ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models

Diego Darriba*[,1,2], David. Posada[3,4,5], Alexey M. Kozlov[2], Alexandros Stamatakis[2,6], Benoit Morel[2], Tomas Flouri[7]

[1]Computer Architecture Group, CITIC, Universidade da A Coruña, Spain
[2]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies
[3]Department of Biochemistry, Genetics, and Immunology, University of Vigo, 36310 Vigo, Spain
[4]Biomedical Research Center (CINBIO), University of Vigo, 36310 Vigo, Spain
[5]Galicia Sur Health Research Institute, 36310 Vigo, Spain
[6]Institute of Theoretical Informatics, Karlsruhe Institute of Technology
[7]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom
***Corresponding author:** E-mail: diego.darriba@udc.es
Associate Editor:

Abstract

ModelTest-NG is a re-implementation from scratch of jModelTest and ProtTest, two popular tools for selecting the best-fit nucleotide and amino acid substitution models, respectively. ModelTest-NG is one to two orders of magnitude faster than jModelTest and ProtTest but equally accurate, and introduces several new features, such as ascertainment bias correction, mixture and FreeRate models, or the automatic processing of partitioned datasets. ModelTest-NG is available under a GNU GPL3 license at https://github.com/ddarriba/modeltest.

Key words: Phylogenetic model selection, high-performance computing, efficient algorithms, phylogenetic inference.

It is well known that the use of distinct probabilistic models of evolution can change the outcome of phylogenetic analyses (Buckley, 2002; Buckley and Cunningham, 2002; Lemmon and Moriarty, 2004). Not surprisingly, a number of bioinformatic tools have been developed in the last 20 years for selecting the best-fit model for the data at hand (Darriba *et al.*, 2011, 2012; Kalyaanamoorthy *et al.*, 2017). Recently, Abadi *et al.* (2019) concluded that using a parameter-rich model for DNA data leads to very similar inferences as the best-fit models. The authors present the results as an average over a number of MSAs. However, looking at individual MSA analyses we may observe substantial topological distances between inferences under the best-fit models and inferences under a parameter-rich GTR model (Arbiza *et al.*, 2011; Hoff *et al.*, 2016). However, continuous advances in sequencing technologies have made possible the assemblage of large multiple sequence alignments (MSA) that require faster and more scalable tools. In particular, our tools ProtTest (Darriba *et al.*, 2011) and jModelTest (Darriba *et al.*, 2012), which are among the most popular tools

for DNA and protein model selection, despite implementing High Performance Computing (HPC) algorithms for parallel execution with dynamic load balancing, still rely on PhyML (Guindon and Gascuel, 2003) for calculating the maximum likelihood (ML) scores of the competing models. This step constitutes by far the most compute-intensive part, >99% of overall execution time. The dependency on PhyML has several disadvantages: First, software maintenance heavily depends on PhyML developers. Thus, bug fixes and modifications affecting the interplay between Prot/jModelTest and PhyML must be continuously taken into account when incorporating PhyML updates. Second, PhyML and hence, ProtTest and jModelTest are relatively inefficient with respect to their likelihood calculations compared to more recent tools such as IQ-TREE (Nguyen *et al.*, 2015). The model selection feature of IQ-TREE, called ModelFinder (Kalyaanamoorthy *et al.*, 2017) is becoming increasingly popular due to its algorithmic and computational efficiency, the wide range of supported evolutionary models, and its user-friendliness.

With all this in mind, here we introduce ModelTest-NG, a new program that outperforms existing model selection tools in terms of the speed/accuracy trade-off. ModelTest-NG offers a completely redesigned graphical user interface (GUI). Recently, we also integrated ModelTest-NG into ParGenes (Morel *et al.*, 2018), a pipeline for massively parallel gene model selection and gene tree inference which is computationally efficient as well as easy-to-use. Moreover, we significantly improved maintainability, user support, and added several new capabilities. Its main features are:

- Data and models supported: ModelTest-NG supports both, nucleotide and amino acid models. It uses statistical criteria for selecting the best-fit substitution models such as AIC (Akaike, 1974), BIC (Schwarz, 1978), and DT (Minin *et al.*, 2003). For DNA, it can select among all 203 possible time-reversible substitution models. For amino acid data, ModelTest-NG compares 19 empirical replacement matrices as well as several recently introduced mixture models such as LG4M and LG4X (Le *et al.*, 2012). ModelTest-NG can also assess the fit of FreeRate models (Soubrier *et al.*, 2012).
- Partitioned data sets: ModelTest-NG can automatically perform model selection on individual non-overlapping partitions as specified by the user (e.g., on a per-gene basis, or by codon position).
- Phylogenetic templates: users can select so-called templates for the most popular phylogenetic inference tools: RAxML (Stamatakis, 2014), RAxML-NG (Kozlov *et al.*, 2018), IQ-TREE, PhyML, PAUP (Swofford, 2002), MrBayes (Ronquist

et al., 2012). When such a template is specified, ModelTest-NG will only evaluate models supported by the corresponding tool, and print out the respective tool-specific command line for phylogenetic reconstruction under the best-fit model.

- Native implementation: ModelTest-NG constitutes a full re-implementation of jModelTest and ProtTest in `C++` that relies on a novel and efficient low-level implementation of the Phylogenetic Likelihood Library (`https://github.com/xflouris/libpll-2`). This library encapsulates all compute- and memory-intensive phylogenetic likelihood computations and fully leverages the capabilities of modern x86 processors by using the AVX and AVX2 vector instruction sets. It also incorporates a recent algorithmic technique for accelerating likelihood calculations (Kobert *et al.*, 2017). All required numerical optimization routines are implemented in the pll-modules library (`https://github.com/ddarriba/pll-modules`).

We benchmarked ModelTest-NG against jModelTest, ProtTest, and ModelFinder (part of IQ-TREE version 1.6.1) using empirical as well as simulated data sets. We measured runtimes for all datasets, and estimated accuracy (recovering the generating model) using the simulated datasets. The data sets and the experimental setup are described in detail in the supplementary material where we also discuss the results more extensively. For the empirical DNA data sets, ModelTest-NG yielded average speedups of 281.54 over jModelTest and of 1.12 over ModelFinder (Figure S1). For the simulated DNA data, ModelTest-NG was 110.77 times faster than jModelTest but slower than ModelFinder (the latter was 1.59 times faster than ModelTest-NG). On the empirical protein data sets, ModelTest-NG yielded average speedups of 36.94 over ProtTest and similar runtimes as ModelFinder. On the simulated protein data, ModelTest-NG was 36.07 times faster than ProtTest and 1.03 times faster than ModelFinder. We observed that ModelTest-NG scales better than ModelFinder and jModelTest/ProtTest on large data sets. In general, the larger the data set is, in terms of number of taxa and number of sites, the better ModelTest-NG performs compared to the competing tools (see Figure 1). In terms of accuracy, ModelTest-NG found the true generating model for 81% of the simulated DNA data sets (jModelTest: 81%, ModelFinder: 70%) and for 85% of the simulated protein data sets (ProtTest: 85%, ModelFinder: 87%) (Figure 1).

These results were obtained under the default model selection parameter settings for both tools. In additional experiments we found that there is a pronounced trade-off between speed and accuracy. Thus, we can expect that the more thoroughly we optimize the likelihood score for a set of substitution model parameters, the more accurate

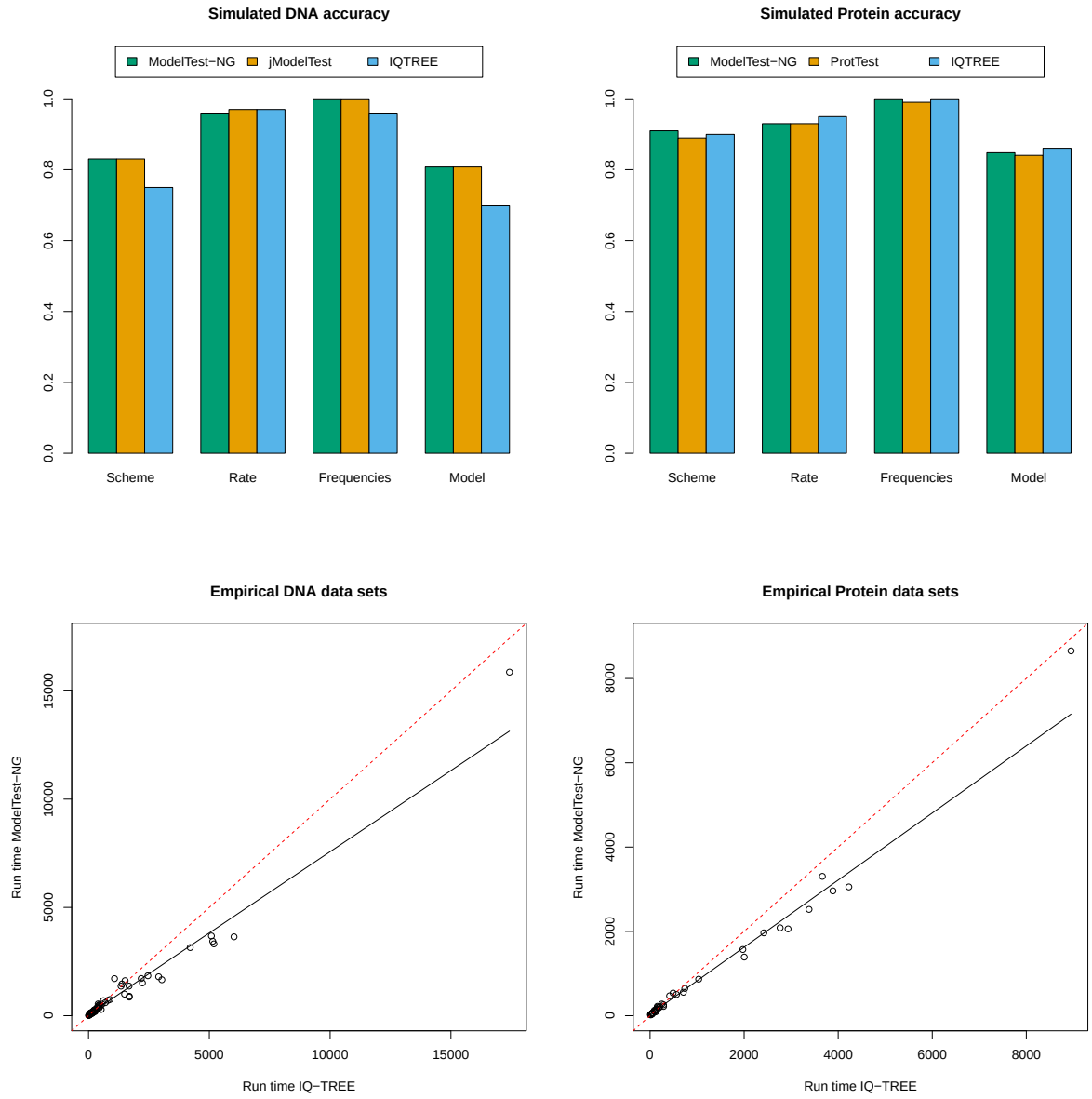

FIG. 1. Model selection accuracy comparison between ModelTest-NG, jModelTest/ProtTest, and IQ-TREE for simulated data (top) and LOESS curved fitted to a scatter plot of ModelTest-NG run times versus IQ-TREE for empirical data (bottom), for DNA (left) and protein (right) MSAs. The dashed red line represents equal run times.

the results will become (see Supplementary Material online). The thoroughness of model parameter optimization routines can be specified in ModelTest-NG.

ModelTest-NG thus represents a substantial improvement over our previous highly-cited tools, jModelTest and ProtTest. It preserves the accuracy of its predecessors, as evaluated against the ground truth on simulated datasets, while the runtime is improved by two orders of magnitude on empirical data. Compared to IQ-TREE we observed similar run times for empirical data sets, but IQ-TREE was faster on synthetic data and particularly so on DNA data. However, the accuracy of IQ-TREE on DNA data was substantially lower than for ModelTest-NG (70% versus 81%, respectively). In future versions of ModelTest-NG, we intend to introduce

methods to dynamically determine the optimal speed/accuracy trade-off for the dataset at hand.

Supplementary Material

Supplementary tables S1–S5 and figures S1–S4 are available at BIORXIV.

Acknowledgments

This work was supported by the Ministry of Economy and Competitiveness of Spain and FEDER funds of the EU (Project TIN2016-75845-P) and by the Galician Government (Xunta de Galicia) under the Consolidation Program of Competitive Research (ref. ED431C 2017/04). Part of this work was funded by the Klaus Tschira foundation and DFG grant STA-860/6.

References

Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature communications*, 10(1): 934.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6): 716–723.

Arbiza, L., Patricio, M., Dopazo, H., and Posada, D. 2011. Genome-wide heterogeneity of nucleotide substitution model fit. *Genome biology and evolution*, 3: 896–908.

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.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. 2011. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27: 1164–1165.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. 2012. jmodeltest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8): 772–772.

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

Hoff, M., Orf, S., Riehm, B., Darriba, D., and Stamatakis, A. 2016. Does the choice of nucleotide substitution models matter topologically? *BMC bioinformatics*, 17(1): 143.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. 2017. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6): 587.

Kobert, K., Stamatakis, A., and Flouri, T. 2017. Efficient detection of repeating sites to accelerate phylogenetic likelihood calculations. *Systematic biology*, 66(2): 205–217.

Kozlov, A., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. 2018. Raxml-ng: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *bioRxiv*, page 447110.

Le, S. Q., Dang, C. C., and Gascuel, O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular biology and evolution*, page mss112.

Lemmon, A. R. and Moriarty, E. C. 2004. The importance of proper model assumption in Bayesian phylogenetic. *Syst Biol*, 53: 265–277.

Minin, V., Abdo, Z., Joyce, P., and Sullivan, J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst Biol*, 52(5): 674–683.

Morel, B., Kozlov, A. M., and Stamatakis, A. 2018. Pargenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *bioRxiv*, page 373449.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2015. Iq-tree: a fast and effective

stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1): 268–274.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3): 539–542.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann Stat*, 6: 461–464.

Soubrier, J., Steel, M., Lee, M. S., Der Sarkissian, C., Guindon, S., Ho, S. Y., and Cooper, A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular Biology and Evolution*, 29(11): 3345–3358.

Stamatakis, A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Swofford, D. L. 2002. *Paup*: Phylogenetic analysis using parsimony (and other methods) Version 4*. Sinauer Associates, Sunderland, Massachusetts.