SHTests v1.0

SHIMODAIRA and HASEGAWA Tests of Phylogenetic Hypotheses

Users Manual

.

Andrew Rambaut

Department of Zoology
University of Oxford
South Parks Road
Oxford. OX1 3PS.
U.K.

e-mail: andrew.rambaut@zoo.ox.ac.uk
tel: +44 1865 271261
fax: +44 1865 271249

This program implements the method of:

> SHIMODAIRA, H., AND M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

It is evaluated and compared with similar methods in:

> Goldman, N., J. P. Anderson, and A. G. Rodrigo. In press. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*.

Version History:

Version 1.0 - First released version.

Introduction

This program implements the test described in:

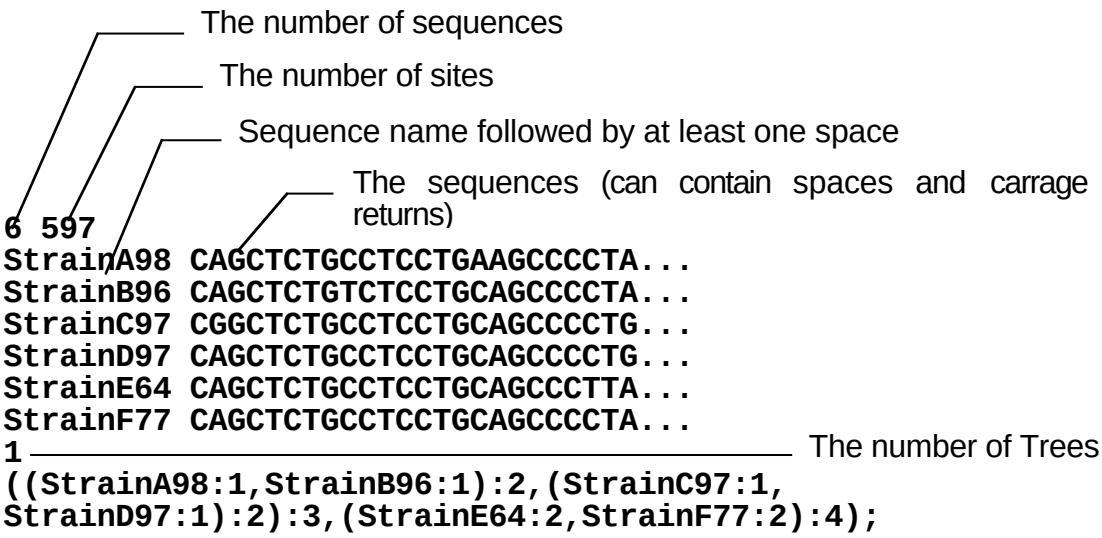
> SHIMODAIRA, H., AND M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

For information about the applicability of this test please refer to Nick Goldman's web site:

http://www.zoo.cam.ac.uk/zoostaff/goldman/index.html

Running SHTests

SHTests requires a sequence alignment together with a set of trees. The sequence alignment should be input in the following format:



The only further information that SHTests requires is given immediately upon running, and is entered as 'command-line' arguments. In other words, various arguments are given on the line after typing the program name on UNIX machines, or in a dialog box that pops up upon running the program on a Macintosh.

UNIX machine:

Compiling

Before running the program on a UNIX machine it needs to be compiled. This can be done by typing **`make`** followed by return. Alternatively in the SHTests folder, type:

```
cc -o shtests *.c -lm
```

You can add optimisation flags or use a different compiler (instead of cc) if required. Contact your system administrator if you have any problem compiling the program.

Running

To run the program on a UNIX machine type the name of the program followed by the desired parameter settings, the input file preceded by <, and the name of the file to which output is to be written preceded by >. For example:

```
shtests -mHKY -t1.05 < input_file > output_file
```

Required parameters and their default settings are given in the Parameters section of this manual.

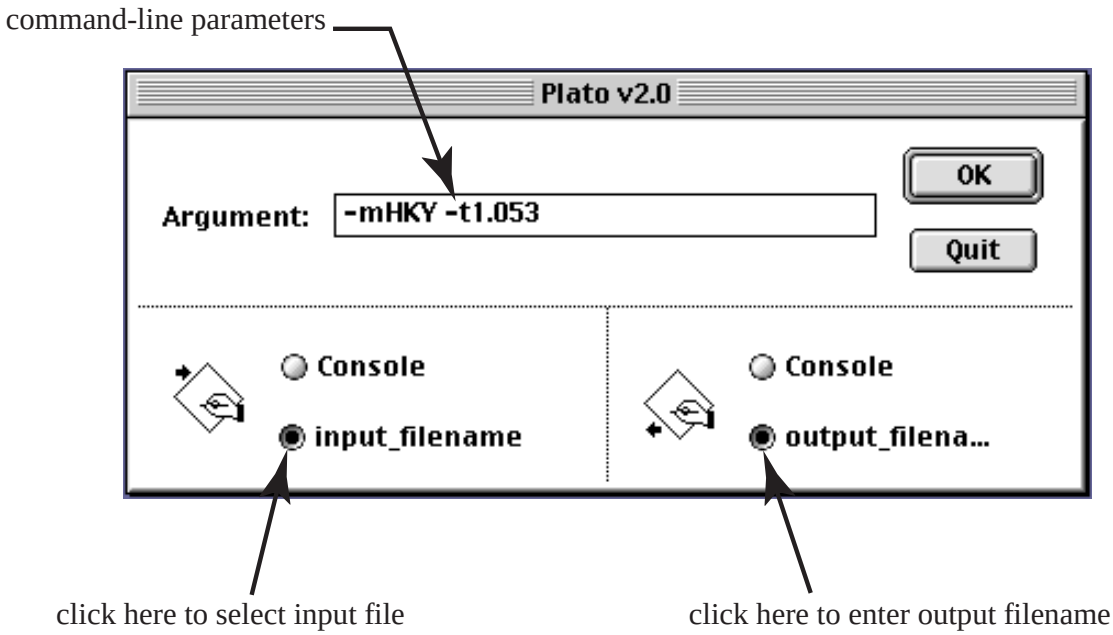
Macintosh

Compiling

Precompiled executables are distributed with the source code in the Macintosh package.

Running

Upon running SHTests on the Macintosh a dialog box will appear (see below). This box simulates a UNIX environment and allows parameter settings, and the input/ output files to be specified.



Command Line Parameters

The parameters that SHTests requires, and their default values, are given below. These parameters specify the substitution model to be used, and also allow other options, such likelihood ratio test, or the amount of information to be output, to be set.

Model

This option sets the model of nucleotide substitution with a choice of either *F84*, *HKY* (also known as HKY85) or REV (markov general reversable model). The first two models are quite similar but not identical. They both require a transition transversion ratio and relative base frequencies as parameters. Other models such as *K2P*, *F81* and *JC69* are special cases

of *HKY* and can be obtained by setting the nucleotide frequencies equal (for *K2P*) or the transition transversion ratio to 1.0 (for *F81*) or both (for *JC69*). The usage is:

`-m <MODEL>`

Where <MODEL> is a three letter code: HKY ,F84 or REV. If no model is specified, the default is F84 which is computationally simpler.

Kind of Test

There are two kinds of SH tests implemented in this program, the RELL test and the FULL test. The FULL model is slower because it re-optimizes the branch lengths and other parameters of each tree for each bootstrap replicate. The RELL model simply resamples the partial likelihoods for each site meaning it is (much) faster but approximate. This distinction is discussed in Nick Goldmans paper (cited above). The usage is:

`-k <KIND>`

Where <KIND> is a four letter code: RELL or FULL. If no kind is specified, the default is RELL.

Number of Bootstrap Replicates

Using this option the user may specify the number of bootstrap replicates to use when performing the test. For the RELL model this can be quite high (the default is 1000) but I would suggest using less when using the FULL model.

`-n <NUM_REPLICATES>`

Where <NUM_REPLICATES> is an integer number that specifies the number of bootstrap replicates to perform.

Codon-Specific Rate Heterogeneity

Using this option the user may specify the relative rates for each codon position. This allows codon-specific rate heterogeneity to be modelled. The default is no site-specific rate heterogeneity.

`-c <CODON_POSITION_RATES>`

Where the codon-specific rates are specified by <CODON_POSITION_RATES>, which are three decimal numbers, separated by commas or spaces.

Discrete Gamma Rate Heterogeneity

Using this option the user may specify the number of categories for the discrete gamma rate heterogeneity model.

`-g <NUM_CATEGORIES>`

Where <NUM_CATEGORIES> is an integer number between 2 and 32 that specifies the number of categories to use with the discrete gamma rate heterogeneity model.

Gamma Rate Heterogeneity

Using this option the user may specify a shape for the gamma rate heterogeneity called alpha. The default is no site-specific rate heterogeneity.

`-a <ALPHA>`

Where <ALPHA> is a real number >0 that specifies the shape of the gamma distribution to use with gamma rate heterogeneity. Only a discrete gamma model is implemented, the number of categories of which are specified by **`-g`**.

Relative Nucleotide Frequencies

This option is used to specify the relative frequencies of the four nucleotides. By default, SHTests will estimate them empirically from the data. If the given values don't sum to 1.0 then they will be scaled so that they do.

`-f <NUCLEOTIDE_FREQUENCIES>`

Where <NUCLEOTIDE_FREQUENCIES> are four decimal numbers for the frequencies of A, C, G and T respectively, separated by spaces or commas.

Transition Transversion Ratio

This option allows the user to set a value for the transition transversion ratio (TS/TV). This is only valid when either the HKY or F84 model has been selected.

`-t <TRANSITION_TRANSVERSION_RATIO>`

Where <TRANSITION_TRANSVERSION_RATIO> is a decimal number greater than zero (default =2.0).

General Reversable Rate Matrix

This option allows the user to set 6 values for the general reversable model's rate matrix. This is only valid when either the REV model has been selected.

`-t <RATE_MATRIX_VALUES>`

Where <RATE_MATRIX_VALUES> are size decimal numbers for the instantaneous rates of change from A to C, A to G, A to T, C to G, C to T and G to T respectively, separated by spaces or commas. The matrix is symmetrical so the reverse changes occur at the same instantaneous rate as forward changes (e.g. C to A equals A to C) and therefore only six values need be set. These values will be scaled such that the last value (G to T) is 1.0 and the others are set relative to this.

Verbose information

On its own this option displays more information to the standard error. This does not alter the results sent to standard output but will result in the same information being sent

`-v`

With an additional characters, other options are specified:

`-vd`

SHTests will output the log likelihood ratios for each bootstrap replicate.

`-vs`

SHTests will output the likelihoods for each site for each tree.

`-vp`

SHTests will provide some information about the progress of the analysis.

`-vm`

SHTests will output further information regarding either memory usage.

Minimum Information

This option prevents any output except the final output and any error messages.

`-q`

Help

This option prints a help message describing the options and then quits.

-h

What does SHTests output?

SHTests will write a table of results to the standard output:

	Tree	lnL	delta	P(delta)
	1	-5169.854182	17.171335	0.095000
ML	2	-5152.682848	0.000000	1.000000
	3	-5169.651721	16.968874	0.080000

In this case, tree 2 is the ML tree but the other trees are not significantly worse.

Problems and Bugs

If you have any problems with the program which are not answered by reading this manual, or if you discover a bug please e-mail Andrew Rambaut at:

andrew.rambaut@zoo.ox.ac.uk

Acknowledgements

Thanks to the Wellcome Trust for funding.