pahadist: Pairwise alpha heterotachy adjusted distances
Version 1.0

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Introduction

The main program `pahadist`, implements the methods described in Wu and Susko (2011); please cite this reference when using the software. The program gives the estimated distance matrix for a sequence alignment and the estimated α parameters when both distances and α are estimated separately for each pair of taxa.

Installation

The main program `pahadist` is compiled from C source code.

To install the `pahadist` routines on other systems

1. Download and unpack the software:

   ```
   $ tar zxf pahadist.tar.gz
   ```

   This will create a directory `pahadistv1.0` that contains the source code and test input files.

2. Change directories to `pahadistv1.0` and create the main program files with the `make` command:

   ```
   $ cd pahadistv1.0
   $ make
   ```

This should produce the program file `pahadist` which can be copied to a location in your PATH.

A brief example

The program can be run at the command line with

```
$ pahadist controlfile
```

As a brief example, consider the control file `mtprot.ctl`:

```
* Example controlfile
seqfile =  data/mtprot.dat  * location of sequence file
nchar =  20                 * amino acid data
model =  3                  * emprical+F model
aaRatefile =  mtREV24.dat   * emprical rate matrix
```

The command

```
$ pahadist mtprot.ctl
```

Output to the screen is a distance matrix in a format that can be understood by standard phylogenetic programs like those in the PHYLIP package (Felsenstein, 2004; Felsenstein, 1989) followed by another ‘distance’ matrix containing the estimated α parameters for the pairs in place of estimated distances.

```
6
Homsa      0.000000 0.300475 0.281697 0.311242 0.356831 0.383840
Phovi      0.300475 0.000000 0.131978 0.195039 0.278213 0.316647
Bosta      0.281697 0.131978 0.000000 0.195436 0.257691 0.312499
Orycu      0.311242 0.195039 0.195436 0.000000 0.262254 0.331654
Musmu      0.356831 0.278213 0.257691 0.262254 0.000000 0.339790
Didvi      0.383840 0.316647 0.312499 0.331654 0.339790 0.000000
6
Homsa      0.000000 2.456631 3.498824 1.402925 2.438682 5.813623
Phovi      2.456631 0.000000 -1.000000 2.290182 3.066239 6.772770
Bosta      3.498824 -1.000000 0.000000 5.767504 6.137967 3.398991
Orycu      1.402925 2.290182 5.767504 0.000000 2.672055 1.909729
Musmu      2.438682 3.066239 6.137967 2.672055 0.000000 5.625852
Didvi      5.813623 6.772770 3.398991 1.909729 5.625852 0.000000
```

The first block gives the estimated distances and the second block gives the estimated α parameters. The -1.0 indicates that the estimated α for the pair of taxa `Phovi` and `Bosta` was infinite; a rates-across-sites parameter is not needed for this pair.

Programs like `neighbor` in the PHYLIP package, which obtains the neighbor-joining tree, will accept the output above as an input distance matrix; they will ignore the second block with the α estimates.

Input

All input to the routine is through a main control file, `controlfile`. The control file is similar in format to the control files used by the programs `baseml` and `codeml` in the PAML package (Yang 1997, 2007). For instance, the `model` variable specifies the substitution model and gives a subset of the models available in PAML, with the same numbering scheme. As with PAML control files, blank lines are allowed and all text following a ’*’ till the end of a line is treated as a comment. The word on the left of an equal sign gives a control variable and the word on the right gives the value of that variable. Spaces are required on both side of an equal sign. The order of variables is unimportant. The control variables are as follows. All variables not indicated as optional are required.

- `nchar`: An optional integer indicating that the model was for nucleotide data (`nchar` = 4) or amino acid data (`nchar` = 20). The default value is 4.

- `model`: An integer code for the substitution model. For nucleotide data (`nchar` = 4), the models currently implemented are

`model`	Model
0	JC
2	F81
3	F84
4	HKY
7	GTR

and for amino acid data (`nchar` = 20) the models currently implemented are

`model`	Model
0	Poisson
1	Proportional
2	Empirical
3	Empirical+F
8	REVaa

The documentation for the PAML package gives a good description of the models listed and can fit all of them.

The GTR and REVaa models refer to the most general time-reversible models in the nucleotide and amino acid case, respectively. The Poisson and Proportional models are the analogues of the JC and F81 models for amino acid data. The Poisson and Proportional models have substitution probabilities

$$P_{ij}(t) = \begin{cases} \pi_j + (1-\pi_j)\exp[-\mu t] & \text{if } i = j \\ \pi_j - \pi_j \exp[-\mu t] & \text{otherwise} \end{cases}$$

where $\mu = [\sum \pi_i(1-\pi_i)]^{-1}$ and π_j gives the stationary of the jth amino acid. In the Poisson model, the frequencies are all 1/20

When `model` = 2 or 3 an empirical model is fit. The model is specified by the variable `aaRatefile`. When `model` = 2, the stationary frequencies are the stationary frequencies of the specified empirical model.

When `model`=3, the stationary frequencies must be specified in a file `frfile`. In this case, the empirical model is used to specify the exchangeabilities. The exchangeability of amino acid i and j is defined as

$$S_{ij} = Q_{ij}/\pi_j$$

where, for the specified empirical model, Q_{ij} is the rate of substitution from i to j and π_j is the stationary frequency j. When `model`=3, the rate of substitution from i to j is

$$\tilde{Q}_{ij} = S_{ij}\tilde{\pi}_j$$

where $\tilde{\pi}_j$ is the frequency of j specified in `frfile`.

- `aaRatefile`: Only required for empirical amino acid models (`model` = 2 or 3 and `nchar` = 20). The name of the empirical model to fit. The models currently implemented are

`dayhoff.dat`	Dayhoff or PAM	Dayhoff et al. (1978)
`jones.dat`	JTT	Jones et al. (1992)
`wag.dat`	WAG	Whelan and Goldman (2001)
`mtREV24.dat`	mtREV	Adachi and Hasegawa (1996)
`lg.dat`	LG	Le and Gascuel (2008)

The naming scheme was chosen to be consistent with PAML. However, `aaRatefile` is not actually the name of file, rather it identifies a model.

Amino acids are ordered alphabetically:

```
alanine, arginine, asparagine, aspartic, cysteine,
glutamine, glutamic, glycine, histidine, isoleucine,
leucine, lysine, methionine, phenylalanine, proline,
serine, threonine, tryptophan, tyrosine, valine
```

which is the same ordering used by most phylogenetic packages including PAML, PHYLIP and TREE-PUZZLE.

- `Qfile`: Only required for the general time reversible model, GTR or REVaa (`model` = 7, `nchar` = 4 or `model` = 8, `nchar` = 20). The name of a file containing the entries of the rate matrix separated by blanks.

- `kappa` or `ttratio`: One of these is required for the F84 and HKY models (`model` = 3 or 4 and `nchar` = 4). A real number giving the κ parameter for the model. The F84 model has a rate matrix proportional to

$$\begin{bmatrix} \cdot & \pi_C & (1+\kappa/\pi_R)\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & (1+\kappa/\pi_Y)\pi_T \\ (1+\kappa/\pi_R)\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & (1+\kappa/\pi_Y)\pi_C & \pi_G & \cdot \end{bmatrix}$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. The HKY model has a rate matrix proportional to

$$\begin{bmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{bmatrix}$$

The transition-transversion ratio (`ttratio`) is related to the κ parameter in the F84 model through

$$R = \kappa \times \frac{\pi_A\pi_G/\pi_R + \pi_C\pi_T/\pi_Y}{\pi_R\pi_Y} + \frac{\pi_A\pi_G + \pi_C\pi_T}{\pi_R\pi_Y}$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. For the HKY model the relationship is

$$R = \kappa \times \frac{\pi_A\pi_G + \pi_C\pi_T}{\pi_R\pi_Y}$$

- `seqfile`: The file should conform to the requirements of the PHYLIP package (Felsenstein, 1989, 2004). Sequence names should be 10 characters long and padded by blanks. The names should match the names used in the input treefile. Input can be either interleaved or sequential with one caveat: The lines 2 through $m+2$, where m is the number of taxa, must contain the name of taxa followed by sequence data. For instance the start of a sequence file might be

```
6 3414
Homsa      ANLLLLIVPI LI...
Phovi      INIISLIIPI LL...
...
```

but not

```
6 3414
Homsa
ANLLLLIVPI LI...
Phovi
INIISLIIPI LL...
...
```

which would be allowed under the sequential format by PHYLIP.

Additional information is available at

```
http://evolution.genetics.washington.edu/phylip/doc/sequence.html
```

References

Felsenstein, J. 2004. PHYLIP Phylogeny Inference Package (version 3.6). Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (version 3.2). Cladistics 5: 164-166.

Yang, Z. (2007). PAML 4: a program for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang, Z. (1997). PAML: a program for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Wu, J. and Susko, E. (2009). General heterotachy and distance method adjustments. Mol. Biol. Evol. 26:2689–2697.