Syst. Biol. 68(2):281–297, 2019

DOI:10.1093/sysbio/syy061
Advance Access publication September 21, 2018

Long-Branch Attraction in Species Tree Estimation: Inconsistency of Partitioned Likelihood and Topology-Based Summary Methods

SEBASTIEN ROCH[1], MICHAEL NUTE[2], AND TANDY WARNOW[3,*]

[1]*Department of Mathematics, University of Wisconsin–Madison, 480 Lincoln Dr, Madison, WI 53706, USA;*
[2]*Department of Statistics, The University of Illinois at Urbana-Champaign, 725 S Wright St #101, Champaign, IL 61820, USA; and*
[3]*Department of Computer Science, The University of Illinois at Urbana-Champaign, 201 North Goodwin Avenue, Urbana, IL 61801-2302, USA*
**Correspondence to be sent to: Department of Computer Science, The University of Illinois at Urbana-Champaign, 201 North Goodwin Avenue, Urbana, IL 61801-2302, USA; E-mail: warnow@illinois.edu.*

Received 16 March 2018; reviews returned 11 September 2018; accepted 12 September 2018
Associate Editor: Laura Kubatko

Abstract.—With advances in sequencing technologies, there are now massive amounts of genomic data from across all life, leading to the possibility that a robust Tree of Life can be constructed. However, "gene tree heterogeneity", which is when different genomic regions can evolve differently, is a common phenomenon in multi-locus data sets, and reduces the accuracy of standard methods for species tree estimation that do not take this heterogeneity into account. New methods have been developed for species tree estimation that specifically address gene tree heterogeneity, and that have been proven to converge to the true species tree when the number of loci and number of sites per locus both increase (i.e., the methods are said to be "statistically consistent"). Yet, little is known about the biologically realistic condition where the number of sites per locus is bounded. We show that when the sequence length of each locus is bounded (by any arbitrarily chosen value), the most common approaches to species tree estimation that take heterogeneity into account (i.e., traditional fully partitioned concatenated maximum likelihood and newer approaches, called summary methods, that estimate the species tree by combining estimated gene trees) are not statistically consistent, even when the heterogeneity is extremely constrained. The main challenge is the presence of conditions such as long branch attraction that create biased tree estimation when the number of sites is restricted. Hence, our study uncovers a fundamental challenge to species tree estimation using both traditional and new methods. [Incomplete lineage sorting; partitioned likelihood analysis; species tree estimation; statistical consistency.]

Species trees are a key aspect of much biological research, including the detection of co-evolution, the inference of ancestral traits, and the dating of speciation events (Posada 2016). The availability of sequence data collected from diverse species representing a broad spectrum of life has led to the expectation that the construction of a robust Tree of Life should be possible using statistical estimation methods, such as maximum likelihood. These estimations are increasingly based on large numbers of loci (sometimes thousands) selected from across the genomes of different species (Meredith et al. 2011; Jarvis et al. 2014; Misof et al. 2014; Wickett et al. 2014; Cannon et al. 2016; Maddison 2016).

Many methods used for species tree estimation, however, have been designed for gene tree estimation, which is a simpler statistical estimation problem. For gene tree estimation, the assumption is that the input sequences have all evolved down a single model tree (called the "gene tree") under a sequence evolution model, such as Cavender–Farris–Neyman (Neyman 1971; Farris 1973; Cavender 1978), Jukes–Cantor (Jukes and Cantor 1969), or the Generalized Time Reversible (GTR) model (Tavaré 1986). The estimation of the gene tree under these models from the aligned sequence data is a well-studied problem, and many statistically consistent methods have been developed under these models (Semple and Steel 2003). Species tree estimation is much more complex, since gene trees can differ from the species tree due to multiple causes, including incomplete lineage sorting (ILS), as modeled by the multi-species coalescent (MSC) model (Maddison 1997). Indeed, many recent phylogenetic analyses of genome-scale biological data sets for birds (Jarvis et al. 2014), land plants (Wickett et al. 2014), worms (Cannon et al. 2016), and other organisms, have revealed substantial heterogeneity across the genes that is consistent with ILS.

The construction of the species tree when there is gene tree heterogeneity due to ILS can be seen as a statistical estimation problem under a two-phase model of sequence evolution where gene trees evolve within a species tree under the MSC model, and then gene sequences evolve down each gene tree under a sequence evolution model. For example, under the MSC+JC model where true gene trees evolve within the species tree under the MSC model and gene sequences evolve down the gene trees under the Jukes–Cantor (JC) model, the estimation of species trees from gene sequence data needs to use the properties of the evolutionary models in order to be statistically consistent. One such approach for species tree estimation is to estimate gene trees for each locus, and then combine these gene trees into a species tree using a coalescent-based summary method (that takes gene tree incongruence due to ILS into account); such approaches can be proven to converge in probability to the true species tree as the number of genes and number of sites per gene both increase. Thus, for example, statistically consistent species tree estimation under these conditions is possible under the MSC+JC model when gene trees are estimated using Jukes–Cantor maximum likelihood and then combined

into a species tree using an appropriate coalescent-based summary method. Examples of these summary methods that enable statistically consistent species tree estimation include MP-EST (Liu et al. 2010), NJst (Liu and Yu 2011), ASTRID (Vachaspati and Warnow 2015), ASTRAL (Mirarab et al. 2014b; Mirarab and Warnow 2015), STEM (Kubatko et al. 2009), STEAC (Liu et al. 2009), STAR (Liu et al. 2009), and GLASS (Mossel and Roch 2010).

In contrast, many species trees are estimated using "unpartitioned maximum likelihood", where the gene sequence alignments are concatenated into a single supermatrix, and a tree is then estimated on that supermatrix under the assumption that all the sites evolve under the same model tree. As shown by Roch and Steel (2015), this approach is not statistically consistent and can even be positively misleading in the presence of gene tree heterogeneity due to ILS.

Although unpartitioned concatenated analysis with maximum likelihood (CA-ML) is known to be statistically inconsistent and coalescent-based species tree methods can be statistically consistent, performance in practice (and in particular on simulated data sets) has been mixed, with CA-ML sometimes more accurate than leading summary methods (Leaché and Rannala 2010; Patel et al. 2013; Mirarab et al. 2014a, Bayzid et al. 2015; Chou et al. 2015; Molloy and Warnow 2017). One of the challenges to using summary methods is gene tree estimation error, resulting in part from limited sequence lengths per gene (Bayzid and Warnow 2013). The "statistical binning" approach (Mirarab et al. 2014a) was designed to improve the accuracy of species trees estimated using summary methods by binning sequences from different genes together using statistical techniques for detecting strongly supported incongruence (e.g., using bootstrap support on estimated gene trees) and then estimating new gene trees on the combined data sets. As shown in Bayzid et al. (2015), weighted statistical binning (WSB) (an improved version of the original statistical binning approach) followed by appropriate summary methods is statistically consistent under the MSC+JC model under the condition that the number of genes and number of sites per gene both increase.

Note however that the guarantees of statistical consistency provided so far have nearly always made the following assumptions: every locus is recombination-free, the number of sites per locus increases without bound, and the number of loci increases without bound. These assumptions are unrealistic, since recombination-free loci are generally short. Therefore, of greater relevance to practice is the question of statistical consistency where the number of recombination-free loci increases, but the number of sites per locus is bounded by some $L \in \mathbf{Z}_+$ (Warnow 2015; Roch and Warnow 2015). We investigate this question for the following methods:

- fully partitioned maximum likelihood,
- topology-based summary methods (i.e., methods that combine gene tree topologies), and
- weighted statistical binning pipelines followed by topology-based summary methods.

We address this question under the MSC+CFN model, where the CFN is the symmetric two-state sequence evolution model (i.e., the two-state version of the Jukes–Cantor model). Perhaps surprisingly, our results are negative: for all L, none of the approaches is statistically consistent under the MSC+CFN model and can even be positively misleading. Furthermore, this problematic behavior occurs *even when* all the genes evolve down a single model CFN tree. Therefore, expectations of accurate species trees using any of these methods given large amounts of data may be unfounded.

The key challenge to species tree estimation is *long branch attraction*, a phenomenon that can confound maximum likelihood tree estimation when sequence lengths for each genomic region are finite. In fact, we show that many species tree estimation methods that are statistically consistent when the number of genomic regions and their lengths both increase become inconsistent when only the number of regions increases, and the sequence length for each genomic region is bounded (however arbitrarily). These results suggest that many of the common approaches to species tree estimation are far from being mathematically rigorous, even under highly simplified model conditions where there is no heterogeneity between the loci. This is a very substantial limitation for multi-locus phylogeny estimation methods in general and shows that new approaches for species tree estimation are needed.

Evolution under the MSC

Our analysis is based on the MSC+CFN model. A CFN model gene tree is an unrooted binary tree $(\mathcal{T}, \Lambda)$ with topology $\mathcal{T}$ and branch lengths Λ. Under the assumption that the tree has n leaves, each site (character) χ refers to the length-n vector of character states corresponding the same homologous site for each taxon. The possible character states are $\{0,1\}$ and evolutionary changes are modeled by a continuous-time Markov process with instantaneous rate matrix $Q = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$. In particular, the probability of a change along a branch of length λ is parameterized as $p = \frac{1}{2}\left(1 - e^{-2\lambda}\right)$. Under the MSC+CFN model, each locus j evolves independently on a random gene tree $(\mathcal{T}_j, \Lambda_j)$, which is derived from the MSC on a species tree $(\mathcal{S}, \Gamma, \theta)$, where the Γ_es are the branch lengths in units of $\theta_e = 2N_e\mu_e$ with N_e and μ_e the effective population size and mutation rate of branch e. That is, on each branch e of $\mathcal{S}$, looking backwards in time, lineages entering the branch coalesce at rate $2/\theta_e$ according to the Kingman coalescent. The remaining lineages at the top of the branch enter the ancestral population, and so on (see Fig. 1 for an illustration).

We assume that all m loci evolve on the same species tree and that each locus has a constant, finite sequence

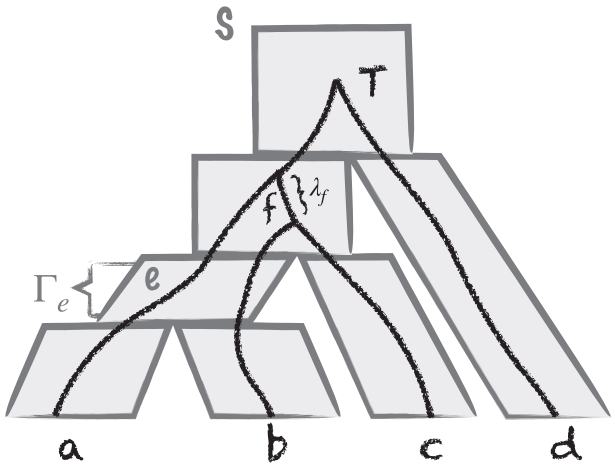

FIGURE 1. A species tree $(\mathcal{S}, \Gamma, \theta)$, represented above by the "box tree," together with a gene tree $(\mathcal{T}, \Lambda)$ inside it. An incomplete lineage sorting event is depicted: in branch e of the species tree the lineages from a and b fail to coalesce, thereby producing an unrooted topology for $\mathcal{T}$ (i.e., $ad|bc$) that differs from the unrooted topology of $\mathcal{S}$ (i.e., $ab|cd$).

length L. Let χ_{ij} represent site i on locus j, where $1 \le i \le L$ and $1 \le j \le m$, and let $\chi_{\cdot j}$ represent the set of all characters for locus j. We refer to the $\chi_{\cdot j}$ as *j-th locus sequences*. Denote the entire set of characters on all loci as X.

Inconsistency of Partitioned Maximum Likelihood

Let $\mathcal{L}(\mathcal{T}^0, \Lambda, \chi)$ denote the likelihood function for a single site χ under the CFN model on $(\mathcal{T}^0, \Lambda)$, and let $\ell = \log \mathcal{L}$ be the log-likelihood. Under fully partitioned maximum likelihood, we seek a single binary tree topology $\mathcal{T}^0$ but allow each locus to have its own branch length parameter Λ_j; hence, the general likelihood function over all sites and all loci is

$$\ell^*(\mathcal{T}^0, \Lambda_1, \ldots, \Lambda_m, X) = \sum_{j=1}^{m} \sum_{i=1}^{L} \ell(\mathcal{T}^0, \Lambda_j, \chi_{ij}),$$

and a maximum likelihood topology is any element of the set

$$\arg\max_{\mathcal{T}^0} \max_{\Lambda_1, \ldots, \Lambda_m} \ell^*(\mathcal{T}^0, \Lambda_1, \ldots, \Lambda_m, X). \qquad (1)$$

Theorem 1 (Inconsistency of partitioned ML)**.** *Under the MSC+CFN model, fully partitioned maximum likelihood on loci with a bounded number of sites is not statistically consistent and is even positively misleading. That is, for any length $L \in \mathbb{N}$, there is a species tree with topology, branch lengths and mutation rates such that, given data generated under the MSC+CFN model, as the number of loci $m \to \infty$, the maximum likelihood topology is unique and is different from the true species tree topology with probability going to 1.*

The proof of this theorem is provided in Section "Proofs of the Main Results" of the Appendix.

Inconsistency of Topology-Based Summary Methods

Summary methods have been developed that are designed to address heterogeneity between gene tree topologies due to ILS and are statistically consistent under the MSC model. We consider topology-based summary methods that take as input unrooted gene trees, and only use their topologies and not any additional information (e.g., sequence data, branch lengths, bootstrap support).

- We assume that the tree provided for a given gene sequence alignment is its maximum likelihood gene tree, and if there is a tie for the best maximum likelihood tree topology, then a random best-scoring tree is selected.

When the number of species is four, then the summary method is selecting the best unrooted tree topology from the three possible unrooted tree topologies, also referred to as quartet trees. By Allman et al. (2011), under the MSC the most probable quartet tree is the true species tree for any four species (i.e., there is no anomaly zone on unrooted four-leaf species trees). Hence, in the four species case, we will make the assumption that the summary method will return the tree topology that appears the most frequently among its input gene trees, as this is a statistically consistent technique for estimating the unrooted species tree on four leaves. We refer to this most frequent quartet tree as the "dominant" quartet tree. That is, we restrict ourselves to the following "reasonable" property of a summary method $\mathcal{A}$:

- When $n = 4$, as the number of loci m increases then with probability converging to 1, $\mathcal{A}(\mathcal{T}_1, \ldots, \mathcal{T}_m) = t$ where t is the quartet tree that appears with the highest frequency in the input $\mathcal{T}_1, \ldots, \mathcal{T}_m$; if there are ties, then $\mathcal{A}$ picks uniformly at random between the most frequent quartet trees.

We will say that the summary method $\mathcal{A}$ is *reasonable* if it satisfies this property. Many of the popular summary methods (e.g., ASTRAL and BUCKy, Larget et al. 2010) are reasonable in that sense.

Theorem 2 (Inconsistency of reasonable summary methods)**.** *Under the MSC+CFN model, any reasonable summary method $\mathcal{A}$ with maximum likelihood input trees on loci with a bounded number of sites is not statistically consistent. That is, for any length $L \in \mathbb{N}$, there is a species tree with topology, branch lengths, and mutation rates, such that given data generated under the MSC+CFN model, as the number of loci $m \to \infty$, the topology produced by $\mathcal{A}$ is unique and is different from the true species tree topology with probability going to 1.*

Inconsistency of Weighted Statistical Binning Followed by a Summary Method

The "statistical binning" method, and its improved version "weighted statistical binning", were developed

to address challenges in species tree estimation that result from gene tree estimation error. In Bayzid et al. (2015), it was shown that statistical binning was inconsistent under the MSC+CFN model but that WSB was statistically consistent. Those proofs depend crucially on the number of sites per locus increasing to infinity, and so this previous work did not address the case we consider here, where each site has length bounded by L.

In a WSB pipeline, estimated gene trees with bootstrap support are provided for every locus, and then an incompatibility graph is computed for that set of gene trees with branch support. The graph is used to partition the genes into sets (called "bins") and then "supergene trees" are computed using a fully partitioned maximum likelihood analysis on each bin. These supergene trees are then given to the selected summary method as input, and a species tree is returned. In a WSB pipeline, each supergene tree is replicated by the number of genes in its associated bin. The incompatibility graph depends on a parameter B, a proportion in $[0,1]$, as follows: two gene trees are considered to be incompatible if there is a pair of edges, one from each tree, each with bootstrap support strictly greater than B, that conflict. Hence, if $B=1$, then no two trees can be considered incompatible.

Theorem 3 (Inconsistency of WSB pipeline followed by reasonable summary method)**.** *Under the multi-locus MSC+CFN model, with a single site evolving down each gene tree, the WSB pipeline followed by a reasonable summary method is not statistically consistent.*

The proof of this theorem is given in the Appendix, and establishes that when each locus has a single site then there is a $B<1$ and a tree with topology, branch lengths, and mutation rates such that, given data generated under the MSC+CFN model, as the number of loci $m\to\infty$, the distribution produced by the WSB pipeline with support threshold B is "approximately flat." Hence, the application of $\mathcal{A}$ to this distribution will *not* converge to the true species tree topology with probability going to 1. Intuitively the WSB pipeline is not statistically consistent under the MSC+CFN model because uninformative genes can "swamp the bins" and produce a flat distribution.

One possible modification to the WSB pipeline, which we refer to as the WSB* pipeline, is to remove all genes that have no branches with "strong" bootstrap support:

- Remove all gene trees that do not support any internal edge above the bootstrap threshold B from the analysis before doing any binning.

This is one way—which can be analyzed rigorously—to address the problem above in that the distribution is no longer made "flat" by uninformative genes. However, we still show:

Theorem 4 (Inconsistency of WSB* pipeline followed by reasonable summary method)**.** *The WSB* pipeline followed by $\mathcal{A}$ is not only not statistically consistent but is positively misleading. That is, for any length $L\in\mathbb{N}$, there is a $B<1$ and a species tree with topology, branch lengths, and mutation rates such that, given data generated under the MSC+CFN model, as the number of loci $m\to\infty$, the topology produced by $\mathcal{A}$ after going through the WSB* pipeline with support threshold B is unique and is different from the true topology with probability going to 1.*

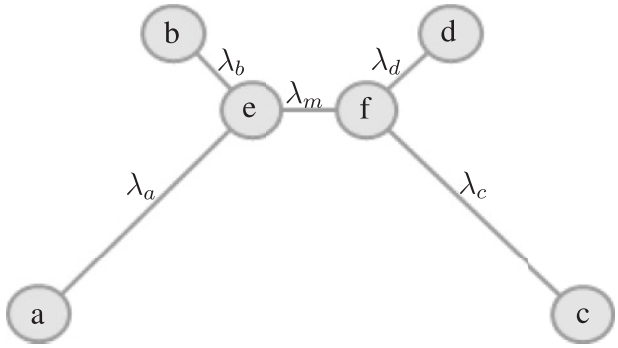

FIGURE 2. A four-taxon tree.

THEORETICAL FRAMEWORK

Our analysis in fact establishes a stronger—perhaps more counter-intuitive—result. We show that partitioned maximum likelihood, topology-based summary methods, and weighted statistical binning pipelines are statistically inconsistent for multi-locus evolution *where there is no gene tree heterogeneity at all* and when all loci have only L sites for any arbitrarily selected L. By a continuity argument, we also establish that these negative results imply that these methods, which were designed to address heterogeneity across the genome resulting from ILS, are also statistically inconsistent under the MSC+CFN model.

Setting for Analysis

Fix $\mathcal{T}^0$ to be the four-taxon topology $ab|cd$ on $\{a,b,c,d\}$ and let Λ^0 denote a vector of branch lengths on $\mathcal{T}^0$ under the CFN model. Specifically, denote the endpoint of the middle edge on the ab side as e, and on the cd side as f (see Fig. 2). For this tree, denote the length of branch ae as λ_a^0, be as λ_b^0, cf as λ_c^0, df as λ_d^0, and ef as λ_m^0.

For a branch length λ, we will also use the parameterization $\phi=-\frac{1}{2}\log\lambda$ in terms of which the probability of a change along this branch is

$$p=\frac{1}{2}\left(1-e^{-2\lambda}\right)=\frac{1}{2}(1-\phi), \tag{2}$$

and the probability of no change is $q=\frac{1}{2}(1+\phi)$. See (Semple and Steel, 2003, Section 8.6) for more details on this standard parameterization. Denote the p-, q-, and ϕ-parameters as defined above for each branch using the same subscripts. We choose Λ^0 to construct a Felsenstein zone tree (i.e., a four-leaf model tree where some tree

estimation methods are positively misleading, as shown in Felsenstein (1978)) where, for a parameter $\rho > 0$, $p_a^0 = p_c^0 = \rho$ and $p_b^0 = p_d^0 = p_m^0 = \rho^3$. Note that for any $\rho > 0$, we can set $\lambda_a^0 = \lambda_c^0 = -\frac{1}{2}\log(1-2\rho)$ and $\lambda_b^0 = \lambda_d^0 = \lambda_m^0 = -\frac{1}{2}\log(1-2\rho^3)$ to satisfy this relationship. We assume that the characters $\chi_{\cdot j}$, $j = 1,2,\ldots$, are generated under the CFN model on $(\mathcal{T}^0, \Lambda^0)$. We also denote the alternate topologies by $\mathcal{T}^* = ac|bd$ and $\mathcal{T}^1 = ad|bc$.

Basic Claims

The key step in proving our main theorems is to establish the following three claims. In the first claim, we show that for any sequence length L, by taking the mutation rate small enough (through choosing ρ) in the four-taxon species tree $(\mathcal{T}^0, \Lambda^0)$ described above we can ensure that the wrong topology is chosen by partitioned ML in the limit of large gene numbers. The idea behind the proof is described in the Section "Analysis of Partitioned ML" below. This claim implies Theorem 1 *in the absence of incomplete lineage sorting*. We then show in the Appendix that the effect of the MSC can be made negligible.

Claim 1 (Partitioned ML: Felsenstein zone)**.** *Assume that the length-L locus sequences $\chi_{\cdot j}$, $j = 1,2,\ldots$, are generated under the CFN model on $(\mathcal{T}^0, \Lambda^0)$ and let $\hat{\mathcal{T}}_j$ be the fully partitioned maximum likelihood topology obtained from the sequences of the first j loci. For any length $L \geq 1$, there is $\rho > 0$ small enough such that, with probability one, $\hat{\mathcal{T}}_j \to \mathcal{T}^*$ as $j \to +\infty$.*

In the second claim, we show that for any sequence length L, by taking the mutation rate small enough in the species tree $(\mathcal{T}^0, \Lambda^0)$ we can ensure that the wrong topology is chosen by any reasonable summary method in the limit of large gene numbers. The proof is similar to that of the previous claim, including the extra step to account for the effect of the MSC, as described in the Appendix.

Claim 2 (Reasonable summary methods: Felsenstein zone)**.** *Assume that the length-L locus sequences $\chi_{\cdot j}$, $j = 1,2,\ldots$, are generated under the CFN model on $(\mathcal{T}^0, \Lambda^0)$ and let $\hat{\mathcal{T}}_j$ be the topology obtained from a reasonable summary method $\mathcal{A}$ on the sequences of the first j loci using maximum likelihood. For any length $L \geq 1$, there is $\rho > 0$ small enough such that, with probability one, $\hat{\mathcal{T}}_j \to \mathcal{T}^*$ as $j \to +\infty$.*

Finally in the third claim, we show that for any sequence length L, by taking the mutation rate small enough in the species tree $(\mathcal{T}^0, \Lambda^0)$ we can ensure that the wrong topology is chosen by the WSB* pipeline (for a range of threshold B) followed by any reasonable summary method in the limit of large gene numbers. The idea of the proof is described in the Section "Analysis of WSB* Pipeline."

Claim 3 (WSB* pipeline: Felsenstein zone)**.** *Let $1 - \frac{2}{3}\left(\frac{1}{L}\right)^L \leq B < 1$. Assume that the length-L locus sequences $\chi_{\cdot j}$, $j = 1,2,\ldots$, are generated under the CFN model on $(\mathcal{T}^0, \Lambda^0)$ and let $\hat{\mathcal{T}}_j$ be the topology obtained from the WSB* pipeline with threshold B followed by a reasonable summary method $\mathcal{A}$ on the sequences of the first j loci. There is $\rho > 0$ small enough such that, with probability one, $\hat{\mathcal{T}}_j \to \mathcal{T}^*$ as $j \to +\infty$.*

While the claims above are established under the multi-locus CFN model with a single tree, we show in the Appendix that these results also apply to the MSC+CFN model by choosing a species tree, which is highly likely to produce gene trees matching the species tree.

Analysis of Partitioned ML

We describe the main ideas used to prove Claim 1. First, we note that the case $L = 1$ of a single site per gene corresponds to the No Common Mechanism model of Tuffley and Steel (1997), under which it was shown that maximum likelihood is equivalent to maximum parsimony, establishing statistical inconsistency (together with Felsenstein (1978)).

So we assume from now on that $L \geq 2$. Extending the results of Tuffley and Steel (1997) to this more general multi-locus setting requires a delicate asymptotic argument. We proceed as follows:

(a) By choosing ρ small enough, we show that we can restrict the analysis to the five most common data set types, which we refer to as *locus patterns*.

(b) We then show that, for these locus patterns, the likelihood on $\mathcal{T}^*$ dominates the likelihood on $\mathcal{T}^0, \mathcal{T}^1$, and that this domination is strict in one case.

Under our choice of branch lengths, as $\rho \to 0$, the five most common locus patterns, which we refer to as *dominant* (see Lemma 1 below for justification), are:

1. *All constant sites:* Every character has the same state on all four taxa, but that state can change from one character to another (e.g., $x^a = x^b = x^c = x^d = 0001010$). We let $\mathcal{X}_0$ be the set of such data sets and we let Q_0 be the probability of observing any $x \in \mathcal{X}_0$ under $(\mathcal{T}^0, \Lambda^0)$.

2. *One singleton site on a or c:* All sites are constant except for one, on which either a or c is different from all others (e.g., $x^a = 0111110$, $x^b = x^c = x^d = 1111110$). We let $\mathcal{X}_{11}$ be the set of such data sets and we let Q_{11} be the probability of observing any $x \in \mathcal{X}_{11}$ under $(\mathcal{T}^0, \Lambda^0)$.

3. *Two identical singleton sites on a or c:* All sites are constant except for two, each of which has the same taxon a or c different from the others (e.g.,

$x^a = 0011110$, $x^b = x^c = x^d = 1111110$). We let $\mathscr{X}_{2=}$ be the set of such data sets and we let $Q_{2=}$ be the probability of observing any $x \in \mathscr{X}_{2=}$ under $(\mathcal{T}^0, \Lambda^0)$.

4. *Two different singleton sites on a and c:* All sites are constant except for two, one of which has a different character state on a and the other a different character state on c (e.g., $x^a = 1001110$, $x^c = 0101110$, $x^b = x^d = 0001110$). We let $\mathscr{X}_{2\neq}$ be the set of such data sets and we let $Q_{2\neq}$ be the probability of observing any $x \in \mathscr{X}_{2\neq}$ under $(\mathcal{T}^0, \Lambda^0)$.

5. *One site with a 2/2-split ac|bd:* $L-1$ sites are constant with a single site having a and c different from b and d (e.g., $x^a = x^c = 1001110$, $x^b = x^d = 0001110$). We let $\mathscr{X}_{12}$ be the set of such data sets and we let Q_{12} be the probability of observing any $x \in \mathscr{X}_{12}$ under $(\mathcal{T}^0, \Lambda^0)$.

Note that above only the last pattern is informative and it supports the split in $\mathcal{T}^*$ rather than $\mathcal{T}^0$. Let $\widetilde{\mathscr{X}}$ be the set of all remaining locus patterns.

The next lemma, which encapsulates the key technical steps in the proof of Theorem 1, has two parts: in (a) the probability of observing the dominant locus patterns is bounded analytically; in (b) we show that the wrong topology has higher expected locus-wise likelihood under these dominant patterns. Claim 1 then follows by the law of large numbers, as detailed in the Appendix.

Lemma 1 (Dominant patterns and their likelihood contributions)**.** *Assume* $L \geq 2$.

(a) The probabilities of observing the dominant locus patterns are bounded as follows:

$$Q_0 = \left(\frac{1}{2}\right)^L - \mathcal{O}(\rho), \quad Q_{11} = \mathcal{O}(\rho), \quad Q_{2=} = \mathcal{O}(\rho^2),$$

$$Q_{2\neq} = \mathcal{O}(\rho^2) \text{ and } Q_{12} = \left(\frac{1}{2}\right)^L \rho^2 + \mathcal{O}(\rho^3).$$

Moreover, for all $x \in \widetilde{\mathscr{X}}$*, the probability of observing* x *under the CFN model on* $(\mathcal{T}^0, \Lambda^0)$ *is* $\mathcal{O}(\rho^3)$.

(b) For all $x \in \mathscr{X}_0 \cup \mathscr{X}_{11} \cup \mathscr{X}_{2=} \cup \mathscr{X}_{2\neq}$*, it holds that*

$$\sup_{\Lambda} \ell(\mathcal{T}^*, \Lambda, x) - \sup_{\Lambda} \ell(\mathcal{T}^0, \Lambda, x) \geq 0,$$

while, for all $x \in \mathscr{X}_{12}$*,*

$$\sup_{\Lambda} \ell(\mathcal{T}^*, \Lambda, x) - \sup_{\Lambda} \ell(\mathcal{T}^0, \Lambda, x) \geq K_{12} > 0,$$

for some positive constant K_{12} *depending only on L. The same holds if one replaces* $\mathcal{T}^0$ *with* $\mathcal{T}^1$ *above.*

Note that the big-O notation implicitly includes the contribution from L, which we treat as a constant. The detailed proofs of Lemma 1 and Claim 1 are provided in the Appendix. Claim 2 follows from a similar argument, which is also detailed in the Appendix.

Analysis of WSB Pipeline*

Our analysis of the WSB* pipeline follows along similar lines. Our key additional observation is that, by choosing an appropriate bootstrap threshold, we ensure that the only loci passed on to the summary method are "saturated," that is all their sites correspond to an equivalent character. The rest of the analysis is similar to Claim 2 and relies on the fact that the loci passed on to the summary method are dominated by the "wrong split." Formally, we say that two characters are equivalent if they are identical up to switching 0s and 1s. We say that a locus pattern x is *saturated* if all characters in x are equivalent. On four taxa, there are only three types of saturated patterns:

1. *All-constant:* Every character has the same value on all four taxa (e.g., $x^a = x^b = x^c = x^d = 0001010$). We let $\mathscr{X}_0^{\mathrm{s}}$ be the set of such data sets and we let Q_0^{s} be the probability of observing any $x \in \mathscr{X}_0^{\mathrm{s}}$ under $(\mathcal{T}^0, \Lambda^0)$.

2. *All-singleton on a fixed taxon:* All sites have the same taxon different from all others (e.g., $x^a = 0101111$, $x^b = x^c = x^d = 1010000$). We let $\mathscr{X}_1^{\mathrm{s}}$ be the set of such data sets and we let Q_1^{s} be the probability of observing any $x \in \mathscr{X}_1^{\mathrm{s}}$ under $(\mathcal{T}^0, \Lambda^0)$.

3. *All-2/2-split with a fixed split:* All sites have two fixed taxa—say, a and c—identical, while being different from the other two taxa—b and d—(e.g., $x^a = x^c = 1010111$, $x^b = x^d = 0101000$). We let $\mathscr{X}_{ac|bd}^{\mathrm{s}}$ be the set of such data sets for the split $ac|bd$ and we let $Q_{ac|bd}^{\mathrm{s}}$ be the probability of observing any $x \in \mathscr{X}_{ac|bd}^{\mathrm{s}}$ under $(\mathcal{T}^0, \Lambda^0)$ (and similarly for the other possible splits). For short, we refer to this type of data sets as *split-saturated genes*.

The next lemma, which encapsulates the key technical steps in the proof of Theorem 4, has two parts: in (a) the probability of observing the saturated locus patterns is bounded analytically, with the wrong topology being more common; in (b) we show that, under each informative saturated pattern, the corresponding topology has higher expected locus-wise likelihood. Claim 3 then follows by the law of large numbers, as detailed in the Appendix.

Lemma 2 (Saturated genes)**.**

(a) Under the WSB pipeline with threshold* $B \geq 1 - \frac{2}{3}\left(\frac{1}{L}\right)^L$*, the only length-L locus sequences passed on to the summary method are the ones in* $\mathscr{X}_{ac|bd}^{\mathrm{s}}$, $\mathscr{X}_{ab|cd}^{\mathrm{s}}$,

and $\mathscr{X}^s_{ad|bc}$. *Moreover,*

$$Q^s_{ac|bd} = \left(\frac{1}{2}\right)^L \rho^{2L} + \mathcal{O}(\rho^{2L+1}),$$

while

$$Q^s_{ab|cd} = \mathcal{O}(\rho^{3L}), \quad Q^s_{ad|bc} = \mathcal{O}(\rho^{3L}).$$

(b) *For any* $x \in \mathscr{X}^s_{ab|cd}$, *the topology* $ab|cd$ *is the unique ML optimizer. And similarly for the other splits.*

The detailed proofs of Lemma 2 and Claim 3 are provided in the Appendix.

Discussion

Our results show that fully partitioned maximum likelihood is inconsistent (even positively misleading) even when there is no gene tree heterogeneity at all (i.e., when all loci evolve down a common CFN model tree), and hence by continuity under the multi-locus MSC+CFN model. The inconsistency result occurs because each locus has at most L sites (for an arbitrarily selected bound L), and the loci all evolve down gene trees that have long branch attraction (LBA). It is well known that maximum likelihood is statistically consistent even in the presence of LBA, but our results show that LBA is sufficient to bias fully partitioned ML towards the same wrong tree on each locus, and hence towards the same wrong tree for the partitioned concatenation analysis.

The same argument is used to establish that reasonable summary methods and weighted statistical binning pipelines that use these reasonable summary methods can be positively misleading when each locus has only L sites, even when there is no gene tree heterogeneity. Hence, summary methods and WSB pipelines do not solve this challenge, either. All the methods we addressed in this study can be seen as partitioned analyses—partitioned maximum likelihood estimates numeric parameters for each locus but keeps the tree topology the same across the loci, and summary methods estimate the gene trees independently across the loci. The fundamental challenge to multi-locus species tree estimation using these partitioned analyses (whether partitioned maximum likelihood or summary methods) is that maximum likelihood tree estimation is impacted by conditions such as LBA when the number of sites is not allowed to increase.

It is interesting to consider unpartitioned maximum likelihood under the same set of conditions. When all the loci evolve down the same CFN model tree, even though each locus has only L sites, as the number of loci increases, the unpartitioned maximum likelihood analysis will converge to the true tree; thus, unpartitioned maximum likelihood analysis is consistent under this setting. On the other hand, when there is gene tree heterogeneity resulting from ILS (as modeled by the MSC), then unpartitioned ML is inconsistent and can be positively misleading (Roch and Steel 2015). Hence, unpartitioned maximum likelihood can be statistically consistent under one setting and inconsistent (and even positively misleading) under another. In other words, unpartitioned maximum likelihood is not the solution to the challenge raised by this study.

Our analysis does not apply to multi-locus methods that estimate the species tree directly from sequence data—without a gene tree reconstruction step. These include for instance METAL (Dasarathy et al. 2015), SNAPP (Bryant et al. 2012), SVDquartets (Chifman and Kubatko 2014; Chifman and Kubatko 2015; Long and Kubatko 2017), and *BEAST (Heled and Drummond 2009). In particular, METAL has been shown to be consistent on finite-length genes under some assumptions on the MSC (Dasarathy et al. 2015). It is also worthwhile pointing out that our results, while being based on the MSC, are likely to hold more generally for other sources of gene tree discordance, including horizontal gene transfer (HGT). Indeed, as long as rates of HGT are low enough, in the Felsenstein zone similar conclusions about inconsistency will follow for partitioned ML and summary-based methods.

Conclusion

Prior to this study, many coalescent-based species tree estimation methods were assumed to be statistically consistent under this regime, but no proofs had been provided. This study now establishes that many of the standard methods used in phylogenomic species tree estimation are statistically inconsistent.

Moreover, only a small number of methods have been proven to be statistically consistent for bounded L. Some of the summary methods described in Roch and Warnow (2015) for instance are statistically consistent for $L = 1$, but the proofs depend on the strict molecular clock.

When the strict molecular clock assumption does not hold, few methods are statistically consistent for bounded L. METAL (Dasarathy et al. 2015) is one of the few coalescent-based methods that does not require a molecular clock, and that has been proven to be statistically consistent under the MSC+CFN model. It should be noted however that the model of evolution in Dasarathy et al. (2015) allows mutation rates to vary across branches of the species tree, but those rates must be the same across loci, a major constraint. More recently, a log-det distance based extension of METAL has been shown consistent under more general models of substitution and population size variation, as well as certain clock-constrained models allowing variation of rates across loci (Allman et al. 2018). SVDquartets (Chifman and Kubatko 2014), a quartet-based method which has also been formally shown to be statistically consistent (Kubatko 2018), builds on identifiability results (Chifman and Kubatko 2015; Long and Kubatko 2017) that allow mutation rate variation across sites on each gene, not necessarily under a molecular clock. Much remains to be understood about

the important theoretical question of fixed locus length consistency of multi-locus methods in general.

Funding

This work was supported by funding by the U.S. National Science Foundation [CCF-1535977 to T.W.] and [DMS-1149312 (CAREER), DMS-1614242, and CCF-1740707 (TRIPODS) to S.R.].

Appendix
Inconsistency of WSB on a Single Site

Here, we show that the weighted statistical binning pipeline (as defined in Bayzid et al. (2015)) is *not statistically consistent* when $L=1$ for some four-species model tree. Recall that a WSB pipeline proceeds along the following main steps:

1. Estimated gene trees with bootstrap support are provided for every locus. To be more specific, bootstrap support is obtained by re-sampling L sites with replacement from each locus and for each split in the estimated gene tree. For the purposes of the proof below, we assume that exact bootstrap support values are used (i.e., in the limit of infinitely many samples).

2. Then an incompatibility graph is computed for that set of gene trees. The incompatibility graph depends on a parameter B, a proportion in $[0,1]$, as follows: two gene trees are considered to be incompatible if there is a pair of edges, one from each tree, each with bootstrap support strictly greater than B, that are not compatible (see Semple and Steel 2003).

3. This incompatibility graph is used to partition the genes into bins using a minimum vertex coloring heuristic. To balance the bins (i.e., vertex colors), the algorithm processes genes one by one in a particular order and adds each gene to the smallest bin that has no incompatibility with it. When two or more bins have the same smallest size, the algorithm breaks the ties arbitrarily.

4. Next, supergene trees are computed using a fully partitioned maximum likelihood analysis on each bin. These supergene trees are then given to the selected summary method as input, and a species tree is returned. In a weighted statistical binning pipeline, each supergene tree is replicated by the number of genes in its associated bin.

We begin our analysis with a lemma characterizing the contents of the bins in the single-site-per-locus case.

Lemma 3. *Let $\mathcal{S}$ be a model species tree with four species a,b,c,d, and suppose every locus has only one site. In a weighted statistical binning pipeline with bootstrap support threshold $B \geq \frac{1}{3}$, there will at most three bins, one for each of the three possible binary topologies on four leaves, and the bin associated with topology $ab|cd$ will have all the ML-informative genes that support $ab|cd$.*

Proof. Because there is only one site for each gene, the ML-informative genes have bootstrap support of 100%. Hence, no two ML-informative genes can be placed in the same bin if they support different tree topologies. Therefore, for any bin, the ML-informative genes placed in the bin will support the same topology. Also, the ML-uninformative genes produce trees with bootstrap support equal to $\frac{1}{3}$, since every tree topology has equal maximum likelihood score. These genes are therefore considered compatible with every other gene, since the bootstrap support threshold $B \geq \frac{1}{3}$.

Since there are only three tree topologies, the incompatibility graph is the union of a complete 3-partite graph (defined by the ML-informative genes) and a collection of isolated vertices (defined by the ML-uninformative genes). Hence, the incompatibility graph can be 3-colored. Since statistical binning seeks the minimum vertex coloring for the incompatibility graph, it will partition the genes into three bins, with one bin for each binary tree topology. Hence, the ML-informative genes are partitioned into three sets based on the tree topology they support. □

We continue with an analysis of WSB pipelines followed by reasonable summary methods in the case of a single site per gene. The following result implies Theorem 3.

Claim 4. *Suppose every gene has only one site, and let $(\mathcal{S},\Gamma,\theta)$ be a MSC+CFN model species tree with leaves a,b,c,d. Let $B \geq \frac{1}{3}$. There is a choice of branch lengths such that, for all binary trees t on a,b,c,d, the probability that a random gene is ML-informative and supports t is $< \frac{1}{3}$. Moreover, in that case, weighted statistical binning followed by a reasonable summary method will not be statistically consistent.*

Proof. The argument will establish that, under the conditions of the theorem, as the number of genes increases, the WSB binning process will roughly speaking converge to a flat distribution on the three possible tree topologies on a,b,c,d, so that any reasonable summary method will be inconsistent. More formally, our analysis is based on three basic observations.

Observation 1 [One bin for each topology]: By Lemma 3, in a weighted statistical binning pipeline, there will be at most three bins, one for each binary tree topology, and the bin for binary tree t (if non-empty) will have all genes that are ML-informative and support the split for t, and may also have ML-uninformative genes. Furthermore, the ML-uninformative genes can be distributed to the bins arbitrarily, since their bootstrap support is exactly $\frac{1}{3}$ and $B \geq \frac{1}{3}$.

Observation 2 [Partitioned ML = MP]: Since every gene has only one site, the supergene alignment associated to the bin for $ab|cd$ will consist of sites that all split

ab|cd. As pointed out at the beginning of the Section "Analysis of Partitioned ML," in the single site per gene case partitioned maximum likelihood is equivalent to maximum parsimony (Tuffley and Steel 1997). Hence, when a fully partitioned ML analysis is applied to the bin for t, the resultant supergene tree will be the tree t. In a WSB pipeline, the supergene trees for each bin will be replicated as many times as the number of genes in the bin for t. These trees are the newly computed gene trees that will be passed to the reasonable summary method.

Observation 3 [Swamped bins are balanced]: As we explained at the beginning of this section, the division of genes into bins attempts to achieve balanced bins, so that the number of genes in each bin will be as close to the same as possible. Therefore, if the probability that a gene is ML-uninformative is sufficiently high and the number of genes is divisible by 3, then it will be possible to achieve perfectly balanced bins, and the distribution of newly computed gene trees will be the flat distribution. But reasonable summary methods cannot infer the species tree from flat distributions.

It remains to quantify the three observations above and formally establish the failure of statistical consistency. Per Observation 3, we take the number m of genes to be a multiple of 3. For a fixed species tree topology $\mathcal{S}$, as long as all branch lengths are strictly positive, each site pattern has strictly positive probability. Let α_Γ be the smallest such probability under branch lengths Γ. Then the probability that any of the six ML-informative patterns is never observed among the m genes is at most $6(1-\alpha_\Gamma)^m$, which rapidly converges to 0 as $m \to +\infty$. Per Observation 1, with probability at least $1-6(1-\alpha_\Gamma)^m$, there will be one bin for each topology, an event we denote by $E^1_{\Gamma,m}$. On the other hand, by taking all branch lengths small enough, one can ensure that the constant site patterns have overall probability $>2/3$, which in particular implies that, for each binary tree t, the probability of an ML-informative pattern corresponding to t is $<1/3$ as claimed. Let Γ be such a choice of branch lengths and let $\beta_\Gamma < 1/3$ be the largest probability of an ML-informative pattern over all topologies t. For any $\epsilon < 1/3-\beta_\Gamma$, by the law of large numbers, the probability of the following event $E^2_{\Gamma,m}$ goes to 1 as $m \to +\infty$: every bin has at most $(\beta_\Gamma+\epsilon)m$ ML-informative patterns. Because m is divisible by three, the uninformative patterns are distributed among the bins to make them exactly of the same size, in the event $E^1_{\Gamma,m} \cap E^2_{\Gamma,m}$. By Observation 2, in that event, each bin produces exactly $m/3$ supergene trees, with one bin for each topology t. Under a reasonable summary method, when all topologies are represented exactly the same number of times, the output is chosen uniformly at random. In particular, the probability of correct reconstruction is 1/3. Since the probability of the event $E^1_{\Gamma,m} \cap E^2_{\Gamma,m}$ is going to 1 as m increases through multiples of 3, the probability of reconstruction converges to 1/3 on that subsequence, and we have established that the probability of correct reconstruction *cannot* converge to 1—hence statistical consistency fails. □

PROOFS OF THE MAIN RESULTS

We provide detailed proofs of the main claims.

Key Lemmas

Proof of Lemma 1. Recall that we assume $L \geq 2$.

(a) Under our choice of branch lengths, as $\mu \to 0$, the five most common locus site patterns are:

1. *All constant sites:* Every character has the same value on all four taxa (e.g., $x^a = x^b = x^c = x^d =$ 000101). For any such $x \in \mathcal{X}_0$, x occurs with probability

$$Q_0 = \left[\frac{1}{2}\left(1-\rho^3\right)^3(1-\rho)^2 + \mathcal{O}(\rho)\right]^L = \left(\frac{1}{2}\right)^L - \mathcal{O}(\rho),$$

where the first term in the brackets corresponds to the case of no substitution, while the second term accounts for all possibilities with at least one substitution. For convenience, we denote the expression in brackets—the probability of a single site being identical on all four taxa—as q_0.

2. *One singleton site on a or c:* All sites are constant except for one, on which either a or c is different from all others (e.g., $x^a = 01\ldots$, $x^b = x^c = x^d = 11\ldots$). Any data set with this locus site pattern occurs with probability

$$Q_{11} = q_0^{L-1}\left[\frac{1}{2}(1-\rho^3)^3(1-\rho)\rho + \mathcal{O}(\rho^2)\right] = \mathcal{O}(\rho),$$

where the first term in the brackets corresponds to the case of a single substitution along the edge leading to the differing taxon, while the second term accounts for all possibilities involving at least two substitutions.

3. *Two identical singleton sites on a or c:* All sites are constant except for two, each of which has the same taxon a or c different from the others (e.g., $x^a =$ 001…, $x^b = x^c = x^d = 111\ldots$). Any data set with this locus site pattern occurs with probability

$$Q_{2=} = q_0^{L-2}\left[\frac{1}{2}(1-\rho^3)^3(1-\rho)\rho + \mathcal{O}(\rho^2)\right]^2 = \mathcal{O}(\rho^2),$$

which follows from the same computation as in the one singleton case.

4. *Two different singleton sites on a and c:* All sites are constant except for two, one of which has a different character on a and the other a different character on c (e.g., $x^a = 100\ldots$, $x^c = 010\ldots$, $x^b = x^d = 000\ldots$). Any data set with this locus site pattern occurs with probability

$$Q_{2\neq} = q_0^{L-2}\left[\frac{1}{2}(1-\rho^3)^3(1-\rho)\rho + \mathcal{O}(\rho^2)\right]^2 = \mathcal{O}(\rho^2),$$

which follows from the same computation as in the one singleton case.

5. *One site with a 2/2-split ac|bd:* $L-1$ sites are constant with a single site having a and c different from b and d (e.g., $x^a = x^c = 100\ldots$, $x^b = x^d = 000\ldots$). Any data set with this locus site pattern occurs with probability

$$Q_{12} = q_0^{L-1}\left[\frac{1}{2}(1-\rho^3)^3\rho^2 + \mathcal{O}(\rho^3)\right] = \left(\frac{1}{2}\right)^L \rho^2 + \mathcal{O}(\rho^3), \tag{A.1}$$

where the first term in the brackets corresponds to the case of substitutions along the edges leading to the differing taxa, while the second term accounts for all possibilities with at least one substitution along the other edges.

Any remaining locus site pattern must include either a change along one of the short branches, which involves multiplication by ρ^3, or three changes along one of the long branches, which also means multiplication by ρ^3. Thus all x in $\widetilde{\mathcal{X}}$ have probability $\mathcal{O}(\rho^3)$. That concludes the proof of the claim in (a).

(b) It remains to prove (b). For each locus site pattern, we will put an upper bound on the maximum of the likelihood function for topology $\mathcal{T}^0 = ab|cd$, and show that in every case the alternate topology $\mathcal{T}^* = ac|bd$ has maximum likelihood greater than or equal to this upper bound, and in at least one case is strictly greater.

Some remarks about notation first. Note that the labels we have used for the branch lengths of $\mathcal{T}^0$ can be used similarly regardless of the topology of the tree: λ_m represents the middle branch in any topology, and the others represent the branch leading to their respective taxon. Also we use Λ and Φ interchangeably, where Φ is the corresponding collection of ϕ-parameters as defined in (2). Finally, we will use the following property of the ϕ-parameterization (Semple and Steel 2003): the ϕ's multiply along paths; indeed, we have for instance,

$$\begin{aligned}
&\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi^0)}[x_1^a \neq x_1^b] \\
&= (1-p_a^0)p_b^0 + p_a^0(1-p_b^0) \\
&= \frac{1}{2}(1+\phi_a^0)\frac{1}{2}(1-\phi_b^0) + \frac{1}{2}(1-\phi_a^0)\frac{1}{2}(1+\phi_b^0) \\
&= \frac{1}{2}(1-\phi_a^0\phi_b^0).
\end{aligned} \tag{A.2}$$

Finally, because by inclusion the probability of observing $\chi_{\cdot 1}$ is at most the probability of observing χ_{a1}, which is simply $\left(\frac{1}{2}\right)^L$ by independence of the sites, we have

$$\sup_{\Lambda} \ell(\mathcal{T}, \Lambda, \chi_{\cdot 1}) \leq \log\left(\frac{1}{2}\right)^L = -L\log 2. \tag{A.3}$$

We divide up the proof of by locus site pattern.

1. *All constant sites:* Recall from (A.3) that, for any $\mathcal{T}$ (and, in particular, for $\mathcal{T}^0$),

$$\sup_{\Lambda} \ell(\mathcal{T}, \Lambda, x) \leq -L\log 2.$$

For $x \in \mathcal{X}_0$, that can always (in particular, for $\mathcal{T}^*$) be achieved by setting all branch lengths to 0.

2. *One singleton site on a or c:* Without loss of generality, assume the non-constant site is site 1 and that it has $(x_1^a, x_1^b, x_1^c, x_1^d) = (1,0,0,0)$. Assume also that $(x_i^a, x_i^b, x_i^c, x_i^d) = (0,0,0,0)$ for all $i = 2,\ldots,L$. We can put the following upper bound on the likelihood function for $\mathcal{T}^0$. Letting $\phi_{ab} = \phi_a\phi_b$ and using (A.2), we have

$$\begin{aligned}
&\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=1, x_1^b=x_1^c=x_1^d=0\right) \\
&\quad\times \mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_i^a=x_i^b=x_i^c=x_i^d=0\right)^{L-1} \\
&\leq \mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=1\neq x_1^b\right) \\
&\quad\times \mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_i^a=0=x_i^b\right)^{L-1} \\
&\leq \frac{1}{2}\left(\frac{1-\phi_{ab}}{2}\right)\left[\frac{1}{2}\left(\frac{1+\phi_{ab}}{2}\right)\right]^{L-1},
\end{aligned} \tag{A.4}$$

where the first inequality follows by inclusion. To derive our upper bound, we maximize the expression on the last line as a function of ϕ_{ab}. Taking the log, differentiating and equating to 0, we get

$$\frac{-1}{1-\phi_{ab}} + (L-1)\frac{1}{1+\phi_{ab}} = 0$$

that is, $\phi_{ab}=\frac{L-2}{L}$. Plugging this back above, we get the upper bound

$$\sup_{\Phi}\ell(\mathcal{T}^0,\Phi,x)$$
$$\le L\log\left(\frac{1}{2}\right)+\log\left(\frac{1}{L}\right)+(L-1)\log\left(1-\frac{1}{L}\right).$$

On the other hand, for $\mathcal{T}^*$ (or, in fact, any topology), setting $\lambda_b=\lambda_c=\lambda_d=\lambda_m=0$ and λ_a so that $p_a=\frac{1}{L}$, we get the matching bound

$$\mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_1^a=1,x_1^b=x_1^c=x_1^d=0\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_i^a=x_i^b=x_i^c=x_i^d=0\right)^{L-1}$$
$$=\frac{1}{2}\left(\frac{1}{L}\right)\left[\frac{1}{2}\left(1-\frac{1}{L}\right)\right]^{L-1},$$

which establishes the required lower bound on $\sup_\Phi\ell(\mathcal{T}^*,\Phi,x)$.

3. *Two identical singleton sites on a or c:* For this locus site pattern, the argument is identical to the previous locus site pattern when $L\ge4$, with the difference that the exponents in (A.4) are 2 and $L-2$, and accordingly throughout, giving an optimal ϕ_{ab} of $\frac{L-4}{L}$ and the upper bound $L\log(1/2)+2\log\left(\frac{2}{L}\right)+(L-2)\log\left(1-\frac{2}{L}\right)$. This can likewise be achieved with topology $\mathcal{T}^*$ (or, in fact, any topology) if $\lambda_b=\lambda_c=\lambda_d=\lambda_m=0$ and λ_a is set so that $p_a=\frac{2}{L}$. When $L=2$ or $L=3$, the optimal ϕ_{ab} is 0 and the upper bound is $2L\log(1/2)$. This can likewise be achieved with topology $\mathcal{T}^*$ (or, in fact, any topology) if $\lambda_b=\lambda_c=\lambda_d=\lambda_m=0$ and λ_a is set so that $p_a=\frac{1}{2}$.

4. *Two different singleton sites on a and c:* Assume that $(x_1^a,x_1^b,x_1^c,x_1^d)=(1,0,0,0)$, $(x_1^a,x_1^b,x_1^c,x_1^d)=(0,0,1,0)$, and $(x_i^a,x_i^b,x_i^c,x_i^d)=(0,0,0,0)$ for all $i=3,\dots,L$, without loss of generality. (Recall that the case of two different singletons not involving a and c has negligible probability of being observed by part (a) and is therefore not considered here.) We will use the following property of the CFN model: on $\mathcal{T}^0$, because the path joining a,b and the path joining c,d are disjoint, the event $\{x_1^c=x_1^d\}$ is independent of the states x_1^a and x_1^b. This is immediate by the symmetry of the CFN model and the Markov property (Semple and Steel 2003). (Indeed, conditioning on the state at f has no effect on the agreement between c and d.) Using this fact as well as inclusion and (A.2), we get

$$\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=1,x_1^b=x_1^c=x_1^d=0\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^c=1,x_1^a=x_1^b=x_1^d=0\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_i^a=x_i^b=x_i^c=x_i^d=0\right)^{L-2}$$
$$\le\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=1\ne x_1^b,x_1^c=x_1^d\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=0=x_1^b,x_1^c\ne x_1^d\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_i^a=0=x_i^b,x_i^c=x_i^d\right)^{L-2}$$
$$=[\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_1^a=1\ne x_1^b)$$
$$\times P_{x\sim(\mathcal{T}^0,\Phi)}(x_1^c=x_1^d)]$$
$$\times[\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_1^a=0=x_1^b)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_1^c\ne x_1^d)]$$
$$\times[\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_i^a=0=x_i^b)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_i^c=x_i^d)]^{L-2}$$
$$=\frac{1}{2}\left(\frac{1-\phi_{ab}}{2}\right)\left(\frac{1+\phi_{cd}}{2}\right)$$
$$\times\frac{1}{2}\left(\frac{1+\phi_{ab}}{2}\right)\left(\frac{1-\phi_{cd}}{2}\right)$$
$$\times\left[\frac{1}{2}\left(\frac{1+\phi_{ab}}{2}\right)\left(\frac{1+\phi_{cd}}{2}\right)\right]^{L-2}$$
$$=\left(\frac{1}{2}\right)^L\left(\frac{1-\phi_{ab}}{2}\right)\left(\frac{1+\phi_{ab}}{2}\right)^{L-1}$$
$$\times\left(\frac{1-\phi_{cd}}{2}\right)\left(\frac{1+\phi_{cd}}{2}\right)^{L-1},$$

where $\phi_{ab}=\phi_a\phi_b$ and $\phi_{cd}=\phi_c\phi_d$. Maximizing this last expression over ϕ_{ab} and ϕ_{cd} proceeds as in (A.4). We then get the upper bound

$$\sup_{\Phi}\ell(\mathcal{T}^0,\Phi,x)$$
$$\le L\log\left(\frac{1}{2}\right)+2\log\left(\frac{1}{L}\right)+2(L-1)\log\left(1-\frac{1}{L}\right).$$

On the other hand, for $\mathcal{T}^*$ (or, in fact, any topology), setting $\lambda_b=\lambda_d=\lambda_m=0$ and $\lambda_a=\lambda_c$ so that $p_a=p_c=\frac{1}{L}$, we get

$$\mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_1^a=1,x_1^b=x_1^c=x_1^d=0\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_1^c=1,x_1^a=x_1^b=x_1^d=0\right)$$

$$\times \mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_i^a=x_i^b=x_i^c=x_i^d=0\right)^{L-2}$$
$$=\frac{1}{2}\left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)\times\frac{1}{2}\left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)$$
$$\times\left[\frac{1}{2}\left(1-\frac{1}{L}\right)^2\right]^{L-2},$$

which establishes the required lower bound on $\sup_\Phi \ell(\mathcal{T}^*,\Phi,x)$.

5. *One site with a 2/2-split ac|bd:* Without loss of generality, we assume that $(x_1^a,x_1^b,x_1^c,x_1^d)=(1,0,1,0)$ and $(x_i^a,x_i^b,x_i^c,x_i^d)=(0,0,0,0)$ for all $i=2,\ldots,L$. Arguing as in the previous case,

$$\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=x_1^c=1,x_1^b=x_1^d=0\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_i^a=x_i^b=x_i^c=x_i^d=0\right)^{L-1}$$
$$\leq\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_1^a=1\neq x_1^b,x_1^c\neq x_1^d\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}\left(x_i^a=0=x_i^b,x_i^c=x_i^d\right)^{L-1}$$
$$=[\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_1^a=1\neq x_1^b)\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_1^c\neq x_1^d)]$$
$$\times[\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_i^a=0=x_i^b)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^0,\Phi)}(x_i^c=x_i^d)]^{L-1}$$
$$=\frac{1}{2}\left(\frac{1-\phi_{ab}}{2}\right)\left(\frac{1-\phi_{cd}}{2}\right)$$
$$\times\left[\frac{1}{2}\left(\frac{1+\phi_{ab}}{2}\right)\left(\frac{1+\phi_{cd}}{2}\right)\right]^{L-1}$$
$$=\left(\frac{1}{2}\right)^L\left(\frac{1-\phi_{ab}}{2}\right)\left(\frac{1+\phi_{ab}}{2}\right)^{L-1}$$
$$\left(\frac{1-\phi_{cd}}{2}\right)\left(\frac{1+\phi_{cd}}{2}\right)^{L-1},$$

where, again, $\phi_{ab}=\phi_a\phi_b$ and $\phi_{cd}=\phi_c\phi_d$. This bound matches the bound we obtained in the previous case. Hence, we once again get the upper bound

$$\sup_\Phi \ell(\mathcal{T}^0,\Phi,x)$$
$$\leq L\log\left(\frac{1}{2}\right)+2\log\left(\frac{1}{L}\right)+2(L-1)\log\left(1-\frac{1}{L}\right).$$

However, in this case, we claim that the maximum likelihood under $\mathcal{T}^*$ is strictly greater. Indeed, letting $\lambda_a=\lambda_b=\lambda_c=\lambda_d=0$ and setting λ_m such that $p_m=\frac{1}{L}$, we get

$$\mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_1^a=x_1^c=1,x_1^b=x_1^d=0\right)$$
$$\times\mathbf{P}_{x\sim(\mathcal{T}^*,\Phi)}\left(x_i^a=x_i^b=x_i^c=x_i^d=0\right)^{L-1}$$
$$=\frac{1}{2}\left(\frac{1}{L}\right)\times\left[\frac{1}{2}\left(1-\frac{1}{L}\right)\right]^{L-1},$$

so

$$\sup_\Phi \ell(\mathcal{T}^*,\Phi,x)$$
$$\geq L\log\left(\frac{1}{2}\right)+\log\left(\frac{1}{L}\right)+(L-1)\log\left(1-\frac{1}{L}\right).$$

Therefore

$$\sup_\Phi \ell(\mathcal{T}^*,\Phi,x)-\sup_\Phi \ell(\mathcal{T}^0,\Phi,x)$$
$$\geq-\log\left(\frac{1}{L}\right)-(L-1)\log\left(1-\frac{1}{L}\right)$$
$$=:K_{12}>0,$$

where the last equality is a definition. Positivity of K_{12} can be seen, for instance, by noticing that dividing it by L gives the entropy of a 2-state random variable.

In all the above cases, a similar argument still applies if one replaces $\mathcal{T}^0$ with $\mathcal{T}^1$ (by exchanging the roles of b and d throughout). That concludes the proof of the claim in (b). □

Proof of Lemma 2. (a) The expressions for $Q^s_{ac|bd}$, $Q^s_{ab|cd}$ and $Q^s_{ad|bc}$ come from taking $L=1$ in Lemma 1 (a) and raising to the power L. Specifically, it was shown in (A.1) that observing a single site splitting a,c from b,d has probability of the form $(1/2)\rho^2+O(\rho^3)$. Since a saturated locus contains L sites with the same probability, we raise this expression to the power L to obtain

$$Q^s_{ac|bd}=\left(\frac{1}{2}\right)^L\rho^{2L}+\mathcal{O}(\rho^{2L+1}).$$

Similarly, it was observed in the proof of Lemma 1 (a) that observing a single site splitting a,b from c,d (or a,d from b,c) has probability $O(\rho^3)$. Raising to the power L gives $Q^s_{ab|cd},Q^s_{ad|bc}=\mathcal{O}(\rho^{3L})$.

For the first part of the claim, we consider several cases.

– Suppose that sequence data set x contains at least one uninformative character (i.e., a constant site or a singleton). Then, in computing bootstrap supports, there is probability at least $(1/L)^L$ of resampling a data set containing *only* that particular uninformative site. We have shown in the proof of Lemma 1 (b) (see cases 1 and 2 with $L=1$) that all topologies have an equal ML score

on such a site and therefore on such a resampled data set (since the probability of observing a data set of this type is the probability of observing a single site to the power L). Hence each topology is supported with probability 1/3. Hence the bootstrap support for the ML-optimizer for x is at most $1-(2/3)(1/L)^L \le B$ and x is rejected by WSB*.

– Suppose that sequence data set x contains two different informative characters (i.e., two different splits). One of those splits is incompatible with the ML-optimizer (possibly random) for x. Then, in computing bootstrap supports, there is probability at least $(1/L)^L$ of resampling a data set containing only that incompatible split. From the argument in Lemma 1 (b) again (case 5 with $L=1$), the incompatible split is then the ML-optimizer of such a resampled data set. Hence the bootstrap support for the ML-optimizer for x is at most $1-(1/L)^L < 1-(2/3)(1/L)^L \le B$ and x is rejected by WSB*.

– Suppose finally that sequence data set x contains only characters equivalent to a given split. Then all resampled data sets are saturated for that split as well. From the argument in Lemma 1 (b) again (case 5 with $L=1$), that split is the unique ML-optimizer for x. Hence the bootstrap support for the ML-optimizer for x is $1 > B$ and x is passed along by WSB* to the summary method.

(b) This was proved in (a). □

Partitioned ML on CFN Model

Proof of Claim 1. Using Lemma 1, we are now ready to prove Claim 1.

We first show that, for a fixed topology, as the number of loci grows to infinity the maximum likelihood value converges almost surely to the expected value of the maximum likelihood value on a single locus.

Lemma 4 (Convergence of the partitioned log-likelihood)**.** *Let $\mathcal{T}'$ be a fixed topology on the four taxa with branch lengths Λ'. Let also $\mathcal{T}''$ be a fixed topology on the four taxa (possibly, but not necessarily, equal to $\mathcal{T}'$). If the length-L locus sequence data sets $\chi_{\cdot j}$, $j=1,2,\ldots$, are generated under the CFN model on $(\mathcal{T}',\Lambda')$, then it holds that*

$$\frac{1}{m}\sum_{j=1}^{m}\sup_{\Lambda_j}\ell(\mathcal{T}'',\Lambda_j,\chi_{\cdot j}) \to \mathbf{E}_{\chi_{\cdot 1}\sim(\mathcal{T}',\Lambda')}\left[\sup_{\Lambda}\ell(\mathcal{T}'',\Lambda,\chi_{\cdot 1})\right] \in [-4L\log 2, -L\log 2], \tag{A.5}$$

almost surely as $m\to+\infty$. Above, the subscript $\chi_{\cdot 1}\sim(\mathcal{T}',\Lambda')$ indicates that the expectation is taken over a single locus under the CFN model on $(\mathcal{T}',\Lambda')$.

Proof. For a given topology and data set there is a unique maximum likelihood value, though the branch lengths at which it is attained may not themselves be unique. For any given locus j, there are a finite number of four-sequence data sets χ_j of length L that can occur under the CFN model. As the number of loci approaches infinity, the frequency of each data set approaches its expected value by the Strong Law of Large Numbers (SLLN) (see e.g., Durrett 1996). To check that the conditions of the SLLN are satisfied, note that the log-likelihood is non-positive. In fact, by taking branch lengths to $+\infty$ under the CFN model, we have for any topology $\mathcal{T}$ on $\{a,b,c,d\}$ and any locus data set $\chi_{\cdot 1}$

$$\sup_{\Lambda}\ell(\mathcal{T},\Lambda,\chi_{\cdot 1}) \ge \log\left(\frac{1}{2}\right)^{4L} = -4L\log 2. \tag{A.6}$$

On the other hand, because by inclusion the probability of observing $\chi_{\cdot 1}$ is at most the probability of observing χ_{a1}, which is simply $\left(\frac{1}{2}\right)^L$ by independence of the sites, we also have

$$\sup_{\Lambda}\ell(\mathcal{T},\Lambda,\chi_{\cdot 1}) \le \log\left(\frac{1}{2}\right)^{L} = -L\log 2. \tag{A.7}$$

So the expectation on the RHS of (A.5) lies in the interval $[-4L\log 2, -L\log 2]$. □

Hence, in view of Lemma 4, our goal is to show that there is $\rho>0$ small enough such that the expected log-likelihood under $(\mathcal{T}^0,\Lambda^0)$ is higher for $\mathcal{T}^*$ than it is for $\mathcal{T}^0$ or $\mathcal{T}^1$. That is, it suffices to establish the following claim.

Lemma 5 (Expected locus-wise maximum likelihood on a fixed topology: key inequality)**.** *There exists $\rho>0$ such that*

$$\mathbf{E}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\sup_{\Lambda}\ell(\mathcal{T}^0,\Lambda,\chi_{\cdot 1})\right] < \mathbf{E}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\sup_{\Lambda}\ell(\mathcal{T}^*,\Lambda,\chi_{\cdot 1})\right], \tag{A.8}$$

and

$$\mathbf{E}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\sup_{\Lambda}\ell(\mathcal{T}^1,\Lambda,\chi_{\cdot 1})\right] < \mathbf{E}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\sup_{\Lambda}\ell(\mathcal{T}^*,\Lambda,\chi_{\cdot 1})\right]. \tag{A.9}$$

Proof. Let $\mathcal{X}$ be the set of all possible single-locus data sets. To prove Lemma 5, we expand the expectations

in (A.8) over $\mathscr{X}$. In other words, we seek to show that

$$\sum_{x\in\mathscr{X}} \mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}[\chi_{\cdot 1}=x] \times\left\{\sup_{\Lambda}\ell(\mathcal{T}^*,\Lambda,x)-\sup_{\Lambda}\ell(\mathcal{T}^0,\Lambda,x)\right\}>0. \quad (A.10)$$

We then use Lemma 1 as follows. By (a),

$$\sum_{x\in\widetilde{\mathscr{X}}} \mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}[\chi_{\cdot 1}=x] \times\left|\sup_{\Lambda}\ell(\mathcal{T}^*,\Lambda,x)-\sup_{\Lambda}\ell(\mathcal{T}^0,\Lambda,x)\right|=\mathcal{O}(\rho^3). \quad (A.11)$$

Indeed, any locus site pattern in $\widetilde{\mathscr{X}}$ has probability $\mathcal{O}(\rho^3)$. Moreover, recall from (A.6) and (A.7) that the expression in absolute value is bounded by $3L\log 2$. In addition, by (a) and (b), we then arrive at

$$\begin{aligned}&\sum_{x\in\mathscr{X}} \mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}[\chi_{\cdot 1}=x] \times\left\{\sup_{\Lambda}\ell(\mathcal{T}^*,\Lambda,x)-\sup_{\Lambda}\ell(\mathcal{T}^0,\Lambda,x)\right\}\\ &\geq K_{12}\left\{\left(\frac{1}{2}\right)^L\rho^2+\mathcal{O}(\rho^3)\right\}+\mathcal{O}(\rho^3)>0,\end{aligned}$$

for $\rho>0$ small enough.

The same argument applies for (A.9). □

Combining Lemmas 4 and 5 gives Claim 1. □

Reasonable Summary Methods on CFN Model

Proof of Claim 2. Using Lemma 1, we are now ready to prove Claim 2.

By definition of a reasonable summary method, on a four-taxon data set, $\mathcal{A}$ outputs the most common quartet topology (breaking ties uniformly at random). We also assume that for genes with multiple optimal ML topologies, a highest scoring topology is picked uniformly at random. We denote by $\hat{\mathcal{R}}(\chi_{\cdot j})$ the ML gene tree on the j-th locus sequence data set. The law of large numbers immediately gives the following.

Lemma 6 (Convergence of frequencies). *Let $\mathcal{T}'$ be a fixed topology on the four taxa with branch lengths Λ'. Let also $\mathcal{T}''$ be a fixed topology on the four taxa (possibly, but not necessarily, equal to $\mathcal{T}'$). If the length-L locus sequence data sets $\chi_{\cdot j}$, $j=1,2,\ldots$, are generated under the CFN model on $(\mathcal{T}',\Lambda')$, then it holds that*

$$\frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\left[\hat{\mathcal{R}}(\chi_{\cdot j})=\mathcal{T}''\right]\longrightarrow \mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}',\Lambda')}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}''\right],$$

almost surely as $m\to+\infty$. Above, $\mathbf{1}[\mathcal{E}]$ is 1 if event $\mathcal{E}$ occurs, and 0 otherwise.

Hence, in view of Lemma 6, our goal is to show that there is $\rho>0$ small enough such that, under $(\mathcal{T}^0,\Lambda^0)$, $\mathcal{T}^*$ is more likely to be the ML gene tree topology than $\mathcal{T}^0$ or $\mathcal{T}^1$. That is, it suffices to establish the following claim.

Lemma 7 (Locus-wise maximum likelihood on a fixed topology: key inequality). *There exists $\rho>0$ such that*

$$\begin{aligned}&\mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}^0\right]\\ &<\mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}^*\right], \quad (A.12)\end{aligned}$$

and

$$\begin{aligned}&\mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}^1\right]\\ &<\mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}^*\right]. \quad (A.13)\end{aligned}$$

Proof. By Lemma 1 (b), for all $x\in\mathscr{X}_0\cup\mathscr{X}_{11}\cup\mathscr{X}_{2=}\cup\mathscr{X}_{2\neq}$, all three topologies are ML-optimal, while for all $x\in\mathscr{X}_{12}$, $\mathcal{T}^*$ alone is ML-optimal. Moreover, by Lemma 1 (a), all other patterns are negligible. Hence, we get

$$\begin{aligned}&\mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}^*\right]\\ &\geq\frac{1}{3}2^L\left[Q_0+2LQ_{11}+2\binom{L}{2}Q_{2=}+L(L-1)Q_{2\neq}\right]\\ &+2^LQ_{12}+\mathcal{O}(\rho^3),\end{aligned}$$

while

$$\begin{aligned}&\mathbf{P}_{\chi_{\cdot 1}\sim(\mathcal{T}^0,\Lambda^0)}\left[\hat{\mathcal{R}}(\chi_{\cdot 1})=\mathcal{T}^0\right]\\ &\leq\frac{1}{3}2^L\left[Q_0+2LQ_{11}+2\binom{L}{2}Q_{2=}+L(L-1)Q_{2\neq}\right]\\ &+\mathcal{O}(\rho^3),\end{aligned}$$

and similarly for $\mathcal{T}^1$. The result then follows from the fact that

$$Q_{12}=\left(\frac{1}{2}\right)^L\rho^2+\mathcal{O}(\rho^3).$$

□

Combining Lemmas 6 and 7 gives Claim 2. □

WSB Pipeline on CFN Model*

Proof of Claim 3. Using Lemma 2, we are now ready to prove Claim 3.

We begin with two basic results.

Lemma 8. *In a WSB* pipeline with bootstrap support threshold $B\geq 1-\frac{2}{3}\left(\frac{1}{L}\right)^L$, there will be at most three bins (one for each of the three possible binary topologies on four leaves), and the bin associated with topology ab|cd will have all the saturated genes that support ab|cd (and similarly for ac|bd and ad|bc).*

Proof. By Lemma 2 (b), the genes saturated for a given split have bootstrap support of 100%. Hence, no two such genes can be placed in the same bin if they support different tree topologies. Therefore, for any bin, the genes placed in the bin will support the same topology. By Lemma 2 (a), all other genes are discarded.

Since there are only three tree topologies, the incompatibility graph is the union of a collection of 3-partite graphs (each 3-partite graph defined by a split-saturated gene). Hence, the incompatibility graph can be 3-colored. Since statistical binning seeks the minimum vertex coloring for the incompatibility graph, it will partition the genes into three bins, with one bin for each binary tree topology. Hence, the split-saturated genes are partitioned into three sets based on the tree topology they support. □

Lemma 9. *(Lemma 2 from Bayzid et al. (2015):) Let S be a set of taxa, and let S_i be a set of DNA sequences for S, with $i=1,2,\ldots p$. Suppose that tree topology t is an optimal solution for GTR maximum likelihood for each S_i (allowing various GTR parameters for different $i=1,2,\ldots p$). Then t will be an optimal solution to a fully partitioned GTR maximum likelihood analysis on a concatenation of $S_1, S_2, \ldots, S_p$.*

Corollary 1. *The set of newly computed gene trees computed during a WSB* pipeline has the same distribution as the original set of ML gene trees obtained from the split-saturated genes.*

Proof. By Lemma 8, the split-saturated genes are partitioned into three bins for the different tree topologies. By Lemma 9, fully partitioned maximum likelihood on each supergene alignment produces the tree topology associated with the bin. In a WSB* pipeline, the supergene tree for each bin is copied by as many genes as in the bin. Hence, the distribution defined by the newly computed gene trees is identical to the distribution defined by original ML gene trees. □

The rest of the argument follows as in the proof of Claim 2. By Lemma 2 (a), under our four-taxon model species tree with topology $ab|cd$, the most probable estimated quartet tree on split-saturated genes is $ac|bd$. After removing all the loci that are not split-saturated, we are left only with genes that split 2/2. As the number of loci increases, with probability going to 1 the most frequent estimated quartet tree will be $ac|bd$. Therefore by Corollary 1, in a WSB* pipeline with bootstrap support threshold $B \geq 1-\frac{2}{3}\left(\frac{1}{L}\right)^L$, the most frequent supergene tree computed by weighted statistical binning is identical to the most frequent estimated quartet tree in the input, and will converge to $ac|bd$ as the number of loci increases by the law of large numbers. Hence, WSB* pipelines followed by reasonable summary methods will be positively misleading under this model. □

Extension to MSC+CFN Model

In this section, we extend the main claims to the MSC+CFN model. The key idea is to choose a species tree that is highly likely to produce, on any given locus, sequence data whose distribution is close to that of a fixed gene tree in the Felsenstein zone.

When a character of length L, $\chi_{\cdot j}$, is generated under the CFN model on $(\mathcal{T}, \Lambda)$, we write $\chi_{\cdot j} \sim \mathcal{D}_g^L[\mathcal{T}, \Lambda]$. Formally, $\mathcal{D}_g^L[\mathcal{T}, \Lambda]$ is a probability distribution over sequence data sets in$\{0,1\}^{n\times L}$, that is, containing n sequences of length L taking values in $\{0,1\}$, where n is the number of leaves in $\mathcal{T}$. The subscript g is meant to refer to the fact that this is a distribution obtained from a single gene tree.

We also consider sequence data sets generated by the MSC+CFN model. Consider a species tree $(\mathcal{S}, \Gamma, \theta)$ with n leaves. Each gene $j=1,\ldots,m$ has a genealogical history represented by its gene tree $\mathcal{T}_j$ distributed according to the following process: looking backwards in time, on each branch e of the species tree, the coalescence of any two lineages is exponentially distributed with rate $2/\theta_e$, independently from all other pairs; whenever two branches merge in the species tree, we also merge the lineages of the corresponding populations, that is, the coalescence proceeds on the *union* of the lineages. More specifically, the probability density of a realization of this model for m independent genes is

$$\prod_{j=1}^{m}\prod_{e\in E}\exp\left(-\binom{O_j^e}{2}\left[\gamma_j^{e,O_j^e+1}-\gamma_j^{e,O_j^e}\right]\frac{2}{\theta_e}\right) \times \prod_{\ell=1}^{I_j^e-O_j^e}\exp\left(-\binom{\ell}{2}\left[\gamma_j^{e,\ell}-\gamma_j^{e,\ell-1}\right]\frac{2}{\theta_e}\right),$$

where, for gene j and branch e, I_j^e is the number of lineages entering e, O_j^e is the number of lineages exiting e, and $\gamma_j^{e,\ell}$ is the ℓ^{th} coalescence time in e; for convenience, we let $\gamma_j^{e,0}$ and $\gamma_j^{e,I_j^e-O_j^e+1}$ be respectively the divergence times of e and of its parent population (which depend on Γ).

When a character of length L, $\chi_{\cdot j}$, is generated under the MSC+CFN model on $(\mathcal{S}, \Gamma, \theta)$, we write $\chi_{\cdot j} \sim \mathcal{D}_s^L[\mathcal{S}, \Gamma, \theta]$. Formally, $\mathcal{D}_s^L[\mathcal{S}, \Gamma, \theta]$ is a probability distribution over sequence data sets in $\{0,1\}^{n\times L}$, where n is the number of leaves in $\mathcal{S}$. The subscript s is meant to refer to the fact that this is a distribution obtained from the MSC on a species tree.

As in the main text, fix $\mathcal{T}^0$ to be the four-taxon topology $ab|cd$ on $\{a,b,c,d\}$ and let Λ^0 denote a vector of branch lengths on $\mathcal{T}^0$. Denote the endpoint of the middle edge on the ab side as e, and on the cd side as f. For this tree, denote the length of branch ae as λ_a^0, be as λ_b^0, cf as λ_c^0, df as λ_d^0 and ef as λ_m^0. For a branch length λ, recall that

we also use the parameterization $\phi = -\frac{1}{2}\log\lambda$ in terms of which the probability of a change along this branch is

$$p = \frac{1}{2}\left(1 - e^{-2\lambda}\right) = \frac{1}{2}(1-\phi),$$

and the probability of no change is $q = \frac{1}{2}(1+\phi)$. We choose Λ^0 to construct a Felsenstein zone tree where, for a parameter $\rho > 0$, $p_a^0 = p_c^0 = \rho$ and $p_b^0 = p_d^0 = p_m^0 = \rho^3$. Note that for any $\rho > 0$, we can set $\lambda_a^0 = \lambda_c^0 = -\frac{1}{2}\log(1-2\rho)$ and $\lambda_b^0 = \lambda_d^0 = \lambda_m^0 = -\frac{1}{2}\log(1-2\rho^3)$ to satisfy this relationship. We also denote the alternate topologies by $\mathcal{T}^* = ac|bd$ and $\mathcal{T}^1 = ad|bc$.

The next claim shows that we can choose a species tree that is highly likely to produce locus-wise sequence data whose distribution is close to that of the desired gene tree $(\mathcal{T}^0, \Lambda^0)$. Note, in particular, that for generic choices of Γ and θ the output gene trees are not ultrametric (see Semple and Steel (2003) for a definition)—and therefore can lie in the Felsenstein zone. Below, one should see the distribution $\mathcal{R}$ as a small approximation error from the desired gene tree distribution.

Claim 5 (Species tree in the Felsenstein zone)**.** *For all $\epsilon > 0$, there is a species tree $(\mathcal{S}^0, \Gamma^0, \theta^0)$ with leaves $\{a,b,c,d\}$ and a probability distribution $\mathcal{R}$ over $\{0,1\}^{4\times L}$ such that*

$$\mathcal{D}_s^L[\mathcal{S}^0, \Gamma^0, \theta^0] = (1-\epsilon)\mathcal{D}_g^L[\mathcal{T}^0, \Lambda^0] + \epsilon\mathcal{R}.$$

Proof. The idea of the proof is simple: we choose a species tree $(\mathcal{S}^0, \Gamma^0, \theta^0)$ that is close to the desired gene tree $(\mathcal{T}^0, \Lambda^0)$. But we make one more crucial observation: each edge e of the species tree has in fact *two* parameters, γ_e^0 and θ_e^0, that jointly control the speed of coalescence on e as well as the amount of substitution on the corresponding edge(s) of the produced gene tree (in a tangled manner; see the definition of the MSC in Section "Evolution Under the MSC"). As a result, we can choose $(\mathcal{S}^0, \Gamma^0, \theta^0)$ to be close to $(\mathcal{T}^0, \Lambda^0)$ while ensuring that the amount of incomplete lineage sorting is arbitrarily negligible.

Formally, we let $\mathcal{S}^0$ be the balanced species tree with split $ab|cd$ and root r. Denote the endpoint of the edge incident to the root on the ab side as e, and on the cd side as f. For this tree, denote the parameters of branch ae as γ_a^0 and θ_a^0, be as γ_b^0 and θ_b^0, cf as γ_c^0 and θ_c^0, df as γ_d^0 and θ_d^0, er as γ_e^0 and θ_e^0, and fr as γ_f^0 and θ_f^0. The branch $r\infty$ above the root r has infinite length and parameter θ_r^0. We take $\theta_a^0 = \theta_b^0 = \theta_c^0 = \theta_d^0 = 1$, $\gamma_a^0 = \lambda_a^0$, $\gamma_b^0 = \lambda_b^0$, $\gamma_c^0 = \lambda_c^0$, $\gamma_d^0 = \lambda_d^0$. Finally, we let $\gamma_e^0 = \gamma_f^0 = \alpha + \lambda_m^0/2$ and $\theta_e^0 = \theta_f^0 = \theta_r^0 = \beta$.

Now take α and β small enough that:

- coalescences in er, fr and $r\infty$ occur within α of e, f, and r, respectively;
- no mutation occurs within α above e, f and r, respectively;

with probability at least $1-\epsilon$. Specifically, for a fixed α, we choose $\beta := \beta(\alpha)$ small enough to ensure that coalescence occurs within α of the divergences e, f and r with probability $1-\epsilon/2$. Then we take α small enough to ensure that no mutations occur within α of the same divergences with probability $1-\epsilon/2$. Conditioned on the event above, the distribution of sequence data sets is precisely $\mathcal{D}_g^L[\mathcal{T}^0, \Lambda^0]$. The result follows. □

We are now ready to prove the main theorems.

Proof of Theorem 1. We take $(\mathcal{S}^0, \Gamma^0, \theta^0)$ as in Claim 5 for $\epsilon > 0$ to be determined below. We think of the first m loci as divided into two subsets: $\mathcal{M}_0^m$ coming from distribution $\mathcal{D}_g^L[\mathcal{T}^0, \Lambda^0]$ and $\mathcal{M}_{\mathcal{R}}^m$ coming from $\mathcal{R}$. By the law of large numbers, we have

$$\frac{|\mathcal{M}_0^m|}{m} \to 1-\epsilon \qquad \text{and} \qquad \frac{|\mathcal{M}_{\mathcal{R}}^m|}{m} \to \epsilon.$$

We then apply the argument in the proof of Claim 1 to the samples in $\mathcal{M}_0^m$ and take ϵ small enough that the contribution of $\mathcal{M}_{\mathcal{R}}^m$ to the partitioned log-likelihood is in the limit $m \to +\infty$ smaller than the expected gap between $\mathcal{T}^*$ and $\mathcal{T}^0$. □

The proofs of Theorems 2 and 4 follow from similar arguments.

References

Allman E., Degnan J., Rhodes J. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62:833–862.

Allman E.S., Long C., Rhodes J.A. 2018. Species tree inference from genomic sequences using the log-det distance. arXiv:1806.04974.

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS One doi:10.1371/journal.pone.0129183.

Bayzid M.S., Warnow T. 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29:2277–84.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Cannon J.T., Vellutini B.C., Smith J., Ronquist F., Jondelius U., Hejnol A. 2016. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530:89–93.

Cavender J.A. 1978. Taxonomy with confidence. Math. Biosci. 40:271?280.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent. Bioinformatics 30:3317–3324.

Chifman J., Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J. Theor. Biol. 374:35–47.

Chou J., Gupta A., Yaduvanshi S., Davidson R., Nute M., Mirarab S., Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16:S2.

Dasarathy G., Nowak R., Roch S. 2015. Data requirement for phylogenetic inference from multiple loci: a new distance method. IEEE/ACM Trans. Comput. Biol. Bioinformatics 12:422–432.

Durrett R. 1996. Probability: theory and examples. 2nd ed. Belmont, CA: Duxbury Press.

Farris J.S. 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22:250–256.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Heled J., Drummond A.J. 2009. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., M. Schneider P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., K. A. Jønsson, Johnson W., K.-P. Koepfli, S. O'Brien, Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., T. C. Glenn, McCormack J., Burt D., Ellegren H., P. Alström, Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York and London: Academic Press, p. 22–132.
Kubatko L. 2018. Personal communication.
Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.
Larget B.R., Kotha S.K., Dewey C.N., Ané, C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910–2911.
Leaché A.D., Rannala B. 2010. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.
Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.
Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.
Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–477.
Long C., Kubatko L. 2017. Identifiability and reconstructibility of species phylogenies under a modified coalescent. arXiv:1701.06871.
Maddison D. 2016. The rapidly changing landscape of insect phylogenetics. Curr. Opin. Insect Sci. 18:77–82.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
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 334:521–4.
Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014a. Statistical binning improves species tree estimation in the presence of gene tree incongruence. Science 6215:1250463.
Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M., Warnow T. 2014b. ASTRAL: Accurate Species TRee ALgorithm. Bioinformatics 30:i541–i548.
Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.
Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspöck U., Aspöck H., Bartel D., Blanke A., Berger S., Böhm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schütte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.
Molloy, E.K., Warnow T. 2017. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.
Mossel, E., Roch S. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Trans. Comput. Biol. Bioinform. 7:166–71.
Neyman J. 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta S., Yackel J., editors. Statistical decision theory and related topics. New York and London: Academic Press, p. 1–27.
Patel, S., Kimball R.T., Braun E.L. 2013. Error in phylogenetic estimation for bushes in the tree of life. J. Phylog. Evol. Biol. 1:110.
Posada D. 2016. Phylogenomics for systematic biology. Syst. Biol. 65:353–356.
Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.
Roch S., Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64:663–676.
Semple C., Steel M.A. 2003. Phylogenetics. Vol. 24. Oxford Lecture Series in Mathematics and Its Applications. Oxford, UK: Oxford University Press.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Lectures on mathematics in the life sciences. Vol. 17. Rhode Island: American Mathematical Society, p. 57–86.
Tuffley C., Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.
Vachaspati P., Warnow T. 2015. ASTRID: Accurate species TRees from internode distances. BMC Genomics 16:S3.
Warnow T. 2015. Concatenation analyses in the presence of incomplete lineage sorting. PLoS Curr. doi: 10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7.
Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., B. Ruhfel R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U.S.A. 111:E4859–E4868.