BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 7 2009, pages 971–973
doi:10.1093/bioinformatics/btp079

Phylogenetics

STEM: species tree estimation using maximum likelihood for gene trees under coalescence

Laura S. Kubatko[1,*], Bryan C. Carstens[2] and L. Lacey Knowles[3]

[1]Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, [2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803 and [3]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

Received on November 28, 2008; revised and accepted on February 4, 2009

Advance Access publication February 10, 2009

Associate Editor: Martin Bishop

ABSTRACT

Summary: STEM is a software package written in the C language to obtain maximum likelihood (ML) estimates for phylogenetic species trees given a sample of gene trees under the coalescent model. It includes options to compute the ML species tree, search the space of all species trees for the k trees of highest likelihood and compute ML branch lengths for a user-input species tree.

Availability: The STEM package, including source code, is freely available at http://www.stat.osu.edu/~lkubatko/software/STEM/.

Contact: lkubatko@stat.osu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The increasing availability of sequence data from multiple loci for inferring phylogenetic trees has led to a growing awareness that the evolutionary histories of individual genes may differ substantially from the underlying species tree. This incongruence can result from numerous process, including horizontal transfer, gene duplication and incomplete lineage sorting (deep coalescence) (Maddison, 1997). When phylogenetic trees representing the species history are of primary interest, it is therefore necessary to either modify standard phylogenetic methods to handle multi-locus data, or to develop new methods that explicitly model the source of discord (Ane *et al*., 2007; Liu, 2008; Liu and Pearl, 2007). Although several recent studies have claimed that the commonly used procedure of concatenating multi-gene data prior to phylogenetic analysis performs well (Chen and Li, 2001; Rokas *et al*., 2003), others have highlighted situations in which such procedures fail (Carstens and Knowles, 2007; Kolaczkowski and Thornton, 2004; Kubatko and Degnan, 2007; Mossel and Vigoda, 2005).

Here, we describe a new software package called STEM that estimates the maximum likelihood (ML) species tree from a sample of gene trees, assuming that discord between the observed gene trees and the species tree arises solely from the coalescent process (Kingman, 1982). As is the case with other available programs for estimating species phylogenies from multilocus data [e.g. BEST, Liu (2008)], STEM assumes no recombination within loci, free recombination between loci and no gene flow following speciation. STEM provides the analytically derived ML estimate of the species trees when only a single estimate is desired. In addition, STEM provides a capability for searching the space of species trees for a collection of k species trees with high likelihood, where k is set by the user. Finally, STEM can compute ML branch lengths on any given species tree, which reduces the search for high-likelihood trees to a discrete (topology only) space, as well as allows evaluation of any species tree of interest.

As noted above, the programs BEST (Liu, 2008; Liu and Pearl, 2007) and BUCKy (Ane *et al*., 2007) are related to STEM in that they also seek to provide a species-level phylogenetic estimate. However, STEM is distinct from these in that (i) it uses a maximum likelihood, rather than Bayesian, framework to obtain an estimate; and (ii) the availability of analytic results in the ML case using gene trees as the data allow computations to be carried out more rapidly than the Markov chain Monte Carlo (MCMC)-based analyses utilized by these programs.

2 DESCRIPTION

2.1 Phylogenetic model

Let g_j denote the gene tree topology and branch lengths for the tree representing locus j ($j=1,2,....N$) in a sample of N loci. Assuming that the N loci are sampled independently throughout the genome, the likelihood function is

$$L(S,\tau)=\prod_{j=1}^{N} f(g_j|S,\tau) \quad (1)$$

where S represents the species tree and τ is the set of branch lengths on that tree. The function $f(.|.)$ is the gene tree density under the coalescent model given by Rannala and Yang (2003). We note that this density is general enough to allow samples of multiple lineages per species-level taxon. Membership of alleles to species-level taxa is specified as input to STEM.

The likelihood in (1) is a function of the parameter $\theta=4N_e\mu$, where N_e is the effective population size and μ is the per-site mutation rate. In the most general case, θ may vary along species tree branches. However, it is not uncommon to assume a single θ for the entire tree. For example, Liu (2006) showed that when it can be assumed that there is a single θ for the entire tree, it is possible to analytically derive the joint ML estimate of θ and of the species

*To whom correspondence should be addressed.

tree topology and branch lengths. He calls the estimator of the tree obtained in this way the Maximum Tree (MT), and shows that it is a consistent estimator of the species tree when the gene trees and branch lengths are known without error.

Mossel and Roch (2009) also consider a sample of gene trees with branch lengths known without error and derive a consistent estimator of the species tree in the case in which θ is known (but not necessarily equal) for all branches of the species tree, which they call the GLASS tree (an acronym for Global LAteSt Split, which is derived from the method used to compute it). The GLASS tree coincides with MT whenever it can be assumed that the θ along all branches of the species tree are the same and take their value from the MLE for θ. The relationship of the ML tree returned by STEM to these methods is noted below.

Input to the STEM program requires a sample of gene trees with branch lengths in units of expected number of nucleotide substitutions per site along with an overall value of θ to be applied to all loci. The value of θ is used to convert gene tree branch lengths into coalescent units (number of $2N_e$ generations) by multiplying all gene tree branch lengths by $1/\theta$. Further, because evolutionary rates may vary across sampled loci, the user may also provide rates to be applied to each locus separately. For example, if rate r_i is specified for locus i, then all branch lengths in gene tree i will be additionally multipied by $1/r_i$. In addition to adjusting for variation in the mutation rate of each locus, the r_i values allow the user to adjust for ploidy in the individual genes (e.g. the rate provided for an mtDNA locus should be divided by 2 to incorporate the haploid status of this marker). While selection of the θ and r_i values is completely at the discretion of the user, reasonable settings for these parameters can be straightforwardly obtained. For example, the θ parameter could be estimated by some available method, such as Watterson's estimator (Watterson, 1975). The r_i values could be estimated by examining average divergence from an outgroup, as suggested by Yang (2002).

2.2 STEM output

When the ML estimate of the species tree is requested, STEM returns the MT of Liu (2006) for the particular user-specified values of θ and the gene-specific rates. STEM is also able to evaluate the likelihood for any given species tree rapidly by incorporating a new result that analytically derives ML branch lengths for an arbitrary species tree under (1). The details of this result, which is an extension of the work of Liu (2006), are provided in Supplementary Material 1. In addition, STEM includes an option to search this space for a set of species trees of high likelihood using a simulated annealing algorithm, similar to that used by Salter and Pearl (2001).

2.3 Performance

We demonstrate the usefulness of the STEM package using simulated data. First, a sample of 10 gene trees is generated from the species tree in Figure 1a using the program COAL (Degnan and Salter, 2005). Branches y and z were set to 1.0 coalescent units, while branch length x was varied between 0.2 and 1.0 in increments of 0.2, to include settings in which inference of the species tree is known to be difficult (Kubatko and Degnan, 2007). The second step is the simulation of DNA sequence data along the sampled gene trees using Seq-Gen (Rambaut and Grassly, 1997).

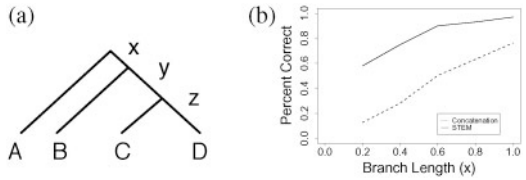

Fig. 1. (a) Model tree used for the simulations; (b) Results of the simulations comparing the performance of STEM to concatenation in terms of the percent of times the true species trees is obtained as a function of x.

Once the data are generated, ML estimates of the individual gene trees are obtained using the program PAUP* (Swofford, 2003) and then used as input to STEM. The entire simulation was repeated 100 times for each value of x. Figure 1b compares the results of the STEM program with the naive method of estimating a single ML tree from the concatenated sequence. For both methods (STEM and concatenation), the same mutation model (JC69) was used to generate data and to perform ML estimation in PAUP* in order to remove model misspecification as a source of error in species tree estimates. STEM clearly shows an improvement over concatenation in this setting, even when species tree branch lengths are short.

3 CONCLUSION

As the availability of multi-locus data for inference of species trees increases, the need for development of software to model relationships between gene and species trees is also increasing. STEM provides a computationally efficient method to estimate ML species phylogenies and to explore the likelihood surface under the coalescent model for a given sample of gene trees that will serve as a useful compliment to the more comptuationally intensive Bayesian methods (Ane *et al*., 2007; Liu, 2008) currently available.

ACKNOWLEDGEMENTS

We thank Liang Liu for generously sharing manuscripts during development of this software, and James Degnan and other anonymous reviewers for helpful comments on an earlier version.

Funding: NSF DMS-07-02277 (L.S.K.); NSF DEB-04-47224 (L.L.K).

Conflict of Interest: none declared.

REFERENCES

Ane,C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.*, **24**, 412–426.

Chen,F.-C. and Li,W.-H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.*, **68**, 444–456.

Carstens,B.C. and Knowles,L.L. (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. *Syst. Biol.*, **56**, 400–411.

Degnan,J.H. and Salter,L.A. (2005) Gene tree distributions under the coalescent process. *Evolution*, **59**, 24–37.

Kingman,J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.*, **13**, 235–248.

Kolaczkowski,B. and Thornton,J.W. (2004) Performance of maximum parsimony and maximum likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.

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

Liu,L. (2006) Reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions. PhD. Dissertation.

Liu,L. (2008) BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, **24**, 2542–2543.

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.

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

Mossel,E. and Roch,S. (2009) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinform.*

Mossel,E. and Vigoda,E. (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, **309**, 2207–2209.

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

Rannala,B. and Yang,Z. (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.

Salter,L. and Pearl,D. (2001) A stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Syst. Biol.*, **50**, 7–17.

Swofford,D.L. (2003) *PAUP* Phylogenetic analysis using parsimony (* and other methods)*, Version 4. Sinauer Associates, Sunderland, MA.

Watterson,G.A. (1975) On the number of segregation sites. *Theor. Popul. Biol.*, **7**, 256–276.

Yang,Z. (2002) Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. *Genetics*, **162**, 1811–1823.