

LETTERS

jModelTest: Phylogenetic Model Averaging

David Posada

Departamento de Genética, Bioquímica e Inmunología, Facultad de Biología, Universidad de Vigo, Vigo, Spain

jModelTest is a new program for the statistical selection of models of nucleotide substitution based on “Phyml” (Guindon and Gascuel 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.). It implements 5 different selection strategies, including “hierarchical and dynamical likelihood ratio tests,” the “Akaike information criterion,” the “Bayesian information criterion,” and a “decision-theoretic performance-based” approach. This program also calculates the relative importance and model-averaged estimates of substitution parameters, including a model-averaged estimate of the phylogeny. jModelTest is written in Java and runs under Mac OSX, Windows, and Unix systems with a Java Runtime Environment installed. The program, including documentation, can be freely downloaded from the software section at <http://darwin.uvigo.es>.

Introduction

Models of nucleotide substitution allow for the calculation of probabilities of change between nucleotides along the branches of a phylogenetic tree. The use of a particular substitution model may change the outcome of the phylogenetic analysis (e.g., Buckley 2002; Buckley and Cunningham 2002; Lemmon and Moriarty 2004), and statistical model selection has become an essential step for the estimation of phylogenies from DNA sequence alignments. In-depth reviews about model selection in phylogenetics are available elsewhere (Johnson and Omland 2003; Posada and Buckley 2004; Sullivan and Joyce 2005). Indeed, the performance of different model selection strategies has been the subject of active research (Posada 2001; Posada and Crandall 2001; Pol 2004; Abdo et al. 2005; Alfaro and Huelsenbeck 2006).

Several programs already exist for the statistical selection of models of nucleotide substitution (e.g., Nylander 2004; Keane et al. 2006). Among these, Modeltest (Posada and Crandall 1998) has been one of the most popular. This note describes a new program called jModelTest that supersedes Modeltest in several aspects. jModelTest allows for the definition of restricted sets of candidate models (table 1), implements customizable “hierarchical likelihood ratio tests” (hLRTs) (Fratini et al. 1997; Huelsenbeck and Crandall 1997; Sullivan et al. 1997) and “dynamic likelihood ratio tests” (dLRTs) (Posada and Crandall 2001), provides a rank of models according to the “Akaike Information Criterion” (AIC) (Akaike 1973), to the “Bayesian Information Criterion” (BIC) (Schwarz 1978) or to a “decision-theoretic performance-based” approach (DT) (Minin et al. 2003) (table 2), calculates the relative importance of every parameter, and computes model-averaged estimates of these, including a model-averaged estimate of the tree topology (Posada and Buckley 2004).

Model Selection with jModelTest

jModelTest is essentially a front-end to a computational pipeline that takes advantage of existing programs for

Key words: model selection, likelihood ratio tests, AIC, BIC, performance-based selection, statistical phylogenetics.

E-mail: dposada@uvigo.es.

Mol. Biol. Evol. 25(7):1253–1256. 2008

doi:10.1093/molbev/msn083

Advance Access publication April 8, 2008

running different tasks. Basically, this pipeline (fig. 1) includes:

- “ReadSeq” (Gilbert 2007): for conversion among different DNA sequence alignment formats.
- “Phyml” (Guindon and Gascuel 2003): for the likelihood calculations, including estimates of model parameters and trees.
- “Ted” (D. Posada): to compute Euclidean distances between trees for performance-based model selection.
- “Consense” (from the PHYLIP package) (Felsenstein 2005): to calculate weighted and strict consensus trees representing model-averaged phylogenies.

Likelihood Calculations

Likelihood calculations, including model parameters and tree estimates, are carried out with Phyml (Guindon and Gascuel 2003). The tree topology used in these calculations can be the same across models (fixed) or optimized for each one. Fixed tree topologies can be estimated with the BIONJ algorithm (Gascuel 1997) upon JC distances (Jukes and Cantor 1969) or user-defined. Alternatively, a BIONJ or an ML tree can be estimated under each model. In all cases, branch lengths are estimated and counted as parameters.

Custom Set of Models

Currently, there are 11 different nucleotide substitution schemes implemented in jModelTest, which combined with equal or unequal base frequencies (+F), a proportion of invariable sites (+I), and rate variation among sites (+G), result in 88 distinct models (table 1). The program offers the possibility of defining to a reasonable extent which models are included in the candidate set.

Sequential Likelihood Ratio Tests

A series of likelihood ratio tests (LRTs) can be implemented under a particular hierarchy (hLRTs), in which the user can specify their order, and whether parameters are added (forward selection) or removed (backward selection). Alternatively, the order of the LRTs can be set dynamically (dLRTs) (Posada and Crandall 2001), by comparing the current model with the one that is one hypothesis away and provides the largest increase (under forward selection) or smallest decrease (under backward

Table 1
Substitution Models Available in jModelTest

Model ^{a-c}	Free Parameters	Base Frequencies	Substitution Rates	Substitution Code
JC	k	Equal	AC = AG = AT = CG = CT = GT	000000
F81	$k + 3$	Unequal	AC = AG = AT = CG = CT = GT	000000
K80	$k + 1$	Equal	AC = AT = CG = GT, AG = CT	010010
HKY	$k + 4$	Unequal	AC = AT = CG = GT, AG = CT	010010
TrNe	$k + 2$	Equal	AC = AT = CG = GT, AG, CT	010020
TrN	$k + 5$	Unequal	AC = AT = CG = GT, AG, CT	010020
TPM1	$k + 2$	Equal	AC = GT, AT = CG, AG = CT	012210
TPM1u	$k + 5$	Unequal	AC = GT, AT = CG, AG = CT	012210
TPM2	$k + 2$	Equal	AC = AT, CG = GT, AG = CT	010212
TPM2u	$k + 5$	Unequal	AC = AT, CG = GT, AG = CT	010212
TPM3	$k + 2$	Equal	AC = CG, AT = GT, AG = CT	012012
TPM3u	$k + 5$	Unequal	AC = CG, AT = GT, AG = CT	012012
TIM1e	$k + 3$	Equal	AC = GT, AT = CG, AG, CT	012230
TIM1	$k + 6$	Unequal	AC = GT, AT = CG, AG, CT	012230
TIM2e	$k + 3$	Equal	AC = AT, CG = GT, AG, CT	010232
TIM2	$k + 6$	Unequal	AC = AT, CG = GT, AG, CT	010232
TIM3e	$k + 3$	Equal	AC = CG, AT = GT, AG, CT	012032
TIM3	$k + 6$	Unequal	AC = CG, AT = GT, AG, CT	012032
TVMe	$k + 4$	Equal	AC, AT, CG, GT, AG = CT	012314
TVM	$k + 7$	Unequal	AC, AT, CG, GT, AG = CT	012314
SYM	$k + 5$	Equal	AC, AG, AT, CG, CT, GT	012345
GTR	$k + 8$	Unequal	AC, AG, AT, CG, CT, GT	012345

NOTE.—The same number of branch lengths (k) needs to be estimated for every model.

^a JC (Jukes and Cantor 1969), F81 (Felsenstein 1981), K80 (Kimura 1980), HKY (Hasegawa et al. 1985), TrN (Tamura and Nei 1993), TPM (“3-parameter model,” = K81) (Kimura 1981), TIM (“transitional model”) (Posada 2003), TVM (“transversional model”) (Posada 2003), SYM (Zharkikh 1994), and GTR (Tavaré 1986).

^b Any of these can include invariable sites (+I), rate variation among sites (+G), or both (+I+G).

^c 5 equal frequencies; 5 unequal frequencies.

selection) in likelihood. The hLRTs and dLRTs will be available only if the likelihood scores were calculated upon a fixed topology, due to the nesting requirement of the χ^2 approximation.

Information Criteria

The program implements 3 different information criteria: the AIC (Akaike 1973), the BIC (Schwarz 1978), and a performance-based approach based on decision theory (DT) (Minin et al. 2003). Under the AIC framework, there is also the possibility of using a corrected version for small samples (AICc) (Sugiura 1978; Hurvich and Tsai 1989), instead of the standard AIC. In this case, sample size has to be specified, which by default is approximated as the number of sites in the alignment (note that the sample size of an alignment is presently an unknown quantity).

Model Selection Uncertainty

The AIC, BIC, and DT methods assign a score to each model in the candidate set, therefore providing an objective function to rank them. Using the differences in scores, the program can calculate a measure of model support called AIC or BIC weights (Burnham and Anderson 2003). For the DT scores, this calculation is not as straightforward, and right now a very gross approach is used instead, where the DT weights are the rescaled reciprocal DT scores. Confidence intervals (CIs) can be defined according to the cumulative weights, including a specified fraction of the models. When the CI includes only partially a given model, this model is included (yes/no) in the CI with a probability equal to the fraction included.

Parameter Importance and Model-Averaged Estimates

The program can also calculate the relative importance of every parameter of the substitution model and

Table 2
Model Selection Strategies Implemented in jModelTest

	Hierarchical Likelihood Ratio Tests	Dynamical Likelihood Ratio Tests	Akaike Information Criterion	Bayesian Information Criterion	Performance-Based Selection
Abbreviation	hLRTs	dLRTs	AIC	BIC	DT
Base tree	Fixed	Fixed	Fixed, optimized	Fixed, optimized	Fixed, optimized
Nesting requirement	Yes	Yes	No	No	No
Simultaneous comparison	No	No	Yes	Yes	Yes
Selection uncertainty	No	No	Yes	Yes	Yes ^a
Parameter importance	No	No	Yes	Yes	Yes ^a
Model averaging	No	No	Yes	Yes	Yes ^a

^a DT weights are simply the rescaled reciprocal DT scores. This is a gross implementation very likely to change.

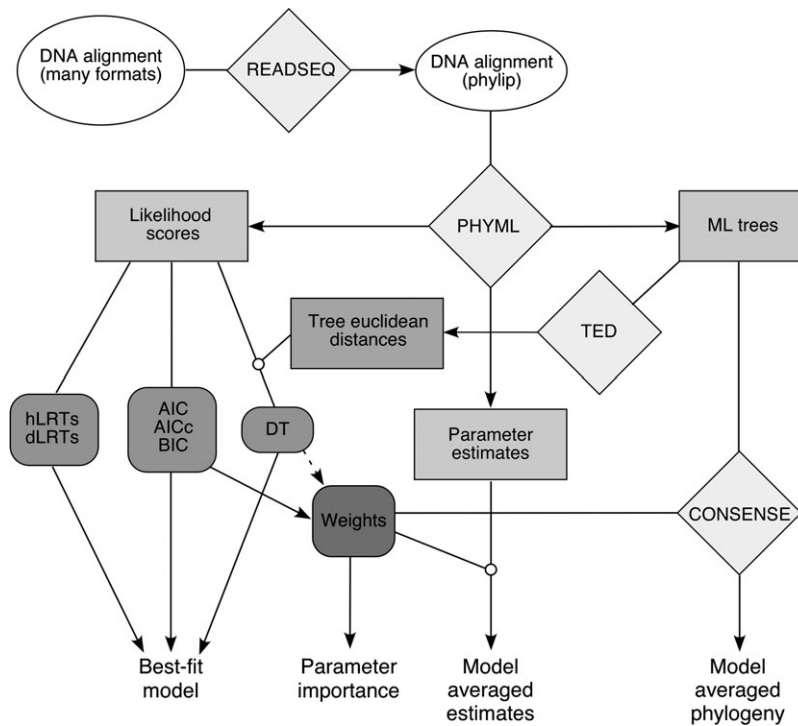


FIG. 1.—jModelTest pipeline. Alignments are loaded using the ReadSeq library (Gilbert 2007). Likelihood calculations, including estimates of model parameters and trees, are carried out with Phylml (Guindon and Gascuel 2003). A custom program called Ted (D. Posada) is used to compute Euclidean distances between trees for performance-based model selection (DT), whereas Consense (Felsenstein 2005) is used to calculate weighted and strict consensus trees representing model-averaged phylogenies.

model-averaged estimates of these, using all the models in the candidate set, or a fraction included in a particular CI (see Posada and Buckley 2004).

of these, providing an increased flexibility for the user to explore the data and the role of the substitution model on the estimation of phylogenetic trees.

Model-Averaged Phylogenies

jModelTest is able to compute an average estimate of the tree topology by building a consensus of the maximum likelihood (ML) trees for every model in the candidate set, weighting them with their model weights (AIC, BIC, or DT) (fig. 2). Indeed, this option is only available when the tree topology has been optimized for every model. The consensus tree is constructed using the Consense program from the PHYLIP package (Felsenstein 2005).

Software Platform and Availability

jModelTest is written in Java and can be started in any operating system with a Java Runtime Environment (see <http://www.java.com>). However, jModelTest uses other programs for different tasks, and these have been compiled for Mac OSX, Windows XP, and Linux. The package, including installation instructions, documentation, executables, and example data, is distributed free of charge for academic use from the software section at <http://darwin.uvigo.es>.

Conclusions

Model selection is an important issue in statistical phylogenetics, around which some questions still remain open (Kelchner and Thomas 2007). jModelTest addresses some

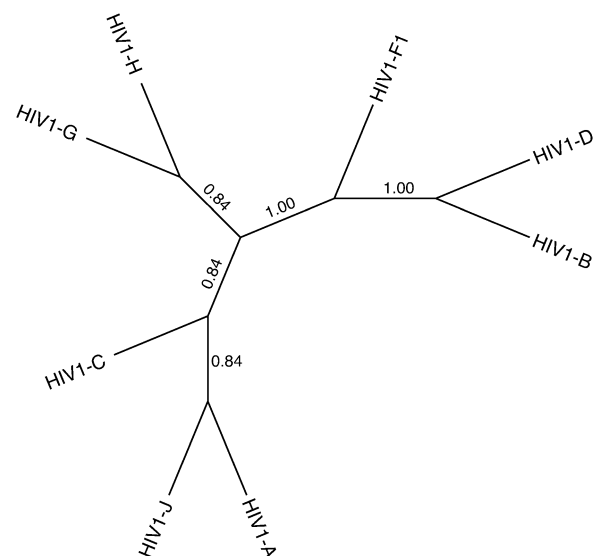


FIG. 2.—Model-averaged tree of HIV-1 *pol* sequences. The topology shown is the consensus of 88 ML tree topologies, one for every model, weighted according to the AIC weights. The numbers on the branches represent uncertainty due to model selection. In this case, clades (AJ), (AJC), and (HG) are supported by the best and fourth best AIC models (GTR + G, AIC weight = 0.83; TIM3 + G, AIC weight = 0.01; respectively) and others, but not by the second or third best AIC models (GTR + I + G, AIC weight = 0.15; GTR + I, AIC weight = 0.01; respectively).

Acknowledgments

I want to thank a number of users of Modeltest that had made numerous comments and suggestions through the years. Special thanks to Stephane Guindon for his generous help with Phyml and to John Huelsenbeck for suggesting the stochastic calculation of CIs. I want to acknowledge Sudhir Kumar for inviting me to present the latest advances in Modeltest at the 2006 SMCBE annual meeting, which finally prompted the completion of jModelTest.

Literature Cited

- Abdo Z, Minin VN, Joyce P, Sullivan J. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Mol Biol Evol.* 22:691–703.
- Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.
- Alfaro ME, Huelsenbeck JP. 2006. Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst Biol.* 55:89–96.
- Buckley TR. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst Biol.* 51:509–523.
- Buckley TR, Cunningham CW. 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Mol Biol Evol.* 19:394–405.
- Burnham KP, Anderson DR. 2003. Model selection and multi-model inference: a practical information-theoretic approach. New York: Springer.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.
- Felsenstein J. 2005. PHYLIP (phylogeny inference package). Seattle (WA): Department of Genome Sciences, University of Washington.
- Frati F, Simon C, Sullivan J, Swofford DL. 1997. Evolution of the mitochondrial cytochrome oxidase II gene in *Collembola*. *J Mol Evol.* 44:145–158.
- Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14:685–695.
- Gilbert D. 2007. ReadSeq. Bloomington (IN): Indiana University.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Hasegawa M, Kishino K, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.
- Huelsenbeck JP, Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Ecol Syst.* 28:437–466.
- Hurvich CM, Tsai C-L. 1989. Regression and time series model selection in small samples. *Biometrika.* 76:297–307.
- Johnson JB, Omland KS. 2003. Model selection in ecology and evolution. *Trends Ecol Evol.* 19:101–108.
- Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HM, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
- Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol.* 6:29.
- Kelchner SA, Thomas MA. 2007. Model use in phylogenetics: nine key questions. *Trends Ecol Evol.* 22:87–94.
- 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.
- Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst Biol.* 53:265–277.
- Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst Biol.* 52:674–683.
- Nylander JA. 2004. MrAIC [Internet]. [cited 2008 April 23]. Available from: <http://www.abc.se/~nylander/>. program distributed by the author
- Pol D. 2004. Empirical problems of the hierarchical likelihood ratio test for model selection. *Syst Biol.* 53:949–962.
- 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. 2003. Using Modeltest and PAUP* to select a model of nucleotide substitution. In: Baxevanis AD, Davison DB, Page RDM, Petsko GA, Stein LD, Stormo GD, editors. Current Protocols in Bioinformatics. New York: John Wiley & Sons. p. 6.5.1–6.5.14.
- Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol.* 53:793–808.
- Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics.* 14:817–818.
- Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Syst Biol.* 50:580–601.
- Schwarz G. 1978. Estimating the dimension of a model. *Ann Stat.* 6:461–464.
- Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun Stat Theory Methods.* A7:13–26.
- Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst.* 36:445–466.
- Sullivan J, Markert JA, Kilpatrick CW. 1997. Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. *Syst Biol.* 46:426–440.
- Tamura K, 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.
- Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM, editor. Some mathematical questions in biology—DNA sequence analysis. Providence (RI): American Mathematical Society. p. 57–86.
- Zharkikh A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol.* 39:315–329.

Sudhir Kumar, Associate Editor

Accepted April 2, 2008