DivBayes and SubT: exploring species diversification using Bayesian statistics

Ryberg, M1,*, Nilsson, R.H.2,3 and Matheny, P.B1
1Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610, USA
2Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden.
3Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St, 51005 Tartu, Estonia

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.

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 data sets featuring extinctions. Default values were used for priors and starting values.

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


