

## A QUICK GUIDE TO GET YOU STARTED USING *PAUP\**(*PHYLOGENETIC ANALYSIS USING PARSIMONY \*AND OTHER METHODS*)

This handout is meant as a practical guide to get you started using the computer package *PAUP\** [1] for building evolutionary trees from DNA sequence data. This guide is no substitute for reading the online manual, which is full of valuable theoretical background as well as detailed information about how to use the program. Some of the features and options available in *PAUP\** are not appropriate for analyzing sequence data, but rather are there for other types of character data (e.g., morphological). Conversely, some features apply only to sequence data.

Before the program will accept your data file, it must be in a fairly strict format, as illustrated in the boxed example below. The **sample data** files in *PAUP\** contain several examples of formats that will be accepted, and the **help menu** explains what most of the statements mean. Some examples and explanations are included in this handout. This should be enough to get you started using *PAUP\**; your understanding will grow as you become more familiar with the program.

### DATA BLOCK:

The most important thing for you to get correct initially is the **data block**. You must put semicolons (;) at the end of each statement and at the end of the data block (after **endblock;** or **end;**). Your data block should look something like the DNA sequence example below. In the first case under the **assumptions block**, transversions (Tv) are weighted twice transitions (Ti), since they are rarer evolutionary events, and gaps (-) are weighted 4 times transitions. In the second case, Tv are weighted 10 times Ti, and gaps are not counted as evolutionary events (in which case, **gapmode** should be "missing", not "newstate"). Choose an appropriate weighting scheme.

```
#NEXUS
[You may put notations to yourself in brackets. Those items in
brackets are "unseen" by the actual computer program, but will
appear on your output for documentation purposes.]

begin data;
  dimensions ntax=4 nchar=7;
  format datatype=dna gap=- missing=?;
  options gapmode=newstate;
  matrix
Yourtaxa1 gatatga
Yourtaxa2 aatacct
Yourtaxa3 tttgcta
Yourtaxa4 ?tagtta
;
endblock;

begin assumptions;
  usertype 2_1_4 = 5 [weights Tv 2X Ti, and gaps 4X Ti]
      a c g t -
[a] . 2 1 2 4
[c] 2 . 2 1 4
[g] 1 2 . 2 4
[t] 2 1 2 . 4
[-] 4 4 4 4 .
;
  usertype 10_1 = 4 [weights transversions 10 times
                    transitions]
      a c g t
[a] . 10 1 10
[c] 10 . 10 1
[g] 1 10 . 10
[t] 10 1 10 .
;
endblock;
```

**DIMENSIONS statement:**

It is critical for you to get the following values correct. If they are not correct, your data file will not "**execute**" correctly, and you will get an error statement that may (or may not) make sense to you. If your data file will not execute, check these values carefully.

**ntax**=[The **number of taxa** — *i.e.*, sequences — in your data matrix.]

**nchar**=[The **number of characters** in each of the sequences. This number should be the same for each taxon. If there is missing sequence at the beginning or end of some sequences, code these characters as "unknown" or "missing" with question marks (?), not as gaps.]

**FORMAT statement:**

Under the **format** statement there are several options that will work. For the parameter **datatype**, it should read **protein**, **DNA**, or **RNA**. You can define **gap**, **missing**, and **matchcharacter** as almost anything that you want, so long as it does not conflict with your data matrix or a statement elsewhere in the program (for example, you can't use a ;). Given above are some common and useful examples: it is easy and conventional to code gaps as dashes (-), missing data as question marks (?), and matching characters as periods (.).

**OPTIONS statement:**

In the above example, **gapmode=newstate** means that each gap (*i.e.*, each character that has a -) in the sequence is treated as a phylogenetically relevant character (*i.e.*, as an individual, real insertion or deletion, that can be treated as parsimony-informative). If you do not want gaps to have this much phylogenetic importance, you can make **gapmode=missing**, which will cause the gap characters to be ignored during the calculations (but not the other characters at this site).

**Data MATRIX:**

The name of your taxa (**Yourtaxa1**, etc., above) can be as long and complicated as you like, but if you include punctuation marks or spaces in the name, you must enclose the name in single quote marks like 'this'. The exception is if you use an underline to separate the parts of the name, such as **Genus\_species**. The full name you put here will appear on the output and on the terminal nodes of the tree, thus you should make these easily understood, but not excessively long. Names are longer than about 20 letters the output can look messy or be truncated, so it is best to use short (about 10 letters or less), but informative, names. You should then put a minimum of one space after the taxon name before starting the sequence characters.

It is best (for your sanity) if you display the sequences aligned with respect to each other, even though the program does not require this (see the help menu about #NEXUS format if you want an explanation). **PAUP\*** will accept a very long sequence on each line. Alternatively, you can "interleave" the taxa; if you do this, you must have the statement "**interleave**" under **format**. See the **PAUP\*** sample data file "Hominoid mtDNA" for an example of interleaved data.

**ASSUMPTIONS:**

There are many complex options that you can put under the assumptions block. Options placed here can be turned on or off from the menu bars while running the program, so you can create various options that you may wish to use later. For example, in the case above you can weight transversions:transitions 2:1, 10:1, or turn off the weighting option (that is, they would be weighted 1:1). You can also make additional weighting matrices using different values and choose which one to use during any particular run. See the **PAUP\*** sample data files for examples. For protein sequences, it is a very good idea to use a Protpars matrix (or other weighting scheme of your choice), which weights the various possible amino acid replacements according to the minimum number of nucleotide substitutions that separate them in the genetic code. The use of such a weighting matrix during parsimony analysis greatly increases the *accuracy* of the parsimony method as compared to unweighted parsimony. Indeed, weighted parsimony often performs almost as well as maximum likelihood methods, without as great of an increase in computational time.

**To run PAUP\*:**

(This assumes that you have PAUP\* loaded properly on a Macintosh computer. Other systems may require slightly different menu commands or require line-driving.)

1. Open the program PAUP\* by double clicking on the icon.
2. Choose "NEW" under the column "file" by dragging the mouse to that position. This will open a new file for you. Enter your information into a data block, as shown in the above example. You can COPY and PASTE from the sample data files and from any data files you may have made. You can also "IMPORT" or "OPEN" a file from the "sample files" folder to play with.
3. "SAVE" your PAUP\* input file under a relevant name (to you) such as "yourfile".
4. Choose "EXECUTE yourfile" under the "FILE" menu column. If you have set up the data block properly you should get a message "PROCESSING OF FILE 'YOURFILE' IS COMPLETED". If so, you are ready to "build" trees. If not, go back and fix what you did wrong and try again until you are successful. [You may need to examine the "sample data files" to discover your error. The most common problems usually have to do with not stating the correct number of taxa or characters, or naming taxa improperly. Please explore these possibilities thoroughly before asking for help.]
5. Once you have successfully "executed" the file, there are many options for you to choose.
  - a. It is a good idea to select "OUTPUT TO DISK" under the "file" column, so your manipulations will be saved to a file on the computer.
  - b. If you have known outgroups (species or gene duplicates), it is a good idea to "define outgroups" under the "data" menu or "Rooting" under the "options" menu. This can be done after the tree search completes if you are using unpolarized characters (which is true for sequence data), since the placement of the root has no effect on the overall tree length.
  - c. You will need to choose the type of analysis you would like to perform under the "Analysis" menu (e.g., Parsimony, Likelihood, or Distance). Set appropriate parameters for your dataset under the "Settings" option for each method (doing this part appropriately and well requires thought and experience).
  - d. If you have few enough taxa (generally less than about 10-12 for PROTPARS; if using DNA sequences often more can be handled), you can do an "EXHAUSTIVE SEARCH" under the "search" column. Exhaustive searches are exact algorithms and will give you a frequency distribution of all possible trees, which can be very informative.
  - e. You can view the trees by the "SHOW TREES" or "DESCRIBE TREES" selections under the "trees" column. "DESCRIBE TREES" gives you more options for tree descriptions. You should explore these options. But be warned: some of them will give you massive files of information that you will not need each time. (You can stop saving to disk under the "OUTPUT TO DISK" option, if you want.)
  - f. You can view and print the tree(s) under "Print trees" option of the "Trees" menu. You can also save the tree as a PICT file that can be opened and edited for presentation.
  - g. There are numerous other options (such as Bootstrap analysis, statistical tests of trees, etc.) that you will want to explore in PAUP\*. This package is currently the easiest and most versatile phylogenetic tree inference package available, but even it has its limits. For example, at present PAUP\* does not support maximum likelihood models of protein evolution.
6. When you are through, "QUIT" the program (under the "file" column).
7. If you want a printed copy of your output, you can print it directly after opening your **.out** file. However, you can get a better-looking copy if you open the file under a word processing program, edit it (by changing the font to Courier 10 or similar font), and print it from this program.

AGAIN, THIS IS JUST A SHORT GUIDE TO PAUP\*. Explore this wonderful program, and have fun!  
 NOTE: This guide does not represent an official endorsement of this program by Caro-Beth Stewart (SUNY-Albany) or the NIH, neither of whom have a commercial interest in this product.

[1] Swofford, D.S. (1998) *PAUP\*(PHYLOGENETIC ANALYSIS USING PARSIMONY \*AND OTHER METHODS)*, Sinauer Associates, Sunderland, MA.  
<http://www.sinauer.com/Titles/frswofford.htm>