BIOINFORMATICS APPLICATIONS NOTE

Vol. 14 no. 9 1998
Pages 817–818

MODELTEST: testing the model of DNA substitution

David Posada and Keith A. Crandall

Department of Zoology, Brigham Young University, 574 WIDB, Provo, UT 84602-5255, USA

Received on May 20, 1998; revised on July 22, 1998; accepted on July 27, 1998

Abstract

Summary: *The program MODELTEST uses log likelihood scores to establish the model of DNA evolution that best fits the data.*

Availability: *The MODELTEST package, including the source code and some documentation is available at http://bioag.byu.edu/zoology/crandall_lab/modeltest.html.*

Contact: *dp47@email.byu.edu*

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 and 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, 1993). 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. 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

$$\Lambda = \frac{\max\ [L_0\ (Null\ Model \mid Data)]}{\max\ [L_1\ (Alternative\ Model \mid Data)]}$$

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). 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 (Kendall and Stuart, 1979). To preserve the nesting of the models, the likelihood scores are estimated using 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 the Monte Carlo simulation (parametric bootstrapping) (Goldman, 1993).

Another way of comparing different models without the nested requirement is the Akaike information criterion (minimum theoretical information criterion, AIC) (Akaike, 1974). The AIC is a useful measure that rewards models for good fit, but imposes a penalty for unnecessary parameters (e.g. Hasegawa, 1990). If L is the maximum value of the likelihood function for a specific model using n independently adjusted parameters within the model, then AIC = $-2\ln L + 2n$. Smaller values of AIC indicate better models.

MODELTEST is a simple program written in ANSI C and compiled for the Power Macintosh using Metrowerks CodeWarrior. 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 δ = Ð2 log Λ 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 hypotheses about the process of DNA substitution. It also calculates the AIC estimate associated with each likelihood score.

The user communicates with the program using a standard console interface, where the input and output files as well as some options and help can be specified. By default, the program will accept two classes of input files: a file containing ordered raw log likelihood scores corresponding to the tested models (see Figure 1) or a PAUP* (Swofford, 1998) file containing a matrix of the same log likelihood scores resulting from the execution of a block of PAUP* (Swofford, 1998) commands. This block of PAUP* commands is available in the documentation. When specified, the program can also read a file with likelihood scores for identifying the minimum AIC estimate. The output of MODELTEST consists of the P-values corresponding to the tests performed. In these tests the null hypotheses are equal base frequencies, transition rate equals transversion rate, equal transition rates and equal transversion rates, rates equal among sites and no invariable sites. Finally, the program interprets these P-values and chooses the model that fits the data best among those tested following the likelihood ratio test and/or AIC 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 an individual alpha value of 0.01), or another value specified by the user.

© Oxford University Press

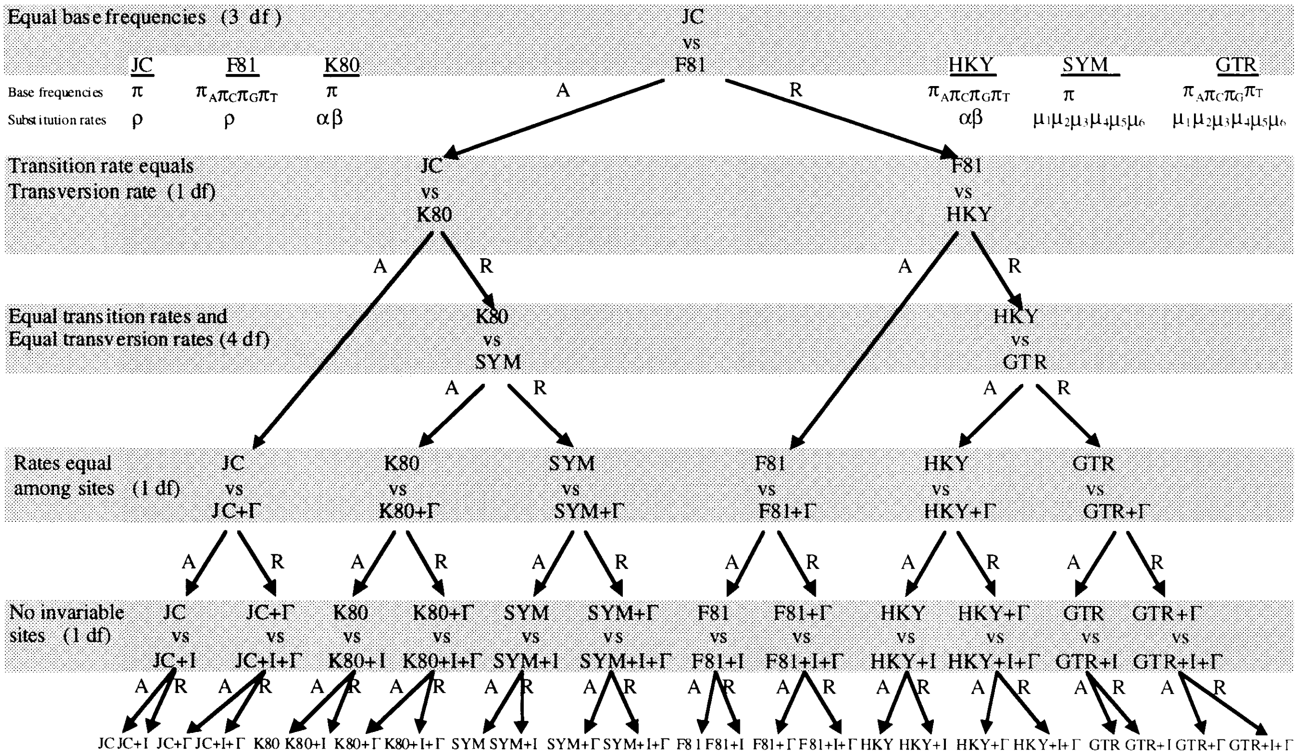

Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al*., 1985), and GTR (Rodríguez *et al*., 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. ^{1}: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G rate, μ_5: C⇒T rate, μ_6: G⇒T rate.

Acknowledgements

This project was supported by a fellowship from Caixagalicia Foundation (D.P.), the Alfred P. Sloan Foundation (K.A.C), and the National Institutes of Health (K.A.C.). We wish to thank the anonymous reviewers for their excellent suggestions.

References

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Contr.*, **19**, 716–723.

Felsenstein,J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.*, **22**, 521–565.

Goldman,N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.*, **36,** 182–198.

Hasegawa,M. (1990) Phylogeny and molecular evolution in primates. *Jpn J. Genet.*, **65,** 243–265.

Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **21,** 160–174.

Huelsenbeck,J.P. and Crandall,K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.*, **28,** 437–466.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro (ed.), *Mammalian Protein Metabolism.* Academic Press, New York, pp. 21–132.

Kendall,M. and Stuart,A. (1979) *The Advanced Theory of Statistics*, Vol. 2, 4th edn. Charles Griffin, London, pp. 240–252.

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.

Rodríguez,F.J., Oliver,J.L., Marín,A. and Medina,J.R. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, **142**, 485–501.

Swofford,D.L. (1998) PAUP*: phylogenetic analysis using parsimony (and other methods). Version 4.0 (prerelease test version). Sinauer, Sunderland, Massachusetts (in press).

Zharkikh,A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, **9**, 315–329.