MATH 660
Allman

An Introduction to PAUP*

This tutorial introduces PAUP*, Phylogenetic Analysis Using Parsimony (* and other methods), which is one of the most commonly used phylogeny inference programs. Developed by David Swofford, now of Duke University, you can download it from the web `http://paup.phylosolutions.com/`. Make sure you download the manuals too. Generally speaking this website is an excellent source of information for beginner users.

Before beginning: While this is not a requirement, you might consider downloading one of the many 'tree viewer' programs from the internet. Popular programs include `TreeView`, `Dendroscope`, and `FigTree`. A simple Google search will lead you to the webpage for any of these programs, and downloading and installing is easy. The latest version of PAUP* now has a GUI interface that saves trees to pdf files.

Starting:

PAUP* takes as input a text file in *nexus* format. These files typically contain data such as sequences, commands indicating how the data should be processed, and perhaps output saved from previous sessions. A nexus file is organized into *blocks*. For example, a tiny sample nexus file that would be useful for solving Problems 2 and 3 from Chapter 3 in the class notes is included at the end of this tutorial. Please refer to this sample nexus file while reading the explanations of the various blocks below.

- the **TAXA** block

  This self-explanatory block contains taxa labels and the number of taxa.

- the **CHARACTERS** block

  This block includes the data matrix. The number `nchar` of characters must be set here, as well as commands to indicate what sort of data will be analyzed (*eg.* `datatype = dna`).

- the **ASSUMPTIONS** block

  We will use this block to enable us to undertake weighted parsimony searches. For example, here weight matrices for a 1:2 and 1:3 transition:transversion weighting scheme are defined. These matrices are called `2_1` and `3_1` respectively.

- the **TREES** block

  If you want to evaluate some optimatility criterion (parsimony, likelihood) on specific user-defined trees, then the trees block allows you to 'read' these trees into PAUP*. You will need to understand *Newick* format for trees (named after a lobster restaurant in New Hampshire). Fortunately, it's fairly intuitive, using parentheses to indicate clades of taxa. PAUP* has the somewhat annoying feature that it numbers the input trees, and despite what the manual may say, you can not assign a name to your favorite tree.

  Pay careful attention to the Newick syntax; in particular, note the semicolons ending commands.

- the **DISTANCES** block

  In the sample files for the class, you will find one named *class_dist_example.nex* and one defined *user-defined-dist.nex*. These files show sample distance blocks and should be enough to get you started. The GUI menu has good options too.

- the **PAUP** block

  The PAUP block contains all the execution commands. You will need to set the tree selection method (parsimony, nj, likelihood, etc.) and instruct PAUP* how to carry out its searches. For beginners, the best way to learn is by following examples. Several examples will be given below; for completing your homework exercises, the PAUP block *must* be modified appropriately to compute the quantities of interest. Most of the commands described below belong in the PAUP block.

Generating and displaying trees, basic commands:

Below are listed some basic PAUP commands, ones that you will need to generate a set of trees, save a tree to a file, start some search, etc. All such commands should be placed in the PAUP block and then pasted into the lower box on the portal interface. Repeat: These commands go into the **command** file.

- Start and end a log file to keep a record of your commands. There is also a menu option to start and end logging a session. Before you do this, make sure you have set PAUP* to your working directory. `File -> Change Default Directory`

  ```
  log start;
  log stop;
  ```

- Random tree generation

  `GenerateTrees all;` This command generates all unrooted trees for the active data.

  `GenerateTrees random;` generate 100 random trees (for large data sets).

  `GenerateTrees random NTrees=`k`;` generate k random trees.

- Automatic searching

  The two simplest commands for finding parsimonious trees are

  `AllTrees;` Searches through all trees for the most parsimonious ones. This may take too long to be feasible for large number of taxa, but it's entertaining to try.

  `hsearch;` Performs a heuristic search to find the most optimal trees. There are many options to control how this is done. To see default settings, use `Hsearch ?`.

  `hsearch nbest=5;` Performs a heuristic search and saves the 5 best in memory.

- Get information about your data matrix and assumptions.

  `cstatus;` Give the status of your active character data. If you execute this command after loading and executing a data file, for example, you will see that the default optimality criterion is set to parsimony.

  `tstatus;` Get information about currently included taxa in analysis.

- Display trees

  `showtrees;` This command displays tree 1.

  `showtrees all;` displays all trees stored in memory.

  `showtrees k;` displays tree k.

  `showtrees k-l;` displays trees k through l

  `describetrees;` This command displays trees with lots of additional information. Use the same tree list options described above.

  ```
  describetrees 1-5 /plot=cladogram;
  describetrees all /plot=phylogram brlens=yes;
  ```

- Save trees to file; upload trees from file.

  `savetrees file=`*treefile*; (Replace *treefile* with the name of your choosing.)
  `gettrees file=`*treefile*; (Only useful if you use the command line version of PAUP*.)

- Removing or restoring particular taxa to the analysis.

  `delete` *taxonname*; Remove a taxon (or set of taxa) from the analysis.

  `undelete` *taxonname*; Restore a previously deleted taxon (or set of taxa) to the analysis.

- Additional helpful commands

  `help;`
  `showmatrix;` Shows the data matrix.

Parsimony for small trees:

Because the number of unrooted binary trees relating n taxa grows so quickly, complete searches for the most parsimonious tree are not possible unless the number of taxa is relatively small. For large n, only heuristic searches are possible and you should use the command `hsearch;`.
Once you have a handful of trees in memory, then a most useful command is

- `pscores;` Compute and display the parsimony scores of all trees currently in memory.

To see how to use PAUP* to compute the weighted parsimony score of a tree, where the weight are given according to some user-defined weight matrix, see the sample nexus file.

Neighbor Joining and distance methods:

For a phylogenetic analysis using distance methods, you might use (any, all, a subset of) the following commands in a PAUP block.

- `set criterion=distance;` This is absolutely necessary if you intend to undertake any type of distance method (*e.g.* least squares fitting) that requires searching over all trees. The default analysis in PAUP* is parsimony.

- `showdist;` Display the distance matrix. PAUP* calls the Hamming distance by another name, the *uncorrected p-distance*. Other distances are possible, and perhaps better, but we need to develop our models of site substitution first. The 'uncorrected' refers to the fact that the Hamming distance does not correct for phenomenon like back substitutions or multiple substitutions on a branch.

- `upgma;`

- `nj;`

- For a least squares fit, the following lines suffice:

  `set criterion=distance;`
  `dset objective=lsfit power=2;` 'dset' stands for 'distance methods settings'
  `hsearch nbest=5;`

Exploring Primate data:

Now let's explore PAUP* with a data set that relates twelve primate mitochondrial DNA sequences. This is an interesting data set in that phylogenetic trees constructed from this molecular data helped solidify the current view of the relationship of humans to other primates. This data is distributed with PAUP* (and other packages) as `primates-mtDNA.nex`. A modified version, in which the scientific taxa names are replaced by their more familiar common names, is available on the course website using the same name.

Exercises:

Please ***only hand in solutions to the numbered exercises***; the others are to get you acquainted with using PAUP*, and start you thinking about the right things.

With your favorite text editor, open `primates-mtDNA.nex` and stare at the nexus file. There are many commands that we did not discuss, but this gives a good idea what a 'typical' data file might look like. Also, note that two taxon sets have been defined, a group of 'hominoids' and a group of 'others'.

On the full dataset:

- Generate 5 random trees, and calculate their parsimony scores. How many informative sites are this in this data? View the most parsimonious of the five trees. How well do you think it describes the evolutionary history of the primates?
- Generate 100 random trees and calculate their parsimony scores. Which tree (or trees) are 'best'? Are you satisfied with the tree with the smallest parsimony score?
- How many trees would you have to consider, if a full parsimony analysis were undertaken for these 12 taxa?

1. **Heuristic parsimony search:** Use `hsearch` to find what are probably the most parsimonious trees. Save the three most parsimonious trees. Look at them. Based on your analysis, which primate is most closely related to humans? Are you satisfied the trees you found describe the evolution of these sequences?

   Now repeat this search with a 1:2 transition/transversion weighting for state changes. Compare these trees to the ones you found using unweighted parsimony. Comment.

2. **Distance methods:** Use PAUP* to compute and display the Hamming distance matrix for the 12-taxa dataset. Then find and display the UPGMA and the NJ tree. Save the UPGMA and NJ trees to a file and include them with this assignment.

With a subset of the taxa:

- Here you will examine a data set with only the 'hominoids': human, chimpanzee, orangutan, gorilla, and gibbon, along with lemur and tarsier for outgroups. The commands

  ```
  delete others;
  undelete lemur tarsier;
  tstatus;
  ```

  accomplish this and must be added to your PAUP block.

3. Try `AllTrees` to perform an exhaustive search of all possible tree topologies. Record the parsimony scores for the ten best trees. Based on your analysis, which primate is most closely related to humans?

   Now repeat this analysis for these seven taxa using 1:2 weighted parsimony.