Modeltest 3.7 (June 2005)

© David Posada
dposada@uvigo.es
http://darwin.uvigo.es/

DISCLAIMER

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

IMPORTANT: Note than only the last versions of Modeltest (3.x) are compatible with the new PAUP* version (4.0). Also consult the PAUP web page (http://paup.csit.fsu.edu/index.html) for potential bugs that may interfere with Modeltest

See MODELTEST Frequently Asked Questions at http://darwin.uvigo.es/

MODELTEST Citation:

Posada D and Crandall KA 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14 (9): 817-818.

Recommended readings

Posada D and Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. Systematic Biology 53: 793-808.

PROGRAM HISTORY

Version 3.7 (June 05):Implemented BIC (option "-b"). Removed AICcalc function (option "-i"). Removed AICfile function (option "-f"). (March 05): Fixed cosmetic bug: In the comments of the AAC PAUP* block it was printing the name of the hlRT model instead of the name of the AIC model. (December 04): Averaged estimates for alpha are now alpha(G) and alpha(IG), instead of alpha(G) and alpha(G+IG) (same thing for pinv) (thanks to Roman Biek, now this is congruent with Posada and Buckley (2004)). Now report confidence level for hLRTs. Minor aesthetic changes

Version 3.7 (November 04): The program includes now model averaged estimates and variable importance calculations. New option (-w) to define confidence interval of models used for model-averaged estimates. By default this interval is 1.0, so all models are included in model-averaged estimates. Use double precision in the AIC calculator (thanks to Renee Park). Likelihoods and number of parameters are now printed in the AIC table. The argument for specifying sample size is now -n (it was -c). Aesthetic changes.

Version 3.5 (May 04): This is a minor update that does not affect the calculations. AIC weights were

sorted by their value, but because these can be almost zero (zero for the computer) for several models, their order would not make sense in the light of the AIC values. Now the program orders the AIC weights by the AIC scores.

Version 3.4 (March 04): There was a typo printing the Rd value for K81uf+I. It was printing *p-inv* instead. (thanks to Michael Sorenson)

Version 3.3 (Nov 03): Added options to include branch length estimates as parameters and to calculate AICc. Changed some option letters accordingly.

Version 3.2 (March 03): Aesthetic changes. TrN+I had 5 df instead of 6.

Version 3.1 (Jan 02): Akaike weights and AIC differences are now calculated. Several minor aesthetic variations.

Version 3.07 (May 01): There was a bug that caused that the wrong set of likelihood scores to be printed in the first column when this option was selected.

Version 3.06 (Apr 01): Print likelihood scores by default. In the windows version there was a bug by which the file scores.txt was always the standard input (Andy Vierstraete). Using GNU license (I should have done this a long time ago) (thanks to Naoki Takebayashi)

Version 3.05 (Feb 01): In the windows version, the AIC[55] gave an AIC of 0 to the GTRIG. Now dimension is AIC[56] (Juan Suarez). TIM+G reported invariable sites instead of gamma shape (Cymon Cox)

Version 3.04 (July 00): The program outputs now a block of commands to implement the likelihood settings in PAUP* for the best-fit model selected. The mixed chi-square distribution is added as default for the I and G tests.

Version 3.0 beta 2 (December 99): Because in the new release of PAUP beta3, the likelihood score file includes now base frequencies estimates, the program was modified accordingly. The likelihood ratio tests are now explained with more detail, and the output of Modeltest is more consistent with the likelihood settings option in PAUP*.

Version 3.0 beta 1 (December 99): 16 new models are added for a total now of 56 models. These models are variations of two main substitution schemes that I called TIM (transitional model) and TIV (transversional model). They are described in figure 1 below.

Version 3.0 (February 99): several cosmetic variations. The output of Modeltest is designed now to specify the model selected accordingly to PAUP* likelihood settings.

Version 2.1 (October 99): A bug in the selection of the minimum AIC, which implied that the model GTR+I+G could not be selected is solved. Also, the number of free parameter is redefined. Now JC is considered to have 0 free parameters and GTR+I+G 10. This would affect only a few AIC calculations.

Version 2.0 (June 99): 40 models are included

Version 1-1.06 (June 98): several aesthetic variations

I really appreciate the input from several users and would like to show my gratitude. Thank you very much!

BACKGROUND

All phylogenetic methods make assumptions, whether explicit or implicit, about the process of DNA substitution (Felsenstein, 1988). For example, an assumption common to many phylogenetic methods is a bifurcating tree to describe the phylogeny of species (Huelsenbeck, Crandall, 1997). Consequently, all the methods of phylogenetic inference depend on their underlying models. To have confidence in inferences it is necessary to have confidence in the models (Goldman, 1993b). Because of this, all the methods based on explicit models of evolution should explore which is the model that fits the data best, justifying then its use.

Likelihood ratio tests (LRT)

In traditional statistical theory, a widely accepted statistic for testing the goodness of fit of models is the likelihood ratio test statistic $\delta = -2 \log \Lambda$, being

$$\Lambda = \frac{\max\left[L_0(NullModel \mid Data)\right]}{\max\left[L_1(AlternativeModel \mid Data)\right]}$$

where L_0 is the likelihood under the null hypothesis (simple model) and L_1 is the likelihood under the alternative hypothesis (more complex, parameter rich, model). The value of this statistic is always equal to or greater than zero, as the likelihood under the more complex model will always be equal or higher than the likelihood under the simpler model. When the models compared are nested (the null hypothesis is a special case of the alternative hypothesis) and the null hypothesis is correct, the δ statistic is asymptotically distributed as χ^2 with q degrees of freedom, where q is the difference in number of free parameters between the two models; equivalently, q is the number of restrictions on the parameters of the alternative hypothesis required to derive the particular case of the null hypothesis (Goldman, 1993b; Kendall, Stuart, 1979). To preserve the nesting of the models, the likelihood scores are estimated upon the same tree, and then, once the models have been compared, a final tree is estimated using the chosen model of evolution.

When the models are not nested, an alternative means of generating the null distribution of the δ statistic is through Monte Carlo simulation (parametric bootstrapping) (Goldman, 1993a). The χ^2 approximation used to represent the underlying distribution of the LRT has been problematic. Goldman (1993b) first pointed out the difficulty in counting the number of degrees of freedom and the problem of the sparseness of the DNA data. Later, Yang et al. (1995) suggested that the LRT was well represented by a χ^2 distribution. However, the χ^2 distribution may not be reliable when the null model is equivalent to fixing some parameters at the boundary of the parameter space of the alternative model, e.g., rate homogeneity test, where the null hypothesis is a special case of the gamma-distribution model with shape parameter (α) equal to infinity (Yang, 1996). Whelan and Goldman (1999) have also shown that for comparisons of rate variation across sites and nucleotide frequencies estimated as the observed base frequencies, the χ^2 distribution was significantly different from the true distribution, and the tests were conservative. To account for the boundary problem Ota et al. (1992)and Goldman and Whelan (2000) suggested the use of a mixed χ^2 (or $\overline{\chi}^2$) distribution consisting of 50% χ^2_0 and 50% χ^2_1 to construct LRT tests for the invariable sites. *This mixed distribution for model fitting is implemented in the Modeltest 3.04 and future versions for the invariable sites and rate heterogeneity among sites LRTs.*

Akaike Information Criterion

The Akaike information criterion (AIC, (Akaike, 1974) is an asymptotically unbiased estimator of the Kullback-Leibler information quantity (Kullback, Leibler, 1951), which is a measure of the information that is lost when a model is used to approximate full reality. Selecting the model with the minimum AIC is approximately equivalent to minimizing the expected Kullback-Leibler distance between the true model and the estimated sample. The AIC penalizes for the increasing number of parameters in the model, so it is taking into account not only the goodness of fit but also the variance of the parameter estimates. The AIC for model i is computed as:

$$AIC_i = -2 \ln L_i + 2 K_i,$$

where K_i is the number of free parameters in the ith model and L_i is the maximum-likelihood value of the data under the ith sample.

When sample size is small compared to the number of parameters (say, $n/K < 40$) the use of a second-order AIC, AIC_c (Hurvich, Tsai, 1989; Sugiura, 1978), is recommended:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1},$$

where sample size is roughly approximated by the number of variable characters in the alignment (although which number to use is really up to the user).

The AIC compares several candidate models simultaneously, it can be used to compare both nested and non-nested models, and model-selection uncertainty can be easily quantified using the AIC differences and Akaike weights. AIC differences (Δ_i or deltas) are rescaled AICs:

$$\Delta_i = AIC_i - \min AIC$$

The AIC differences are easy to interpret and allow a quick comparison and ranking of candidate models. As a rough rule of thumb, models having Δ_i within 1-2 of the best model have substantial support and should receive consideration. Models having Δ_i within 3-7 of the best model have considerably less support, while models with $\Delta_i > 10$ have essentially no support. Akaike weights (w_i) are the normalized relative AIC for each candidate model, and they can be interpreted, from a Bayesian perspective, as the probability that a model is the best approximation to the truth given the data:

$$w_i = \frac{\exp(-1/2\Delta_i)}{\sum_{r=1}^{R} \exp(-1/2\Delta_r)}$$

for R candidate models. Indeed, we could also think of estimating phylogenies under the best models and combine these trees according to their Akaike weights. Burnham and Anderson (2003) provide an excellent introduction to the AIC and model selection.

Bayesian Information Criterion
All of the above can be implemented using the Bayesian Information Criterion (BIC) (Schwarz, 1978):

$$BIC_i = -2 \ln L_i + K_i \log n,$$

where K_i is the number of free parameters in the ith model and L_i is the maximum-likelihood value of the data under the ith sample, and n is sample size (could be roughly approximated by the number of characters in the alignment, although which value to use is really up to the user).

The BIC was developed as an approximation to the log marginal likelihood of a model, and therefore, the difference between two BIC estimates may be a good approximation to the natural log of the Bayes factor (Kass, Wasserman, 1995). Given equal priors for all competing models, choosing the model with the smallest BIC is equivalent to selecting the model with the maximum posterior probability. In this way, BIC weights can be seen as approximate model posterior probabilities (Wasserman, 2000). The BIC tends to select models that are less complex than Bayes factors (for discussion see Raftery, 1999; Weakliem, 1999), and if n > 8, the BIC selects simpler models than the AIC (Forster, Sober, 2004). The BIC approximation might not be appropriate when the posterior mode occurs at the boundary of the parameter space (Hsiao, 1997; Ota et al., 2000).

Model Uncertainty
Akaike weights and BIC weights are very useful for assessing model-selection uncertainty. For example, we can establish an approximate 95% confidence set of models for the best K-L model by summing the Akaike weights from largest to smallest until the sum is just ≥ 0.95; the corresponding subset of models is a type of confidence set on the best K-L model (Burnham, Anderson, 1998, pp. 169-171; Burnham, Anderson, 2003).

Model Averaging or Multimodel Inference
In some cases there is quit a bit of uncertainty in selecting the best candidate model. In such cases, or just one when does not want to rely on a single model, inferences can be drawn from all models (or a best subset) simultaneously. This is known as model averaging or multimodel inference. See Posada and Buckley (2004) and references therein for an explanation of application of these techniques in the context of phylogenetics.

Within the AIC framework (it would be the same for the BIC), it is straightforward to obtain a **model-averaged estimate** of any parameter (Posada, 2003b). For example, a model-averaged estimate of the substitution rate between adenine and cytosine (φA-C) using the Akaike weights (w_i) for R candidate models would be:

$$\hat{\bar{\varphi}}_{A-C} = \frac{\sum_{i=1}^{R} w_i \, I_{\varphi_{A-C}}(M_i) \, \varphi_{A-Ci}}{w_+(\varphi_{A-C})},$$

where

$$w_+(\varphi_{A-C}) = \sum_{i=1}^{R} w_i I_{\varphi_{A-C}}(M_i),$$

and

$$I_{\varphi_{A-C}}(M_i) = \begin{cases} 1 \text{ if } \varphi_{A-C} \text{ is in model } M_i \\ 0 \text{ otherwise} \end{cases},$$

We also need to be careful when interpreting the relative importance of parameters. When the number of candidate models is less than the number of possible combinations of parameters, the presence-absence of some pairs of parameters can be correlated, and so their relative importances. Indeed, the averaged parameter could be the topology itself, so we could construct a **model–averaged estimate of phylogeny**. For example, one could estimate a ML tree for all models (or a best subset) and with those one could build a weighted consensus tree using the corresponding Akaike weights. See Posada and Buckley (2004) for a practical example. Likewise, one could estimate the **relative importance** of any parameter by summing the Akaike weights across all models that include the parameters we are interested in. For example, the relative importance of the substitution rate between adenine and cytosine across all candidate models is simply the denominator above, $w_+(\varphi_{A-C})$.

Performance of model selection procedures
Posada and Crandall (2001a; 2001b) and Posada (2001) have evaluated the performance of different model selection strategies.

THE PROGRAM
MODELTEST is a simple calculator written in ANSI C and compiled for the Macintosh and Windows platforms using Metrowerks CodeWarrior. Source code in ANSI C is provided, along with a Makefile for compilation with *gcc* (or *cc*) in Unix-like environments. It is designed to compare different nested models of DNA substitution in a hierarchical hypothesis-testing framework (Figure 1). MODELTEST calculates the likelihood ratio test statistic $\delta = -2 \log \Lambda$ and its associated *P*-value using a χ^2 distribution with *q* degrees of freedom in order to reject or fail to reject different null hypothesis about the process of DNA substitution. It also calculates the AIC and BIC estimate associated with each likelihood score, the AIC and BIC differences and the Akaike and BIC weights.

Usage
The user communicates with the program using a standard console interface (Figure 2).

In **Macintosh** machines where the input and output files can be specified, By clicking with the mouse in the left **File** button, the user can select an INPUT FILE. By clicking in the right **File** button, the user can specify an OUTPUT FILE (the default output is to the **Console** in the screen). In the **Argument** line the user can interact with the program.

In **Windows** (Figure 3) and **Unix** operative systems the users interact with the program through a command or argument line (Figure 4), so a terminal prompt (console, MS-DOS window) needs to be used. Doing a double click on the executable will, therefore, not work.

These are the options:

-d : debug level (e.g. -d2)");
-a : alpha level (e.g., -a0.05) (default is a=0.01)
-n : sample size (e.g., number of characters). Forces the use of AICc (e.g., -c345) (default is to use AIC)
-t : number of taxa. Forces to include branch lengths as parameters (e.g., -t28) (default is not to count them)
-w : confidence interval for averaging (e.g., -w0.95) (default is w=1.0)

-l : LRT calculator mode (e.g., -l)
-b : Use BIC instead of AIC for all calculations (default is to use AIC)
-? : help

An example command line for Windows would be:
```
modeltest3.7 -n896 -t18 < mydata.scores > mydata.modeltest
```

An example command line for Unix would be:
```
./modeltest3.7 -n896 -t18 < mydata.scores > mydata.modeltest
```

Default Mode
By default, the program will accept two classes of input files: a file containing ordered raw log likelihood scores and parameter estimates corresponding to the tested models (JC, JC+I, JC+G, JC+I+G, K80, K80+I, K80+G, K80+I+G, TrNef, TrNef+I, TrNef+G, TrNef+I+G, K81, K81+I, K81+G, K81+I+G, TVMef, TVMef+I, TVMef+G, TVMef+I+G, TIMef, TIMef+I, TIMef+G, TIMef+I+G, SYM, SYM+I, SYM+G, SYM+I+G, F81, F81+I, F81+G, F81+I+G, HKY, HKY+I, HKY+G, HKY+IΓ, TrN, TrN+I, TrN+G, TrN+I+G, K81uf, K81uf+I, K81uf+G, K81uf+I+G, TVM, TVM+I, TVM+G, TVM+I+G, TIM, TIM+I, TIM+G, TIM+I+G GTR, GTR+I, GTR+G, GTR+I+G; I: invariable sites; G: gamma distribution; see Figure 1 for abbreviations) and a PAUP* (Swofford, 1998) file containing a matrix of the same log likelihood scores resulting from the execution of a block of PAUP* commands. This block of PAUP (modelblock) commands is included in the package.

Alpha level (-a)
The user can set the alpha level of significance (by default this is 0.01) in the command line, inputting –a followed by the desired value. The program will use this level of significance in all its calculations.

Sample size (-n)
If this option is specified, the corrected AIC (AICc) will be use instead of the standard AIC (AIC). At the same time, the sample size (e.g., -n500) needs to be specified.

A note on sample size: sample size is a difficult concept for an alignment of DNA sequences, as it will depend on the number of characters, on the number of taxa and on the correlations within those. One could use the number of characters or the number of characters times the number of taxa as sample size, but probably none of these options is correct most of the time. Which value to use is up to the user.

BIC (-b)
If this option is specified, the BIC is used instead of AIC for all calculations. Under the BIC, the user needs to specify also a sample size (-n).

Number of taxa (-t)
If this option is specified, the number of branches will be included when counting which free parameters have been estimated. This will not change the order of the AIC scores, but possibly their relative differences.

Confidence interval for model averaging (-w)
This options specifies which models are included in the confidence interval, and therefore in the calculation of model-averaged estimates and parameter importance. It specifies the minimum cumulative Akaike weight (when models ar raked from best Akaike weight to worst) of the models

included in the confidence interval. By default all models are included in this interval. For example, in the table below, if the confidence interval was set to 0.8, only the first 6 models (HKY+G, HKY+I+G, TrN+G, K81uf+G, TrN+I+G and K81uf+I+G) would be included in the confidence interval.

```
Model             -lnL   K        AIC       delta      weight  cumWeight
---------------------------------------------------------------------
HKY+G         2888.2053   5   5786.4106     0.0000     0.2919     0.2919
HKY+I+G       2887.8667   6   5787.7334     1.3228     0.1507     0.4426
TrN+G         2888.0442   6   5788.0884     1.6777     0.1262     0.5687
K81uf+G       2888.1130   6   5788.2261     1.8154     0.1178     0.6865
TrN+I+G       2887.6951   7   5789.3901     2.9795     0.0658     0.7523
K81uf+I+G     2887.7705   7   5789.5410     3.1304     0.0610     0.8133
TIM+G         2887.9448   7   5789.8896     3.4790     0.0513     0.8646
TVM+G         2887.4417   8   5790.8833     4.4727     0.0312     0.8958
....
```

LRT calculator Mode (-l)
This is a useful mode when the user wants to calculate likelihood ratio tests and their associate probability for different hypotheses, or only some of the hypothesis tested by default by Modeltest. As the P-value is calculated using a chi-square distribution, **the models tested should be nested** (the null hypothesis is a special case of the alternative hypothesis). The user is guided by prompts for inputting a pair of likelihood scores and the number of degrees of freedom. Modeltest performs the likelihood ratio test and calculates its associated P-value using a chi-square distribution. In this case the user is responsible for calculating the appropriate number of degrees of freedom and log likelihood scores for testing the intended hypotheses.

Help (-?)
By typing "-?" in the command line, the user can have access to some help. Nevertheless, the lecture of the manual is encouraged…

Debug level (-d)
This is a programming feature that does not affect the calculations. This option is for development, and not for usage, of the program. If you don't know what is this, don't worry; simply do not use it…

Output
The output of MODELTEST consists of a description of the likelihood ratio tests performed, and their associated P-values. The null hypotheses tested are described in Figure 1. The program interprets the resulting P-values and chooses the model that fits the data best among those tested following the likelihood ratio test, AIC or BIC criteria, using a default individual alpha value of 0.01 (for maintaining an overall alpha value of 0.05, the standard Bonferroni correction -alpha/number of tests-results in a individual alpha value of 0.01), or another value specified by the user. The program also calculates AIC values, indicating the smallest, the AIC differences (deltas) and the Akaike weights In addition, it will use those weights to calculate model-averaged estimates and parameter importance. All these calculations are also implemented for the BIC. Note that when the equal ti/tv (transition/transversion) rates hypothesis is not rejected, the equal ti and equal tv rates hypotheses are automatically rejected, and then will not be tested. See also Posada (2003a; 2003b).

Often, the hLRT and the AIC result in the selection of different models. **It is up to the user to decide which model selection criteria is going to use.**

PACKAGE
The MODELTEST package includes several files in different subdirectories:

README.html: quick instructions and comments for the users.
/paupblock/modelblockPAUPb10: the batch file with PAUP* commands to obtain likelihood scores for the competing models in the proper format for MODELTEST
/bin/Modeltest3.7.macX.app: A Macintosh (OS X) executable
/bin/Modeltest3.7.mac9.app: A Macintosh (OS 9) executable
/bin/Modeltest3.7.win.exe: A Windows executable
/doc/Modeltest3.7.pdf: Documentation in PDF format
/license/gpl.html: GNU general public license in HTML format
/examples/example.pauplog: a log file describing the calculations performed by PAUP* to obtain *example.scores*
/examples/example.nex: an example data file in NEXUS format
/examples/example.modeltest.out: the output file of MODELTEST resulting from the analysis of *examples.scores*.
/examples/example.mac.scores: file with likelihood scores produces by *PAUP after loading *examples.nex* and executing the *modelblockPAUPb10*batch file. In macintosh format.
/examples/example.unix.scores: file with likelihood scores produces by *PAUP after loading *examples.nex* and executing the *modelblockPAUPb10*batch file. In unix format.
/examples/example.win.scores: file with likelihood scores produces by *PAUP after loading *examples.nex* and executing the *modelblockPAUPb10*batch file. In windows format.
/source/Makefile: Makefile for compilation of MODELTEST in UNIX-like environments
/source/Modeltest3.7.c: C source code

Example file
The example file (*example.nex*) included in MODELTEST it is a simulated data set with 10 aligned DNA sequences 1000 bp long. This alignment was simulated on a tree obtained from coalescent process and under the HKY+I model, with the next parameter values:

> Effective population size = 10000
> Mutation rate per nucleotide per site = 5e-5
> Base frequencies (A, C, G, T) = 0.4, 0.2, 0.1, 0.3
> Transition/transversion rate = 4
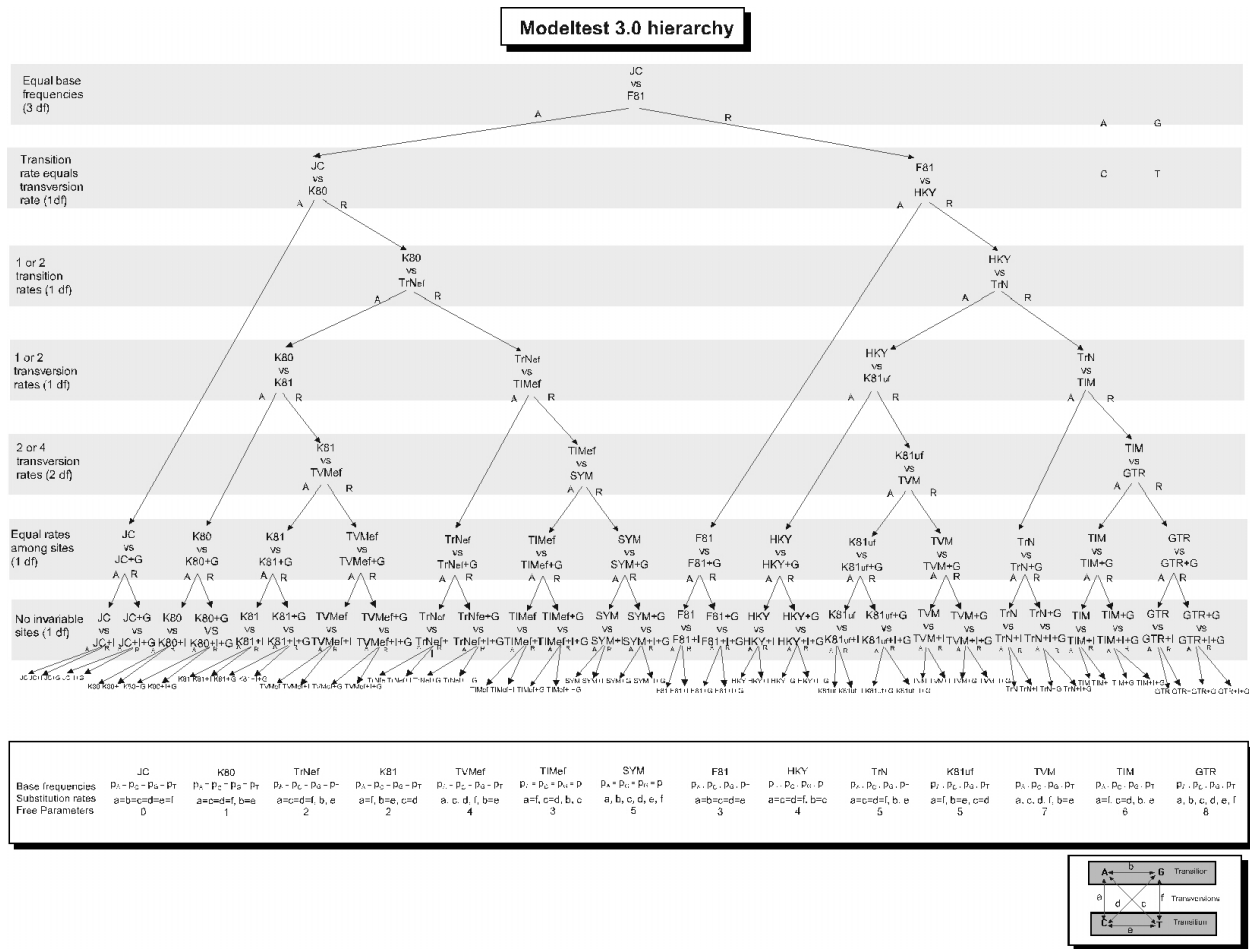> Alpha parameter of the gamma distribution = 0.4

Modeltest hierarchy

Figure 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes, Cantor, 1969), K80 (Kimura, 1980), TrNef (TrN equal base frequencies; see below), K81 (Kimura, 1981), TIMef (TIM with equal base frequencies), TIV (TIV with equal base frequencies), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa et al., 1985), TrN (Tamura, Nei, 1993), K81uf (K81 unequal base frequencies; see above), TIM, TIV, and GTR (Tavaré, 1986). G: shape parameter of the gamma distribution; I: proportion of
invariable sites. df: degrees of freedom.

10

Table 1. Model names. Some models have no reference (TNef, K81uf, TIMef, TIM, TVMef, TVM), they are just some variations of some existing models, and they were no developed, only named, by D. Posada.

Model	Name
JC	Jukes and Cantor (Jukes and Cantor, 1969)
F81	Felsenstein 81 (Felsenstein, 1981)
K80	Kimura 80 (=K2P) (Kimura, 1980)
HKY	Hasegawa, Kishino, Yano 85 (Hasegawa, Kishino and Yano, 1985)
TNef	Tamura-Nei equal frequencies
TN	Tamura-Nei (Tamura and Nei, 1993)
K81	Two transversion-parameters model 1 (=K81=K3P) (Kimura, 1981)
K81uf	Two transversion-parameters model 1 unequal frecuencies
TIMef	Transitional model equal frequencies
TIM	Transitional model
TVMef	Transversional model equal frequencies
TVM	Transversional model
SYM	Symmetrical model (Zharkihk, 1994)
GTR	General time reversible (=REV) (Tavaré, 1986)

Table 2. Model parameters. The substitution codes are just two ways of indicating the substitution scheme. Any of these models can ignore rate variation or include invariable sites (+I), rate variation among sites (+G), or both (+I+G).

Model	Free parameters	Base frequencies	Substitution rates	Substitution code 1	Substitution code 2
JC	0	equal	a=b=c=d=e=f	000000	aaaaaa
F81	3	unequal	a=b=c=d=e=f	000000	aaaaaa
K80	1	equal	a=c=d=f, b=e	010010	abaaba
HKY	4	unequal	a=c=d=f, b=e	010010	abaaba
TNef	2	equal	a=c=d=f, b, e	010020	abaaca
TN	5	unequal	a=c=d=f, b, e	010020	abaaca
K81	2	equal	a=f, c=d, b=e	012210	abccba
K81uf	5	unequal	a=f, c=d, b=e	012210	abccba
TIMef	3	equal	a=f, c=d, b, e	012230	abccda
TIM	6	unequal	a=f, c=d, b, e	012230	abccda
TVMef	4	equal	a, c, d, f, b=e	012314	abcdbe
TVM	7	unequal	a, c, d, f, b=e	012314	abcdbe
SYM	5	equal	a, c, d, f, b, e	012345	abcdef
GTR	8	unequal	a, c, d, f, b, e	012345	abcdef

Command line
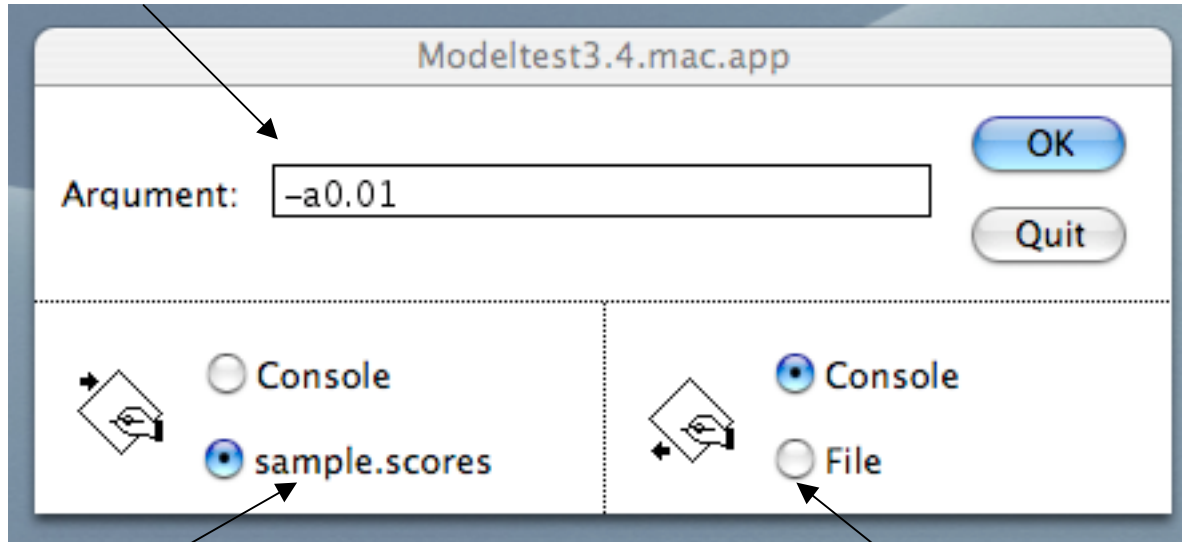
Select input file   Select output file
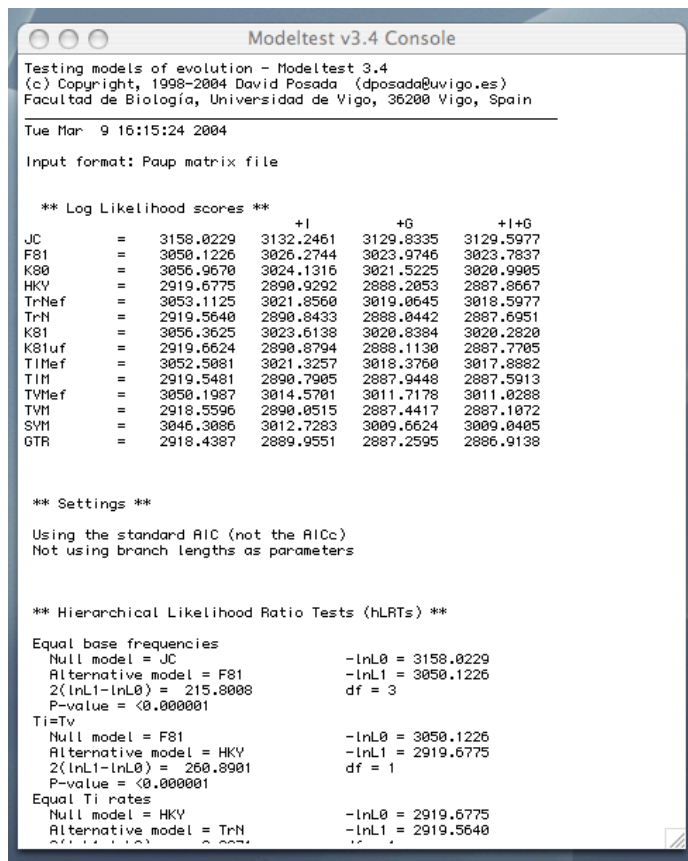
Figure 2. Console Interface for Macintosh

Figure 3. Macintosh output

Figure 4. Command prompt in Windows (you need to run the program from a terminal window)

Figure 5. Unix console (Unix, Linux, MacOS X console...)

RUNNING THE PAUP COMMANDS BLOCK
The input of Modeltest is likelihood scores corresponding to the specific data set and each one of 40 models. The easiest way of obtaining these scores from an alignment of DNA sequences is using PAUP*. A block of commands for PAUP* is provided below in the commands file ("modelblock3"). Follow these steps:

1) Open your data file and execute it in PAUP
2) Open the command file (modelblock) and execute it
3) Paup starts to work in the data following the commands. Once is finished you will see a file called "model.scores" in the same directory as the command file
4) Run the file "model.scores" through modeltest

Alternatively, if you are familiar with PAUP, you can add the PAUP* commands after your data block directly in your data file and execute it.*

REFERENCES

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.
Burnham KP, Anderson DR (1998) *Model selection and inference: a practical information-theoretic approach* Springer-Verlag, New York, NY.
Burnham KP, Anderson DR (2003) *Model selection and multimodel inference: a practical information-theoretic approach* Springer-Verlag, New York, NY.
Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376.
Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**, 521-565.
Forster MR, Sober E (2004) Why likelihood? In: *Likelihood and Evidence* (eds. Taper M, Lele S). University of Chicago Press, Chicago.
Goldman N (1993a) Simple diagnostic statistical test of models of DNA substitution. *J. Mol. Evol.* **37**, 650-661.
Goldman N (1993b) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182-198.
Goldman N, Whelan S (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol Biol Evol* **17**, 975-978.
Hasegawa M, Kishino K, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160-174.
Hsiao CK (1997) Approximate Bayes factors when a mode occurs on the boundary. *Journal of the American Statistical Association* **92**, 656-663.
Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* **28**, 437-466.
Hurvich CM, Tsai C-L (1989) Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.
Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism* (ed. Munro HM), pp. 21-132. Academic Press, New York, NY.
Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928-934.
Kendall M, Stuart A (1979) *The Advanced Theory of Statistics*, 4th edn. Charles Griffin, London.
Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111-120.
Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, U.S.A.* **78**, 454-458.

Kullback S, Leibler RA (1951) On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79-86.

Ohta T (1992) Theoretical study of near neutrality. II. Effect of subdivided population structure with local extinction and recolonization. *Genetics*, 917-923.

Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H (2000) Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution* **17**, 798-803.

Posada D (2001) The effect of branch length variation on the selection of models of molecular evolution. *J Mol Evol* **52**, 434-444.

Posada D (2003a) Selecting models of evolution. In: *The Phylogenetic Handbook* (eds. Vandemme A, Salemi M), pp. 256-282. Cambridge University Press, Cambridge, UK.

Posada D (2003b) Using Modeltest and PAUP* to select a model of nucleotide substitution. In: *Current Protocols in Bioinformatics* (eds. Baxevanis AD, Davison DB, Page RDM*, et al.*), pp. 6.5.1-6.5.14. John Wiley & Sons, Inc.

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**, 793-808.

Posada D, Crandall KA (2001a) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* **18**, 897-906.

Posada D, Crandall KA (2001b) Selecting the best-fit model of nucleotide substitution. *Systematic Biology* **50**, 580-601.

Raftery AE (1999) Bayes Factors and BIC: Comment on "A critique of the Bayesian Information Criterion for model selection". *Sociological Methods & Research* **27**, 411-427.

Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics–Theory and Methods* **A7**, 13-26.

Swofford DL (1998) PAUP* Phylogenetic analysis using parsimony and other methods. Sinauer Associates, Sunderland, MA.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512-526.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: *Some mathematical questions in biology - DNA sequence analysis* (ed. Miura RM), pp. 57-86. Amer. Math. Soc., Providence, RI.

Wasserman L (2000) Bayesian Model Selection and Model Averaging. *J Math Psychol* **44**, 92-107.

Weakliem DL (1999) A Critique of the Bayesian Information Criterion for Model Selection. *Sociological Methods & Research* **27**, 359-397.

Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Molecular Biology and Evolution* **16**, 1292-1299.

Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution* **42**, 587-596.

Yang Z, Goldman N, Friday A (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **44**, 384-399.

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol* **39**, 315-329.