Journal of Theoretical Biology 289 (2011) 96–106

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Determining species tree topologies from clade probabilities under the coalescent

Elizabeth S. Allman [a], James H. Degnan [b], John A. Rhodes [a,*]

[a] Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, AK 99775, USA
[b] Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 24 February 2011
Received in revised form
8 June 2011
Accepted 2 August 2011
Available online 19 August 2011

Keywords:
Coalescent
Statistical consistency
Identifiability
Consensus
Phylogenetics

ABSTRACT

One approach to estimating a species tree from a collection of gene trees is to first estimate probabilities of clades from the gene trees, and then to construct the species tree from the estimated clade probabilities. While a greedy consensus algorithm, which consecutively accepts the most probable clades compatible with previously accepted clades, can be used for this second stage, this method is known to be statistically inconsistent under the multispecies coalescent model. This raises the question of whether it is theoretically possible to reconstruct the species tree from known probabilities of clades on gene trees.

We investigate clade probabilities arising from the multispecies coalescent model, with an eye toward identifying features of the species tree. Clades on gene trees with probability greater than 1/3 are shown to reflect clades on the species tree, while those with smaller probabilities may not. Linear invariants of clade probabilities are studied both computationally and theoretically, with certain linear invariants giving insight into the clade structure of the species tree. For species trees with generic edge lengths, these invariants can be used to identify the species tree topology. These theoretical results both confirm that clade probabilities contain full information on the species tree topology and suggest future directions of study for developing statistically consistent inference methods from clade frequencies on gene trees.

1. Introduction

A fundamental problem in evolutionary biology is to determine relative relatedness of species, usually by seeking a rooted tree that diagrammatically depicts these relationships. Although phylogenetic methods of inferring relationships between genes sampled from individuals in the different species are now highly developed, such gene trees are not species trees. Even in the absence of errors due to estimating gene trees from DNA sequences, gene tree topologies need not match the underlying species tree. In recent years, various methods have been proposed for inferring species trees from genetic data (Degnan and Rosenberg, 2009; Edwards, 2009; Knowles and Kubatko, 2010). Many of these methods first estimate gene trees, and then resolve the possible conflicts among them to obtain an overall estimate of the species tree.

An important cause of gene tree conflict is the population effect of *incomplete lineage sorting*, in which gene lineages coalesce in ancestral populations earlier than the time these lineages first enter a common ancestral population. The *multispecies coalescent model* (Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005; Degnan and Rosenberg, 2009) is commonly used to model this process, producing a distribution of rooted gene trees given a rooted species tree topology and branch lengths (a measure of time and population size on each edge of the species tree). The multispecies coalescent provides a natural framework for incorporating population effects, allowing gene trees to possibly be discordant with the species tree (see Fig. 1), a phenomenon that is very common in multilocus studies (Rokas et al., 2003; Ebersberger et al., 2007; Cranston et al., 2009).

Although the distribution of gene tree topologies from the multispecies coalescent determines the species tree (Allman et al., 2011), estimating this distribution is difficult because there are so many possible topologies: $(2n-3)!!$ when n species are under study. Thus most topologies are unlikely to be observed among a moderate number of gene trees. An alternative is to estimate a smaller set of probabilities which is a function of gene tree probabilities but that still retains enough information to identify the species tree. Other works have considered rooted triples (Degnan et al., 2009; Ewing et al., 2008; Liu et al., 2010) and unrooted gene tree topologies (Allman et al., 2011; Larget et al., 2010). Another possibility, which is

[*] Corresponding author.
E-mail address: j.rhodes@alaska.edu (J.A. Rhodes).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.08.006

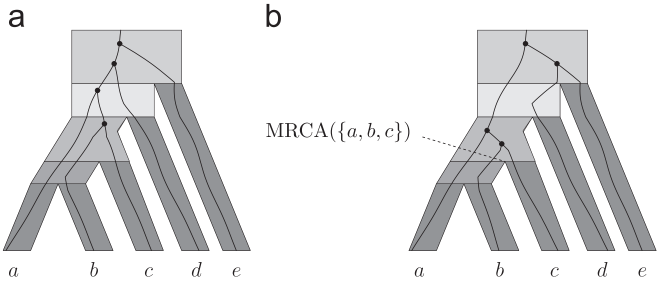

Fig. 1. Gene trees within a species tree. In the multispecies coalescent, gene lineages sampled from species are assumed to coalesce (form nodes in the gene tree) no more recently than their most recent common ancestor (MRCA) in the species tree. Coalescence of lineages in populations more ancient than their MRCA can lead to gene tree topologies that are discordant with the species tree topology. Using upper case letters for gene lineages sampled from their corresponding species, failure of the *A* and *B* lineages to coalesce in their MRCA population makes any of the $\binom{3}{2}$ coalescences between *A*, *B*, and *C* equally likely under the model in the MRCA population of *a*, *b*, and *c*. (a) The gene tree is ((((*B*,*C*),*A*),*D*),*E*). (b) The gene tree is (((*B*,*C*),*A*), (*D*,*E*)).

our focus here, is to use probabilities that a gene tree has a given *clade*, a set of leaves descended from a node of the gene tree that is not ancestral to any other leaves in the gene tree. The probability of a clade under the multispecies coalescent (or any model of gene tree generation) is obtained by simply adding the probabilities of all gene trees that display the given clade (Degnan et al., 2009).

The probability of a clade can be estimated from a collection of gene trees by considering the proportion of gene trees displaying the clade. Since this procedure does not take into account uncertainty in the gene trees, which are themselves estimates from genetic data, a more sophisticated method would quantify the uncertainty in the clades by using posterior probabilities or bootstrap support values for clades obtained from Bayesian or maximum likelihood analyses of the gene trees. The software BUCKy (Ané et al., 2007), for example, takes this approach, using posterior probabilities for clades and additionally incorporating a prior distribution for the amount of gene tree conflict to yield a *concordance factor* for each clade.

One of the most straightforward methods for constructing a species tree from clade probabilities is to use *greedy consensus*, in which the clade with the highest probability (or concordance factor) is accepted, provided it is compatible with previously accepted clades. This process is repeated until a fully resolved tree is formed (Bryant, 2003). This procedure is implemented in BUCKy to construct a *concordance tree*, which is sometimes interpreted as an estimated species tree (Cranston et al., 2009).

To justify a greedy approach, one needs to investigate whether the most probable clades tend also to be clades on the species tree. Indeed, we show in Section 4 that under the multispecies coalescent, any clade with probability greater than 1/3 must be on the species tree, suggesting that the standard majority-rule consensus (which only accepts clades occurring more than 50% of the time) is very conservative in this setting. If the greedy consensus approach is used for clades with probability greater than 1/3 (leaving the tree unresolved with respect to clades with lower probability), then this "not-too-greedy" consensus approach is not misleading, in the sense that it asymptotically cannot return a false species tree clade as the number of loci approaches infinity.

In contrast, previous results have shown that when greedy consensus is applied without restrictions on clade probabilities, the returned tree can be misleading (*i.e.*, for some species trees, as the number of loci increases, the greedy consensus method is increasingly likely to produce a tree that disagrees with the true species tree) for some sets of branch lengths (Degnan et al., 2009). These "too-greedy zones" of edge lengths occur on 4-taxon asymmetric species trees and on any species tree topology with five or more leaves. Thus, caution must be used when probabilities of clades are less than 1/3; it is not obvious how to determine which low-probability clades are on the species tree, even if clade probabilities are known exactly. Other examples show that the most probable k-clade (a clade of $k \geq 2$ elements), is not necessarily a clade on the species tree, even if the species tree is known to have a k-clade.

Undeterred by these negative results, we show in Sections 5 and 6 that under the multispecies coalescent with one lineage sampled per species, the set of clade probabilities does identify the species tree topology for generic branch lengths for any number of species. The proof is based on discovering a linear combination of clade probabilities (a linear invariant) that is equal to zero for any branch lengths on any species tree with a given clade. In theory, if clade probabilities are known, it is therefore possible to identify the species tree by determining all of its clades.

Finally, in Section 6 we extend our results, in part, to cases where the species tree is non-binary and where an arbitrary number of lineages is sampled per species.

Although we frame our questions within the framework of the multispecies coalescent, a careful reading of our arguments reveals that the essential feature of the model that we use is that lineages are *exchangeable*. If two gene lineages are present in the same population at a particular point in time on the species tree, then above that point, the model assumes that both lineages behave the same way. Much of this work, then, should be robust to variations on the coalescent model that preserve exchangeability. Though we do not pursue this here, one could, for instance, consider versions of the multispecies coalescent model in which more than two lineages coalesce simultaneously, as in the Λ-coalescent (Eldon and Wakeley, 2006; Pitman, 1999).

While one might be tempted to use the vanishing of clade invariants for direct inference of clades on a species tree, doing so would require overcoming several obstacles. First, evaluating these invariants on empirical clade probabilities from previously inferred gene trees will rarely yield zero exactly, due to both sampling and gene tree inference errors. Thus it would be necessary to understand the variance of these polynomial values, in order to formulate an appropriate way of determining when values are sufficiently close to zero to indicate a likely clade. Second, the clade invariants we present are not all the constraints on clade probabilities arising from a given species tree. Our clade invariants are all linear equalities, and higher degree equalities can be shown to exist computationally. Moreover, one should expect the existence of non-trivial inequality constraints as well. Ignoring these additional constraints is likely to degrade performance of any such method.

Thus while our linear clade invariants suggest a statistically consistent method of identifying a species tree, how they would perform in practice is unclear. It remains a challenge to incorporate the insight they provide into a practical method that outperforms greedy consensus on most finite data sets. Nonetheless, our results demonstrate that sound statistical inference from clade probabilities is possible.

On a more technical note, there is a key difference in understanding clade probabilities versus many other sets of probabilities related to gene trees or species trees: the failure of marginalization arguments. As this difference plays an important, but unspoken, part throughout this work, we highlight it here.

The problem of establishing identifiability of a species tree from unrooted gene tree probabilities that was taken up previously (Allman et al., 2011) is superficially similar to the clade problem of this paper. Both unrooted gene tree probabilities and clade probabilities can be obtained by summing probabilities of