Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis

David Bryant,*[,1] Remco Bouckaert,[2] Joseph Felsenstein,[3] Noah A. Rosenberg,[4] and Arindam RoyChoudhury[5]

[1]Department of Mathematics and Statistics and the Allan Wilson Centre for Molecular Ecology and Evolution, University of Otago, Dunedin, New Zealand

[2]Computational Evolution Group, Department of Computer Science, University of Auckland, Auckland, New Zealand

[3]Department of Genome Sciences and Department of Biology, University of Washington

[4]Department of Biology, Stanford University

[5]Department of Biostatistics, Mailman School of Public Health, Columbia University

***Corresponding author:** E-mail: david.bryant@otago.ac.nz.

Associate editor: Rasmus Nielsen

Research article

Abstract

The multispecies coalescent provides an elegant theoretical framework for estimating species trees and species demographics from genetic markers. However, practical applications of the multispecies coalescent model are limited by the need to integrate or sample over all gene trees possible for each genetic marker. Here we describe a polynomial-time algorithm that computes the likelihood of a species tree directly from the markers under a finite-sites model of mutation effectively integrating over all possible gene trees. The method applies to independent (unlinked) biallelic markers such as well-spaced single nucleotide polymorphisms, and we have implemented it in SNAPP, a Markov chain Monte Carlo sampler for inferring species trees, divergence dates, and population sizes. We report results from simulation experiments and from an analysis of 1997 amplified fragment length polymorphism loci in 69 individuals sampled from six species of *Ourisia* (New Zealand native foxglove).

Key words: multispecies coalescent, species trees, SNP, AFLP, effective population size, SNAPP.

Introduction

Biallelic markers such as single nucleotide polymorphisms (SNPs) and amplified fragment length polymorphisms (AFLPs) are potentially rich sources of information about species radiations, species divergences, and historical demographics. However, extracting this information is not always straightforward. Patterns of genetic variation at these markers are not just a product of the relationships between the species; they also reflect inheritance patterns within each species. Any full-likelihood (or full-Bayesian) method for inferring species histories from genetic markers needs to model the random distribution of gene tree histories for each marker. To date, this task has often meant implementing massive Monte Carlo simulation-based sampling of both species trees and the gene trees at every locus (Rannala and Yang 2003; Wilson et al. 2003; Hey and Nielsen 2007; Liu and Pearl 2007; Heled and Drummond 2010).

In this paper, we describe an algorithm that allows us to bypass the gene trees and compute species tree likelihoods directly from the markers. The likelihood values, or posterior probabilities, computed by the algorithm are identical to those that would be obtained by sampling every possible gene tree topology and every possible set of gene tree branch lengths at each locus. The algorithm makes use of new formulae for lineage and allele probabilities under the coalescent and employs recently developed numerical techniques (Sidje 1998; Schmelzer and Trefethen 2007) to evaluate these formulae.

Our approach makes the following assumptions of the data:

(A1) Each marker is a single biallelic character (e.g., a biallelic SNP or AFLP banding pattern);

(A2) The genealogies for separate markers are conditionally independent given the species tree. In practice, this assumption applies to unlinked markers or linked markers that have so little linkage that they do not possess a discernible excess of linkage disequilibrium.

This latter assumption is clearly not valid for sites in a single-gene sequence. However, it is satisfied for SNPs that are well spaced along the genome. If the independence assumption (A2) is only partially violated, the effect of linkage could be investigated by subsampling sets of markers with varying degrees of independence. Even when there is linkage between sites, treating the markers as independent often still provides statistically responsible inferences (Gutenkunst et al. 2009; RoyChoudhury 2011).

In principle, the full-likelihood methods of Liu and Pearl (2007) and Heled and Drummond (2010), which are designed primarily for linked sequence data, could be applied to data satisfying assumptions (A1) and (A2) by encoding each marker as a separate locus. This strategy would quickly become computationally infeasible as the number of markers increased. Nielsen et al. (1998) demonstrated that a full-likelihood approach is tractable

for data satisfying (A1) and (A2), presenting an algorithm that uses biallelic characters directly to compute the likelihood of the species tree, though their method was computationally feasible only for small species trees. RoyChoudhury (2006) and RoyChoudhury et al. (2008) made a substantial advance on the computational problem. They took the approach of Nielsen et al. (1998) and placed it within a dynamic programming framework, thereby giving an efficient algorithm for computing the likelihood of a tree with an arbitrary number of species.

The methods developed by Nielsen et al. (1998) and RoyChoudhury et al. (2008) both make a significant and mathematically convenient assumption about mutation. Under their models, mutation can only occur within the population at the root of the species tree. It cannot occur within the populations represented by the branches of the species tree. This assumption is reasonable when comparing closely related populations for which recent mutations are sufficiently rare that they can be ignored. It is less appropriate when analyzing rapidly mutating markers or when comparing more distantly related populations or species.

Here, we extend the dynamic programming structure of RoyChoudhury et al. (2008) to allow mutations within the populations represented by the species tree. In many ways, this is a more parsimonious model: The mutation model in the root population is the same as the mutation model used for the populations along the branches. We address the algorithmic, mathematical, and computational challenges resulting from this deceptively minor change in model assumptions.

Our algorithm implements a "finite-sites" model for mutation. Nielsen (1998) derived a recursion for computing the likelihood of a tree under the "infinite-sites" model for mutation. In general, the recursive formula has too many terms to be evaluated directly, so a Markov chain Monte Carlo (MCMC) method was used instead. However, if the data satisfy (A1) and (A2), then the recursion described by Nielsen (1998) can be evaluated efficiently using algorithms similar to those described here.

One issue that arises when modeling mutation is that some of parameters might not be identifiable from data. Under the infinite-sites model and the "no-branch-mutation" model of Nielsen et al. (1998) and RoyChoudhury et al. (2008), the length of a branch in the species tree and the corresponding population size are confounded: Doubling the effective population size has the same effect on the likelihood as halving the branch length. A similar issue arises when inferring species trees from gene tree topologies without branch lengths (Degnan and Salter 2005; Wu 2011). We show that fully including mutations in the finite-sites model permit the identification of both branch lengths (times) and population sizes (θ), at least in situations where sufficiently many mutations have occurred throughout the species tree.

The coalescent process is often viewed as a dual process to the Wright–Fisher diffusion (Donnelly and Kurtz 1996), and each has some practical advantages over the other. Diffusion-based approaches for analyzing SNP data from multiple populations have been proposed by Gutenkunst et al. (2009) and Siren et al. (2010). The main difference is that coalescent-based methods such as ours' model the history of the ancestral lineages, whereas diffusion-based approaches model variation in the continuous allele frequencies. In practice, it is not clear which approach is preferable for a given data set: both require some level of approximation and each has advantages and disadvantages computationally.

We have implemented the new finite-sites model likelihood algorithm and incorporated it within a Bayesian MCMC sampler, which we call SNAPP ("SNP and AFLP Phylogenies"). SNAPP, which interfaces with the BEAST package (Drummond and Rambaut 2007), takes a range of biallelic data types as input and returns a sample of species trees with (relative) divergence times and population sizes. We have tested and validated the algorithm and software using a range of techniques, and we report results of several experiments with simulated data. The software is open source and is available for download from http://snapp.otago.ac.nz.

To illustrate the application of SNAPP, we analyze AFLP loci in 69 individuals sampled from 6 species of New Zealand *Ourisia* or native foxglove. The New Zealand *Ourisia* form a relatively recent species radiation and inference of branching patterns between these species has proven difficult (Meudt et al. 2009). Meudt et al. propose that the difficulties are due in part to "incomplete lineage sorting," which occurs when the coalescence of lineages within species predates the divergence of different species. Our Bayesian analysis, which models lineage sorting explicitly, provides a relatively clear picture of ancestral species relations in the group and, up to a scale constant, effective population sizes.

Materials and Methods

The Multispecies Coalescent

Our models are all based on the assumption that the lineage dynamics within populations (or species) are well described by the conventional Wright–Fisher model. The distribution of the gene trees within each population is approximated by the "coalescent process" (reviewed in Felsenstein 2004; Hein et al. 2005; Wakeley 2009). This process models the number of ancestral lineages of the sample from a single population as a Markov process that goes backward in time. Initially, the number of ancestral lineages equals the size of the sample. Going backward in time (upward in a branch), lineages meet at common ancestors, and the number of ancestral lineages decreases.

It is customary in coalescent theory to rescale time in terms of effective population size, so that two lineages coalesce at rate 1. This rescaling is not generally possible in the multispecies coalescent since different species can have different effective population sizes. Instead, we adopt the standard practice from phylogenetics and rescale time in terms of expected mutations (as in Rannala and Yang 2003). Hence, the expected time to a coalescence for two lineages is $\theta/2$ and the expected time to a coalescence for k lineages is $\theta/[k(k-1)]$, where θ denotes the expected number of

mutations separating two randomly chosen individuals in the population.

At the first coalescent event, two lineages are selected at random and combined, and we are left with $k - 1$ lineages. This coalescence of lineages continues until the top of the branch is reached, at which anywhere from 1 to k lineages could be present.

The nodes in the species tree represent species divergences or population splits. The individuals in each of the child populations are descendants of individuals in the parent population. In terms of the coalescent process, the lineages coming upward from the child population become lineages at the base of the parent population. This process continues upward in the species tree until the species tree root is reached. At that point, any remaining lineages coalesce according to the standard single-population coalescent model.

See Felsenstein (2004), Degnan and Rosenberg (2009), and Heled and Drummond (2010) for general introductions to the multispecies coalescent. Early contributions to the development of multispecies models built on the branches of a species tree were made by Hudson (1983), Tajima (1983), Takahata and Nei (1985), Nei (1987), Pamilo and Nei (1988), and Takahata (1989).

The multispecies coalescent determines a distribution for gene trees and their branch lengths, conditional on a species tree. The parameters of the distribution are the shape of the species tree, the divergence times within the species tree, and the population sizes along the branches of the species tree (one parameter for each branch). We bundle these parameters into the single composite parameter S, so that the probability of a gene tree G given the species tree is $P(G|S)$. We treat this quantity as a density rather than a discrete probability because of the continuous branch lengths of G.

Let X denote the alignment of sequences for a locus. Conventional phylogenetic models (e.g., Felsenstein 2004) give us the probability that X evolved along a specified gene tree G. These models provide the distribution of states at the root and the mutation probabilities down the edges of the tree. Accordingly, they determine $P(X|G)$, the probability of the data (alignment) given the gene tree. Note that once the gene tree is chosen, the species tree has no further influence on the probability of the data.

Putting $P(G|S)$ and $P(X|G)$ together, we obtain the "joint" probability (or density) of the alignment X and the gene tree G:

$$P(X, G|S) = P(X|G)P(G|S). \tag{1}$$

The gene tree G is not observed directly and it can be difficult to estimate. Since our focus is on the species tree and the features of the species tree, we work with the "marginal" probability of the data. Let Ψ denote the set of all possible genealogies for the individuals incorporating both the topologies and branch lengths. The marginal probability for the data is then found by integrating over Ψ:

$$P(X|S) = \int_{\Psi} P(X|G)P(G|S)\mathrm{d}G. \tag{2}$$

Equation (2) is sometimes called the "Felsenstein equation" (Felsenstein 1988; Rosenberg and Nordborg 2002; Hey and Nielsen 2007).

Generally, we consider multiple genetic markers. We assume that the gene trees for separate markers are independent (given the species tree). Let X_i be the alignment for the ith gene and let G_i be a corresponding gene tree. Under the independence assumption, the total probability of the m alignments at m genes is a product over all the genes:

$$\begin{aligned} P(X_1, X_2, \ldots, X_m|S) &= \prod_{i=1}^{m} P(X_i|S) \\ &= \prod_{i=1}^{m} \int_{\Psi} P(X_i|G_i)P(G_i|S)\mathrm{d}G_i. \end{aligned} \tag{3}$$

If we were to plug this formula into a Bayesian analysis, we would specify a prior distribution $P(S)$ on the species trees and then sample from the posterior distribution

$$P(S|X_1, \ldots, X_m) \propto \left(\prod_{i=1}^{m} \int_{\Psi} P(X_i|G_i)P(G_i|S)\mathrm{d}G_i \right) P(S). \tag{4}$$

Sampling from $P(S|X_1, \ldots, X_m)$ is equivalent to sampling from the joint posterior distribution

$$P(S, G_1, \ldots, G_m|X_1, \ldots, X_m) \propto \left(\prod_{i=1}^{m} P(X_i|G_i)P(G_i|S) \right) P(S) \tag{5}$$

and only considering the marginal distribution of the species trees S. This is the approach taken by BATWING (Wilson et al. 2003), BEST (Liu and Pearl 2007), and STAR-BEAST (Heled and Drummond 2010), among others. Note that if the actual gene trees G_i are provided or if they can be inferred with high accuracy, they can be treated as data and the species tree can be inferred directly (Degnan and Salter 2005; Kubatko et al. 2009).

At this point, it is appropriate to reflect on what exactly is required when applying equations (3) or (4) to large numbers of unlinked biallelic markers. To evaluate the likelihood exactly, we would need to sum (or integrate) over all possible gene trees of all loci. In a Bayesian setting, we would need to sample over a space containing not only every possible choice of species tree but also every possible choice of gene tree for every locus. Furthermore, the marginal probabilities for the gene trees depend not only on the data but also on the species tree, and so the analyses for the separate genes are all interdependent. An analysis of 1,000 independent loci then amounts to 1,001 interlinked Bayesian analyses (1,000 gene trees and one species tree). Even with modern Monte Carlo algorithms, this scale of this analysis is computationally daunting.

Overview of the Likelihood Algorithm

We circumvent these computational difficulties by calculating the integral in equation (3) analytically. In the following sections, we describe a pruning algorithm that we use to compute the likelihood of a species tree given genotype data

at unlinked biallelic markers. The algorithm works in a similar manner to Felsenstein's pruning algorithm (Felsenstein 1981) for computing the likelihood of a gene tree: we define partial likelihoods that focus only on a specific subtree; the partial likelihoods are then computed starting at the leaves (of the species tree), working upward to the root.

There are two major differences. In Felsenstein's pruning algorithm, one partial likelihood is defined for every node and every state (i.e., amino acid or nucleotide). In our algorithm, we have separate partial likelihoods for the top and bottom of each branch in the species tree, for every possible number of ancestral lineages at each point, and for every possible count of the number among these lineages carrying each allele.

Second, we need to deal with the complication that the coalescent process works backward in time (and is not reversible), whereas the mutation process works forward in time. We were not able to define a simple transition process taking numbers of ancestral lineages to numbers of descendant lineages. Instead, we first compute probability distributions for the numbers of ancestral lineages at each node in the species tree. We then define partial likelihoods for subtrees in the species tree and derive the equations required to compute them efficiently. Finally, we show how to handle the probabilities at the root of the species tree when computing the full probability of the genotype data of a marker.

We orient trees so that the ancestral nodes are at the top and time travels downward. Thus, the base of a branch in the species tree corresponds to the population at the time nearest to the present, whereas the top of a branch corresponds to the population just after it has diverged from its ancestral population. In a similar fashion, the genotypic state in a gene tree evolves from the top of the gene tree (the common ancestor) downward to the leaves.

Red and Green Alleles

The multispecies coalescent model for the evolution of markers (SNPs, AFLPs etc.) has two components: the model for the gene trees in the species tree and the model for the markers evolving down the gene tree (i.e., forward in time). The model for gene trees uses a coalescent process that works backward in time, whereas the mutation model for genetic markers (SNPs, AFLPs, etc.) typically works forward in time.

Given a gene tree with branch lengths specified, we model the evolution of a genetic marker using standard phylogenetic machinery. Suppose that there are two alleles, which for ease of illustration we label "red" and "green." Let u be the rate of mutation from the red allele to the green allele per unit time (forward in time), and let v be the corresponding rate of mutating from green to red. We say that a lineage is a red lineage if it has the red allele and a green lineage otherwise.

The allele of the most recent common ancestor at the root of the gene tree is red with stationary probability $v/(u+v)$ and green with probability $u/(u+v)$. The marker evolves down the gene tree as a continuous-time Markov chain whose instantaneous rate matrix has rate u of mutating from red to green and rate v for mutating from green to red. The alleles at the leaves of the gene tree are then the observed alleles. The probability of the allele frequencies at a marker, given the species tree, is therefore the probability of the site given a gene tree multiplied by the probability of the gene tree given the species tree, summed over all possible gene tree topologies and integrated over all possible gene tree branch lengths (eq. [3]).

Ancestral Lineage Counts and the Likelihood

The multispecies coalescent can be used to generate a random gene tree conditional on a species tree. If we take any node or point in the species tree, we can count the number of lineages in the gene tree in that species at that point in time. We say that at a specified time point, this quantity is the number of "ancestral lineages." The count of ancestral lineages is a random variable with distribution determined by the multispecies coalescent process and its resulting distribution of gene trees. The first step in our likelihood algorithm is the calculation of these lineage count distributions. See RoyChoudhury et al. (2008) and Efromovich and Kubatko (2008) for similar computations.

Let x be a branch (i.e., ancestral species) in the species tree. Let $\mathbf{n}_x^{\mathrm{B}}$ denote the number of gene tree lineages at the base of the branch x. Let $\mathbf{n}_x^{\mathrm{T}}$ denote the number of ancestral lineages at the top of the branch and let t be the length of the branch, measured in units of expected number of mutations (see fig. 1). The minimum possible value for $\mathbf{n}_x^{\mathrm{B}}$ and $\mathbf{n}_x^{\mathrm{T}}$ is 1, whereas the maximum possible value is the total number of individuals sampled in populations at or below x, a quantity that we denote by m_x. The distribution of $\mathbf{n}_x^{\mathrm{T}}$ given $\mathbf{n}_x^{\mathrm{B}}$ is given by the probability in the standard coalescent model of going from n ancestors to k ancestors over time t (measured in units of expected mutations):

$$\Pr[\mathbf{n}_x^{\mathrm{T}}=k|\mathbf{n}_x^{\mathrm{B}}=n] = \sum_{r=k}^{n} e^{\frac{-r(r-1)t}{\theta}} \frac{(2r-1)(-1)^{r-k}k_{(r-1)}n_{[r]}}{k!(r-k)!n_{(r)}}, \quad (6)$$

where $n_{[r]}=n(n-1)(n-2),\ldots,(n-r+1)$ and $n_{(r)}=n(n+1),\ldots,(n+r-1)$ (Tavaré 1984).

When x is an "external" branch (adjacent to a leaf) in the species tree, $\mathbf{n}_x^{\mathrm{B}}$ equals the number of samples from the species corresponding to that branch. Let n_x denote this number of samples. Then

$$\Pr[\mathbf{n}_x^{\mathrm{B}}=n] = \begin{cases} 1 & \text{if } n=n_x, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Suppose that x is an internal or external branch in the species tree. Let m_x denote the maximum possible value for $\mathbf{n}_x^{\mathrm{B}}$ or $\mathbf{n}_x^{\mathrm{T}}$, equal to the total number of sampled individuals summed across populations at or below x in the species tree. Suppose that $\Pr[\mathbf{n}_x^{\mathrm{B}}=k]$ has been computed for all k from 1 to m_x. The distribution of $\mathbf{n}_x^{\mathrm{T}}$ is determined by the value of $\mathbf{n}_x^{\mathrm{B}}$ using the conditional probabilities in equation (6):

$$\Pr[\mathbf{n}_x^{\mathrm{T}}=n] = \sum_{k=n}^{m_x} \Pr[\mathbf{n}_x^{\mathrm{B}}=k]\Pr[\mathbf{n}_x^{\mathrm{T}}=n|\mathbf{n}_x^{\mathrm{B}}=k]. \quad (8)$$