

## **Selecting a model of nucleotide substitution**

David Posada

*Variagenics Inc., Cambridge, MA 02139, USA*

*and*

*Center for Cancer Research, Massachusetts Institute of Technology,  
Cambridge, MA 02139, USA*

[dposada@variagenics.com](mailto:dposada@variagenics.com)

### **Abstract**

Models of nucleotide substitution are commonly used in the analysis of DNA sequences. In this unit a protocol is described for the use of the program MODELTEST (coupled with PAUP\*) to find the best-fit model of substitution for the sequence alignment at hand. An example data file is analyzed and the interpretation of the results is discussed. Some background theory on model selection and a discussion of the relevance of models is included at the end of the unit.

## **Choosing the best-fit model of nucleotide substitution**

Models of nucleotide substitution are commonly used in the analysis of DNA sequences, especially in the estimation of evolutionary parameters and in the construction of phylogenetic trees. A common problem that a researcher has to face is the objective selection of such a sample. The program MODELTEST (Posada and Crandall, 1998) offers two sound statistical procedures to carry out this choice.

This protocol describes the joint use of the program MODELTEST and PAUP\* (Swofford, 2000) to select the best-fit model of nucleotide substitution for a given set of aligned DNA sequences. The three basic steps within this protocol consist of the estimation of a phylogenetic tree, the estimation of likelihood scores for the candidate models, and the selection of a particular model given these likelihood scores. The first two procedures are carried in the program PAUP\* with a batch file included in the MODELTEST package. The discussion in this unit summarizes the theory behind the procedure and provides several caveats and guides for interpretation of the results. Particular comments about the implementation different platforms are also given. Users are encouraged to download the most current versions from the MODELTEST and PAUP\* Web site

## **USING MODELTEST AND PAUP\* TO SELECT A MODEL OF NUCLEOTIDE SUBSTITUTION**

### **Necessary Resources**

#### *Hardware*

The program MODELTEST is a standalone application compiled for Linux, Unix, Windows and Macintosh platforms. The C code is freely available, and therefore the program can be compiled in any platform and ran in any computer. The program PAUP\* is also available in a variety of platforms.

#### *Software*

MODELTEST was developed to work with the output provided by the program PAUP\* although alternative programs might be used for the calculation of likelihood scores. The interaction between PAUP\* and MODELTEST is described in Figure 6.5.1. MODELTEST and PAUP can be

obtained from the web sites described at the end of this unit. MODELTEST is freely available, while PAUP\* is commercial software licensed by Sinauer.

### *Data Files*

The initial data file with aligned DNA sequences should be in any of the formats recognized by PAUP\* (e.g., NEXUS, PHYLIP, PIR). Figure 6.5.2 shows an example file in sequential NEXUS format. A PAUP\* batch file is included in the MODELTEST package in order to obtain an output file from PAUP\* that is becomes the input for MODELTEST. The MODELTEST package includes several files in different subdirectories:

*README.html*: quick instructions and comments for the users.  
*/batch/modelblock*: the batch file with PAUP\* commands to obtain likelihood scores for the competing models in the proper format for MODELTEST  
*/bin/Modeltest3.1.mac*: A Macintosh (OS 9) executable  
*/bin/Modeltest3.1.macX*: A Macintosh (OS X) executable  
*/bin/Modeltest3.1.win.exe*: A Windows executable  
*/doc/Modeltest3.1.pdf*: Documentation in PDF format  
*/license/gpl.html*: GNU general public license  
*/sample/sample.nex*: an example data file in NEXUS format  
*/sample/sample.scores*: file with likelihood scores produces by \*PAUP after loading *sample.nex* and executing the *modelblock* batch file.  
*/sample/sample.log*: a log file describing the calculations performed by PAUP\* to obtain *sample.scores*  
*/sample/sample.out*: the output file of MODELTEST resulting from the analysis of *sample.scores*.  
*/source/modeltest3.1.c*: ANSI C source code  
*/source/Makefile*: Makefile for compilation of MODELTEST in UNIX-like environments

The example file (*sample.nex*) included in MODELTEST and used also as an example here 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

### **Loading the data file in PAUP\***

PAUP\* is commonly used in Macintosh OS 9, where a complete GUI application is available. In Windows PAUP\* has a much simpler GUI, but enough for the purposes here. This part protocol will refer to the GUI versions of PAUP\* (Macintosh and Windows), but it will also describe how the same can be accomplished in the command line versions (Unix-like environments).

1. Start PAUP\* by double-clicking on the application file

In Unix type *paupx.x* in the command line.

2. Select Open from the File Menu in the PAUP\* interface.

*In the browser window that appears set the Initial mode to Execute, and then select your data file (in this example use *sample.nex*), which should be in NEXUS format. In Unix type *execute sample.nex*. PAUP\* will display information indicating whether the file has been correctly processed. If the file is not in NEXUS format, the GUI version of PAUP\* can still import it if it is in PHYLIP, MEGA, PIR or MSF format (other less common formats are also admitted). Go to the File Menu and select Import Data. Set the options to the proper Format, Data type and check the Interleaved button if your sequences are in interleaved format (see elsewhere). The file will be then opened automatically in the \*PAUP editor in NEXUS format. You can now save this file (File > Save) or execute it (File > Execute).*

### **Executing the *modelblock* batch file in PAUP\***

3. Select Open from the File Menu in the PAUP\* interface. This will make PAUP to run for some time (minutes to hours) and to generate a file with likelihood scores and parameter estimates (*sample.scores*) (Figure 6.5.3).

In the browser window that appears set the Initial mode to Execute, and then select the batch file *modelblock*. In Unix type *execute modelblock*. PAUP\* will start processing the commands *modelblock* first

estimate a neighbor-joining tree, and then to calculate likelihood scores for 56 substitution models. The likelihood calculations can take from a few minutes (like for *sample.nex*) to even days, depending on the complexity of the data. Remember that at this point you are just running PAUP\* (you have not used MODELTEST yet). Two files will be created during this run in the PAUP\* directory. The file *sample.scores* will be the input file for MODELTEST, and it is often convenient to rename it to *yourdata.scores*. The other file produced by PAUP\*, *sample.log*, it is just intended to check that everything has proceeded normally.

## Running MODELTEST

Again, this part protocol will refer to the GUI versions of MODELTEST (Macintosh), but it will also indicate how the command-line versions of MODELTEST can be used (Windows and Unix-like environments).

4. Double click the MODELTEST application file. This will bring up a GUI (Figure 6.5.4) where the input and output files can be selected, and where arguments can be specified.
5. Click on the left File button. This will bring up a browser window to select the file with the likelihood scores (i.e., *sample.scores*)
6. Click on the right File button. This will bring up a browser window to select the file where MODELTEST will print out the results (e.g., *yourdata.modeltest.out*). Alternatively, the output can be saved after execution.
7. Type in the Argument line any of the arguments to change the default options. Possible arguments (explained in the program documentation) are:
  - a : alpha level for the LRT tests (e.g. -a0.01)
  - c : Likelihood Ratio calculator mode
  - i : AIC calculator mode
  - f : Input from a file for AIC calculation
  - ? : Help
8. Click on the Run button. The analysis will be finished in a few seconds. If you have not previously selected an output file, go to the File menu and select Save. The MODELTEST analysis is done.

In Unix and Windows MODELTEST has to be executed from the command line. In the case of Windows this command line is accessed through a DOS window (MS-prompt, Console). Consult specific Windows documentation for information on command-line features. To automatically carry out steps 4-7 of this protocol in Windows or Unix type in the command line *modeltest3.1 < sample.scores > modeltest.out*. Options can be entered, for example, by typing *modeltest3.1 -a0.01 < sample.scores > modeltest.out*.

## THE MODELTEST OUTPUT

The first section of the MODELTEST output consists of the log likelihood scores from the file *sample.scores*. This feature provides an easy check of the procedure and a simultaneous display of the likelihoods for all the 56 substitution models compared in MODELTEST 3.1 (Figure 6.5.5). The other sections of the output relates to two main statistical procedures to select a best-fit model, the hierarchical likelihood ratio tests (hLRT) and the Akaike Information criterion (AIC).

In the case of the hLRT procedure (Figure 6.5.6), MODELTEST prints first information about the likelihood ratio tests performed, including the corresponding P-value. Next, MODELTEST describes the model selected, indicating the values of the parameter of this sample. It is important to note that these estimates were obtained by PAUP\*, not by MODELTEST. MODELTEST does not estimate any parameter. Finally, MODELTEST prints out a block of commands for PAUP\*. These commands can be very useful if the user wants to use the best-fit model for further analyses, for example to estimate a phylogeny (see elsewhere in the Unit) in PAUP\*. In order to use it, this command block should be copied and pasted after the data block in the NEXUS file (i.e., *sample.nex*). When such file is then executed in PAUP\*, these commands will automatically set up the selected sample.

In the AIC section (Figure 6.5.7) MODELTEST prints again a description of the selected model and a block of PAUP\* commands. The last part of the AIC output is the AIC differences (called deltas) and the Akaike weights for every model, which can be used to get an idea of model selection uncertainty. The models are ordered according to their Akaike weights, and the first model it is always the model selected by the AIC.

## INTERPRETING MODELTEST RESULTS

There are two different statistical approaches to model selection implemented in MODELTEST (LRT and AIC), and in some cases they will indeed select different models. Believing more the results from one or the other technique is really up to the philosophy of the particular user, and such a choice should be made upon a good understanding of both approaches. An introduction to these model selection strategies and some useful references can be found at the end of the unit.

## Results from the example data set (*sample.nex*)

The hierarchical likelihood ratio tests selected the HKY+ $\Gamma$  (= HKY+G) model as the best-fit model for the *sample.nex* data set (Figure 6.5.6). There were a total of six LRT performed, two of them rejected, and four of them accepted. The hypothesis tested by each LRT is indicated, as well as the log likelihood corresponding to the models contrasted, the number of degrees of freedom, the value of the LRT and the associated P-value. Parameter estimates corresponding to the best-fit model (these estimates were calculated by PAUP\*) are also indicated. If the goal is to estimate a phylogenetic tree, these parameters can be fixed and a heuristic search performed under the maximum likelihood criterion. The commands require to load this model in PAUP are below in the MODELTEST window (not shown in this figure, but see Figure 6.5.7).

The AIC criterion also selected the HKY+ $\Gamma$  model as the best-fit model (Figure 6.5.7). The corresponding log likelihood, AIC and parameters estimates are also indicated. In Figure 6.5.7 we can also see the set of PAUP\* commands that the user can attach to its NEXUS file to specify this model in PAUP\* for further analysis. The Akaike weights are also indicated. In this case we can see that there is not that much confidence in the best-fit model, but it rather seems that there is a set of perhaps four plausible models (models having deltas ( $\Delta_i$ ) within 1-2 of the best model have substantial support; see the background section) that could be considered.

In this case both strategies have identified the *true model* as the best-fit model (remember that the *sample.nex* data set was simulated under the HKY+ $\Gamma$  model). In general, this seems the case with model selection strategies (Posada, 2001; Posada and Crandall, 2001b), which suggests that model selection strategies can successfully identify features from the data. However, a very important consideration is that in real life the true model will not be one of our candidate models. We are trying to approximate an unknown complex model, not to find the true sample.

## COMMENTARY

### Background Information

## Selecting Models of Evolution

Phylogenetic estimation is a problem of statistical inference. When using DNA sequences to estimate phylogenetic relationships we need to specify some probability model that describes the different probabilities of change from one nucleotide to another. However, this model is not always made explicit, like in the case of maximum parsimony. Indeed, models are not just interesting from the point of phylogenetic reconstruction, but also because models of substitutions are themselves descriptions of the evolutionary process at the molecular level. Importantly, models do not try to mimic the exact process of molecular evolution, which indeed is unknown, but they rather try to approximate it.

Common models of nucleotide substitution include parameters that describe base frequencies, the substitution rates among the four nucleotides or the distribution of the rate of evolution among sites (rate variation) and among lineages (the molecular clock). Some models assume that the base frequencies are equal, while other models allow them to vary freely, with the only constraint that they have to add to 1. When reversibility of change is assumed, i.e., the probability of changing from nucleotide  $i$  to nucleotide  $j$  is the same as the probability of changing from nucleotide  $j$  to nucleotide  $i$ , there are six possible substitution rates among the four nucleotides ( $r_{AC}$ ,  $r_{AG}$ ,  $r_{AT}$ ,  $r_{CG}$ ,  $r_{CT}$ ,  $r_{GT}$ ). Complex models allow these six rates to vary freely, while less complex models place some constraints on their variation (for example, that all transitions and all transversions have the same rate;  $r_{AG} = r_{CT} \neq r_{AC} = r_{AT} = r_{CG} = r_{GT}$ ). Rate variation among sites can be included in the model by simply assuming that there is a proportion of invariable sites ( $p$ -inv) while the rest of the sites evolve at the same rate. Or we can assign to each site a certain probability of belonging to a specific rate category. This set of probabilities is conveniently described by a discrete gamma distribution with four categories ( $\Gamma$ ) (Yang, 1994; Yang, 1996). When the shape parameter of the gamma distribution ( $\alpha$ ) is small most of the sites evolve very slowly, but a few sites have moderate-to-fast rates. When  $\alpha$  increases, most of the sites evolve at medium rates, and a few at slow and fast rates. When  $\alpha$  is infinity, all the sites evolve at the same rate. Also, we can consider that rates of evolution may be different in different parts of the tree. Figure 6.5.5 describes the basic substitution models included in MODELTEST. For those interested in details of the models, Swofford et al. (Swofford et al., 1996) provide a comprehensive review of common models. More complex models indeed exist (e.g. Goldman and Yang, 1994; Huelsenbeck, 2002;

Huelsenbeck and Nielsen, 1999; Muse and Gaut, 1994; Schöniger and von Haeseler, 1994; Thorne et al., 1998; Tuffley and Steel, 1998), but their application is still uncommon, mainly because these models have not yet been implemented in software packages.

## Goodness of fit

It is important to understand how well the models we use to make inferences fit the data. A way of assessing the fit of a single model to the data is to calculate the maximum value of the likelihood function under the multinomial distribution as an upper bound to which the likelihood of any model can be compared as a test for model fit (Goldman, 1993). The likelihood function under the multinomial distribution refers to an unconstrained model of evolution, and for  $n$  aligned DNA sequences of length  $N$  sites (excluding gapped sites) it has the form

$$L = \prod_{b \in \Omega} (p_b)^{n_b}$$

where  $\Omega$  is a set of  $4^N$  possible nucleotide patterns that may be observed at each site,  $p_b$  is the probability that any site exhibits the pattern  $b$  in  $\Omega$  given the tree and a substitution model, and  $n_b$  is the number of times the pattern  $b$  is observed out of the  $N$  sites. We should realize, however, that this test is very stringent, and most models offer a significantly worse fit than the multinomial. This does not imply that the models we use today are inadequate to provide reasonable estimates, but rather that current models do not provide a perfect description of the underlying evolutionary process. Since we never expect a model of evolution to be correct in every detail, this test is perhaps best used to estimate how far the assumed model deviates from the underlying process that generated the data (Swofford et al., 1996).

## The Likelihood Ratio Test

The likelihood ratio test (LRT) statistic is one of the most widely used tools for comparing the fit of two competing models:

$$\text{LRT} = 2 (\ln L_1 - \ln L_0)$$

Here  $L_1$  is the likelihood maximized under the more complex model (which is the alternative hypothesis) and  $L_0$  is likelihood maximized under a simpler model (null hypothesis). The value of the LRT is always equal to or greater than zero, even if the simpler model is the true one, simply because the superfluous parameters in the complex model will always provide a better explanation of the stochastic variation in the data than the simpler sample. When the models compared are nested (the simple model is a special case of the complex model) twice this statistic is asymptotically distributed as  $\chi^2$  with a number of degrees of freedom equal to the difference in number of free parameters between the two models. When the P-value associated to the LRT is significant (i.e. smaller than some predetermined value, like 0.05) we conclude that the additional parameters in the complex model increase significantly the fit to the data. On the other hand, a LRT close to zero suggests that the complex model does not fit the data significantly better than the simple sample. The  $\chi^2$  distribution approximation for the LRT statistic is not appropriate when the null model is equivalent to fixing some parameter at the boundary of its parameter space in the alternative model (Whelan and Goldman, 1999). In this case, the use of a mixed  $\chi^2$  distribution (50%  $\chi^2_0$  and 50%  $\chi^2_1$ ) is appropriate. An example of this situation is a LRT between two models that differ only in that the complex model includes a parameter for the proportion of invariable sites, which ranges from 0 to 1. The simple model is a special case of the complex model where the proportion of invariable sites is fixed to 0, which is at the boundary of the range of the parameter in the complex sample. The use of LRTs in phylogenetics is reviewed by Huelsenbeck and Crandall (1997) and Huelsenbeck and Rannala (1997).

## **Hierarchical Likelihood Ratio Tests**

We can consider that when we compare two different nested models through a LRT we are actually testing hypotheses about our data, represented by the difference in the assumptions among the models compared. Therefore, we think of testing several hypotheses in a hierarchical manner to “arrive” to the best-fit model for the data set at hand (Fratini et al., 1997; Huelsenbeck and Crandall, 1997; Posada and Crandall, 1998). For example, to test the equal base frequencies hypothesis, we could do a LRT comparing JC vs. F81, as these models only differ in the fact that F81 allows for unequal base frequencies (alternative hypothesis), while JC assumes equal base frequencies

(null hypothesis). This is one of the strategies implemented in MODELTEST (Figure 6.5.8).

### **Akaike Information Criterion**

The Akaike information criterion (AIC, Akaike, 1974) is an asymptotically unbiased estimator of the Kullback-Leibler information quantity (Kullback and Leibler, 1951), which is a measure of the information that is lost when a model is used to approximate full reality. The smaller the AIC, the better the fit of the model to the data. This 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. It is computed as:

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

where  $N_i$  is the number of free parameters in the  $i$ th model and  $L_i$  is the maximum-likelihood value of the data under the  $i$ th sample.

The AIC offers several interesting advantages. First, the AIC compares several candidate models simultaneously (while the LRT is a pairwise comparison). Second, it can be used to compare both nested and non-nested models. And third, model-selection uncertainty can be easily quantified using the AIC differences and Akaike weights. AIC differences ( $\Delta_i$ ) are rescaled AICs, where the model with the minimum AIC has a value of 0:

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

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  $\Delta_i$  within 1-2 of the best model have substantial support and should receive consideration. Models having  $\Delta_i$  within 3-7 of the best model have considerably less support, while models with  $\Delta_i > 10$  have essentially no support (Burnham and Anderson, 1998). Akaike weights are the normalized relative AIC for each candidate model, and can be interpreted as the probability that a model is the best approximation to the truth given the data:

$$w_i = \frac{\exp\left[-\frac{1}{2}\chi^2_i\right]}{\sum_{r=1}^R \exp\left[-\frac{1}{2}\chi^2_r\right]}$$

for  $R$  candidate models. A very interesting application of the Akaike weights is that inference can be averaged from those models where the Akaike weights are nontrivial. For example, a model-averaged estimate of  $\gamma$  (the shape of the gamma distribution for rate variation among sites) would be:

$$\hat{\gamma} = \sum_{i=1}^R w_i \hat{\gamma}_i$$

Indeed, we could also think of estimating phylogenies under the best models and combine these trees according to their Akaike weights. Burnham and Anderson (1998) provide an excellent introduction to the AIC and model selection.

## The Importance of Models Selection

The relevance of models of nucleotide substitution in evolutionary studies has been extensively discussed. It is clear that the use of one model of evolution or another may change the outcome of the phylogenetic analysis (Cunningham et al., 1998; Kelsey et al., 1999; Leitner et al., 1997; Posada and Crandall, 2001a; Sullivan and Swofford, 1997). The performance of a method is maximized when its assumptions are satisfied, and therefore some indication of the fit of the data to the phylogenetic model is necessary (Huelsenbeck, 1995). In general, phylogenetic methods perform worse when the model of evolution assumed is incorrect (Bruno and Halpern, 1999; Felsenstein, 1978; Huelsenbeck, 1995; Huelsenbeck and Hillis, 1993). Cases where the use of wrong models increases phylogenetic performance (see Posada and Crandall, 2001c; Xia, 2000; Yang, 1997) are exceptional and rather represent a bias towards the true tree associated with violated assumptions (Bruno and Halpern, 1999). However, the relationship between the fit of the model to the data and the ability of the model to correctly predict topology is not completely straightforward (Buckley, 2002; Gaut and Lewis, 1995). Topology estimation by methods such as maximum-likelihood is relatively robust to the model used (Fukami-Kobayashi and Tateno, 1991; Sullivan and Swofford, 2002; Yang et al., 1995). The evaluation of reliability of the estimated trees depends critically on the model; false or simple

models tend to suggest that a tree is significantly supported when it cannot be (Buckley and Cunningham, 2002; Buckley et al., 2001; Yang et al., 1994). The use of appropriate models is especially critical for parameter estimation. When a relatively simple model of substitution is assumed, the transition/transversion ratio, branch lengths, and sequence divergence are underestimated, while the shape parameter of the gamma distribution is overestimated (Adachi and Hasegawa, 1995; Buckley et al., 2001; Tamura, 1992; Wakeley, 1994; Yang et al., 1994; Yang et al., 1995). Moreover, the outcome of different tests of evolutionary hypotheses, like the molecular clock, may depend on the model of evolution assumed (Zhang, 1999). Also, simple models of substitution can increase the number of false positives when comparing tree topologies (Buckley, 2002).

### **Critical Parameters and Troubleshooting**

MODELTEST is a fairly simple program, and the other parameter that can be set is the alpha level for the individual LRT tests, which by default is set to 0.01. This value aims to provide a family alpha level of 0.05, and results from a standard Bonferroni correction considering that there will be 5 or 6 LRT performed (the exact number of LRT cannot be known *a priori*). Sometimes problems arise because of the interaction with PAUP\*, and the user should check PAUP\* web site in case of problems with the format of the likelihood scores file (*sample.scores*). It is important to distinguish whether we are running PAUP\* or MODELTEST. To keep up to date with problems and or updates of MODELTEST, the user needs to register its copy.

### **Literature Cited**

- Adachi, J. and Hasegawa M. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J. Mol. Evol.* 40:622-628.
- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* 19:716-723.
- Bruno, W.J. and Halpern A.L. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* 16:564-566.
- Buckley, T.R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51:509-523.

- Buckley, T.R. and Cunningham C.W. 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Mol. Biol. Evol.* 19:394-405.
- Buckley, T.R., Simon C., and Chambers G.K. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: The effects of model assumptions on estimates of topology, edge lengths, and bootstrap support. *Syst. Biol.* 50:67-86.
- Burnham, K.P. and Anderson D.R. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, NY.
- Cunningham, C.W., Zhu H., and Hillis D.M. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* 52:978-987.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401-410.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368-376.
- Frati, F., Simon C., Sullivan J., and Swofford D.L. 1997. Gene evolution and phylogeny of the mitochondrial cytochrome oxidase gene in Collembola. *J. Mol. Evol.* 44:145-158.
- Fukami-Kobayashi, K. and Tateno Y. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J. Mol. Evol.* 32:79-91.
- Gaut, B.S. and Lewis P.O. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* 12:152-162.
- Goldman, N. 1993. Simple diagnostic statistical test of models of DNA substitution. *J. Mol. Evol.* 37:650-661.
- Goldman, N. and Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725-736.
- Hasegawa, M., Kishino K., and Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174.
- Huelsenbeck, J.P. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17-48.
- Huelsenbeck, J.P. 2002. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* 19:698-707.
- Huelsenbeck, J.P. and Crandall K.A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28:437-466.
- Huelsenbeck, J.P. and Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42:247-264.

- Huelsenbeck, J.P. and Nielsen R. 1999. Variation in the pattern of nucleotide substitution across sites. *J. Mol. Evol.* 48:86-93.
- Huelsenbeck, J.P. and Rannala B. 1997. Phylogenetic methods come of age: testing hypothesis in a evolutionary context. *Science* 276:227-232.
- Jukes, T.H. and Cantor C.R. 1969. Evolution of protein molecules. *In* Mammalian Protein Metabolism (H.M. Munro, eds.) pp. 21-132. Academic Press, New York, NY.
- Kelsey, C.R., Crandall K.A., and Voevodin A.F. 1999. Different models, different trees: The geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* 13:336-347.
- Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.
- Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 78:454-458.
- Kullback, S. and Leibler R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22:79-86.
- Leitner, T., Kumar S., and Albert J. 1997. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* 71:4761-4770.
- Muse, S.V. and Gaut B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11:715-724.
- 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. and Crandall K.A. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* 14:817-818.
- Posada, D. and Crandall K.A. 2001a. Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* 18:897-906.
- Posada, D. and Crandall K.A. 2001b. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50:1-22.
- Posada, D. and Crandall K.A. 2001c. Simple (wrong) models for complex trees: empirical bias. *Mol. Biol. Evol.* 18:271-275.
- Rodríguez, F., Oliver J.F., Marín A., and Medina J.R. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142:485-501.
- Schöniger, M. and von Haeseler A. 1994. A stochastic model for the evaluation of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* 3:240-247.

- Sullivan, J. and Swofford D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *J. Mamm. Evol.* 4:77-86.
- Sullivan, J. and Swofford D.L. 2002. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* 50:723-729.
- Swofford, D.L., Olsen G.J., Waddell P.J., and Hillis D.M. 1996. Phylogenetic Inference. *In* Molecular Systematics (D.M. Hillis, C. Moritz, and B.K. Mable, eds.) pp. 407-514. Sinauer Associates, Sunderland, MA.
- Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Mol. Biol. Evol.* 9:678-687.
- Tamura, K. and Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512-526.
- Thorne, J., Kishino H., and Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15:1647-1657.
- Tuffley, C. and Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci* 147:63-91.
- Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11:436-442.
- Whelan, S. and Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 16:1292-1299.
- Xia, X. 2000. Phylogenetic relationships among horseshoe crab species: Effect of substitution models in phylogenetic analysis. *Syst. Biol.* 49:87-100.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306-314.
- Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11:367-372.
- Yang, Z. 1997. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* 14:105-108.
- Yang, Z., Goldman N., and Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11:316-324.
- Yang, Z., Goldman N., and Friday A. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* 44:384-399.

- Zhang, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* 16:868-875.
- Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* 39:315-329.

### **Key References**

Burnham and Anderson. 1998. See above.

*This book provides a very clear and accessible explanation of different issues around model selection, particularly for the AIC. The book is written by ecologists and it includes many biological examples. A fundamental reference for any biologist doing data analysis.*

Swofford et al., 1996. See above.

*This chapter is still the most comprehensive review of phylogenetic inference today. It provides a detailed description of several substitution models and their use in phylogenetics.*

Posada and Crandall. 2001b. See above.

*A simulation study of the performance of different strategies for selecting models of substitution. Includes a detailed description of the different selection strategies.*

### **Internet Resources**

<http://www.evolgenics.com/software>

*MODELTEST Web site*

<http://paup.csit.fsu.edu/index.html>

*PAUP\* Web site*

## Figure legends

Figure 6.5.1. Flow chart of the model selection procedure using PAUP\* and MODELTEST.

Figure 6.5.2. Aligned sequences in sequential Nexus format in the file *sample.nex*. This file can be read by PAUP\*.

Figure 6.5.3. Likelihood scores for different models of substitution in the file *sample.scores*. This is the format required by MODELTEST.

Figure 6.5.4. MODELTEST 3.1 GUI in Macintosh OS 9. The input file can be chosen by selecting the left File button, while MODELTEST output can be redirected to a file by selecting the right File button. MODELTEST arguments can be entered in the Argument field.

Figure 6.5.5. Basic 14 models of substitution included in MODELTEST. Circles represent base frequencies, while rectangles represent symmetric substitution rates among nucleotides. Frequencies or rates with the same color are constrained within the same parameter. The number of free parameters for each model is the number of different colors minus two (because base frequencies have to add to one and the substitution rate  $G \leftrightarrow T$  is arbitrarily set to 1). Each one of these models can also include rate heterogeneity by specifying a proportion of invariable sites (+ $\square$ ), gamma distributed rates (+ $\square$ ), or both (+ $\square$ + $\square$ ). This makes  $14 \times 4 = 56$  models in MODELTEST. For example, there will be JC, JC+ $\square$ , JC+ $\square$ , and JC+ $\square$ + $\square$  models.

Figure 6.5.6. Section of the MODELTEST output corresponding to the hierarchical likelihood ratio tests (hLRT).

Figure 6.5.7. Section of the MODELTEST output corresponding to the Akaike Information Criterion (AIC).

Figure 6.5.8. MODELTEST hierarchical likelihood ratio tests. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and 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 and Nei, 1993), K81uf (K81 unequal base frequencies; see above), TIM, TIV, and GTR (Rodríguez et al., 1990).

G: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom.

model.nex



Execute NEXUS file with aligned sequences in PAUP\*



PAUP\* loads the data

Execute the batch file modelblock in PAUP\*



PAUP\* run produces

model.scores



Input the scores file in MODELTEST

Run MODELTEST



Intepret the results

BEST-FIT MODEL OF NUCLEOTIDE SUBSTITUTION

#NEXUS

Begin data;

Dimensions ntax=10 nchar=1000;

Format datatype=nucleotide gap=- missing=? matchchar=.;

Matrix

```
seq01  TACACCAATGTAATCTTTCCCTCTTAACTTGTCCCTCCTCCAACTTATTCTCTATATCGAGCCACTATAATAGACTAAGAATTACACAGACC
seq02  TACACTAATGTAATCTTTCCCTCTTAAATTGTCCCTCCTCCAACTTATTCTCTATATCGAGCCACTATGACAGACTAAAAATTACACAGACC
seq03  TACACCAATGTAATCTTTCCCTCATAAATTGTCCCTCCTCCAACTTATTCTCTATATCGAGCCACTATGACAGACGAAAAGATTACACAGACC
seq04  TACACCAATGTAATCTTTCCCTCTTAAATTGTCCCTCCTCCAACTTATTCTCTATATCGAGCCACTATGACAGACGAAAAGATTACACAGACC
seq05  TACACCAATGTAATCTTTCCCTCTTAAAGTTGTCCCTCCTCCAACTTATTCTCTATATCGAGCCACTATGACAGACTAAAAATTACACAGACC
seq06  TACACCAATGTAATCTTTCCCTCTTAAATTCGTCCCTCCTCCGACTCATTCTCTATATCGAGCCACTATGACAGACTAAAAATTACACAGACC
seq07  TACACTAATGTAATTTTTCCCTCTTAACTCGTCCTACCTCCAACTCATTCTCTATATCGAGCCACTATGAAAACCTAGAAATTCATTAGACC
seq08  TACACTAATGTAATTTTTCCCTCTTAACTCGTCCTACCTCCAACTCATTCTCTATATCGAGCCACTATGAAAACCTAGAAATTCATTAGACC
seq09  TACACTAATGTAATTTTTCCCTCTTAACTCGTCCTACCTCCAAATAATTTCTATATCGAGCCACTATGAAAACCTAGAAATTCATTAGACC
seq10  TACACCAATATAATTTTTCCCCCTAACTTATCCTACCCCCAACTTACATTCTAAAACGAGCCACTATGATAAATAAAGATTACACAGACC
```

;

End;

```

|Tree   -lnL
1   3158.02291033
Tree   -lnL   p-inv
1   3132.24597321   0.53267955
Tree   -lnL   gamma shape
1   3129.83347926   0.466880
Tree   -lnL   p-inv   gamma shape
1   3129.59756390   0.23243526   0.841659
Tree   -lnL   freqA   freqC   freqG   freqT
1   3050.12249205   0.37840456   0.21562890   0.11652509   0.28944145
Tree   -lnL   freqA   freqC   freqG   freqT   p-inv
1   3026.27436571   0.37492740   0.21681977   0.11676653   0.29148631   0.51718980
Tree   -lnL   freqA   freqC   freqG   freqT   gamma shape
1   3023.97469930   0.37458084   0.21701267   0.11673135   0.29167514   0.500169
Tree   -lnL   freqA   freqC   freqG   freqT   p-inv   gamma shape
1   3023.78378475   0.37445320   0.21706786   0.11672940   0.29174953   0.21858409   0.876024
Tree   -lnL   ti/tv ratio
1   3056.96706982   2.61371609
Tree   -lnL   ti/tv ratio p-inv
1   3024.13167094   3.15394853   0.57263799
Tree   -lnL   ti/tv ratio gamma shape
1   3021.52258096   3.19655170   0.363121
Tree   -lnL   ti/tv ratio p-inv   gamma shape
1   3020.99059503   3.23705392   0.31098888   0.796682
Tree   -lnL   freqA   freqC   freqG   freqT   ti/tv ratio
1   2919.67740232   0.40198211   0.20568791   0.10570152   0.28662846   2.69949302
Tree   -lnL   freqA   freqC   freqG   freqT   ti/tv ratio p-inv
1   2890.92916808   0.40044930   0.20602487   0.10601805   0.28750778   3.24067492   0.52310068
Tree   -lnL   freqA   freqC   freqG   freqT   ti/tv ratio gamma shape
1   2888.20529255   0.40093404   0.20596690   0.10586658   0.28723247   3.34235255   0.433724
Tree   -lnL   freqA   freqC   freqG   freqT   ti/tv ratio p-inv   gamma shape
1   2887.86667193   0.40083784   0.20604291   0.10592133   0.28719793   3.37341656   0.26050305   0.838836
Tree   -lnL   R(a)   R(b)   R(c)   R(d)   R(e)
1   3053.11248635   1.00000000   4.26891364   1.00000000   1.00000000   6.18942282

```

Modeltest3.1

Argument:

OK

Quit



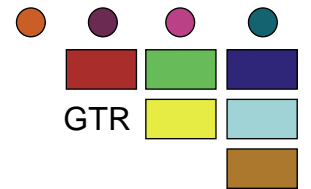
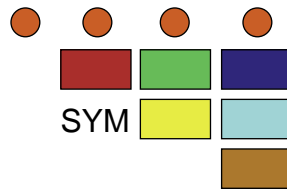
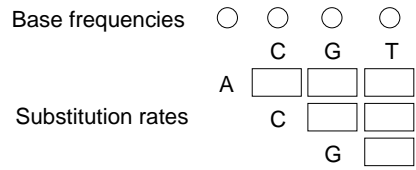
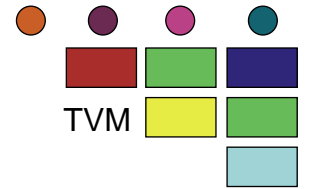
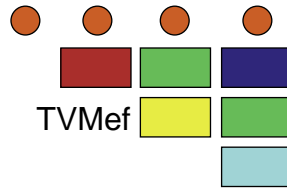
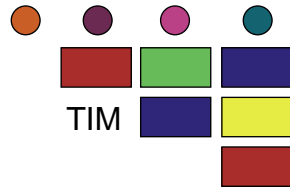
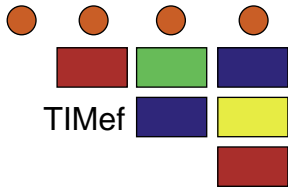
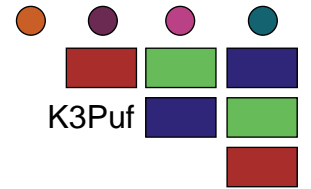
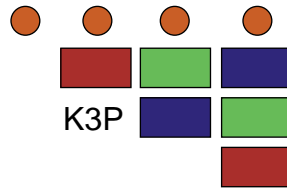
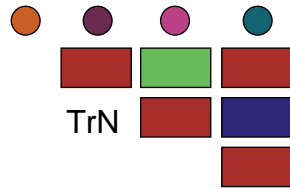
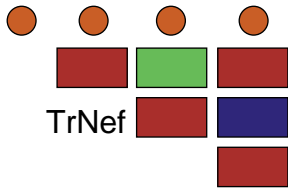
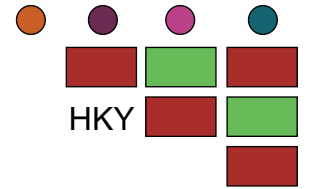
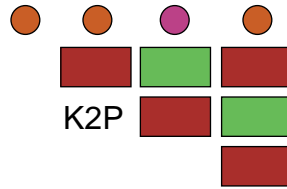
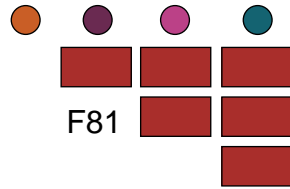
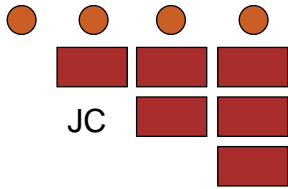
Console

File



Console

File



## \*\* Hierarchical Likelihood Ratio Tests (hLRTs) \*\*

## Equal base frequencies

Null model = JC	-lnL0 = 3158.0229
Alternative model = F81	-lnL1 = 3050.1226
$2(\ln L_1 - \ln L_0) = 215.8008$	df = 3
P-value = <0.000001	

## Ti=Tv

Null model = F81	-lnL0 = 3050.1226
Alternative model = HKY	-lnL1 = 2919.6775
$2(\ln L_1 - \ln L_0) = 260.8901$	df = 1
P-value = <0.000001	

## Equal Ti rates

Null model = HKY	-lnL0 = 2919.6775
Alternative model = TrN	-lnL1 = 2919.5640
$2(\ln L_1 - \ln L_0) = 0.2271$	df = 1
P-value = 0.633719	

## Equal Tv rates

Null model = HKY	-lnL0 = 2919.6775
Alternative model = K81uf	-lnL1 = 2919.6624
$2(\ln L_1 - \ln L_0) = 0.0303$	df = 1
P-value = 0.861871	

## Equal rates among sites

Null model = HKY	-lnL0 = 2919.6775
Alternative model = HKY+G	-lnL1 = 2888.2053
$2(\ln L_1 - \ln L_0) = 62.9443$	df = 1
Using mixed chi-square distribution	
P-value = <0.000001	

## No Invariable sites

Null model = HKY+G	-lnL0 = 2888.2053
Alternative model = HKY+I+G	-lnL1 = 2887.8667
$2(\ln L_1 - \ln L_0) = 0.6772$	df = 1
Using mixed chi-square distribution	
P-value = 0.205268	

## Model selected: HKY+G

-lnL = 2888.2053

## Base frequencies:

freqA =	0.4009
freqC =	0.2060
freqG =	0.1059
freqT =	0.2872

## Substitution model:

Ti/tv ratio = 3.3424

## Among-site rate variation

Proportion of invariable sites = 0

Variable sites (G)

Gamma distribution shape parameter = 0.4337

\*\* Akaike Information Criterion (AIC) \*\*

Model selected: HKY+G

-lnL = 2888.2053

AIC = 5786.4106

Base frequencies:

freqA = 0.4009

freqC = 0.2060

freqG = 0.1059

freqT = 0.2872

Substitution model:

Ti/tv ratio = 3.3424

Among-site rate variation

Proportion of invariable sites = 0

Variable sites (G)

Gamma distribution shape parameter = 0.4337

--

PAUP\* Commands Block: If you want to implement the previous estimates as likelihood settings in PAUP\*, attach the next block of commands after the data in your PAUP file:

```
[!
Likelihood settings from best-fit model (HKY+G) selected by AIC in Modeltest
3.1
]
```

BEGIN PAUP;

Lset Base=(0.4009 0.2060 0.1059) Nst=2 TRatio=3.3424 Rates=gamma

Shape=0.4337 Pinvar=0;

END;

--

\* Model selection uncertainty : Akaike Weights

Model	AIC	Delta	Weight	CumWeight
HKY+G	5786.4106	0.0000	0.2881	0.2881
HKY+I+G	5787.7334	1.3228	0.1487	0.4368
TrN+G	5788.0884	1.6777	0.1245	0.5613
K81uf+G	5788.2261	1.8154	0.1162	0.6775
TrN+I+G	5789.3901	2.9795	0.0649	0.7425
K81uf+I+G	5789.5410	3.1304	0.0602	0.8027
TIM+G	5789.8896	3.4790	0.0506	0.8533
TVM+G	5790.8833	4.4727	0.0308	0.8841
TIM+I+G	5791.1826	4.7720	0.0265	0.9106
TrN+I	5791.6865	5.2759	0.0206	0.9312
HKY+I	5791.8584	5.4478	0.0189	0.9501
TVM+I+G	5792.2144	5.8037	0.0158	0.9659
GTR+G	5792.5190	6.1084	0.0136	0.9795
K81uf+I	5793.7588	7.3481	0.0073	0.9868
GTR+I+G	5793.8276	7.4170	0.0071	0.9939
TIM+I	5795.5811	9.1704	0.0029	0.9968
TVM+I	5796.1030	9.6924	0.0023	0.9991
GTR+I	5797.9102	11.4995	0.0009	1.0000

