in a separate line. Note that some programs (e.g. FITCH) require unrooted user trees whereas others (e.g. DNAPARS) need rooted user trees instead. See the manuals for details.

PHYLIP does not provide programs to generate multiple alignments or distance matrices. These must be built by hand (remember to save the file in "*Text Only*" or *ASCII* format) or through the use of other programs like, e.g. *Clustal* or *TreeAlign* (be sure to select PHYLIP output format for the data file, inspect it for correctness and rename it to infile). Many PHYLIP programs use the output of other programs in the package to further process the results. You should rename the appropriate output file (outfile or treefile) to infile before continuing.

A few programs are intended exclusively for converting or preprocessing PHYLIP input data for use with other programs:

FACTOR recodes multistate characters into binary datasets.

SEQBOOT bootstraps input data sets (molecular sequences, binary characters, restriction sites or gene frequencies).

Using the programs

When run, each program will display on your terminal its role, its version and a menu. Your screen will usually look like this:

```
Nucleic acid sequence Maximum Likelihood method, version 3.572c

 Settings for this run:
 U               Search for best tree?  Yes
 T       Transition/transversion ratio:  2.0000
 F      Use empirical base frequencies?  Yes
 C  One category of substitution rates?  Yes
 G              Global rearrangements?  No
 J  Randomize input order of sequences?  No.Use input order
 O                        Outgroup root?  Yes,at seq.Number 1
 M          Analyze multiple data sets?  No
 I         Input sequences interleaved?  Yes
 0  Terminal type (IBM PC, VT52, ANSI)?  ANSI
 1   Print out the data at start of run  No
 2 Print indications of progress of run  Yes
 3                        Print out tree  Yes
 4      Write out trees onto tree file?  Yes
Are these settings correct? (type Y or the letter for one to change)
```

To change any option, type the corresponding letter from the left column in the menu. You will be prompted for additional information as needed. Once you are happy with the options selected press "Y" to go on.

PHYLIP © 1986-1995 by Joseph Felsenstein and the University of Washington
http://evolution.genetics.washington.edu/

This document is © by José R. Valverde from EMBnet/CNB and is being distributed by the P&PR PC of EMBnet.

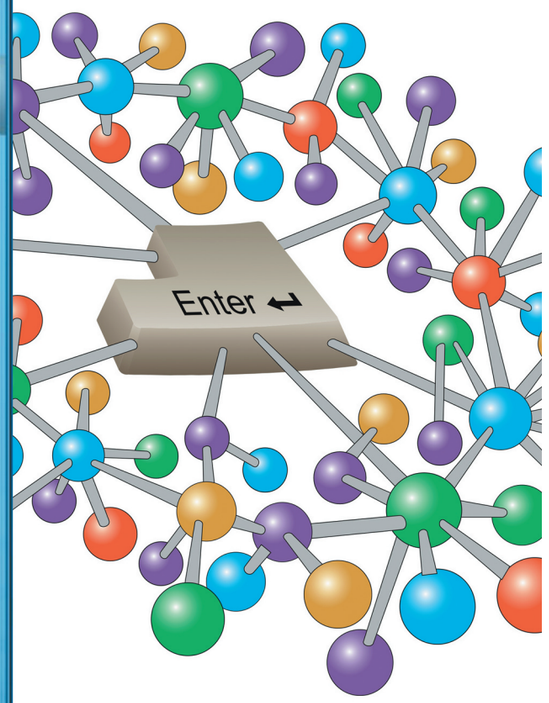
EMBnet is a bioinformatics support network situated primarily in Europe. Most countries have a national node which can provide training courses and user support. Further information on PHYLIP is available from your national node, which is reachable through

http://www.embnet.org/

If you have found this publication useful or have any ideas for similar documents, please, contact us: emb-pr@dl.ac.uk

A Quick Guide to PHYLIP
1998
reprinted in 2003

folding instruction



EMBnet

A Quick Guide PHYLIP

A Quick Guide to PHYLIP

This is an introduction to the PHYLIP package of phylogeny inference programs. PHYLIP is a public domain package written by *Joe Felsenstein* and composed of a large number of programs, which makes it versatile and very powerful.

This Guide is only a *simple roadmap* to PHYLIP. Useful tips and reminders will be dropped along the way. A basic understanding of phylogeny, evolutive mechanisms and the different theoretical approaches employed in evolutionary analysis is assumed.

Preparing input data

PHYLIP programs read their data from a file that should be named "infile". The general format of this file is:

```
   4   40  W                                   [1]
W          0101001111 0101110101 0101110011 [2]
           1101010110
dmras1     GTCGTCGTTG GACCTGGAGG CGTGGGCAAG [3]
spras      GTAGTTGTAG GAGATGGTGG TGTTGGTAAA
scras1     GTAGTTGTCG GTGGAGGTGG CGTTGGTAAA
scras2     GTCGTCGTTG GTGGTGGTGG TGTTGGTAAA
           TCCGCGCTCA
           AGTGCTTTGA
           TCTGCTTTAA
           TCTGCTTTGA
1                                              [4]
((dmras1,ddrasa),((hschras,spras),(scras1,scras2)));
```

1. The first line contains the number of species, the number of characters and, possibly, one or more program options. Check the manual of each program to see which options are available.

2. If any option requires extra information, add it using lines that start with the option's letter followed by *all* the data.

3. Next comes the species and character data in separate lines. Each line starts with 10 letters or symbols reserved for the species name and is followed by the characters to analyze. If the characters require more than one line you may use either a *sequential or interleaved* (like the above example) format.

4. You may provide one or more user trees to guide the analysis process (option "U" in the program menu): put the number of trees in one line and each subsequent tree

Constructing phylogenies

Next you have to select the program to use, depending on the type of characters with which you are working, what you want to find out and your preferred evolutionary theory/analysis method:

Molecular sequence data: DNA

Parsimony analysis of phylogeny:

DNAPARS unrooted parsimony counting base changes by the method of Fitch.

DNAPENNY finds most parsimonious trees by branch and bound.

DNAINVAR computes phylogenetic invariants (evolutionary parsimony) for four species using Cavender's method.

Maximum likelihood methods:

DNAML implements the maximum likelihood method.

DNAMLK applies ML assuming trees must be consistent with a molecular clock, i.e. the leaves are all equidistant from the root.

Compatibility methods of phylogeny inference:

DNACOMP chooses the tree and topology to maximize the number of sites with minimum number of substitutions.

Compute distances between species:

DNADIST builds a matrix using either of Jukes-Cantor, Kimura, Jin-Nei or maximum likelihood methods.

Molecular sequence data: Proteins

Parsimony analysis of phylogeny:

PROTPARS unrooted parsimony using a mixed approach of the methods of Fitch and Eck-Dayhoff.

Maximum likelihood methods:

PROTML unsupported program developed by © Jun Adachi and Masami Hasegawa for ML analysis of amino acid data

Compute distances between species:

PROTDIST builds a distance matrix by using either Dayhoff's PAM matrix, Kimura's distance or a categories distance.

Molecular sequence data: Restriction enzymes

Maximum likelihood analysis of phylogeny:

RESTML phylogeny reconstruction from restriction sites (not restriction fragments) data by the method of Smouse and Li.

Gene frequencies and continuous characters

Maximum likelihood analysis of phylogeny:

CONTML uses a Brownian motion model to reconstruct phylogeny by a restricted maximum likelihood method.

Compute distances between species:

GENDIST computes a distance matrix using Nei's genetic distance, Cavalli-Sforza's chord measure or Reynolds, Weir, and Cockerham's genetic distance.

0-1 Discrete data

Parsimony analysis of phylogeny:

MIX general program mixing Wagner and Camin-Sokal criteria for each character separately.

PENNY finds all most parsimonious trees applying a branch and bound search strategy.

DOLLOP applies the Dollo and Polymorphism analysis methods.

DOLPENNY finds all most parsimonious trees of your data using the Dollo and polymorphism parsimony methods applying a branch and bound search.

Compatibility methods of phylogeny inference:

CLIQUE compatibility method for unrooted 2-state characters.

Distance Matrix data

FITCH infers phylogenies using Fitch-Margoliash, least-squares of Cavalli-Sforza and Edwards or other similar methods of the same family.

KITSCH builds phylogenies using same methods as FITCH but assuming an evolutionary clock and that all leaves of the tree are contemporaneous.

NEIGHBOR uses the neighbor-joining method of Nei and Saitou and the UPGMA methods to successively cluster lineages.

Displaying results

Most programs write a report into a file named "outfile" and a representation of the trees found in a file named "treefile". You should rename these files if you want to preserve the results, otherwise they will be overwritten by the next program you run.

Contents of outfile vary from program to program and depending on the output options selected. Usually it contains the name of the program, the input data and the phylogenies plus associated information. Note that in most cases, the produced trees are unrooted.

The trees contained in the treefile are represented in a standard format that uses parentheses to indicate grouping of species. PHYLIP provides programs to construct, manipulate and print trees in graphic format:

RETREE allows interactive construction and manipulation of trees (topology, branch lengths, labels, etc..).

DRAWGRAM draw a cladogram or phenogram of a *rooted* tree.

DRAWTREE draws *unrooted* phylogenies on a variety of output devices.

After you select a text font, you will be able to play with the tree interactively using these programs; once you are satisfied with the tree, write it out to a "plotfile" in the format of your choice. This file can then be printed on the appropriate device.

Estimating phylogenies

Estimating phylogenies by hand:

DNAMOVE interactive reconstruction of DNA evolution using parsimony.

MOVE interactive reconstruction of 0-1 discrete data evolution using Wagner and Camin-Sokal mixed parsimony (as in MIX).

DOLMOVE interactive phylogeny reconstruction using the Dollo and Polymorphism parsimony criteria.

CONTRAST using quantitative characters and a tree, produce independent contrasts, correlations and statistics for those characters; results can be further analyzed in any multivariate statistics package.

Bootstrapping

SEQBOOT allows to resample data sets by the bootstrap, jackknife or permutation methods.

CONSENSE finds the consensus of a number of trees using M methods (including majority-rule and strict cohsensus).

To do a bootstrap analysis follow these steps:

1) Run SEQBOOT on the input dataset, select a shuffling method and specify at least 100 (better 1000 or more) replicates. Do **not** select option "1" to avoid interference with programs run at the next step. Rename outfile to infile.

2) Run the phylogeny analysis program of your choice depending on your data type, desired analysis and preferred method with your desired options. Select option "M" (multiple data sets) and enter the number of replicates generated with SEQBOOT.

3) If you generated distance matrices in the previous step you need to rename outfile to infile and run one of the Distance Matrix programs (FITCH, KITSCH or NEIGHBOR) now to generate the trees.

4) Rename the treefile to infile and run CONSENSE to evaluate the significance of your analysis.

Note that CONSENSE gives the bootstrap support for branching order at each branch of the produced tree and not actual branch lengths.