STAT 882 - STATISTICAL PHYLOGENETICS SPRING 2010

LAB 1: PARSIMONY ANALYSIS USING PAUP*

1 Data

Throughout the lab portion of this course, we will make use of at least three published data sets:

- **primates.nex:** This data set consists of mtDNA sequences for 12 primate species, published by Hayasaka et al. (*Mol. Biol. Evol.* 5: 626-644, 1988).
- **hpv.nex:** This data sets consists of sequence data for the L1 gene in 38 papillomaviruses. It was used by Morrison in a paper comparing methods for maximum likelihood phylogenetic inference (*Syst. Biol.* 56(6): 988-1010, 2007).
- **rokas14.nex:** This data sets consists of 106 genes for 14 species of yeast. It was originally published by Rokas and Carroll (*Mol. Biol. Evol.* 22: 1337-1344, 2005), and was used by Morrison as a test data set in the paper mentioned above.

These data sets are all available at my webpage: www.stat.osu.edu/∼lkubatko/stat882/
.

2 Introduction to PAUP*

The program PAUP* (Phylogenetic Analysis Using Parsimony (*and other methods)) is a package written by David Swofford and distributed by Sinauer Associates. It has only been released in a beta version, but is available for purchase online (http://paup.csit.fsu.edu/). Documentation can also be found at the web address above (click on Downloads and then Command Reference Document).

We'll be using PAUP* through the Department of Statistics. To access PAUP*, you'll need to log in to the department server. Using a secure connection (e.g., ssh – see), connect to username@mordor.stat.osu.edu. At the prompt, type `paup` to launch the program. Typing `quit` exits the program.

You can work through the set of steps below to learn to use the PAUP* program. Before we begin, however, let's take a minute to understand the input file formats for PAUP*. To read in sequence data, PAUP* uses the NEXUS file format. This is a fairly standard file format for phylogenetics (e.g., it is the format used for MrBayes and BEST, two of the program we'll discuss later this quarter, as well). Together, we'll take a look at the NEXUS input file for the primates data. Using your ssh login, type `more primates.nex` at the prompt.

If we want to perform an analysis using the parsimony criterion in PAUP*, there are two main things we might consider. First, we might read in the data and carry out a search for the MP tree. Second, we might read in the data and a set of trees we'd like to compare, and evaluate the parsimony scores of those trees. We'll consider doing both types of analysis below. To prepare for the second type of analysis, let's look at the tree file format. PAUP* (and most other phylogenetics programs) uses the Newick format for trees. It's easiest to understand this format by taking a look at an example – type `more primates.tre` at the prompt to see an example.

Finally, let's spend a few minutes talking about the search strategies available in PAUP*. Once we complete that, please feel free to work through the questions below on your own. I'll walk around and help where needed.

3 On Your Own

3.1 Primate data

1. Read the data stored in the file primates.nex into PAUP*. To do this, launch PAUP*, and issue the command

   ```
   >exe primates.nex
   ```

   If PAUP* successfully reads the file, you should see a message like the following:

   ```
   Processing of file ‘‘primates.nex’’ completed.
   ```

   This means that the data has been read in and stored by PAUP*.

2. Read in the two trees in the file `primates.tre` using the command

   ```
   >gettrees file=primates.tre
   ```

3. Compute the parsimony scores of the two trees:

   ```
   >set criterion=parsimony
   >pscores
   ```

   Now try the commands

   ```
   >pscores / single=all
   >pscores / single=var
   ```

   What information does this give you? (Look at the documentation if you need to).

4. Now carry out a heuristic search for the most parsimonious tree. Do this using the `hsearch` command. You can do this simply by using the command

   ```
   > hsearch
   ```

   You should also read through the documentation for this command and try some other settings. Experiment with more thorough and less thorough searches, and see how that affects your results. Some subcommands within `hsearch` to think about modifying are `nreps, swap, addseq`.

5. Some other useful commands are:

   `showtrees 2` — prints the second tree stored in memory to the screen

   `savetrees file=mytrees.tre from=1 to=2` — saves trees 1 and 2 to the file mytrees.tre in the current directory in Newick format

3.2 HPV Data

This data set is interesting because it is known to have local optima when the likelihood criterion (to be discussed next week) is used (see Morrison, *Syst. Biol.* 56(6): 988-1010, 2007, for details). We'll explore here whether there are also local optima in the parsimony score. To do this, try each of the following commands:

```
hsearch addseq=random nreps=20 swap=nni;
hsearch addseq=random nreps=20 swap=spr;
hsearch addseq=random nreps=20 swap=tbr;
```

Try each command several times (you'll get a different random number seed each time, so the results will be different). Think about consistency between runs with a particular branch-swapping strategy and also how the branch swapping strategies compare to one another. Make sure the PAUP* output makes sense to you. What would you recommend to someone to search for the MP tree?

3.3 Yeast Data

If you have any time left, you can play with the yeast data as well. Try the commands you used above for the HPV data. Notice that this data set is fairly large in the number of sites (characters), but since the number of taxa isn't very large, searches still run quickly. Think about why increases in the number of characters aren't as costly as increases in the number of taxa would be (hint: consider computation of the parsimony score for two site patterns that are identical).