Terraces in Species Tree Inference from Gene Trees

Mursalin Habib[1], Atif Hasan Rahman[1], and Md. Shamsuzzoha Bayzid[1,*]

[1]Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka-1205, Bangladesh.
[*]Corresponding author: shams_bayzid@cse.buet.ac.bd

Abstract

A terrace in a phylogenetic tree space is a region where all trees contain the same set of subtrees, due to certain patterns of missing data among the taxa sampled, resulting in an identical optimality score for a given data set. This was first investigated in the context of phylogenetic tree estimation from sequence alignments using maximum likelihood (ML) and maximum parsimony (MP). The concept of terraces was later extended to the species tree inference problem from a collection of gene trees, where a set of equally optimal species trees was referred to as a "pseudo" species tree terrace. Pseudo terraces do not consider the topological proximity of the trees in terms of the induced subtrees resulting from certain patterns of missing data. In this study, we mathematically characterize species tree terraces and investigate the mathematical properties and conditions that lead multiple species trees to induce/display an identical set of locus-specific subtrees owing to missing data, therefore, as a consequence of combinatorial structure, resulting in the same optimality score. We present analytical results and combinatorial characteristics of species tree terraces. We report that species tree terraces are agnostic to gene tree topologies and the discordance therein, considering which is central to developing statistically consistent species tree estimation techniques using gene tree distributions. Therefore, we introduce and characterize a special type of gene tree topology-aware terrace which we call "peak terrace", highlight its impact and importance in species tree inference, and investigate conditions on the patterns of missing data that give rise to peak terraces.

1 Introduction

Species tree estimation is frequently based on phylogenomic approaches that use multiple genes from throughout the genome. The estimation of species trees from multiple genes is necessary since true gene trees can differ from each other and from the true species tree due to various processes, including gene duplication and loss, horizontal gene transfer, and incomplete lineage sorting (ILS) [1]. A traditional approach to species tree estimation from multi-locus data is called concatenation (also known as combined analysis), where alignments are estimated for each gene and concatenated into a supermatrix, which is then used to estimate the species tree using a sequence based tree estimation technique (e.g., maximum parsimony, maximum likelihood etc.). The concatenation approach, which is agnostic to the topological differences among the gene trees, can be statistically inconsistent [2] and can return incorrect trees with high confidence [3–6].

As a result, "summary methods", that operate by summarizing estimated gene trees and can explicitly take gene tree discordance/heterogeneity into account are becoming increasingly popular [7–18]. Fundamental to most of these summary methods is the ability to search the tree space under certain optimization criteria (e.g., maximizing pseudo likelihood score [12], maximizing quartet score [11, 18, 19], maximizing triplet score [17], minimizing deep coalescence [20]). As the size of the tree space grows exponentially with the number of taxa, finding the optimal species tree with respect to a particular optimization criterion is challenging. Moreover, the presence of local optima and multiple optimal solutions make the tree search even more complicated.

A concept related to the presence of multiple optimal solutions is called "phylogenetic terraces" – regions of the tree space where all trees have the same score due to certain patterns of missing data. Sanderson *et al.* [21] first formally investigated this and showed that when phylogenetic trees are estimated from sequence alignments using maximum likelihood (ML), multiple distinct trees can have exactly the same likelihood score due to missing data (i.e., missing genes) – a phenomenon which was referred to as terraces

and was further investigated in subsequent studies [22–25]. Farah *et al.* [26] showed that similar phenomenon can arise when species trees are estimated by summarizing a collection of gene trees. They introduced the concept of "pseudo species tree terrace", where potentially large numbers of distinct species trees may have exactly the same optimality score with respect to a set of input gene trees. There could be many reasons for multiple trees to have an identical score, but the trees in a terrace are indistinguishable in an important way: they "display" the same set of subtrees which subsequently results in identical optimality scores. For a species tree T and and a locus/gene tree gt, the locus-specific induced subtree of T is obtained by pruning the taxa in T that are missing in gt. Two topologically different trees T_1 and T_2 can induce the same locus-specific subtree due to certain patterns of missing data. This type of topological closeness in terms of identical sets of induced subtrees was not considered in pseudo species tree terraces (this is why it was called a "pseudo" terrace). Therefore, pseudo species tree terraces can arise even without the presence of missing data.

The discovery of species tree terraces has implications for summary methods that navigate through and score the trees within the tree space. Because all of the trees within a species tree terrace have the same optimization score, recognizing a terrace may help reduce computing efforts by avoiding computation time that would otherwise be spent evaluating many trees with identical scores. However, it is possible that some trees in a terrace are topologically more correct than the other ones, which was systematically analyzed and empirically demonstrated in [26]. As a result, in the presence of potentially large sets of equally optimal trees, detecting terraces and identifying relatively more reliable trees within the terraces and their neighborhoods may improve the performance of tree search algorithms. Indeed, *terrace-aware* data structures led to substantial speedup of RAxML [27, 28] and IQ-tree [29] for estimating ML trees from alignments [23].

The conditions for datasets to have phylogenetic terraces, described in [21], are general and extensible to the gene tree-species context [26, 30]. However, the combinatorial properties and mathematical characterizations of species tree terraces and the charac-

teristics of the input gene trees and missing data patterns that lead to the presence of species tree terraces have not been elucidated in the gene tree-species tree context. In this paper, we mathematically characterize the species tree terrace and investigate various combinatorial properties of terraces. Moreover, we show that species tree terraces are not sensitive to gene tree topologies and their discordance and as a result, one set of species tree acts as a species tree terrace for an extremely large number of different sets of input gene trees. Therefore, we introduce and formalize a special type of gene tree topology-aware species tree terrace "peak terrace", describe its importance, argue why it suffices to only look at them to understand the properties of species tree terraces in general, and investigate conditions on the patterns of missing data and taxon coverage that give rise to peak terraces.

2 Preliminaries

We now define some general terminology we will use throughout this paper; other terminology will be introduced as needed. All trees T that follow are full binary trees with node set $V(T)$, edge set $E(T)$, and leaf set $L(T)$. Let T be a full binary tree and let $X \subseteq L(T)$. The homeomorphic subtree of T induced by X, denoted by $T|X$, is the unique tree obtained by restricting T to the leaf set X and then suppressing all the nodes of degree two in the resulting tree (see Figure 1). If T' is a homeomorphic subtree of T, then we say that T *displays* T'. We consider the *restriction-based* approach [31–33] where an incomplete gene tree (i.e., a tree that can miss some taxa) gt is reconciled with a species tree T by taking the homeomorphic subtree $T' = T|L(gt)$.

The problem of inferring species trees from multi-locus data using summary methods generally involves the following setup: we have as input a sequence $\mathcal{G} = (g_1, g_2, \ldots, g_k)$ of k gene trees such that $L(g_i) \subseteq \mathcal{X}$ for each $i \in \{1, 2, \ldots, k\}$ and we wish to find a species tree T that is optimal with respect to $\mathcal{G}$ according to a predefined scoring function $s_{\mathcal{G}}$. There are many scoring functions of interest such as the extra lineage (due to deep coalescence)

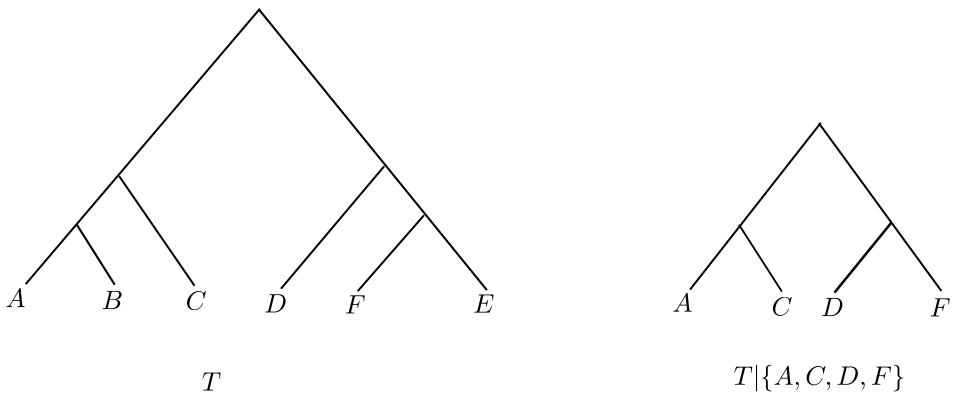

Figure 1: A full binary tree T and its homeomorphic subtree $T|X$ where $X = \{A, C, D, F\}$

score [20], the pseudo-likelihood score [12], triplet score [17], and the quartet score [11]. In this article, we will focus on the quartet score, but our results are general and extensible to other optimization criteria as well. A *quartet* is a binary tree on four leaves. We denote by $ab|cd$ the unrooted quartet tree in which the pair a, b is separated from the pair c, d by an edge. Given a binary tree T with at least four leaves, we denote by $Q(T)$ the set of all quartets displayed by T. The *quartet score* of a species tree T with respect to an input sequence $\mathcal{G} = (g_i)_{i=1}^k$ of gene trees is given by $q_{\mathcal{G}}(T) = \sum_{i=1}^k |Q(T) \cap Q(g_i)|$.

3 Species tree terraces

The concept of phylogenetic terraces [21], originally described for tree estimation from sequence data using maximum likelihood, was later extended to species tree estimation from gene trees using summary approaches in [26], which showed that for a fixed sequence $\mathcal{G}$ of gene trees, there can be potentially large sets of species trees with identical optimality scores. These sets of equally good species trees can arise regardless of the presence of missing data and were referred to as pseudo species terraces (see Definition 3.1).

Definition 3.1 (Pseudo Species Tree Terrace)**.** *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees and let $s_{\mathcal{G}}(\cdot)$ be a scoring function. A pseudo species tree terrace is a set $\mathcal{S}$ of the*

following form.

$$\mathcal{S} = \{T : L(T) = \bigcup_{i=1}^{k} L(g_i) \text{ and } s_{\mathcal{G}}(T) = y \text{ for some } y \in \mathbb{R}_+\}$$

See Figure 2 for an example of a pseudo quartet terrace. Note that every tree in this pseudo terrace has the same quartet score with respect to the input sequence of gene trees. However, some of these trees have the same score due to a very specific reason – they display the same set of subtrees. In order to highlight this point, we define *species tree terraces* which are pseudo species tree terraces with an additional condition.

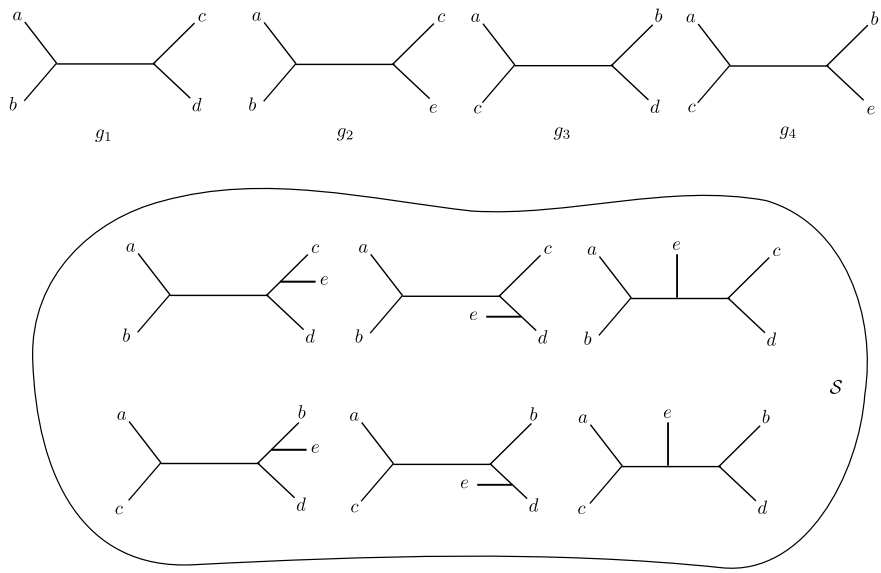

Figure 2: A pseudo quartet terrace $\mathcal{S}$ for the sequence $\mathcal{G} = (g_1, g_2, g_3, g_4)$ of gene trees. Any tree in $\mathcal{S}$ has a quartet score of 2 with respect to $\mathcal{G}$.

Definition 3.2 (Species Tree Terrace)**.** *Given an input sequence $\mathcal{G} = (g_i)_{i=1}^{k}$ of gene trees, two species trees T and T' are said to be in a species trees terrace if both of the following conditions hold.*

1. *T and T' reside in the same pseudo species trees terrace with respect to $\mathcal{G}$, and*
2. *the sequences $(T|L(g_i))_{i=1}^{k}$ and $(T'|L(g_i))_{i=1}^{k}$ are equal.*

In other words, trees in a species tree terrace not only have the same score with respect to $\mathcal{G}$ but also display the same homeomorphic subtrees when restricted to the leaves of

the input gene trees. A few comments about Definition 3.2 are in order. Firstly, most scoring functions we care about turn out to have certain common properties. Given a sequence $\mathcal{G} = (g_i)_{i=1}^k$ of gene trees, we call a scoring function $s_{\mathcal{G}}(\cdot)$ *additive* if the following holds for every species tree T.

1. There exists a function c such that $c(T, g_i) = c\left(T|L(g_i), g_i\right)$ for all $1 \leq i \leq k$, and
2. $s_{\mathcal{G}}(T) = \sum_{i=1}^k c(T, g_i)$.

Clearly, the quartet score $q_{\mathcal{G}}(\cdot)$ is additive with $c(T, g_i)$ being the number of quartets common to both T and g_i i.e., $|Q(T) \cap Q(g_i)|$. The reader is encouraged to verify that other scores such as the extra-lineage score, the triplet score, the gene duplication and loss score [32, 33], and the pseudo likelihood score [12] are also additive.

It turns out that if we are dealing with only additive scoring functions, condition 1 in Definition 3.2 is redundant. In other words, we have the following.

Proposition 3.1 (From [30]). *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees and let $s_{\mathcal{G}}(\cdot)$ be an additive scoring function. If T and T' are species trees with $L(T) = L(T') = \bigcup_{i=1}^k L(g_i)$ that satisfy $(T|L(g_i))_{i=1}^k = (T'|L(g_i))_{i=1}^k$, then $s_{\mathcal{G}}(T) = s_{\mathcal{G}}(T')$.*

Proposition 3.1 allows us to define an "additive version" of a species tree terrace.

Definition 3.3 (Additive Species Tree Terrace). *Given an input sequence $\mathcal{G} = (g_i)_{i=1}^k$ of gene trees, two species trees T and T' with $L(T) = L(T') = \bigcup_{i=1}^k L(g_i)$ are said to be in an additive species trees terrace if $(T|L(g_i))_{i=1}^k = (T'|L(g_i))_{i=1}^k$.*

Since our focus is the quartet score which is an additive scoring function, for the remainder of article we use "species tree terrace" and "additive species tree terrace" interchangeably.

Another point worth noting is that Definition 3.2 does not disallow terraces of cardinality one. In fact, given $\mathcal{G} = (g_i)_{i=1}^k$, every species tree T with $L(T) = \bigcup_{i=1}^k L(g_i)$ is part of some terrace (containing possibly only T itself). Indeed, let T be an arbitrary species tree with $L(T) = \bigcup_{i=1}^k L(g_i)$ and consider the set $\mathcal{S} = \{T' : L(T') = L(T) \text{ and } (T|L(g_i))_{i=1}^k = (T'|L(g_i))_{i=1}^k\}$.

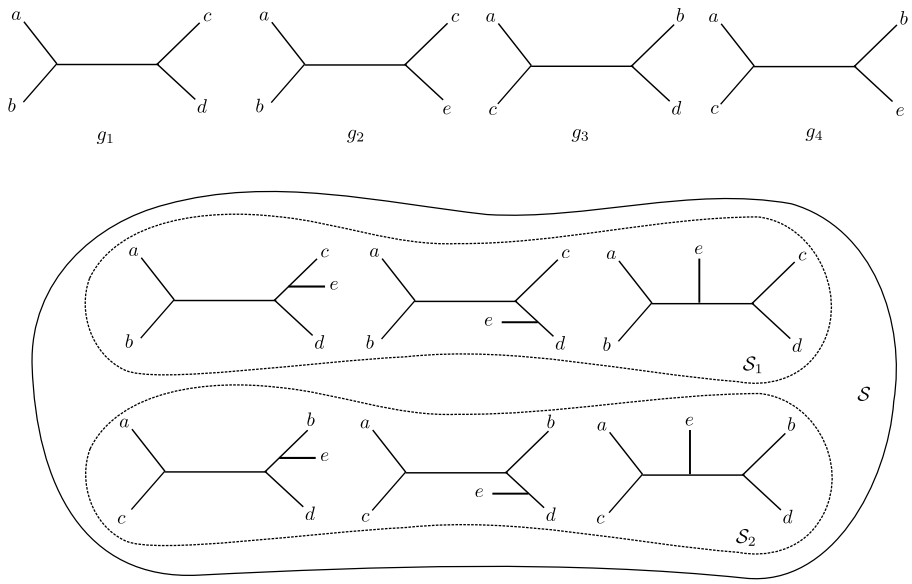

Figure 3: The pseudo quartet terrace of Figure 2 partitioned into two species tree quartet terraces $\mathcal{S}_1$ and $\mathcal{S}_2$.

Clearly, $T \in \mathcal{S}$ and therefore, $|\mathcal{S}| \geq 1$. If $|\mathcal{S}| > 1$, then $\mathcal{S}$ is clearly a species tree terrace. However, note that even if $|\mathcal{S}| = 1$, the conditions in Definition 3.2 are vacuously true and $\mathcal{S}$ is technically a terrace. Such terraces of cardinality one are not very interesting in the context of phylogenomic analyses. Rather a much more interesting task is to investigate the conditions for which there exist more than one tree in a particular terrace.

From Definitions 3.1 and 3.2, it is clear that every pseudo species tree terrace can be partitioned into a set of species trees terraces. In other words, there may be multiple terraces imbedded in a larger pseudo terrace at the same "elevation" in the landscape of tree space. Figure 3 shows how the pseudo quartet terrace $\mathcal{S}$ in Figure 2 is composed of two different quartet terraces $\mathcal{S}_1$ and $\mathcal{S}_2$. All of the trees in $\mathcal{S}_1$ (and likewise in $\mathcal{S}_2$) display identical trees when restricted to the leaf set of a particular gene tree. For example, when restricted to $L(g_1)$, all the trees in $\mathcal{S}_1$ and $\mathcal{S}_2$ display $((a, b), (c, d))$ and $((a, c), (b, d))$, respectively.

Note that due to the extra condition that trees in a species tree terrace must satisfy, terraces of size greater than one can only exist if there is missing data in the input gene trees i.e. for every $1 \leq i \leq k$, $L(g_i) \neq \bigcup_{j=1}^{k} L(g_j)$. This is because if there exists i such that g_i contains data from all the taxa, then for any species tree T, $T|L(g_i) = T$, and the

only tree with leaf set $\bigcup_{j=1}^{k} L(g_j)$ that displays T is T itself. Pseudo species tree terraces have no such restrictions: large pseudo-terraces exist even without the presence of any missing data [26]. Some lower bounds on the sizes of such large pseudo terraces can also be derived as shown below.

Theorem 3.1. *Let $\mathcal{G} = (g_i)_{i=1}^{k}$ be a sequence of gene trees such that for all $1 \leq i \leq k$, $L(g_i) = L$, i.e., there is no missing data in any of the gene trees. If $|L| = n$, then there exists a pseudo quartet terrace of size at least $\frac{(2n-5)!!}{k\binom{n}{4}+1}$.*

Proof. Let $\mathcal{T}_L$ be the set of all unrooted species trees on L. Note that we have $|\mathcal{T}_L| = (2n-5)!!$. There are a total of $\binom{n}{4}$ quartets in a tree with n taxa. Therefore, for a gene tree $g_i \in \mathcal{G}$ and a species tree T, both on the same set of n taxa, T can satisfy at most $\binom{n}{4}$ quartets (this is when g_i and T have an identical topology). Hence, for a set $\mathcal{G}$ of k gene trees, the maximum number of consistent quartets with respect to a species tree T is $k\binom{n}{4}$. Therefore, we have $q_{\mathcal{G}}(T) \in \left\{0, 1, 2, \ldots, k\binom{n}{4}\right\}$. Now for $0 \leq m \leq k\binom{n}{4}$, let $\mathcal{T}_m = \{T \in \mathcal{T}_L : q_{\mathcal{G}}(T) = m\}$. Clearly, the sets $\mathcal{T}_m$ form a partition of $\mathcal{T}_L$ and by the pigeonhole principle, there exists m such that $|\mathcal{T}_m| \geq \frac{(2n-5)!!}{k\binom{n}{4}+1}$. This set $\mathcal{T}_m$ is our required pseudo quartet terrace.

□

Since the numerator in $\frac{(2n-5)!!}{k\binom{n}{4}+1}$ grows much faster with n than the denominator, the sizes of such pseudo terraces can grow very large very quickly even without missing data.

We also note that terraces – unlike pseudo terraces – are agnostic to gene tree topologies and depend solely on taxon coverage (i.e., the taxa sampled in gene trees). This is a consequence of Proposition 3.1 which says that if two trees display the same set of trees, then they have the same score for an additive scoring function. In other words, even if we replace each gene tree g_i with a different gene tree g_i', as long as the leaf sets remain the same, i.e. $L(g_i) = L(g_i')$, two trees T_1 and T_2 that were once in a terrace will remain in a terrace. So, in a very real sense, for species tree terraces, the topologies of the gene trees in the input sequence $\mathcal{G} = (g_i)_{i=1}^{k}$ do not matter. The only thing that matters is

the taxon coverage, i.e., the sequence of leaf sets $(L(g_i))_{i=1}^k$. As long as this sequence of leaf sets in the gene trees is fixed, any two trees that are once in a species tree terrace with respect to $\mathcal{G}$ will always remain in a species tree terrace for $\mathcal{G}'$ (albeit the optimality scores of the terraces could be different due to the topological differences of the gene trees in $\mathcal{G}$ and $\mathcal{G}'$). Therefore, we have the following theorem.

Theorem 3.2. *Let $\mathcal{S}$ be a species tree terrace with respect to $\mathcal{G} = (g_i)_{i=1}^k$ of gene trees and an additive scoring function $s_{\mathcal{G}}(\cdot)$. Then $\mathcal{S}$ is a species tree terrace with respect to every sequence $\mathcal{G}' = (g_i')_{i=1}^k$ where for $1 \leq i \leq k$, $L(g_i) = L(g_i')$.*

Since the input gene tree topologies do not matter for determining whether a species tree belongs to a terrace, one set of species tree acts as a species tree terrace for an extremely large number of sequences of input gene trees as shown in the proposition below.

Proposition 3.2. *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees and let $\mathcal{S}$ be a species tree terrace with respect to $\mathcal{G}$. Let $\mathcal{M}_{\mathcal{S}}$ be the set of all k gene tree sequences $\mathcal{G}'$ for which $\mathcal{S}$ is a species tree terrace. More formally,*

$$\mathcal{M}_{\mathcal{S}} = \{\mathcal{G}' = (g_i')_{i=1}^k : \mathcal{S} \text{ is a species tree terrace with respect to } \mathcal{G}'\}.$$

Then $|\mathcal{M}_{\mathcal{S}}| \geq \prod_{i=1}^k (2|L(g_i)| - 5)!!$.

Proof. Clearly, if $L(g_i) = L(g_i')$ for all $1 \leq i \leq k$, then $\mathcal{S}$ is a species tree terrace with respect to $\mathcal{G}'$. Since there are $(2n-5)!!$ full binary trees on n leaves, the result follows. □

The phylogenetic terrace, originally described by Sanderson *et al.* [21] in the context of estimating phylogenetic trees from sequence alignments, also depends solely on the taxon coverage. This is not a limiting factor for phylogenetic terraces as it is defined for a problem which takes sequence alignments as input, making the taxon coverage the only relevant information from the input data. Summary methods in the gene tree-species tree context, however, take a collection of gene trees – representing the taxon coverage of the

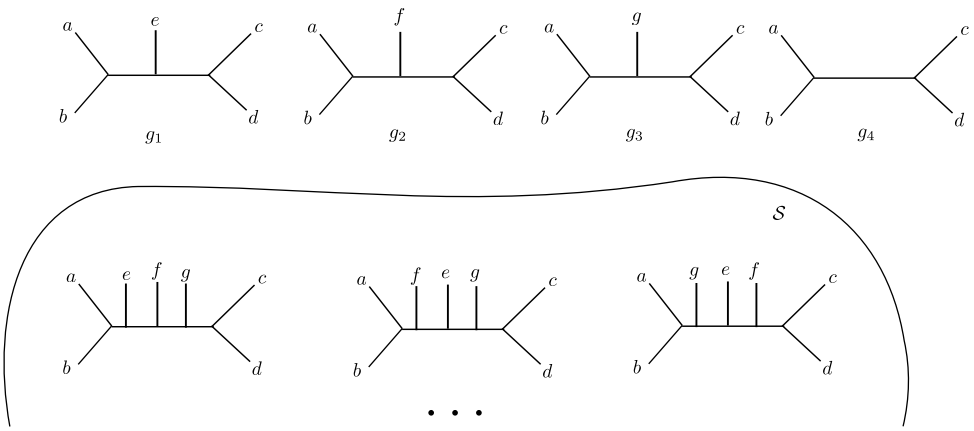

Figure 4: Portion of the peak terrace $\mathcal{S}$ with respect to the sequence $\mathcal{G} = (g_1, g_2, g_3, g_4)$ of gene trees. Every tree in $\mathcal{S}$ displays each one of g_1, g_2, g_3, and g_4.

input gene trees as well as gene tree topologies and their discordance. Taking the gene tree discordance into account is fundamental for estimating species trees from a collection of gene trees using a statistically consistent way. Therefore, it is desirable to have variants of species tree terraces that are sensitive to gene tree topologies in addition to the taxon coverage. In this regard, we introduce a special type of gene tree topology-aware terraces, which we call *peak terraces*.

4 Peak Terraces

Peak terraces are species tree terraces with one extra condition: for any tree T in a peak terrace, the sequence $(T|L(g_i))_{i=1}^k$ must equal $\mathcal{G}$. More formally, we have the following definition.

Definition 4.1 (Peak Terrace)**.** *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees. A set $\mathcal{S}$ is called a peak terrace with respect to $\mathcal{G}$, if for every $T \in \mathcal{S}$, $L(T) = \bigcup_{i=1}^k L(g_i)$ and $(T|L(g_i))_{i=1}^k = \mathcal{G}$.*

Peak terraces are named as such due to the fact that any tree in a peak terrace achieves the optimal score with respect to the input sequence of gene trees. For maximization problems (e.g., maximizing quartet score), the trees in a peak terrace have the highest score, whereas for minimization criteria (e.g., minimizing deep coalescence, minimizing

gene duplication and loss), the trees in a peak terrace achieve the lowest "cost" (i.e., the highest score, where score = −cost). We note that, in the landscape sense, the peak terraces for minimization problems are actually "basins" rather than "peaks". The following proposition shows the optimality of the trees in a peak terrace in the context of quartet scores.

Proposition 4.1. *Given a sequence $\mathcal{G} = (g_i)_{i=1}^k$ of gene trees, any tree T in a quartet peak terrace with respect to $\mathcal{G}$ must have the maximum possible quartet score with respect to $\mathcal{G}$ i.e., $q_{\mathcal{G}}(T) = \sum_{i=1}^k \binom{|L(g_i)|}{4}$.*

Proof. Since T is in a quartet peak terrace, for every $1 \leq i \leq k$, $T|L(g_i) = g_i$. Therefore, $Q(T) \supseteq Q(g_i)$, and so, $q_{\mathcal{G}}(T) = \sum_{i=1}^k |Q(T) \cap Q(g_i)| = \sum_{i=1}^k |Q(g_i)| = \sum_{i=1}^k \binom{|L(g_i)|}{4}$. □

Note that Proposition 4.1 implies that for a fixed sequence $\mathcal{G}$ of gene trees – unlike species tree terraces – there exists exactly one quartet peak terrace. However, we would like to point out that this is true for not only quartet peak terraces but also any s-terrace where s is an additive scoring function. More interestingly, for a particular sequence of gene trees, there is exactly one peak terrace for all possible additive scoring functions.

The concept of peak terraces has impacts on species tree inference from a collection of gene trees using summary methods. Summary methods attempt to find a tree with the optimal score (e.g., quartet score), meaning that when the problem is solved exactly, the output species tree will reside in a pseudo species terrace with the optimal score. The trees in such a terrace are all optimal in terms of an optimization criterion, but they are topologically different and hence have different topological accuracies – posing a challenge for the search algorithms to find comparatively reliable tree from a pool of equally optimal trees. This phenomenon was systematically analyzed and demonstrated by [26], where it was observed that Phylonet [34,35] (a method for estimating species trees by minimizing the number of extra lineages resulting from deep coalescence events) can produce trees with identical or competitive quartet scores as ASTRAL, but ASTRAL is typically substantially more accurate than Phylonet. Note that a pseudo species terrace

with the optimal score is not necessarily a terrace or a peak terrace, but a peak terrace is a pseudo terrace with the optimal score, and both are sensitive to gene tree topologies. Moreover, as we will show in the following, characterizing peak terraces is sufficient for characterizing terraces. Understanding peak terraces may thus potentially help in the development of terrace-aware data structures and algorithms to circumvent the challenges and ambiguity posed by equally good trees, thereby improving tree search strategies for summary methods.

Another reason why peak terraces are important is the fact that every species tree terrace for some input sequence of gene trees is a peak terrace for possibly a different sequence of input gene trees. This different sequence is simply the sequence of gene trees displayed by the trees in the terrace when restricted to the leaves of the original sequence of input gene trees. In other words, we have the following fact.

Fact 4.1. *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees. If $\mathcal{S}$ is a non-empty species tree terrace with respect to $\mathcal{G}$, then $\mathcal{S}$ is also a peak terrace with respect to the sequence $(T|L(g_i))_{i=1}^k$ where $T \in \mathcal{S}$.*

Clearly, by definition, every peak terrace is a species tree terrace. Fact 4.1, however, tells us that every species tree terrace is also a peak terrace. It then follows that the set of all species tree terraces over all sequences of input gene trees is exactly the same as the set of all peak terraces over all sequences of input gene trees. So, if we want to understand the structural properties of terraces, it suffices to focus on the structural properties of peak terraces.

Although gene tree topologies do not matter for species tree terraces, they do matter for pseudo-terraces. Indeed, if one of the input gene trees is changed while keeping the leaf set unchanged, two trees that were once in a pseudo-terrace may cease to remain in it. The same is true for peak terraces too. Unlike species tree terraces which only depends on the leaf sets of the input gene trees, peak terraces are sensitive to the topologies of the gene trees. Moreover, it is not guaranteed that any species tree is a part of a peak

terrace, meaning that the peak terrace can be necessarily empty for some set of inputs. The following fact highlights a simple sufficient condition for this to happen.

Fact 4.2. *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees such that there exists $1 \leq i < j \leq k$ for which $L(g_i) = L(g_j) = \bigcup_{m=1}^k L(g_m)$ but $g_i \neq g_j$. Then the peak terrace of $\mathcal{G}$ is necessarily empty.*

Fact 4.2 tells us if there are two different gene trees that both contain all the taxa, then the peak terrace is necessarily empty. In other words, missing data is a necessity for peak terraces to be non-empty unless the input gene trees are all identical. However, that alone is not sufficient. The condition that all input gene trees be displayed is a rather strong one. A collection of trees is called *compatible* if there exists at least one tree that displays every tree in the collection. We have the following fact about peak terraces.

Fact 4.3. *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of gene trees. The peak terrace of $\mathcal{G}$ is non-empty if and only if $\mathcal{G}$ is compatible.*

Whether a collection of gene trees is compatible can be decided using the BUILD algorithm from [36].

4.1 Finding Patterns of Missing Data that Give Rise to Non-Trivial Peak Terraces

We now focus on the case where the input gene trees do not have any missing data. In other words, for each $1 \leq i \leq k$, $L(g_i) = L = \bigcup_{j=1}^k L(g_j)$. In this case, unless the input gene trees are not all identical, the peak terrace is necessarily empty. Therefore, we seek conditions on the patterns of missing data that cause the peak terrace to be non-empty. More formally, for $1 \leq i \leq k$, we seek $X_i \subseteq L(g_i)$ with $|X_i| \geq 4$ such that there exists at least one tree T with leaf set $\bigcup_{i=1}^k L(g_i)$ that displays each tree in the sequence $(g_i|X_i)_{i=1}^k$. We call such a sequence $(X_i)_{i=1}^k$ a *kernel*. Each kernel specifies a pattern of missing data for which a non-empty peak terrace exists.

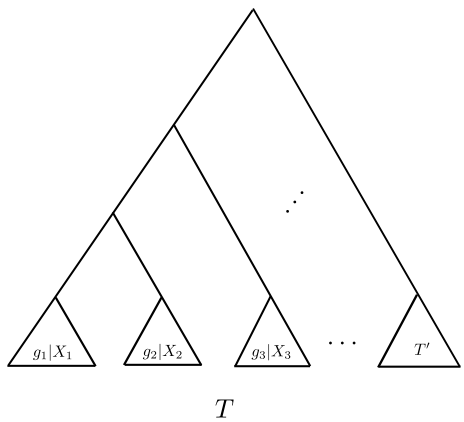

Figure 5: The tree T used in the proof of Proposition 4.1.

We first show that if there are not too many gene trees in the input, then such a pattern of missing data can always be found.

Theorem 4.1. *Let $L = \bigcup_{i=1}^{k} L(g_i)$ be the set of all taxa present in the data. If the number of input gene trees, k, satisfies the inequality $k \leq \frac{|L|}{4}$, then there exists a kernel i.e. a sequence $(X_i)_{i=1}^{k}$ with $|X_i| \geq 4$ such that there exists a T with $L(T) = L$ that displays $(g_i|X_i)_{i=1}^{k}$.*

Proof. Let $n = |L|$ and let $L = \{L_1, L_2, \ldots, L_n\}$. For $1 \leq i \leq k$, we set $X_i = \{L_{4i-3}, L_{4i-2}, L_{4i-1}, L_{4i}\}$ and let $L' = L \setminus \bigcup_{i=1}^{k} X_i$. Note that for each $1 \leq i \leq k$, $g_i|X_i$ is a quartet. Finally, let T' be any full binary tree with leaf set L'. Now consider the tree T in Figure 5, which contains the quartets $g_1|X_1, g_2|X_2, \ldots, g_k|X_k$, and the subtree T'. Clearly, T displays $(g_i|X_i)_{i=1}^{k}$.

□

Theorem 4.1 gives a simple sufficient condition for the existence of a kernel based on the number of input gene trees. One might ask if there is an analogous necessary condition. It turns out there is not: no matter how many distinct gene trees are in the input sequence, there always exists a kernel provided that the gene trees have at least six leaves.

Theorem 4.2. *Let L be a set with $|L| \geq 6$ and let $\mathcal{G} = (g_i)_{i=1}^{k}$ be a sequence of gene trees*

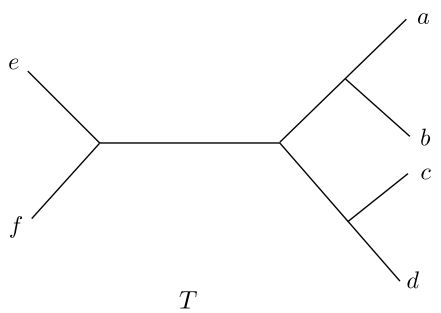

Figure 6: The tree T used in the proof of Lemma 4.1. For every tree T' with $L(T') = L(T)$, $Q(T) \cap Q(T') \neq \emptyset$.

such that for all $1 \leq i \leq k$, $L(g_i) = L$. *Then there exists a kernel i.e. a sequence* $(X_i)_{i=1}^k$ *with* $|X_i| \geq 4$ *such that there exists a tree* T *with* $L(T) = L$ *that displays* $(g_i|X_i)_{i=1}^k$.

To prove Theorem 4.2, we first need the following lemma.

Lemma 4.1. *Let* L *be a set such that* $|L| = 6$. *There exists a tree* T *with* $L(T) = L$, *such that for every tree* T' *with* $L(T') = L$, $Q(T) \cap Q(T') \neq \emptyset$.

Proof. Let $L = \{a, b, c, d, e, f\}$. We claim that the tree T in Figure 6 works.

Assume, for the sake of contradiction, that there exists T' with $L(T') = L$ such that $Q(T) \cap Q(T') = \emptyset$. Then there exists a pair of leaves in T' that are siblings (i.e., the number of edges on the path between these two leaves is two). Without loss of generality, let a be one of these leaves and let x be its sibling. Note that $x \neq b$ since otherwise, every quartet of the form $ab \mid yz$, where $y, z \in \{c, d, e, f\}$, would be in $Q(T) \cap Q(T')$. Let x' be the sibling of x in T. Now choose leaves y and z from $L \setminus \{a, x, b, x'\}$. Clearly, since $|L| = 6$, this can always be done. Now the quartet $ax \mid yz \in Q(T) \cap Q(T')$, a contradiction. □

Now we can proceed to prove Theorem 4.2.

Proof of Theorem 4.2. Let $L' \subseteq L$ such that $|L'| = 6$. By Lemma 4.1, there exists a tree T' with $L(T') = L'$ such that $Q(L') \cap Q(g_i|L') \neq \emptyset$ for each $1 \leq i \leq k$. Now for each $1 \leq i \leq k$, choose $q_i \in Q(L') \cap Q(g_i|L')$ and set $X_i = L(q_i)$. Clearly, T' displays $(g_i|X_i)_{i=1}^k = (q_i)_{i=1}^k$, and so any tree T with $L(T) = L$ that displays T' completes the proof. □

The bound in Theorem 4.2 is tight. If the gene trees all have fewer than six leaves, then for certain input sequences, it is possible for no kernel to exist. This is because the existence of a tree similar to the one in Lemma 4.1 is not guaranteed if $|L| = 5$.

Lemma 4.2. *Let L be a set such that $|L| = 5$. Then for every tree T with $L(T) = L$, there exists a tree T' with $L(T') = L$ such that $Q(T) \cap Q(T') = \emptyset$.*

Proof. Let T be any tree with $L(T) = L$ and let $x \in L$ have no sibling. Since $|L| = 5$, such an x can always be found. Let $q = T|L \setminus \{x\}$ and q' be a quartet on the same set of leaves as q (i.e., $L(q') = L(q) = L \setminus \{x\}$) but is topologically different from q. Let T' be the tree obtained by attaching x to the internal branch of q'. It can be very easily seen that $Q(T) \cap Q(T') = \emptyset$. □

Due to Lemma 4.2, if the input sequence $(g_i)_{i=1}^k$ contains, say, all the gene trees on five leaves, then no kernel is possible. In other words, we have the following theorem.

Theorem 4.3. *Let X be a set with $|X| = 5$. If $\mathcal{G} = (g_i)_{i=1}^k$ is a sequence of gene trees containing every gene tree on X, then there does not exist a sequence $(X_i)_{i=1}^k$ with $|X_i| \geq 4$ such that there is a tree T_1 that displays $(g_i|X_i)_{i=1}^k$.*

Proof. For the sake of contradiction, assume there exists a sequence $(X_i)_{i=1}^k$ with $|X_i| \geq 4$ such that there is a tree T_1 which displays $(g_i|X_i)_{i=1}^k$. Let $T = T_1|X$. Using Lemma 4.2, we can now obtain T' such that $L(T') = X$ and $Q(T) \cap Q(T') = \emptyset$. Since $\mathcal{G}$ contains every gene tree on X, it also contains T'. And since T and T' have no quartets in common, neither do T_1 and T'. So, T_1 can not display T', a contradiction. □

It is worth noting that even though Theorem 4.2 is stated in a way that disallows missing data, a similar result also holds when there is missing data as long as there is a set of at least six leaves common to every input gene tree.

Corollary 4.1. *Let $\mathcal{G} = (g_i)_{i=1}^k$ be a sequence of (possibly incomplete) gene trees. If $|\bigcap_{i=1}^k L(g_i)| \geq 6$, then there exists $(X_i)_{i=1}^k$ with $|X_i| \geq 4$ such that there exists a T with $L(T) = \bigcup_{i=1}^k L(g_i)$ that displays $(g_i|X_i)_{i=1}^k$.*

Proof. We start with picking $L' \subseteq \bigcap_{i=1}^{k} L(g_i)$ such that $|L'| = 6$. The rest of the proof is almost identical to that of Theorem 4.2. □

4.2 Extending to Other Scoring Functions

So far, we have only been focusing on quartet scores. For example, we wanted our X_i's to be of size at least four because according to our definition, the quartet score of a tree with fewer than four leaves is undefined. However, our results are general and can easily be extended to other additive scoring functions too. Consider, for example, the extra-lineage score for which we have the following analogous propositions. The proofs of these are left to the reader (a proof sketch for Theorem 4.4 is provided). Note that, unlike the quartet score, computing the extra-lineage score requires the gene trees and species trees to be rooted.

Proposition 4.2. *Given a sequence $\mathcal{G} = (g_i)_{i=1}^{k}$ of rooted gene trees, any tree T in an extra-lineage peak terrace with respect to $\mathcal{G}$ must have an extra-lineage score of zero with respect to $\mathcal{G}$.*

Theorem 4.4. *Let L be a set such that $|L| \geq 6$ and let $\mathcal{G} = (g_i)_{i=1}^{k}$ be a sequence of rooted gene trees such that for all $1 \leq i \leq k$, $L(g_i) = L$. Then there exists $(X_i)_{i=1}^{k}$ with $|X_i| \geq 3$ for each i such that the extra-lineage peak terrace of $(g_i|X_i)_{i=1}^{k}$ is non-empty.*

Proof sketch for Theorem 4.4. The main idea is to find a *rooted* tree T such that for any rooted tree T' with $L(T) = L(T')$, there exists at least one rooted triplet that is common to both T and T'. It turns out that rooting the tree in Figure 6 on any branch results in a tree that works. Then we proceed similarly as the proof for Theorem 4.2.

5 Conclusions

The mathematical characterizations of species tree terraces and the new results on terraces in the gene tree-species tree context presented here are timely and important as

large genome-scale phylogenomic studies with substantial amounts of missing data are becoming increasingly common [37, 38]. The multiplicity of equally good species trees in terraced landscapes, stemming from certain patterns of missing data, poses various challenges as well as opens up several important research avenues. The ability to detect if a tree resides on a terrace, computing the size of the terrace, and enumerating the trees in a terrace can potentially improve the scalability and accuracy of summary methods. However, detecting and exploring species tree terraces, and leveraging that to improve the tree search strategies are challenging. Discovering various combinatorial properties of terraces and conditions for the presence of multiple equally good trees can contribute towards developing terrace-aware data structures and tree search algorithms.

We formally characterized species tree terraces and contrasted them with pseudo species tree terraces. We showed that, unlike pseudo terraces, species tree terraces depend only on the taxon coverage and are agnostic about the gene tree topology. However, considering different genes having different topologies is central to developing statistically consistent species tree estimation methods. In this study, we introduced a new type of terrace called peak terrace which requires one additional condition than terraces, making it sensitive to the distribution of gene tree topologies. Moreover, despite the differences in their mathematical definitions and conditions for existence, we showed that the set of species terraces is identical to the set of peak terraces. Therefore, we argued that understanding the structural properties of peak terraces suffices to understand species tree terraces in general. We proved various combinatorial properties of peak terraces and investigated patterns of missing data that lead to the existence of peak terraces. Although we explicitly considered the quartet score for investigating some of the properties of terraces, our results are general and apply to other additive scoring functions. Thus, we believe that this study will prompt more analytical and experimental studies to better understand the terraced landscapes of species trees and pioneer new terrace-aware methods, data structures, and optimization criteria for computing species trees from gene trees despite missing data and gene tree heterogeneity.

This study can be extended in several directions. Efficient terrace-aware data structures and algorithms for systematically navigating trees both inside a species tree terrace and its neighborhood would contribute to the improvement of the summary methods both in terms of accuracy and scalability. Therefore, investigating how to adapt the summary methods and algorithms to the existence of terraces is one of the most interesting research avenues. Developing efficient tools to identify species tree terraces and enumerating trees in them (similar to the existing tools [39] for counting trees in phylogenetic terraces) for different optimality criteria (e.g., quartet score, extra lineage score, etc.) is another important research direction. There are many questions of theoretical interest as well. For example, Section 4.1 considers the problem of finding the existence of a kernel given a gene tree sequence. A natural extension to this would be to find a kernel of the maximum size. Such a kernel would maximize the sum of the number of leaves remaining in the gene trees. One might also try to generalize Lemma 4.1, which asserts the existence of a binary tree T on six leaves whose quartets intersect every other binary tree on the same leaf set. One way to generalize this would be to consider trees larger than quartets. In other words, given k, let $n(k)$ be the smallest number for which there exists a binary tree T on $n(k)$ leaves such that for every binary tree T' with $L(T') = L(T)$, there is at least one k-leaf binary tree that is displayed by both T and T'. By Lemma 4.1, we have $n(4) = 6$. One might be interested in $n(k)$ for $k > 4$ and ask how it grows with k. Overall, future studies need to investigate further combinatorial properties of species tree terraces, the challenges they pose, and solutions to these problems.

References

[1] W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46:523–536, 1997.

[2] Sebastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, 2015.

[3] L S Kubatko and J H Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56:17, 2007.

[4] Scott V Edwards, Liang Liu, and Dennis K Pearl. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941, 2007.

[5] A D Leaché and B Rannala. The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, 60(2):126–137, 2011.

[6] Michael DeGiorgio and James H Degnan. Fast and consistent estimation of species trees using supermatrix rooted triples. *Molecular Biology and Evolution*, 27(3):552–569, 2009.

[7] Md Shamsuzzoha Bayzid and Tandy Warnow. Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18):2277–2284, 2013.

[8] J Heled and A J Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27:570–580, 2010.

[9] E Mossel and S Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):166–171, 2011.

[10] L S Kubatko, B C Carstens, and L L Knowles. Stem: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25:971–973, 2009.

[11] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

[12] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10:302, 2010.

[13] Liang Liu and Lili Yu. Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5):661–667, 2011.

[14] B Larget, S K Kotha, C N Dewey, and C Ané. BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 2010.

[15] David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah A Rosenberg, and Arindam RoyChoudhury. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932, 2012.

[16] Julia Chifman and Laura Kubatko. Quartet from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, 2014.

[17] Mazharul Islam, Kowshika Sarker, Trisha Das, Rezwana Reaz, and Md Shamsuzzoha Bayzid. Stelar: A statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. *BMC Genomics*, 21(1):1–13, 2020.

[18] Mahim Mahbub, Zahin Wahab, Rezwana Reaz, M Saifur Rahman, and Md Shamsuzzoha Bayzid. wQFM: highly accurate genome-scale species tree estimation from weighted quartets. *Bioinformatics*, 37(21):3734–3743, 2021.

[19] Rezwana Reaz, Md Shamsuzzoha Bayzid, and M Sohel Rahman. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS ONE*, 9(8):e104008, 2014.

[20] Y. Yu, T. Warnow, and L. Nakhleh. Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11):1543–1559, 2011.

[21] Michael J Sanderson, Michelle M McMahon, and Mike Steel. Terraces in phylogenetic tree space. *Science*, 333(6041):448–450, 2011.

[22] Michael J Sanderson, Michelle M McMahon, Alexandros Stamatakis, Derrick J Zwickl, and Mike Steel. Impacts of terraces on phylogenetic inference. *Systematic biology*, 64(5):709–726, 2015.

[23] Olga Chernomor, Arndt Von Haeseler, and Bui Quang Minh. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, 65(6):997–1008, 2016.

[24] Katherine St. John. The shape of phylogenetic treespace. *Systematic Biology*, 66(1):e83–e94, 2017.

[25] Barbara H Dobrin, Derrick J Zwickl, and Michael J Sanderson. The prevalence of terraced treescapes in analyses of phylogenetic data sets. *BMC Evolutionary Biology*, 18(1):46, 2018.

[26] Ishrat Tanzila Farah, Muktadirul Islam, Kazi Tasnim Zinat, Atif Hasan Rahman, and Shamsuzzoha Bayzid. Species tree estimation from gene trees by minimizing deep coalescence and maximizing quartet consistency: a comparative study and the presence of pseudo species tree terraces. *Systematic Biology*, 70(6):1213–1231, 2021.

[27] Alexandros Stamatakis and Michael Ott. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):3977–3984, 2008.

[28] Alexandros Stamatakis and Nikolaos Alachiotis. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics*, 26(12):i132–i139, 2010.

[29] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.

[30] Michael J Sanderson, Michelle M McMahon, and Mike Steel. Terraces in gene tree reconciliation-based species tree inference. *bioRxiv*, 2020.

[31] L. Zhang. From gene trees to species trees II: Species tree inference by minimizing deep coalescence events. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(9):1685–1691, 2011.

[32] M. S. Bayzid, S. Mirarab, and T. Warnow. Inferring optimal species trees under gene duplication and loss. In *Proc. of Pacific Symposium on Biocomputing (PSB)*, volume 18, pages 250–261, 2013.

[33] M. S. Bayzid and T. Warnow. Gene tree parsimony for incomplete gene trees: addressing true biological loss. *Algorithms for Molecular Biology*, 13:1, 2018.

[34] C. V. Than, D. Ruths, and L. Nakhleh. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9:322, 2008.

[35] Yun Yu, Tandy Warnow, and Luay Nakhleh. Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11):1543–1559, 2011.

[36] Alfred V. Aho, Yehoshua Sagiv, Thomas G. Szymanski, and Jeffrey D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, 10(3):405–421, 1981.

[37] Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho, Brant C Faircloth, Benoit Nabholz, Jason T Howard, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.

[38] Norman J Wickett, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, Michael S Barker, J Gordon Burleigh, Matthew A Gitzendanner, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868, 2014.

[39] Rudolf Biczok, Peter Bozsoky, Peter Eisenmann, Johannes Ernst, Tobias Ribizel, Fedor Scholz, Axel Trefzer, Florian Weber, Michael Hamann, and Alexandros Stamatakis. Two c++ libraries for counting trees on a phylogenetic terrace. *Bioinformatics*, 34(19):3399–3401, 2018.