Molecular Phylogenetics and Evolution 70 (2014) 63–69

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Short Communication

Gene tree rooting methods give distributions that mimic the coalescent process

Yuan Tian, Laura S. Kubatko *

Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, United States

ARTICLE INFO

Article history:
Received 13 March 2013
Revised 28 August 2013
Accepted 6 September 2013
Available online 18 September 2013

Keywords:
Coalescent model
Molecular clock rooting
Outgroup rooting
Gene tree distribution

ABSTRACT

Multi-locus phylogenetic inference is commonly carried out via models that incorporate the coalescent process to model the possibility that incomplete lineage sorting leads to incongruence between gene trees and the species tree. An interesting question that arises in this context is whether data "fit" the coalescent model. Previous work (Rosenfeld et al., 2012) has suggested that rooting of gene trees may account for variation in empirical data that has been previously attributed to the coalescent process. We examine this possibility using simulated data. We show that, in the case of four taxa, the distribution of gene trees observed from rooting estimated gene trees with either the molecular clock or with outgroup rooting can be closely matched by the distribution predicted by the coalescent model with specific choices of species tree branch lengths. We apply commonly-used coalescent-based methods of species tree inference to assess their performance in these situations.

1. Introduction

In multi-locus phylogenetic analysis, a phylogenetic tree based on a single gene (e.g., a gene tree) may not agree with the species tree, the tree that represents the actual evolutionary pathway (Pamilo and Nei, 1988; Maddison, 1997). The many possible causes of this discord are well-known, and include processes such as incomplete lineage sorting (ILS), gene duplication, hybridization, non-neutral evolution, and horizontal gene transfer (Hein, 1993; Maddison, 1997; Sang and Zhong, 2000; Bayzid and Warnow, 2012; Kubatko, 2009). An appropriate probabilistic model that links gene trees and species trees should involve the phylogenetic relationships of species as well as the genealogical history of each gene (Anderson et al., 2012). Such models are necessary to carry out accurate inference of the species phylogeny from a multi-locus data set.

The coalescent process is a retrospective model of population genetics that is commonly used to model ILS. The coalescent model is based on tracing the evolutionary history of sampled genes by considering the time from the present back to their most recent common ancestor (Kingman, 2000), and can be derived as the limiting distribution (as the population size becomes large) that results from the Wright-Fisher and other commonly-used population genetics models (Wakeley, 2008; Kingman, 1982; Takahata and Nei, 1985). Under the coalescent model, the probability distribution of gene trees given a fixed species tree topology and branch lengths can be computed (Degnan and Salter, 2005). The coalescent model is also used as the basis for different methods to estimate species trees using either multi-locus DNA sequence data or a set of observed gene trees (Kubatko et al., 2009; Than et al., 2008; Liu and Pearl, 2007; Liu et al., 2010; Heled and Drummond, 2010). These methods have been widely applied to the multi-locus data sets that are commonly produced by next-generation sequencing techniques.

The coalescent process has been so widely applied in part because it is believed that ILS is a predominant cause of the incongruence observed between gene trees and species trees (Liu et al., 2010). Indeed, the predictions made by the coalescent model in terms of the distribution of gene trees are consistent with several observed data sets (Ebersberger et al., 2007; Ané, 2010; Kubatko et al., 2011). However, a recent study of seven large multi-partition genome-level data sets has suggested that random rooting might be another potential explanation for the apparent fit to the coalescent model (Rosenfeld et al., 2012). Here, we consider whether the signature evidenced by this gene tree topology distribution is unique to the coalescent process. In particular, we show that in the case of four taxa, this distribution can be mimicked by generating gene trees from a single species tree and then rooting these gene trees using either the assumption of a molecular clock or outgroup rooting.

Below we briefly review the coalescent process, focusing on the gene tree distribution in the four-taxon case. We then use simulation to show that both the coalescent and non-coalescent models described above lead to nearly identical distributions on the set

* Corresponding author. Address: Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, United States. Fax: +1 614 292 2866.
E-mail addresses: tian.52@osu.edu (Y. Tian), kubatko.2@osu.edu (L.S. Kubatko).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.09.004

of gene trees. Further, we study the behavior of several coalescent-based methods of species tree estimation when the data come from a single tree. We conclude that a model without ILS can produce a distribution that mimics that of the coalescent process. Furthermore, under different gene tree distributions not all of the methods of species tree inference examined here can correctly estimate the true species tree with high frequency.

1.1. Gene tree distribution under coalescent process

Consider a 4-taxon, asymmetric bifurcating species tree as shown in Fig. 1(a) (thick lines). A gene tree that has the same coalescent history as the species tree is nested within the species tree (Fig. 1(a), thin lines). Note that this species tree contains two internal branch lengths that we denote by x and y that are given in coalescent units (number of $2N_e$ generations). In the gene tree, the coalescent times are denoted by t_1, t_2, and t_3, from the most recent to the oldest, respectively. The probability of observing gene tree g given a particular species tree S can be calculated using the formula (Degnan and Salter, 2005):

$$P\{G = g|S\} = \sum_{histories} P\{G = g, history|S\} = \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b),$$

where the product on the right-hand side is taken over all branches, b, w_b is the probability of getting a sequence of coalescent events that is consistent with g, and $P_{u(b),v(b)}(t_b)$ is the probability that $u(b)$ lineages coalescent into $v(b)$ lineages along branch b which has length t_b (see, e.g., Rosenberg, 2007). This formula can be used to obtain expressions for the probabilities of all 15 gene tree topologies that are possible for the species tree in Fig. 1(a) when one lineage per species is sampled. Fig. 1(b)–(e) shows the entire probability distribution on the set of 15 gene trees for several choices of the species tree branch lengths x and y. Note that when x and y are long, only one tree has substantial probability (Fig. 1(b)). As y becomes shorter, there are three asymmetric trees that have substantial probability (Fig. 1(c)). If y is long but x becomes shorter, the distribution of mass shifts to two asymmetric trees as well as a symmetric tree (Fig. 1(d)). When both x and y are very short, many of the 15 trees have nearly equal probability (Fig. 1(e)).

The fact that under the coalescent model for four taxa one tree will have most of the probability while two others occur with smaller, but equal, probability (and all other trees have much lower probability) has been noted to be a signature of the coalescent model. This type of distribution has been observed in empirical studies as well (Ebersberger et al., 2007; Ané, 2010; Kubatko et al., 2011; Rosenfeld et al., 2012), and for this reason has been used as evidence of the importance of incorporating the coalescent model into species tree inference methodology. It is thus of interest to determine whether other features of the process of molecular evolution coupled with tree estimation methodology can produce a distribution on gene trees that shows this feature. This is the question that we examine here.

2. Methods

2.1. Sequence simulation and gene tree estimation

Five 4-taxon, asymmetric bifurcating gene trees with topology (((AB)C)D) and different branch length parameters t_1, t_2, and t_3 as depicted in Fig. 2(a) were chosen as "true gene trees" in our simulation study. All five of the true gene trees used satisfied the molecular clock assumption. Multiple sequence alignments of 500 base pairs (bp) were generated with the program Seq-Gen (Rambaut and Grassly, 1997) under the HKY85 substitution model (Hasegawa, Kishino and Yano, 1985) for each of the following five input true gene trees:

(((A:0.2,B:0.2):0.15,C:0.35):0.4,D:0.75),
(((A:0.25,B:0.25):0.015,C:0.265):0.15,D:0.415),
(((A:0.05,B:0.05):0.0005,C:0.0505):0.15,D:0.2005),
(((A:0.15,B:0.15):0.025,C:0.175):0.015,D:0.19), and
(((A:0.2,B:0.2):0.0075,C:0.2075):0.0075,D:0.215).

This procedure was repeated 10,000 times for each tree, to generate 10,000 alignments of length 500 for each model tree.

From these alignments, gene trees were then estimated under the molecular clock assumption and rooted using maximum likelihood (ML) as implemented in the program PAUP* 4.0b10 (Swofford, 2002). PAUP* was also used to obtain ML gene tree estimates without the molecular clock assumption, and the resulting trees were rooted by the outgroup rooting method with taxon D specified as the outgroup. In both the molecular clock rooting and outgroup rooting methods, the frequencies of the resulting rooted gene tree topology estimates were recorded. These were used both to examine the induced gene tree distribution and to assess the performance of several current coalescent-based methods of species trees estimation.

2.2. Species trees estimation

The simulated data sets were used to estimate species trees under the coalescent model. Each data set was randomly divided into 100 groups, with 100 genes in each group. Each group could be considered as a multi-locus data set of 100 gene trees for inference of species trees. Three coalescent-based software packages were used to estimate species trees: STEM (Kubatko et al., 2009), MP-EST (Liu et al., 2010), and the minimize deep coalescences method (MDC), as implemented in PhyloNet (MDC-PhyloNet; Than and Nakhleh, 2009). For each of these three methods, the input data were the trees estimated for each simulated gene alignment under the ML method with the assumption of a molecular clock. For MP-EST and MDC-PhyloNet, the gene trees estimated using outgroup rooting were also used for analysis (note that this is not possible for STEM, because branch lengths that satisfy the molecular clock are required). For each method, we recorded the proportion of the 100 data sets (each consisting of 100 genes) that correctly recovered the tree that generated the data.

3. Results

3.1. Estimation of the gene tree distribution under different rooting methods

Fig. 2(b)–(f) shows the proportion of times each gene tree (red bars) was estimated in the five data sets that are simulated with different choices of the true gene tree branch lengths t_1, t_2, and t_3. We see that estimating ML gene trees under the molecular clock assumption can lead to an estimated gene tree distribution that mimics that which would be expected under the coalescent model. For example, Fig. 2(b) also shows the expected distribution under the coalescent model for the species tree in Fig. 1(a) (tree in thick lines) with $x = 10.0$ and $y = 10.0$ (green[1] bars). Note the agreement between this expected distribution and the observed distribution.

The proportion of times that each gene is estimated when using outgroup rooting is also shown in Fig. 2(b)–(f) (blue bars), along with the predicted gene tree distribution under the coalescent model for the species tree in Fig. 1(a) with a different choice of species tree branch lengths x and y (purple bars). As was the case with

[1] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.