BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 9 2005, pages 2104–2105
doi:10.1093/bioinformatics/bti263

*Phylogenetics*

# ProtTest: selection of best-fit models of protein evolution

Federico Abascal[1,2,*], Rafael Zardoya[2] and David Posada[1]

[1]Department of Biochemistry, Genetics and Immunology, Universidad de Vigo, 36310 Vigo, Spain and
[2]Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales, 28006 Madrid, Spain

Received on October 7, 2004; revised on December 15, 2004; accepted on January 5, 2005
Advance Access publication January 10, 2005

**ABSTRACT**

**Summary:** Using an appropriate model of amino acid replacement is very important for the study of protein evolution and phylogenetic inference. We have built a tool for the selection of the best-fit model of evolution, among a set of candidate models, for a given protein sequence alignment.

**Availability:** ProtTest is available under the GNU license from http://darwin.uvigo.es

**Contact:** fabascal@uvigo.es

## 1 INTRODUCTION

### 1.1 Models of protein evolution

Models of protein evolution, or amino acid replacement, describe the probabilities of change from one amino acid to another, and therefore become indispensable tools for the characterization of the process of protein evolution (Thorne, 2000; Thorne and Goldman, 2003). Indeed, these models provide the foundation for the reconstruction of protein phylogenies under distance, maximum likelihood and Bayesian methods. Dayhoff *et al*. (1978) introduced the most influential model of amino acid replacement, a 20-state time reversible homogeneous Markov model. Because the large number of parameters in a 20-state replacement matrix, estimates of these parameters are usually obtained from large datasets prior to the analysis of the dataset of interest. In this way, different empirical matrices with fixed relative rates of amino acid replacement have been already proposed, like the Dayhoff matrix (Dayhoff *et al*., 1978), the JTT matrix (Jones *et al*., 1992), the mtREV matrix (Adachi and Hasegawa, 1996) or the WAG matrix (Whelan and Goldman, 2001). While these models generally assume that the process of amino acid replacement is very similar across all positions, conservation of protein function and structure imposes constraints on which positions can change. This evolutionary information can be inferred by considering a fraction of amino acids to be invariable ('+I') (Reeves, 1992), or assigning each site a probability to belong to given rate categories ('+G') (Yang, 1993). Additionally, observed amino acid frequencies can also be considered ('+F') (Cao *et al*., 1994).

### 1.2 Model selection and inference

Model selection may be seen as a way of identifying the model that, among a set of candidates, is closest to reality. Looking for a balance between accuracy and simplicity, Akaike (1973) found a simple relationship between the likelihood ($L$) and the number of parameters ($K$):

$$\mathrm{AIC} = -2\ln L + 2K$$

to estimate the expected distance of a given model from truth. When the sample size ($n$) is small compared to the number of parameters (e.g. $n/K < 40$), the AIC might not be accurate and then the use of the corrected AIC (AICc) (Sugiura, 1978) is recommended as,

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2K(K+1)}{n-K-1}.$$

A different, but simple approach is the Bayesian Information Criterion (BIC) (Schwarz, 1978), formulated as

$$\mathrm{BIC} = -2\ln L + K\log n.$$

Scaled AIC (Akaike weights) and the BIC can be easily used to asses model selection uncertainty, for model-averaging and to estimate parameter importance in an evolutionary context (Burnham and Anderson, 2002; Posada and Buckley, 2004).

## 2 THE PROGRAM: PROTTEST

Although widely-used software exists for the selection of the best-fit nucleotide models (Posada and Crandall, 1998), no program has been developed until now for protein models. ProtTest is a java program to find the best model of amino acid replacement for a given protein alignment. It is based on the Phyml program (Guindon and Gascuel, 2003) for the ML optimizations, modified to support +F and four extra substitution matrices and uses the PAL library (Drummond and Strimmer, 2001) for handling protein alignments and trees. ProtTest is available for Mac OSX, Linux and Windows, and it can be run in three ways: using a GUI, at the command-line and through the web. Its basic workflow is summarized in Figure 1.

Given a protein alignment and a tree topology the program calculates the likelihood under each candidate model, and estimates model parameters. The current version 1.2 implements 64 empirical models: the eight matrices WAG, mtREV, Dayhoff, JTT, VT, Blosum62, CpREV and RtREV under +F, +G, +I and their combinations. Other models exist, particularly mechanical models, that are not implemented in ProtTest. For each model, the tree topology can be fixed [provided by the user or calculated by BIONJ (Gascuel, 1997)] or optimized under ML. After this, the user can choose a model selection strategy (AIC, AICc, BIC), and obtain a rank of model fits, model-averaged parameter estimates or measures of parameter importance. For the AICc and the BIC, sample size is set by

*To whom correspondence should be addressed.

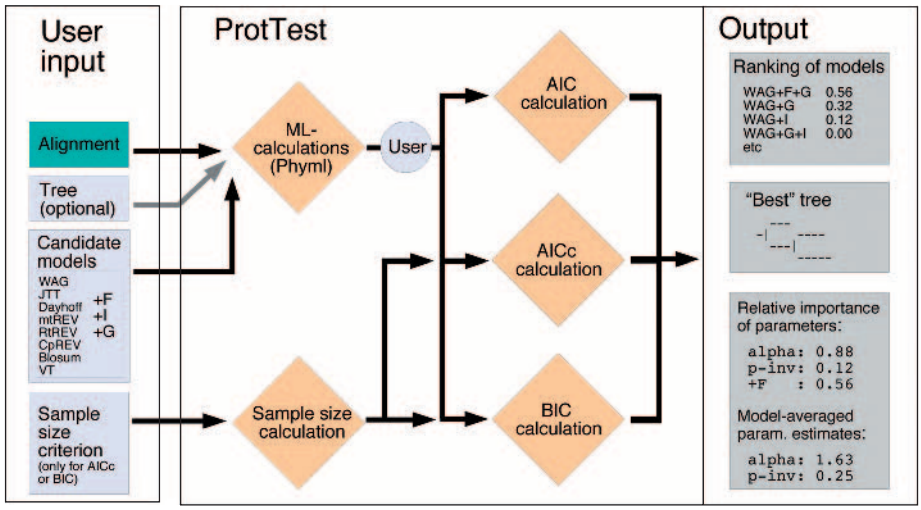


**Fig. 1.** The basic workflow of ProtTest Program. This figure can be viewed in colour on *Bioinformatics* online.

default to the number of positions in the alignment. Other options to define sample size attempt to take into account both the number of sequences and their redundancy. Other valuable features include the ability to restrict the set of candidate models (only in the GUI version) and the possibility to output the tree corresponding to the best model.

## ACKNOWLEDGEMENTS

Special thanks to Stephane Guindon (Phyml) and Matthew Goode (PAL) for their help. This work was financially supported from a research grant in bioinformatics from the Fundación BBVA (Spain).

## REFERENCES

Adachi,J. and Hasegawa,M. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, **42**, 459–468.

Akaike,H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proceedings of 2nd International Symposium on Information Theory*, Budapest, Hungary, pp. 267–281.

Burnham,K.P. and Anderson,D.R. (2002) *Model Selection and Multi-Model Inference. A Practical Information—Theoretic Approach*, 2nd edn. Springer-Verlag, New York.

Cao,Y., Adachi,J., Janke,A., Paabo,S. and Hasegawa,M. (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.*, **39**, 519–527.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In: Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure* National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Drummond,A. and Strimmer,K. (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, **17**, 662–663.

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.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.*, **8**, 275–282.

Posada,D. and Buckley,T.R. (2004) Model selection and model averaging in phylogenetics: advantages of AIC and Bayesian approaches over likelihood ratio tests. *Syst. Biol.*, **53**, 793–808.

Posada,D. and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Reeves,J.H. (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J. Mol. Evol.*, **35**, 17–31.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Sugiura,N. (1978) Further analysis of the data by Akaike's information criterion and the finite correction. *Comm. Statist. A-Theory. Meth.*, **7**, 13–26.

Thorne,J.L. (2000) Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.*, **10**, 602–605.

Thorne,J.L. and Goldman,N. (2003) Probabilistic models for the study of protein evolution. In: (ed.), *Handbook of Statistical Genetics* (Balding,D.J., Bishop,M. and Cannings,C.) John Wiley & Sons, Ltd, Chichester, England, pp. 209–226.

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Yang,Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10**, 1396–1401.