BIOINFORMATICS

Vol. 00 no. 00 ????
Pages 1–2

Multilocus species delimitation in BEAST

Graham Jones [1]*

[1]21e Balnakeil, Durness, Lairg, IV27 4PT, UK.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

This note describes a method for species delimitation using molecular sequence data which is based on the multispecies coalescent model in a Bayesian framework. It does not require a guide tree or prior assignment of individuals to clusters or species, but instead explores the full space of possible clusterings and tree topologies. It uses an approximation to avoid the need for reversible-jump MCMC. It is implemented as part of BEAST and requires only a few changes from a standard *BEAST analysis.

Availability: Software available at http://code.google.com/p/beast-mcmc/

Contact: art@gjones.name, www.indriid.com

Supplementary information: Supplementary material is available at *Bioinformatics* online.

DRAFT VERSION 2

1 INTRODUCTION

Many methods have been proposed for the task of species delimitation (e.g., see Miralles and Vences (2013)). The present method (called ‘Dissect’) is closest to that of Yang and Rannala (2010), which also uses the multispecies coalescent model. Their method employs a user-supplied guide tree in which some nodes may be collapsed (i.e., all descendants of these nodes assigned to one species). Collapsing a node is equivalent to setting its height to zero, since the multispecies coalescent density is the same for a single population and a population which has just split at time zero. When a node is collapsed, the dimensionality of the parameter space changes, so a reversible-jump MCMC algorithm is needed to sample the species trees. The basic idea behind Dissect is to sample trees in which each tip represents a single individual (or a cluster of individuals which definitely belong in one species), but replace the usual prior density on node heights with one which includes a spike near zero. The dimensionality of the parameter space is fixed, but nodes whose heights have a high posterior probability of being below a threshold can be regarded as collapsed.

2 DESCRIPTION

2.1 The model

In Bayesian phylogenetic analysis, a prior distribution over species trees is needed, and for rooted trees as used here, the reconstructed birth-death process (Gernhard, 2008) is often used. In this process, the densities of the unordered node heights, when conditioned on the origin time of the tree t, the speciation rate λ, and extinction rate μ, are i.i.d. and are also independent of the number of tips n (Gernhard, 2008, Theorem 2.5). This nice mathematical property makes the present model tractable. Let the density of a node height s be $f(s|n,t,\lambda,\mu) = f(s|t,\lambda,\mu)$. Suppose f is replaced with a mixture $(1-w)f(s|t,\lambda,\mu) + wm(s)$ where $w \in [0,1]$ is a user-chosen value which expresses a prior opinion about the probable number of species present: the prior mean for the number of species is $1+(n-1)(1-w)$. Suppose first that $m(s)$ is the Dirac delta function $\delta(s)$. Then the distribution over trees is one in which the trees with k external branches of nonzero length have total probability mass $\binom{n-1}{k-1}(1-w)^{k-1}w^{n-k}$. These collapsed trees have the same distribution as the reconstructed birth-death process on k tips, conditioned on t, λ, and μ. Using $m(s) = \delta(s)$ would require a reversible-jump MCMC method. However, if $m(s) = \epsilon^{-1}\mathbf{1}_{[0,\epsilon]}(s)$, where ϵ is small, ordinary MCMC operators can be used. The model is completed by assuming a prior density for the origin time t which does not depend on k. The supplementary information contains more details.

2.2 Using the software

The analysis can be run in version 1.8pre of BEAST (Drummond *et al.*, 2012). BEAUTi can be used to set up most of the analysis, as if for a *BEAST analysis. The word ‘species’, as it appears in BEAUTi and in the BEAST XML file, is regarded as a minimal cluster of individuals. Each species should only be assigned individuals which definitely belong together, since Dissect will consider merging but never splitting these minimal clusters. Two changes need to be made to the XML file. The birth-death model must be replaced with a birth-death-collapse model, where ϵ can be set, and an operator must be added for the origin height. The parameter w can either be given a fixed value, or estimated by adding a hyperprior and an operator. The supplementary information includes instructions for obtaining the required version of BEAST, an example XML file, and an R script for visualizing the results.

The parameter ϵ, which is measured in units of substitutions per site per generation, should be set to a small value such as 0.0001. This appears to be small enough for most analyses, in the sense that

*to whom correspondence should be addressed smaller values will give similar results more slowly, but more experience is needed. Extremely small values may lead to poor mixing. If ϵ is too large it will not be possible to distinguish very recent divergences.

The trees sampled from the posterior can be analyzed with a tool SpeciesDelimitationAnalyser. The user supplies a threshold (either ϵ or larger) for assigning individuals to clusters. The choice of this threshold is inevitably subjective to some extent, and depends on one's definition of 'species'. SpeciesDelimitationAnalyser produces a table of possible clusterings with posterior probabilities.

2.3 Caveats

The multispecies coalescent model entails several assumptions: no recombination within genes; free recombination between genes; no paralogs; no speciation via hybridization; no horizontal gene transfer. Perhaps most importantly in the context of species delimitation, speciation is regarded as instantaneous with no introgression afterwards. Violation of these assumptions may produce misleading results. The MCMC chain explores a huge space of possible clusterings of individuals and tree topologies, so several long runs with different seeds should be performed to guard against poor mixing.

3 EXAMPLE

The example uses a data set from the rodent genus *Thomomys* (pocket gophers), which was previously analyzed by Belfiore *et al.* (2008) using BEST and by Heled and Drummond (2010) using *BEAST. For the analysis here, the 26 individuals were assigned to separate 'species' in the BEAST XML file, w was set to 0.7, and ϵ to 0.0001. The threshold in SpeciesDelimitationAnalyser was set to 0.0005. The results are summarized in Fig. 1. One can see for example that two clusters (rows 2-6 and 7-12) within *Thomomys bottae* emerge, but the *Thomomys bottae bottae* individual (row 1) appears closer to both clusters than they are to one another, a pattern suggestive of migration between populations, or if one is a 'splitter', introgression between species. The supplementary information contains some results on simulated data.

Acknowledgement: I thank Bengt Oxelman for valuable comments.

Funding: None.

REFERENCES

Belfiore, N. M., Liu, L., and Moritz, C. (2008) Multilocus Phylogenetics of a Rapid Radiation in the Genus Thomomys *Syst. Biol.*, **57(2)**, 294-310.

Drummond, Alexei J., Suchard, Marc A., Xie, Dong, and Rambaut, Andrew (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Gernhard, T. (2008) The conditioned reconstructed process. *J. Theo. Biol.*, **253**, 769–778.

Heled, J. and Drummond, A. (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, **27**, 570–580.

Miralles, A. and Vences, M. (2013) New Metrics for Comparison of Taxonomies Reveal Striking Discrepancies among Species Delimitation Methods in Madascincus Lizards. *PLoS ONE*, **8(7)**, e68242.

Yang, Z. and Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of U.S.A*, **107**, 9264–9269.

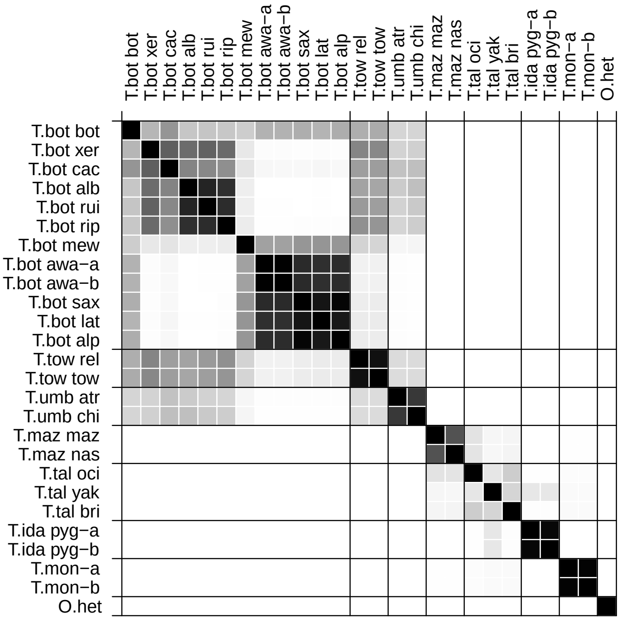

Fig. 1. Similarity matrix for 26 individuals. The squares represent posterior probabilities (white=0, black=1) for pairs of individuals to belong to the same cluster. In the labels, 'T' stands for *Thomomys*, 'O' for *Orthogeomys* and species and subspecies designations are abbreviated to the first three letters.