2-1-2009

PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy

Jeremy M. Brown
The University of Texas at Austin

Robert Eldabaje
The University of Texas at Austin

Recommended Citation

Brown, J., & Eldabaje, R. (2009). PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. *Bioinformatics, 25* (4), 537-538. https://doi.org/10.1093/bioinformatics/btn651

BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 4 2009, pages 537–538
doi:10.1093/bioinformatics/btn651

Phylogenetics

PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy

Jeremy M. Brown[1,2,*] and Robert ElDabaje[1]

[1]Section of Integrative Biology and [2]Center for Computational Biology and Bioinformatics, University of Texas – Austin, Austin, TX 78712, USA

Received on August 14, 2008; revised and accepted on December 17, 2008

Advance Access publication December 19, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: The accuracy of Bayesian phylogenetic inference using molecular data depends on the use of proper models of sequence evolution. Although choosing the best model available from a pool of alternatives has become standard practice in statistical phylogenetics, assessment of the chosen model's adequacy is rare. Programs for Bayesian phylogenetic inference have recently begun to implement models of sequence evolution that account for heterogeneity across sites beyond variation in rates of evolution, yet no program exists to assess the adequacy of these models. PuMA implements a posterior predictive simulation approach to assessing the adequacy of partitioned, unpartitioned and mixture models of DNA sequence evolution in a Bayesian context. Assessment of model adequacy allows empirical phylogeneticists to have appropriate confidence in their results and guides efforts to improve models of sequence evolution.

Availability: This program is available as source code, a Java .jar application, and a native Mac OS X application. It is distributed under the terms of the GNU General Public License at http://code.google.com/p/phylo-puma.

Contact: jembrown@mail.utexas.edu

1 INTRODUCTION

Probabilistic approaches to phylogenetic inference require the specification of explicit models of sequence evolution. The dependence of resulting phylogenetic estimates on the underlying model of sequence evolution is well established (Lemmon and Moriarty, 2004; Swofford *et al*., 2001; Yang *et al.*, 1994). Much work has been done to develop models of sequence evolution that incorporate the complexities of the evolutionary process important in empirical datasets [see Swofford *et al.* (1996) and references therein]. In particular, approaches that incorporate heterogeneity in the evolutionary process across sites have recently received much attention (Nylander *et al*., 2004; Pagel and Meade, 2004).

As empiricists have faced a rapidly increasing pool of models from which to choose, many studies have explored objective methods for model choice (Minin *et al.*, 2003; Posada and Buckley, 2004; Sullivan and Joyce, 2005). However, far less attention has been paid to whether the best model adequately accounts for the processes important in the generation of a given dataset. This paucity of interest has occurred despite the development of such approaches over 15 years ago (Goldman, 1993). One hindrance to the widespread use of model adequacy tests is a lack of software able to perform such tests for recently developed models that incorporate heterogeneity in process across sites, although model adequacy tests that include heterogeneity in rates can be performed in MAPPS (Bollback, 2002). Here, we describe PuMA, software that implements tests of model adequacy in a Bayesian framework using posterior predictive simulation (Bollback, 2002). PuMA allows model adequacy tests to be performed for partitioned and mixture models of DNA sequence evolution. PuMA will facilitate much broader application of posterior predictive simulation tests of model adequacy, including much-needed benchmarking.

2 PuMA

2.1 Posterior predictive simulation

PuMA implements a posterior predictive simulation approach to the testing of model adequacy (Gamerman, 1997; Gelman *et al.*, 1995; Rubin, 1984), first introduced to phylogenetics by Bollback (2002). Posterior predictive simulation begins with a collection of parameter values and trees resulting from Markov chain Monte Carlo (MCMC) sampling of the posterior distribution during Bayesian phylogenetic analysis (Fig. 1). PuMA currently accepts input from unpartitioned and a priori partitioned analyses performed in MrBayes (Huelsenbeck and Ronquist, 2001), as well as mixture model analyses from BayesPhylogenies (Pagel and Meade, 2004). Each set of sampled parameter values and tree topologies is used to simulate a predictive dataset of the same size as the original, employing the same model of sequence evolution assumed during analysis. If the model of sequence evolution adequately captures the salient features of the evolutionary process, the simulated datasets should 'look' very similar to the original dataset. The 'look' of a dataset is summarized by a test statistic [given by $T(\mathbf{X})$, with $\mathbf{X}$ denoting a given dataset]. Well-designed test statistics can probe the adequacy of different assumptions underlying the model. PuMA saves all simulated datasets, allowing users to apply test statistics of their own choosing. PuMA's current implementation uses the unconstrained likelihood as a test statistic, which aims to assess model adequacy very generally (Bollback, 2002; Goldman, 1993). The unconstrained model interprets the data as a series of site patterns, each sampled with some fixed probability. The maximum

*To whom correspondence should be addressed.

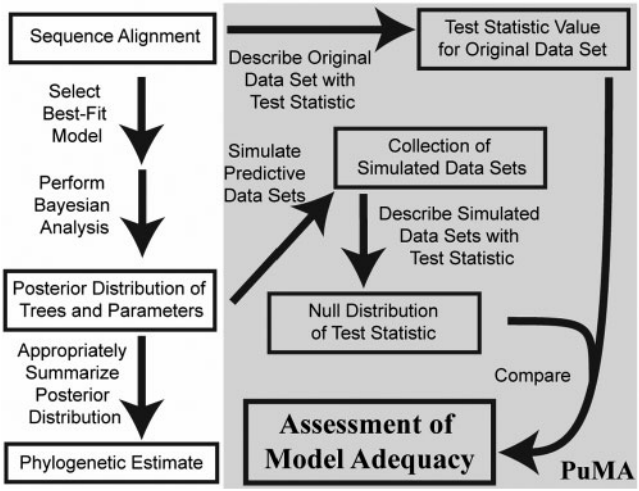

Fig. 1. Flowchart for Bayesian phylogenetic analysis, including posterior predictive simulation for the assessment of model adequacy. Shaded analyses are implemented in PuMA.

likelihood estimate of the sampling probability for any given site pattern is simply the frequency with which that pattern is observed in the data (Goldman, 1993). Therefore, the unconstrained likelihood of an entire dataset is calculated as

$$L(M|X) = \prod_{i=1}^{n} \left(\frac{N_{\Theta(i)}}{N} \right)^{N_{\Theta(i)}}$$

where M is the unconstrained model, **X** is the dataset, n is the number of unique site patterns, $\Theta(i)$ is the i-th unique site pattern, $N_{\Theta(i)}$ is the number of instances of $\Theta(i)$ in the dataset and N is the total number of sites. For convenience, the natural log of this likelihood is taken to be the test statistic. The posterior predictive distribution of T(**X**) consists of the set of T(**X**) values calculated from the datasets simulated using the posterior distribution of trees and parameter values. The posterior predictive P-value is the percentage of the posterior predictive distribution with T(**X**) values greater than or equal to the value of T(**X**) given by the original dataset. Example assessments of model adequacy for empirical data are given in Table 1. Note that model adequacy analyses may produce results that differ from standard model choice tests, due to effects of priors, the chosen test statistic, and the relative power of the tests.

2.2 Implementation details

PuMA is written in Java, extending the JPanel class, and uses Unix commands to manipulate output files. Therefore, it currently requires a Unix-based system (e.g. Mac OS X) that supports a GUI. PuMA calls Seq-Gen (Rambaut and Grassly, 1997) to simulate individual partitions and then combines all partitions into one dataset, if necessary. Analyses can be started using either the GUI interface or PuMA batch input files. PuMA is distributed both as a Java .jar application, as well as a native Mac OS X application. PuMA can also call MrConverge (by A. R. Lemmon; available from http://www.evotutor.org/MrConverge.html).

Table 1. Model adequacy tests using the multinomial likelihood and comparison to Bayes factors (BFs), for example datasets

Data	Taxa	Sites	Part. no.	P	2ln(BF)
Actinopterygii	42	3214	1	0.38	
Actinopterygii	42	3214	12	0.11	3759
Tetrapods	88	1728	1	0.04	
Tetrapods	88	1728	3	0.11	2238
Ants	163	3809	1	<0.01	
Ants	163	3809	4	<0.01	1483

BFs give the support in favor of the partitioned analysis for each dataset. Data are from Li *et al*. (2008) (actinopterygii), Hugall *et al*. (2007) (tetrapods) and Rabeling *et al*. (2008) (ants).

ACKNOWLEDGEMENTS

The comments of A.R. Lemmon, T.A. Heath and D.M. Hillis greatly improved this article.

Funding: National Science Foundation graduate research fellowship (to J.M.B.); Donald D. Harrington fellowship from the University of Texas – Austin (to J.M.B.).

Conflict of Interest: none declared.

REFERENCES

Bollback,J.P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, **19**, 1171–1180.

Gamerman,D. (1997) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, New York.

Gelman,A. *et al*. (1995) *Bayesian Data Analysis*. Chapman and Hall, New York.

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

Huelsenbeck,J.P. and Ronquist,F. (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Hugall,A.F. *et al*. (2007) Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Syst. Biol.*, **56**, 543–563.

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

Li,C. *et al*. (2008) Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. *Syst. Biol.*, **57**, 519–539.

Minin,V. *et al.* (2003) Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.*, **52**, 674–683.

Nylander,J.A.A. *et al*. (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, **53**, 47–67.

Pagel,M. and Meade,A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, **53**, 571–581.

Posada,D. and Buckley,T.R. (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.

Rabeling,C. *et al*. (2008) Newly discovered sister lineage sheds light on early ant evolution. *Proc. Natl Acad. Sci. USA*, **105**, 14913–14917.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Rubin,D.B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.*, **12**, 1151–1172.

Sullivan,J. and Joyce,P. (2005) Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.,* **36**, 445–466.

Swofford,D.L. *et al*. (1996) Phylogenetic inference. In Hillis,D.M. *et al*. (eds) *Molecular Systematics.* 2nd edn. Sinauer Associates, Sunderland, MA, USA, pp. 407–514.

Swofford,D.L. *et al*. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, **50**, 525–539.

Yang,Z. *et al*. (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.*, **11**, 316–324.