

-Enter the rest of your taxa on separate lines. Your finished data matrix should look like :

```
     5    6
Ancestor  000000
L.longipes000000
L.leyseroi111111
L.tennella111111
```

```
L.gnaphalo111100
```

-Save your file as a text file in the following PHYLIP folder:
Applications> IB200A> Phylip>exeA

- Name the file with a .txt extension. Repeat this process with your data matrix from the compatibility exercise.

MIX

Now that we've made our data files, we'll move onto analyzing them with some of the programs. MIX is a program that uses Wagner or Camin-Sokal parsimony to create cladograms. Wagner parsimony is the default, and it allows character state transitions from 0 to 1 and from 1 to 0. Camin-Sokal parsimony allows character state transitions from 0 to 1, but not from 1 to 0 (i.e., reversals are prohibited). Both methods search for the most parsimonious cladogram or cladograms. What assumptions about ancestral states do the two methods require?

-Open MIX in the exe folder. The program should ask you to enter a file name. Enter the name of your input file here (e.g., infile.txt).

-Once you've done that, a screen with a list of options should appear. All of the options are described in the documentation that comes with PHYLIP. We'll just change a few.

-First, we will randomize the input order of the species. This will allow MIX to search a greater portion of tree of tree-space, giving us more confidence that we have found the most parsimonious result. Enter j and hit enter. Enter any odd number to act as a random number seed. Enter 10 as the number of times to jumble (this will tell the program to randomly add the taxa together to form a starting tree 10 times; each of the 10 starting trees will then be used in its searches for a most parsimonious cladogram).

-Next, we'll change the outgroup. Enter the number of you 'ancestor' taxon here (i.e. is it taxon 1 in the data matrix or taxon 2 or taxon 5?)-Finally note that we are using Wagner parsimony in this run. Make sure you also try Camin-Sokal parsimony.

When the analysis is finished, the program window will close, and two new files, 'outfile' and 'outtree' will be added to your PHYLIP folder. You can open these files in Word or Text Edit to check the results. The outfile shows the cladogram(s) you have inferred, and gives some details of the results. The outtree gives you a simple, parenthetical representation of your cladogram(s).

Repeat this process using Camin-Sokal parsimony. Make sure to change the names of your output files after each run so that they don't get overwritten by your next run. Do these methods give you the same results?

CONSENSE

If you got the same results as I did, when you ran the Camin-Sokal parsimony option on the Wagner tree homework data set, you ended up with three most parsimonious trees. Maybe you would like to compute a consensus cladogram to summarize your results and see what parts of the three trees are in agreement. We'll talk more about consensus cladograms in lecture. For now, be aware the CONSENSE will compute strict and majority-rule consensus cladograms.

-First, rename your outtree file from your Camin-Sokal run. This is a seemingly minor point, but it is important to get the program to run. If you accidentally erased you Camin-Sokal results from the Wagner Tree exercise, simply rerun the analysis.

-Open CONSENSE, type the new name of your tree file.

-A screen of options should appear again. The only one you need to change is the outgroup root. Change it to '5.' Although the ancestor was first in your matrix, it is now the fifth taxa in your input trees.

-The CONSENSE results will be written into the outtree file and outfile. You can check your results by opening the files in Word or notepad.

Are you surprised by your results? Make sure you understand why you got the results that you did. Part of the problem is an idiosyncrasy of the programs, but another is an important point. (Hint: rerun the Camin-Sokal analysis and look at the trees you ended up with.)

CLIQUE

CLIQUE uses compatibility methods to find the largest set of compatible, binary, unrooted characters and the tree(s) they suggest. You can also set a minimum clique size and the program will then list all cliques (and trees) that are that size or larger.

-Open CLIQUE. The program should ask you for a file name. Enter the name of your compatibility data matrix.

-An options screen should appear. You may want to change your ancestor.

-The results should be written to an outfile and an outtree file.

- Examine the results. Are they the same as we got in class.

DISTANCE

Now let’s try using a distance matrix to create a tree. The first thing we need to do is create a distance matrix from DNA data. This is an estimate of the number of changes between pairs of DNA sequences. Rather than just counting the number of bases at which two sequences differ it accounts for the fact that a second substitution at one site will mask the first substitution.

-Open the SEQUENCE documentation (it should be an HTML file in the Phylip/doc folder). Scroll down and find the first data set that looks like this:

```
     5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
Salmo gairAAGCCTTGGC AGTGCAGGGT
H. SapiensACCGGTTGGC CGTTCAGGGT
Chimp     AAACCCTTGC CGTTACGCTT
Gorilla   AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

-Copy this into a new text file and save it in the Phylip/exe folder. Make sure that there are still five spaces before the first number.

-Now open DNADIST. The program should ask for the name of your data set. Enter it.

-Let's change the distance measure to Kimura. This allows for only two parameters, transition and transversion rates.

-Run the program and look at your outfile in word. It has a very different formula than the files we have used before. It is a matrix of the distances between the species' DNA sequences.

NEIGHBOR

NEIGHBOR is a program that allows you to group taxa together based on the distances between species. You can use Neighbor-Joining or UPGMA to create your phenogram. We'll talk a bit more about these methods and their assumptions in lecture.

Let's use it to construct a tree from the distance matrix we just created.

-First rename the outfile from DNADIST.

-Now open NEIGHBOR. The program should ask for the name of your data set. Enter it.

-A list of options should now appear. Run a neighbor-joining analysis. When you run the program, the results will be placed in an outfile and an outtree (as usual). Be sure to also run a UPGMA analysis and compare the results.