Gene Tree Parsimony for Incomplete Gene Trees*

Md. Shamsuzzoha Bayzid[1] and Tandy Warnow[2]

1 Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
shams_bayzid@cse.buet.ac.bd

2 Department of Computer Science, University of Illinois Urbana-Champaign, Illinois, IL, USA
warnow@illinois.edu

Abstract

Species tree estimation from gene trees can be complicated by gene duplication and loss, and "gene tree parsimony" (GTP) is one approach for estimating species trees from multiple gene trees. In its standard formulation, the objective is to find a species tree that minimizes the total number of gene duplications and losses with respect to the input set of gene trees. Although much is known about GTP, little is known about how to treat inputs containing some *incomplete gene trees* (i.e., gene trees lacking one or more of the species). We present new theory for GTP considering whether the incompleteness is due to gene birth and death (i.e., true biological loss) or taxon sampling, and present dynamic programming algorithms that can be used for an exact but exponential time solution for small numbers of taxa, or as a heuristic for larger numbers of taxa. We also prove that the "standard" calculations for duplications and losses exactly solve GTP when incompleteness results from taxon sampling, although they can be incorrect when incompleteness results from true biological loss. The software for the DP algorithm is freely available as open source code at https://github.com/shamsbayzid/DynaDup.

1998 ACM Subject Classification G.2.1 Combinatorics, G.2.2 Graph Theory, J.3 Life and Medical Sciences

Keywords and phrases Gene duplication and loss, gene tree parsimony, deep coalescence

Digital Object Identifier 10.4230/LIPIcs.WABI.2017.2

1 Introduction

The estimation of species trees is often performed by estimating multiple sequence alignments for some collection of genes, concatenating these alignments into one super-matrix, and then estimating a tree (often using maximum likelihood or a Bayesian technique) on the resultant super-matrix. However, this approach cannot be used when the species' genomes contain multiple copies of some gene, which can result from gene duplication. Since gene duplication and loss is a common phenomenon, the estimation of species trees requires a different type of approach in this case.

Gene Tree Parsimony (GTP) is an optimization problem for estimating species trees from a set of gene trees (estimated from individual gene sequence alignments). In its most typical formulations, only gene duplication and loss are considered, so that GTP depends upon two parameters: c_d (the cost for a duplication) and c_l (the cost for a loss). The two most popular versions of GTP are MGD (minimize gene duplication), for which $c_d = 1$ and $c_l = 0$, and

* This work was supported by US NSF (1062335) to TW and Fulbright Fellowship to SB. The authors thank Siavash Mirarab for his helpful suggestions and help in implementing the software.

MGDL (minimize gene duplication and loss), for which $c_d = c_l = 1$. The version of GTP that seeks the tree minimizing the total number of losses has also been studied; for this, $c_d = 0$ and $c_l = 1$. These variants of GTP are NP-hard optimization problems [14], but software such as DupTree [24] and iGTP [4] for GTP are in wide use.

Basic to all these problems is the ability to compute the number of duplications and losses implied by a species tree and gene tree. This problem is called the "reconciliation problem", surveyed in [7], and intensively studied in the literature (see, for example, [9, 12, 18, 16, 19, 17, 20, 27, 13, 14, 10, 28]). The mathematical formulation of the reconciliation problem was derived for the case where the gene tree and the species tree have the same set of taxa, and then extended to be able to be used on *incomplete* gene trees, i.e., trees that can miss some taxa.

Incomplete gene trees are quite common, and can arise for two different reasons: (1) *taxon sampling*: the gene may be available in the species' genome, but was not included for some reason in the dataset for that gene, or (2) *gene birth/death*: as a result of gene birth and death (true biological gene loss), the species does not have the gene in its genome.

Given a gene tree gt and a species tree ST, two formulations for the number of losses have been defined. The most commonly used one computes the number of losses by first computing the "homeomorphic subtree" $ST(gt)$ of ST induced by gt, and then computing the number of losses required to reconcile gt with $ST(gt)$ (see, for example, [12, 14, 28]). Although this second formulation is in wide use (and is the basis of both iGTP [4] and Duptree [24], two popular methods for "solving" GTP), we will show that this can be incorrect when incompleteness is due to true biological loss. We refer to this formulation as the "standard" approach because of this widespread use in both software and the theoretical literature on GTP. The other, described in [5, 23], correctly computes the number of losses when incompleteness is a result of true gene loss, as we will prove.

This paper addresses the GTP problem for the case where some of the input gene trees may be incomplete due to either sampling or true biological loss. The main results are as follows:

- We formalize the duploss reconciliation problem when gene trees are incomplete due to taxon sampling as the "optimal completion of a gene tree" (Section 2.2), and we prove (Theorem 1) that the standard calculation correctly computes losses for this case.
- We show by example that the standard calculation for losses in GTP can be incorrect when incompleteness is due to true biological loss (Section 2.3).
- We show how to compute the number of losses implied by a gene tree and species tree, when incompleteness is due to true biological loss (Section 3).
- We formulate variants of the GTP problem (when gene tree incompleteness is due to true biological loss) as minimum weight maximum clique problems (see Theorem 10 for one duploss variant), and show how to solve these problems efficiently using dynamic programming (Section 4). We show that these optimal cliques can be found in polynomial time in the number of vertices of the graph, because of the special structure of the graphs. We also show that a constrained version of these problems, where the subtree-bipartitions of the species tree are drawn from the subtree-bipartitions of the input gene trees, can be solved in time that is polynomial in the number of gene trees and taxa.

2 Basics

2.1 Notation and terminology

Throughout this paper we will assume that gene trees and species trees are rooted binary trees, with leaves drawn from the set $\mathcal{X}$ of n taxa, and we allow the gene trees to have multiple copies of the taxa, and even to miss some taxa. We let gt denote a gene tree and ST denote a species tree. We let $L(t)$ denote the set of taxa at the leaves of the tree t, and require that $L(gt) \subseteq L(ST)$. If $L(gt) = L(ST)$ we say that gt is complete, and otherwise we say that gt is incomplete.

We now define some general terminology we will use throughout this paper; other terminology will be introduced as needed. Let T be a rooted binary tree. We denote the set of vertices of T by $V(T)$, the set of edges of T by $E(T)$, the root by $r(T)$, the internal nodes by $V_{int}(T)$, and the set of taxa that appear at the leaves by $L(T)$. Note that if T is a gene tree, it can be incomplete, and so it is possible for $|L(T)|$ to be smaller than the number of leaves in T. A *clade* in T is a subtree of T rooted at a node in T, and the set of leaves of the clade is called a *cluster*. Given a node v in T, the cluster of leaves below v is denoted by $c_T(v)$, and the subtree of T rooted at v is denoted by T_v. The most recent common ancestor (MRCA) of a set A of leaves in T is denoted by $MRCA_T(A)$. Given a gene tree gt and a species tree ST, we define $\mathcal{M} : V(gt) \to V(ST)$ by $\mathcal{M}(v) = MRCA_{ST}(c_{gt}(v))$. Finally, given a node u in a rooted binary tree, we let r denote the right child of u and l denote the left child of u.

For a rooted gene tree gt and a rooted species tree ST, where $L(gt) \subseteq L(ST)$, an internal node v in gt is called a duplication node if $\mathcal{M}(v) = \mathcal{M}(v')$ for some child v' of v, and otherwise v is a speciation node [28, 12, 14, 1].

$ST(gt)$ is the homeomorphic subtree of ST induced by the taxon set of gt, and is produced as follows: ST is restricted to the taxon set of gt, and then nodes with in-degree and out-degree 1 are suppressed. $ST^*(gt)$ is the tree obtained by restricting ST to the taxon set of gt, but not suppressing nodes of in-degree and out-degree 1.

We say that clade cl in ST is a *missing clade* with respect to gt if $L(gt) \cap L(cl) = \emptyset$, and a *maximal missing clade* if it is not contained in any other missing clade. Maximal missing clades that are descendants of $\mathcal{M}(r(gt))$ are called the “lower” maximal missing clades, and those that are not descendants of $\mathcal{M}((r(gt))$ are called the “upper” maximal missing clades. We denote by $LMMC(gt, ST)$ (or $LMMC$), the set of lower maximal missing clades, and $UMMC(gt, ST)$ (or $UMMC$), the set of upper maximal missing clades. Note $UMMC(gt, ST) = \emptyset$ iff $\mathcal{M}((r(gt)) = r(ST)$.

2.2 The standard formula for computing losses

The *standard* formula (see, for example, [12, 14, 10, 28, 1]) for computing the minimum number of losses of a (potentially incomplete) gene tree gt with respect to a species tree ST is denoted $L_{std}(gt, ST)$, and is defined to be $L_{std}(gt, ST) = \sum_{u \in V_{int}(gt)} F(u, ST(gt))$, where $F(u, T)$ is defined for internal nodes u with children l and r (which can be interchanged in the formula below) by:

$$F(u,T) = \begin{cases} d(\mathcal{M}(r),\mathcal{M}(u)) + 1 & \text{if } \mathcal{M}(r) \neq \mathcal{M}(u) \ \& \ \ \mathcal{M}(l) = \mathcal{M}(u), \\ d(\mathcal{M}(l),\mathcal{M}(u)) + 1 & \text{if } \mathcal{M}(l) \neq \mathcal{M}(u) \ \& \ \ \mathcal{M}(r) = \mathcal{M}(u), \\ d(\mathcal{M}(r),\mathcal{M}(u)) & \\ +d(\mathcal{M}(l),\mathcal{M}(u)) & \text{if } \mathcal{M}(r) \neq \mathcal{M}(u) \ \& \ \ \mathcal{M}(l) \neq \mathcal{M}(u), \\ 0 & \text{if } \mathcal{M}(r) = \mathcal{M}(l) = \mathcal{M}(u). \end{cases} \tag{1}$$

where $d(s, s')$ is the number of internal nodes in T on the path from s to s'. When gt is complete, then $ST(gt) = ST$, and this formula follows from [5].

Optimal completion of a gene tree

- Input: rooted binary gene tree gt and rooted binary species tree with $L(gt) \subseteq L(ST)$.
- Output: complete gene tree $T_{samp}(gt, ST)$ that is an extension of gt such that $T_{samp}(gt, ST)$ implies a minimum number of losses with respect to ST.

In other words, we add all the missing taxa into gt (each taxon added at least once, but perhaps several times) so as to produce a complete binary gene tree that has a minimum number of losses with respect to ST. Let $L_{samp}(gt, ST) = L_{std}(T_{samp}(gt, ST), ST)$. Thus, $L_{samp}(gt, ST)$ denotes the total number of losses needed for an optimal completion of gt. Similarly, we can define $DL_{samp}(gt, ST)$ to be the total number of duplications and losses needed for a completion of gt that minimizes the duploss score.

▶ **Theorem 1.** *Given a binary rooted gene tree gt and a binary rooted species tree ST such that $L(gt) \subseteq L(ST)$, the MRCA mapping defines a reconciliation that minimizes the number of duplications, the number of losses, and hence also the total number of duplications and losses, where we treat losses as due to sampling. Furthermore, $L_{std}(gt, ST) = L_{samp}(gt, ST)$, which means the standard formula correctly computes the number of losses when we treat incompleteness as due to sampling.*

Proof omitted due to space constraints.

2.3 Incompleteness due to gene birth and death

As we will see, while the MRCA mapping is still an optimal reconciliation when gene trees are incomplete due to gene birth and death (implied from [5, 11]), the standard formula does not correctly compute the number of losses.

Consider the simple example $gt = ((a, b), c)$ and $ST = ((a, (b, d)), c)$. Under the standard formula, $L_{std}(gt, ST) = 0$, since $ST(gt) = gt$. Under the assumption that incompleteness is due to true biological loss, the genome for d does not have the gene. Because d is sister to b and all the other taxa have the gene, the gene must have been present in the parent of d, and lost on the branch leading to d. *Therefore, the standard formula for the number of losses can be incorrect when gene trees are incomplete due to gene birth and death (i.e., true biological loss).*

3 How to calculate losses

We now show how to calculate the number of losses for an incomplete gene tree gt and species tree ST, treating incomplete gene trees as due to gene birth and death. How this is defined will depend upon whether one assumes, *a priori*, that the gene is present in the genome of the common ancestor of the species in ST (i.e., at the root of ST). Thus, this section shows how to calculate the following values:

- $L^*_{bd}(gt, ST)$, the minimum number of losses, under the assumption the gene is present in the common ancestor of the species in ST ($DL^*_{bd}(gt, ST)$ is defined similarly for the total number of duplications and losses), and
- $L_{bd}(gt, ST)$ the minimum number of losses *without* assuming the gene is present in the common ancestor of the species in ST ($DL_{bd}(gt, ST)$ is defined similarly for duplications and losses).

We now show how to compute the *number* of losses (i.e., $L_{bd}(gt, ST)$ and $L^*_{bd}(gt, ST)$), using the fact that the MRCA mapping defines an optimal reconciliation.

▶ **Theorem 2.** *Let gt be a gene tree and ST a species tree such that $L(gt) \subseteq L(ST)$. Then, $L_{bd}(gt, ST) = \sum_{u \in V_{int}(gt)} F(u, ST)$, and $L^*_{bd}(gt, ST) = L_{bd}(gt, ST) + |UMMC(gt, ST)|$. Furthermore, these values can be calculated in $O(n + n')$ time, where ST has n leaves and gt has n' leaves.*

Proof. Note that we use a modification of the standard formula, $F(u, ST)$, so that we do not replace ST by $ST(gt)$ as was done in [5, 23]. The equality for L_{bd} is implied from [5] and we omit the proof concerning L_{bd} due to space constraints.

Derivation of $L^*_{bd}(gt, ST)$. By definition of $L^*_{bd}(gt, ST)$, the gene is assumed to be present at the root of the species tree ST. If $\mathcal{M}((r(gt)) = r(ST)$, then $UMMC(gt, ST) = \emptyset$, and the result follows. However, if $\mathcal{M}((r(gt)) \neq r(ST)$, the gene must be present on the path between $r(ST)$ and $\mathcal{M}((r(gt))$. Since the gene is not present in any leaf that is not below $\mathcal{M}((r(gt))$, to minimize losses, the gene must be lost on every edge off that path, since such edges lead to subtrees that do not have the gene present in any leaf. Note that if $\mathcal{M}((r(gt)) \neq r(ST)$, then the number of edges that lead off that path is $|UMMC(gt, ST)| = d(\mathcal{M}((r(gt)), r(ST)) + 1$. Since the gene must be lost on each of those edges, and the total number of losses is the sum of this value and the number of losses that occur within the subtree rooted at $\mathcal{M}((r(gt))$, it follows that $L^*_{bd}(gt, ST) = L_{bd}(gt, ST) + |UMMC(gt, ST)|$.

The running time follows easily from the fact that the MRCA mapping can be computed in linear time [8]. ◀

4 Algorithms to find species trees

Here we address the problem of finding a species tree that has a minimum total number of duplications and losses, treating incompleteness as due to true biological loss. Prior results on GTP include a branch-and-bound algorithm in [6], based on techniques from [5], a randomized hill climbing based heuristic presented in [24], a probabilistic and computationally expensive method for coestimating gene and species trees [2], and dynamic programming based solutions by Hallett and Lagergren [13], Bayzid *et al.* [1] and Chang *et al.* [3]. However, none of these works takes the reasons of incompleteness into account, and we have already shown in Sec 2.3 that the standard calculation for losses can be incorrect when incompleteness is due to true biological loss.

In this section, we derive a different approach for the GTP problems, treating incomplete gene trees as due to true biological loss (i.e., minimizing $L_{bd}(gt, ST)$ or $L^*_{bd}(gt, ST)$). The techniques we propose can be used to solve GTP exactly for small datasets, or approximately (though without any guaranteed error bounds) on larger datasets. The approach we take here is based on [1] (see also [21, 13, 26, 25], which use very similar techniques). Bayzid *et al.* [1] provided a graph-theoretic formulation for $MGDL_{std}$, whereby an optimal solution to $MGDL_{std}$ corresponded to finding a minimum weight maximum clique inside a graph called the “Compatibility Graph”. The nodes of the compatibility graph correspond to “subtree-bipartitions”, a concept Bayzid *et al.* [1] introduced and we will also use. [1] showed how to find a minimum weight max clique using a dynamic programming approach. We will use the same graph-theoretic formulation as in [1], but modify the weights appropriately, to show that the optimal solution to $MGDL^*_{bd}$ still corresponds to a minimum weight max clique. The DP algorithm in [1] can then be used directly to find the optimal solution to $MGDL^*_{bd}$. To achieve this, we first derive an efficient formula for $L_{bd}(gt, ST)$ (and $L^*_{bd}(gt, ST)$, similar to the one derived in [28] for $L_{std}(gt, ST)$, but somewhat more involved.

We will let $\mathcal{D}_{gt,ST}$ denote the set of duplication nodes in gt with respect to ST and $\mathcal{S}_{gt,ST}$ denote the set of speciation nodes in gt with respect to ST. When gt and ST are known, we may write these as $\mathcal{D}$ and $\mathcal{S}$. The calculation for the number of losses depends on how we interpret incompleteness in gene trees. Therefore, rather than having a single optimization problem like $MGDL$, we have variants of this problem depending on how we treat incompleteness. As shown in Theorem 1, the term $MGDL$ in the literature refers to $MGDL_{std}$, which (by Theorem 1) is identical to $MGDL_{samp}$. Here, we consider the optimization problems $MGDL^*_{bd}$, where we treat incompleteness as due to gene birth and death. And therefore, we also consider $MGDL_{bd}$, MGL^*_{bd}, and MGL_{bd}.

4.1 Basic material

4.1.1 Subtree-bipartitions

Let T be a rooted binary tree and u an internal node in T. The *subtree-bipartition* of u, denoted by $\mathcal{SBP}_T(u)$, is the unordered pair $(c_T(l)|c_T(r))$, where l and r are the two children of u. Note that subtree-bipartitions are not defined for leaf nodes. The set of subtree-bipartitions of a tree T is denoted by $\mathcal{SBP}_T = \{\mathcal{SBP}_T(u) : u \in V_{int}(T)\}$. Furthermore, any pair A and B of disjoint subsets of $\mathcal{X}$ also define a subtree-bipartition (though we may refer to these as *candidate* subtree-bipartitions to emphasize this).

Subtree-bipartition domination: Let $BP_i = (P_{i_1}|P_{i_2})$ and $BP_j = (P_{j_1}|P_{j_2})$ be two subtree-bipartitions. We say that BP_i is *dominated by* BP_j (and conversely that BP_j *dominates* BP_i) if either of the following two conditions holds: (1) $P_{i_1} \subseteq P_{j_1}$ and $P_{i_2} \subseteq P_{j_2}$, or (2) $P_{i_1} \subseteq P_{j_2}$ and $P_{i_2} \subseteq P_{j_1}$. We say that subtree-bipartition $(A|B)$ is dominated by a species tree T if one of T's subtree-bipartitions dominates $(A|B)$. Bayzid *et al.* showed that an internal node u in a gene tree gt is a duplication node with respect to a species tree ST if $\mathcal{SBP}_{gt}(u)$ is dominated by ST [1]. Finally, for a set $\mathcal{G}$ of gene trees on taxon set $\mathcal{X}$ and for any candidate subtree-bipartition $(A|B)$, we let $W_{dom}(A|B)$ be the total number of subtree-bipartitions in $\mathcal{G}$ that are dominated by $(A|B)$.

Due to space constraints, we refer to Bayzid *et al.* [1] for discussions on subtree-bipartition "domination", "containment" and "compatibility", and the compatibility graph.

4.1.2 Deep coalescence and the MDC problem

Deep coalescence (also called *incomplete lineage sorting*, or ILS) refers to the failure of alleles to coalesce (looking backwards in time) into a common ancestral allele until deeper than the most recent speciation events [15]. One of the measures for incongruence between a gene tree and a species tree under ILS is $XL(gt, ST)$, the number of extra lineages defined for the pair ST and gt [15]. For a gene tree gt and a species tree ST such that $L(gt) \subseteq L(ST)$, the number of extra lineages (summing over all edges) is defined to be $XL(gt, ST) = \sum_{e' \in E(ST^*(gt))} XL(gt, e')$, where $XL(gt, e')$ is the number of extra lineages on e'.

MDC ("minimize deep coalescence") is an optimization problem for estimating species trees in the presence of ILS. The input to MDC is a set $\mathcal{G}$ of gene trees and the output is a species tree ST such that $\sum_{gt \in \mathcal{G}} XL(gt, ST)$ is minimized. This problem is also NP-hard [28], and software for the problem exists in Phylonet [22] and iGTP [4], among others. We now describe theoretical material leading to the algorithmic approach in Phylonet [26].

▶ **Definition 3** (From [26]). For $B \subseteq \mathcal{X}$ and gene tree gt, we set $k_B(gt)$ to be the number of B-maximal clusters in gt, where a **B-maximal cluster** is a cluster $Y \subseteq L(gt)$ such that $Y \subseteq B$ but no other cluster of gt containing Y is a subset of B.

▶ **Definition 4.** We define $W_{xl}(x, gt)$ for x either a subtree-bipartition or a subset of $\mathcal{X}$, as follows. If $x \subseteq \mathcal{X}$, then we set $W_{xl}(x, gt) = 0$ if $x \cap L(gt) = \emptyset$ and otherwise

$W_{xl}(x, gt) = k_x(gt) - 1$. If x is a subtree-bipartition, then we let $B = p \cup q$ for $x = (p|q)$, and we set $W_{xl}(x, gt) = 0$ if $B \cap L(gt) = \emptyset$, and otherwise $W_{xl}(x, gt) = k_B(gt) - 1$. For a set $\mathcal{G}$ of gene trees and ST a species tree, we set $W_0 = \sum_{gt\in\mathcal{G}} \sum_{x\in\mathcal{X}} W_{xl}(\{x\}, gt)$.

Yu *et al.* [26] showed that for any edge e in ST, where B is the cluster below e, then $k_B(gt)$ is the number of lineages going through edge e, and so $k_B(gt) - 1$ is the number of extra lineages going through e. They defined weights on potential species tree clusters B by $W_{mdc}(B, gt) = 0$ if $B \cap L(gt) = \emptyset$ and otherwise $W_{mdc}(B, gt) = k_B(gt) - 1$ (i.e., W_{mdc} is defined for clusters while W_{xl} is defined for subtree-bipartitions), and extended this to a set $\mathcal{G}$ of gene trees by $W'_{mdc}(B) = \sum_{gt\in\mathcal{G}} W_{mdc}(B, gt)$, and then to a set C of clusters by $W''_{mdc}(C) = \sum_{B\in C} W'_{mdc}(B)$. From this, it follows easily that a set C of $n-1$ compatible clusters minimizing $W''_{mdc}(C)$ defines a rooted binary species tree with a minimum MDC score.

4.2 Deriving $L_{bd}(gt, ST)$ and $L^*_{bd}(gt, ST)$

We begin with the following theorem:

▶ **Theorem 5** (From [28]). *Let gt be a rooted binary gene tree, ST a rooted binary species tree and $\mathcal{D}$ the set of duplication nodes in gt with respect to ST. Then*

$$L_{std}(gt, ST) = XL(gt, ST(gt)) + 2|\mathcal{D}| + |V(gt)| - |V(ST(gt))|.$$

We now derive formulas for $L_{bd}(gt, ST)$ and $L^*_{bd}(gt, ST)$; to obtain formulas for $DL_{bd}(gt, ST)$ and $DL^*_{bd}(gt, ST)$, simply add $|\mathcal{D}_{gt,ST})|$.

Recall that in the definition of $F(u, T)$ given in Eqn. 1, losses are associated with internal nodes, and the total number of losses is defined as the sum of losses associated to each internal node. However, the definition of the number of losses corresponding to a node can be rewritten in terms of edges, as we now show. Let $D(s, s')$ be the number of edges in the path in ST between s and s'. Therefore, $D(s, s')$ can be defined as follows.

$$D(s, s') = \begin{cases} d(s, s') + 1 & \text{if } d(s, s') \geq 1, \\ d(s, s') & \text{if } d(s, s') = 0. \end{cases}$$

Then, for a vertex u in gt with children r and l, we can rewrite Eqn. 1 as follows:

$$F(u, ST) = \begin{cases} D(\mathcal{M}(r), \mathcal{M}(u)) + D(\mathcal{M}(l), \mathcal{M}(u)) & \text{if } \mathcal{M}(r) \neq \mathcal{M}(u) = \mathcal{M}(l), \\ (D(\mathcal{M}(r), \mathcal{M}(u)) - 1) + (D(\mathcal{M}(l), \mathcal{M}(u)) - 1) & \text{if } \mathcal{M}(u) \notin \{\mathcal{M}(l), \mathcal{M}(r)\}, \\ D(\mathcal{M}(r), \mathcal{M}(u)) + D(\mathcal{M}(l), \mathcal{M}(u)) & \text{if } \mathcal{M}(r) = \mathcal{M}(u) = \mathcal{M}(l). \end{cases}$$

It is easy to see that in all three branches of the equation above, the two terms of the sum correspond to the edges connecting u to its two children l and r. (The second term in the first branch and both terms in the third branch are 0, but we wrote them in terms of the function $D(.,.)$ for convenience.) Let $p(x)$ be the parent of x in a tree T. Therefore, we can associate gene losses to edges $e = (x, p(x))$ instead of nodes, as follows:
$\mathcal{MD}(e) = D(\mathcal{M}(x), \mathcal{M}(p(x)), and$

$$edgeloss_{ST}(e) = \begin{cases} \mathcal{MD}(e) & \text{if } p(x) \in \mathcal{D}_{gt,ST}, \\ \mathcal{MD}(e) - 1 & \text{otherwise.} \end{cases}$$

We use the subscript ST in $edgeloss_{ST}(e)$ to emphasize the fact that the distance is taken within the tree ST and not within $ST(gt)$. Note therefore $\sum_{u\in V_{int}(gt)} F(u, ST) = \sum_{e\in E(gt)} edgeloss_{ST}(e)$.

▶ **Lemma 6.** *For all gene trees gt and species trees ST with $L(gt) \subseteq L(ST)$,*

$$L_{bd}(gt, ST) = \sum_{e \in E(gt)} \mathcal{M}D(e) - |E(gt)| + 2|\mathcal{D}|, \tag{2}$$

and for a set $\mathcal{G}$ of gene trees,

$$\begin{aligned} L_{bd}(\mathcal{G}, ST) &= \sum_{gt \in \mathcal{G}} L_{bd}(gt, ST) \\ &= \sum_{gt \in \mathcal{G}} \sum_{e \in E(gt)} \mathcal{M}D(e) - \sum_{gt \in \mathcal{G}} |E(gt)| + 2 \sum_{gt \in \mathcal{G}} |\mathcal{D}_{gt,ST}|. \end{aligned} \tag{3}$$

Finally, equalities concerning $DL_{bd}(gt, ST)$ and $DL_{bd}(\mathcal{G}, ST)$ can be obtained from these equalities by adding $|\mathcal{D}_{gt,ST}|$ and $|\mathcal{D}_{\mathcal{G},ST}|$, where $|\mathcal{D}_{\mathcal{G},ST}| = \sum_{gt \in \mathcal{G}} |\mathcal{D}_{gt,ST}|$.

Proof. We partition all the non-root nodes in gt into two sets: CD (children of duplications), consisting of those nodes whose parents are duplication nodes, and CS (children of speciations), consisting of those nodes whose parents are speciation nodes. Note that every edge $(x, p(x)) \in E(gt)$ can be associated with the set containing x. Therefore,

$$\begin{aligned} L_{bd}(gt, ST) &= \sum_{e \in E(gt)} edgeloss_{ST}(e) \\ &= \sum_{x \in CD} \mathcal{M}D(x, p(x)) + \sum_{x \in CS} (\mathcal{M}D(x, p(x)) - 1) \\ &= \sum_{e \in E(gt)} \mathcal{M}D(e) - |CS|. \end{aligned} \tag{4}$$

Since each internal node has two children, clearly the number of vertices x for which $p(x)$ is a speciation node is twice the number $|\mathcal{S}|$ of speciation nodes; therefore $L_{bd}(gt, ST) = \sum_{e \in E(gt)} \mathcal{M}D(e) - 2|\mathcal{S}|$. Since each internal node is a speciation node or a duplication node, it follows that $2(|\mathcal{D}| + |\mathcal{S}|) = |E(gt)|$, and the result follows. ◀

Let $L(gt, e)$ be the number of lineages that go through edge $e \in E(ST)$; thus, $XL(gt, e) = L(gt, e) - 1$, and so

$$XL(gt, ST) = \sum_{e' \in E(ST^*(gt))} L(gt, e') - |E(ST^*(gt))|. \tag{5}$$

▶ **Lemma 7.** *For any gene tree gt and species tree ST,*
$\sum_{e \in E(gt)} \mathcal{M}D(e) = \sum_{e' \in E(ST^*(gt))} L(gt, e')$, and (by Equation 5)

$$XL(gt, ST) = \sum_{e \in E(gt)} \mathcal{M}D(e) - |E(ST^*(gt))|. \tag{6}$$

Thus, for a set $\mathcal{G}$ of gene trees and species tree ST,

$$XL(\mathcal{G}, ST) = \sum_{gt \in \mathcal{G}} XL(gt, ST) = \sum_{gt \in \mathcal{G}} \sum_{e \in E(gt)} \mathcal{M}D(e) - \sum_{gt \in \mathcal{G}} |E(ST^*(gt))|.$$

Proof. We establish the first equality, since the remaining ones follow directly from it. Consider the lists of edges in paths in ST from $\mathcal{M}(x)$ to $\mathcal{M}(p(x))$, as x ranges over the internal vertices in gt. It is easy to see that the number of occurrences of an edge $e' \in E(ST^*(gt))$ in these lists is $L(gt, e')$ (the number of lineages through e'). Also, the edges $e \in E(ST) - E(ST^*(gt))$ will not be present in these lists, since these are the edges incident on the missing clades in ST with respect to gt. Therefore, the sum of the lengths of these lists is equal to $\sum_{e \in E(gt)} \mathcal{M}D(e)$ and also equal to $\sum_{e \in ST^*(gt)} L(gt, e)$. ◀

▶ **Theorem 8.** *For all gene trees gt, sets $\mathcal{G}$ of gene trees, and species trees ST, $L_{bd}(gt, ST) = XL(gt, ST) + 2|\mathcal{D}| + |E(ST^*(gt))| - |E(gt)|$, and*

$$L_{bd}(\mathcal{G}, ST) = XL(\mathcal{G}, ST) + 2\sum_{gt\in\mathcal{G}} |\mathcal{D}_{gt,ST}| + \sum_{gt\in\mathcal{G}} (|E(ST^*(gt))| - |E(gt)|). \tag{7}$$

Proof. Follows from Lemma 6 and Lemma 7. ◀

▶ **Corollary 9.** *For all gene trees gt and species trees ST,*

$$\begin{aligned} L^*_{bd}(gt, ST) &= L_{bd}(gt, ST) + |UMMC(gt, ST)| \\ &= XL(gt, ST) + 2|\mathcal{D}_{gt,ST}| + |E(ST^*(gt))| - |E(gt)| + |UMMC(gt, ST)|. \end{aligned}$$

$$\begin{aligned} DL^*_{bd}(gt, ST) &= L_{bd}(gt, ST) + |UMMC(gt, ST)| + |\mathcal{D}_{gt,ST}| \\ &= XL(gt, ST) + 3|\mathcal{D}_{gt,ST}| + |E(ST^*(gt))| - |E(gt)| + |UMMC(gt, ST)| \end{aligned}$$

Proof. The equalities concerning L^*_{bd} follow from Thm. 2 and Thm. 8. The equalities concerning DL^*_{bd} follow by adding $|\mathcal{D}_{gt,ST}|$. ◀

4.3 Assigning weights to subtree-bipartitions

To use the graph-theoretic formulation of $MGDL^*_{bd}$, we have to assign weights to each node in the compatibility graph, $CG(\mathcal{G})$, where $\mathcal{G}$ is the input set of gene trees, so that a minimum weight clique of $n-1$ vertices defines an optimal solution to $MGDL^*_{bd}(\mathcal{G})$. We will define weights $W_{xl}(v), W_{dom}(v), W_{EC}(v)$, and $W_{MMC}(v)$ to each subtree-bipartition (i.e., node in the compatibility graph), and set

$$W_{MGDL^*_{bd}}(v) = W_{xl}(v) - 3W_{dom}(v) + W_{EC}(v) + W_{MMC}(v).$$

We then prove (see Theorem 10) that a set of $n-1$ compatible subtree-bipartitions that has minimum total weight defines a species tree that optimizes $MGDL^*_{bd}$. Note that weights $W_{xl}(v)$ and $W_{dom}(v)$ have already been defined (in Section 4.1.1 and Section 4.1.2, respectively). Hence, all that remains is to define $W_{EC}(v)$ and $W_{MMC}(v)$, and then to prove Theorem 10.

Calculating $W_{EC}(v)$ and $|E(ST^*(gt))|$

We now show how to define weight $W_{EC}(v, gt)$ for every vertex v in the compatibility graph $CG(\mathcal{G})$ so that for all species trees ST, $|E(ST^*(gt))|$ is the sum of the vertex weights for the $n-1$ clique $\mathcal{C}$ in $CG(\mathcal{G})$ corresponding to ST. To count the number of edges in $E(ST^*(gt))$, we need to exclude those edges from $E(ST)$ that are incident on a clade that is missing in gt. For a vertex v associated with the subtree-bipartition $(p|q)$, we define $W_{EC}(v, gt)$ as follows (swapping p and q as needed):

$$W_{EC}(v, gt) = \begin{cases} 0 & \text{if } p \cap L(gt) = \emptyset \text{ and } q \cap L(gt) \in \{L(gt), \emptyset\} \\ 1 & \text{if } p \cap L(gt) = \emptyset \text{ and } \emptyset \neq q \cap L(gt) \subsetneq L(gt) \\ 2 & \text{otherwise.} \end{cases} \tag{8}$$

Then, $|E(ST^*(gt))| = \sum_{u\in\mathcal{SBP}_{ST}} W_{EC}(u, gt)$. We set $W_{EC}(v) = \sum_{gt\in\mathcal{G}} W_{EC}(v, gt)$. Then, for any species tree ST and set $\mathcal{G}$ of gene trees,

$$\sum_{gt\in\mathcal{G}} |E(ST^*(gt))| = \sum_{v\in\mathcal{C}} W_{EC}(v), \tag{9}$$

where $\mathcal{C}$ is the clique in $CG(\mathcal{G})$ that corresponds to ST.

Calculating $W_{MMC}(v)$ and $|UMMC(gt, ST)|$

We now show how to assign the weight $W_{MMC}(v, gt)$ to each vertex v of the compatibility graph so that for all species trees ST, $|UMMC(gt, ST)|$ is the sum of the weights over all the vertices of the clique $\mathcal{C}$ in $CG(\mathcal{G})$ corresponding to ST. Recall that $UMMC(gt, ST)$ is the set of upper maximal missing clades in ST. For a vertex v associated with the subtree-bipartition $(p|q)$, we define $W_{MMC}(v, gt)$ as follows (swapping p and q as needed):

$$W_{MMC}(v, gt) = \begin{cases} 1 & \text{if } p \cap L(gt) = \emptyset \text{ and } q \cap L(gt) = L(gt) \text{ (or vice-versa)} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Then $|UMMC(gt, ST)| = \sum_{u \in \mathcal{SBP}_{ST}} W_{MMC}(u, gt)$. Finally, we set $W_{MMC}(v) = \sum_{gt \in \mathcal{G}} W_{MMC}(v, gt)$. Then, for any species tree ST and set $\mathcal{G}$ of gene trees,

$$\sum_{gt \in \mathcal{G}} |UMMC(gt, ST)| = \sum_{v \in \mathcal{C}} W_{MMC}(v), \tag{11}$$

where $\mathcal{C}$ is the clique in $CG(\mathcal{G})$ that corresponds to ST.

We can extend the $MGDL^*_{bd}$ techniques to allow for losses and duplications to have different costs, as follows. Let c_d be the cost of a duplication and assume the cost of a loss (c_l) is 1. (Note that, our techniques work for any arbitrary c_d and c_l.) Let $|\mathcal{D}_{\mathcal{G},ST}| = \sum_i^k |\mathcal{D}_{gt_i,ST}|$, and set $DL^*_{bd}(\mathcal{G}, ST, c_d) = c_d * |\mathcal{D}_{\mathcal{G},ST}| + L^*_{bd}(\mathcal{G}, ST)$. Let $MGDL^*_{bd}(\mathcal{G}, c_d)$ be the problem that takes a set $\mathcal{G}$ of gene trees and duplication cost c_d as input, and finds the species tree that minimizes the weighted duploss score $DL^*_{bd}(\mathcal{G}, ST, c_d)$. Let $W^{c_d}_{MGDL^*_{bd}}(v) = W_{xl}(v) - (c_d + 2)W_{dom}(v) + W_{EC}(v) + W_{MMC}(v)$. (If $c_d = 1$, we omit the superscript c_d and write $W_{MGDL^*_{bd}}(v)$.)

▶ **Theorem 10.** *Let $\mathcal{G} = \{gt_1, gt_2, \ldots, gt_k\}$ be a set of binary rooted gene trees on set $\mathcal{X}$ of n species, and set the weights on the vertices in the compatibility graph using $W^{c_d}_{MGDL^*_{bd}}(v)$.* (a) *A set of subtree-bipartitions in an $(n-1)$-clique of minimum weight in $CG(\mathcal{G})$ defines a binary species tree ST that minimizes $DL^*_{bd}(\mathcal{G}, ST, c_d)$. Furthermore, the weighted duploss score of ST is given by $W_0 + W^{c_d}_{MGDL^*_{bd}}(\mathcal{C}) + c_d(N - k)$, where $N = \sum_{i=1}^k n_i$.* (b) *If we reset the weights to be $W_{MGL^*_{bd}}(v) = W_{MGDL^*_{bd}}(v) + W_{dom}(v)$, then a set of subtree-bipartitions in an $(n-1)$-clique of minimum weight in $CG(\mathcal{G})$ defines a binary species tree ST that minimizes $L^*_{bd}(\mathcal{G}, ST)$.*

Proof. We prove (a), since (b) follows directly from (a). Let $\mathcal{C}$ be a clique of size $n-1$ in $CG(\mathcal{G})$ and ST the associated species tree. Let $\mathcal{SBP}_{dom}(gt, ST)$ be the set of subtree-bipartitions in gt that are dominated by a subtree-bipartition in ST. Note that $|\mathcal{SBP}_{dom}(gt, ST)|$ is the number of speciation nodes in gt with respect to ST [1]. Therefore, the total number of speciation nodes in $\mathcal{G}$ is $\sum_{i=1}^k |\mathcal{SBP}_{dom}(gt_i, ST)| = \sum_{v \in V_{int}(ST)} W_{dom}(v)$. Also, $\sum_{v \in \mathcal{C}} W_{xl}(v) = \sum_{i=1}^k XL(gt_i, ST)$, and $\sum_{i=1}^k |\mathcal{D}_{gt_i,ST}| = \sum_{i=1}^k (n_i - 1) - \sum_{v \in \mathcal{C}} W_{dom}(v)$, where n_i is the number of leaves in gt_i. Finally, since all gene trees are rooted binary trees, $|E(gt_i)| = 2n_i - 2$ and $|V_{int}(gt_i)| = n_i - 1$. Recall that W_0 is the number of extra lineages contributed by the leaf set of the species tree (Definition 4). Therefore,

$$\begin{aligned}
DL^*_{bd}(\mathcal{G}, ST, c_d) &= \sum_{i=1}^{k}(c_d * |\mathcal{D}_{gt_i,ST}| + L^*_{bd}(gt_i, ST)) \\
&= \sum_{i=1}^{k}[XL(gt_i, ST) + (c_d+2)|\mathcal{D}_{gt_i,ST}| + \ |UMMC(gt_i, ST)| \\
&\quad + |E(ST^*(gt_i))| - |E(gt_i)|] \text{ (by Cor. 9)} \\
&= W_0 + \sum_{v\in\mathcal{C}} W_{xl}(v) + \sum_{i=1}^{k}(c_d+2)(n_i-1) - (c_d+2)\sum_{v\in\mathcal{C}} W_{dom}(v) \\
&\quad + \sum_{v\in\mathcal{C}} W_{MMC}(v) \ + \sum_{v\in\mathcal{C}} W_{EC}(v) - \sum_{i=1}^{k}(2n_i-2) \text{ (by Eqns. 9 and 11.)} \\
&= W_0 + W^{c_d}_{MGDL^*_{bd}}(\mathcal{C}) + c_d(N-k).
\end{aligned}$$

Note that W_0 does not depend on the topology of the species tree. Hence, the $(n-1)$-clique $\mathcal{C}$ with minimum weight defines a tree ST that minimizes $DL^*_{bd}(\mathcal{G}, ST, c_d)$. The proof for (b) follows trivially. ◀

4.4 Dynamic programming algorithm

Let $\mathcal{SBP}$ be a set of subtree-bipartitions, with $\mathcal{SBP}$ equal to all possible subtree-bipartitions if an exact solution is desired, and otherwise a proper subset if a faster algorithm is desired or necessary. We present the DP algorithm for the $MGDL^*_{bd}(\mathcal{G}, c_d)$ problem. We compute $score(A)$ in order, from the smallest cluster to the largest cluster $\mathcal{X}$.

Algorithm $MGDL^*_{bd}(\mathcal{G}, c_d)$

```
if $|A| = 1$ then $score(A) = W_{XL}(A)$
else
$score(A) = max\{score(A_1) + score(A - A_1) + W^{c_d}_{MGDL^*_{bd}}(A_1|A - A_1) : (A_1|A - A_1) \in \mathcal{SBP}\}$
```

If there is no $(A_1|A - A_1) \in \mathcal{SBP}$, we set its $score(A)$ to $-\infty$, signifying that A cannot be further resolved. At the end of the algorithm, if $\mathcal{SBP}$ includes at least one clique of size $n-1$, we have computed $score(\mathcal{X})$ as well as sufficient information to construct the optimal set of compatible clusters and hence the optimal species tree (subject to the constraint that all the subtree bipartitions in the output tree are in $\mathcal{SBP}$). If subtree bipartitions in $\mathcal{SBP}$ are not sufficient for building a fully resolved tree on $\mathcal{X}$, then $score(\mathcal{X})$ will be $-\infty$, and our algorithm returns FAIL.

The running time is $O(n|SBP|^2)$. The optimal number of duplications and losses is given by $score(\mathcal{X}) + c_d(N-k)$, by Theorem 10. If $\mathcal{SBP}$ contains all possible subtree-bipartitions, we have an exact but exponential time algorithm. However, if $\mathcal{SBP}$ contains only those subtree-bipartitions from the input gene trees, then the algorithm finds the optimal constrained species tree in time that is polynomial in the number of gene trees and taxa.

4.5 Extensions

It is trivial to extend the theory for $MGDL^*_{bd}$ and MGL^*_{bd} to $MGDL_{bd}$ and MGL_{bd}, as we now show. Recall that $L_{bd}(gt, ST) = L^*_{bd}(gt, ST) - |UMMC(gt, ST)|$ and that

$DL_{bd}(gt, ST) = DL^*_{bd}(gt, ST) - |UMMC(gt, ST)|$. Therefore, to extend the algorithmic approach to solve MGL_{bd} and $MGDL_{bd}$, we define $W_{MGL_{bd}}(v, gt) = W_{MGL^*_{bd}}(v, gt) - W_{MMC}(v, gt)$ and $W_{MGDL_{bd}}(v, gt) = W_{MGDL^*_{bd}}(v, gt) - W_{MMC}(v, gt)$, and then seek a minimum weight maximum clique in the compatibility graph with these modified weights.

5 Conclusion

In this paper we investigated how different reasons for gene tree incompleteness affects the mathematical formulation of gene loss. We present the first mathematical formulation to model gene loss due to true biological loss, and distinguish this from incompleteness due to taxon sampling. We proposed exact and heuristic algorithms to infer species trees from a set of incomplete gene trees by minimizing gene duplications and losses when the incompleteness is due to true biological loss.

References

1 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.

2 B. Boussau, G. J. Szöllősi, L. Duret, M. Gouy, E. Tannier, and V. Daubin. Genome-scale coestimation of species and gene trees. *Genome research*, 23(2):323–330, 2013.

3 W. C. Chang, A. Wehe, P. Górecki, and O. Eulenstein. Exact solutions for classic gene tree parsimony problems. In *Proc. of the 5th Int. Conf. on Bioinformatics and Computational Biology*, pages 225–230, 2013.

4 R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernández-Baca, and O Eulenstein. iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinf.*, pages 574–574, 2010.

5 C. Chauve, J. P. Doyon, and N. El-Mabrouk. Gene family evolution by duplication, speciation, and loss. *J. Comp. Biol.*, 15(8):1043–1062, 2008.

6 J. P. Doyon and C. Chauve. Branch-and-bound approach for parsimonious inference of a species tree from a set of gene family trees. *Adv. Exp. Med. Biol.*, 696:287–295, 2011.

7 J. P. Doyon, V. Ranwez, V. Daubin, and V. Berry. Models, algorithms and programs for phylogeny reconciliation. *Brieif. Bioinf.*, 12(5):392–400, 2011.

8 H. N. Gabow and R. E. Tarjan. A linear-time algorithm for a special case of disjoint set union. In *Proc. 15th ACM Symp. Theory of Comp. (STOC)*, pages 246–251, 1983.

9 M. Goodman, J. Czelusniak, G. Moore, E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, 28:132–163, 1979.

10 P. Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. In *Proc. 8th Ann. Int. Conf. on Computational Molecular Biology*, pages 316 – 325, 2004.

11 P. Górecki and J. Tiuryn. DLS-trees: A model of evolutionary scenarios. *Theor. Comput. Sci.*, 359(8):378–399, 2006.

12 R. Guigo, I. Muchnik, and T. Smith. Reconstruction of ancient molecular phylogeny. *Mol. Phylog. and Evol.*, 6(2):189–213, 1996.

13 M. T. Hallett and J. Lagergren. New algorithms for the duplication-loss model. In *Proc RECOMB*, pages 138–146, 2000.

14 B. Ma, M. Li, and L. Zhang. From gene trees to species trees. *SIAM J. on Comput.*, 30(3):729–752, 2000.

15 W. P. Maddison. Gene trees in species trees. *Syst Biol*, 46:523–536, 1997.

16 B. Mirkin, I. Muchnik, and T. Smith. A biologically consistent model for comparing molecular phylogenies. *J. Comput. Biol.*, 2(4):493–507, 1995.

17 R. Page and M. Charleston. Reconciled trees and incongruent gene and species trees. In B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzehtsky, editors, *Mathematical hierarchies in biology*, volume 37. American Math. Soc., 1997.

18 R. D. M. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. *Systematic Biology*, 43(1):58–77, 1994.

19 R. D. M. Page. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14(9):819–820, 1998. doi:10.1093/bioinformatics/14.9.819.

20 U. Stege. Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable. In *Proc. of the 6th Int. Workshop on Algorithms and Data Structures (WADS'99)*, pages 166–173, 1999.

21 C. V. Than and L. Nakhleh. Species tree inference by minimizing deep coalescences. *PLoS Comp. Biol.*, 5(9), 2009.

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

23 B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *J. Comp. Biol.*, 15(8):981–1006, 2008.

24 A. Wehe, M. S. Bansal, J. G. Burleigh, and O. Eulenstein. Duptree: A program for large-scale phylogenetic analyses using gene tree parsimony. *Amer. Jour. Bot.*, 24(13):1540–1541, 2008.

25 Y. Yu, T. Warnow, and L. Nakhleh. Algorithms for MDC-based multi-locus phylogeny inference. In *Proc. RECOMB*, 2011.

26 Y. Yu, T. Warnow, and L. Nakhleh. Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *J. Comp. Biol.*, 18(11):1543–1559, 2011.

27 L. Zhang. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J. Comp. Biol.*, 4(2):177–188, 1997.

28 L. Zhang. From gene trees to species trees II: Species tree inference by minimizing deep coalescence events. *IEEE/ACM Trans. Comp. Biol. Bioinf.*, 8(9):1685–1691, 2011.