BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 7 2010, pages 962–963
doi:10.1093/bioinformatics/btq062

Phylogenetics

Advance Access publication February 15, 2010

Phybase: an R package for species tree analysis

Liang Liu[1,*] and Lili Yu[2]

[1]Department of Agriculture and Natural resources, Delaware State University, Dover, DE 19901 and [2]Department of Biostatistics, Georgia Southern University, Statesboro, GA 30458, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Phybase is an R package for phylogenetic analysis using species trees. It provides functions to read, write, manipulate, simulate, estimate, summarize and plot species trees, which contain not only the topology and branch lengths but also population sizes.

Availability: The Phybase package is available at the R repository. The manual and supporting materials including source code, sample R code and sample data files for the species tree analysis are available at http://stat.osu.edu/~liuliang/research/phybase.html

Contact: lliu@desu.edu

Received on January 20, 2010; revised on February 9, 2010; accepted on February 10, 2010

1 INTRODUCTION

Species trees are the fundamental tools in studying biodiversity and evolutionary history. It has been appreciated that phylogenies of genes (gene trees) are distinct from phylogenies of species (species trees) (Doyle, 1992; Edwards, 2009; Maddison, 1997; Maddison and Knowles, 2006). However, most phylogenetic programs are designed for estimation of gene trees and lack functions to perform species tree analysis. As an example, when species trees are recorded as modified Newick files, the population size parameter (θ) in the species tree is not readable by most phylogenetic programs including PAUP (Swofford, 2003), PHYLIP (Felsenstein, 2005) and MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), making it difficult to summarize species trees in these phylogenetic programs. It is desirable to have a phylogenetic program that can provide functions necessary for the analysis of species trees. Phybase is an R package designed specifically for species tree analysis, although it is also capable of analyzing gene trees. It provides functions to read, write, manipulate, simulate, estimate and summarize species trees.

The input/output functions in the package can read and write DNA sequences and phylogenetic trees (gene trees and species trees) in the Newick format. Phylogenetic trees are read in as a string and then transformed to a matrix, which describes the ancestral relationships of nodes and branch lengths. The nodes matrix provides an easy access for developers to further manipulate the tree, while the tree string provides a useful interface to other phylogenetic packages in R such as the package APE (Paradis *et al.,* 2004). Basic functions are available in the package to manipulate phylogenetic trees including deleting and swapping nodes, rooting and unrooting trees. Phybase includes functions to summarize collections of species trees such as those that arise from bootstrap analysis or the posterior distribution

*To whom correspondence should be addressed.

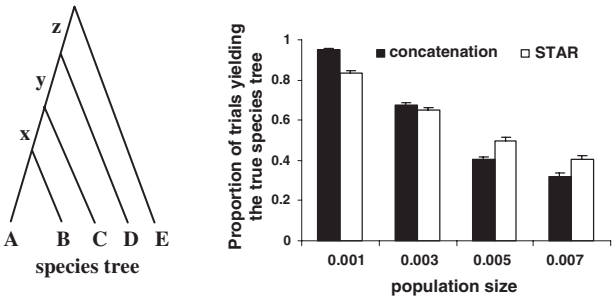

Fig. 1. The performance of STAR and concatenation in estimating the species tree as the population size changes. The lengths of internal branches in the species tree are $x = y = z = 0.001$ (in mutation units). The population size is constant across all populations in the species tree. The bars represent the standard errors of the proportions.

in Bayesian analysis of species trees (Liu *et al.,* 2008). Phybase is able to simulate DNA sequences from a prespecified species tree. Gene trees can be generated from a species tree under the coalescent model (Rannala and Yang, 2003), which assumes a molecular clock for gene trees and the species tree. Alternatively, gene trees can be generated from a non-clock-like species tree which allows variable mutation rates along the branches (populations) in the species tree (Mossel and Roch, 2007). The variable mutation rate is most likely due to changes in generation time as well as changes to the overall rate of mutation along the branches (populations) in the species tree. Phybase has functions to generate DNA sequences from gene trees under a variety of substitution models (Felsenstein, 2004). The species tree can be simulated from another phylogenetic package PhySim in R.

2 A PHYBASE APPLICATION: COMPARING THE PERFORMANCE OF THE STAR METHOD WITH CONCATENATION IN ESTIMATING SPECIES TREES

As an example of the utility of Phybase, we used Phybase to simulate DNA sequences from a 5-taxon species tree (Fig. 1). For simplicity, we assume a constant population size across all populations in the species tree. The population size and branch lengths in the species tree are in mutation units. The population size is set to be 0.001, 0.003, 0.005 and 0.007 to investigate the effect of the population size on the performance of STAR (Liu *et al.,* 2009) and concatenation (Huelsenbeck *et al.,* 1996) in estimating species trees. We generated 10 gene trees from the species tree (an allele per species) under the coalescent model (Rannala and Yang, 2003). DNA sequences of

500 bp were generated from the 10 gene trees with the HKY model (Hasegawa *et al.,* 1985). The simulated DNA sequences were then treated as data to estimate the species tree using the concatenation and STAR methods. Phybase does not have functions to calculate maximum likelihood (ML) gene trees. For the STAR method, we used the phylogenetics program PhyML (Guindon and Gascuel, 2003) to estimate ML gene trees without a molecular clock. We do not need to estimate coalescence times in gene trees, because STAR estimates species trees on the basis of the ranks of coalescence times. The simulation was repeated 1000 times. The performance of STAR or concatenation is measured by the proportion of trials yielding the true species tree. According to coalescent theory, the proportion of gene trees matching the species tree decreases as the population size increases. Higher proportion of gene trees matching the species tree implies that it is easier to recover the true species tree. Thus, we expect that the proportion of trials yielding the true species tree for both methods will increase as the population size decreases. The results of the simulation agree with our expectation, showing that both methods perform better as the population size decreases (Fig. 1). In addition, the results indicate that concatenation outperforms STAR at the population sizes 0.001 and 0.003, while STAR outperforms concatenation at the population sizes 0.005 and 0.007. For the population size 0.007, 110 and 160 genes are required, respectively, for STAR and concatenation to successfully recover the true species tree for all trials. Since the concatenation method assumes concordant gene trees, it may perform poorly for the cases when the concordant gene trees assumption is seriously violated (Kubatko and Degnan, 2007). This explains the poor performance of concatenation when the population size is equal to 0.005 and 0.007, and the corresponding proportions of concordant gene trees are 0.124 and 0.089. In contrast, concatenation performs well when the population size is 0.001 and the corresponding proportion of concordant gene trees is 0.717. This simulation study suggests that the concatenation method for estimating species trees is suitable for the cases when the majority of gene trees are similar to each other, although we expect the confidence in the tree generated by concatenation to be inflated nonetheless (Liu *et al.,* 2008). If the data indicate highly incongruent gene trees, the STAR method may be more appropriate than concatenation for the species tree estimation.

3 CONCLUSION

Phybase is available under the GNU General Public License (GPL). As an R package, Phybase inherits the open source nature of R. We invite others to continue to contribute phylogenetics inference packages in the R framework.

ACKNOWLEDGEMENTS

We thank Scott Edwards and reviewers for helpful suggestions on the first draft of the manuscript and the manual of the package.

Funding: National Science Foundation (grant Deb 0743616) to Scott Edwards and Dennis Pearl.

Conflict of Interest: none declared.

REFERENCES

Doyle,J.J. (1992) Gene trees and species trees—molecular systematics as one-character taxonomy. *Syst. Bot.*, **17**, 144–163.

Edwards,S.V. (2009) Is a new and general theory of molecular systematics emerging? *Evolution*, **63**, 1–19.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Felsenstein,J. (2005). PHYLIP. Department of Genome Science, University of Washington, Seattle.

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

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–74.

Huelsenbeck,J.P. *et al.* (1996) Combining data in phylogenetic analysis. *Trends Ecol. Evol.*, **11**, 152–158.

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

Kubatko,L and Degnan,J (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.*, **56**, 17–24.

Liu,L. *et al.* (2008) Estimating species trees using multiple-allele DNA sequence data. *Evolution*, **62**, 2080–2091.

Liu,L. *et al.* (2009) Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, **58**, 468–477.

Maddison,W.P. (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 523–536.

Maddison,W.P. and Knowles,L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.*, **55**, 21–30.

Mossel,E. and Roch,E. (2010) Incomplete lineage sorting: consistent phylogeny estimation from multiple Loci. *IEEE/ACM Trans Comput. Biol. Bioinform.*, **7**, 166–171.

Paradis,E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Rannala,B. and Yang,Z.H. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

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

Swofford,D.L. (2003) *Phylogenetic analysis using parsimony*. Sinauer Associates, Sunderland, Massachusetts.