

TipDate v1.2



Users Manual

A program to estimate the rate of molecular evolution and time-scale of a phylogeny from dated sequences.

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

If you use this program, please cite:

Andrew Rambaut, 2000. Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*. **16**: 395-399.

Version History:

Version 1.2 - New feature - variable rate model. Improved printing of results. On Macintosh the user can now interrupt the optimization process.

Version 1.1 - Major fix - previous version gave erroneous results when the REV model was used with gamma rate heterogeneity. Improved result reporting slightly.

Version 1.01 - Minor fixes - fixes an occasional crash.

Version 1.0 - First released version.

Introduction

TipDate is an application for estimating the rate molecular evolution (and hence a time-scale) for a phylogeny consisting of dated tips. These will most frequently be from viruses or other fast-evolving pathogens that have been isolated over a range of dates. The program can also return the likelihood for the simple molecular clock model (i.e., assuming that all sequences are contemporary) or the non-clock model. These are useful for likelihood ratio tests of the fit of the model to the data.

The purpose of this manual is to describe how to run the program, rather than to describe the method. The method is described in:

Andrew Rambaut, 2000. Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*. **16**: 395-399.

Running TipDate

TipDate requires a sequence alignment together with a description of the quartets and dates to be used. The idea is that the user may supply an alignment of a large number of species and then define multiple quartets to be constructed from them. The sequence alignment should be input in the following format:

```
6 597
StrainA98 CAGCTCTGCCTCCTGAAGCCCCTA...
StrainB96 CAGCTCTGTCTCCTGCAGCCCCTA...
StrainC97 CGGCTCTGCCTCCTGCAGCCCCTG...
StrainD97 CAGCTCTGCCTCCTGCAGCCCCTG...
StrainE64 CAGCTCTGCCTCCTGCAGCCCTTA...
StrainF77 CAGCTCTGCCTCCTGCAGCCCCTA...
1 _____ The number of Trees
((StrainA98:1,StrainB96:1):2,(StrainC97:1,
StrainD97:1):2):3,(StrainE64:2,StrainF77:2):4);
_____ Rooted tree
```

Note that the sequences names have a date at the end (in this case in years). TipDate looks for any number at the end of the names and assumes these are dates. They can be decimal (i.e. 98.2) is 98 and 1/5 (if you want months specify the whole date in months). The units are arbitrary because the rates and dates estimated by the program will be specified in the same units – the program doesn't need to know what they are.

The only further information that TipDate 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 TipDate folder, type:

```
cc -o TipDate *.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:

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

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

Macintosh

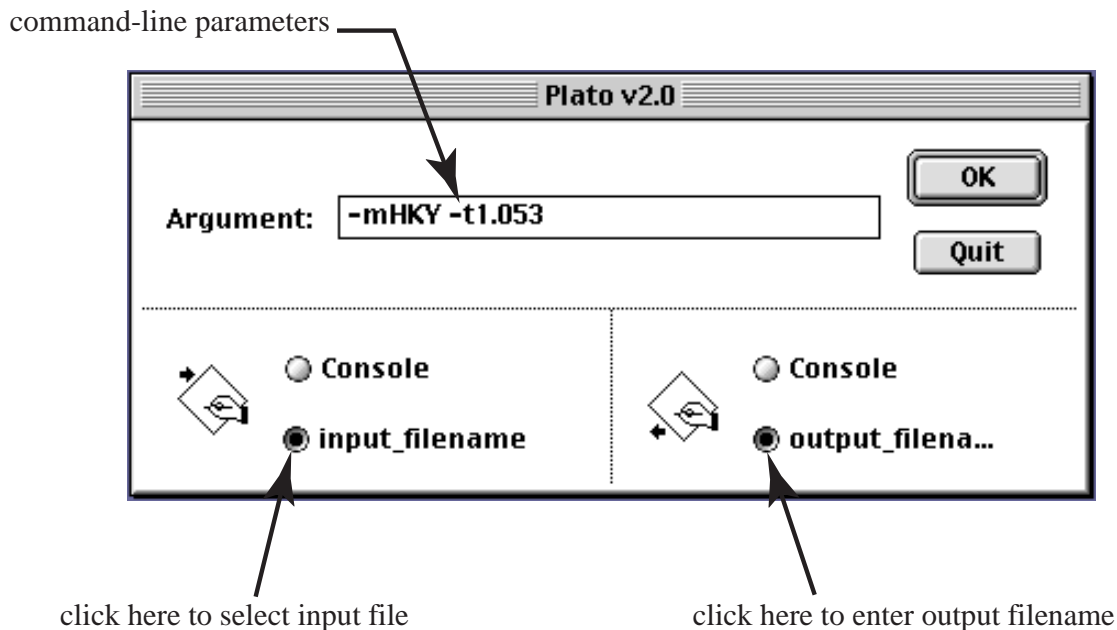
Compiling

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

Running

Upon running TipDate 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.

If the progress indicators are chosen (the **-vp** option), pressing the 'esc' key will interrupt the optimisation progress without quitting the program. This allows you to cut short an optimisation when sufficient precision is reached.



Command Line Parameters

The parameters that TipDate requires, and their default values, are given below. These parameters specify the substitution model to be used, and also allow other options, such as 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 reversible 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.

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**.

Molecular Clock Model

This option specifies the Molecular Clock model (Single Rate, SR model). This model is the equivalent of the DNAMLK program in PHYLIP or specifying the molecular clock option in PAUP*. The default (i.e., not specifying **-k**) gives the Non-Clock model (Different Rate, DR model).

-k

Tip Date Model

This option specifies the Single Rate Date Tips (SRDT) model. The default value is the Non-Clock model. The input tree and sequences must have names that end with dates – see above.

+s

Using this option will estimate the rate of evolution as a maximum likelihood parameter. Alternatively you can specify this rate using:

-s <RATE_OF_EVOLUTION>

Where < RATE_OF_EVOLUTION > is an real number that gives the rate of molecular evolution in substitutions per site per unit time (whatever the units of time that are represented by the input tip dates).

Variable Rate Tip Date Model

This option specifies the Variable Rate Dated Tip (VRDT) model. This model assumes that the rate of substitution changes linearly as we go back through time. The rate of change of rate is given as a proportion of the rate of substitution at the present. This rate can be positive or negative but not all data sets will have the power to estimate this parameter. This model must be used in conjunction with the **+s** or **-s** options, above.

+w

Using this option will estimate the rate of evolution as a maximum likelihood parameter. You can also constrain the estimation of this parameter to be only positive or only negative by adding a **+** or **-** sign after the **-w**:

+w+ OR +w-

Alternatively you can specify the rate of change of rate using:

-w <RATE_OF_CHANGE_OF_RATE>

Where < RATE_OF_CHANGE_OF_RATE > is an real number that gives the rate of change of rate as proportion of rate at present per unit time.

Estimate confidence intervals

This option specifies which parameters should have confidence intervals estimated. The default is not to estimate confidence intervals. These options can be used in combination.

-is

Estimate confidence intervals for the absolute rate of substitution (requires +s option).

-id

Estimate confidence intervals for the date of the root of the tree (requires +s option).

-iw

Estimate confidence intervals for the rate of change of rate parameter in the Variable Rate Dated Tip model (requires +w option).

Value to obtain confidence intervals

This option specifies the value to use to obtain the confidence intervals around the estimate of rate of molecular evolution (and corresponding date of root of the tree).

-l <VALUE>

Where <VALUE> is a real number ≥ 0 that specifies the log likelihood ratio that gives the confidence interval. The default is 1.92 which corresponds to half χ^2 with 1 degree of freedom. A value of 0 will disable the calculation of confidence intervals.

Specify rooting of tree

To perform the molecular clock and tip date models, the input tree must be rooted. This option is used to specify an outgroup sequence to root the tree with (sorry this is not very sophisticated – if you need to use more than one outgroup, root the tree before hand).

-r <VALUE>

Where <VALUE> is an integer number which refers to the sequence that will be used to root the tree (starting at 1). Alternatively TipDate can find the maximum likelihood position of the root – this tries all possible positions (2n-3) so increases the duration of analysis:

+r

Relative Nucleotide Frequencies

This option is used to specify the relative frequencies of the four nucleotides. By default, TipDate 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 Reversible Rate Matrix

This option allows the user to set 6 values for the general reversible 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:

-vP

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

-vm

TipDate 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 TipDate output?

TipDate will write a tree to the standard output, which is the maximum likelihood tree under the specified model. It will also print estimates of the parameters of the models to the console. This will include the estimate of substitution model parameters (TS/TV, alpha etc.), the rate of molecular evolution (with confidence intervals unless specified otherwise) and the date of the root of the tree (also with confidence intervals).

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 (AR: Grant No. 50275) for funding this work. Many thanks to Paul Harvey and Eddie Holmes. Also thanks to Nick Grassly for help creating the program.