Theoretical Population Biology 110 (2016) 12–24

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure

Michael DeGiorgio[a,*], Noah A. Rosenberg[b]

[a] Department of Biology, Pennsylvania State University, 502 Wartik Laboratory, University Park, PA 16802, USA

[b] Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 28 January 2014
Available online 13 April 2016

Keywords:
Coalescent
Consensus methods
Incomplete lineage sorting
Phylogenetics
Statistical consistency

ABSTRACT

In the last few years, several statistically consistent consensus methods for species tree inference have been devised that are robust to the gene tree discordance caused by incomplete lineage sorting in unstructured ancestral populations. One source of gene tree discordance that has only recently been identified as a potential obstacle for phylogenetic inference is ancestral population structure. In this article, we describe a general model of ancestral population structure, and by relying on a single carefully constructed example scenario, we show that the consensus methods Democratic Vote, STEAC, STAR, R* Consensus, Rooted Triple Consensus, Minimize Deep Coalescences, and Majority-Rule Consensus are statistically inconsistent under the model. We find that among the consensus methods evaluated, the only method that is statistically consistent in the presence of ancestral population structure is GLASS/Maximum Tree. We use simulations to evaluate the behavior of the various consensus methods in a model with ancestral population structure, showing that as the number of gene trees increases, estimates on the basis of GLASS/Maximum Tree approach the true species tree topology irrespective of the level of population structure, whereas estimates based on the remaining methods only approach the true species tree topology if the level of structure is low. However, through simulations using species trees both with and without ancestral population structure, we show that GLASS/Maximum Tree performs unusually poorly on gene trees inferred from alignments with little information. This practical limitation of GLASS/Maximum Tree together with the inconsistency of other methods prompts the need for both further testing of additional existing methods and development of novel methods under conditions that incorporate ancestral population structure.

1. Introduction

Recently, much attention has been given to the development of methods that consistently infer the correct species tree from the discordant gene trees produced under incomplete lineage sorting—the failure of lineages from two different species to coalesce in the population immediately ancestral to the divergence of the two species (Degnan and Rosenberg, 2009). Consensus approaches, each of which takes a set of gene trees as input and returns a species tree estimate according to a specific rule (Bryant, 2003), have provided one important source of methods for species tree inference in this context.

A consensus method $\widehat{C}$ is a statistically consistent estimator of a species tree topology under some model if for each species tree σ, $\widehat{C}$ applied to a set of gene trees randomly generated under the model, assuming that the species tree is σ, converges in probability to the topology of σ as the number of gene trees approaches ∞. Statistical consistency is a desirable property because it is reasonable to expect that as more data are gathered, evidence should accumulate in support of the true value of the parameter being estimated.

Degnan and Rosenberg (2006) showed that when gene trees are distributed according to the multispecies coalescent model for the evolution of gene lineages conditional on a species tree, an extreme case of incomplete lineage sorting can arise in which the most likely gene tree topology does not match the species tree topology. This inconsistency implies that species tree estimation methods must use information other than the most frequently occurring gene tree topology in order to accurately infer the

[*] Corresponding author.
E-mail address: mxd60@psu.edu (M. DeGiorgio).

http://dx.doi.org/10.1016/j.tpb.2016.02.002
0040-5809/© 2016 Elsevier Inc. All rights reserved.

Table 1
Notation.

Notation	Definition
$\mathbf{D}$	$(n-1)$-dimensional vector of the numbers of demes in the $n-1$ ancestral populations
$\mathbf{N}$	$(n-1)$-dimensional vector with vector-valued elements for the deme sizes in each of the $n-1$ ancestral populations
$\mathbf{M}$	$(n-1)$-dimensional vector with matrix-valued elements for the backward migration matrices in each of the $n-1$ ancestral populations
$\boldsymbol{\Psi}$	Matrix that describes how demes connect across species boundaries
$\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$	Ancestral population structure model with parameters σ, $\mathbf{D}$, $\mathbf{N}$, $\mathbf{M}$, and $\boldsymbol{\Psi}$
$\mathbb{P}[E\,;\ \mathcal{S}]$	Probability of event E under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$
λ_{A}	Subtree of species tree σ that contains species A and that descends from the divergence of species A and B
λ_{B}	Subtree of species tree σ that contains species B and that descends from the divergence of species A and B
λ_{C}	Subtree of species tree σ that contains species C and that descends from the divergence of species (AB) and C
$\Gamma_{\mathrm{A}}, \Gamma_{\mathrm{B}}, \Gamma_{\mathrm{C}}$	Sets of taxa at the leaves of subtrees λ_{A}, λ_{B}, and λ_{C}, respectively
$\mathcal{L}$	Set of taxa
$\mathcal{T}\vert\mathcal{L}$	Tree displayed by phylogenetic tree $\mathcal{T}$ restricted to the set of taxa $\mathcal{L}$
$\mathrm{top}(\mathcal{T})$	Topology of phylogenetic tree $\mathcal{T}$
$p_{\mathcal{S}}(\mathrm{X}, \mathrm{Y})$	Probability that a lineage sampled from species X and a lineage sampled from species Y are in the same deme at the speciation time of X and Y under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$
$P_{\mathcal{S}}[\mathcal{T}]$	Probability of gene tree topology $\mathcal{T}$ under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$
$\widehat{P}[\mathcal{T}]$	Sample proportion of topology $\mathcal{T}$ in a set of gene trees
T_{XY}^{ℓ}	Random coalescence time at locus ℓ for a lineage sampled from species X and a lineage sampled from species Y
$\mathbb{E}_{\mathcal{S}}[T_{XY}^{\ell}]$	Expected coalescence time under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ for a lineage sampled from species X and a lineage sampled from species Y at locus ℓ
$\overline{T}_{\mathrm{XY}}$	Mean coalescence time across all sampled gene trees between one lineage sampled from species X and one lineage sampled from species Y
R_{XY}^{ℓ}	Rank of the coalescent event at locus ℓ for a lineage sampled from species X and a lineage sampled from species Y
$\mathbb{E}_{\mathcal{S}}[R_{\mathrm{XY}}^{\ell}]$	Expected rank under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ of the coalescent event for a lineage sampled from species X and a lineage sampled from species Y at locus ℓ
$\overline{R}_{\mathrm{XY}}$	Mean rank of coalescent events across all sampled gene trees between one lineage sampled from species X and one lineage sampled from species Y
$t_{\mathrm{XY}}^{\min}$	Minimum coalescence time across all sampled gene trees between one lineage sampled from species X and one lineage sampled from species Y
$\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T})$	Number of extra lineages contributed by the topology of fixed species tree σ for a fixed gene tree topology $\mathcal{T}$
$\mathrm{xl}(\mathrm{top}(\sigma))$	Number of extra lineages contributed by the topology of fixed species tree σ for a fixed set of gene trees

species tree topology. Indeed, many consensus methods relying on other principles provide statistically consistent estimators of the species tree topology under the multispecies coalescent model. This collection of methods includes STEAC (Liu et al., 2009), STAR (Liu et al., 2009), R* Consensus (Degnan et al., 2009), GLASS (Mossel and Roch, 2010), and Maximum Tree (Liu et al., 2010), as well as extensions of some of these methods that preserve the consistency property (Helmkamp et al., 2012; Jewett and Rosenberg, 2012; Allman et al., 2013).

In its simplest form, the multispecies coalescent model assumes that each modern species and each ancestral species have a constant population size, each pair of lineages within a given ancestral species has an equal chance of coalescing, and each species is an unstructured population. Because the multispecies coalescent assumes that random mating occurs within species, when ancestral species are structured, as has been argued for various species (e.g., Garrigan et al., 2005; Thalmann et al., 2007; White et al., 2009), it is unclear whether methods that are consistent under the multispecies coalescent continue to be consistent.

The difficulty of species tree estimation in the presence of ancestral population structure lies in the way that population structure alters the probability distribution of gene trees given a species tree compared to the unstructured case. Using a three-taxon example, Slatkin and Pollack (2008) showed that with ancestral population structure, the probability distribution of gene tree topologies can have a certain asymmetry, and the most likely three-taxon gene tree topology need not match the species tree topology. These consequences of the multispecies coalescent with ancestral structure do not occur in the standard multispecies coalescent.

Here, we describe an extension of the ancestral population structure model considered by Slatkin and Pollack (2008). Using our extended model, we evaluate the consistency of several consensus methods, employing a single example scenario to show that many methods are inconsistent. We show that each of the inconsistent methods is in fact "misleading" in the sense that for a certain fixed species tree σ and a particular set of parameters, the probability that the consensus tree contains a clade not present on σ approaches 1 as the number of loci approaches ∞. To evaluate the speed at which methods converge to or diverge from the correct bifurcating species tree topology, we perform simulations of our model. As predicted by our theoretical results, the only method that does not strongly support incorrect species tree topologies is GLASS/Maximum Tree. However, in accord with past simulations using model species trees (Liu et al., 2009; Leaché and Rannala, 2011; Wu, 2012; DeGiorgio and Degnan, 2014), we show that GLASS/Maximum Tree performs poorly when an absence of substitutions causes little information to exist in sequence alignments. We conclude with a discussion of the implications of the results for understanding evolutionary relationships.

2. Model

We use the notation in Table 1. Suppose time is measured in generations, and that generation time is constant throughout the tree. Consider an ultrametric n-taxon bifurcating species tree σ with $n \geq 3$ taxa (i.e., each leaf has an identical sum of branch lengths to the root). Then we can always find a set of species A, B, and C on σ with relationship $((\mathrm{A}{:}\tau_3, \mathrm{B}{:}\tau_3){:}\tau_2 - \tau_3, \mathrm{C}{:}\tau_2)$, where $\tau_2 > \tau_3 > 0$.

Each internal branch along the species tree specifies an ancestral population. An n-taxon species tree contains $n-1$ such populations, including the branch above the root. Label these populations of σ by recursively visiting the root, then the left subtree, and finally the right subtree (a pre-order traversal of σ). Each ancestral population is allowed to be structured; the population structure model is identical across L independent loci, so that each of L gene trees is a random variate conditional on the same species tree.

In ancestral population i, let $D^{(i)}$ be the number of demes, let $\mathbf{N}^{(i)}$ be the vector of population sizes for the $D^{(i)}$ demes, and let $\mathbf{M}^{(i)}$ be the backward migration matrix between demes (Fig. 1). Denote the ancestral population structure model by $\mathcal{S} = \mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, where $\mathbf{D} = [D^{(1)}, D^{(2)}, \ldots, D^{(n-1)}]$, $\mathbf{N} = [\mathbf{N}^{(1)}, \mathbf{N}^{(2)}, \ldots, \mathbf{N}^{(n-1)}]$, $\mathbf{M} = [\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(n-1)}]$, and $\boldsymbol{\Psi}$ is an $(n + \sum_{i=1}^{n-1} D^{(i)}) \times$