Software for
Testing for Differences in Rates Across Sites Distributions inPhylogenetic Subtrees

Edward Susko
Department of Mathematics and Statistics, Dalhousie University

Portions of the user interface and translations from the statistical programming language R to C were implemented by Karin Stoner.

Introduction

The main routine `bivar` fits the bivariate rate distribution as discussed in Susko et al (2003).

The routine modifies of the protdist routine of the alpha PHYLIP distribution, version 3.6 and requires the some of the same user input as the protdist routine. Some additional routines from the Harwell Software Libraries (HSL) need to be downloaded.

Unix Installation

To unpack the software, type

```
$ gzip -d bivar_soft.tar.gz
$ tar xvf bivar_soft.tar
```

This will create a directory `bivar`.

The Harwell Software Library Routines

The main program `bivar` uses Harwell Software Libraries (HSL) optimization routines. Each user of HSL software is required to register to use the library and so the source code for these routines must be downloaded separately at

`http://www.cse.clrc.ac.uk/nag/hsl/`

The routine required is the `VE11AD` FORTRAN 77 subroutine with any other dependent functions and subroutines. This routine is part of the freely available HSL Archives. The `VE11AD` routine and dependencies should be copied to the file `ve11.f`. The file should be placed in the `bivar` directory.

To create the executables type

```
$ cd bivar
$ make
```

After copying the following files to a directory in your PATH, you can remove the `bivar` directory and its contents.

```
bivar
promlratec
cratedist
jratedist
```

The input

The routine `bivar` should be called at the command line with

```
$ bivar
```

The user will then be prompted for input in the following order.

```
infile1
treefile1
infile2
treefile2
number of rates (m)
upper bound (u)
number of sites in the sequence (n)
```

Here the infiles `infile1` and `infile2` should be standard PHYLIP format data file and the treefiles `treefile1` and `treefile2` should be standard PHYLIP format (Newick format) treefiles. The data file `infile1` should be the data file for the `treefile1` and the data file `infile2` should be the data file for the `treefile2`

The ordering of the infiles determines the ordering of the differences and other output. Rate differences are for the rates for `treefile1` minus the rates for `treefile2`. The rate distributions are output as `rate1 rate2 prob`, where `rate1` is the rate for `treefile1` and `rate2` is the rate for `treefile2`.

The rates for the bivariate distribution are m equally spaced rates from 0 to u. For instance with $m = 11$ and $u = 10$, the rates are 0, 0.1, ..., 1. The bivariate rate

distribution specifies `rate1`, `rate2` and the probability that at a given site, `rate1` will be assigned to `tree1` and `rate2` will be assigned to `tree2`. For the example above with $m = 11$ and $u = 10$, probabilities would be assigned for each of the $m^2 = 121$ possible pairs of rates, $(0, 0)$, $(0, 0.1)$, ..., $(0, 1)$, $(0.1, 0)$, ..., $(1, 1)$.

The output

There are three output files that are created by the routines, `confidenceIntervals.dat`, `ratew.dat` and `SeparateRates.dat`. The file `ratew.dat` gives the bivariate rate distribution and has rows of the form

```
rate1 rate2 prob
```

For each row, `prob` specifies the probability that at a given site, `rate1` will be assigned to `tree1` and `rate2` will be assigned to `tree2`.

The output file, `confidenceIntervals.dat`, has rows of the form

```
Site     Lower       Upper      Rated        PValue
```

Here `Site` is the site number, `Lower` and `Upper` give the lower and upper bounds for a 95% confidence interval for the rate difference at the site. This rate difference is the rate for the first tree input minus the rate for the second tree input. The posterior mean rate difference is given in `Rated` and `Pvalue` gives the p-value for a test of the null hypothesis that the rate difference is 0 at the site.

The final output file `SeparateRates.dat` has two columns. The ith row gives the rates for the two subtrees at site i in the sequence. The first entry for a row is the rate for the first tree and the second for the second tree.

Susko, E., Inagaki, Y., Field, C., Holder, M.E. and Roger, A.J. (2002). Testing for Differences in Rates Across Sites Distributions in Phylogenetic Subtrees. *Molecular Biology and Evolution*, **19**, 1514–1523.