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.

Theorem 1. *Under the model described above, there exist tree topologies T with branch lengths L for T, and a site substitution rate θ sufficiently small, for which the following holds: For any $\delta > 0$, there is a value g_0 so that*

$$P(T, L, r, g, \ell, \theta) \leq \delta$$

for all $g \geq g_0$, and for all sequence length functions $\ell = \ell(g)$.

3. Heuristic argument and a key preliminary result

The formal proof of Theorem 1 is presented in the next section. Here we describe the idea of the proof, and establish a preliminary result that is central to the proof. Notice that the although our proof involves the most frequent gene tree topology differing from the species tree topology, that in itself, does not imply that maximum likelihood on the concatenation of data sets will pick the wrong tree. However, what we show is that the wrong topology does indeed lead to a higher expected likelihood.

An outline of the proof of the main theorem is as follows: We show that the expected proportion of sites that are constant can be made arbitrary large with low rates of evolution (the lower bounds are formalized in Claim 4) and that the empirical frequencies of site patterns is concentrated around the expected values (Claim 2). When there are a large enough number of invariable sites, it can be shown that likelihood scores and parsimony scores converge to the same answer (formalized in Claim 1). Thus trees that have better parsimony score have better likelihood under these scenarios. Therefore, it suffices to show that parsimony is not statistically consistent under arbitrary low rates of evolution. Furthermore, it can be argued that if all the branches of the species tree are small enough, chances of coalescence events and mutations both becomes exceedingly small for all branches except the ancestral population; therefore, the expected parsimony score of the concatenated aligned sequence data sets is close to the parsimony score of the coalescent tree in the ancestral population (formalized in Claim 5).

Moreover, it can be shown that the difference between the expected parsimony score for a coalescent tree under the r-state models of evolution and the infinite alleles model (i.e. which implies no back mutations) can be bounded (formalized in Claim 6). Claims 3, 5 and 6 (given the right bounds) imply that the expected parsimony score of the concatenated aligned sequence data sets is arbitrarily close to the expected parsimony score of a coalescent tree under infinite alleles model. A key insight (Proposition 1) is that assuming an infinite alleles model, the balanced coalescent tree on six taxa has a lower expected parsimony score than an unbalanced tree with six taxa. Therefore, if the true tree is the unbalanced tree, the balanced tree has better parsimony score according to infinite alleles model, and by extension (under the conditions imposed by bounds), according to r-state model. Therefore, parsimony would pick the balanced tree under these scenarios.

We state and prove this preliminary proposition in this section, as it plays a key role in the final step (Claim 7) of the proof of the theorem.

Given allele frequencies $(a_1, a_2, \ldots, a_n)$ where $\sum_{j=1}^{n} ja_j = n$, the celebrated 'Ewens' Sampling Formula' describes the probability of generating such an allele distribution in a coalescent tree, with scaled mutation rate $\theta = 4N\mu$ under an infinite alleles model:

$$P_{\theta,n}(a_1, a_2, \ldots, a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!},$$

where $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$. (for details, see Durrett, 2008, p.18). We will apply this in the current setting, where $n = 6$ and $\theta = \epsilon$ a small positive constant (to be determined later).

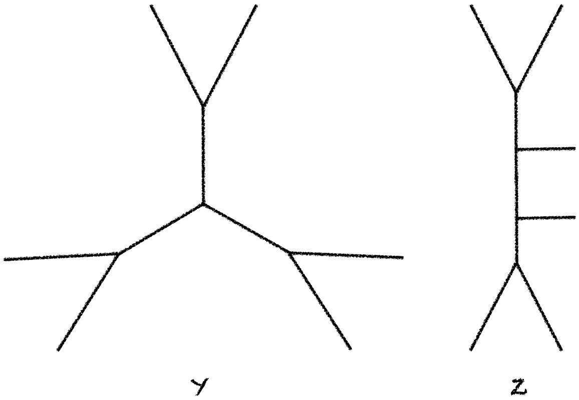

Fig. 1. The two binary tree shapes on six leaves: (a) the shape Y; (b) the shape Z. There are 15 and 90 phylogenetic trees on a given leaf set that have the shapes Y and Z, respectively.

Let $\mathbf{x} = (0, 1, 0, 1, 0, 0)$ and $\mathbf{y} = (0, 0, 2, 0, 0, 0)$. Then

$$P_{\epsilon,6}(\mathbf{x}) = \frac{6!}{\epsilon_{(6)}} \frac{(\epsilon/2)^1}{1!} \frac{(\epsilon/4)^1}{1!} = \frac{3}{4}\epsilon + O(\epsilon^2). \quad (1)$$

Similarly,

$$P_{\epsilon,6}(\mathbf{y}) = \frac{6!}{\epsilon_{(6)}} \frac{(\epsilon/3)^2}{2!} = \frac{1}{3}\epsilon + O(\epsilon^2). \quad (2)$$

Note that in both the equations, $O(\epsilon^2)$ refers to any values – negative or positive – which in absolute value is less than a constant times ϵ^2.

Consider the two unrooted binary tree shapes on six leaves, shown in Fig. 1, and denote these as Y (the symmetric tree with three cherries) and Z (the caterpillar tree with two cherries), where a *cherry* refers to a pair of leaves that are adjacent to a shared vertex.

We apply the above calculations to establish the following result, where an *unrooted binary phylogenetic tree* is a leaf-labelled tree, in which each non-leaf vertex has degree 3, and where a *site pattern* refers to the partition of the leaf set.

Proposition 1. *Let T and T' be two unrooted binary phylogenetic trees of shapes Z and Y respectively. Consider a site pattern that is randomly generated on a coalescent tree on the same leaf set under the infinite alleles model with scaled mutation rate $\theta (= 4N\mu) = \epsilon$. For a binary tree topology W, let $\mathcal{P}_W(\chi)$ denote the parsimony score of a site pattern χ on W. Then*

$$\mathbb{E}_{\mathrm{ESF}}[\mathcal{P}_Z(\chi) - \mathcal{P}_Y(\chi)] = \frac{1}{60}\epsilon + O(\epsilon^2)$$

where $\mathbb{E}_{\mathrm{ESF}}$ denotes the expectation under the infinite-alleles model.

Proof. We first note that we need only consider binary site patterns (which correspond to $\mathbf{x}$ and $\mathbf{y}$) to establish Proposition 1. This is because, apart from $\mathbf{x}$ and $\mathbf{y}$, the only allele distributions $\mathbf{z} = (a_1, a_2, \ldots, a_n)$ for which $P_{\epsilon,n}(\mathbf{z})$ is not $O(\epsilon^2)$ are $\mathbf{z} = (1, 0, 0, 0, 1, 0)$ and $\mathbf{z} = (0, 0, 0, 0, 0, 1)$, and each of these distributions corresponds to a character that has equal parsimony scores on the two tree topologies Y and Z. Thus it suffices to consider just the two binary patterns.

We refer to a binary pattern on the leaf set $\{1, 2, 3, 4, 5, 6\}$ as a *k-clade* if there are k leaves in one state, and $6-k$ in another ($k \leq 3$). Given such a binary pattern, the *additional penalty* of this clade is its homoplasy score (i.e. the parsimony score minus 1, unless the clade is a 0-clade in which case the penalty is 0).

For a phylogenetic tree having shape Y there are:

- $\binom{3}{2} \cdot 2 \cdot 2 = 12$ in total 2-clades that cost an additional penalty of $+1$;
- $\binom{3}{1} \cdot \binom{2}{1} = 6$ in total 3-clades that cost an additional penalty of $+1$;
- $\frac{1}{2} 2 \cdot 2 \cdot 2 = 4$ in total 3-clades that cost an additional penalty of $+2$.

Thus the expected value of the additional parsimony penalty Δ_Y for a tree phylogenetic tree having shape Y is:

$$1 \cdot P_{\epsilon,6}(\mathbf{x}) \cdot \frac{12}{\binom{6}{2}} + 1 \cdot P_{\epsilon,6}(\mathbf{y}) \cdot \frac{6}{\frac{1}{2}\binom{6}{3}} + 2 \cdot P_{\epsilon,6}(\mathbf{y}) \cdot \frac{4}{\frac{1}{2}\binom{6}{3}}.$$

Substituting Eqs. (1) and (2) into this last expression gives:

$$\mathbb{E}_{\mathrm{ESF}}[\Delta_Y] = \frac{16}{15}\epsilon + O(\epsilon^2). \tag{3}$$

A similar analysis for a Z-shape tree shows that there are:

- 13 in total 2-clades that cost an additional penalty of $+1$;
- 5 in total 3-clades that cost an additional penalty of $+1$;
- 4 in total 3-clades that cost an additional penalty of $+2$.

Thus the expected value of the additional parsimony penalty Δ_Z for a phylogenetic tree having shape Z is:

$$1 \cdot P_{\epsilon,6}(\mathbf{x}) \cdot \frac{13}{\binom{6}{2}} + 1 \cdot P_{\epsilon,6}(\mathbf{y}) \cdot \frac{5}{\frac{1}{2}\binom{6}{3}} + 2 \cdot P_{\epsilon,6}(\mathbf{y}) \cdot \frac{4}{\frac{1}{2}\binom{6}{3}}.$$

Substituting Eqs. (1) and (2) into this last expression gives:

$$\mathbb{E}_{\mathrm{ESF}}[\Delta_Z] = \frac{13}{12}\epsilon + O(\epsilon^2). \tag{4}$$

Combining Eqs. (3) and (4) gives:

$$\mathbb{E}_{\mathrm{ESF}}[\Delta_Z - \Delta_Y] = \frac{1}{60}\epsilon + O(\epsilon^2). \tag{5}$$

Proposition 1 now follows from (5). □

4. Proof of Theorem 1

To establish Theorem 1 it suffices to do so for any number n of taxa, and we do so for $n = 6$. For the species tree $\mathcal{T}$, take any rooted tree that has the unrooted topology of the Z-shaped tree (caterpillar). Make the branch length L of all the non-root edges of this tree less than β (see Fig. 2). We use the following notation:

- Denote by $G_1, \ldots, G_g$ the gene trees generated by the multi-species coalescent on $\mathcal{T}$.
- Let $\mathbb{E}_{\mathcal{T}}$ denote the expectation operation under $\mathcal{T}$ (i.e. for the expectation of quantities dependent on the gene tree, which are sampled from $\mathcal{T}$ under the multispecies coalescent model) and let G be a gene tree generated under $\mathcal{T}$.
- Let $\mathcal{C} = [r]^n$ be the set of r-state characters on the set of n taxa.
- Let $\chi^f_k \in \mathcal{C}$ be the k-th character of the f-th aligned sequence data set, where $1 \le k \le \ell$ and $1 \le f \le g$, and let $\mathcal{X} = \{\chi^f_k\}_{k,f}$.
- For a character $\chi \in \mathcal{C}$, let N^f_χ be the number of times character χ appears in the f-th aligned sequence data set and let N_χ be the number of times it appears overall.
- For an n-leaf tree with mutation probabilities $\{q_e\}$, let p^U_χ denote the probability that χ is produced by U under the symmetric r-state site substitution model.
- Let p^U_0 denote the probability that the characters produced by U under the symmetric r-state site substitution model assigns all leaves the same state.

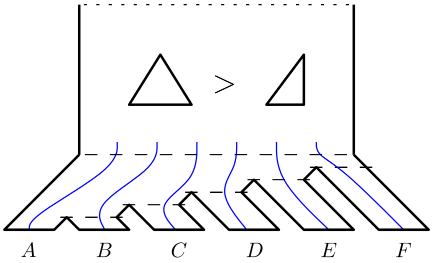

Fig. 2. Anomalous gene trees on a 6-taxon species tree with (unrooted) shape Z. The event of deepest coalescence is depicted. For the associated unrooted phylogenetic tree, any balanced tree (shape Y) has higher probability than any unbalanced tree (shape Z).

Observe that the (mis-specified, i.e., not taking into account the coalescent) empirical minus log-likelihood under tree U is given by

$$\mathcal{L}_U(\mathcal{X}) = -\frac{1}{g\ell}\sum_{k,f}\log\left[rp^U_{\chi^f_k}\right] = -\frac{1}{g\ell}\sum_{\chi\in\mathcal{C}} N_\chi \log\left[rp^U_\chi\right].$$

We want to show that with high probability $\mathcal{L}_U(\mathcal{X})$ is *not* minimized on the species tree topology. We follow the proof sketched in Section 3.

For a binary tree topology W and a character $\chi \in \mathcal{C}$ we let $\mathcal{P}_W(\chi)$ denote the parsimony score of χ on W. Let

$$\mathcal{P}_W(\mathcal{X}) = \frac{1}{g\ell}\sum_{k,f}\mathcal{P}_W(\chi_{k,f}) = \frac{1}{g\ell}\sum_{\chi\in\mathcal{C}} N_\chi \mathcal{P}_W(\chi).$$

Let $E(W)$ and $V(W)$ be the edges and vertices of W. We assume that W is binary and has n leaves, hence $|E(W)| = 2n - 3$ and $|V(W)| = 2n-2$. Let $\mathcal{L}^*_W(\mathcal{X})$ be the minus log-likelihood under an optimal choice of branch lengths (in $[0, +\infty]$) for W. Let N_0 denote the number of constant characters and $N_{\neq 0} = g\ell - N_0$.

Claim 1 (*Parsimony-Based Approximation of the Likelihood*). *If*

$$N_0 > 1, \quad \frac{g\ell\mathcal{P}_W(\mathcal{X})}{N_0} \le 1 \tag{6}$$

then, for all $q_0 \in (0, 1)$,

$$\mathcal{L}^*_W(\mathcal{X}) \le -\mathcal{P}_W(\mathcal{X}) \log\left(\frac{q_0}{r-1}\right) - 2n\log(1-q_0) \tag{7}$$

and

$$\mathcal{L}^*_W(\mathcal{X}) \ge -\mathcal{P}_W(\mathcal{X}) \log\left(\frac{g\ell\mathcal{P}_W(\mathcal{X})}{(r-1)N_0}\right) - \frac{N_{\neq 0}}{g\ell} n \log r. \tag{8}$$

Proof. First we recall some further notation: given a binary tree topology W and a character $\chi \in \mathcal{C}$ a *minimal extension* of χ is an assignment of states to the interior vertices of T that minimizes the number of edges of W with different states at the ends of the edge (thus $\mathcal{P}_W(\chi)$ is the number of such edges).

We adapt several bounds derived in Tuffley and Steel (1997, Lemmas 5 and 6). Letting U have topology W with all transition probabilities equal to q_0, by considering a minimal extension of χ (see Tuffley and Steel, 1997, Eq. (52)) we have

$$\begin{aligned} rp^U_\chi &\ge \left(\frac{q_0}{r-1}\right)^{\mathcal{P}_W(\chi)} (1-q_0)^{2n-3-\mathcal{P}_W(\chi)} \\ &\ge \left(\frac{q_0}{r-1}\right)^{\mathcal{P}_W(\chi)} (1-q_0)^{2n}, \end{aligned}$$

and therefore

$$\begin{aligned}\mathcal{L}_W^*(\mathcal{X}) &\le -\frac{1}{g\ell}\left\{\sum_{\chi\neq 0} N_\chi \log\left[\left(\frac{q_0}{r-1}\right)^{\mathcal{P}_W(\chi)}(1-q_0)^{2n}\right]\right.\\ &\quad\left.+N_0\log\left[(1-q_0)^{2n}\right]\right\}\\ &= -\mathcal{P}_W(\mathcal{X})\log\left(\frac{q_0}{r-1}\right) - 2n\log(1-q_0),\end{aligned}$$

where we used that $N_0 + \sum_{\chi\neq 0} N_\chi = g\ell$. This proves (7).

For the other direction, let U be the tree with topology W and optimal mutation probabilities $(q_e^*)_e$. Let $\bar{q} = \max_e q_e$. Then, summing over all minimal extensions of χ (see Tuffley and Steel, 1997, Eq. (63)),

$$rp_\chi^U \le r^{n-2}\left(\frac{\bar{q}}{r-1}\right)^{\mathcal{P}_W(\chi)} \le r^n\left(\frac{\bar{q}}{r-1}\right)^{\mathcal{P}_W(\chi)},$$

and by considering two leaves whose connecting path goes through an edge with probability $\bar{q}$ (see Tuffley and Steel, 1997, Eq. (9))

$$p_0^U \le 1-\bar{q}.$$

Hence

$$\begin{aligned}\mathcal{L}_U(\chi) &\ge -\frac{1}{g\ell}\left\{\sum_{\chi\neq 0} N_\chi\log\left[r^n\left(\frac{\bar{q}}{r-1}\right)^{\mathcal{P}_W(\chi)}\right] + N_0\log(1-\bar{q})\right\}\\ &\ge -\frac{N_{\neq 0}}{g\ell} n\log r - \mathcal{P}_W(\mathcal{X})\log\left(\frac{\bar{q}}{r-1}\right) + \frac{N_0}{g\ell}\bar{q},\end{aligned}$$

where we used $-\log(1-\bar{q}) \ge \bar{q}$. Minimizing $\mathcal{L}_U(\chi)$ over $\bar{q}$ (see Tuffley and Steel, 1997, Eqs. (65) and (66)), a lower bound is obtained by fixing $\bar{q}$ to $g\ell\mathcal{P}_W(\mathcal{X})/N_0$. □

In order for the approximation in Claim 1 to be useful, we need that N_0 is asymptotically larger than $\max\{nN_{\neq 0}, r\}$ and that $\mathcal{P}_W(\mathcal{X})$ is not too small. We proceed to prove that these two properties hold when the mutation rate is low enough.

We begin by showing that the empirical frequencies of characters are close to their expectation when $g \to +\infty$.

Claim 2 (*Concentration of Empirical Frequencies*)**.** *For every $\zeta_1 > 0$, with probability exceeding $1 - 2r^n\exp(-2g\zeta_1^2)$, for all $\chi \in \mathcal{C}$,*

$$\left|\frac{1}{g\ell}N_\chi - \mathbb{E}_{\mathcal{T}}[p_\chi^G]\right| < \zeta_1. \tag{9}$$

Proof. For all $\chi \in \mathcal{C}$,

$$\begin{aligned}\frac{1}{g\ell}N_\chi &= \frac{1}{g\ell}\sum_{k,f}\mathbf{1}_{\{\chi_k^f=\chi\}} = \frac{1}{g}\sum_f \frac{1}{\ell}N_\chi^f\\ &= \frac{1}{g}\sum_f\left(\frac{1}{\ell}\sum_k \mathbf{1}_{\{\chi_k^f=\chi\}}\right).\end{aligned}\tag{10}$$

Noting that the $\ell^{-1}N_\chi^f$s are in [0, 1] and independent, Hoeffding's inequality implies for all $\zeta_1 > 0$

$$\mathbb{P}_{\mathcal{T}}\left[\left|\frac{1}{g\ell}N_\chi - \frac{1}{g\ell}\mathbb{E}_{\mathcal{T}}[N_\chi]\right| \ge \zeta_1\right] \le 2\exp\left(-2g\zeta_1^2\right).$$

Moreover by Eq. (10)

$$\frac{1}{g\ell}\mathbb{E}_{\mathcal{T}}[N_\chi] = \frac{1}{g\ell}\sum_{k,f}\mathbb{E}_{\mathcal{T}}[\mathbf{1}_{\{\chi_k^f=\chi\}}] = \mathbb{P}_{\mathcal{T}}[\chi_k^f = \chi] = \mathbb{E}_{\mathcal{T}}[p_\chi^G].$$

The result follows from the fact that $|\mathcal{C}| = r^n$. □

An immediate corollary is the concentration of the parsimony score.

Claim 3 (*Concentration of Parsimony Score*)**.** *Under Eq. (9),*

$$|\mathcal{P}_W(\mathcal{X}) - \mathbb{E}_{\mathcal{T}}[\mathcal{P}_W(\chi)]| \le nr^n\zeta_1.$$

Proof. By definition,

$$\begin{aligned}&|\mathcal{P}_W(\mathcal{X}) - \mathbb{E}_{\mathcal{T}}[\mathcal{P}_W(\chi)]|\\ &= \left|\frac{1}{g\ell}\sum_\chi \mathcal{P}_W(\chi)N_\chi - \sum_\chi \mathcal{P}_W(\chi)\mathbb{E}_{\mathcal{T}}[p_\chi^G]\right|\\ &\le \sum_\chi \mathcal{P}_W(\chi)\left|\frac{1}{g\ell}N_\chi - \mathbb{E}_{\mathcal{T}}[p_\chi^G]\right|\\ &\le r^n n\zeta_1.\end{aligned}$$

□

The next two claims relate the multispecies coalescent to the standard coalescent. We will refer to the population of $\mathcal{T}$ ancestral to all taxa as the *master population*. We let $\mathcal{D}$ be the gene tree event that no coalescence occurs before the master population, which we refer to as *deepest coalescence* (*c.f* Fig. 2). We let $\mathcal{T}_{\mathcal{D}}$ be the coalescent model on the master population (i.e., the standard n-coalescent). We further let $\mathcal{D}'$ be the site event such that $\mathcal{D}$ occurs and further no mutation occurs below the master population. Let M be the number of mutations on a site.

Claim 4 (*Lower Bound on the Number of Constant Characters*)**.** *There is ζ_2 (depending only on n and β) such that, for any $\theta > 0$,*

$$\mathbb{E}_{\mathcal{T}}[p_0^G] \ge 1 - \zeta_2\theta.$$

Proof. Note that when no mutations occur in the tree then all leaves have the same state, and so:

$$\mathbb{E}_{\mathcal{T}}[p_0^G] \ge \mathbb{P}_{\mathcal{T}}[M = 0]. \tag{11}$$

The number of mutations on a site is stochastically dominated by the same quantity *conditioned on* $\mathcal{D}$. Indeed deepest coalescence ensures the highest total length of the gene tree. Hence

$$\begin{aligned}\mathbb{P}_{\mathcal{T}}[M = 0] &\ge \mathbb{P}_{\mathcal{T}}[M = 0 \mid \mathcal{D}] = \mathbb{E}_{\mathcal{T}}\left[\exp\left(-\theta H_G\right) \mid \mathcal{D}\right]\\ &\ge \mathbb{E}_{\mathcal{T}}\left[1 - \theta H_G \mid \mathcal{D}\right],\end{aligned}\tag{12}$$

where H_G is the total length of gene tree G. Note that, on $\mathcal{D}$,

$$H_G \le n\cdot n\beta + H_G',$$

where H_G' is the total length of the gene tree inside the master population. Letting $h_n^{(1)}$ be the expected length of the standard coalescent on n samples, we have, from (11) and (12):

$$\mathbb{E}_{\mathcal{T}}[p_0^G] \ge 1 - \theta(n^2\beta + h_n^{(1)}).$$

Therefore we can take $\zeta_2 = n^2\beta + h_n^{(1)}$. □

Claim 5 (*Reduction to Standard Coalescent*)**.** *For any θ and $\zeta_3 > 0$, there is β small enough (depending only on ζ_3, n, and θ), such that*

$$\left|\mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi)\right] - \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}\left[\mathcal{P}_W(\chi)\right]\right| \le \zeta_3.$$

Proof. Note that

$$\mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi) \mid \mathcal{D}'\right] = \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}\left[\mathcal{P}_W(\chi)\right].$$

Further

$$\begin{aligned}\mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi)\right] &= \mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi) \mid \mathcal{D}'\right]\mathbb{P}_{\mathcal{T}}[\mathcal{D}']\\ &\quad+\mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi) \mid (\mathcal{D}')^c\right]\mathbb{P}_{\mathcal{T}}[(\mathcal{D}')^c]\\ &\le \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}\left[\mathcal{P}_W(\chi)\right] + n\frac{\zeta_3}{n},\end{aligned}$$

by choosing β small enough to make the probability

$$\mathbb{P}_{\mathcal{T}}[\mathcal{D}'] \geq (e^{-\binom{n}{2}\beta})^n \exp(-\theta(n \cdot n\beta)) \geq 1 - \frac{\zeta_3}{n}.$$

Above we used that $\mathcal{P}_W(\chi) \leq n$. Similarly,

$$\begin{aligned}\mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi)\right] &= \mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi) \mid \mathcal{D}'\right]\mathbb{P}_{\mathcal{T}}[\mathcal{D}'] \\ &\quad + \mathbb{E}_{\mathcal{T}}\left[\mathcal{P}_W(\chi) \mid (\mathcal{D}')^c\right]\mathbb{P}_{\mathcal{T}}[(\mathcal{D}')^c] \\ &\geq \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}\left[\mathcal{P}_W(\chi)\right]\left(1 - \frac{\zeta_3}{n}\right) \\ &\geq \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}\left[\mathcal{P}_W(\chi)\right] - n\frac{\zeta_3}{n}. \quad \square\end{aligned}$$

Recall that $\mathbb{E}_{\text{ESF}}$ is the expectation under the infinite-alleles model on $\mathcal{T}_{\mathcal{D}}$.

Claim 6 (*Infinite-Alleles Approximation*)**.** *There is ζ_4 depending only on n such that, for any $\theta > 0$,*

$$\left|\mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi)] - \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi)]\right| \leq \zeta_4\theta^2.$$

Proof. Note that

$$\mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi) \mid M \leq 1] = \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi) \mid M \leq 1],$$

as a single mutation has the same effect on the characters of r-state symmetric and infinite-alleles models. Moreover, because both models are run with the same parameters, they have the same distribution of number of mutations. In particular,

$$\mathbb{P}_{\mathcal{T}_{\mathcal{D}}}[M \leq 1] = \mathbb{P}_{\text{ESF}}[M \leq 1].$$

Now, the number of mutations (M) follows a Poisson distribution with a mean that is proportional to the total tree length and so:

$$\begin{aligned}\mathbb{P}_{\text{ESF}}[M > 1] &= \mathbb{E}_{\text{ESF}}\left[\sum_{i\geq 2} e^{-\theta H_G}\frac{(\theta H_G)^i}{i!}\right] \\ &\leq \mathbb{E}_{\text{ESF}}\left[(\theta H_G)^2 \sum_{i\geq 0} e^{-\theta H_G}\frac{(\theta H_G)^i}{i!}\right] \\ &= \theta^2 h_n^{(2)},\end{aligned}$$

where $h_n^{(2)} = \mathbb{E}_{\text{ESF}}[H_G^2]$. Hence, since

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi)] &= \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi) \mid M \leq 1]\mathbb{P}_{\mathcal{T}_{\mathcal{D}}}[M \leq 1] \\ &\quad + \mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi) \mid M > 1]\mathbb{P}_{\mathcal{T}_{\mathcal{D}}}[M > 1],\end{aligned}$$

we have on the one hand

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi)] &\leq \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi) \mid M \leq 1]\mathbb{P}_{\text{ESF}}[M \leq 1] \\ &\quad + (\mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi) \mid M > 1] + n)\mathbb{P}_{\text{ESF}}[M > 1] \\ &\leq \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi)] + n\theta^2 h_n^{(2)}.\end{aligned}$$

And, on the other hand, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_{\mathcal{D}}}[\mathcal{P}_W(\chi)] &\geq \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi) \mid M \leq 1]\mathbb{P}_{\text{ESF}}[M \leq 1] \\ &\quad + (\mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi) \mid M > 1] - n)\mathbb{P}_{\text{ESF}}[M > 1] \\ &\geq \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi)] - n\theta^2 h_n^{(2)}. \quad \square\end{aligned}$$

Claim 7 (*Final Argument*)**.** *Let $\theta = \epsilon$. There are ϵ and β small enough (depending on n and r) such that $\mathcal{L}_Z^*(\mathcal{X}) > \mathcal{L}_Y^*(\mathcal{X})$, with probability exceeding $1 - 2r^n \exp(-2g\epsilon^4)$.*

Proof. Choosing β small enough, $\zeta_1 = \zeta_3 = \epsilon^2$. Claims 2 and 4 imply that, with probability exceeding $1 - 2r^n \exp(-2g\zeta_1^2)$,

$$N_0 \geq g\ell[1 - \zeta_2\epsilon - \epsilon^2] = g\ell(1 - O(\epsilon)), \quad N_{\neq 0} = O(g\ell\epsilon). \tag{13}$$

By Claims 3, 5 and 6,

$$|\mathcal{P}_W(\mathcal{X}) - \mathbb{E}_{\text{ESF}}[\mathcal{P}_W(\chi)]| = O(\epsilon^2). \tag{14}$$

Together with Proposition 1, this implies that

$$\mathcal{P}_Z(\mathcal{X}) - \mathcal{P}_Y(\mathcal{X}) = \frac{1}{60}\epsilon + O(\epsilon^2). \tag{15}$$

We finally return to the likelihood. Note that (6) in Claim 1 is satisfied by (13) and

$$\mathcal{P}_Y(\mathcal{X}) \leq nN_{\neq 0}/g\ell = O(n\varepsilon). \tag{16}$$

Hence, taking

$$q_0 = g\ell\mathcal{P}_Y(\mathcal{X})/N_0 = O(n\varepsilon), \tag{17}$$

in Claim 1 yields

$$\begin{aligned}\mathcal{L}_Z^*(\mathcal{X}) - \mathcal{L}_Y^*(\mathcal{X}) &\geq -[\mathcal{P}_Z(\mathcal{X}) - \mathcal{P}_Y(\mathcal{X})]\log\left(\frac{g\ell\mathcal{P}_Y(\mathcal{X})}{(r-1)N_0}\right) \\ &\quad + 2n\log\left(1 - \frac{g\ell\mathcal{P}_Y(\mathcal{X})}{N_0}\right) - \frac{N_{\neq 0}}{g\ell}n\log r \\ &\geq \left[\frac{1}{60}\epsilon + O(\epsilon^2)\right]\log[\Omega((n\epsilon)^{-1})] \\ &\quad - 2n \cdot O(n\epsilon) - n\log r \cdot O(\epsilon) \\ &> 0,\end{aligned}$$

by (16) and (17), when ϵ is small enough (depending on n and r). $\square$

Theorem 1 now follows immediately from Claim 7 by noting that the lower bound $1 - 2r^n \exp(-2g\epsilon^4)$ converges to 1 as g grows; consequently, the probability that Y has a higher likelihood than Z (i.e. a lower minus log-likelihood) converges to 1 as the number of aligned sequence data sets g increases.

5. Concluding comments

Our statistical inconsistency result applies for the particular case of a tree with six leaves. While this suffices to establish inconsistency in general, we conjecture that for any number $n > 6$ of leaves there is some species tree for which our inconsistency claim holds. However a detailed proof of this assertion is beyond the scope of this short note.

Acknowledgments

The authors thank the Simons Institute at UC Berkeley, where this work was carried out. S.R. is supported by National Science Foundation grants DMS-1248176 and DMS-1149312 (CAREER), and an Alfred P. Sloan Research Fellowship. M.S. would like to thank the NZ Marsden Fund (12-UOC-016) and the Allan Wilson Centre for funding support. In addition, we thank the three anonymous reviewers of this paper, and editor, for numerous helpful suggestions.

References

Bayzid, M.S., Warnow, T., 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29 (18), 2277–2284.

Chifman, J., Kubatko, L., 2014. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes. arXiv:1406.4811.

Dasarathy, G., Nowak, R., Roch, S., 2014. New sample complexity bounds for phylogenetic inference from multiple loci. In: 2014 IEEE International Symposium on Information Theory, ISIT, June, pp. 2037–2041.

DeGiorgio, M., Degnan, J.H., 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. Mol. Biol. Evol. 27 (3), 552–569.

Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N., 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58 (1), 35–54.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24 (6), 332–340.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution, second ed. Springer.

Felsenstein, J., 2004. Inferring Phylogenies, Vol. 2. Sinauer Associates, Sunderland.

Gadagkar, S.R., Rosenberg, M.S., Kumar, S., 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exp. Zool. B Mol. Dev. Evol. 304, 64–74.

Gatesy, J., Springer, M.S., 2013. Concatenation versus coalescence versus concatalescence. Proc. Natl. Acad. Sci. USA 110 (13), E1179.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27 (3), 570–580.

Huang, H., He, Q., Kubatko, L.S., Knowles, L.L., 2010. Sources of error for species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59 (5), 573–583.

Knowles, L.L., Kubatko, L.S., 2010. Estimating Species Trees: Practical and Theoretical Aspects. Wiley–Blackwell.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25 (7), 971–973.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56 (1), 17–24.

Larget, B.R., Kotha, S.K., Dewey, C.N., Ané, C., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26 (22), 2910–2911.

Leaché, A.D., Rannala, B., 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60 (2), 126–137.

Liu, L., Yu, L., Edwards, S.V., 2010a. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10 (1), 302.

Liu, L., Yu, L., Kubatko, L., Pearl, D.K., Edwards, S.V., 2009a. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53 (1), 320–328.

Liu, L., Yu, L., Pearl, D., 2010b. Maximum tree: a consistent estimator of the species tree. J. Math. Biol. 60, 95–106.

Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58 (5), 468–477.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46 (3), 523–536.

Mayden, R.L., 1997. A hierarchy of species concepts: the denoument in the saga of the species problem. In: Dawah, H.A., Claridge, M.F., Wilson, M.R. (Eds.), Species: The Units of Diversity. Chapman and Hall, London, pp. 381–423.

Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L.L., Stadler, T., Rabosky, D.L., Honeycutt, R.L., Flynn, J.J., Ingram, C.M., Steiner, C., Williams, T.L., Robinson, T.J., Burk-Herrick, A., Westerman, M., Ayoub, N.A., Springer, M.S., Murphy, W.J., 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334 (6055), 521–524.

Mirarab, S., Bayzid, M.S., Warnow, T., 2014. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. http://dx.doi.org/10.1093/sysbio/syu063. in press.

Mossel, E., Roch, S., 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) 7 (1), 166–171.

Nichols, R., 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16 (7), 358–364.

Roch, S., 2013a. An analytical comparison of multilocus methods under the multispecies coalescent: the three-taxon case. In: PSB'13—Proceedings of the Pacific Symposium on Biocomputing 2013, pp. 297–306.

Roch, S., 2013b. An analytical comparison of multilocus methods under the multispecies coalescent: the three-taxon case. In: Pacific Symposium on Biocomputing. World Scientific, pp. 297–306.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.

Song, S., Liu, L., Edwards, S.V., Wu, S., 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. 109 (37), 14942–14947.

Steel, M., 2013. Consistency of Bayesian inference of resolved phylogenetic trees. J. Theoret. Biol. 336, 246–249.

Tajima, F., 1983. Evolutionary relationships of DNA sequences in finite populations. Genetics 105, 437–460.

Tuffley, C., Steel, M.A., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59 (3), 581–607.

Wu, S., Song, S., Liu, L., Edwards, S.V., 2013. Reply to Gatesy and Springer: the multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. Proc. Natl. Acad. Sci. USA 110 (13), E1180.