Biology 545 - Phylogenetics

Laboratory 2: Searching Tree Space in PAUP*

This exercise has three goals:

1. Provide you with some experience using PAUP* in the command line.
2. Teach you about tree searching (regardless of program).
3. Illustrate the effects that decisions in search strategy can have on estimates of phylogeny.

One reason PAUP* is so useful is that is has the ability to evaluate trees under each of the three optimality criteria we've discussed. In addition, it has the flexibility to run both fast and approximate "algorithmic" analyses (such as neighbor joining, quartet puzzling, UPGMA, etc.) as well as searches of tree space (both exact and heuristic).

Section 1: Exploring Tree Space

1a. The NEXUS file format

You were briefly introduced to the NEXUS file format in Lab 1. PAUP* (and other programs) requires NEXUS files as input. NEXUS files are composed of several pieces of information, usually separated into blocks, including the taxa, characters, and some information about the sequences. Trees can be input or output in parenthetical (Newick) notation. In PAUP*, these are accompanied by a TAXA block that associates taxa with identifiers used in the parenthetical description of the tree within the file.

NEXUS files start with #NEXUS. They are traditionally followed by a data block, which looks something like the following. Notice the notation used for beginning and ending the block.

```
#NEXUS
begin data;
dimensions ntax=234 nchar=4703;
format DATATYPE=DNA interleave=yes gap=- missing=?;
[anon:1-751, acr:752-1845, zan:1846-2703, zp2:2704-3584,
cytB:3585-4703]
[all data model: TrN+I+G]
Matrix
Species1  atgcatgcatgc
Species2  atgcatgcatgc
Species3  atgcatgcatgc
.
.
.
Species233     atcgatgcatgc
```

```
Species234      atcgatcgatgc
;
End;
```

This lets an interpreter know that some information about the data is about to be read in. There are 234 taxa, and the dataset is 4703 characters. The characters are DNA sequence data, they are interleaved, gaps are represented by hyphens, and missing data are represented by question marks. Each line MUST end with a semicolon, or the program trying to execute the file will crash. Comments can be included in NEXUS files and are enclosed within square brackets. Here, the scientist has made note that this is a concatenated dataset composed of five genes whose start and stop points are listed. The model of nucleotide sequence evolution is also listed: TrN, plus a proportion of invariable sites, plus gamma-distributed rate heterogeneity among sites.

The data come next, defined with the heading MATRIX. The block ends with an END;.

There are many variations on the NEXUS format, and it is extremely common to need a NEXUS file for one part of an analysis and a PHYLIP or FASTA file for another. The EMBOSS suite of programs can also provide this functionality; the program seqret may prove helpful: https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/

The dataset has been emailed to the class mailing list and posted on the course website. Download the file to your desktop and open it up to take a look at it. Identify the parts of the NEXUS data blocks mentioned above.

In the next section, you will use this NEXUS formatted dataset. For this you will need to move the nexus file on your local computer to your account on the cluster. For this, use the `scp` command.

```
scp ~/path_to_file/biol545ParDat.nex # path on personal computer
usrname@ford.hpc.uidaho.edu:/mnt/ceph/usrname/ # path on UIdaho
HPC
```

Alternatively, if you are on a windows OS, you can use the secure file transfer protocol (`sftp`). Open a terminal window and move into the directory where you have saved your NEXUS file and use the following function:

```
sftp usrname@ford.hpc.uidaho.edu

put biol545ParDat.nex
```

1b. Heuristic Searching

PAUP* is invoked by calling it from the command line. Login to one of the classroom servers and load the PAUP* module.

```
ssh usrname@ford.hpc.uidaho.edu
```

Agree to the connection and enter your password. Next, you can see all of the modules that are available to load with '`module avail`' and you can load the paup module with `module load paup`

To run PAUP* and immediately load your data, the command is:
`paup [Your Input NEXUS File]`
You can also just run paup using 'paup' and then load the data using:

```
execute [Your Input NEXUS File];
```

This will load the information from the NEXUS file into PAUP* in preparation for analysis. If you have included a PAUP block (a list of commands you want to execute in PAUP*), the analysis will begin automatically.

PAUP will output everything to the standard output (STOUT) which is lost upon ending the session. To save your PAUP STOUT to a log file for future reference, you can run the following command in PAUP.

```
log file=biol545_lab2.log
```

In this exercise, we'll demonstrate the need to search tree space thoroughly.

First, we'll do a very greedy search, by only saving one tree at any time and using NNI branch swapping. NNI stands for nearest-neighbor interchange, and branch swapping is used to create alternative topologies to be included in the search of tree space. We will go over this and other branch swapping techniques in lecture, but NNI is the least rigorous.

Type the following commands.

```
set criterion=parsimony maxtree=1 increase=n;
hs swap=nni;
```

The first line designates the optimality criterion as parsimony, limits the program to a single tree in the tree buffer, and prohibits the program from increasing that limit if > 1 equally parsimonious tree is found, either during construction of the starting tree or branch swapping. To see other options for these command type '`set ?`' or '`hs ?`'.

The second line tells PAUP* to conduct a heuristic search with branch swapping conducted using nearest-neighbor interchanges. With default settings, the starting tree is attained by stepwise addition with a simple addition sequence.

After the search has completed, read through the output. Notice that search settings are explicitly stated (this would be a good time to confirm that you told the program what to do

correctly) and you can get a basic rundown of the entire analysis. Furthermore, you can see basic information about the results of the search.

Answer assignment question 1.

Next, we'll do a moderately rigorous search using the defaults. Type the following.

```
Set criterion=parsimony maxtree=100 increase=auto;
hs swap=tbr;
```

Here, we'll allow PAUP* to begin by storing up to 100 trees in the tree buffer, and allow the program to increase that limit as needed (if > 100 equally parsimonious trees are found). Again, the second line initiates a heuristic search, but now with branch swapping conducted using tree bisection and reconnection, the most rigorous branch-swapping technique in terms of computational time, but also the most likely to shorter alternative solution than NNI.

Answer assignment question 2.

This represents the level of rigor one would attain by simply running a parsimony analysis with PAUP* at its **default settings**. In this search, we kept multiple equally parsimonious trees at every stage, but only the stepwise addition tree(s) generated from a single addition sequence as the starting tree(s) for branch swapping.

Now, we'll search the tree space a little more rigorously. Type the following:

```
hs addseq=rand nrep=500;
```

Since maxtree, increase, and swap are persistent settings, they're kept the same as above. Now we're conducting 500 replicate heuristic searches, each starting with a stepwise addition tree generated using a different random addition sequence.

The tree-island profile reported at the end of the search provides some details about the parts of treespace that have been explored in this search. For example, the score of best trees from a particular island and how many times (out of the 500 reps) that a particular island was found during branch swapping.

Answer assignment question 3, 4, and 5.

Lab 2

Biology 545: Phylogenetics

Name: _____

1. What was the length of the tree found by the first search?

2. What was the length of the tree found by the second search?

3. Did you find a better tree with a more rigorous search?

4. There should be a large number of tree islands in this dataset. For the best island, how many addition sequence replicates hit it?

5. Based on this analysis, make a statement about your confidence with respect to finding the globally shortest tree.