



## PHYLIP

### Phylogenetic analysis workshop

The purpose of this course is to demonstrate how to run the phylogenetic analysis software and some of the options that are available. This workshop is not intended as course in phylogenetics, and due to the short time available, phylogenetic concepts will only be discussed briefly.

Phylip is a comprehensive phylogenetic analysis package created by Joseph Felsenstein at the University of Washington. This package can do almost all phylogenetic analysis available in the literature today. **Phylogenetic analysis depends on an accurate multiple sequence alignment.** The best tool for sequence alignment is the PileUp module of the GCG programs. However, no program is perfect and it may require considerable effort in adjusting the alignments before a satisfactory result is obtained. (See the Multi-sequence alignment workshop for more details.)

### Starting PHYLIP

DRAWTREE and DRAWGRAM need a fontfile in the same directory where they are executed.

To create a fontfile in your local directory type

```
cp /seqprg/phylip/font1 fontfile
```

If you want to select alternative fonts the following are available:

font1	(simple sans-serif Roman)
font2	(medium quality sans-serif Roman)
font3	(high quality serified Roman)
font4	(medium quality sans-serif Italic)
font5	(high quality serified Italic)
font6	(Russian Cyrillic)

The PHYLIP analysis package operates within the same molecular biology shell as GCG. To activate the shell type

```
initmb
```

### Converting sequences between GCG and PHYLIP

#### Manual formatting

Any multi-sequence alignment may be manually reformatted with a text editor such as jove or even Microsoft word. The format requirements for PHYLIP are very stringent and any deviation will result in a program that hangs, usually with the error message "Unable to allocate memory."

The file must conform to the following:

**Only** non-ambiguous sequence characters are allowed, e.g. A, T, G, and C for DNA sequences.

All characters **must** be in the upper case.

Sequence names **must** be **exactly** 10 characters long.

Gaps in sequences **must** be indicated with "-" or "?"

The **exact** number of sequences must be in line 1.

The **exact** number of bases in the sequence must be in line 1.



To reformat an MSF file produced by GCG proceed as follows:  
 Run readseq by typing  
 readseq filename

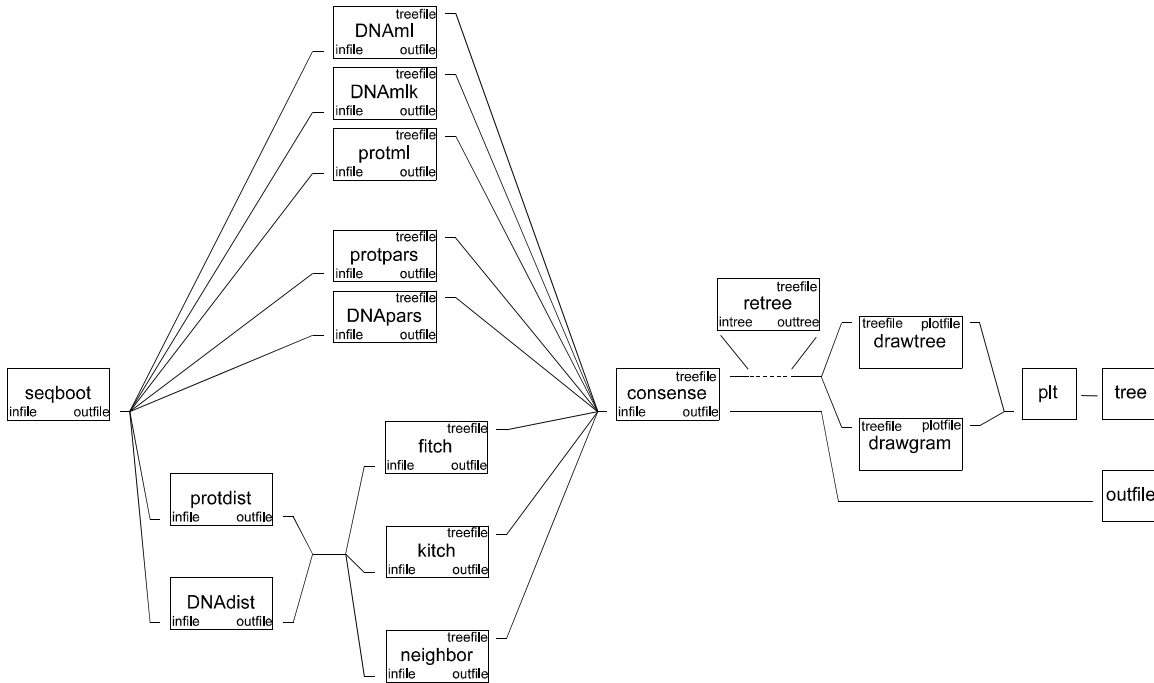
Reply to all the prompts.

Choose as output file format option 12, PHYLIP.

Hint: If you have formatted a set of sequences manually and it does not work, reformat them again with readseq. This will often solve the problem.

## Running PHYLIP modules

Most PHYLIP programs run in the same way. The input for a program is taken from a file called “infile” and the results are written in a file called “outfile” Some programs may write both “outfile” and a file called “treefile”, or “plotfile”



This table represents an approximate “road map” on how some of the modules can be used together. Each module contains several options, so the number of possible types of analysis is almost limitless. This large number of options and the fact that there is no consensus on what type of analysis is most appropriate, makes any phylogenetic analysis a daunting prospect.

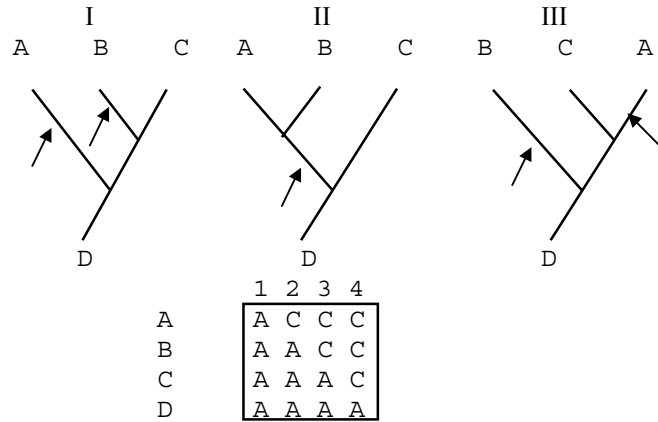
Sequence analysis can be divided into two main types: discrete character, based analysis and matrix analysis.

## Character Analysis

There are two main methods of analyzing character based data, parsimony and maximum likelihood.

### Parsimony analysis

Parsimony analysis carried out by DNAPARS and PROTPARS is based on the concept of the least number of changes that can resolve a tree. For the following sequences there are three trees possible.



In the above example positions 2 and 4 are easily explained by a change in a single branch. Position 3 can be explained by two changes in tree I and tree III or a single change in tree II (indicated by the arrows). Since the changes in tree II is the smallest, it is considered the most parsimonious and therefore favored. Position 3 is called an indicative site. Only indicative sites are considered in parsimony analysis.

### Maximum Likelihood Analysis

This analysis uses statistical tables to calculate the likelihood of a change occurring in a particular position. It considers all positions and also includes the likelihood of a change not occurring, for example an A changing to another character and then back to an A.

### Matrix Analysis

A matrix analysis creates a matrix consisting of the number of differences between the sequences. There are several models that score the differences between sequences, these models vary in the weight they assign to different substitutions, for example how heavily transversions should be weighed against transitions in DNA sequences. Since distance matrix methods sum all the data there is an “averaging” effect and the results are relatively stable and unaffected by bootstrapping. All changes, not only indicative sites, are taken into account. Distance matrix can also produce a tree with branch lengths proportional to the evolutionary distances between sequences.

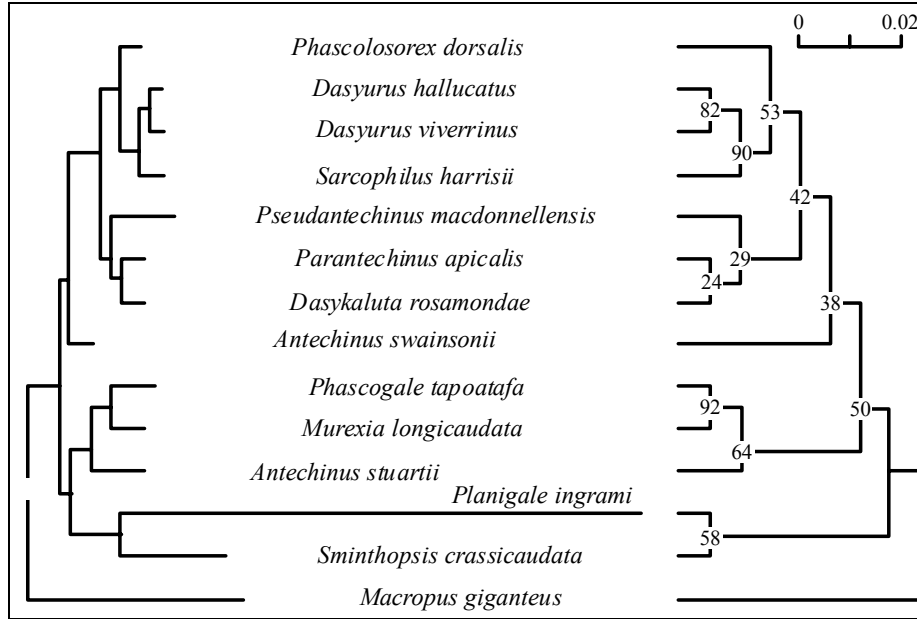
### An example of a distance matrix

14

	dorsa	harri	viver	apica	rosam	macdo	swain	longi	stuar	tapoa	crass	grey	pingra	mole
dorsa	0.0000	0.0145	0.0161	0.0195	0.0220	0.0161	0.0229	0.0377	0.0378	0.0378	0.0596	0.1708	0.1345	0.2669
harri	0.0145	0.0000	0.0112	0.0195	0.0237	0.0177	0.0280	0.0428	0.0412	0.0446	0.0579	0.1683	0.1384	0.2677
viver	0.0161	0.0112	0.0000	0.0227	0.0220	0.0128	0.0296	0.0443	0.0445	0.0479	0.0631	0.1757	0.1465	0.2773
apica	0.0195	0.0195	0.0227	0.0000	0.0116	0.0112	0.0281	0.0292	0.0260	0.0309	0.0436	0.1643	0.1316	0.2617
rosam	0.0220	0.0237	0.0220	0.0116	0.0000	0.0100	0.0325	0.0304	0.0271	0.0339	0.0328	0.1458	0.1286	0.2408
macdo	0.0161	0.0177	0.0128	0.0112	0.0100	0.0000	0.0310	0.0321	0.0321	0.0354	0.0533	0.1654	0.1377	0.2678
swain	0.0229	0.0280	0.0296	0.0281	0.0325	0.0310	0.0000	0.0144	0.0112	0.0178	0.0559	0.1615	0.1335	0.2758
longi	0.0377	0.0428	0.0443	0.0292	0.0304	0.0321	0.0144	0.0000	0.0094	0.0126	0.0566	0.1674	0.1389	0.2784
stuar	0.0378	0.0412	0.0445	0.0260	0.0271	0.0321	0.0112	0.0094	0.0000	0.0126	0.0550	0.1653	0.1391	0.2788
tapoa	0.0378	0.0446	0.0479	0.0309	0.0339	0.0354	0.0178	0.0126	0.0126	0.0000	0.0621	0.1692	0.1394	0.2747
crass	0.0596	0.0579	0.0631	0.0436	0.0328	0.0533	0.0559	0.0566	0.0550	0.0621	0.0000	0.1760	0.1298	0.2591
grey	0.1708	0.1683	0.1757	0.1643	0.1458	0.1654	0.1615	0.1674	0.1653	0.1692	0.1760	0.0000	0.2556	0.2480
pingra	0.1345	0.1384	0.1465	0.1316	0.1286	0.1377	0.1335	0.1389	0.1391	0.1394	0.1298	0.2556	0.0000	0.3840
mole	0.2669	0.2677	0.2773	0.2617	0.2408	0.2678	0.2758	0.2784	0.2788	0.2747	0.2591	0.2480	0.3840	0.0000

## Building trees

The results obtained in matrix analysis is very dependent on the tree building program used. Both the Fitch and Margoliash and the Neighbor-joining protocols are available in the FITCH and NEIGHBOR programs respectively. You are strongly urged to try both methods and see which one is more appropriate for your dataset.



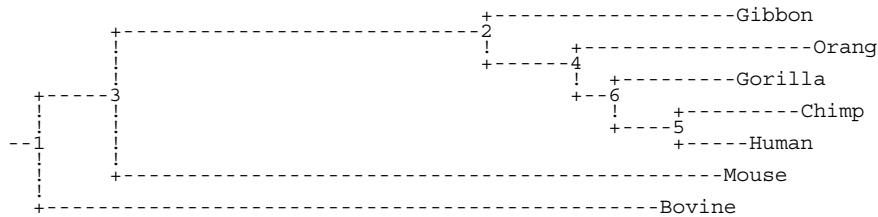
The tree on the left is a distance matrix tree and the tree on the right is a parsimony tree.

## Interpreting Trees

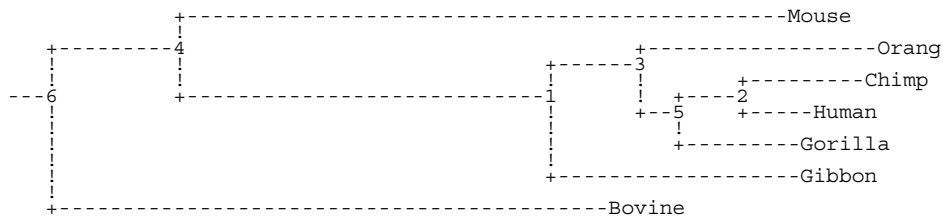
A tree "grows" from left to right. The numbers at the forks are arbitrary and are used (if present) merely to identify the forks. In some of the programs asterisks ("\*") are used instead of numbers. For many of the programs the tree produced is unrooted. It is printed out in nearly the same form, but with a warning message: "remember: this is an unrooted tree!"

Trees can be rearranged without changing the interpretation of the data. The following two trees are identical

Tree 1



Tree 2



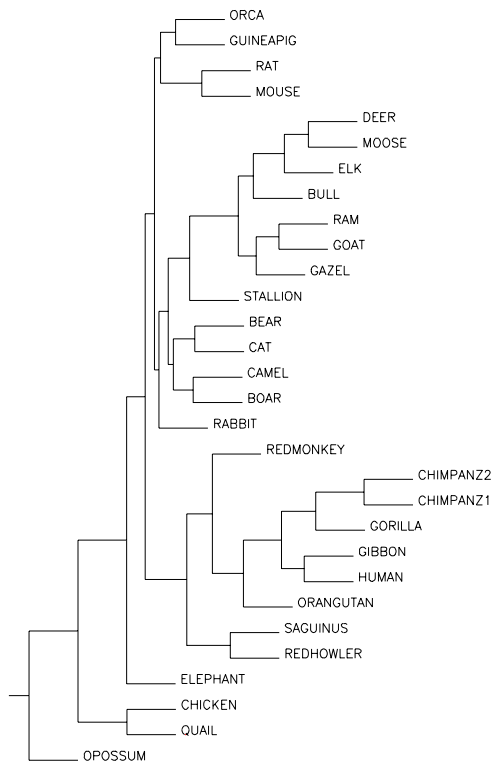


## Bootstrapping

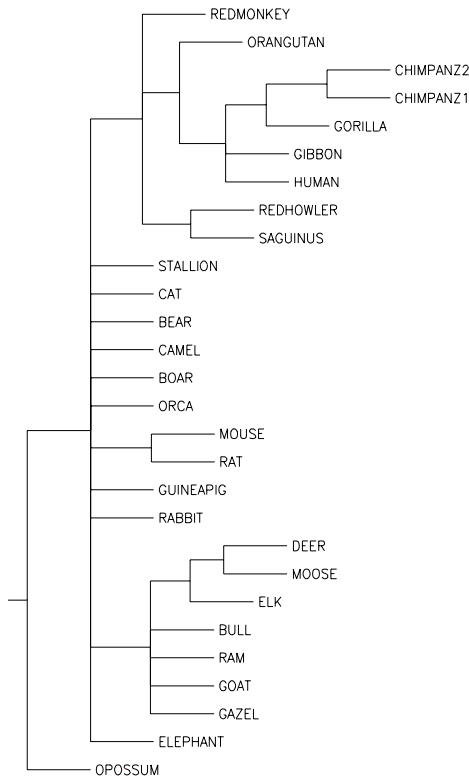
Bootstrapping involves the random resampling of the dataset. In practice SEQBOOT creates 100 copies of a sequence, each of which has some data randomly removed and some data randomly inserted to maintain the size of the dataset. The principle of the analysis is simple. If the association between two sequences is robust, randomly changing some of the bases would not affect this association. On the other hand, if an association is tenuous, changing some bases would have a dramatic effect on the association.

In the following example the value of bootstrapping is illustrated. The tree on the left is shown without bootstrap values. Some of the associations in the tree are extremely unusual. Human's closest relatives are the gibbons, a primitive arboreal monkey, and orcas are placed with the rodents. These artifacts are due to a very unstable tree produced by the protein sequence of a small gene. Without the bootstrap values there is no way of telling how stable the tree is. The tree on the right is produced from the same data, but all nodes with bootstrap values below 50 were collapsed into polytomies. (This is simple to do by editing the tree file with your favorite editor.) Note that the very distant birds were also removed and that a marsupial was used as an outgroup. It is now clear that the tree is largely unresolved and that the unusual associations were due to a lack of resolution.

Tree 1



Tree 2



## Tree 1

```
(((((ORCA:100.0, GUINEAPIG:100.0):30.3, (RAT:100.0, MOUSE:100.0):83.7):12.5, (((((DEER:100.0, MOOSE:100.0):49.1, ELK:100.0):62.8, BULL:100.0):30.9, ((RAM:100.0, GOAT:100.0):46.6, GAZEL:100.0):36.2):99.0, STALLION:100.0):43.8, (BEAR:100.0, CAT:100.0):43.3, (CAMEL:100.0, BOAR:100.0):40.1):10.6):19.0, RABBIT:100.0):9.0):19.1, (REDMONKEY:100.0, (((CHIMPANZ2:100.0, CHIMPANZ1:100.0):98.0, GORILLA:100.0):69.9, (GIBBON:100.0, HUMAN:100.0):46.2):77.6, ORANGUTAN:100.0):63.9):52.3, (SAGUINUS:100.0, REDHOWLER:100.0):89.1):84.3):38.0, ELEPHANT:100.0):99.0, (CHICKEN:100.0, QUAIL:100.0):100.0):100.0, OPOSSUM:100.0);
```

## Tree2

```
((REDMONKEY:100.0, (ORANGUTAN:100.0, ((CHIMPANZ2:100.0, CHIMPANZ1:100.0):96.3, GORILLA:100.0):63.8, GIBBON:100.0, HUMAN:100.0):73.1):58.9, (REDHOWLER:100.0, SAGUINUS:100.0):76.7):81.8, STALLION:100.0, CAT:100.0, BEAR:100.0, CAMEL:100.0, BOAR:100.0, ORCA:100.0, (MOUSE:100.0, RAT:100.0):96.3, GUINEAPIG:100.0, RABBIT:100.0, ((DEER:100.0, MOOSE:100.0):53.2, ELK:100.0):62.6, BULL:100.0, RAM:100.0, GOAT:100.0, GAZEL:100.0):95.0, ELEPHANT:100.0):100.0, OPOSSUM:100.0);
```

# Considerations when doing phylogenetic analysis

## Selecting an outgroup

An outgroup greatly effects the result of building a tree. An outgroup is the marker against which the sequences are analyzed. Usually an outgroup is the closest, distant relation to the selected dataset. The closer an outgroup is the sequence you are analyzing the more accurate your results will be, while you have to be sure that the outgroup does not actually belong to your group of sequences. Examples of outgroups would be platypus (a primitive egg-laying mammal) for a group of mammals, or mouse as an outgroup for a group of primates.

## Rooting trees

You can only use the rooted tree option when you know the ancestor of your sequences. That sequence can form the root of your tree. Even though an outgroup serves as a root in the construction of the tree it is not a rooted tree in the true sense of the word.

## Protein or DNA sequences

The choice between DNA and protein sequences for analysis is often not clear. The evolutionary pressure for a gene occurs at the protein level. The coded proteins are therefore the first choice to study the associations between genes. This is particularly appropriate for looking at genes over long evolutionary distances. Over shorter periods, there are often not enough information in the protein sequences to resolve the trees and it is more appropriate to analyze the DNA sequences.

## Saturation of mutations

The more changes there are between the sequences in a group, the more indicative sites there are and the fore information is available to analyze. However, as the number of mutations increase, the chances increase that a character may change to another state, just to be changed back to its original state. This is usually manifest is a decrease in the number of mutations that occur over a period of time. You should be aware of this possibility and it may be necessary to correct for this.

## Gaps

Gaps in sequences represent one of the most intractable problems in phylogenetic analysis. In most programs in this package gaps are treated as a fourth character. As far as the algorithm is concerned, in a DNA sequence there are A, T, G, C, and -(gap) characters. This works well for small gaps. However, large gaps may be introduced by a single event. Say, for example, a single event produces a gap 10 characters long. The program will interpret this as 10 separate events. Any two sequences that share large gaps will be associated very strongly even though they may actually share relatively few mutational events.

There is no easy cure for this problem. Looking at the effect that gaps have on your trees should alert you to any problems. In parsimony analysis you can manually reduce the dataset to indicative sites only and score gaps as a single event.

Note that uneven sequence lengths will also be scored as gaps. All sequences analyzed should be of the same length.

## Tutorial

The following tutorial is an exercise to familiarize you with the way the program modules work. There are many different programs and this tutorial selected only one type of analysis. The best analysis for your sequence will depend on the gene itself, your preferences and the traditions in your field.

### Initialize the Molecular Biology shell

To start the shell type

```
gcg
```

### Set up your directory

Create a directory to work in by typing

```
mkdir tutorial4
```

Now move there with

```
cd tutorial4
```

Copy the font file to the new directory with

```
cp /seqprg/phylip/font1 fontfile
```

You can use any MSF file, such as those created by PILEUP for this exercise. A file called marsupial.msf is provided and you can fetch it by typing

```
fetch marsupial.msf
```

The final formatting must be carried out with a text editor, e.g. Jove. Type

```
jove infile.ori
```

Replace any “.” gap characters with “-”.

To do this type

```
<esc> <ctrl>e and answer with .<return> and then -<return>)
```

In some sequence file it may be necessary to:

1. Replace “~” characters with “?” or “-“ Usually necessary when files were created in SEQLAB.
2. Replace any ambiguous characters with non-ambiguous characters or gaps.
3. Replace any lower case characters with upper case characters.

Save the file with <ctrl>x s

Quit Jove with <ctrl>x <ctrl>c

## Reformat the MSF file

Run readseq by typing  
readseq marsupial.msf

Respond to “Name the output file” with infile.ori<return>

Respond to “Choose an output format” with 12<return>

Respond to “Choose a sequence” with all<return>

## Bootstrapped parsimony analysis

The first step is to generate 100 sequence samples

Copy the file you want to analyze to infile as follows

```
cp infile.ori infile
```

NOTE: Never keep valuable data in the infile, outfile, or treefile. These files are overwritten often and you will lose the data.

```
seqboot
```

Supply a random number, e.g. 9<return>

Accept all the defaults with y<return>.

Copy the outfile of seqboot to the infile for the next program

```
cp outfile infile
```

Depending on your account, you may have to press y<return> to overwrite the file

Run the parsimony program as follows

```
dnapars
```

Change the outgroup to 12

Set “Analyze multiple datasets” to “Yes” for 100 datasets.

This program produced 100 trees.

You need to know the position of your outgroup. Type

```
more treefile
```

We want to make “grey” the outgroup. Make a note of its position in the first tree. (Hint: touch **q** if you want to get out of **more**)

You must now calculate the consensus for all the trees. To do this type

```
cp treefile infile
```

```
consense
```

Set the following options:

“Change Outgroup root?” to “Yes, at species number 1”

Accept all the other defaults.

The results of the analysis is written to an outfile. You can inspect you results by typing

```
more outfile
```

Note: You should keep a copy of the outfile. The bootstrap values are not written to the graphic tree plot and you will have to include them manually.



There are many options to set here and you may want to experiment to find what works best for you.

When the selections are completed type “y” to accept the defaults.

If your default printer is set, you can plot the file to the default by typing  
`lpr plotfile`

If your default printer is **not** set you can direct to printing to any printer, e.g. achs\_l3 by typing  
`lpr -Pachs_l3 plotfile`

If you want to manipulate the figure in a graphic program, copy the plotfile to your local directory with Rapidfiler or FTP. Remember, only HPGL files will be accepted by most programs.

If you have time, edit the treefile and collapse all polytomies less than 5.

For the example given above, the treefile should look as follows:

```
(grey:10.0,(((crass:10.0,pingra:10.0):10.0,((macdo:10.0,((dorsa:10.0,harri:10.0):7.0,
viver:10.0):6.7,(apica:10.0,rosam:10.0):7.0):5.9,((longi:10.0,(tapoa:10.0,stuar:10.0):7.5):5.5,
swain:10.0):9.0):6.9):10.0,mole:10.0):10.0);
```

## Viewing the tree

If you are working on an Xwindows terminal, you may view the plotfile with **ghostview**.

Start **ghostview** by opening a terminal window (Do **NOT** start gcg!) and type **ghostview**. (This assumes your display options are set the same as for seqlab)

Select **File** and then **Open...** Navigate to the plotfile, select it, and click on **Okay**.

The tree will now be displayed in the window.

## Documentation

### On-line

The complete documentation for these programs is available from the ACHS molecular biology web page <http://www.med.virginia.edu/achs/molbio.html>. All program algorithms contain compromises and you are urged to read the documentation carefully.

### Publications

Fundamentals of Molecular Evolution by W-H Li and D Grauer, published by Sinauer Associates, Sunderland, Massachusetts.

Molecular Evolution by W-H Li (1997), published by Sinauer Associates, Sunderland, Massachusetts.

Felsenstein, J. (1982) Numerical methods for inferring evolutionary trees. *Quart. Review of Biology* 57:379-404.

Felsenstein, J. (1988) Phylogenies from molecular sequences: Inference and reliability. *Ann. Rev. Genet.* 22:521-565.

Swofford, D. L. and Olsen, G. J. (1990) Phylogeny reconstruction. In *Molecular Systematics*, D. M. Hillis and C. Moritz, ed. (Sunderland, MA: Sinauer Assoc.), pp. 411-501.

## Appendix

### Common Options

**U (User tree)** option. By default the program will search for the best tree. If the option is selected the user tree specified in the input file will be used. When more than one tree is specified, the statistically most significant one will be selected. It is possible to provide your own tree file, for example a parsimony tree, in a distance tree analysis so that the distance analysis will add the branch-length values to the parsimony tree.

Example of an input file:

```
6 13
Archaeopt 0011001110000
Hesperorni0001101101101
Baluchithe1111011011101
B. virginii1111011101101
Brontosaur0110100111011
B. subtilis0000000011010
1
((B.subtilis,Baluchithe),((Brontosaur,B._virgini),
(Hesperorni,Archaeopt)));
```

The number of trees included

**Note** that a tree file may continue from one line to the next, usually after a comma.

**G (Global)** option toggles between the default of local rearrangement and global rearrangement.

**J (Jumble)** option. Most tree construction programs are depend on the order of input of species. When selected The input order of the sequences in randomized. The random number generator requires a "seed" in the form of an integer between 1 and 32767, and should be  $4n+1$ , which means that it must give a remainder of 1 when divided by 4. This can be judged by looking at the last two digits of the number. By simply changing the random number seed and re-running the programs one can look for other, and better trees.

**O (Outgroup)** option. This specifies which species is to be used to root the tree by having it become the outgroup.

**T (Threshold)** option. This sets a threshold such that if the number of steps counted in a character is higher than the threshold, it will be taken to be the threshold value rather than the actual number of steps. The default is a threshold so high that it will never be surpassed.

**M (Multiple data sets)** option. This allows multiple datasets for bootstrapped analysis