

Phylogenetics

PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating

Nicolas Lartillot^{1,*}, Thomas Lepage¹ and Samuel Blanquart²¹Département de Biochimie, Université de Montréal, Montréal, Québec, Canada and²Département d'Informatique, LIRMM, Montpellier, France

Received on January 26, 2009; revised on June 9, 2009; accepted on June 10, 2009

Advance Access publication June 17, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: A variety of probabilistic models describing the evolution of DNA or protein sequences have been proposed for phylogenetic reconstruction or for molecular dating. However, there still lacks a common implementation allowing one to freely combine these independent features, so as to test their ability to jointly improve phylogenetic and dating accuracy.

Results: We propose a software package, PhyloBayes 3, which can be used for conducting Bayesian phylogenetic reconstruction and molecular dating analyses, using a large variety of amino acid replacement and nucleotide substitution models, including empirical mixtures or non-parametric models, as well as alternative clock relaxation processes.

Availability: PhyloBayes is freely available from our web site <http://www.phylobayes.org>. It works under Linux, Mac OS X and Windows operating systems.

Contact: nicolas.lartillot@umontreal.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The field of phylogenetics has been particularly prolific over the recent years. Many new models have been proposed, such as mixture models, accounting for site-specific effects (Huelsenbeck and Suchard, 2007; Lartillot and Philippe, 2004; Le *et al.*, 2008a, b; Pagel and Meade, 2004; Wang *et al.*, 2008), or flexible molecular clocks (Drummond *et al.*, 2006; Lepage *et al.*, 2007; Thorne *et al.*, 1998). However, many of these developments are often available through distinct implementations, sometimes under different statistical paradigms (maximum likelihood versus Bayesian), making comparative evaluations more difficult, and preventing potentially powerful model combinations to be applied to empirical data. In particular, mixture models of amino acid replacement have resulted in important advances in model fit and phylogenetic reconstruction (Lartillot *et al.*, 2007; Le *et al.*, 2008b; Wang *et al.*, 2008), suggesting that their use in molecular dating analyses may also result in fundamental improvements. Yet, current molecular dating software packages do not implement such sophisticated substitution models.

In this direction, we propose a Bayesian phylogenetic reconstruction program, PhyloBayes 3. As its two main distinguishing features, this program gathers a large class of recently published models accounting for variations of the substitution patterns along the sequences, and rate variations along lineages (relaxed clocks). Overall, PhyloBayes 3 makes it possible to use a large spectrum of substitution models for both phylogenetic reconstruction and molecular dating analyses.

2 METHODS

2.1 Substitution models

Gathering several recent developments about the use of mixtures in statistical phylogenetics, PhyloBayes 3 proposes a wide range of models accounting for site-specific variations of several features such as:

- the rate of substitution, using either a discretized (Yang, 1994) or continuous (Mateiu and Rannala, 2006) gamma distribution, or a Dirichlet process infinite mixture (Huelsenbeck and Suchard, 2007);
- the equilibrium frequencies of the substitution process, using either a Dirichlet process (Lartillot and Philippe, 2004) or pre-specified empirical mixture of profiles (Le *et al.*, 2008a; Wang *et al.*, 2008);
- the entire substitution matrix again using a Dirichlet process, or pre-specified empirical mixtures of matrices (Le *et al.*, 2008b).

More classical site-homogeneous models are also implemented, such as JTT (Jones *et al.*, 1992), WAG (Whelan and Goldman, 2001) or LG matrices (Le and Gascuel, 2008), for proteins, and the general time-reversible (GTR) model for protein and nucleic acid data. In addition to these prespecified settings, PhyloBayes 3 allows users to enter their own matrix or mixture of matrices or profiles.

2.2 Molecular dating

Currently available molecular dating software programs propose either branch i.i.d. models of rate variations (Akerborg *et al.*, 2008; Drummond *et al.*, 2006), or autocorrelated model of rate variations (Kishino *et al.*, 2001; Lepage *et al.*, 2007; Rannala and Yang, 2007). In the case of branch i.i.d. models, the rate of evolution on each branch is independent from that of neighboring branches. In autocorrelated models, on the other hand, the rate follows a diffusion process along the lineages, so that trends in the overall substitution rate may last over several successive branches of the tree.

Differences also exist in the way the likelihood is computed. A first approach consists in using a normal approximation around the maximum

*To whom correspondence should be addressed.

likelihood estimate (Thorne *et al.*, 1998). A computationally more intensive but more exact approach requires to combine relaxed clock models with the classical pruning algorithm for computing the likelihood. Dynamic programming algorithms have been proposed to reduce the computational burden entailed by such exact computations (Akerborg *et al.*, 2008). In our case, we use data augmentation methods (see below), which also result in much more tractable computations in practice.

Finally, there are several ways in which fossil calibrations can be enforced. In the hard constraint approach, calibrations are considered as totally certain, so that calibrated nodes are never allowed to fall outside the specified intervals provided by fossil evidence (Kishino *et al.*, 2001). Alternatively, the soft bound approach allows for smoothly decreasing probability outside the intervals provided by the fossil calibrations (Inoue *et al.*, 2009; Rannala and Yang, 2007; Yang and Rannala, 2006).

Subsuming all these developments, PhyloBayes 3 implements both autocorrelated and non-autocorrelated models, and allows for molecular dating analyses both with or without the normal approximation, thereby making all substitution models currently implemented in PhyloBayes accessible to molecular dating analyses. It also accepts either hard or soft fossil constraints, thus allowing extensive comparisons of alternative approaches for estimating divergence dates.

2.3 Monte Carlo methods

On the algorithmic side, PhyloBayes relies on classical Metropolis–Hastings Monte Carlo methods, combined with more sophisticated sampling algorithms based on data augmentation and conjugate Gibbs sampling (Lartillot, 2006; Mateiu and Rannala, 2006). These latter algorithmic developments proved essential for proper mixing in the high-dimensional spaces entailed by the more complex models proposed by the program, in particular the infinite mixtures. They are also a key ingredient of the molecular dating analyses without the normal approximation.

Correctly assessing convergence is a particularly important aspect of Markov chain Monte Carlo, in particular under complex non-parametric models. In this respect, we propose several convergence diagnostics, consisting in measuring the discrepancy between the posterior averages obtained from several independent runs, as well as estimating the decorrelation time and the effective size of several summary statistics. These diagnostics are also available as automated stopping rules, telling the program to stop once the diagnostics indicate a sufficiently good convergence.

The implementation was checked against several alternative software programs under equivalent models, as well as using self-consistency tests based on simulated data (see Supplementary Material).

2.4 Model comparison and assessment

PhyloBayes 3 offers several approaches for Bayesian model comparison and assessment, such as Bayes factor computation for comparing relaxed molecular clock models under the normal approximation, cross-validation and posterior predictive testing. Based on our experience thus far, tentative guidelines for model choice can be provided: the CAT-GTR model, which is a Dirichlet process mixture of profiles of equilibrium frequencies combined with general exchange rates (i.e. an infinite mixture of matrices sharing the same set of exchange rates, and differing only by their equilibrium frequencies) is the best overall model, although its fit breaks down for smaller datasets (<1000 aligned positions), for which empirical mixtures then provide good alternatives. On the other hand, for very large datasets, the computational cost of CAT-GTR may be prohibitive, in which case combining an infinite mixture of profiles with flat exchange rates (the CAT model) offers a good compromise between computational speed and model fit.

Table 1. Cross-validation scores (averaged over 10 replicates) and posterior predictive tests

Model	Cross-validation		Saturation	Diversity	
	Score	SD	<i>P</i> -value	<i>P</i> -value	<i>z</i> -score
WAG ^a	0	0	<0.001	<0.001	79.8
LG ^b	398	21	<0.001	<0.001	63.7
WLSR5 ^c	496	76	0.281	<0.001	47.0
GTR	689	52	<0.001	<0.001	49.5
UL3 ^d	1413	53	<0.001	<0.001	49.0
C60 ^e	2003	74	<0.001	<0.001	35.0
CAT	2475	101	0.847	0.005	2.4
CAT-GTR	2863	100	0.670	0.046	1.6

^aWhelan and Goldman (2001); ^bLe and Gascuel (2008); ^cWang *et al.* (2008); ^dLe *et al.* (2008b); ^eLe *et al.* (2008a).

The cross-validation scores are relative to the WAG model (a higher score meaning a higher fitness). All *P*-values >0.05 are in bold-face.

3 ILLUSTRATION

We conducted an analysis on a previously published phylogenomic dataset encompassing 15 553 aligned positions for 52 eukaryotic taxa (Philippe *et al.*, 2007). We first compared alternative models by cross-validation (the procedure to follow is described in the manual, see Supplementary Material). The CAT-GTR model appears to be the best model, followed by the CAT model.

Using posterior predictive tests, the models were assessed for their ability to account for saturation and site-specific amino acid propensities (diversity). The CAT-GTR model is the only model not rejected by both tests, while CAT is only marginally rejected for the diversity test, and passes the saturation test (Table 1). Interestingly, the WLSR model (Wang *et al.*, 2008) also passes the saturation (but not the diversity) test, and this, in spite of its low cross-validation fit. All other models are rejected by the two tests.

As both cross-validation and posterior predictive assessments strongly suggest the use of CAT-GTR, this model was used to infer the tree and perform a molecular dating analysis. The tree was identical to that obtained under the CAT model, although with globally higher posterior probability support values (Fig. S1 in the Supplementary Material; see also Philippe *et al.*, 2007). For the molecular dating analysis, we used a log-normal autocorrelated clock relaxation model, a birth–death prior on divergence times, combined with soft calibrations (Rannala and Yang, 2007; Yang and Rannala, 2006). The resulting chronogram, automatically provided as a postscript file by the program, is displayed in Figure 1. Imposing hard bounds for the calibrations, using a uniform prior on divergence times, or alternative clock relaxation models, did not result in significant changes in the divergence date estimates (Fig. 2 in the Supplementary Material).

The overall analysis took approximately 1 month, occupying 32 nodes on a distributed memory cluster (Intel Q9550 bi-processors), >70% of the CPU being used by the cross-validation analysis. The inference of the tree topology took 2 weeks on two independent processors, and the molecular dating analysis required ~1 week on eight nodes.

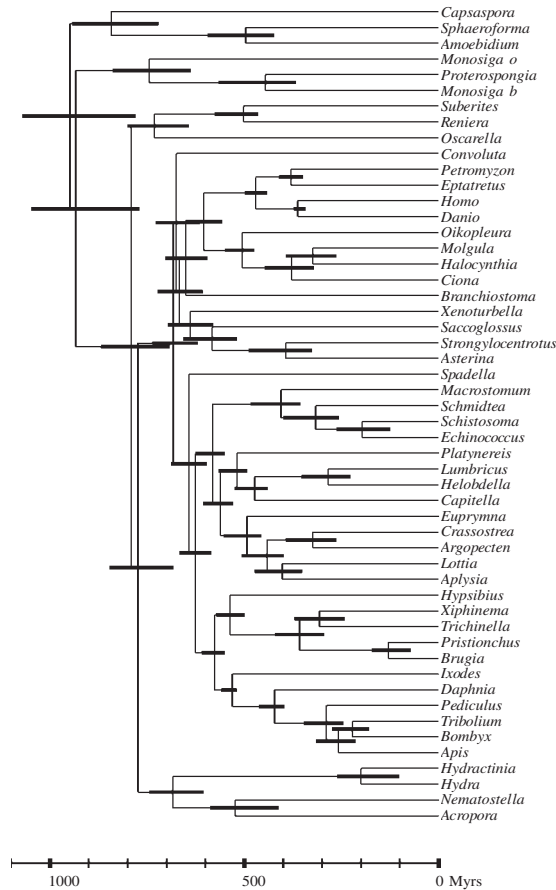


Fig. 1. Chronogram obtained for the dataset of Philippe *et al.* (2007). Two fossil calibration were used, taken from Douzery *et al.* (2004): the ancestor of vertebrates (354–417 Mya), and of arthropods (490–543 Mya). These calibrations were proposed as soft bounds under a birth–death prior (Inoue *et al.*, 2009).

ACKNOWLEDGEMENTS

We wish to thank Le Si Quang, Olivier Gascuel, Huai-Chun Wang and Andrew Roger for providing the models; Sebastien Bigaret for his technical help in compiling under windows; Hervé Philippe, Frédéric Delsuc and Emmanuel Douzery for their extensive beta testing of the program; Nicolas Rodrigue and three anonymous referees for their comments on the manuscript.

Funding: French ANR Mitosys program; Université de Montréal; Canada Foundation for Innovation.

Conflict of interest: none declared.

REFERENCES

Akerborg, O. *et al.* (2008) Birth-death prior on phylogeny and speed dating. *BMC Evol. Biol.*, **8**, 77.

Douzery, E.J.P. *et al.* (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc. Natl Acad. Sci. USA*, **101**, 15386–15391.

Drummond, A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS*, **4**, 699–710.

Huelsenbeck, J.P. and Suchard, M.A. (2007) A nonparametric method for accommodating and testing across-site rate variation. *Syst. Biol.*, **56**, 975–987.

Inoue, J. *et al.* (2009) The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.*, in press.

Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, **8**, 275–282.

Kishino, H. *et al.* (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.*, **18**, 352–361.

Lartillot, N. (2006) Conjugate gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.*, **13**, 1701–1722.

Lartillot, N. and Philippe, H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.

Lartillot, N. *et al.* (2007) Suppressing long branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.*, **7**, S4.

Le, S.Q. and Gascuel, O. (2008) An improved general amino-acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.

Le, S.Q. *et al.* (2008a) Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, **24**, 2317–2323.

Le, S.Q. *et al.* (2008b) Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **363**, 3965–3976.

Lepage, T. *et al.* (2007) A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.*, **24**, 2669–2680.

Mateiu, L. and Rannala, B. (2006) Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Syst. Biol.*, **55**, 259–269.

Pagel, M. and Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, **53**, 561–581.

Philippe, H. *et al.* (2007) Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *Mol. Biol. Evol.*, **2**, e717.

Rannala, B. and Yang, Z. (2007) Inferring speciation times under an episodic molecular clock. *Syst. Biol.*, **56**, 453–466.

Thorne, J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.

Wang, H. *et al.* (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.*, **8**, 331.

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. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Yang, Z. and Rannala, B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.*, **23**, 212–226.