Theoretical Population Biology 166 (2025) 56–69

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Inconsistency of parsimony under the multispecies coalescent

Daniel A. Rickert [a,b,*], Louis Wai-Tong Fan [a,1,2], Matthew W. Hahn [b,c]

[a] Department of Mathematics, Indiana University, 831 E 3rd St, 47405, Bloomington, IN, USA
[b] Department of Biology, Indiana University, 1001 E 3rd St, 47405, Bloomington, IN, USA
[c] Department of Computer Science, Indiana University, 700 N Woodlawn Ave, 47405, Bloomington, IN, USA

ARTICLE INFO

Dataset link: https://github.com/darickert/exp-branch-lengths

Keywords:
Coalescent
Incomplete lineage sorting
Concatenation
Parsimony
Species tree

ABSTRACT

While it is known that parsimony can be statistically inconsistent under certain models of evolution due to high levels of homoplasy, the consistency of parsimony under the multispecies coalescent (MSC) is less well studied. Previous studies have shown the consistency of concatenated parsimony (parsimony applied to concatenated alignments) under the MSC for the rooted 4-taxa case under an infinite-sites model of mutation; on the other hand, other work has also established the inconsistency of concatenated parsimony for the unrooted 6-taxa case. These seemingly contradictory results suggest that concatenated parsimony may fail to be consistent for trees with more than 5 taxa, for all unrooted trees, or for some combination of the two. Here, we present a technique for computing the expected internal branch lengths of gene trees under the MSC. This technique allows us to determine the regions of the parameter space of the species tree under which concatenated parsimony fails for different numbers of taxa, for rooted or unrooted trees. We use our new approach to demonstrate that while parsimony succeeds in the unrooted 5-taxa case, there are regions of statistical inconsistency for concatenated parsimony for rooted 5+-taxa cases and unrooted 6+-taxa cases. Our results therefore suggest that parsimony is not generally dependable under the MSC.

1. Introduction

One of the major goals of phylogenetics is to describe the relationships among organisms. We suppose the evolutionary relationship among n species or taxa can be described by a rooted, binary, and ultrametric species tree $S = (T_*, \boldsymbol{x})$ with n tips, where T_* denotes the rooted binary topology of the species tree, and $\boldsymbol{x}$ gives the branch lengths of S. The goal is to be able to infer the species tree S, or some component of it, such as the topology T_*, using data available from the tip species.

The most common data used to infer species trees come from DNA sequences. DNA sequences are available from every gene (or locus) in a genome. Coalescent-based models give a probability distribution on the gene tree G that represents the evolutionary history of a given locus among sampled individuals (Kingman, 1982; Hudson, 1990). The gene tree topology, G, at a locus is conditionally random given the species tree, S, when sampled individuals come from different species. The sequence data at this locus is then conditionally random given G, depending on any mutation events that have occurred on it. It is well-understood that the gene tree can be discordant (i.e. have internal branches that disagree) with the species tree for a number of biological reasons, such as introgression or horizontal gene transfer (see for instance Maddison, 1997; Edwards, 2009). However, arguably the most well-studied cause of gene tree discordance is incomplete lineage sorting (ILS), in which lineages in a population do not coalesce until entering a further ancestral population. In our analysis, we take ILS to be the sole cause of gene tree discordance, owing to the simplicity of mathematical models of ILS under the standard multispecies coalescent (MSC) model (Pamilo and Nei, 1988; Rannala and Yang, 2003; Rannala et al., 2020). Recombination events along a chromosome allow neighboring loci to take on different gene tree topologies, all affected by the same biological processes. Accordingly, we assume that the gene tree G at any given locus has a distribution given by the MSC for species tree S. This distribution describes the probability distribution of the gene tree of a locus uniformly picked at random among a large number of loci.

ILS is particularly common when the internal branch lengths of the species tree are short. In some regions of the parameter space of the species tree (called the anomaly zone, or AZ for short), a discordant rooted gene tree topology can be more likely to occur than one that matches the species tree topology (Degnan and Rosenberg, 2006); a

* Correspondence to: Statistics and Operations Research, University of North Carolina, 18 E Cameron Ave, 27599, Chapel Hill, NC, USA.
E-mail addresses: rickertd@unc.edu, daricker@iu.edu (D.A. Rickert), louisfan@unc.edu (L.W.-T. Fan), mwh@iu.edu (M.W. Hahn).
[1] Present affiliation: Statistics and Operations Research, University of North Carolina, 18 E Cameron Ave, 27599, Chapel Hill, NC, USA.
[2] Present affiliation: School of Data Science and Society, University of North Carolina, 211 Manning Dr, 27514, Chapel Hill, NC, USA.

https://doi.org/10.1016/j.tpb.2025.09.004
Received 2 July 2024
Available online 27 September 2025

similar result also holds for unrooted gene tree topologies (Degnan, 2013). Further work (Rosenberg, 2013) has demonstrated that the AZ arises as the result of ILS on consecutive short branches of the species tree. Therefore, even in ideal world where we can infer gene tree topologies directly – essentially ignoring the randomness of sequences evolving on gene trees and the errors involved in inferring gene trees – the 'democratic vote' method that attempts to infer the species tree topology by simply returning the most common gene tree topology over many independent loci will be statistically inconsistent in some areas of parameter space (Degnan and Rosenberg, 2009), though more sophisticated methods using gene tree topology frequencies can give statistically consistent estimators of the species tree topology (Allman et al., 2011, 2018). The fact that the most probable gene tree topology under the MSC is not in general the species tree topology might appear to doom so-called concatenation methods, which combine data from multiple loci into a single alignment and essentially assume that all loci have evolved along the same gene tree. These concatenated alignments can be analyzed using a range of methods (e.g. parsimony, maximum likelihood, neighbor joining [NJ]) in order to return an estimated tree or topology. However, the regions of statistical consistency of such concatenation methods may differ from the AZ. For instance, simulations done in Kubatko and Degnan (2007) showed that, under the MSC, concatenated maximum-likelihood (ML) for 4 taxa could be consistent inside the AZ and inconsistent outside of it. This was shown more exhaustively in Mendes and Hahn (2018), who sampled a far greater number of points in parameter space.

Perhaps surprisingly, Liu and Edwards (2009) and Mendes and Hahn (2018) found that concatenated parsimony was statistically consistent for the rooted 4-taxa case across parameter space, assuming an infinite-sites mutation model and the MSC model. These findings contrast with the well-known results described in Felsenstein (1978), which found an area of parameter space of statistical inconsistency of parsimony (sometimes called the Felsenstein zone). These two sets of results do not conflict, as inconsistency in the Felsenstein zone is caused by similarity due to homoplasy (multiple substitutions at a site), a phenomenon that does not occur in the infinite-sites model. It should also be noted that the analysis in Felsenstein (1978) does not incorporate gene tree discordance, while the results of Liu and Edwards (2009) and Mendes and Hahn (2018) do.

In this work, we focus on concatenated parsimony and similar concatenation methods ('concatenated counting methods'), all of which take a concatenated alignment, $\mathcal{A}$, as input, and associate a 'cost' $c(T \mid V)$ to each candidate topology, T, and site pattern, $V \in \mathcal{A}$. These methods then attempt to infer the species tree topology by returning the candidate topology T that minimizes the total cost across the entire concatenated alignment. Such methods are similar to the idea behind concatenated ML, with the difference that concatenated ML attempts to minimize the total negative log-likelihood of a candidate tree (with branch lengths included) rather than just the total cost of a candidate topology (which encodes no information about branch lengths). The idea of cost minimization is also common in various gene tree methods using a collection, $\mathcal{G}$, of (estimated) gene trees and a cost function, $c(T \mid G)$, for each $G \in \mathcal{G}$. For instance, taking a cost function representing whether the rooted topology of G matches with T gives the 'democratic vote' method, while taking a cost function that returns the number of shared quartets between T and the unrooted topology of G motivates ASTRAL (Mirarab et al., 2014). Other choices of cost function have also been examined, for instance the minimize-deep-coalescence (MDC) criterion (Maddison, 1997; Maddison and Knowles, 2006; Than and Rosenberg, 2011).

In the same manner that examining the statistical consistency of gene tree methods as more loci are sampled usually requires calculating the frequencies of gene tree topologies under the MSC, examining the statistical consistency of concatenated counting methods involves calculating the expected lengths of branches of gene trees under the MSC. We begin by presenting a novel combinatoric technique to calculate these expected lengths, and demonstrate that the technique correctly recovers known results in the 4-taxa case. We then apply this technique to further understand the success and failure of concatenated parsimony methods for cases with 5 or more taxa ("5+ taxa"), both in the rooted and unrooted cases. While Roch and Steel (2015) have demonstrated inconsistency of concatenated parsimony for the unrooted 6-taxa case under a general r-state mutation model (even when homoplasy is negligible), their results do not characterize the precise regions of parameter space where parsimony fails; instead, their model assumed the probability of coalescence in internal branches is sufficiently small, justifying the use of Ewens' sampling formula (Ewens, 1972) in computing the probabilities of site patterns. Moreover, Bryant and Hahn (2020) have argued that the results of Roch and Steel (2015) only directly demonstrate inconsistency for biologically unrealistic species tree branch lengths. With our method of computing expected branch lengths, we demonstrate that for the previously unexplored unrooted 5-taxa case, concatenated parsimony is consistent under a MSC + infinite-sites model of evolution. We also show that under the same modeling assumptions, concatenated parsimony always has a region of inconsistency for the rooted 5+-taxa case and the unrooted 6+-taxa case, and find that this anomalous region is non-trivial, including many biologically realistic species trees. We conclude by discussing the implications of our results for the accurate inference of species trees.

2. Definitions

Let $\mathbb{T}_n$ denote the set of all rooted, binary, and labeled tree topologies on n taxa, with tips labeled by the label set $[n] = \{1, 2, \ldots, n\}$. For clarity, we will often use uppercase letters $\{A, B, C, \ldots, \}$ as the label set in place of $[n]$ when discussing specific examples. For each rooted topology $T \in \mathbb{T}_n$, we let $\overline{T}$ be its unrooted analogue, and define $\overline{\mathbb{T}}_n$ to be the collection of all such unrooted n-taxa topologies.

We think of a rooted, binary, and ultrametric species tree with n taxa as a pair $S = (T_*, \boldsymbol{x})$, where $T_* \in \mathbb{T}_n$ denotes the rooted topology of S and $\boldsymbol{x}$ is a vector of non-negative branch lengths of the species tree. For convenience, we will assume each branch of the species tree has a constant, large effective population size of N_e diploid individuals, and that all time and lengths are measured in coalescent units of $2N_e$ generations. For our data, we will assume an MSC + infinite-sites model of evolution, as follows:

- Loci are labeled by $i \in \{1, 2, 3 \ldots\}$, with G_i denoting the gene tree at locus i. The collection of gene trees $(G_i)_{i=1}^{\infty}$ is assumed to be independently and identically distributed according to the multispecies coalescent (MSC) on the species tree S, with one sequence sampled per taxa. In particular, any pair of lineages of the gene tree that exist in the same ancestral population coalesce independently at rate 1. We think of each sampled gene (or lineage) as being labeled by the label of the taxa it is sampled from. For example, the collection of lineages in the set $\{A, B, C\}$ refers to the lineages of the gene tree that are sampled from taxa A, B, C.
- Each locus consists of infinitely many sites, and mutations fall on the gene tree G_i according to an infinite sites model with rate $\theta/2$, where $\theta = 4N_e\mu$ is the scaled mutation rate, and μ is the per-generation mutation rate per locus, assumed to be constant throughout time and across loci.
- The alignment at locus i is denoted A_i. The jth row of A_i corresponds to the sequence of the sampled gene from taxa j, and each column corresponds to the nucleotide data for each of the n samples at a particular site. See Fig. 1a for an illustration.
- The ancestral allelic state (i.e. the allelic state immediately prior to a mutation event) at a site is denoted by 0, and the derived allelic state at a site is denoted by 1.

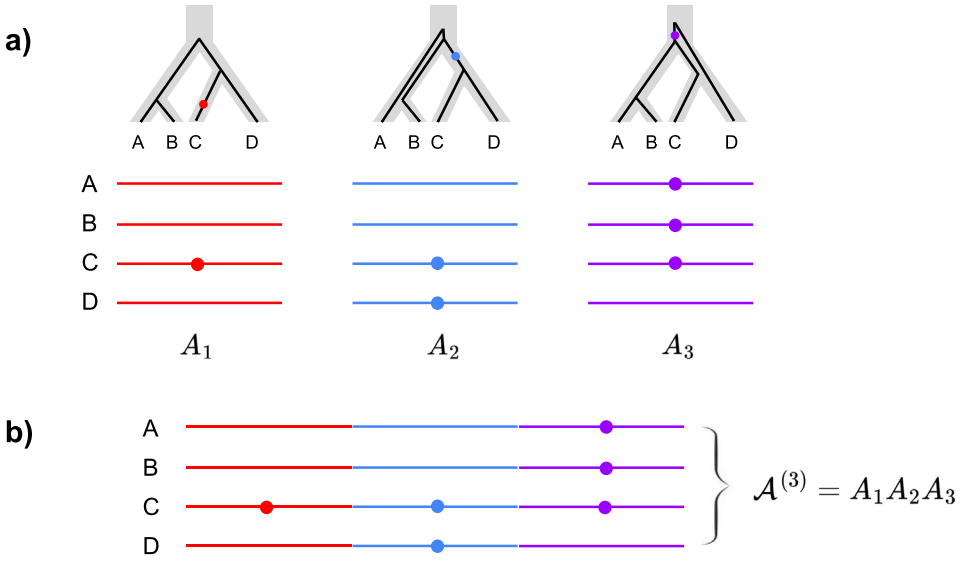

Fig. 1. Illustration of the concatenation procedure. (a) Three gene trees G_1, G_2, G_3 (black lines) and alignments A_1, A_2, A_3 for $k = 3$ independently evolving loci on the same species tree (shaded gray). Each mutation (circle) represents a $0 \to 1$ (ancestral to derived) transition at a new site in the locus. Only one segregating site per locus is shown for simplicity. (b) The resulting concatenated alignment $\mathcal{A}^{(3)} = A_1A_2A_3$, with three segregating sites.

- Each site in the alignment is summarized by a site pattern $V \subseteq [n]$ denoting the sequences at that site that have the derived allelic type 1. We will often write a site pattern by a concatenated string of its elements, for example $V = AB$ in place of $V = \{A, B\}$.

Assuming that we can distinguish between the derived and ancestral allelic states at each site, we may attempt inference of the rooted species tree topology T_*. A *rooted concatenated counting method* assigns a cost $c(T \mid V)$ to each pair of rooted candidate topology $T \in \mathbb{T}_n$ and site pattern V. The total cost of the candidate topology T on an concatenated alignment $\mathcal{A}^{(k)}$, obtained by combining alignments $A_1, \ldots, A_k$ (Fig. 1b), is given by

$$\mathrm{Cost}(T \mid \mathcal{A}^{(k)}) = \sum_V c(T \mid V)\#(V \mid \mathcal{A}^{(k)}) \tag{2.1}$$

where the sum is over all $V \subseteq [n]$, and $\#(V \mid \mathcal{A}^{(k)})$ is the number of occurrences of the site pattern V in $\mathcal{A}^{(k)}$. The resulting estimator is given by the topology T that minimizes the total cost; if there is a tie between multiple elements of $\mathbb{T}_n$, then we pick one uniformly at random.

On the other hand, when it is not possible to distinguish between the derived and ancestral allelic states, we will need to arbitrarily assign the derived and allelic states at each site. Hence, the site pattern V and its complement $[n] \setminus V$, resulting from the two possible choices of assignment, must be treated equally in any method of inference. Usually, this limits us to attempting inference of the unrooted species tree topology only. An *unrooted concatenated counting method* assigns a cost $c(\overline{T} \mid V)$ to each pair of unrooted candidate topology $\overline{T} \in \overline{\mathbb{T}}_n$ and site pattern V, with the symmetry requirement $c(\overline{T} \mid V) = c(\overline{T} \mid [n] \setminus V)$, and attempts to minimize

$$\mathrm{Cost}(\overline{T} \mid \mathcal{A}^{(k)}) = \sum_V c(\overline{T} \mid V)\#(V \mid \mathcal{A}^{(k)}) \tag{2.2}$$

For both rooted and unrooted concatenated counting methods, we will also make the following regularity assumptions on the choice of costs, which are met by the all the concatenated counting methods we discuss.

Parsimony uninformative sites are ignored A site with site pattern V is said to be parsimony informative (or simply informative) if $2 \le |V| \le n-1$ in the rooted case, and if $2 \le |V| \le n-2$ in the unrooted case. All other sites/site patterns are considered uninformative and are excluded when calculating the total cost as in (2.1). Note that a site pattern V with $|V| = n-1$ is no longer considered informative in the unrooted case, because it can be explained by a single mutation event on an external branch of any unrooted gene tree, regardless of its topology. See Fig. 2 for an illustration of the distinction between an informative and uninformative site pattern.

Implication: Informative sites arise as the result of mutations occurring on internal branches of the rooted/unrooted gene tree. Since no coalescent events on the gene trees $(G_i)_i$ occur on external branches of the species tree, the lengths of these external branches are irrelevant to the behavior of the estimator. Therefore, we will write the species tree branch lengths $\boldsymbol{x} = (x_1, \ldots, x_{n-2})$ to be a vector of $n-2$ internal branch lengths throughout the paper.

Exchangeability of labels of taxa The cost function should not favor any particular taxa or group of taxa over any other based solely on their (arbitrary) labeling. In particular, if π is any permutation of $[n]$, we require that $c(\pi(T) \mid \pi(V)) = c(T \mid V)$, where $\pi(T) \in \mathbb{T}_n$ is the topology obtained by permuting the labels of the tips of T by π, and $\pi(V) \subseteq [n]$ is the site pattern obtained by applying π to each of the elements of V.

Implication: The behavior of a concatenated counting method does not depend on the choice of labeling. Therefore, in any exploration of species tree space, it suffices to examine one representative of each possible unlabeled n-taxa topology, instead of examining all possible labeled n-taxa topologies.

The most well-known group of concatenated counting methods are parsimony methods, such as Camin–Sokal parsimony (Camin and Sokal, 1965), Wagner (unordered) parsimony (Kluge and Farris, 1969; Farris, 1970), and Dollo parsimony (Le Quesne, 1974, 1977; Farris, 1977). See Felsenstein (1983) for an overview of these methods. Each of these methods assigns costs $c(T \mid V)$ by counting the minimal number of mutations needed to explain a site pattern V on the topology T, with variation among the methods due to different assumptions on the allowed transitions from the ancestral allelic state (0) to the derived allelic state (1). Among these parsimony methods, we will focus our attention on Wagner parsimony (henceforth just referred to as parsimony), so we give a quick definition here:

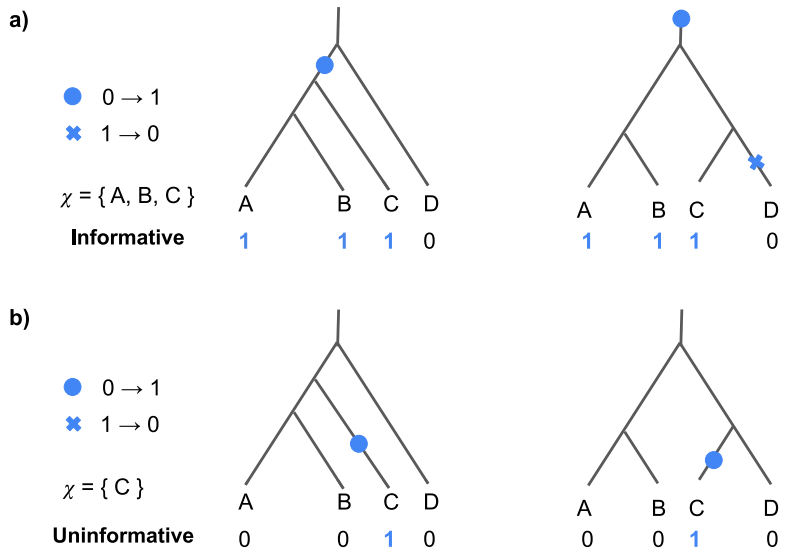

Fig. 2. (a) An informative site pattern (ISP), $V = \{A, B, C\}$ resolved on two different candidate tree topologies requiring different numbers of mutations; (b) an uninformative site pattern, $V = \{C\}$ resolved on two different topologies requiring only one mutation in both cases.

Rooted parsimony The cost $c(T \mid V)$ is the minimum total number of $0 \to 1$ and $1 \to 0$ mutations needed to resolve V on the topology T, assuming an ancestral state of 0. By convention, a $0 \to 1$ mutation may be placed above the root of the topology. See Fig. 2a for an example.

Unrooted parsimony The cost $c(\overline{T} \mid V)$ is the minimum total number of $0 \to 1$ and $1 \to 0$ mutations needed to resolve V on a rooted representative T of $\overline{T}$, allowing the ancestral state to be 0 or 1.

Parsimony methods are not the only concatenated counting methods, with recent work highlighting quartet-based approaches, both in the case of an infinite-sites model of mutation (e.g. SDPQuartets / ASTRAL-BP (Springer et al., 2020; Molloy et al., 2022)) and more general mutation models (e.g. CASTER Zhang et al., 2025). Both SDPQuartets and ASTRAL-BP are closely related to parsimony — using the fact that parsimony is statistically consistent for the unrooted 4-taxa case under a MSC + infinite-sites model of mutation (Mendes and Hahn, 2018; Molloy et al., 2022) – though the exact implementation used to search for the optimal candidate tree topology varies between these methods. CASTER is also closely related to parsimony, but involves a negative weighting of some parsimony-uninformative sites to compensate for the possibility of multiple mutations at a single site. Since our focus is on an infinite-sites model of mutation, we give a brief framing of the methods SDPQuartets/ASTRAL-BP as concatenated counting methods.

For any four distinct elements a, b, c, d of the label set $[n]$, we say that unrooted topology $\overline{T} \in \overline{\mathbb{T}}_n$ displays the quartet $ab|cd$ if the restriction of T to taxa a, b, c, d has unrooted topology $((ab)(cd))$, i.e. taxa a, b are sisters in $\overline{T}$ as are taxa c, d. Similarly, we say that a site pattern V supports the quartet $ab|cd$ if $a, b \in V$ but $c, d \notin V$ or vice versa.

SDPQuartets/ASTRAL-BP The cost $c(\overline{T} \mid V)$ is taken to be $-q(\overline{T} \mid V)$, where $q(\overline{T} \mid V)$ is the number of the $\binom{n}{4}$ unrooted quartets $ab|cd$ implied by T that the site pattern V supports.

To examine the statistical consistency of concatenated counting methods as the number of sampled loci included in the alignment grows large, we note that the expected number of sites with site pattern V in the alignment A_i is proportional to the expected length of the branch in G_i which subtends exactly the lineages in V. Hence, the expected contribution of the alignment A_i to the total cost in (2.1) is proportional to

$$C(T \mid S) := \sum_V c(T \mid V)\ell(V \mid S) \tag{2.3}$$

where $\ell(V \mid S)$ is the expected branch length of a branch ancestral to (subtending) exactly the sampled lineages in V for a gene tree generated under the MSC on S. One may analogously define the expected cost per locus $C(\overline{T} \mid S)$ in the unrooted case. Therefore, the strong law of large numbers gives the following criteria for consistency/inconsistency of the given concatenated counting estimator under the species tree S (with an analogous criterion holding in the unrooted case):

(Consistency) $C(T_* \mid S) < C(T \mid S)$ for all $T \neq T_* \in \mathbb{T}_n$.

(Inconsistency) $C(T_* \mid S) \geq C(T \mid S)$ for some $T \neq T_* \in \mathbb{T}_n$.

The main objectives of this paper are to present a novel method of computing the quantity $\ell(V \mid S)$, and analyze the statistical consistency of concatenated counting methods (in particular parsimony) as follows:

1. In Section 3.1, we demonstrate a new, simple manner of computing expected branch lengths $\ell(V \mid S)$ in the gene tree under the MSC.
2. In Section 3.2, we discuss a decomposition of expected branch lengths $\ell(V \mid S)$ and related quantities, which is useful both conceptually and computationally.
3. In Section 3.3, we provide sample calculations of finding $\ell(V \mid S)$ in the 4-taxa case, and show that these results agree with existing work;
4. In Section 4.1, we prove the statistical consistency of parsimony under a MSC + infinite-sites model of evolution in the previously analyzed unrooted 5-taxa case;
5. In Sections 4.2 and 4.3, we show that parsimony is statistically inconsistent under a MSC + infinite-sites model of evolution in the rooted 5-taxa and unrooted 6-taxa cases, and characterize the exact regions of statistical consistency.

3. Expected branch lengths under the MSC

In Section 2, we saw that establishing the statistical (in)consistency of concatenated counting methods requires calculating the expected branch lengths of gene trees. In this section, we demonstrate how existing work on the expected height of gene trees under the MSC, in particular that of Efromovich and Salter Kubatko (2008), may be used to compute the expected branch lengths in gene trees under the

MSC. The idea of using the expected height of the MRCA of sampled tips (usually, by sampling multiple tips per taxa) has been used more directly in species tree inference by methods such as GLASS (Mossel and Roch, 2008), iGLASS (Jewett and Rosenberg, 2012), STEAC (Liu et al., 2009), and MAC (Helmkamp et al., 2012); however, here we will use the expected heights as a means to an end. We note that some calculations regarding expected branch lengths have already been done in the 3- and 4-taxa cases, though the approach taken by existing methods makes extensions to 5+-taxa difficult. For instance, Mendes and Hahn (2018) did so by computing the expected branch length conditional on each possible gene tree history, a method that quickly becomes infeasible for large n since there are $(2n-3)!! = (2n-3)\times(2n-1)\times\cdots\times 1$ possible gene tree topologies and even more possible gene tree histories. Alternatively, a diffusion approximation has been used (Doronina et al., 2017), but this approach requires similar amounts of work. Our new approach involves a relatively simple combinatoric trick (the inclusion–exclusion principle) that minimizes individual calculations and enables easier analysis. It should be noted that the same fundamental combinatorical reasoning we apply here has previously been used to derive branch lengths in a population-genetic context in order to calculate entries of the expected site frequency spectrum (Jewett, 2020).

3.1. Gene tree lengths, subtending lengths, and heights

We first introduce notation for certain branch lengths of interest in a (random) gene tree G. We will not continue the gene tree above its root (i.e. the MRCA of all sampled lineages), since mutations that occur above the root do not give rise to parsimony informative sites.

Definition (*Subtending*)**.** A branch in a gene tree G is said to subtend a collection of sampled lineages $j_1,\dots,j_r \in [n]$ if it extends from the common ancestor of a clade containing $j_1,\dots,j_r$, or equivalently, if it lies on the path connecting j_i to the root of G for each $i = 1,\dots,r$. If the branch does not subtend any additional sampled lineages, we say the branch subtends exactly $j_1,\dots,j_r$.

Definition (*Length Corresponding to a Site Pattern*)**.** For $V \subseteq [n]$, the random variable $L(V)$ is the length of the branch subtending exactly the lineages in V in the random gene tree G if such a branch exists; otherwise it is defined to be 0.

Definition (*Subtending Length Corresponding to a Site Pattern*)**.** For $V \subseteq [n]$, the random variable $L^+(V)$ is the total length of branches that subtend at least the lineages in V in the random gene tree G:

$$L^+(V) = \sum_{W\supseteq V} L(W). \tag{3.1}$$

Ignoring branch lengths that are zero in the sum $\sum_{W\supseteq V} L(W)$, we see that the subtending length $L^+(V)$ amounts to the length of the path from the MRCA of the lineages in V to the root of the gene tree. In particular,

$$L^+(V) = H([n]) - H(V)$$

where $H([n])$ is the height of the MRCA of all n lineages, and $H(V)$ is the height of the MRCA of the sampled lineages in V. An example of the relationships connecting $L(V), L^+(V)$ and $H(V)$ is given in Fig. 3. By taking expected values,

$$\ell^+(V \mid S) := \mathbb{E}_S[L^+(V)] = h([n] \mid S) - h(V \mid S), \tag{3.2}$$

where $h([n] \mid S)$ is defined to be the expected height of a gene tree with one sample from each taxon, and $h(V \mid S)$ is defined to be the expected height of a gene tree with one sample from each of the taxa in V only. Both expectations are under the MSC on S. In computing $h(V \mid S)$, we may restrict S to just the taxa in V, since we only sample lineages from these taxa. The computation of these expected heights may be done by a standard dynamic programming method; see for instance (Efromovich and Salter Kubatko, 2008, Eq. (6)).

To get a feel for how the lengths $(L(V))_{V\subseteq[n]}$ can be related back to the subtending lengths $(L^+(V))_{V\subseteq[n]}$, it can be helpful to make a connection with indicator functions and the inclusion–exclusion principle. Consider the four-taxa case with a site pattern $V = \{a,b\}$, letting c,d denote the other two taxa of the species tree not in V. Then we can write

$$L(ab) = \sum_{V\subseteq\{a,b,c,d\}} 1\{a,b\in V, c,d\notin V\}L(V)$$
$$L^+(ab) = \sum_{V\subseteq\{a,b,c,d\}} 1\{a,b\in V\}L(V)$$

and similarly for other site patterns. We then expand the indicator function $1\{a,b\in V, c,d\notin V\}$:

$$\begin{aligned}
&1\{a,b\in V, c,d\notin V\}\\
&= 1\{a,b\in V\}\cdot 1\{c\notin V\}\cdot 1\{d\notin V\}\\
&= 1\{a,b\in V\}\cdot\big(1-1\{c\in V\}\big)\cdot\big(1-1\{d\in V\}\big)\\
&= 1\{a,b\in V\} - 1\{a,b,c\in V\} - 1\{a,b,d\in V\} + 1\{a,b,c,d\in V\}
\end{aligned}$$

The terms on the last line, after multiplying by $L(V)$ and summing over all $V\subseteq\{a,b,c,d\}$, correspond exactly to $+L^+(ab)$, $-L^+(abc)$, $-L^+(abc)$, and $+L^+(abcd)$ respectively. (Note that $L^+(abcd)$ is 0 in this example, since no internal branch of G subtends all four sampled lineages). This gives an idea for how we may 'invert' the relationship given in (3.1); also see Fig. 4 for a further visual example of this inversion in the 5-taxa case. In general, we have

$$L(V) = \sum_{W\supseteq V} (-1)^{|W|-|V|} L^+(W). \tag{3.3}$$

By taking expected values in (3.3), we obtain a practical method of computing $\ell(V \mid S)$ in general:

$$\ell(V \mid S) = \sum_{W\supseteq V} (-1)^{|W|-|V|} \ell^+(V \mid S) \tag{3.4}$$

or, alternatively, after applying (3.2) and canceling terms $h([n] \mid S)$,

$$\ell(V \mid S) = -\sum_{W\supseteq V} (-1)^{|W|-|V|} h(V \mid S) \tag{3.5}$$

The generality of the above ideas connecting the expected length $\ell(V \mid S)$ to the expected heights of gene trees should be noted. For instance, if we extend the algorithm of Efromovich and Salter Kubatko (2008) to calculate expected gene tree heights for a given phylogenetic network (i.e. in models involving introgression), we can still use (3.5) to analyze the regions of inconsistency for a given concatenated counting method (cf. Hibbins and Hahn (2022)). It is also relatively straightforward to create a similar inversion formula connecting expected heights to expected branch lengths in the case where more than one lineage is sampled per taxa; see for instance (Jewett, 2020, Section 3.5). The technique also applies in the context of population genetics for general coalescent processes with multiple mergers, such as in the Λ-coalescent (Pitman, 1999) or the Ξ-coalescent (Schweinsberg, 2000), and this idea has been explored in Spence et al. (2016), though in an indirect manner that did not fully explore the underlying connection to the inclusion–exclusion principle.

It is also possible that a similar combinatoric approach could be applied to find the probability that a randomly sampled gene tree has a particular branch (or collection of branches), rather than to find the expected lengths of branches. This may allow alternative derivations of results on monophyly and paraphyly as examined in several previous works (Rosenberg, 2003; Mehta et al., 2016; Mehta and Rosenberg, 2019; Mehta et al., 2022).

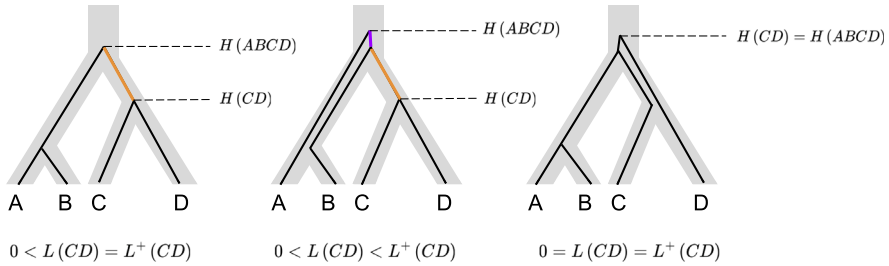

Fig. 3. Three different realizations of a gene tree G within a species tree with topology $T_* = ((AB)(CD))$. Branches of the gene tree that contribute to the length $L(CD)$ and the subtending length $L^+(CD)$ are highlighted in orange, whereas branches of the gene tree that contribute to $L^+(CD)$ but not $L(CD)$ are highlighted in purple. The heights of the overall gene tree $H(ABCD)$ and the height $H(CD)$ of the gene tree restricted to tips C, D are also shown.

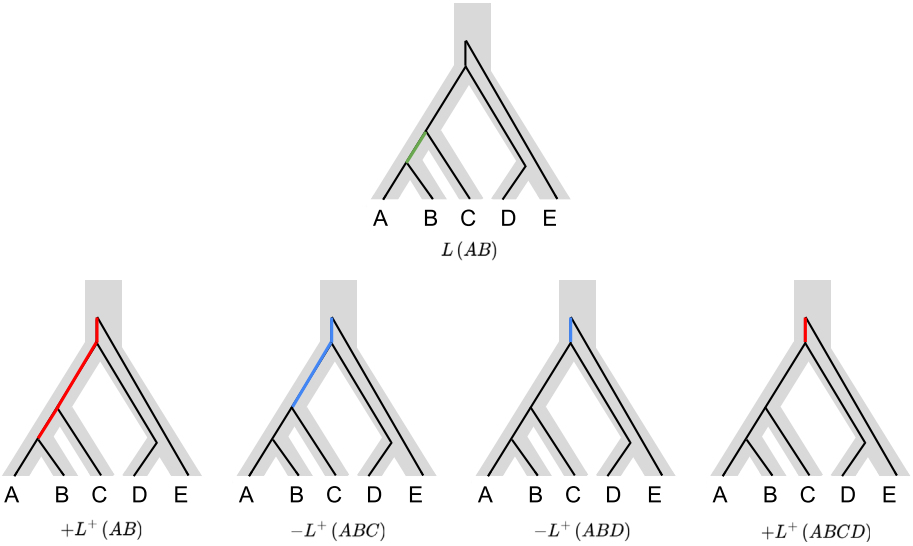

Fig. 4. Illustration of recovering the length $L(V)$ for $V = AB$ from the collection of subtending lengths $(L^+(W))_{W \supseteq V}$ on a gene tree. Terms with an subtending length $L^+(W) = 0$ for this realization of the gene tree, namely those with $E \in W$, are omitted.

3.2. Decomposition into species and coalescent terms

In performing analysis with the expected branch lengths, it can be helpful to first decompose the expected heights $h(V \mid S)$ as the height $h_{\text{sp}}(V \mid S)$ of the species MRCA of the taxa in V in the species tree S, plus the expected height $h_{\text{coa}}(V \mid S)$ of the MRCA of the lineages in V above the species MRCA of the taxa in V. We will call these terms the 'species' and 'coalescence' terms respectively; Efromovich and Salter Kubatko (2008) refers to the coalescence term as the 'species-gene coalescent time'.

$$h(V \mid S) = h_{\text{sp}}(V \mid S) + h_{\text{coa}}(V \mid S)$$

The species term $h_{\text{sp}}(V \mid S)$ is the sum of a collection of branch lengths of S. Meanwhile the coalescent term $h_{\text{coa}}(V \mid S)$ is a polynomial in the transformed branch lengths $X_i := \exp(-x_i)$, since the coalescent transition probabilities are a polynomial function of $\exp(-t)$ (Tavaré, 1984).

By grouping species and coalescence terms that appear together in the defining relationships between quantities, we can recursively define a similar decomposition for the expected subtending length $\ell^+(V \mid S)$, the expected length $\ell(V \mid S)$, and the expected cost per locus $C(T \mid S)$. That is, we write

$$\ell^+_{\text{sp}}(V \mid S) = h_{\text{sp}}([n] \mid S) - h_{\text{sp}}(V \mid S)$$
$$\ell_{\text{sp}}(V \mid S) = \sum_{W \supseteq V} (-1)^{|W|-|V|} \ell^+_{\text{sp}}(V \mid S)$$

and we make analogous definitions for the coalescence terms. These decompositions often have nice interpretations. For instance, in the decomposition

$$\ell(V \mid S) = \ell_{\text{sp}}(V \mid S) + \ell_{\text{coa}}(V \mid S) \tag{3.6}$$

the species term $\ell_{\text{sp}}(V \mid S)$ is nothing but the length of the branch in S that subtends exactly the tips in V, if such a branch in S exists; otherwise it is 0. Similarly, we look can look at the decomposition for the expected cost per locus:

$$C(T \mid S) = \underbrace{\sum_V c(T \mid V)\ell_{\text{sp}}(V \mid S)}_{C_{\text{sp}}(T|S)} + \underbrace{\sum_V c(T \mid V)\ell_{\text{coa}}(V \mid S)}_{C_{\text{coa}}(T|S)} \tag{3.7}$$

It is possible to interpret $C_{\text{sp}}(T \mid S)$ as the expected cost per locus in the absence of gene tree heterogeneity, while $C_{\text{coa}}(T \mid S)$ can be interpreted as an additional cost that arises due to ILS and gene tree discordance.

$C_{\text{coa}}(T \mid S)$ is easy to compute in the limit where S is a star tree (i.e. all internal branch lengths of S are 0). We denote a star tree by $\star$. Note once again that the external branch lengths of the star tree are irrelevant, since no coalescent events occur along these branches. For a star tree, the coalescence term in fact comprises the entirety of the expected cost. Under our modeling assumptions, the coalescent process in the singular ancestral population (where all internal branches of the gene trees arise) is that of the Kingman coalescent, and so the expected length of internal branches of the gene tree subtending exactly $2 \le i \le n-1$ tips is $2/i$ (Fu, 1995). We then have

$$\ell_{\text{coa}}(V \mid \star) = \frac{2}{|V|}\binom{n}{|V|}^{-1}$$
$$C_{\text{coa}}(T \mid \star) = \sum_V c(T \mid V)\,\frac{2}{|V|}\binom{n}{|V|}^{-1}$$

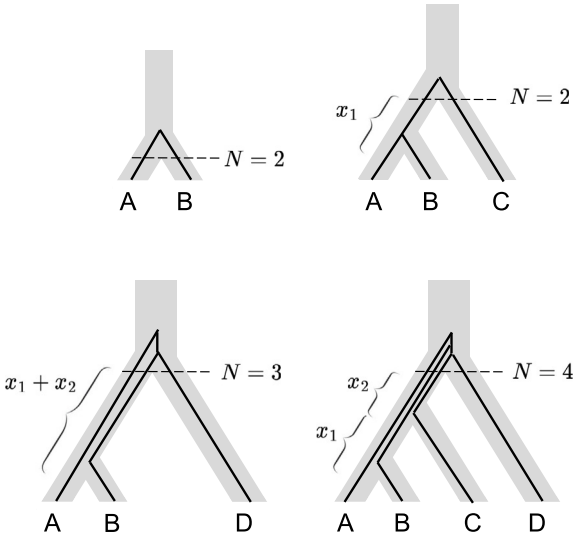

Fig. 5. The restriction of a single species tree S with asymmetric topology $(((AB)C)D)$ to taxa AB, ABC, ABD, and $ABCD$. Example realizations of a gene tree for each restricted species tree and the random variable N (the number of lineages entering the root of the restricted species tree) are shown.

where the expression for $\ell_{\text{coa}}(V \mid \star)$ follows by symmetry, as there are $\binom{n}{|V|}$ possible branches subtending exactly $|V|$ tips. This provides a simple criterion to show inconsistency: if there exist two topologies $T, T' \in \mathbb{T}_n$ such that

$$C(T \mid \star) < C(T' \mid \star)$$

then, owing to the continuity of $C(T \mid S)$ as a function of $\boldsymbol{x}$, the concatenated counting estimator will always prefer T over T' for sufficiently short branch lengths $\boldsymbol{x}$, even if the true species topology is $T_* = T'$, i.e. $C(T \mid (T', \boldsymbol{x})) < C(T' \mid (T', \boldsymbol{x}))$ when $\boldsymbol{x}$ is in some neighborhood of $\mathbf{0}$.

3.3. A detailed analysis for the 4-taxa case

To demonstrate the procedure of finding $\ell(V \mid S)$ in general, we work out two examples in the 4-taxa case. For these examples, we will take S to have the asymmetric topology $T_* = (((AB)C)D)$ and internal branch lengths x_1, x_2 subtending taxa A, B and taxa A, B, C, respectively. Our goal will be to find the expected branch lengths $\ell(AB \mid S)$ (Example 1) and $\ell(CD \mid S)$ (Example 2). We will see that we can reuse much of the work from Example 1 in Example 2, so we will provide most of the detailed exposition of the procedure in Example 1. We will then finish by observing that the sum $\ell(AB \mid S) + \ell(CD \mid S)$ has a particularly nice form, and use it to motivate a useful lemma (Lemma 1), originally shown by Molloy et al. (2022).

Example 1. To compute $\ell(AB \mid S)$, we start with the decomposition into species and coalescence terms as in (3.6): we know that $\ell_{\text{sp}}(AB \mid S) = x_1$, because the length of the internal branch of the species tree that subtends exactly AB is x_1. Meanwhile, we can use (3.5) to find that

$$\begin{aligned}\ell_{\text{coa}}(AB \mid S) = &-h_{\text{coa}}(AB \mid S) + h_{\text{coa}}(ABC \mid S) \\ &+ h_{\text{coa}}(ABD \mid S) - h_{\text{coa}}(ABCD \mid S)\end{aligned}$$

Our next step will be to construct four sub-species trees of S, obtained by restricting S to the taxa AB, ABC, ABD and $ABCD$, respectively. We will denote the species tree obtained by restricting to the taxa in V by $S_{|V}$. For each of these restricted trees, we must find the distribution of N, the number of lineages of the gene tree that enter the root of the restricted species tree. See Fig. 5 for an illustration. To do so, we use the coalescent transition probability functions $g_{ij}(t)$ from Tavaré (1984, Equation 6.1). In this example, we will only need the specific formulae for $i = 2, 3$, listed below for reference.

$$\begin{aligned}&g_{21}(t) = 1 - \exp(-t), \quad g_{22}(t) = \exp(-t) \\ &g_{31}(t) = 1 - \frac{3}{2}\exp(-t) + \frac{1}{2}\exp(-3t), \quad g_{32}(t) = \frac{3}{2}\exp(-t) \\ &\qquad - \frac{3}{2}\exp(-3t), \quad g_{33}(t) = \exp(-3t)\end{aligned}$$

Once the distribution of N is found, we may apply (Efromovich and Salter Kubatko, 2008, Equation 3) to find the expected height above the root of the restricted tree:

$$h_{\text{coa}}(V \mid S) = 2 - \sum_{i=2}^{|V|} \frac{2}{i}\,\mathbb{P}(N = i \mid S_{|V})$$

We now carry out the calculation for each of the restricted species trees:

- For restricted species tree $S_{|AB}$: Clearly, $\mathbb{P}(N = 1 \mid S_{|AB}) = 1$, so $h_{\text{coa}}(AB \mid S) = 1$.
- For restricted species tree $S_{|ABC}$: $\mathbb{P}(N = 2 \mid S_{|ABC}) = g_{21}(x_1) = 1 - \exp(-x_1)$ (i.e. if lineages A, B coalesce in the ancestral population of species A, B) and $P(N = 3 \mid S_{|ABC}) = g_{22}(x_1) = \exp(-x_1)$. Thus, $h_{\text{coa}}(ABC \mid S) = 2 - \frac{2}{2}(1 - \exp(-x_1)) - \frac{2}{3}\exp(-x_1) = 1 + \frac{1}{3}\exp(-x_1)$.
- For restricted tree $S_{|ABD}$: the same logic as above (with the only difference that $S_{|ABD}$ has a single internal branch of length $x_1 + x_2$) shows that $h_{\text{coa}}(ABD \mid S) = 1 + \frac{1}{3}\exp(-x_1 - x_2)$.
- For restricted species tree $S_{|ABCD}$: we first must consider the number of lineages entering and exiting the ancestral population of taxa A, B, C. Two lineages of the gene tree enter the ancestral population of taxa A, B, C with probability $g_{21}(x_1) = 1 - \exp(-x_1)$ and three lineages of the gene tree enter the ancestral population with probability $g_{22}(x_1) = \exp(-x_1)$. Therefore, by considering the number of coalescence events that occur, we have either 1, 2 or 3 lineages of the gene tree at the top of the ancestral population of taxa A, B, C. This leads rise to 2, 3, or 4 lineages that enter the root population when also considering lineage D:

$$\begin{aligned}\mathbb{P}(N = 2 \mid S) &= g_{31}(x_2)g_{22}(x_1) + g_{21}(x_2)g_{21}(x_1) = 1 - \exp(-x_2) \\ &\quad - \frac{1}{2}\exp(-x_1 - x_2) + \frac{1}{2}\exp(-x_1 - 3x_2) \\ \mathbb{P}(N = 3 \mid S) &= g_{32}(x_2)g_{22}(x_1) + g_{22}(x_2)g_{21}(x_1) = \exp(-x_2) \\ &\quad + \frac{1}{2}\exp(-x_1 - x_2) - \frac{3}{2}\exp(-x_1 - 3x_2) \\ \mathbb{P}(N = 4 \mid S) &= g_{33}(x_2)g_{22}(x_1) = \exp(-x_1 - 3x_2)\end{aligned}$$

  Putting this together gives $h_{\text{coa}}(ABCD \mid S) = 1 + \frac{1}{3}\exp(-x_2) + \frac{1}{6}\exp(-x_1 - x_2)$.

With these expected heights, we find that

$$\begin{aligned}\ell_{\text{coa}}(AB \mid S) &= -1 + \left[1 + \frac{1}{3}\exp(-x_1)\right] + \left[1 + \frac{1}{3}\exp(-x_1 - x_2)\right] \\ &\quad - \left[1 + \frac{1}{3}\exp(-x_2) + \frac{1}{6}\exp(-x_1 - x_2)\right] \\ &= \frac{1}{3}\exp(-x_1) - \frac{1}{3}\exp(-x_2) + \frac{1}{6}\exp(-x_1 - x_2)\end{aligned}$$

and adding back the species term gives

$$\ell(AB \mid S) = x_1 + \frac{1}{3}\exp(-x_1) - \frac{1}{3}\exp(-x_2) + \frac{1}{6}\exp(-x_1 - x_2)$$

or, using the transformed variables $(X_1, X_2) = (\exp(-x_1), \exp(-x_2))$ as shorthand,

$$\ell(AB \mid S) = x_1 + \frac{1}{3}(X_1 - X_2) + \frac{1}{6}X_1X_2$$

Example 2. To compute $\ell(CD \mid S)$, we again decompose into species and coalescence terms, noting $\ell_{sp}(CD \mid S) = 0$ since no branch of the species tree subtends exactly the taxa C, D. Therefore, the entire expected length $\ell(CD \mid S)$ consists of the coalescence term:

$$\ell(CD \mid S) = -h_{coa}(CD \mid S) + h_{coa}(ACD \mid S) + h_{coa}(BCD \mid S) - h_{coa}(ABCD \mid S)$$

Here, we can reuse many of the calculations from Example 1, both in a direct and indirect manner. For instance, we may reuse our calculation for $h_{coa}(ABCD \mid S)$ directly, while applying the same logic as in our calculations for $h_{coa}(ABC \mid S)$ to determine $h_{coa}(ACD \mid S)$ and $h_{coa}(BCD \mid S)$. In particular, $S_{|ABC}$ is a 3-taxa caterpillar tree with the singular internal branch length x_1, whereas $S_{|ACD}$ and $S_{|BCD}$ are both 3-taxa caterpillar trees with internal branch length x_2. So $h_{coa}(ACD \mid S)$ can be found by substituting x_2 in for x_1 in the expression for $h_{coa}(ABC \mid S)$. Putting these observations together, we get

$$\begin{aligned}\ell(CD \mid S) &= -1 + \left[1 + \frac{1}{3}\exp(-x_2)\right] + \left[1 + \frac{1}{3}\exp(-x_2)\right] \\ &\quad - \left[1 + \frac{1}{3}\exp(-x_2) + \frac{1}{6}\exp(-x_1 - x_2)\right] \\ &= \frac{1}{3}\exp(-x_2) - \frac{1}{6}\exp(-x_1 - x_2)\end{aligned}$$

or, once again using the transformed variables $(X_1, X_2) = (\exp(-x_1), \exp(-x_2))$ as shorthand,

$$\ell(CD \mid S) = \frac{1}{3}X_2 - \frac{1}{6}X_1X_2$$

Using the results of Examples 1 and 2, we can make the observation that $\ell(AB \mid S) + \ell(CD \mid S)$, which is proportional to the expected number of informative site patterns that support the unrooted quartet $AB|CD$, is in fact independent of x_2:

$$\ell(AB \mid S) + \ell(CD \mid S) = x_1 + \frac{1}{3}X_1 \tag{3.8}$$

This result can be thought of in terms of the unrooted species tree $\overline{S}$, i.e. the unrooted analogue of S. Indeed, $\overline{S}$ has a single internal branch of length x_1, with the previous internal branch length of x_2 in the rooted tree S being collapsed into an external branch of $\overline{S}$. Moreover, we can use this result to quickly compute the expected number of informative site patterns that support the alternative quartets $AC|BD$ or $AD|BC$. To see how, we note that for any permutation a, b, c, d of A, B, C, D, we have

$$\begin{aligned}\ell(ab \mid S) + \ell(cd \mid S) = &-h(ab \mid S) + h(abc \mid S) + h(abd \mid S) - h(abcd \mid S) \\ &-h(cd \mid S) + h(acd \mid S) + h(bcd \mid S) - h(abcd \mid S)\end{aligned}$$

We observe that regardless of the choice of the permutation a, b, c, d of A, B, C, D, the terms $+h(ABC)$, $+h(ABD)$, $+h(ACD)$, $+h(BC)$, and $-h(ABCD)$ (twice) will appear in this expansion in some order. That is,

$$\ell(ab \mid S) + \ell(cd \mid S) = -h(ab \mid S) - h(cd \mid S) + \text{const.} \tag{3.9}$$

where the constant is the same regardless of the choice of permutation a, b, c, d of A, B, C, D, depending only on S. Further, since the coalescent terms $h_{coa}(ab \mid S) = 1$ regardless of the choice of a, b, we have that

$$\ell_{coa}(ab \mid S) + \ell_{coa}(cd \mid S) = \text{const.}$$

and we can read off this constant as $\frac{1}{3}X_1$ from (3.8). Meanwhile the species terms are also 0 for the alternative quartets $AC|BD$ and $AD|BD$:

$$\ell_{sp}(AC \mid S) + \ell_{sp}(BD \mid S) = \ell_{sp}(AD \mid S) + \ell_{sp}(BC \mid S) = 0$$

because no internal branches of S subtend exactly AC, BD, AD, or BC. Therefore, we conclude

- $\ell(AB \mid S) + \ell(CD \mid S) = x_1 + \frac{1}{3}X_1$;
- $\ell(AC \mid S) + \ell(BD \mid S) = \ell(AD \mid S) + \ell(BC \mid S) = \frac{1}{3}X_1$

While we have only examined the case where S has an asymmetric topology, it turns out this result can be extended to the case where S has the symmetric topology $T_* = ((ab)(cd))$ as well, as in the following lemma, previously stated and proven in Molloy et al. (2022) (though we rewrite it in our notation):

Lemma 1. *Suppose S is a 4-taxa rooted species tree with taxa A, B, C, D, and let a, b, c, d be a permutation of A, B, C, D. Then*

- $\ell(ab \mid S) + \ell(cd \mid S) = \tau + \frac{1}{3}\exp(-\tau)$ *if $\overline{S}$ displays the quartet $ab|cd$;*
- $\ell(ab \mid S) + \ell(cd \mid S) = \frac{1}{3}\exp(-\tau)$ *if $\overline{S}$ displays does not display the quartet $ab|cd$;*

where τ is the length of the single internal branch of the unrooted species tree $\overline{S}$.

In words, this lemma tells us that in the 4-taxa case, (1) most informative site patterns support the quartet displayed by the unrooted species tree over the two alternative quartets; and (2) the magnitude of this support is an increasing function of the internal branch of the unrooted species tree. We will make use of this observation in Section 4.1 when discussing the consistency of parsimony in the unrooted 5-taxa case.

4. Where parsimony succeeds and fails across tree space

4.1. Parsimony for the unrooted 5-taxa case

There is only one possible shape for an unrooted 5-taxa topology — all topologies are of form $((ab)(cd)e)$ (for some permutation a, b, c, d, e of A, B, C, D, E), i.e. the two pairs of sister taxa a, b and c, d are separated by taxon e (Fig. 6a). Therefore, we will examine one labeled representative of this shape only, assuming that species tree has true unrooted topology $\overline{T}_* = ((AB)(CD)E)$. The most direct method to analyze the consistency of concatenated parsimony (and a method can be generalized to all unrooted concatenated counting methods) would proceed as follows:

1. Find all rooted topologies T_* that have unrooted topology $((AB)(CD)E)$;
2. Compute $C(\overline{T} \mid S)$ as a function of $\boldsymbol{x}$ for $S = (T_*, \boldsymbol{x})$ for each $\overline{T} \in \overline{\mathbb{T}}_5$;
3. Check if $C(\overline{T} \mid S)$ is minimized at $\overline{T}_* = ((AB)(CD)E)$ for all choices of nonzero branch lengths $\boldsymbol{x}$.

We will ultimately follow a procedure analogous to this in the unrooted 6-taxa case (Section 4.3). However, such an approach would overlook the particularly simple structure of parsimony in the unrooted 5-taxa case. For the candidate unrooted topology $\overline{T} = ((ab)(cd)e)$, parsimony assigns a cost of 1 to the site patterns ab, abe, cd, cde and a cost of 2 to all other site patterns. As a result, as the number of loci sampled grows large, concatenated parsimony will prefer the topology $((ab)(cd)e)$ for which the sum of lengths

$$P(\overline{T} \mid S) := \ell(ab \mid S) + \ell(abe \mid S) + \ell(cd \mid S) + \ell(cde \mid S)$$

is maximal. Noting the irrelevance of taxon e in $P(\overline{T} \mid S)$, it should not be difficult to believe that $P(\overline{T} \mid S)$ depends only on the restricted species tree $S_{|abcd}$:

$$P(\overline{T} \mid S) = \ell(ab \mid S_{|abcd}) + \ell(cd \mid S_{|abcd})$$

For a more rigorous argument of this fact, one may expand each of the lengths appearing in $P(\overline{T} \mid S)$ using (3.5). For instance, when expanding $\ell(ab \mid S)$ and $\ell(abe \mid S)$, we get (after canceling terms)

$$\ell(ab \mid S) + \ell(abe \mid S) = -h(ab \mid S) + h(abc \mid S) + h(abd \mid S) - h(abcd \mid S)$$

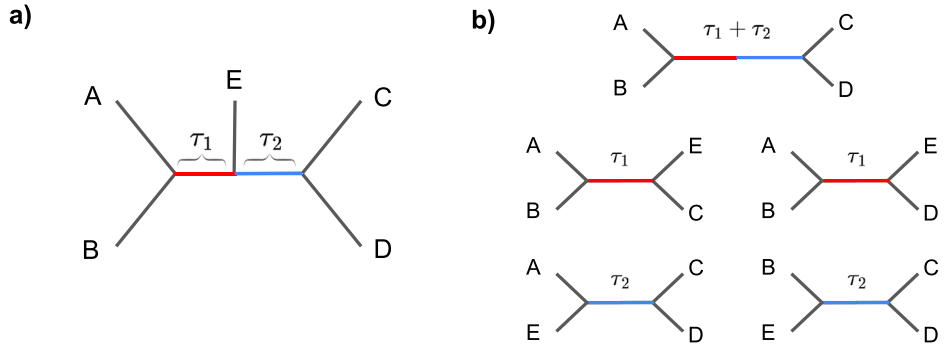

Fig. 6. (a) An unrooted 5-taxa species tree $\overline{S}$ with unrooted topology $\overline{T}_* = ((AB)(CD)E)$ and internal branch lengths τ_1, τ_2. (b) The five quartet trees displayed by the unrooted species tree $\overline{S}$, showing that $\overline{S}_{|ABCD}$ has the longest internal branch among them.

$$
\begin{aligned}
&= -h(ab \mid S_{|abcd}) \\
&+ h(abc \mid S_{|abcd}) + h(abd \mid S_{|abcd}) - h(abcd \mid S_{|abcd}) \\
&= \ell(ab \mid S_{|abcd})
\end{aligned}
$$

where the second line follows as e does not appear in any term $\pm h(V \mid S)$ in the first line (hence, taxon e may be omitted from the species tree). A similar argument shows $\ell(cd \mid S) + \ell(cde \mid S) = \ell(cd \mid S_{|abcd})$.

Since we are now dealing only with the 4-taxa tree $S_{|abcd}$, we are in a position to apply Lemma 1. In particular, $P(\overline{T} \mid S)$ is indeed maximized at $\overline{T}_* = ((AB)(CD)E)$, because (1) $\overline{S}_{|ABCD}$ displays the quartet $AB|CD$, and (2) the length of the internal branch of $\overline{S}_{|abcd}$ is maximal when $\{a, b, c, d\} = \{A, B, C, D\}$. (See Fig. 6 for an illustration.) We therefore conclude that concatenated parsimony is statistically consistent in the unrooted 5-taxa case under the MSC + infinite-sites model of evolution as presented in Section 2.

We stress that if the assumptions of the MSC + infinite-sites model of evolution are not met, then we cannot use the above result to make a conclusion about the consistency of parsimony in this case. However, we expect that many of the strict assumptions that we made on the particular details of the infinite-sites model of mutation are not necessary for consistency. For example, in our proof of consistency we have used the fact that as the number of loci sampled grows large, most informative site patterns support the quartet displayed by $\overline{S}_{|abcd}$ over the two alternative quartets. Let us generalize this case to the scenario where the scaled mutation rate $\theta = 4N_e\mu$ is allowed to vary within and across species tree branches (though we will assume it remains bounded between two positive numbers). We could also allow variability across loci, but this requires some additional theoretical care. We consider the expected contribution of the ith locus to the overall cost in (2.2); we want to show that this expected contribution is minimal when the candidate topology is taken to be the quartet displayed by $\overline{S}_{|abcd}$ as compared to the two alternative quartets.

We begin by rescaling all time to be in mutation units, such that mutations fall on the gene tree G_i at constant rate 1. Note that the species tree may no longer be an ultrametric tree with respect to the branch lengths given in mutation units, which complicates our notion of height used throughout the paper. However, we can easily resolve this issue by appending additional length to the external branches of the species tree so that it becomes ultrametric again — this is possible since no informative site patterns result from mutations occurring in these external branches. The rescaling of time also causes the pairwise coalescent rate to vary inversely to scaled mutation rate across the species tree, apparently further complicating our analysis. However, when time is measured in mutation units, sites in the alignment A_i for locus i that support a particular quartet $ab|cd$ within a 4-taxa tree $S_{|abcd}$ is still proportional (or in this case, equal) to $\ell^\theta(ab \mid S_{|abcd}) + \ell^\theta(cd \mid S_{|abcd})$, where we use the "$\theta$" superscript to denote that these expected lengths are in mutation units. We then can apply (3.9), which tells us that $\ell^\theta(ab \mid S_{|abcd}) + \ell^\theta(cd \mid S_{|abcd})$ is maximal when $h^\theta(ab \mid S_{|abcd}) + h^\theta(cd \mid S_{|abcd})$ is minimal. Using this observation, it is easy to verify (without any explicit calculation) that most informative site patterns will still support the quartet displayed by $\overline{S}_{|abcd}$ over the two alternative quartets. For example, if $S_{|abcd}$ has an asymmetric topology $(((ab)c)d)$, then

$$
\begin{aligned}
&h^\theta(ab \mid S_{|abcd}) < h^\theta(bc \mid S_{|abcd}) = h^\theta(ac \mid S_{|abcd}) \\
&h^\theta(cd \mid S_{|abcd}) = h^\theta(ad \mid S_{|abcd}) = h^\theta(bd \mid S_{|abcd})
\end{aligned}
$$

which shows that $h^\theta(ab \mid S_{|abcd}) + h^\theta(cd \mid S_{|abcd})$ is less than $h^\theta(ac \mid S_{|abcd}) + h^\theta(bd \mid S_{|abcd})$ and $h^\theta(ad \mid S_{|abcd}) + h^\theta(bc \mid S_{|abcd})$, as desired.

4.2. The parsimony anomaly zone for the rooted 5-taxa case

There are three possible shapes for rooted, 5-taxa topologies (i.e. three possible topologies up to relabeling). To analyze the consistency/inconsistency of parsimony across all of parameter space, it suffices to choose one labeled representative of each such shape. Accordingly, we let $T_1 = ((((AB)C)D)E), T_2 = (((AB)C)(DE)), T_3 = (((AB)(CD))E)$ be labeled representatives of these three shapes for 5-taxa, and define $[T_i]$ for $i = 1, 2, 3$ to be the collection of all rooted topologies which agree in unlabeled topology with T_i, i.e. $T \in [T_i]$ if and only if there is a permutation π of $\{A, B, C, D, E\}$ such that $\pi(T) = T_i$. We then consider three species trees $S_{5,i} := (T_i, \boldsymbol{x})$ i.e. $S_{5,i}$ is a species tree with topology T_i and the particular branch length assignment $\boldsymbol{x}$ demonstrated in Fig. 7.

To begin, we can show that concatenated parsimony fails to be statistically consistent when the species tree has sufficiently short internal branch lengths. Applying the decomposition introduced earlier under a star tree (Section 3.2), T_3 has a lower average cost per locus than T_1 and T_2:

$$C(T_3 \mid \star) = \frac{43}{10} < \frac{13}{3} = C(T_1 \mid \star) = C(T_2 \mid \star)$$

Owing to the symmetry of the star tree, the choice of labeled representatives T_1, T_2, T_3 from $[T_1], [T_2], [T_3]$ is clearly irrelevant in the above relations. Thus, all 15 candidate topologies with $T \in [T_3]$ are preferred over a true species tree topology $T_* \in [T_1] \cup [T_2]$ for sufficiently short internal branch lengths $\boldsymbol{x}$. However, this argument does not characterize the exact regions where concatenated parsimony will fail to be consistent.

To visualize the region of inconsistency, we define $K(S)$ to be the number of candidate topologies preferred over the true species tree topology T_*:

$$K(S) := \sum_{T \neq T_* \in \mathbb{T}_5} 1\{C(T \mid S) \leq C(T_* \mid S)\}$$

We call a candidate topology T that contributes $+1$ to $K(S)$ as a parsimony anomalous gene tree (PAGT) for S. The values of $K(S)$, considered for a fixed species tree topology T_* as a function of $\boldsymbol{x}$, form what we will call the parsimony anomaly zone for the topology T_*.

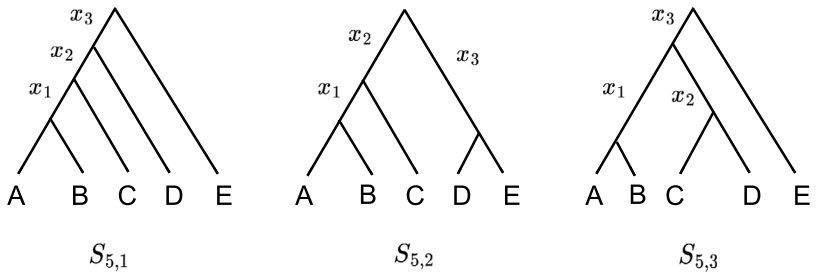

Fig. 7. The species trees $S_{5,1}, S_{5,2}, S_{5,3}$, with respective topologies T_1, T_2, T_3 and branch lengths $\boldsymbol{x} = (x_1, x_2, x_3)$. External branch lengths are irrelevant for analysis and are thus omitted.

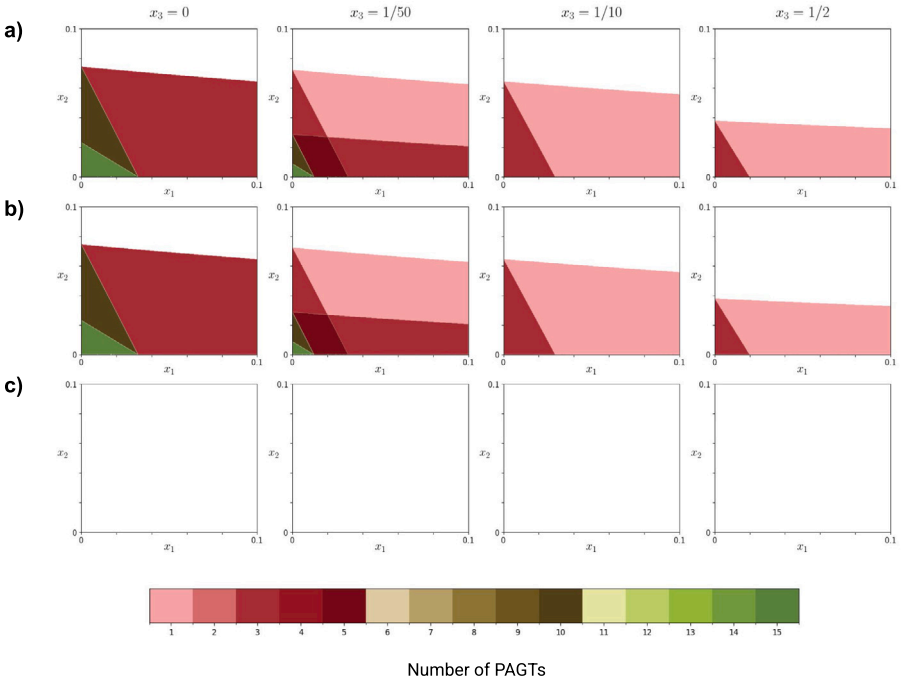

Fig. 8. Number of topologies preferred by parsimony over the true species tree topology for (a) T_1, (b) T_2, (c) T_3, visualized a function of the internal branch lengths $\boldsymbol{x}$ of the corresponding species trees $S_{5,1}$, $S_{5,2}$ and $S_{5,3}$. Each column has a fixed value of x_3 (given in coalescent units), with x_1 and x_2 variable across $[0, 0.1]$ coalescent units in each plot.

To visualize the parsimony anomaly zone for the three representative species tree topologies T_1, T_2, T_3, we computed $K(S_{5,i})$ as a function of $\boldsymbol{x}$ for $i = 1, 2, 3$, varying $x_3 \in \{0, 1/50, 1/10, 1/2\}$ and varying x_1, x_2 each across 400 uniformly spaced values in $[0, 0.1]$. We interpolated between grid points to generate filled contours using the *contourf()* function in *matplotlib* (Hunter, 2007). The results are given in Fig. 8.

From Fig. 8, we see that when the true species tree topology T_* has an unlabeled topology agreeing with T_3, parsimony always prefers the true topology T_*, regardless of the branch lengths chosen. Unfortunately, researchers do not know when this topology is the true one, and so this fact on its own is not useful for applications. Meanwhile, when the species tree topology T_* is in $[T_1] \cup [T_2]$, there are regions in which parsimony prefers other topologies over the true species tree topology, i.e. regions in which parsimony will be statistically inconsistent.

By coincidence, the species tree topologies T_1 and T_2 appear to share an identically structured parsimony anomaly zone with the choice of branch lengths assignment for $S_{5,1}$ and $S_{5,2}$, as seen in Fig. 7. However, the particular topologies that are anomalous may differ by region. In either case, we can observe there are exactly 15 PAGTs near $\boldsymbol{x} = \boldsymbol{0}$, exactly as suggested by the star tree argument. It is then interesting to ask which (if any) of the 15 topologies in $[T_3]$ are maximally anomalous for T_i $(i = 1, 2)$, in the sense that they are always preferred over any other labeled variant in $[T_3]$ regardless of the choice of branch lengths $\boldsymbol{x}$ in $S_{5,i}$. It turns out that this may be answered relatively easily, due to the expected cost per locus $C(T \mid S_{5,i})$ for $i = 1, 2$ taking a surprisingly similar form for all $T \in T_3$. Recall the decomposition of the expected cost per locus into a 'species' and 'coalescence' term as in Section 3.2:

$$C(T \mid S) = C_{\text{sp}}(T \mid S) + C_{\text{coa}}(T \mid S)$$

We have verified (see the Jupyter notebook code in Data availability) that for a species tree topology $T_* \in [T_1] \cup [T_2]$, the coalescent term $C_{\text{coa}}(T \mid S_{5,i})$ is identical for all $T \in [T_3]$ for fixed $i = 1, 2$, and this term even agrees between the two species trees $S_{5,1}, S_{5,2}$ (which helps explain the identically shaped parsimony anomaly zones). In particular, letting $(X_1, X_2, X_3) = (\exp(-x_1), \exp(-x_2), \exp(-x_3))$, we have that for all $T \in [T_3]$,

$$\begin{aligned} C_{\text{coa}}(T \mid S_{5,1}) &= C_{\text{coa}}(T \mid S_{5,2}) \\ &= X_1 + X_2 + X_3 + \frac{1}{2}X_1X_2 + \frac{1}{2}X_2X_3 + \frac{1}{4}X_1X_2X_3 + \frac{1}{20}X_1X_2^3X_3 \end{aligned}$$

In words, this result tells us that on average, ILS and gene tree discordance causes an equal additional parsimony cost to each $T \in [T_3]$ for a species tree with topology $T_* \in [T_1] \cup [T_2]$. Therefore, the maximally anomalous topology $T \in [T_3]$ for $T_* = T_i$ $(i = 1, 2)$ will be the one that minimizes the species term $C_{\text{sp}}(T \mid S_{5,i})$, i.e. the topology in $[T_3]$ that would be the most parsimonious in the absence of gene

tree heterogeneity. For $S_{5,1}$, this implies that the maximally anomalous topology is T_3, with $C_{sp}(T_3 \mid S_{5,1}) = x_1+2x_2+x_3$. This expression follows since the site patterns $AB, ABCD$ both have a parsimony cost of 1 on T_3, while the site pattern ABC has a parsimony cost of 2 on T_3. Meanwhile, for the species tree $S_{5,2}$, the topology $\widetilde{T}_3 = (((AB)(DE))C)$ is maximally anomalous.

Using the fact that T_3 is maximally anomalous for T_1, finding the exact shape of the parsimony anomaly zone may be found by solving $C(T_1 \mid S_{5,1}) \geq C(T_3 \mid S_{5,1})$, obtaining an inequality on x_1 in terms of x_2, x_3:

$$x_1 \leq \log\left[\frac{1 - X_2 + X_2^3 X_3/10}{3x_2 - X_3 + X_3 X_2}\right] \tag{4.1}$$

Note the unexponentiated branch lengths x_1, x_3 do not appear on the right hand side of (4.1) while x_2 does, as a result of the fact that $C_{sp}(T_3 \mid S_{5,1}) - C_{sp}(T_1 \mid S_{5,1}) = x_2$. Similar bounds may be found that define the exact regions in which other candidate topologies $T \in [T_3]$ are anomalous. We can also partially visualize the overall geometry of the anomaly zone by calculating $K(S_{5,1}), K(S_{5,2})$ in the degenerate cases when one internal branch length $x_i = 0$. In particular, we vary the other two internal branch lengths x_j, x_k across $[0, 0.1]$ coalescent units, sampling 400 uniformly spaced values for each x_j, x_k. The results are shown in Fig. 9. We report the results for $S_{5,1}$ only, since we once again observed in our data that both $S_{5,1}$ and $S_{5,2}$ gave an identically shaped parsimony anomaly zone. Of particular interest in Fig. 9 is that when $x_2 = 0$, there is always at least one PAGT (namely, the maximally anomalous topology T_3) regardless of how large x_1 and x_3 are. This demonstrates that it is in general necessary for all branch lengths x_i to exceed some critical threshold $x_{\min}$ in order to guarantee the consistency of parsimony. To find this threshold, assume we have a species tree $S = (T_1, \boldsymbol{x})$ with all internal branches having the same length ($x_1 = x_2 = x_3 > 0$). Making this substitution in (4.1) and setting both sides equal, we find a solution of $x_{\min} \approx 0.062205$ coalescent units. In comparison, for a species tree with topology T_1, Rosenberg and Tao (2008) showed that the critical threshold of minimum branch length for the most probable gene tree topology to agree with the species tree topology is $x_{\min} \approx 0.1935$ coalescent units, so parsimony still outperforms the democratic vote method by a non-trivial margin in the rooted 5-taxa case.

4.3. The parsimony anomaly zone for the unrooted 6-taxa case

When computing the number of PAGTs in the unrooted 6-taxa case, the definition of the number of PAGTs, $K(S)$, must be updated to use unrooted topologies of $\overline{\mathbb{T}}_6$:

$$K(S) = \sum_{\overline{T} \neq \overline{T}_* \in \overline{\mathbb{T}}_6} 1\{C(\overline{T} \mid S) \leq C(\overline{T}_* \mid S)\}$$

We may again use the idea of reducing to a star tree to show the inconsistency of parsimony in this case: consider two unrooted topologies $\overline{T}_1, \overline{T}_2$ with shapes as shown in Fig. 10. The expected costs that arise when S is a star tree amount to

$$C(\overline{T}_2 \mid \star) = \frac{43}{10} < \frac{13}{3} = C(\overline{T}_1 \mid \star)$$

Therefore, parsimony fails to be statistically consistent for the unrooted 6-taxa case; if the true unrooted topology is $\overline{T}_* = \overline{T}_1$, then parsimony will always prefer one of the 15 possible unrooted topologies with shape agreeing with $\overline{T}_2$ for sufficiently short branch lengths $\boldsymbol{x}$. This preference was already known (Roch and Steel, 2015, Equation 5): in particular, it has been demonstrated that under a JC69 model of mutation, the expected difference in parsimony score per locus between $\overline{T}_1$ and $\overline{T}_2$ in the limit of a star tree is $\frac{\theta}{60} + O(\theta^2)$, where $\theta/2$ is the scaled mutation rate. This matches with our result, since we have

$$\frac{\theta}{2} \cdot \left[C(\overline{T}_1 \mid \star) - C(\overline{T}_2 \mid \star)\right] = \frac{\theta}{2} \cdot \frac{1}{30} = \frac{\theta}{60}.$$

To visualize where exactly parsimony fails, we work with the six possible shapes of rooted 6 taxa topologies. For each, we may in theory compute the number of PAGTs for each for any given choice of branch lengths x_1, x_2, x_3, x_4. However, since we have four degrees of freedom in choosing these branch lengths, performing a full analysis of parameter space (say by fixing two of the x_i and varying the other two) to get a plot analogous to Fig. 8 is impractical. Instead, we create a plot of $K(S)$ in the case where all branch lengths are identical ($x_1 = x_2 = x_3 = x_4 > 0$) for each of the six shapes of topologies. The result is given in Fig. 11.

We can again see that, identically to the rooted 5-taxa case, once the minimum branch length exceeds $x_{\min} \approx 0.062205$, unrooted parsimony is guaranteed to be consistent for any unrooted 6-taxa topology. The numerical agreement of the $x_{\min}$ needed for consistency between the rooted 5-taxa and unrooted 6-taxa case is perhaps not terribly surprising: we conjecture that such a result holds true between the rooted n-taxa and unrooted $(n+1)$-taxa cases for all $n \geq 5$.

5. Discussion

Prior to the publication of Kubatko and Degnan (2007), maximum likelihood (ML) analyses of concatenated datasets dominated phylogenetics. While statistically inconsistent estimators may still be preferred over consistent ones, particularly when the main concern is model error and/or data is limited, the demonstration in Kubatko and Degnan (2007) that concatenated ML was inconsistent when ILS was high nonetheless caused a huge explosion of research into methods that are robust to gene tree discordance (e.g. ASTRAL Mirarab et al., 2014; Zhang et al., 2018, MP-EST Liu et al., 2010, STAR Liu and Edwards, 2009). Rather than concatenate all loci, these methods instead consider each gene tree separately. Despite theoretical guarantees of consistency, such gene tree-based methods may suffer because of errors in inferring individual tree topologies from short sequences (Molloy and Warnow, 2018). Because longer (and likely therefore concatenated) alignments offer several advantages over shorter sequences (reviewed in Bryant and Hahn, 2020), there is still a desire for concatenation methods that are robust to ILS.

Results in both Liu and Edwards (2009) and Mendes and Hahn (2018) found that concatenated parsimony under an infinite-sites model was consistent when applied to rooted 4-taxa trees, and we have confirmed the consistency of concatenated parsimony in the unrooted 5-taxa case under the same conditions. While concatenated parsimony had shown to be inconsistent when applied to unrooted trees with 6+ taxa in Roch and Steel (2015), it was not entirely clear if this result was applicable to any biologically realistic species trees, as argued in Bryant and Hahn (2020). The results presented here confirm that concatenated parsimony is inconsistent for a small, but non-trivial region of tree space for 5 or more taxa in the rooted case, or 6 or more taxa in the unrooted case. In this sense, the results of consistency for concatenated parsimony in the rooted 4-taxa/ unrooted 5-taxa cases appear to be a coincidence owing to the both the low-dimensionality of tree space and simplicity of the parsimony cost function in these scenarios. For instance, in the unrooted 5-taxa case, there is only one possible species tree topology up to relabeling, and to infer this topology it suffices to find the quartet $ab|cd$ with the longest internal branch length in the unrooted species tree. Concatenated parsimony does this successfully by determining the quartet $ab|cd$ supported by the most informative site patterns.

Although directly applying parsimony no longer appears to be a viable option for consistent inference from concatenated data, there are other options for statistically consistent estimation of the species tree topology. Since there is no anomalous region for parsimony in the unrooted 4-taxa case under a MSC + infinite-sites model of mutation (Mendes and Hahn, 2018; Molloy et al., 2022), it is possible to use parsimony to estimate the quartet $ab|cd$ for each choice of four taxa a, b, c, d, and then to find a tree that agrees with the greatest

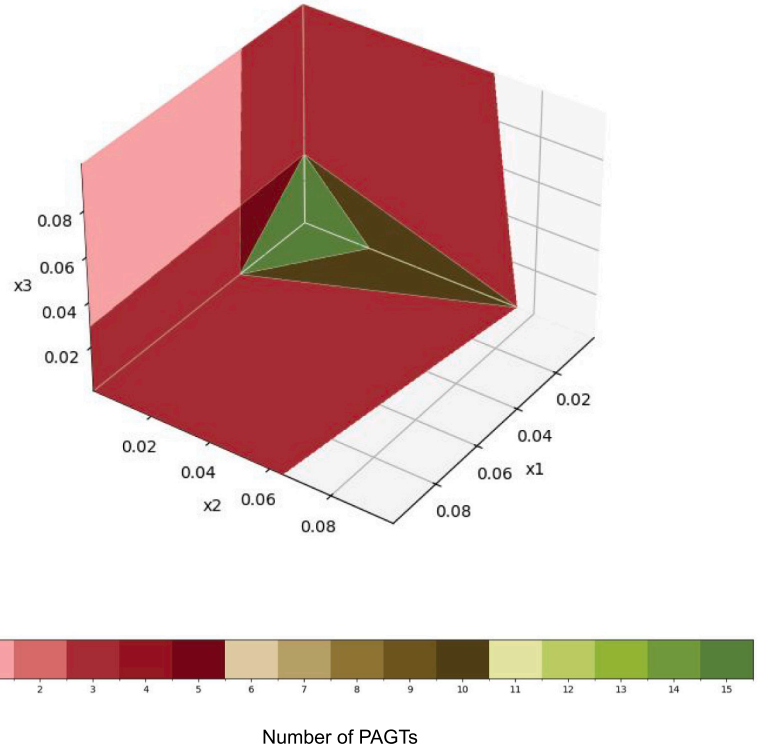

Fig. 9. Number of topologies preferred by parsimony over the true species tree topology for the species tree $S_{5,1}$ as shown in Fig. 7. The three surfaces shown correspond to the cases $x_i = 0$ for $i = 1, 2, 3$, with the other two internal branch lengths x_j, x_k variable across $[0, 0.1]$ coalescent units.

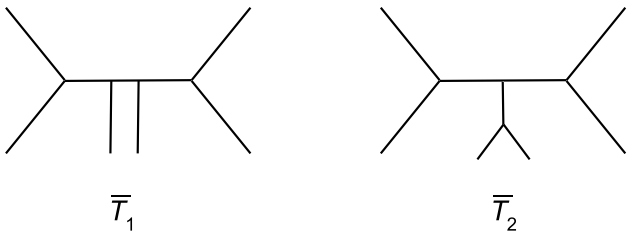

Fig. 10. The two possible unrooted binary tree shapes for 6 taxa. Labels are omitted: $\overline{T}_1$ and $\overline{T}_2$ may be taken to be any unrooted labeled topologies with these respective shapes.

number of inferred trees, which is the approach taken by the methods SDPQuartets and ASTRAL-BP (Springer et al., 2020). Alternatively, for more general models of mutation, one may use SVDQuartets (Chifman and Kubatko, 2014, 2015) to infer quartets, but this requires direct iteration through a significant portion of the $\binom{n}{4}$ possible choices of four taxa in the n-taxa case to guarantee accurate reconstruction, which may become prohibitively computationally expensive for large n. A comparable approach is CASTER (Zhang et al., 2025), a generalization of the concatenated counting methods examined in this paper to more general models of mutation. CASTER estimates quartets in a similar manner to parsimony, though with the addition of a negative weight to some parsimony uninformative site patterns. One major advantage of CASTER over SVDQuartets is that the optimization step of CASTER avoids listing all $\binom{n}{4}$ possible quartets.

Meanwhile, concatenated distance methods, which only require iterating through the $\binom{n}{2}$ pairs of taxa in the concatenated alignment, have shown great theoretical promise for the consistent estimation. Both (Liu and Edwards, 2009) and Mendes and Hahn (2018) found that concatenated neighbor joining (NJ) was consistent on rooted 4-taxa

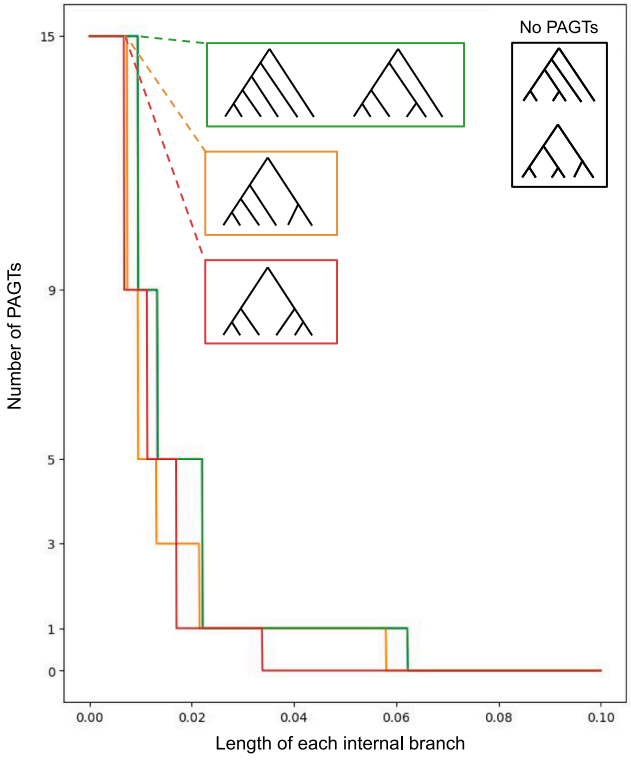

Fig. 11. The number of parsimony anomalous gene trees (PAGTs) (vertical axis) for each possible rooted 6-taxa topology up to labeling, when all branch lengths have the same length (horizontal axis, coalescent units). Four topologies have any PAGTs, while two do not.

trees in the presence of ILS, and positive theoretical results regarding the consistency of concatenated distance methods were later examined in Dasarathy et al. (2015) (which proposed the method METAL) and Allman et al. (2019) (which examined the use of the log-det distance). Despite these encouraging results, a recent study has found that both SVDQuartets and METAL do not perform as well as concatenated ML when using similar amounts of data in a real-life avian dataset (Braun et al., 2024). As an alternative to SVDQuartets and concatenated distance methods, the mixtures across sites and trees (MAST) model of Wong et al. (2024) has all of the advantages of concatenated ML, but allows the alignment to come from a set of alternative topologies. This approach has been found to be consistent in simulations, but more theoretical work is needed to prove its consistency more broadly.

In order to examine the regions of inconsistency for concatenated parsimony, we have introduced a new mathematical method for estimating the total length of branches subtending a given sub-tree over a large number of independent and identically distributed gene trees. Under an infinite-sites model, this total length is proportional to the expected number of the corresponding informative site pattern in a concatenated alignment across all gene trees (Mendes and Hahn, 2018). While this method still involves the calculation of an exponentially large number of terms and is not currently feasible for usage with large numbers of taxa, it nonetheless simplifies the derivation of analytical results for relatively small numbers of taxa. It is also important to point out that its relevance here depends on the infinite-sites assumption, though not necessarily on the assumption of a strict molecular clock. Future work using alternative mutation models (e.g. Jukes–Cantor) may allow our approach to be used in a wider range of scenarios and to be compared directly with work using alternative approaches (Roch and Steel, 2015; Zhang et al., 2025).

CRediT authorship contribution statement

Daniel A. Rickert: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Louis Wai-Tong Fan:** Writing – review & editing, Funding acquisition, Formal analysis. **Matthew W. Hahn:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by National Science Foundation, United States grants DBI-2146866, DMS-2534011, DMS-2532674, DMS-2152103, DMS-2348164. We thank Dr. Noah Rosenberg and several anonymous reviewers whose comments helped improve this work.

Data availability

Code used in the creation of quantitative figures and results is available through the folowing GitHub repository: https://github.com/darickert/exp-branch-lengths.

References

Allman, E.S., Degnan, J.H., Rhodes, J.A., 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., Degnan, J.H., Rhodes, J.A., 2018. Split probabilities and species tree inference under the multispecies coalescent model. Bull. Math. Biol. 80, 64–103.

Allman, E.S., Long, C., Rhodes, J.A., 2019. Species tree inference from genomic sequences using the log-det distance. SIAM J. Appl. Algebra Geom. 3 (1), 107–127.

Braun, E.L., Oliveros, C.H., White Carreiro, N.D., Zhao, M., Glenn, T.C., Brumfield, R.T., Braun, M.J., Kimball, R.T., Faircloth, B.C., 2024. Testing the mettle of METAL: A comparison of phylogenomic methods using a challenging but well-resolved phylogeny. http://dx.doi.org/10.1101/2024.02.28.582627, bioRxiv.

Bryant, D., Hahn, M.W., 2020. The concatenation question. In: Scornavacca, C., Delsuc, F., Galtier, N. (Eds.), Phylogenetics in the Genomic Era. pp. 3.4:1–3.4:23, chapter 3.4, Self Published.

Camin, J.H., Sokal, R.R., 1965. A method for deducing branching sequences in phylogeny. Evolution 19 (2), 311–326.

Chifman, J., Kubatko, L., 2014. Quartet inference from snp data under the coalescent model. Bioinformatics 30 (23), 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. Theoret. Biol. 374, 35–47.

Dasarathy, G., Nowak, R., Roch, S., 2015. Data requirement for phylogenetic inference from multiple loci: A new distance method. IEEE/ACM Trans. Comput. Biology Bioinform. 12 (2), 422–432. http://dx.doi.org/10.1109/TCBB.2014.2361685.

Degnan, J.H., 2013. Anomalous unrooted gene trees. Syst. Biol. 62 (4), 574–590.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2 (5), e68.

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

Doronina, L., Churakov, G., Kuritzin, A., Shi, J., Baertsch, R., Clawson, H., Schmitz, J., 2017. Speciation network in Laurasiatheria: retrophylogenomic signals. Genome Res. 27 (6), 997–1003.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63 (1), 1–19.

Efromovich, S., Salter Kubatko, L., 2008. Coalescent time distributions in trees of arbitrary size. Stat. Appl. Genet. Mol. Biol. 7 (1), 2.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3 (1), 87–112.

Farris, J.S., 1970. Methods for computing Wagner trees. Syst. Biol. 19 (1), 83–92.

Farris, J.S., 1977. Phylogenetic analysis under Dollo's law. Syst. Biol. 26 (1), 77–88.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27 (4), 401–410.

Felsenstein, J., 1983. Parsimony in systematics: biological and statistical issues. Annu. Rev. Ecol. Syst. 14 (1), 313–333.

Fu, Y.-X., 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 172–197.

Helmkamp, L.J., Jewett, E.M., Rosenberg, N.A., 2012. Improvements to a class of distance matrix methods for inferring species trees from gene trees. J. Comput. Biol. 19 (6), 632–649.

Hibbins, Mark S., Hahn, Matthew W., 2022. Distinguishing between histories of speciation and introgression using genomic data, biorxiv. http://dx.doi.org/10.1101/2022.09.07.506990, https://www.biorxiv.org/content/early/2022/09/09/2022.09.07.506990, https://www.biorxiv.org/content/early/2022/09/09/2022.09.07.506990.full.pdf.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7 (1), 1–44.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95. http://dx.doi.org/10.1109/MCSE.2007.55.

Jewett, E.M., 2020. Fast and accurate approximation of the joint site frequency spectrum of multiple populations. http://dx.doi.org/10.1101/2020.05.01.073213, bioRxiv, URL https://www.biorxiv.org/content/early/2020/05/28/2020.05.01.073213.

Jewett, E.M., Rosenberg, N.A., 2012. iGLASS: an improvement to the GLASS method for estimating species trees from gene trees. J. Comput. Biol. 19 (3), 293–315.

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13 (3), 235–248.

Kluge, A.G., Farris, J.S., 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18 (1), 1–32.

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

Le Quesne, W.J., 1974. The uniquely evolved character concept and its cladistic application. Syst. Zool. 23 (4), 513–517.

Le Quesne, W.J., 1977. The uniquely evolved character concept. Syst. Zool. 26 (2), 218–220.

Liu, L., Edwards, S.V., 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol. 58 (4), 452–460.

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 (5), 468–477.

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

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55 (1), 21–30.

Mehta, R.S., Bryant, D., Rosenberg, N.A., 2016. The probability of monophyly of a sample of gene lineages on a species tree. Proc. Natl. Acad. Sci. 113 (29), 8002–8009.

Mehta, R.S., Rosenberg, N.A., 2019. The probability of reciprocal monophyly of gene lineages in three and four species. Theor. Popul. Biol. 129, 133–147.
Mehta, R.S., Steel, M., Rosenberg, N.A., 2022. The probability of joint monophyly of samples of gene lineages for all species in an arbitrary species tree. J. Comput. Biol. 29 (7), 679–703.
Mendes, F.K., Hahn, M.W., 2018. Why concatenation fails near the anomaly zone. Syst. Biol. 67, 158–169.
Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30 (17), i541–i548.
Molloy, E.K., Gatesy, J., Springer, M.S., 2022. Theoretical and practical considerations when using retroelement insertions to estimate species trees in the anomaly zone. Syst. Biol. 71 (3), 721–740.
Molloy, E.K., Warnow, T., 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67 (2), 285–303.
Mossel, E., Roch, S., 2008. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Trans. Comput. Biol. Bioinform. 7 (1), 166–171.
Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5 (5), 568–583.
Pitman, J., 1999. Coalescents with multiple collisions. Ann. Probab. 27, 1870–1902.
Rannala, B., Leache, A., Edwards, S., Yang, Z., 2020. The multispecies coalescent model and species tree inference. In: Scornavacca, C., Delsuc, F., Galtier, N. (Eds.), Phylogenetics in the Genomic Era. pp. 3.3:1–3.3:21, chapter 3.3, Self Published.
Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164 (4), 1645–1656.
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.
Rosenberg, N.A., 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57 (7), 1465–1477.
Rosenberg, N.A., 2013. Discordance of species trees with their most likely gene trees: a unifying principle. Mol. Biol. Evol. 30 (12), 2709–2713.
Rosenberg, N.A., Tao, R., 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57 (1), 131–140.
Schweinsberg, J.R., 2000. Coalescents with Simultaneous Multiple Collisions, vol. 5.
Spence, J.P., Kamm, J.A., Song, Y.S., 2016. The site frequency spectrum for general coalescents. Genetics 202 (4), 1549–1561.
Springer, M.S., Molloy, E.K., Sloan, D.B., Simmons, M.P., Gatesy, J., 2020. Ils-aware analysis of low-homoplasy retroelement insertions: inference of species trees and introgression using quartets. J. Hered. 111 (2), 147–168.
Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26, 119–164.
Than, C.V., Rosenberg, N.A., 2011. Consistency properties of species tree inference by minimizing deep coalescences. J. Comput. Biol. 18 (1), 1–15.
Wong, T.K., Cherryh, C., Rodrigo, A.G., Hahn, M.W., Minh, B.Q., Lanfear, R., 2024. MAST: Phylogenetic inference with mixtures across sites and trees. Syst. Biol. 73, syae008.
Zhang, C., Nielsen, R., Mirarab, S., 2025. CASTER: Direct species tree inference from whole-genome alignments. Science eadk9688.
Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 15–30.