BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 21 2008, pages 2542–2543
doi:10.1093/bioinformatics/btn484

Phylogenetics

BEST: Bayesian estimation of species trees under the coalescent model

Liang Liu

Musuem of Comparative Zoology, Harvard University, Cambridge, MA, USA

Received on August 8, 2008; revised and accepted on September 9, 2008

Advance Access publication September 17, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: BEST implements a Bayesian hierarchical model to jointly estimate gene trees and the species tree from multilocus sequences. It provides a new option for estimating species phylogenies within the popular Bayesian phylogenetic program MrBayes. The technique of simulated annealing is adopted along with Metropolis coupling as performed in MrBayes to improve the convergence rate of the Markov Chain Monte Carlo algorithm.

Availability: http://www.stat.osu.edu/~dkp/BEST.

Contact: lliu@oeb.harvard.edu

The correct estimation of species trees (phylogenies of species) is one of the central problems in evolutionary biology. It is widely accepted that the evolutionary history of species is a stochastic process and should be modeled in a statistical framework (Maddison and Knowles, 2006). In addition, it has recently become better appreciated that the phylogenies of genes (gene trees) are distinct (and often different) from the phylogenies of species in which gene trees are embedded (Felsenstein, 2004). BEST takes advantage of the information from multiple gene trees and performs a Bayesian analysis to estimate the topology of the species tree, divergence times and population sizes, while another popular program, MCMCcoal (Rannala and Yang, 2003), can only estimate species divergence times and population sizes. The Bayesian hierarchical model that BEST implements consists of three components: sequences, gene trees and the species tree (Edwards *et al*., 2007; Liu and Pearl, 2007; Liu *et al*., 2008). The model assumes that discrepancies between gene trees and the species tree are due exclusively to lineage sorting with free recombination between genes and no recombination within genes. Thus, it is not appropriate to implement BEST to reconstruct species trees when discrepancies between gene trees and the species tree are caused by other biological phenomena, such as horizontal transfers or gene duplications/deletions. The algorithm samples from the joint posterior distribution over a set of gene trees (**G**) and the species tree (S), that is,

$$f(S,\mathbf{G},\lambda|D)=\frac{f(D|\mathbf{G},\lambda)f(\lambda)f(\mathbf{G}|S)f(S)}{f(D)}.$$

in which $f(D|\mathbf{G},\lambda)$ is the probability density of sequence data D given gene trees $\mathbf{G}$ and parameters λ of the substitution model, $f(\mathbf{G}|S)$ is the probability density of gene trees $\mathbf{G}$ given the species tree S, $f(\lambda)$ is the prior of parameters λ, and $f(S)$ is the prior distribution of the species tree S. Marginalizing over S then yields an estimate of the posterior distribution over the species tree. BEST employs the Metropolis–Hastings algorithm (Hastings, 1970; Metropolis *et al*., 1953) to estimate the posterior distribution of parameters. The two-step Markov chain Monte Carlo (MCMC) algorithm performed in version 1.6 and 1.7 of BEST (Liu *et al*., 2008) has been merged into a single algorithm (in version 2.0) in which gene trees and the species tree are both updated with each new proposal.

Under the coalescent model, the species divergence times are restricted by the gene coalescence times (Liu *et al*., 2008; Rannala and Yang, 2003). As noted by many authors, if the species tree is fixed, it arbitrarily rules out those gene trees in which coalescence times are smaller (more recent) than the corresponding species divergence times. On the other hand, the fixed gene trees restrict the species tree space by gene coalescence times. A chain with the traditional strategy of iteratively updating gene trees and the species tree is unable to move rapidly in the parameter space because gene trees and the species tree are highly correlated. BEST 2.0 updates gene trees and the species tree jointly. Given the current state ($\mathbf{G}_i$, S_i), a new state of gene trees $\mathbf{G}_{i+1}$ is proposed by the tree rearrangement schemes used in the MrBayes without any topological or temporal restriction and then a species tree S_{i+1} is proposed within the space restricted by the gene trees $\mathbf{G}_{i+1}$ by modifying a Poisson number of nodes of the maximum tree (MT_{i+1}) derived from the gene trees $\mathbf{G}_{i+1}$. The MT_{i+1} is the largest tree (in terms of the branch length) within the species tree space restricted by the gene trees $\mathbf{G}_{i+1}$ (Liu, 2006). It has been shown that if gene trees are given, the MT is a consistent estimator of the species tree (Mossel and Roch, 2008). Additionally, the MT is the maximum likelihood estimate of the species tree if populations in the species tree have equal population size θ (Liu, 2006). Thus, sampling species trees from a small neighborhood of MT (small Poisson mean) can increase the acceptance rate, while sampling from a larger neighborhood (large Poisson mean) may improve mixing but reduce the acceptance rate. The new state ($\mathbf{G}_{i+1}$. S_{i+1}) is either accepted or rejected according to the Metropolis–Hastings ratio

$$\min\left(1,\frac{f\left(\mathbf{G}_{i+1},S_{i+1}\right)}{f\left(\mathbf{G}_i,S_i\right)}\frac{f\left(D|\lambda,\mathbf{G}_{i+1},S_{i+1}\right)}{f\left(D|\lambda,\mathbf{G}_i,S_i\right)}\frac{q\left(\mathbf{G}_i,S_i|\mathbf{G}_{i+1},S_{i+1}\right)}{q\left(\mathbf{G}_{i+1},S_{i+1}|\mathbf{G}_i,S_i\right)}\right)$$

in which $f(\mathbf{G},\ S)=f(\mathbf{G}|S)*f(S)$ and $f(D|\lambda,\ \mathbf{G},\ S)=f(D|\lambda,\ \mathbf{G})$ because the sequences D are assumed conditionally independent of the species tree S given the gene trees $\mathbf{G}$. A simulated annealing technique is used to facilitate fast mixing in the parameter space (Kirkpatrick *et al*., 1983). The temperature is different for the prior of gene trees and the species tree $f(\mathbf{G},\ S)$. High temperature alleviates the prior effect and the chain is guided mainly by the likelihood

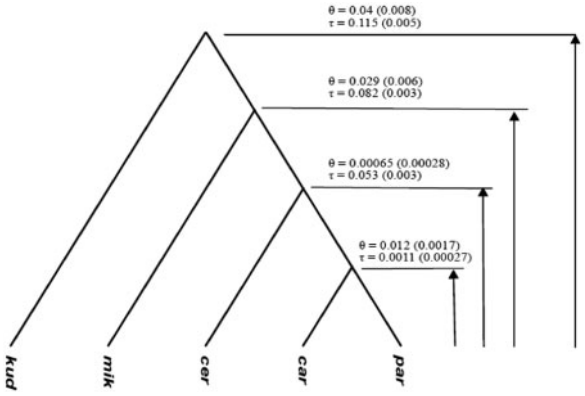

Fig. 1. The estimate of the species tree for the yeast data. The tree is plotted without the outgroup. Abbreviations for species: par, *S.paradoxus*; car, *S.cariocanus*; cer, *S.cerevisiae*; mik, *S.mikatae*; kud, *S.kudriavzevii*. θ, population size; and τ, divergence times. The posterior probability for this tree is 1.0.

$f(D|\mathbf{G}, \lambda)$. The temperature gradually cools down and reaches 1.0 at $P\%$ of the total number of generations. Thereafter, the chain becomes the regular Metropolis–Hastings algorithm and will converge to the posterior distribution. The annealing scheme can be turned off by setting $P = 0$.

A consensus tree is constructed from the estimated posterior distribution and used as the estimate of the species tree. BEST 2.0 estimates the divergence time and population size for each population by their posterior means. BEST 2.0 has been integrated into the popular Bayesian phylogenetic program MrBayes (Ronquist and Huelsenbeck, 2003). The BEST analysis is performed by setting BEST = 1 in the command 'prset' and it is switched to the traditional MrBayes by setting BEST = 0. For more details about the commands used in BEST 2.0, please refer to the BEST manual available at the BEST website.

The BEST analysis is demonstrated on a sample of 22 sequences collected for four genes from six Yeast species—*Saccharomyces cerevisiae* ($n = 9$ sequences), *S.paradoxus* ($n = 1$), *S.cariocanus* ($n = 5$), *S.mikatae* ($n = 3$), *S.kudriavzevii* ($n = 3$), and *S.bayanus* ($n = 1$) (Liti *et al*., 2006). The data were analyzed under the GTR + γ model for two genes and the HKY85 model for the other two genes. The prior distribution of population sizes was inverse gamma with $\alpha = 3$ and $\beta = 0.003$. The prior of the species tree was the uniform distribution. The MCMC algorithm was run for 20 000 000 generations and sampled every 2000 generations. The first 10 million generations were discarded as burn-in. The estimates of the population sizes and divergence times across populations are shown in Figure 1. The estimated posterior distribution of the species tree was summarized by a consensus tree which is consistent with previous results (Rokas *et al*., 2003). This yeast dataset is imbalanced with respect to the number of alleles across species. To investigate the effect of uneven number of alleles on the species tree estimation for this yeast dataset, an allele was randomly chosen from each yeast species and used as the data to estimate the species tree. Another single allele yeast dataset was also analyzed to reconstruct the species tree for the six yeast species (Rokas *et al*., 2003). Both analyses produced the same tree as that in Figure1, suggesting that the uneven number of alleles in the yeast dataset does not introduce systematic biasness in the species tree estimation.

ACKNOWLEDGEMENTS

I sincerely thank BEST users for reporting bugs and their useful suggestions for improving the performance of BEST. I am indebted to Patricia Brito for writing the manual for BEST1.6 and BEST1.7. I thank Justin Slauson and Dennis Pearl for constructing a nice website for BEST and Scott Edwards for suggestions on improving the user interface of the program. I thank the reviewers for their constructive comments on the first draft of the article.

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

Conflict of Interest: none declared.

REFERENCES

Edwards,S.V. *et al*. (2007) High-resolution species trees without concatenation. *Proc. Natl Acad. Sci. USA*, **104**, 5936–5941.

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

Hastings,W. (1970) Monte Carlo sampling methods using Markov chain and their applicaitons. *Biometrika*, **57**, 97–109.

Kirkpatrick,S. *et al*. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Liti,G. *et al*. (2006) Sequence diversity, reproductive isolation and species concepts in Saccharomyces. *Genetics*, **174**, 839–850.

Liu,L. (2006) Reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions. PhD dissertation. Department of Statistics, The Ohio State University, Columbus, pp. 46–49.

Liu,L. and Pearl,D.K. (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, **56**, 504–514.

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

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

Metropolis,N. *et al*. (1953) Equations of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087–1092.

Mossel,E. and Roch,S. (2008) Incomplete lineage sorting: consistent phylogeny estimation from multiple Loci. Available at http://arxiv.org/abs/0710.0262.

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.

Rokas,A. *et al*. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

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