Bioinformatics, 31(7), 2015, 991–998
doi: 10.1093/bioinformatics/btu770
Advance Access Publication Date: 23 November 2014
Original Paper

Phylogenetics

DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent

Graham Jones[1], Zeynep Aydin[1,2] and Bengt Oxelman[1,*]

[1]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE 405 30 Göteborg, Sweden and [2]Department of Biology, Faculty of Sciences, University of Dicle, 21280 Diyarbakir, Turkey

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on March 5, 2014; revised on November 14, 2014; accepted on November 17, 2014

Abstract

Motivation: The multispecies coalescent model provides a formal framework for the assignment of individual organisms to species, where the species are modeled as the branches of the sp tree. None of the available approaches so far have simultaneously co-estimated all the relevant parameters in the model, without restricting the parameter space by requiring a guide tree and/or prior assignment of individuals to clusters or species.

Results: We present DISSECT, which explores the full space of possible clusterings of individuals and species tree topologies in a Bayesian framework. It uses an approximation to avoid the need for reversible-jump Markov Chain Monte Carlo, in the form of a prior that is a modification of the birth–death prior for the species tree. It incorporates a spike near zero in the density for node heights. The model has two extra parameters: one controls the degree of approximation and the second controls the prior distribution on the numbers of species. It is implemented as part of BEAST and requires only a few changes from a standard *BEAST analysis. The method is evaluated on simulated data and demonstrated on an empirical dataset. The method is shown to be insensitive to the degree of approximation, but quite sensitive to the second parameter, suggesting that large numbers of sequences are needed to draw firm conclusions.

Availability and implementation: http://tree.bio.ed.ac.uk/software/beast/, http://www.indriid.com/dissectinbeast.html.

Contact: bengt.oxelman@gu.se, www.indriid.com

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

1 Introduction

Despite its alleged status as a fundamental concept in biology, the species category has lacked a definition allowing explicit testing of particular species limits (e.g. de Queiroz, 2007). In recent years however, several methods have been proposed for the task of delimiting species based on molecular data (see Fujita *et al*. [2012] and Miralles and Vences [2013] for reviews). Multispecies coalescent (Rannala and Yang, 2003) species delimitation (MSCSD) methods make use of multi-locus sequence data to make inferences in the presence of incomplete lineage sorting.

All current MSCSD methods are either heuristic (e.g. O'Meara, 2010), dependent on a guide tree (e.g. Satler *et al*., 2013; Yang and Rannala, 2010; note however that a paper by Yang and Rannala [2014] appeared during the revision of this article, where the requirement of a user-supplied guide tree is eliminated) or are validation methods, which require prior assignment of individuals to clusters or species. Knowles and Carstens (2007) devised a maximum-likelihood approach, which uses fixed gene trees as input data and hierarchical likelihood ratio tests to compare different species classifications. These are treated as different stochastic models with different sets of parameters, and the hierarchical likelihood ratio tests require the models to be nested. Thus, for example, the classification of putative species A, B and C into AB and C or A and BC cannot be compared in this way, whereas ABC can be compared with either. A Bayesian alternative which takes uncertainty in gene tree estimation and does not require compared classifications to be nested is to use Bayes factors, which can be achieved from accurate marginal likelihood estimates (Baele *et al.*, 2012; Xie *et al.*, 2011). Grummer *et al.* (2014) and Aydin *et al.* (2014) used this approach to choose among species classifications, and Leaché *et al.* (2014b) extended the approach to be used for single-nucleotide polymorphism data.

O'Meara (2010) devised parametric and non-parametric heuristic methods to simultaneously find an optimal assignment of individuals to species and their tree relationships. Yang and Rannala (2010, 2014; Rannala and Yang, 2013) developed the idea in a Bayesian framework, in which the gene trees are co-estimated with a constrained species tree. In the simplest option, species are inferred by setting a threshold on the posterior node heights of the species tree, with small heights interpreted as evidence for collapsing a node. This is similar to using *BEAST (Heled and Drummond, 2010) with each individual in its own 'species' in the XML file, and estimating the actual species afterwards. The dimensionality of the parameter space does not change, and there is no special prior involved.

Here, we take a Bayesian approach, which has the advantage that nuisance parameters can be integrated out, and also that prior taxonomic knowledge can be taken to account. Our approach does not require prior assignments of individuals to putative species, and may be viewed as species tree inference while taking uncertainties in MSCSD into account. We present Division of Individuals into Species using Sequences and Epsilon-Collapsed Trees (DISSECT) for species delimitation which requires no prior assignment of individuals to clusters or species, but instead explores the full space of possible clusterings and tree topologies. It is along the lines of the method of Yang and Rannala (2010) which employs a user-supplied guide tree in which some nodes may be collapsed (i.e. all descendants of these nodes assigned to one species). In the most recent version of BP&P (Yang and Rannala, 2014), there is no need for the user to supply a guide tree. Instead, the 'guide tree space' is explored using nearest-neighbor interchange moves. The two operations of collapsing a node, and of setting its height to zero, have the same effect on the likelihood, since the multispecies coalescent density is the same for a single population and a population which has just split at time 0. When a node is collapsed, the dimensionality of the parameter space changes, so a reversible-jump Markov Chain Monte Carlo (rjMCMC) 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 within the spike can be interpreted as 'probably collapsed'.

2 Methods

A set of individual organisms will be called a cluster. Each possible cluster of individuals in the analysis is a candidate for constituting a species. A set of clusters which do not overlap one another and which together include all the individuals in the analysis will be referred to as a clustering. In an analysis using DISSECT, some sets of individuals may be grouped by the user as minimal clusters: these may be merged but never split. We use 'gene' in a loose sense, to mean an alignment of a sequence region which is assumed to be homologous and unlinked to other such regions. A 'gene copy' is a single row from such an alignment.

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 most often used. It includes the Yule process as a special case. The process is assumed to begin at some time t in the past with a single species, and is conditioned on producing the observed number of species at present. The time t is called the 'origin time' or 'origin height.' Theorem 2.5 of Gernhard (2008), following Thompson (1975) shows that, conditioned on t, the speciation rate λ, and the extinction rate μ, the density of the unordered node heights are independently and identically distributed and are also independent of the number of tips k. This nice mathematical property makes the present model tractable. Let the density of a node height s be $f(s|k, t, \lambda, \mu) = f(s|t, \lambda, \mu)$. In the present model, $f(s)$ is replaced with a mixture of $f(s)$ and another density $m(s)$ for s:

$$(1-\omega)f(s|t,\lambda,\mu)+\omega m(s), \tag{1}$$

where ω is a user-chosen weight in [0,1], and this density is used for all the n–1 node heights in a tree with n tips. The joint density is then

$$\prod_{i=1}^{n-1}((1-\omega)f(s_i|t,\lambda,\mu)+\omega m(s_i)), \tag{2}$$

where $s_1, \ldots, s_{n-1}$ are unordered node heights. This can be expanded as

$$\sum_{k=1}^{n}(1-\omega)^{k-1}\omega^{n-k}\sum_{X\in C(k)}\prod_{i\in X}f(s_i|t,\lambda,\mu)\prod_{i\notin X}m(s_i),$$

where $C(k)$ is the set of subsets of $\{1, \ldots, n-1\}$ of size $k-1$. If $m(s)$ was the Dirac delta function $\partial(s)$ (Dirac, 1958), the result would be a distribution in which the trees with k external branches of non-zero length (i.e. the trees with k 'real' tips) have total probability mass

$$|C(k)|(1-\omega)^{k-1}\omega^{n-k}=\binom{n-1}{k-1}(1-\omega)^{k-1}\omega^{n-k}. \tag{3}$$

Note that the product $\prod_{i\in X}f(s_i|t,\lambda,\mu)$ is the density for a reconstructed birth–death process with k tips whose node heights are the $k-1$ non-zero s_i. In practice one cannot sample from such a distribution without implementing rjMCMC, but it can be approximated it using

$$m(s)=\varepsilon^{-1}1_{[0,\varepsilon]}(s), \tag{4}$$

where ε is small.

Figures 1 and 2 illustrate the densities f and $(1-\omega)f+\omega m$, respectively, for the case $n=3$, where there are two internal node heights. One way of sampling trees from the reconstructed birth–death process for $n=3$ is to pick a point (x, y) from a density such as the one in Figure 1; then choose a random ordering of the tip labels from left to right; then insert x and y between them; and finally join the nodes to form the tree. The same process is shown in Figure 2 for the mixture density m. If the point (x, y) is in one of the

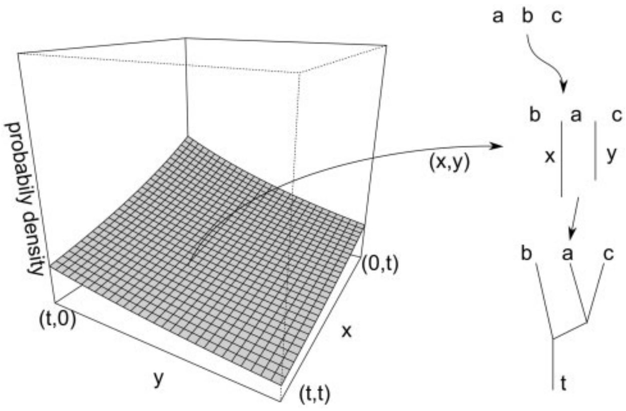

Fig. 1. Sampling trees from the usual birth–death density

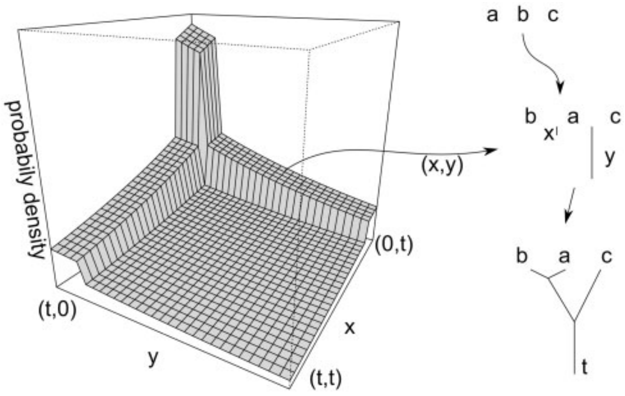

Fig. 2. Sampling trees from the mixture density

two 'walls' along the axes, one node will be collapsed. If the point (x, y) is in the 'pillar' near the origin, both nodes will be collapsed. The approximation means that there is a possibility that a true speciation which is more recent than ε will be missed.

This is very similar to a model in which a separate reconstructed birth–death process is assumed for each k and a rjMCMC is used to sample from the clusterings and trees. Apart from the approximation involving ε, the other difference is that the density $q(t|k)$ for t would normally depend on k in the reversible-jump version, whereas in Equation (2) there is no such dependence: a single density for t for all k is needed. It seems reasonable to assume a density for $q(t)$ which mixes $q(t|k)$ using the probabilities from Expression (3). In a normal BEAST or *BEAST analysis using the birth–death prior, an improper uniform prior on $[0, \infty)$ is assumed for the origin time t of the tree, and the process is then conditioned on the number of species k. The conditional density for t is shown in Theorem 3.2 of Gernhard (2008) to be

$$q(t|k) = k\lambda^k(\lambda - \mu)^2 \frac{\left(1 - e^{-(\lambda-\mu)t}\right)^{(k-1)} e^{-(\lambda-\mu)t}}{\left(\lambda - \mu e^{-(\lambda-\mu)t}\right)^{k+1}}. \qquad (5)$$

Using the probabilities from Expression (3), the prior density for t is

$$q(t) = \sum_{k=1}^{n} \binom{n-1}{k-1} (1-\omega)^{k-1}\ \omega^{n-k} q(t|k). \qquad (6)$$

This can be simplified as shown in the Supplementary Information.

The model was implemented in BEAST by adding a class BirthDeathCollapseModel, which is similar to the usual BirthDeathModel. It contains a parameter for the origin height t as well as for the diversification rate and relative death rate as in the usual birth–death model. An additional MCMC operator is needed to sample from t. This can be added using one of the existing operators in the XML. We used a ScaleOperator. No new MCMC operators were added to explore the space of species trees: the existing NodeReHeight operator explores the posterior as modified by the prior in Equation (2).

2.2 DISSECT workflow

The analysis can be run in BEAST (Drummond *et al.*, 2012) version 1.8.1 and later, see Supplementary Data for instructions. 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 interpreted as a minimal cluster. 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 ω can either be given a fixed value, or estimated by adding a hyperprior and an operator. The trees sampled from the posterior can be analyzed with a tool called SpeciesDelimitationAnalyser. It is similar to TreeAnnotator, but instead of summarizing the clade frequencies from the BEAST output, it summarizes the posterior frequencies of clusterings, and produces a table Ω of clusterings $Z_1, Z_2, \ldots, Z_z$ with corresponding posterior probabilities clusterings $p_1, p_2, \ldots, p_z$ which sum to 1. The clusterings are sorted in order of decreasing posterior probability. An R script for producing a similarity matrix (see Section 3.4) and detailed instructions on how to use SpeciesDelimitationAnalyser are provided in the Supplementary Information.

2.3 Advice on choosing parameters and priors

The parameter ω can be chosen to reflect prior knowledge about the likely number of species. As a consequence of the structure of the model, even when ω is fixed, the prior on the number of species k is somewhat diffuse: it is not possible to insist on exactly seven species for example. In the case of fixed ω, the number of trees with k 'real' tips in the prior has the distribution of $1 + X$, where X is a random variable having the binomial distribution with size parameter $n-1$ and probability parameter $1-\omega$. Its mean is thus $1+(n-1)(1-\omega)$. If the individuals have been assigned in previous work to k_0 species, then $\omega = (n-k_0)/(n-1)$ seems a reasonable choice. If the value of ω is estimated, and a beta prior is used, the prior distribution on $k - 1$ is a beta-binomial distribution, which can be explored using the R package VGAM (Yee [2010] see also Supplementary Information). If a flat prior on the number of species is desired, a Beta distribution with parameters (1,1) will ensure this. If ω is fixed at zero, the value of ε becomes irrelevant, and the model becomes equivalent to the birth–death model as used in *BEAST, except that the origin height is estimated instead of being integrated out analytically.

The parameter ε should be set to a small value such as 1e–4 or 1e–5. The value is a compromise between exactly matching a particular model and the practicalities of computation. Extremely small values may lead to poor mixing, although we have only observed a substantial effect for ε below 1e–6. If ε is too large it will not be possible to distinguish very recent divergences. For most analyses, there will not be enough data to distinguish speciations with node heights below 1e–4, since the expected number of mutations separating the species is only 1 per 5000 sites, so the choice of ε will not be at all critical. Note that these values are based on the premise that the mutation rate is set to 1, or that one is set to 1, and the others are estimated relative to this.

When the number of individuals per species is small, it becomes difficult to estimate the population size parameters in each branch in the multispecies coalescent model. In such a case, care must be taken to use a sensible prior on these parameters, especially the 'species.popMean' parameter. We recommend that the prior should be proper, and diffuse enough to accommodate extreme but possible values, but not absurdly diffuse. This is good advice anyway when using *BEAST, but it becomes more critical when using DISSECT, since it will typically be harder to ensure that the number of individuals per species is not small.

3 Evaluation

3.1 Simulated scenarios and parameter settings

Two sets of simulations were run. The first set evaluates the performance of DISSECT as the number of genes and the amount of incomplete lineage sorting varies, and assesses the sensitivity of the method to choices of ε and ω. The second set focuses on the case of one true species. We use N_e to mean the effective number of (diploid) individuals in a population. If N_e is constant, this means that the expected time for two gene copies to coalesce is $2N_e$ generations. We denote the mutation rate per site per generation by μ. Node heights and ε are in the same units as the product μN_e. Note that the topology and node heights of the gene trees only depend on the product μN_e, so a scenario with $\mu = 1e–8$ and $N_e = 50\,000$ is equivalent to one with $\mu = 1e–9$ and $N_e = 500\,000$ and so on. There are two sources of 'noise' in the data: one comes from coalescences which are deeper than species tree node height and the other from the randomness of mutations. For a node height of 0.001 and a gene length of 500, the expected number of substitutions separating two species is 1, so around 37% of pairs of gene copies from different species would be identical if they coalesced at the species node height.

The first set (SIM-5x5) of simulations all use 25 individuals, 5 assigned to each of 5 species, with one gene copy per individual. The species tree has a comb topology with node heights at 0.001, 0.002, 0.004, and 0.008. These heights are chosen to roughly approximate those in the empirical dataset (see below). The value of $2N_e$ was 69 000 at the tips and at the rootward ends of branches, and 138 600 at the root and tipwards ends of internal branches, varying linearly along the branches. The length of the genes was set to 500 sites, and the number of genes G was set to 3, 9, or 27. The mutation rate μ was set to 2e–9, 1e–8, or 5e–8, representing small, moderate, and large amounts of incomplete lineage sorting. In coalescent units $2\mu N_e$, the height of the most recent speciation is 5, 1, and 0.2, respectively. The root of the species tree is at $0.008/T$ generations, and is therefore 4 000 000 generations when $\mu = 2e–9$, 800 000 generations when $\mu = 1e–8$ and 160 000 generations when $\mu = 5e–8$. To get some idea of the amount of signal and noise due deep to coalescences in the data, consider the $G = 9$ case, where there are 4500 sites. For $\mu = 2e–9$, the number of variable sites is about 100. In the case $\mu = 1e–8$, the number of variables site is around 200, and in the case $\mu = 5e–8$, it is around 500. The increase in variable sites as μ increases is due to deeper coalescences.

We explored the accuracy of the method with respect to changes in ε by using a beta prior for ω with shape parameters 8 and 2, and setting ε to 0.0001 = 1e–4, 3e–5, and 1e–5. We also explored the behavior with respect to different priors for ω by fixing ε to 1e–4, and setting ω to 11/12, 5/6, and 17/24, corresponding to prior means for k of 3, 5, and 8. In addition, two sets of runs performed with a Beta hyperprior on ω with parameters (8,2) and (1,1), respectively. The former distribution has a peak at 4, and the latter means that the probability is uniformly distributed.

The second set SIM-1 of simulations all use $\mu = 1e–8$ and $N_e = 100\,000$, meaning that the simulated genealogies span one coalescent unit. In this case a single species was simulated, so the gene trees are all the result of a coalescence process only. The product μN_e scales the number of substitutions, and thus affects the accuracy with which genes trees can be estimated, but does not change the underlying 'shape' of the problem. The value of ε was 1e–4. A beta prior with shape parameters 8 and 2 was used for ω, which means that the prior was biased toward more than one species. We used $n = 4$, 8, and 16 individuals and G was set to 3, 9, and 27 to examine how these variables affect the rate of false splits.

3.2 Implementation of simulations

The simulated data was generated and analyzed using R (R Development Core Team, 2011) and the R packages APE (Paradis *et al.*, 2004) and phangorn (Schliep, 2011). Gene trees were simulated according to the multispecies coalescent model for each scenario and parameter choice, for 50 replicates. Sequence alignments with 500 sites were generated for these gene trees using Seq-Gen (Rambaut and Grassly, 1997) called with command

```
seqgen.exe -mHKY -t3.0 -f0.3,0.2,0.2,0.3
```

This uses a strict clock and the HKY substitution model, and all genes have the same mutation rate. There is no site rate heterogeneity. These sequences were then incorporated into BEAST XML files, and DISSECT was run for 50 million generations with the first 25 million discarded as burn-in. The priors for `species.popMean`, `meanGrowthRate` and the relative clock rates were all lognormals, with means and standard deviations in log space equal to −7 and 2, 4.6 and 2 and 0 and 1, respectively. The prior for `relativeDeathRate` was uniform in [0,1]. Species DelimitationAnalyser was run after DISSECT with the first 25 million discarded as burn-in.

3.3 Empirical data

Species delimitation in the pocket gopher genus *Thomomys* subgenus *Megascapheus* has been controversial, with a large number of species described by early taxonomists. Most of these have been reduced to subspecific rank by twentieth century taxonomists inspired by the biological species concept (e.g. Wilson and Reeder, 2005). According to opinion of these recent authors, the number of species in the dataset of Belfiore *et al.* (2008), also used by Heled and Drummond (2010), vary between 6 and 8, depending on how the species *T. bottae*, *T. umbrinus* and *T. townsendii* are delimited. We explored the dataset, which consists of 26 individuals which were defined as 'species' and 7 non-coding nuclear sequence regions (Belfiore *et al.*, 2008), by varying ε from 1e–7 to 1e–3, and by setting ω to 0.12 or 0.68 (corresponding to subspecies elevated to species rank, and eight species, as classified in Belfiore *at al.* [2008], respectively). We also used a Beta hyperprior with parameters 4 and 2 (Fig. 3). Each combination of ε and ω was run for 200 million generations, saving parameter values and species trees every 5000th generation. SpeciesDelimationAnalyser was run with τ equal to ε.

The second dataset is an extension of the data used by Aydin et al. (2014) for a group of species in the flowering plant genus *Silene* L. They used marginal likelihood estimates as well as the software BP&P (Yang and Rannala, 2010) to compare species

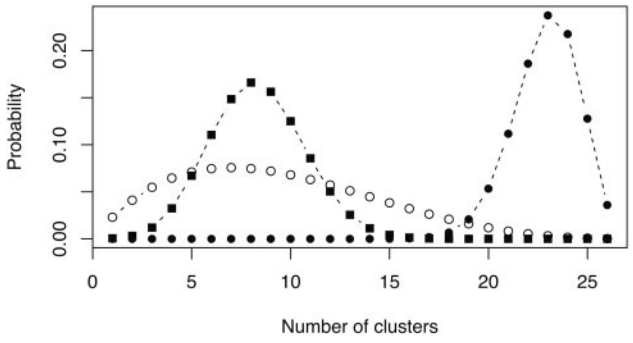

Fig. 3. Prior distribution for the number of clusters when ω is 0.12 (black circles), 0.68 (black squares), and has a Beta distribution with parameters 4 and 2 (open circles)

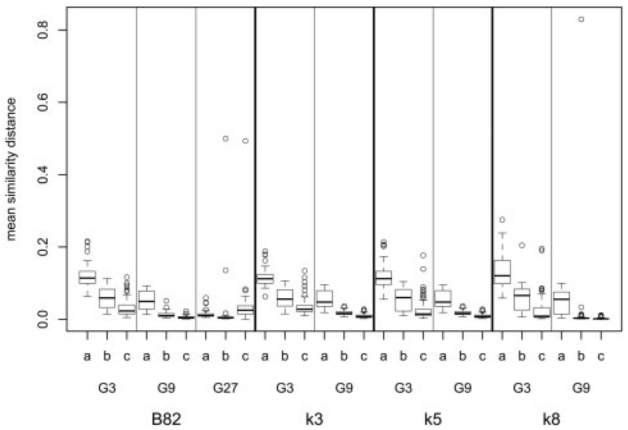

Fig. 4. The boxplots show the values of the error metric over 50 replicates as the number of genes ($G=3$, 9, or 27), the amount of lineage sorting (shortest branches of species tree $a=0.2$, $b=1.0$, $c=5.0$ coalescent units) and prior on ω (B82: Beta $\sim$ (8,2) hyperprior, $k=3$, 5 or 8) vary. Epsilon is 0.0001, and 50 000 000 MCMC generations for each replicate, with first 50% discarded as burn-in

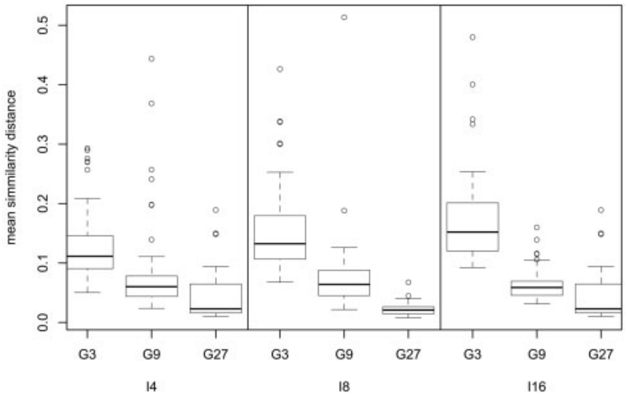

Fig. 5. The boxplots show the values of the error metric over 50 replicates as the number of genes (G), and the number of individuals (I) vary when there is only one species. There is a $\omega \sim$ Beta(8,2) hyperprior. Epsilon is 0.0001, and 50 000 000 MCMC generations for each replicate, with first 50% discarded as burn-in

delimitations in section *Cryptoneurae* Aydin and Oxelman, and concluded that there was strong support for the recognition of the recently described species *S. ertekinii* Aydin and Oxelman, and most probably four species in total. Here, we use data from 20 individuals and sequence data from the same six loci as Aydin et al. (2014) to compare results from a Beta(4,2) prior on ω, to those from a Beta(1,1) prior, which means that all values of ω have uniform prior probability.

3.4 Evaluation metrics

The number of possible clusterings of n individuals (known as the Bell number B_n) increases rapidly with n. For example $B_2=2$, $B_3=5$, $B_4=15$, $B_5=52$, $B_{10}=115\,975$ and $B_{25}\approx 4.6e\ 18$. (See O'Meara, 2010, for more details.) The accuracy of the estimated number of species is not a good way to judge the method, since the number may be correct despite false splits and false merges which cancel out, or despite major mis-assignments, or incorrect due to a single individual being incorrectly merged or separated from a cluster. There are approximately 2.4e 15 ways in which 25 individuals can be grouped into five clusters. The situation is similar to that of inferring phylogenies, where we typically do not expect every clade to be correctly inferred if the number of species is large. In order to assess the accuracy of DISSECT, we therefore want a metric analogous to tree metrics such as the Robinson–Foulds distance.

Rand index. We chose the Rand index (Rand, 1971), which measures the similarity $R(X,Y)$ between two clusterings X and Y of the same set (e.g. the set of individuals). It is convenient to use for accuracy evaluations. The Rand index is always between 0 and 1, and is 1 when the match is perfect. We also define $\overline{R}(X,Y) = 1 - R(X,Y)$ which is a metric in the mathematical sense, and which we will refer to as the Rand metric. Firstly, in order to evaluate the posterior distribution as a whole, we weight the Rand metric between each clustering Z_m in the table Ω produced by DISSECT and the true clustering Z^* by its posterior probability p_m, and thus produce an overall measure of the distance from the posterior distribution to Z^*:

$$D(\Omega, Z^*) = \sum_{m=1}^{z} p_m \overline{R}(Z_m, Z^*).$$

This is our main tool for evaluating DISSECT on simulated data. In the Supplementary Information, we present several other ways to explore the results.

3.5 Results

Results for the first set SIM-5x5 are shown in Figure 4. In general, increasing the number of genes from 3 to 9 increases accuracy, and it is almost perfect for nine genes in the easy cases, where the shortest branches of the species trees are five coalescent units (c in Fig. 4). However, with 27 genes, accuracy goes down at least for the easy case. Checking the trace files and effective sample sizes (ESSs) revealed poor convergence of these runs. The effect of varying the prior for ω shows no obvious effect on accuracy, except possibly a small increase in the posterior probability with increasing k (see Supplementary Data).

Results for the second set SIM-1 are shown in Figure 5. The point estimates were always correct, except in one replicate with nine genes, medium amount of lineage sorting, and eight individuals. There were no false splits with high posterior probabilities. The posterior probability of the correct clustering increases with G, but shows no clear effect with varying number of individuals I. However, when running the same analyses with $\omega \sim$ Beta(N–1,1), meaning that the prior probability of one species is 0.5, accuracy increases with the number of individuals (Supplementary Data).

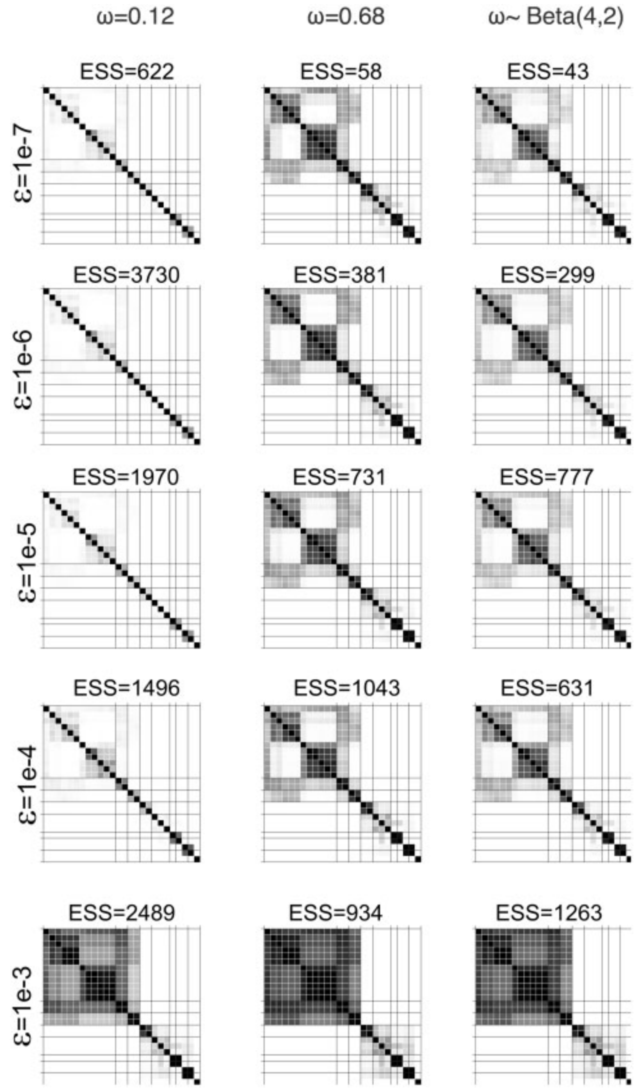

Fig. 6. Similarity matrices for the *Thomomys* dataset under various ε and collapse weight (ω) values. The squares represent posterior probabilities (white = 0, black = 1) for pairs of individuals to belong to the same cluster. The ESS values are ESSs for speciation.likelihood. The right column shows results when a Beta prior distribution with parameters 4 and 2 was used

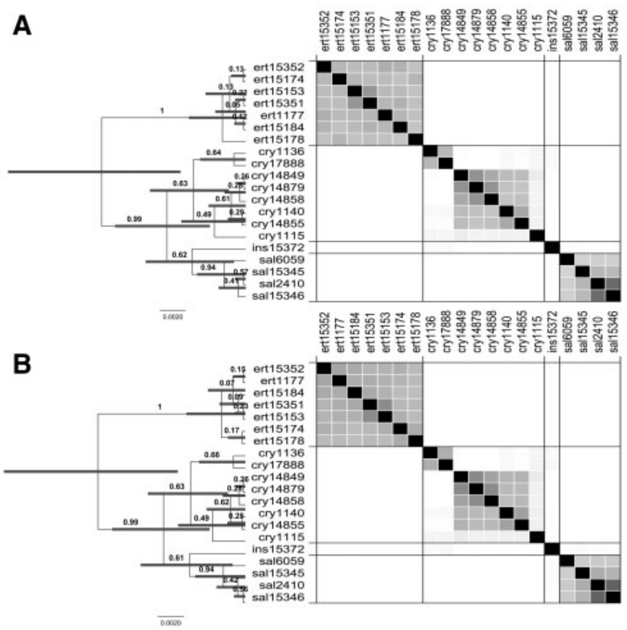

Fig. 7. Species tree and similarity matrices for the *Silene* dataset under ε = 1e–4 and A) ω ∼ Beta(4,2) and B) ω ∼ Beta(1,1). The squares represent posterior probabilities (white = 0, black = 1) for pairs of individuals belonging to the same cluster. The lines in the matrix denote species delimitations as used by Aydin *et al.* (2014). Labels on branches denote posterior probabilities for clades of individuals, bars represent the 95% highest posterior densities for node heights

Varying ε between 1e–4 and 1e–7 on the *Thomomys* data did not have any noticeable effects on the similarity matrices generated from SpeciesDelimitationAnalyzer (Fig. 6). Varying ω had clear effects, with more and smaller clusters for the small ω = 0.12. The posterior mean values for ω when estimated with a Beta(4,2) prior distribution varied between 0.53 and 0.55 when ε was in the range not affecting the posteriors. ESSs for most parameters were well above 300, except for some population size parameters for individual branches, and speciation.likelihood, where the smallest ε values gave low ESSs for ω = 0.68 and ω estimated with Beta(4,2). The results of the analyses of the *Silene* data showed little sensitivity to whether the prior on ω was informative or not (Fig. 7).

4 Discussion

4.1 Simulations

As expected, the accuracy increases with the number of unlinked loci and the ability to detect species increases as the height of the nodes increases. The insensitivity of the method to varying ε suggests that the approximation is unlikely to bias the results.

The results on the scenarios SIM-1 with one true species show the method does not often infer false splits, but it is also clear that a substantial number of sequences are required in order to draw a firm conclusion even in this simplest of cases. A full evaluation of the method on more complex cases is beyond the scope of this article. Note that even with two true species, there is a four-dimensional space of scenarios to explore (node height, effective population size, number of individuals, and number of loci).

In occasional replicates of the simulated data, convergence was poor, something that potentially can affect the accuracy of the method. Therefore, we strongly advice users of the method to carefully review convergence of the MCMC runs.

In general, the number of species was over-estimated in the scenarios used here (results not shown). However, one could add very recent nodes to the scenarios which would tend to be falsely merged and result in an under-estimate instead. It would be interesting to evaluate the method on a large number of scenarios produced by sampling from a birth–death process. For the moment we suggest that estimates of numbers of species are treated with caution.

4.2 Empirical examples

The insensitivity of DISSECT to ε suggested by the results from the simulated data seems corroborated by the *Thomomys* data. In datasets of the size evaluated here, there is far too little information to detect node heights smaller than 0.0001 substitutions per site. On the other hand, the impact of ω on the data was noticeable, indicating that the data are not informative enough to be strongly conclusive about species delimitations.

The ambiguous assignment of several individuals in the *Thomomys* dataset may indicate violations to the assumptions of the model (e.g. no hybridizations), or that the data are not informative enough. To assess absolute fit of the data to the model, posterior predictive simulation-based model checks may give clues to the reasons for this (Reid *et al.*, 2013). Indeed, the *Thomomys* data showed poor fit to the multispecies coalescent model in the survey by Reid *et al.* (2013), and one possible reason to this might be mis-assignment of alleles to species.

The results of our analysis of the data from *Silene* sect. *Cryptoneurae* (Fig. 7) are in agreement with the conclusions of Aydin et al. (2014). The posterior probabilities for clades and MSC membership are almost identical from the analysis with an informative Beta prior (Fig. 7A) to that with an uninfirmative prior (Fig. 7B). However, the support for pairs of individuals belonging to the same species (cluster) is often low. We propose that a possible reason for this pattern, also seen in the *Thomomys* data, might be structured populations due to for example, gradual speciation (see also below). If that is the case, it might be convenient to use a combination of species tree support for clades and the pairwise similarity matrix to aid further considerations about species memberships. High posterior support for clades in the species and congruence among gene trees, but low resolution below that may indicate occasional migration/hybridization. The multispecies coalescent model assumes no migration after speciation, which is instantaneous. This is probably violated in most cases.

Zhang *et al.* (2011) found that low rates (<0.1 migrant per generation) of migration had virtually no effect on the accuracy of BP&P in a simulation study. However, at least when sample size is small, a single sampled recent migrant can cause severe effects. The coalescent prior on the gene trees will affect them in a way that single recent introgressions will be ‘pushed back’ by other gene trees that reflect the ‘true’ speciation event, such that the coalescent time for the migrant may be biased. More research is needed to evaluate the robustness of the model to hybridization, and in particular perhaps, to gradual isolation of species, which may be the most common form of speciation (e.g. Barton and Charlesworth, 1984). Recent studies (Heled *et al.*, 2013; Leaché *et al.*, 2014a) indicate relative robustness of the MSC model to gradual speciation models when it comes to species tree topology inference, but severe effects regarding polation size and divergence time estimates. The effect on MSCSD methods remains to be explored using both simulated and empirical datasets.

4.3 Conclusion

‘Given the intrinsic theoretical and empirical difficulties of the problem, any success would be surprising.’ (O’Meara, 2010). We believe that DISSECT is a useful step forward on the theoretical and computational side. The multispecies coalescent model has assumptions that are likely to be violated and it remains to be seen how important these are for empirical data.

We have not formally evaluated the accuracy of the species trees produced by DISSECT. However, apart from the approximation involving ε, and the slightly different prior on the tree root height, the DISSECT model, when conditioned on a particular clustering Z, is equivalent to *BEAST using Z to assign individuals to species. This means that DISSECT can be used as in a regular *BEAST analysis, taking uncertainties in species delimitation into account. The new version of BP&P, which appeared as advance access (Yang and Rannala, 2014) late in the review process of this article has similar properties, and the accuracy of the two approaches may now be compared.

Acknowledgements

The authors thank four anonymous reviewers and the editors for valuable comments on earlier versions of this article.

Funding

This work was supported by the Swedish Research Council [grant 2012-3719 to B.O.].

Conflict of Interest: none declared.

References

Aydin,Z. *et al.* (2014) Marginal likelihood estimate comparisons to obtain optimal species delimitations in *Silene* sect. *Cryptoneurae* (Caryophyllaceae). *PLoS One*, **9**, e106990.

Baele,G. *et al.* (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.*, **30**, 239–243.

Barton,N.H. and Charlesworth,B. (1984) Genetic revolutions, founder effects and speciation. *Annu. Rev. Ecol. Syst.*, **15**, 133–146.

Belfiore,N.M. *et al.* (2008) Multilocus phylogenetics of a rapid radiation in the genus *Thomomys*. *Syst. Biol.*, **57**, 294–310.

de Queiroz,K. (2007) Species concepts and species delimitation. *Syst. Biol.*, **56**, 879–886.

Dirac,P. (1958) *Principles of Quantum Mechanics*, 4th edn. Clarendon Press, Oxford, UK.

Drummond,A.J. *et al.* (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.

Fujita,M.K. *et al.* (2012) Coalescent-based species delimitation in an integrative taxonomy. *Trends Ecol. Evol.*, **27**, 480–488.

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

Grummer,J.A. *et al.* (2014) Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). *Syst. Biol.*, **63**, 119–133.

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

Heled,J. *et al* (2013) Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evol. Biol.*, **13**, 44.

Knowles,L.L. and Carstens,B.C. (2007) Delimiting species without monophyletic gene trees. *Syst. Biol.*, **56**, 887–895.

Leaché,A.D. *et al.* (2014a) The influence of gene flow on species tree estimation: a simulation study. *Syst. Biol.*, **63**, 17–30.

Leaché,A. *et al.* (2014b) Species delimitation using genome-wide SNP data. *Syst. Biol.*, **63**, 534–542.

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**, e68242.

O’Meara,B.C. (2010) New heuristic methods for joint species delimitation and species tree inference. *Syst. Biol.*, **59**, 59–73.

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

R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

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.

Rannala,B. and Yang,Z. (2013) Improved reversible jump algorithms for Bayesian species delimitation. *Genetics*, **194**, 245–253.

Reid,N.M. *et al*. (2013) Poor fit to the multispecies coalescent is widely detectable in empirical data. *Syst. Biol.*, **63**, 322–333.

Satler,J.D. *et al*. (2013) Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). *Syst. Biol.*, **62**, 805–823.

Schliep,K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

Thompson,E.A. (1975) *Human Evolutionary Trees*. Cambridge, UK: Cambridge University Press.

Wilson,D.E. and Reeder,D.M. (2005) *Mammal Species of the World. A Taxonomic and Geographic Reference*. 3rd edn. Johns Hopkins University Press. Baltimore, Maryland, USA.

Xie,W. *et al*. (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, **60**, 150–160.

Yang,Z. and Rannala,B. (2010) Bayesian species delimitation using multilocus sequence data. *Proc. Natl Acad. Sci. USA*, **107**, 9264–9269.

Yang,Z. and Rannala,B. (2014) Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, **31**, 3125–3135.

Yee,T.W. (2010) The VGAM package for categorical data analysis. *J. Stat. Softw.*, **32**, 1–34.

Zhang,C. *et al*. (2011) Evaluation of a Bayesian coalescent method of species delimitation. *Syst. Biol.*, **60**, 747–761.