Molecular Phylogenetics and Evolution 106 (2017) 144–150

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Short Communication

Expected pairwise congruence among gene trees under the coalescent model

Yuan Tian [a], Laura S. Kubatko [a,b,*]

[a] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, United States
[b] Department of Statistics, The Ohio State University, United States

ARTICLE INFO

Article history:
Received 25 April 2016
Revised 7 September 2016
Accepted 23 September 2016

Keywords:
Coalescent
Incomplete lineage sorting
Gene tree
Species tree
Incongruence
RF distance
Mapping traits
Comparative methods

ABSTRACT

Although it is widely appreciated that gene trees may differ from the overall species tree and from one another due to various evolutionary processes (e.g., incomplete lineage sorting (ILS), horizontal gene transfer, etc.), the extent of this incongruence is rarely quantified and discussed. Here we consider the expected amount of incongruence arising from ILS, as modeled by the coalescent process. In particular, we compute the probability that two gene trees randomly sampled from the same species tree agree with one another as well as the distribution of the Robinson-Foulds distance between them, for species trees with three to eight taxa. We demonstrate that, as expected under the coalescent model, the amount of discordance is affected by species tree-specific factors such as speciation times and effective population sizes for the species under consideration. Our results highlight the fact that substantial discordance may occur, even when the number of species is very small, which has implications both for larger taxon samples and for any method that uses estimated gene trees as the basis for further statistical inference. The amount of incongruence is substantial enough that such methods may need to be modified to account for variability in the underlying gene trees.

1. Introduction

Incomplete lineage sorting (ILS) has long been recognized to be a predominant cause of substantial variation in the evolutionary trees for individual genes, leading to incongruence among gene trees (Pamilo and Nei, 1988; Takahata, 1989; Hein, 1993; Maddison, 1997; Sang and Zhong, 2001; Kubatko, 2009; Liu et al., 2010; Bayzid and Warnow, 2012). Coalescent theory (Kingman, 1982a,b; Tajima, 1983; Tavaré, 1984; Takahata and Nei, 1985; Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005) is commonly used to model ILS, and the predictions concerning agreement among gene trees made by the coalescent model are becoming increasingly well-understood. For example, it is widely appreciated that the extent of gene tree incongruence depends on characteristics of the species tree, such as branch lengths and effective population sizes, suggesting the importance of considering the coalescent process in phylogenetic studies (Degnan and Salter, 2005; Degnan and Rosenberg, 2006). Indeed, most current methods of multi-locus phylogenetic inference incorporate the coalescent process to model ILS (Edwards et al., 2016). In this study, we consider the extent of congruence between a pair of gene trees that is expected under the coalescent process for a small number of taxa. Our study is motivated by four current challenges in empirical multilocus phylogenetics.

First, we consider the problem of assessing the extent to which empirical data fit the predictions made by the coalescent model. In particular, a recent study (Simmons et al., 2016) examined eight empirical data sets and presented the pairwise congruence among gene trees for each of these studies. Simmons et al. (2016) reported that the average pairwise congruence among gene trees varied greatly both between studies and sometimes within a study. However, the study did not consider the extent of pairwise congruence that would be *expected* among gene trees under the coalescent model, making it difficult to assess fit to the coalescent model. Here we consider this problem from an analytic perspective by computing the probability of sampling two identical gene tree topologies, as well as the probability distribution of the pairwise Robinson-Foulds distance (hereafter "RF distance", Robinson and Foulds, 1979) among possible gene trees, for model species trees with three to eight taxa.

Second, note that a well-known example of incongruent gene trees is the commonly observed conflict between mitochondrial and nuclear phylogenies (Ferris et al., 1983; Moore, 1995; Sota

* Corresponding author at: The Ohio State University, Department of Statistics, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, United States.
E-mail address: kubatko.2@osu.edu (L.S. Kubatko).

http://dx.doi.org/10.1016/j.ympev.2016.09.023
1055-7903/

and Vogler, 2001; Shaw, 2002; McCracken and Sorenson, 2005; Leaché and McGuire, 2006; Robertson et al., 2006; Peters et al., 2007; Good et al., 2008). Conflicting mitochondrial and nuclear gene trees are often attributed to the low mutation rate of nuclear DNA sequences, hybridization or other horizontal gene transfer, and natural selection (Avise et al., 1987; Rand, 2001; Sanderson and Shaffer, 2002; Funk and Omland, 2003; Ballard and Whitlock, 2004; Ballard and Rand, 2005; Spinks and Shaffer, 2009; Roos et al., 2011), while the effect of ILS is often overlooked. Incongruence among gene trees is also extensively observed in other phylogenetic settings, such as conflicting plastid and nuclear gene trees, or different gene trees among nuclear genes (Cranston et al., 2009; Moyer et al., 2009; Bell and Hyvönen, 2010). Our computation of the expected extent of pairwise gene tree incongruence for specific examples serves to highlight the important role that ILS is likely playing in these empirical observations more generally.

Third, we consider gene tree incongruence as a potential source of bias when mapping character traits onto phylogenetic trees. In this setting, it is common to use a single tree (normally an estimated species tree) as the phylogenetic framework onto which character traits (e.g., nucleotides, morphological traits, behavioral traits, etc.) are mapped. For example, Hahn and Nakhleh (2015) recently discussed the risk of ignoring variation in gene tree topologies and mapping characters onto a single representation of the species tree. Quantifying the extent to which the two underlying phylogenies may vary using the RF distance for specific examples, as we have done here, gives insights into the extent of the potential bias in mapping character traits onto a fixed phylogeny.

Fourth, we note that a similar issue may exist in phylogenetic comparative studies. A number of comparative studies look for correlation in two or more traits after adjusting for a single species tree. Often this single tree is a well-resolved species tree that is believed to be a reliable estimate of the species-level evolutionary relationships (Wiegmann et al., 2009; Misof et al., 2014). It is clear, however, that each of the traits under consideration may have its own gene tree, and these gene trees may vary substantially from the species tree. Although the uncertainty in phylogenetic relationships has been taken into account in several approaches (Richman and Price, 1992; Huelsenbeck et al., 2000; Lutzoni et al., 2001; Pagel and Lutzoni, 2002; Huelsenbeck and Rannala, 2003; Pagel et al., 2004), most of them focus on a lack of resolution in the overall species tree estimate (Hahn and Nakhleh, 2015) or gene tree estimation error (Simmons et al., 2016), rather than on genuine variation in the underlying histories. Again, quantification of the extent of incongruence can give insights into how robust comparative procedures might be to gene tree variation.

We next briefly describe our computational methods for assessing the extent of gene tree incongruence under the coalescent model, and then provide our results. While we consider only small trees here (in the range of three to eight taxa), these examples serve to highlight the fact that gene tree incongruence is widespread, even when the number of taxa under consideration is not large. We discuss the implications of our findings in the Discussion.

2. Methods

In our study, the probability that two topologically identical gene trees are observed from the same species tree under the coalescent model is computed assuming no recombination within the two loci, and free recombination between the two loci. We further assume that the only evolutionary process generating discord between the two gene trees is the coalescent process (i.e., there is no hybridization or other horizontal gene transfer, and no gene duplication/loss). We consider the case in which one gene lineage is sampled for each species – when additional samples within a species are sampled, the situation will be even more complex in the sense that there is the possibility of even greater discord between gene trees.

Varying population sizes and branch lengths of the species tree are examined to study their effect on congruence among gene trees. Furthermore, the probability distribution of pairwise RF distances between the two gene trees under the coalescent model is computed. Note that the expected gene tree congruence and the distribution of RF distances are computed based on the study of Degnan and Salter (2005). In their paper, they derived a method for computing the distribution of gene tree topologies given a bifurcating species tree with an arbitrary number of taxa, when one gene is sampled for each species. This method was implemented in the computer program COAL. When a given species tree (with branch lengths in coalescent units) is input into COAL, the probability of each possible gene tree topology will be calculated. All calculations presented here are carried out using COAL (Degnan and Salter, 2005).

2.1. Gene tree incongruence for species trees with varying numbers of taxa

Completely asymmetric species trees and selected symmetric species trees with up to eight taxa are used to compute the probability of observing two topologically identical gene trees. We select three different levels for the internal branch lengths (time between one speciation event to the following speciation event) for each species tree. Species trees with all internal branch lengths equal to 2.0 (all branch lengths in coalescent units) are labeled "Long species trees", species trees with all internal branch lengths equal to 1.0 are labeled "Medium species trees", and species trees with all internal branch lengths equal to 0.5 are labeled "Short species trees". Under the coalescent model, it is expected that gene trees sampled from species trees with "long" internal branch lengths will have little ILS, while species trees with "short" internal branch lengths are expected to generate gene trees with extensive ILS. It is impossible to explore the space of all species trees exhaustively, even for small numbers of taxa, because there are infinitely many choices for internal branch lengths. Our choice of "long", "medium" and "short" settings here is meant to span the range of possibilities. The sum of squares of all possible gene tree probabilities gives the total probability of sampling two topologically identical gene trees for a given species tree.

2.2. Gene tree incongruence probability for species trees with varying internal branch lengths

To examine the effect of different branch lengths on the extent of gene tree incongruence, we use species trees of four taxa and five taxa. For the four-taxon case, the species tree is denoted as $(((A:T_0,B:T_0):T_1,C:(T_0+T_1)):T_2,D:(T_0+T_1+T_2))$, where T_1 and T_2 are the two internal branch lengths. Note that the value of T_0 does not affect the probability of gene trees that are embedded in the species tree, because only one individual is sampled in each species. In our calculations, we use 100 equally spaced values from 0.0 to 5.0 for T_1 and T_2. For the five taxa case, the species tree is denoted as $((((A:T_0,B:T_0):T_1,C:(T_0+T_1)):T_2,D:(T_0+T_1+T_2)):T_3,E:(T_0+T_1+T_2+T_3))$. Three levels (0.1, 1.0, and 5.0) are selected for T_3, and 100 equally spaced values from 0.0 to 5.0 are used for T_1 and T_2, similar to the four-taxon case. The probability of getting two topologically identical gene trees for each choice of branch lengths is calculated as described in the last section.