Journal of Theoretical Biology 374 (2015) 35–47

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites

Julia Chifman [a], Laura Kubatko [b,c,*]

[a] Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157, United States
[b] Department of Statistics, The Ohio State University, Columbus, OH 43210, United States
[c] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, United States

HIGHLIGHTS

- Phylogenetic species tree estimation is considered for analyzing next-gen data.
- The coalescent model is used as a model for single-gene phylogenies.
- The general time-reversible (GTR) substitution model is used for sequence evolution.
- Models allowing site-specific rate variation and invariable sites are considered.
- Identifiability of the n-taxon unrooted species tree topology is proven.

ARTICLE INFO

Article history:
Received 11 July 2014
Received in revised form
24 January 2015
Accepted 5 March 2015
Available online 17 March 2015

Keywords:
Phylogenetics
Identifiability
Algebraic statistics

ABSTRACT

The inference of the evolutionary history of a collection of organisms is a problem of fundamental importance in evolutionary biology. The abundance of DNA sequence data arising from genome sequencing projects has led to significant challenges in the inference of these phylogenetic relationships. Among these challenges is the inference of the evolutionary history of a collection of species based on sequence information from several distinct genes sampled throughout the genome. It is widely accepted that each individual gene has its own phylogeny, which may not agree with the *species tree*. Many possible causes of this gene tree incongruence are known. The best studied is the incomplete lineage sorting, which is commonly modeled by the coalescent process. Numerous methods based on the coalescent process have been proposed for the estimation of the phylogenetic species tree given DNA sequence data. However, use of these methods assumes that the phylogenetic species tree can be identified from DNA sequence data at the leaves of the tree, although this has not been formally established. *We prove that the unrooted topology of the n-leaf phylogenetic species tree is generically identifiable given observed data at the leaves of the tree that are assumed to have arisen from the coalescent process under a time-reversible substitution process with the possibility of site-specific rate variation modeled by the discrete gamma distribution and a proportion of invariable sites.*

Published by Elsevier Ltd.

1. Introduction

The field of evolutionary genetics has benefitted enormously from recent advances in sequencing technology that have led to the availability of DNA sequence information for hundreds or thousands of species. These data are commonly used to study evolutionary patterns and processes. A fundamental problem in this area is the inference of a *phylogenetic species tree* that describes the evolutionary relationships among a collection of species for which data have been collected. A species phylogeny is a tree with all internal nodes of degree three, except for a root node of degree two. All edges are treated as directed from the root toward the leaves. The tree represents the biological process of speciation, in which one population splits into two populations which then evolve independently, with no subsequent exchange of genetic material. The root node represents the most ancestral population to all sampled species, while the leaves (also called

* Corresponding author at: Department of Statistics, The Ohio State University, Columbus, OH 43210, United States.
E-mail addresses: jchifman@wakehealth.edu (J. Chifman), lkubatko@stat.osu.edu (L. Kubatko).

http://dx.doi.org/10.1016/j.jtbi.2015.03.006
0022-5193/Published by Elsevier Ltd.

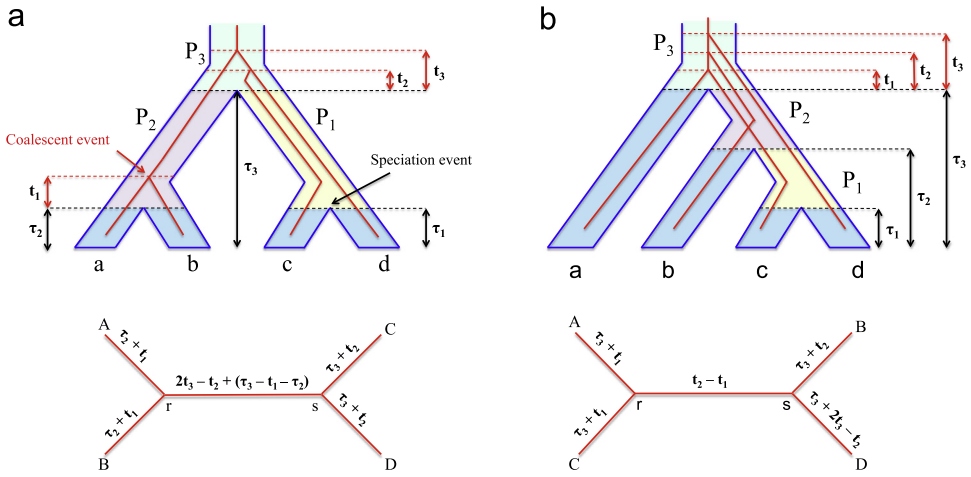

Fig. 1. Example of two gene trees (red) nested within a species tree (blue). In (a) the gene tree and the species tree have the same topology, while in (b) they do not agree with one another. In both subfigures, the dotted black horizontal lines represent speciation events in the species tree. The times of these events, denoted by τ_i, are measured backward from the present time. Portions of the tree that fall between species events represent ancestral populations, denoted by P_i. The dotted horizontal red lines represent coalescent events that occur in the gene trees. The times of coalescent events, denoted by t_i, are measured from the most recent (looking forward in time) speciation event.

taxa; singular: *taxon*) represent present-day populations. An example of a phylogenetic species tree for four species, *a*, *b*, *c*, and *d*, is shown by the outlined tree in Fig. 1.

Although the goal is generally to estimate the species phylogeny from available DNA sequence data, these sequence data are only directly informative about the *gene tree* – the phylogenetic tree underlying the gene for which the DNA sequences are available. It is well accepted that gene trees and species trees may not agree with one another (see, e.g., Maddison, 1997; Pamilo and Nei, 1988), with many evolutionary processes known to give rise to variability in gene phylogenies within a fixed species phylogeny. Examples of such processes are incomplete lineage sorting (ILS), hybridization, horizontal gene transfer, and gene duplication and loss (Maddison, 1997). The best studied of these processes is the ILS, which results when two lineages fail to share a most recent common ancestor (MRCA; represented by an internal node in the gene tree) until further back in time than the immediate ancestral population. For example, in Fig. 1(a), the gene tree embedded within the species tree represents the phylogenetic history of the lineages sampled from species *a*, *b*, *c*, and *d*, which are denoted by *A*, *B*, *C*, and *D*, respectively. Throughout the text, we use uppercase letters to refer to gene lineages, and the corresponding lowercase letters to refer to the species from which these lineages are sampled. Although it is possible for lineages *C* and *D* to share their most recent common ancestor in the population labeled P_1, they remain distinct in this population, and instead share their MRCA in population P_3, thus providing an example of ILS. Note that the topology (branching pattern) of the gene tree in Fig. 1(a) matches the topology of the species tree. Fig. 1(b) gives another example of ILS, but in this case lineage *A* coalesces with lineage *C* in their ancestral population, and the gene tree topology does not match the species tree topology.

One of the reasons that ILS has been well studied is that it can be modeled by the coalescent process. The coalescent process can be derived as the large sample limit (as the population size goes to ∞) of the Wright–Fisher and other common population genetics models (Kingman, 1982a,b; Tavaré, 1984). The key property of the coalescent model is that the waiting time back into the past for a pair of lineages to find their MRCA follows an exponential distribution, with a parameter that depends on the sample size. The coalescent model thus provides a link between the phylogenetic species tree and the set of gene trees embedded within the species tree that give rise to the actual data. For this reason, numerous methods based on the coalescent process have recently been proposed for estimation of the phylogenetic species tree. One group of methods (e.g., BEST, Liu and Pearl, 2007; ⁎BEAST, Heled and Drummond, 2010; STEM, Kubatko et al., 2009; and MP-EST, Liu et al., 2010) assumes that multi-locus data are available for inference, with the assumption that each locus has a single underlying (unobserved) gene tree. Alternatively, single nucleotide polymorphism (SNP) data are sometimes used for inference. SNP data represent sites sampled throughout the genome that are known to be variable, with the assumptions that the sites are unlinked and that they each have their own phylogenetic history. The software package SNAPP (Bryant et al., 2012) has recently been developed for species tree estimation from biallelic SNP data. Use of any of these methods assumes that the phylogenetic species tree can be identified from DNA sequence data at the leaves of the tree, but this has not formally been established (note, however, that Allman et al., 2011b have established identifiability given a collection of gene tree topologies; Allman et al., 2011c have considered identifiability given clade probabilities; and Liu and Edwards, 2009 have established identifiability when the order of ancestral populations, and hence the relationships among all rooted triples, can be consistently estimated).

Here, we prove that the unrooted topology of the phylogenetic species tree is identifiable given observed SNP data at the leaves of the tree that are assumed to have arisen from the coalescent process. Our results hold for data for which a single observation corresponds to recording which of κ possible states occurs at each leaf. These data are modeled by a continuous-time Markov process that specifies the rates of transitions between states along the phylogeny and that satisfies the condition of time-reversibility. We also consider models that allow rate variation across sites as modeled by the discrete gamma distribution, as well as the possibility of invariable sites. For the special case of DNA sequence data, there are four states (i.e., $\kappa=4$) corresponding to the four nucleotides *A*, *C*, *G*, and *T*. In this case, our results hold for the general time reversible (GTR; Tavare, 1986) model with discrete-gamma distributed rate variation and a proportion of invariable sites, and all associated sub-models.

In the next section, we give the necessary background on the coalescent process and on the process of mutation for general κ-state models, pointing out the application to DNA sequence data where relevant. We also examine some common modifications to site-independent sequence substitution models to allow for