PHYLIP

Joe Felsenstein

Depts. of Genome Sciences and of Biology, University of Washington

Software for this lab

This lab is intended to introduce the PHYLIP package and a number of major phylogeny methods.

1. You should download the programs from the as-yet-unreleased version 4.0 of PHYLIP. A table of links to them will be found at

   `http://evolution.gs.washington.edu/sisg/2016/programs/index.html`

   To run the GUI front end of the v4.0 programs on Windows you might need to install Java on your machine. If you have a 64-bit Windows machine maybe best to run the programs directly, and use their character-mode menus. (Mac OS X Java and the Java that comes with Linux should be good enough).
2. If you have come with a tablet with the iPad or Android operating systems, there is no version of PHYLIP available for that.

PHYLIP

- Distributed since 1980
- Originally in Pascal, now in C
- Intended to provide “basic transportation”
- Intended to provide a wide variety of methods
- Freely available (unless you try to charge others for it)

Advantages of PHYLIP

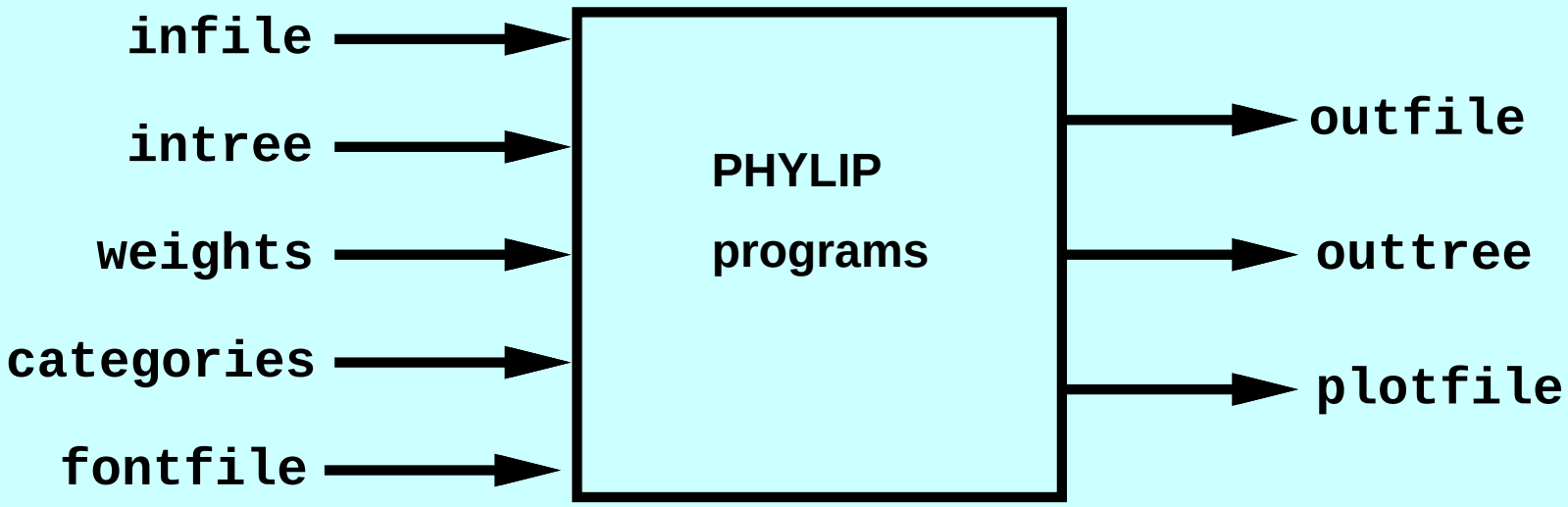
1. Free (in the sense of “free beer”), easily obtainable
2. Runs on all major platforms
3. Very good documentation
4. Lots of people around who know how to use it
5. Often used in teaching about phylogenies.
6. Runs can be automated by using input redirection and command files
7. Support for PHYLIP-format files by many other programs including phylogeny programs and sequence-alignment programs.

Over 31,000 registered users in over 50 countries including: Fiji, Cuba, Papua New Guinea, Iran, Iceland. Large numbers of users in countries such as India, Brazil, Argentina, Russia, and China where even modest cash prices for software can be a major burden.

Disadvantages of PHYLIP

1. Tree search less thorough than some other packages such as PAUP*.
2. Much, much slower than packages such as PAUP* and RAxML
3. Manual steps such as renaming file names can be tedious
4. Still no: codon model (but coming very soon), Bayesian inference.
5. Not as many options available as in other programs
6. Cannot read NEXUS standard files

PHYLIP programs



These are the default file names. If the input files do not exist (or if the output files exist and you choose not to overwrite them), you will be asked for the file name. **This is not a bug.**

Input format for PHYLIP (DNA, Interleaved)

```
    7  112
Bovine     CCAAACCTGT CCCCACCATC TAACACCAAC CCACATATAC AAGCTAAACC AAAAATACCA
Mouse      CCAAAAAAAC ATCCAAACAC CAACCCCAGC CCTTACGCAA TAGCCATACA AAGAATATTA
Gibbon     CTATACCCAC CCAACTCGAC CTACACCAAT CCCCACATAG CACACAGACC AACAACCTCC
Orang      CCCCACCCGT CTACACCAGC CAACACCAAC CCCCACCTAC TATACCAACC AATAACCTCT
Gorilla    CCCCATTTAT CCATAAAAAC CAACACCAAC CCCCATCTAA CACACAAACT AATGACCCCC
Chimp      CCCCATCCAC CCATACAAAC CAACATTACC CTCCATCCAA TATACAAACT AACAACCTCC
Human      CCCCACTCAC CCATACAAAC CAACACCACT CTCCACCTAA TATACAAATT AATAACCTCC

           CCCCAGCCCA ACACCCTTCC ACAAATCCTT AATATACGCA CCATAAATAA CA
           TCCCACCAAA TCACCCTCCA TCAAATCCAC AAATTACACA ACCATTAACC CA
           GCACGCCAAG CTCTCTACCA TCAAACGCAC AACTTACACA TACAGAACCA CA
           ACACCCTAAG CCACCTTCCT CAAAATCCAA AACCCACACA ACCGAAACAA CA
           ACACCTCAAT CCACCTCCCC CCAAATACAC AATTCACACA AACAATACCA CA
           ACATCTTGAC TCGCCTCTCT CCAAACACAC AATTCACGCA AACAACGCCA CA
           ACACCTTAAC TCACCTTCTC CCAAACGCAC AATTCGCACA CACAACGCCA CA
```

Format for trees in tree files (Newick standard)

(Mouse:0.87231,Bovine:0.49807,(Gibbon:0.25930,(Orang:0.24166,
(Gorilla:0.12322,(Chimp:0.13846,
Human:0.08571):0.06026):0.04405):0.10815):0.39538);

More than such tree can be placed end-to-end in the same tree file.

The Newick standard was defined by an informal standards committee in 1986. It is described on this web page:

`http://evolution.gs.washington.edu/phylip/newicktree.html`

It is very widely used by phylogeny programs. For example, although the tree block format of the NEXUS file format is a competitor, inside of it one will actually find ... a Newick tree.

PHYLIP guide

A useful guide to using PHYLIP with molecular sequences has been produced by Jarno Tuimala. It can be downloaded as a PDF from

`http://koti.mbnet.fi/tuimala/oppaat/phylip2.pdf`

or using the link to it on the main PHYLIP web page.

For more information on many other programs

... at my PHYLIP web site there is a master list of over 390 phylogeny programs, with descriptions and links.

To find it simply put the phrase “Phylogeny Programs” into your favorite search engine.

However, it is not really up-to-date. I have had to stop work on it as I have no one to help me on that.

What to do in the PHYLIP likelihood lab exercise

1. Get a DNA or protein sequence data set of aligned sequences. You can use one of the ones provided by the course if you wish. They are also at
`http://evolution.gs.washington.edu/sisg/2016/data/`
2. Make a copy the data file as file `infile` , and then run either `Dnaml` or `Proml`, whichever is appropriate. Use the `R` to do a "Gamma distributed rates" analysis and then the `A` options to set it to a mean block length of about 3. After you accept the menu settings, you will be asked for a coefficient of variation of rates (you could set this at 2.0) and for the number of rate categories used to approximate the Gamma distribution (about 5-6 would be good).
3. Look at the tree by looking at the output file `outfile` (when you examine that file, you will need to make sure the font is a fixed-width one such as Courier) and also by renaming `outtree` to `intree` and then using `Drawgram` (perhaps with font file `font1`). You can also try `Drawtree`. (In using these, when you get a preview of the graph, use the `File` menu to choose whether you want to change settings. The final plot will be called `plotfile` .

More to do: the PHYLIP distance lab exercise

Use your data set and analyze it by the Neighbor-Joining method:

1. Make a copy of your sequences and call that file `infile`
2. Run `Dnadist` or `Protdist`, whichever is appropriate.
3. The distance matrix is in file `outfile`
4. Rename that `infile`
5. Run `Neighbor`, using the default options except maybe the outgroup-rooting option.
6. The output file `outfile` will show your tree, and the output tree file `treefile` has the Newick-format representation of it. Save them by renaming them. When examining the output file, use a constant-width font to avoid distortion of the tree.

More to do: the PHYLIP bootstrap lab exercise

Use that distance matrix method to do a bootstrap analysis:

1. (use `Seqboot`, then renaming `outfile` to `infile`, (You can use 1000 replicates if you have DNA sequences (use menu option R), but don't do 1000 replicates for a protein data set as this will be too slow). When asked for the random number seed, provide any odd number whose last two digits give a remainder of 1 when divided by 4 (for example, they might be 45).
2. Use that `infile` of many data sets as an input for `Dnadist` or `Protdist`, using the M (Multiple input data sets) option (with multiple data sets, not weights).
3. The multiple distance matrices are now in file `outfile`. Rename that to `infile`.
4. Now run program `Neighbor`, making sure to set the multiple data sets option M and provide the number of the bootstrap replicate distance matrices.
5. Rename the output file `outtree` (which will contain multiple bootstrap estimates of the tree) to `intree`.
6. Run program `Consense` which makes an Extended Majority-Rule Consensus Tree from these trees.
7. Look at the consensus tree by examining `outfile`, or renaming `outtree` to `intree` and running either `Drawgram` or `Drawtree`.
8. The branch lengths of this consensus tree are weird (they reflect levels of bootstrap support rather than amounts of change. Can you figure out a way, using the original sequences and the consensus tree and menu option U (User-defined tree) in the likelihood program, to get more reasonable branch lengths in that tree?