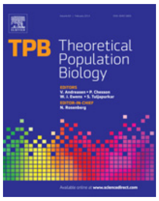
Theoretical Population Biology 100 (2015) 56–62

Contents lists available at ScienceDirect

Theoretical Population Biology

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



Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent

Sebastien Roch[a], Mike Steel[b,*]

[a] Department of Mathematics, University of Wisconsin–Madison, Madison, WI, USA

[b] MS Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 6 September 2014
Available online 26 December 2014

Keywords:
Phylogenetic reconstruction
Incomplete lineage sorting
Maximum likelihood
Consistency

ABSTRACT

The reconstruction of a species tree from genomic data faces a double hurdle. First, the (gene) tree describing the evolution of each gene may differ from the species tree, for instance, due to incomplete lineage sorting. Second, the aligned genetic sequences at the leaves of each gene tree provide merely an imperfect estimate of the topology of the gene tree. In this note, we demonstrate formally that a basic statistical problem arises if one tries to avoid accounting for these two processes and analyses the genetic data directly via a concatenation approach. More precisely, we show that, under the multispecies coalescent with a standard site substitution model, maximum likelihood estimation on sequence data that has been concatenated across genes and performed under the incorrect assumption that all sites have evolved independently and identically on a fixed tree is a statistically inconsistent estimator of the species tree. Our results provide a formal justification of simulation results described of Kubatko and Degnan (2007) and others, and complements recent theoretical results by DeGIorgio and Degnan (2010) and Chifman and Kubtako (2014).

1. Introduction

Modern molecular sequencing technology has provided a wealth of data to help biologists infer evolutionary relationships between species. Not only is it possible to quickly sequence a single gene across a wide range of species, but hundreds, or even thousands of genes can also be sequenced across those taxa. But with this abundance of data comes new statistical and mathematical challenges. These arise because tree inference requires dealing with the interplay of at least two random processes, as we now explain.

For each gene, the associated aligned sequence data provides an estimate of the evolutionary *gene tree* that describes the ancestry of this gene as one traces back its ancestry in time (each copy being inherited from one parent in the previous generation). Moreover, given sufficiently long sequences, several methods (e.g. maximum likelihood and corrected distance methods) have been shown to be statically consistent estimators of the gene tree topology under various site substitution models (Felsenstein, 2004). 'Statistical consistency' here refers to the usual notion in molecular phylogenetics, namely that as the sequence length grows, the probability that the correct gene tree topology is returned from the data converges to 1 as the number of sites grows. Here the site patterns are assumed to be generated independently and identically (i.i.d.) under the substitution model on a binary (fully-resolved) gene tree.

But inferring a gene tree is only part of the puzzle of reconstructing the main evolutionary object of interest in biology—namely a *species tree*. This latter tree describes, on a broad (macroevolutionary) scale, how lineages (consisting of populations of a species) successively separated and diverged from each other over evolutionary time scales, with some lineages forming new species, ultimately leading to the given taxa observed at the present (a precise definition of a species-level phylogenetic tree is problematic as it requires first agreeing on a definition of 'species', for which there are multitude of differing opinions) (Maddison, 1997; Mayden, 1997; Nichols, 2001). A species tree, together with the length (time-scale) and width (population size) of its branches, induces a probability distribution on the possible gene trees and, when the discordance between gene trees is attributed to incomplete lineage sorting, this probability distribution can be described by the so-called *multispecies coalescent* process (details are provided in the recent book by Knowles and Kubatko, 2010). This process extends the celebrated *Kingman coalescent* process from a single population

[*] Corresponding author.
E-mail address: mathmomike@gmail.com (M. Steel).

http://dx.doi.org/10.1016/j.tpb.2014.12.005
0040-5809/© 2014 Elsevier Inc. All rights reserved.

to a phylogenetic tree, where the latter can be viewed as a 'tree of populations'.

The relationship between gene trees and species trees has attracted a good deal of attention from mathematicians and statisticians over the last decade or so (Degnan and Rosenberg, 2009; Huang et al., 2010; Liu et al., 2009a,b; Roch, 2013b; Rosenberg, 2002). An early and easily verified result is that for three taxa, the most probable gene tree topology under the multispecies coalescent matches the species tree (the other two competing binary topologies have equal but lower probability) (Tajima, 1983). Consequently, estimating the species tree by the gene tree that appears most frequently is a statistically consistent method (under the multispecies coalescent) when we have just three taxa. Moreover, when there are more than three taxa, one can still estimate a species tree consistently, for example, by estimating all the rooted triples, and using these to reconstruct the species tree topology (Degnan et al., 2009).

However, the alternative simple 'majority rule' strategy of estimating the species tree by merely taking the most frequent gene tree falls apart when we have more than three species. With four taxa, the most probable gene tree topology can differ from certain (unbalanced) species tree topologies, while for five or more taxa a more striking result applies—*every* species tree topology has branch lengths for which the most probable gene tree topology differs from that of the species tree (for details, see Degnan and Rosenberg, 2009). Nevertheless, one can still infer a species tree in a statistically consistent manner from a series of gene trees generated i.i.d. by the multispecies coalescent process, and several techniques have been developed for this (see e.g. Dasarathy et al., 2014, DeGiorgio and Degnan, 2010, Degnan et al., 2009, Liu et al., 2009b, Liu et al., 2010a, Liu et al., 2010b, Mossel and Roch, 2010 and Roch, 2013a).

There are also additional mechanisms that can lead to conflict between gene trees and species trees, including reticulate evolution (e.g. the formation of hybrid species), lateral gene transfer (in prokaryotic taxa such as bacteria) and gene duplication and loss, but we do not consider these processes here.

We have so far discussed these two random processes – the evolution of sequence site patterns on a gene tree under a site-substitution model, and the random generation of gene trees from the species tree under the multispecies coalescent process – as separate process. But in reality these two processes work in concert, a gene tree will have a random topology (determined by the multispecies coalescent on the species tree) and on this random gene tree sequences will evolve according to a substitution process. Thus, it is not immediately obvious whether methods exist for inferring a species tree topology directly from a series of aligned sequences (one for each gene) which would be statistically consistent as the number of genes grows. Using techniques from algebraic statistics, Chifman and Kubatko (2014) recently established that the species tree topology (up to the placement of the root) is an identifiable discrete parameter under the combined substitution–coalescence process. Moreover they describe an explicit method for estimating the species tree based on phylogenetic invariants and singular value decomposition techniques. For Bayesian inference of species trees directly from sequence data (e.g. via the program *BEAST, Heled and Drummond, 2010) the statistical consistency has also been formally established (Steel, 2013).

In this paper we consider a simpler and alternative strategy that has been used widely for inferring the species tree directly from sequence data, namely concatenation of sequences (e.g. Meredith et al., 2011 and Rokas et al., 2003). In its simplest form, this strategy simply concatenates all the sequences, and treats them as though each site had evolved i.i.d. on a fixed tree. Kubatko and Degnan (2007) used simulations to study the performance of such a concatenation approach, and their finding suggested that it could lead to misleading phylogenetic estimates. Nevertheless, the accuracy of concatenation methods is still very much under debate (e.g. Gatesy and Springer, 2013, Song et al., 2012 and Wu et al., 2013). While many simulation studies have concluded that concatenation methods are significantly less accurate than ILS-based methods or are prone to producing erroneous estimates with high confidence (Heled and Drummond, 2010; Kubatko and Degnan, 2007; Kubatko et al., 2009; Larget et al., 2010; Leaché and Rannala, 2011), others have found that they can be more accurate under some conditions (such as low phylogenetic signal) (Bayzid and Warnow, 2013; Gadagkar et al., 2005; Mirarab et al., in press). Moreover, a formal proof of whether or not a standard statistical method, such as maximum likelihood (ML), is statistically consistent as an estimator of tree topology based on concatenated sequences has never been presented, with the exception of the work of DeGiorgio and Degnan (2010) who established the consistency of ML in the special case of three taxa under a molecular clock under the 2-state symmetric model of site substitution.

This is the motivation for our current paper. We consider what happens when ML is applied under the assumption that the sites evolve i.i.d. on a fixed tree (in keeping with the concatenation approach). Our main result (Theorem 1) shows that ML is statistically inconsistent as an estimator of tree topology, for certain fully-resolved trees on six leaves. Indeed the probability that the true species tree is an ML tree can be made as small as we wish in the limit as the number of genes grows (even with six taxa). What makes this result non-trivial is that studying the behaviour of misspecified likelihoods can be challenging. Our proof of inconsistency involves combining a number of arguments and results, including a classic result in populations genetics (the 'Ewens' Sampling formula'), a formal linkage between likelihood and parsimony, and the interplay of various concentration and approximations bounds.

2. Definitions and main result

Consider:

- a species tree topology T together with branch lengths L (which, for each edge e of T, combine temporal branch lengths (t_e) and an effective population size for that edge N_e—note the subscript e here refers to the edge e not 'effective').
- g aligned sequence data sets $A_1, A_2, \ldots, A_g$, where each data set A_i consists of sequences of the same length ℓ evolved i.i.d. under a symmetric r-state site substitution model at substitution rate θ on the random gene tree (with associated branch lengths) that is generated by (T, L) via the multispecies coalescent model. That is, on each branch of T, looking backwards in time, lineages entering the branch coalesce at constant rate according to the Kingman coalescent with fixed population size. The remaining lineages at the top of the branch enter the ancestral population. For each locus, conditioned on the generated gene tree, each site in the aligned sequence data set is generated according to the symmetric r-state model.

  The sequence length ℓ may in turn depend on the number of data sets g, and so we write $\ell = \ell(g)$.
- maximum likelihood tree(s) T_{ML} for the concatenated aligned sequence data sets $A_1A_2 \cdots A_g$ inferred under the assumption that all sites evolve i.i.d. on a tree according to the symmetric r-state site substitution model (for branch lengths that are optimized, as usual, as part of the ML estimation).

Let $P(T, L, r, g, \ell, \theta)$ be the probability that T has the same unrooted topology as (at least one) ML tree T_{ML}. Our main result can be stated as follows.