BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 24 2008, pages 2932–2933
doi:10.1093/bioinformatics/btn556

Phylogenetics

AUGIST: inferring species trees while accommodating gene tree uncertainty

Jeffrey C. Oliver[1,2]

[1]Interdisciplinary Program in Insect Science, University of Arizona, Tucson, AZ 85721 and
[2]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA

Received on September 3, 2008; revised and accepted on October 23, 2008

Advance Access publication October 25, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: AUGIST (accomodating uncertainty in genealogies while inferring species tress) is a new software package for inferring species trees while accommodating uncertainty in gene genealogies. It is written for the Mesquite software system and provides sampling procedures to incorporate uncertainty in gene tree reconstruction while providing confidence estimates for inferred species trees.

Availability: http://www.lycaenid.org/augist/

Contact: jeffrey.oliver@yale.edu

1 INTRODUCTION

As gene sequence data for more loci become available, it is becoming increasingly easier to base estimates of phylogeny on multiple gene genealogies. Because a single gene tree does not necessarily reflect the evolutionary relationships of the species being investigated (Maddison, 1997; Pamilo and Nei, 1988), it is essential to base phylogenetic inferences on multiple, unlinked loci. Incorporating data from multiple loci has been implemented in a variety of methods. One widely used approach concatenates the sequences from the different loci, to form a single, 'super' sequence for phylogenetic analyses; however, this may lead to erroneous inferences (Kubatko and Degnan, 2007) and unnecessarily constrains, sometimes unreasonably, all gene genealogies to share the same topology and branch lengths. Ideally, unlinked loci should be analyzed independently, to accommodate the potentially different histories and rates of change among loci (Maddison, 1997; Takahata, 1989).

There are increasingly more methods to incorporate independent loci in phylogenetic inference, allowing each locus a unique evolutionary history. One approach is to analyze each locus separately, then generate a consensus tree as a best estimate; however, the utility of approach is limited, given the low probability of reciprocal monophyly in gene genealogies for recently diverged taxa (Hudson and Coyne, 2002; Rosenberg, 2003). That gene genealogies do not always reflect species' histories is a product of the population genetic processes underlying the transmission of genes through time (Tajima, 1983). Methods accommodating those population genetic processes responsible for the discordance between gene trees and species trees would allow more informed inferences of species phylogenies (Carstens and Knowles, 2007; Liu and Pearl, 2007; Maddison, 1997).

One general tree inference approach that has recently been applied to species tree estimation uses gene genealogies to evaluate species trees based on some objective function. These include minimizing the number of deep coalescence events (MDC; Maddison, 1997; Maddison and Knowles, 2006) and maximizing the coalescence probabilities of gene trees for a given species tree (ESP-COAL; Carstens and Knowles, 2007). These approaches are useful for inferring species trees from gene genealogies (e.g. Carling and Brumfield, 2008), but currently treat the underlying gene genealogies as known. Methods of phylogenetic inference should accommodate the uncertainty inherent in gene genealogies in estimates of species phylogeny (Maddison and Knowles, 2006).

Here, I present a method extending the objective function approach of species tree inference, simultaneously accommodating uncertainty in gene genealogies and providing measures of confidence in species tree estimates. AUGIST (accommodating uncertainty in genealogies while inferring species trees), a new package for the Mesquite software system (Maddison and Maddison, 2008), is available for implementation of the method described here.

2 DESCRIPTION

2.1 Method

I first present an informal description of the method, followed by an explicit description of implementation using the deep coalescences criterion. Figure 1 provides a schematic illustration of the method.

2.1.1 Informal description Multiple, unlinked loci are sampled to generate gene sequence data for specimens. These data are analyzed separately to create a distribution of gene genealogies for each locus. Each distribution of gene genealogies will ideally reflect the confidence in individual partitions by frequency of their occurrence in the distribution (Holder and Lewis, 2003). To estimate a species tree, a single gene genealogy for each locus is drawn, at random with replacement, and used to infer the species tree based on an optimality criterion. Topological variation among gene genealogies represents uncertainty in the gene tree, and is incorporated into analyses by resampling these gene genealogies to generate a distribution of species trees. The distribution of species trees can be summarized, for example, as a consensus of all 'optimal' species trees recovered in the resampling procedure. The consensus should reflect the frequency at which partitions in the species tree were encountered

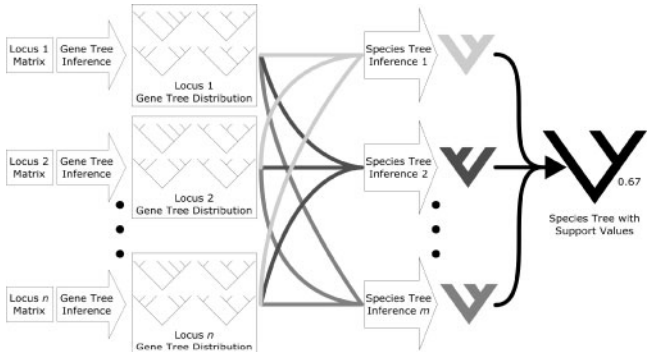

Fig. 1. Illustration of species tree interface methods. Gene genealogy distributions for *n* loci are generated by MCMC sampling. A single gene genealogy is drawn from these distributions for each locus, and the set of *n* genealogies is then used to infer species tree(s) based on some optimality criterion. This process is repeated *m* times, and summarized in a consensus species tree, shown in black, with node values reflecting frequency at which a partition is recovered. See text for explicit implementation details of gene genealogy distribution generation and species tree inference.

during the re-sampling procedure, and thus the uncertainty of species tree topology.

2.1.2 Explicit example Multiple loci are sequenced for multiple individuals and species. In independent Bayesian MCMC analyses in MrBayes (Huelsenbeck and Ronquist, 2001), these DNA sequences are used to generate posterior distributions of gene genealogies. Those genealogies sampled before log likelihood scores stabilize and runs converge are discarded. A single gene genealogy for each locus is selected from the posterior distributions of gene genealogies, and a species tree is inferred using the Tree Search procedure in Mesquite (Maddison and Maddison, 2008), minimizing the number of deep coalescences for multiple loci. This species tree inference step is repeated multiple times, drawing a new gene genealogy for each locus during each species tree search. All inferred species trees are summarized in a single tree, using the Majority Rule Consensus Tree function in Mesquite; the frequency of a partition recovered in the species tree inference step is used as a measure of uncertainty for the relationships in the consensus species tree.

2.2 Implementation details

AUGIST is available as a package for the Mesquite software system (Maddison and Maddison, 2008), and runs on Linux, Windows and Mac OS X. It requires at least version 2.5 build j55 of Mesquite to run properly. The package, instructions, and recommendations for use are available at http://www.lycaenid.org/augist/ or by request from the author. Results of a simulation study demonstrating the accuracy of the AUGIST approach are contained in Oliver (2007), and can be obtained by request from the author.

3 EXENTIONS

The method described here could be used for any tree inference method employing an optimality criterion based on the relationship between gene trees and species trees. For example, instead of deep coalescent events, the number of gene duplication and extinction events could be minimized to infer a species tree (Maddison, 1997). A likelihood approach, using gene tree likelihood based on coalescent probabilities (e.g. ESP-COAL, Carstens and Knowles, 2007) could also be used, although the current version of Mesquite does not calculate this statistic. The flexibility of Mesquite and the AUGIST package will accommodate other optimality criteria as additional Mesquite modules become available. Finally, this method could easily be extended to test phylogenetic hypotheses by comparing constrained versus unconstrained tree searches for a given set of gene genealogies. Phylogenetic hypotheses, represented by a constraint tree, could be rejected if constrained tree searches consistently infer species trees with more deep coalescences than unconstrained tree searches.

ACKNOWLEDGEMENTS

David Maddison provided invaluable insight and assistance in Mesquite programming. Four anonymous reviewers also provided useful comments on this article and AUGIST implementation.

Funding: Center for Insect Science at the University of Arizona and an NSF DDIG (#0412447).

Conflict of Interest: none declared.

REFERENCES

Carling,M.D. and Brumfield,R.T. (2008) Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in *Passerina* buntings. *Genetics*, **178**, 363–377.

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

Holder,M. and Lewis,P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, **4**, 275–284.

Hudson,R.R. and Coyne,J.A. (2002) Mathematical consequences of the genealogical species concept. *Evolution*, **56**, 1557–1565.

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

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

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

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.

Maddison,W.P. and Maddison,D.R. (2008) Mesquite: a modular system for evolutionary analysis. Version 2.5, build j55. Available at http://mesquiteproject.org.

Oliver,J.C. (2007) Inferring species trees from deep coalescences while accommodating gene tree uncertainty. PhD Thesis, University of Arizona, Tucson.

Pamilo,P. and Nei,M. (1988) Relationships between gene trees and species trees. *Mol. Biol. Evol.*, **5**, 568–583.

Rosenberg,N.A. (2003) The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, **57**, 1465–1477.

Tajima,F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Takahata,N. (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, **122**, 957–966.