Bioinformatics Advance Access published July 7, 2011

DivBayes and SubT: exploring species diversification using Bayesian statistics

Ryberg, M[1,*], Nilsson, R.H.[2,3] and Matheny, P.B[1]

[1]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610, USA

[2]Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden.

[3]Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia

Associate Editor: Dr. David Posada

ABSTRACT

Summary: DivBayes is a program to estimate diversification rates from species richness and ages of a set of clades. SubT estimates diversification rates from node heights within a clade. Both programs implement Bayesian statistics and provide the ability to account for uncertainty in the ages of taxa in the underlying data, an improvement over more commonly used maximum likelihood methods.

Availability: DivBayes and SubT are released as C++ source code under the GNU GPL v. 3 software license in Supplementary information 1 and 2, respectively, and at http://web.utk.edu/~kryberg/. They have been successfully compiled on various Linux, MacOS X, and Windows systems.

Contact: kryberg@utk.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Bayesian inference has become a major part of the phylogenetic analysis tool kit, including applications such as topological inference and ancestral state reconstruction (Ronquist and Huelsenbeck, 2003; Pagel et al., 2004; Lartillot et al., 2009; Larget et al., 2010). However, methods of Bayesian inference are still not widely applied to infer species diversification rates, a major field within macroevolution (Nee et al., 1994; Rabosky, 2006; Alfaro et al., 2009; FitzJohn et al., 2009). Here we present two new programs that use Bayesian statistics to estimate net diversification rates (speciation minus extinction rate) and relative extinction rate (extinction rate divided by speciation rate) under the Yule or birth-death model based on two different types of data sets. DivBayes uses the number of species and clade ages of a set of taxa (likelihood formula in Bailey, 1964), while SubT uses node height data from within a clade (likelihood formula in Nee et al., 1994). Here, clade age data refer to the stem age of a group and may be derived from fossils, dated phylogenies, or other sources. Node height data refer to the split between each lineage in a group and are most likely to derive from dated phylogenies.

*To whom correspondence should be addressed.

2 IMPLEMENTATION

Both DivBayes and SubT use a Metropolis-Hastings MCMC algorithm (e.g., Gilks et al., 1995) to estimate the posterior distribution of parameters. In DivBayes it is possible to assume equal or different diversification rates for all clades, or define which clades should have the same rates. The method implemented in SubT is based largely on Bokma (2008) and uses his algorithm to update parameters. Both programs are command line-based and will read commands and data from a text-formatted input file. The output can be summarized with a separate command. The MCMC is presented in a tab-delimited file that can be read by many other applications such as R (R Development Core Team, 2010) or spreadsheet programs.

In Bayesian analysis it is easy to incorporate uncertainty in the underlying data by treating such data as parameters to estimate (Gilks et al., 1995). In DivBayes it is possible to use normal distributions as priors for the clade ages instead of using absolute dates. Similarly, in SubT it is possible to include "substitute taxa" (Ryberg and Matheny, 2011) for taxa with unknown node heights, using a uniform distribution between zero and the oldest included node as a prior. SubT can also handle distributions of data by reading several sets of node heights and performing estimates on each set separately. The final estimate can then be averaged over the given distribution.

3 TEST DATASETS

Program performance was evaluated on data simulated in Geiger (Harmon et al., 2008) under Yule and birth-death models (Supplementary 3 and 4). The DivBayes estimates were compared to maximum likelihood (ML) estimates performed in R. Different standard deviations were applied to the age priors (Supplementary 4). Each analysis was done assuming equal diversification rates for all clades and using a run length of one million generations with the MCMC sampled every one thousand generations. The performance of SubT was compared to that of the ML-based R packages LASER (Rabosky, 2006) and APE (Paradis et al., 2004). SubT was tested on trees simulated to include different number of species as well as different numbers of randomly selected species, with known node heights - replaced by "substitute taxa" (Supplementary 4). Each dataset was analyzed using ten million genera-

tions sampling every ten thousandth. In both DivBayes and SubT analyses, the Yule model was used for data sets simulated under the Yule model while the birth-death model was used for the datasets featuring extinctions. Default values were used for priors and starting values.

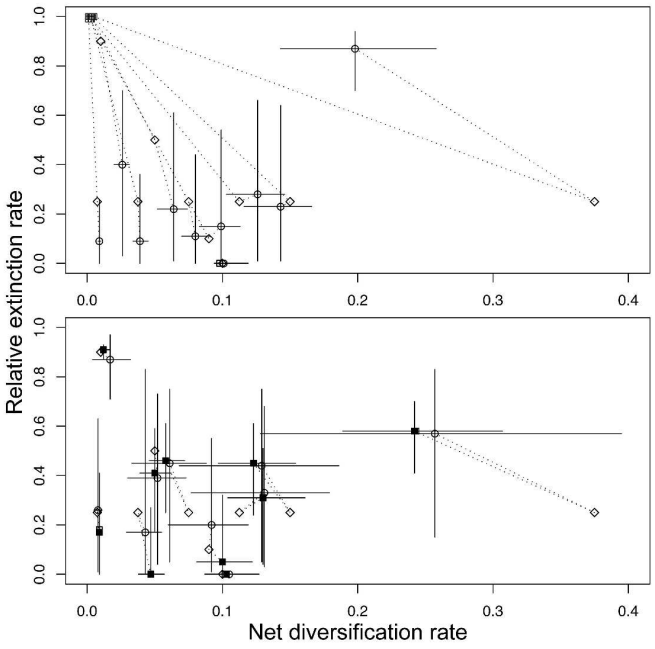

Fig. 1. Top graph shows DivBayes median estimates (circles) for 10 simulations with 20 data points (diamonds) and corresponding ML estimates (squares). Lower graph shows SubT median estimates (circles) for 10 simulations with 100 species (diamonds), and corresponding LASER (open squares) and APE (filled squares) estimates (LASER and APE estimates generally overlap). Solid lines show 95% credibility/confidence intervals. Dashed lines link each estimates with the value used to simulate the underlying data (i.e., the true value).

Both DivBayes and SubT were found to perform well under the various conditions examined. Both analyze reasonably large datasets within minutes, and the execution time scales linearly with number of data points (clades/nodes). In our test datasets, the addition of uncertainty in clade ages results in DivBayes execution times 1.5-2 times longer than runs with exact dates. The extra execution time implementing "substitute taxa" in SubT depends on the number and proportion of taxa. It scales reasonably linearly when <50% of the data are "substitute taxa", but run time increases considerably in our test datasets when "substitute taxa" represent 75% of the data (Supplementary 4).

Parameter values used in simulations of DivBayes fall within the estimated 95%-credibility interval both for the Yule datasets and datasets with relative extinction rate of 0.1. The net diversification rate was overestimated - and extinction rates underestimated - for the data simulated under the highest and lowest net diversification rates (0.375 and 0.0075) and the higher relative extinction rates (0.5 and 0.9; Fig. 1; Supplementary 4). These simulated values differ markedly from the default net diversification prior (mean 0.1). It is only for the highest relative extinction rate that the ML estimate is closer to the parameter values used for the simulation than the median of the posterior distribution estimate in DivBayes.

Parameter values used in simulations of SubT fall within the 95% credibility interval for all but three estimations (for different number of "substitute taxa") made on one tree. In APE the net diversification parameter used in the simulations occurs outside the 95% confidence interval in four cases and in two cases for the relative extinction rate. The ML estimates for net diversification rates from LASER and APE were never more than 0.015 units different from the median of the posterior distribution in SubT (Supplementary 4).

4 CONCLUSIONS

We present two freely available, stand-alone programs for Bayesian analyses of diversification rates: DivBayes and SubT. Both produce reasonable estimates of net diversification and relative extinction rates (comparable to ML estimates) with decent computational times. Both programs also incorporate uncertainty in age estimates of taxa, a significant advantage in dealing with datasets characterized by incomplete taxon sampling.

ACKNOWLEDGEMENTS

We are grateful for comments made by three reviewers on previous versions of this manuscript. *Funding*: This work was supported by a National Science Foundation grant [grant number DEB-0949517]; and the University of Tennessee.

REFERENCES

Alfaro,M.E. (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. U.S.A 106, 13410-13414.

Bailey,N.T.J. (1964) The elements of stochastic processes with applications to the natural sciences. John Wiley and Sons , New York.

Bokma,F. (2008) Bayesian estimation of speciation and extinction probabilities from (in)complete phylogenies. Evolution, 62, 2441-2445.

FitzJohn,R.G., et al. (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol., 58, 595-611.

Gilks,W., et al. eds. (1995) Markov chain Monte Carlo in practice. Chapman & Hall/CRC, Boca Raton, FL.

Harmon,L.J., et al. (2008) GEIGER: investigating evolutionary radiations. Bioinformatics, 24, 129-131.

Larget,B.R., et al. (2010) BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics, 26, 2910-2911.

Lartillot,N., et al. (2009) PhyloBayes 3. A Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics, 25, 2286-2288.

Nee,S., et al. (1994). Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. Lond., B, Biol. Sci., 344, 77-82.

Pagel,M., et al. (2004) Bayesian estimation of ancestral character states on phylogenies. Syst. Biol., 53, 673–684.

Paradis,E. et al. (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics, 20, 289-290.

R Development Core Team. (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky,D. (2006) LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Evol. Bioinformatics Online, 2, 247-250.

Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19, 1572–1574.

Ryberg,M. and Matheny,P.B. (2011) Dealing with incomplete taxon sampling and diversification of a large clade of mushroom-forming fungi. Evolution, in press DOI: 10.1111/j.1558-5646.2011.01251.x.