Reconciliation Feasibility of Non-Binary Gene Trees under a Duplication-Loss-Coalescence Model

Ricson Cheng[1], Matthew Dohlen[2,⋆], Chen Pekker[3,⋆], Gabriel Quiroz[3], Jincheng Wang[3], Ran Libeskind-Hadas[3], and Yi-Chieh Wu[3(✉)]

[1] Department of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA ricsonc@andrew.cmu.edu
[2] Department of Computer Science, California Polytechnic University, 3801 W Temple Ave, Pomona, CA 91768, USA mdohlen@cs.hmc.edu
[3] Department of Computer Science, Harvey Mudd College, 301 Platt Blvd, Claremont, CA 91711, USA
{gpekker,gquiroz,jiwang}@hmc.edu, {hadas,yjw}@cs.hmc.edu

Abstract. Phylogenetic tree reconciliation is a widely-used method to understand gene family evolution. For eukaryotes, the duplication-loss-coalescence (DLC) model seeks to explain incongruence between gene trees and species trees by postulating gene duplication, gene loss, and deep coalescence events. While efficient algorithms exist for inferring optimal DLC reconciliations, they assume that only one individual is sampled per species. In recent work, we demonstrated that with additional samples, there exist gene tree topologies that are impossible to reconcile with any species tree. However, our algorithm required the gene tree to be binary whereas, in practice, gene trees are often non-binary due to uncertainty in the reconstruction process. In this work, we consider for the first time reconciliation under the DLC model with non-binary gene trees. Specifically, we describe an efficient algorithm that takes as input an arbitrary gene tree with an arbitrary number of samples per species and either (1) determines that there is a valid reconcilable binary resolution of that tree and constructs one such resolution or (2) determines that there exists no valid reconcilable binary resolution of that tree. Our work makes it possible to systematically analyze non-binary gene trees and will help biologists identify incorrect gene tree topologies and thus avoid incorrect evolutionary inferences.

Keywords: phylogenetics, reconciliation, gene duplication and loss, coalescence, non-binary trees

1 Introduction

Phylogenetic tree reconciliation is a fundamental technique for understanding the evolutionary histories of genes found across a set of species. Given a gene tree,

⋆ These authors contributed equally to this work.

species tree, and the association between their leaves, a *reconciliation* postulates evolutionary events to explain the incongruence, or topological differences, between those trees. These events may include gene duplication [21], gene loss [2], horizontal gene transfer [20], and incomplete lineage sorting [11], among others. Accurate reconciliations can provide important insights into centrally important questions on gene evolution and the introduction of new gene functions [16, 30].

Reconciliations rely on an underlying evolutionary model. Some widely-used models include the *duplication-loss* (DL) model [14, 22, 7, 35, 15, 6, 1, 24], which allows for gene duplication and gene loss; the *duplication-transfer-loss* (DTL) model [9, 10, 13, 29, 3, 8], which considers horizontal gene transfers as well; and the *multispecies coalescent* (MSC) model [19, 23, 31, 33], which allows for incomplete lineage sorting through deep coalescence. However, the DL and DTL models cannot address population effects, and MSC models cannot address paralogous gene families. Thus, each model has limited accuracy and applicability.

Recently, a unified *duplication-loss-coalescence* (DLC) model was proposed that combines the DL and MSC models [25], thereby addressing the most common events in eukaryotic gene evolution. Given a single haploid sample per species, two algorithms exist for solving the DLC reconciliation problem: DLCoalRecon finds the reconciliation with highest posterior probability [25], and DLCpar finds a most parsimonious reconciliation (one that minimizes the total cost of the constituent events) [32]. More recently, we extended the DLC model to allow for multiple samples per species and demonstrated that these multiple samples impose additional constraints such that gene trees may have no feasible reconciliation. Such infeasible gene trees can occur, for example, due to noisy sequencing, reconstruction error, or violations of model assumptions. To address this problem, we presented a polynomial-time algorithm for determining reconciliation feasibility of gene trees under the DLC model [26].

A significant limitation of these formulations is that they require the gene and species trees to be binary. In practice, species trees for several clades are binary since their reconstruction can benefit from well-behaved gene families as well as multigene phylogeny construction methods [12, 4]. When a species tree is non-binary, the non-binary nodes, or *polytomies*, are often "hard" and represent the simultaneous speciation of a common ancestor into multiple species. In contrast, gene trees are often non-binary due to lack of phylogenetic signal [27]. Their *polytomies* are "soft" in the sense that better data would allow us to resolve such nodes to yield a binary gene tree. Note that the number of binary resolutions is exponential in the number of non-binary nodes and their maximum out-degree. When given a non-binary gene tree and a binary species tree, reconciliation algorithms under the simpler DL and DTL models often seek to find a binary resolution of the gene tree that minimizes the reconciliation cost [5, 18, 34, 17].

In this work, we consider the problem of binary resolution under the DLC model with multiple samples per species. We present an efficient new algorithm that finds a valid binary resolution when such a resolution exists. Note that a brute-force approach of enumerating each binary resolution and testing it for reconcilability would take exponential time and thus be impractical. Using our

algorithm, we also prove that there exist non-binary gene trees for which there is no valid binary resolution. This work generalizes existing results on reconciliation feasibility of binary gene trees and is thus an important step towards a full reconciliation algorithm for non-binary gene trees under the DLC model.

2 Background

2.1 Reconciliation Feasibility

We previously studied reconciliation feasibility under the DLC model [26] and review that work here.

We start with some basic tree and graph definitions. Let T be an unrooted, full, binary tree[1] with a set $V(T)$ of nodes (or vertices) and a set $E(T)$ of branches (or edges). Let $L(T) \subset V(T)$ denote the set of leaves, and for nodes u and v, let $path(u, v)$ denote the set of branches along the unique simple path from u to v in T. Similarly, let $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$ be an undirected graph with a set $V(\mathcal{G})$ of vertices and a set $E(\mathcal{G})$ of edges. Let $\mathcal{C}(\mathcal{G})$ denote the set of connected components of $\mathcal{G}$, where $C \in \mathcal{C}(\mathcal{G})$ is a subgraph of $\mathcal{G}$ denoting a single connected component.

A *species tree* S is a tree that depicts the evolutionary history of a set of species, and a *gene tree* G is a tree that depicts the evolutionary history of a set of genes sampled from these species. Gene trees may be either binary or non-binary while the species tree is always assumed to be binary. A *species leaf map* $Le : L(G) \to L(S)$ associates each leaf of G with the leaf of S in which that gene is found. Note that more than one gene may be sampled from the same species; these genes could correspond to either multiple loci or multiple haploid samples. A gene tree is associated with a finite *locus set* $\mathbb{L}$ of species-specific loci that have evolved within the gene family. A *locus leaf map* $Le^L : L(G) \to \mathbb{L}$ associates each leaf of G with the species-specific locus at which that gene is found. For example, two genes map to the same species-specific locus if they are mapped to the same location on a reference genome. Note that the relationship between loci in different species is assumed to be unknown. Furthermore, there may exist copy number variations resulting in different samples from the same species containing different loci.

The *labeled coalescent tree (LCT)* formalizes the notion of a reconciliation in the DLC model [32]. In brief, the LCT is an annotated gene tree that simultaneously describes the gene tree topology and its reconciliation to the species tree. As a full description of the LCT is not necessary to characterize the reconciliation *feasibility* problem, we present only the necessary concepts and terminology. First, duplications occur along branches in the LCT, denoting that the locus has changed at some point along the branch. Second, the LCT labels each node and branch with the locus in which the gene evolves; for branches with a duplication, one side of the branch (before the duplication) is labeled with the original locus and the other side (after the duplication) with the new locus.

[1] Branch lengths are not used in this work, so a tree always refers to a tree topology.

Multiple species-specific loci may be related through speciation events alone and thus correspond to the same evolutionary locus. This notion is formalized and used to define reconcilable gene trees as follows:

Definition 1 (Locus Class). *Let a collection* $\mathbb{LC} = \{C_i\}$ *of nonempty sets form a partition over* $\mathbb{L}$ *such that each locus* $l \in \mathbb{L}$ *belongs to a single* locus class $C_i \in \mathbb{LC}$.

Definition 2 (Reconcilable Gene Tree[2]). *Given gene tree* G*, species leaf map* Le*, and locus leaf map* Le^L*,* G *is said to be* reconcilable *if there exists some map* $\mathcal{L}\colon L(G) \cup E(G) \to \mathbb{LC}$ *of each leaf and edge of the gene tree to a single locus class, such that for each pair of genes* $g_1 \in L(G), g_2 \in L(G), g_1 \neq g_2$*,* $\mathcal{L}$ *is subject to the following constraints:*

1. *If* $Le^L(g_1) = Le^L(g_2)$*, then* $\mathcal{L}(g_1) = \mathcal{L}(g_2)$ *and for each* $e \in path(g_1, g_2)$*,* $\mathcal{L}(e) = \mathcal{L}(g_1)$*. (Allele Constraint)*
2. *If* $Le(g_1) = Le(g_2)$ *but* $Le^L(g_1) \neq Le^L(g_2)$*, then* $\mathcal{L}(g_1) \neq \mathcal{L}(g_2)$*. (Paralog Constraint)*

Constraint 1 ensures that genes from the same species-specific locus are assigned the same locus class and, because duplications create a unique new locus, that genes and edges assigned the same locus class form a subtree of the gene tree. Constraint 2 ensures that genes from paralagous loci are assigned different locus classes. Note that reconcilability of the gene tree depends on its topology and the mapping of its leaves to the leaves of the species tree and to species-specific loci, but reconcilability does not depend on the actual topology of the species tree.

Problem 3 (Reconciliation Feasibility). Given gene tree G, species leaf map Le, and locus leaf map Le^L, determine whether G is reconcilable.

The reconciliation feasibility problem can be solved using two structures, the Partially Labeled Coalescent Tree (PLCT, Figure 1B) and the Locus Equivalence Graph (LEG, Figure 1C), defined formally below:

Definition 4 (Partially Labeled Coalescent Tree). *Let* $\mathbb{P}(\mathbb{L})$ *denote the power set of* $\mathbb{L}$*. Given* G *and* Le^L*, the* partially labeled coalescent tree (PLCT) *is a map* $\mathcal{P} : E(G) \to \mathbb{P}(\mathbb{L})$ *constructed as follows: Consider each pair of genes* $g_1 \in L(G), g_2 \in L(G), g_1 \neq g_2$ *such that* $Le^L(g_1) = Le^L(g_2) = l$*. For each gene tree edge* $e \in path(g_1, g_2)$*, add* l *to* $\mathcal{P}(e)$*.*

Definition 5 (Locus Equivalence Graph). *Given a PLCT* $\mathcal{P}$ *for* G *and* Le^L*, the* locus equivalence graph (LEG) *is a graph* $\mathcal{G}$ *constructed as follows: Set* $V(\mathcal{G}) = \mathbb{L}$*. For each gene tree edge* $e \in E(G)$ *and each pair of loci* $l_1 \in \mathcal{P}(e), l_2 \in \mathcal{P}(e), l_1 \neq l_2$*, add* (l_1, l_2) *to* $E(\mathcal{G})$*.*

The PLCT captures the allele constraints for each species-specific locus by labeling edges of the gene tree with the species-specific locus or loci to which the edge must belong. If an edge is labeled with multiple loci, these multiple loci must

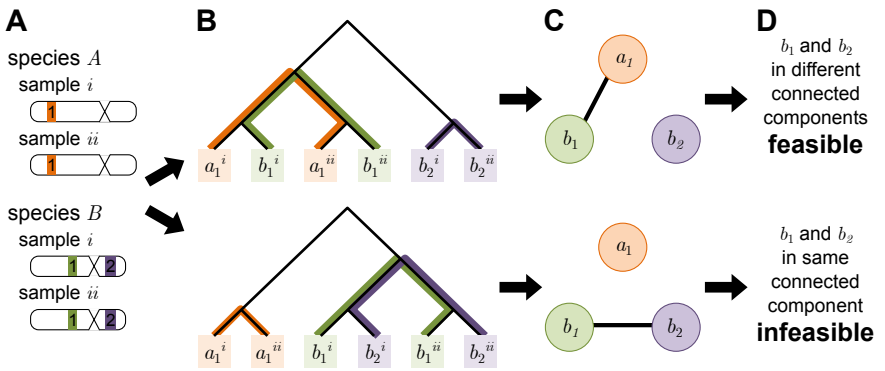

Fig. 1. Reconciliation feasibility for binary gene trees. (**A**) The sampled species (capital letters), loci (numbers), and haploid samples (roman numerals). We assume knowledge of the species-specific locus from which each gene is sampled. Within a species, genes at the same locus (across multiple samples must be alleles), and genes at different loci (regardless of sample) must be paralogs. (**B**) For a gene tree (black), the PLCT uses alleles to label branches along which no duplications are allowed (colored lines). (**C**) The LEG contains one node per species-specific locus and encodes overlapping labels in the PLCT as edges in the LEG. (**D**) A gene tree has a feasible reconciliation if and only if every connected component of the LEG contains no paralogs, that is, no more than one locus from each species. [Figure and caption adapted with permission from Rogers et al. [26].]

correspond to the same locus class. This equivalency constraint is captured as an edge between loci in the LEG. Rogers et al. [26] provide a formal description of the algorithm for constructing the PLCT and LEG, describe an optimization, and derive their time complexities of $O(nk)$ and $O(nk^2)$, respectively, where $n = |L(G)|$ and $k = |\mathbb{L}|$. Next, paralog constraints are used to define reconcilable LEGs:

Definition 6 (Reconcilable Locus Equivalence Graph). *For each $l \in \mathbb{L}$, let map $Le^S : \mathbb{L} \to L(S)$ associate each species-specific locus with the leaf of S in which the locus is found. That is, for each $g \in L(G)$, if $l = Le^L(g)$, then $Le^S(l) = Le(g)$. Given G, Le, and Le^L, a LEG $\mathcal{G}$ for G and Le^L is said to be* reconcilable *if for each $C \in \mathcal{C}(\mathcal{G})$ and for each $s \in L(S)$, there exists no more than one locus $l \in C$ such that $Le^S(l) = s$.*

The LEG enforces the paralog constraints for each species by requiring that each connected component contain no more than one locus from any species. LEG reconcilability can be determined in $O(k^3)$ time and related to gene tree reconcilability [26]:

Theorem 7. *A gene tree is reconcilable if and only if its locus equivalence graph is reconcilable.*

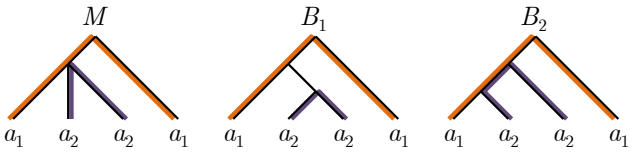

Fig. 2. Reconciliation feasibility for non-binary gene trees. A multifurcating gene tree M and two binarizations B_1 and B_2. Superscripts indicating haploid samples have been omitted. For B_1, a_1 and a_2 can be mapped to distinct locus classes, so the gene tree is reconcilable. For B_2, a_1 and a_2 must be mapped to the same locus class, but a_1 and a_2 are paralogs, so the gene tree is irreconcilable.

3 Reconciliation Feasibility for Non-Binary Gene Trees

In the previous section, all definitions and theorems applied only to *binary* gene trees. In this section, we consider the reconcilability of non-binary gene trees.

Let M be a non-binary, or *multifurcating*, gene tree. Each node with more than two children is called a *multifurcation*. Without loss of generality, and to simplify our discussion, we root M arbitrarily along any branch. A *binarization* $B(M)$ of M is a binary tree in which each multifurcation v with $k > 2$ children is replaced by a binary tree rooted at v with k leaves. These k leaves represent the k original children of v and thus may themselves be the roots of subtrees with their own descendants. The binary tree rooted at v is said to *resolve* the multifurcation, and we call that binary tree an *expansion tree* for v.

We now formalize the notion of reconcilable multifurcating gene trees:

Definition 8. *A multifurcating gene tree M is said to be* reconcilable *if there exists a binarization $B(M)$ of M that is reconcilable.*

Note that for a multifurcating gene tree, not all binarizations may be reconcilable. For example, two binarizations may induce different paths between two genes such that allele and paralog constraints are satisfiable in one binarization but not in another (Figure 2). Rather than enumerate all binarizations and evaluate each for reconcilability, we propose to evaluate the reconcilability of multifurcating gene trees directly.

We start by applying the definitions of the PLCT and LEG (Definitions 4 and 5) directly to multifurcating gene trees. However, Theorem 7, which relates reconcilability of gene trees to reconcilability of LEGs requires that the gene tree be binary.[3] Our goal is to extend Theorem 7 to multifurcating gene trees:

Theorem 9. *A multifurcating gene tree is reconcilable if and only if its locus equivalence graph is reconcilable.*

For a multifurcating gene tree M, let the associated LEG be $\mathcal{G}_M$. We then reformulate Theorem 9 as two separate theorems, one for each direction of the "if and only if" statement:

[3] The proof considers only the single unique path between two genes in a binary tree.

Theorem 9a. *If $\mathcal{G}_M$ is reconcilable, then there exists a binarization $B(M)$ of M that is reconcilable.*

Theorem 9b. *If $\mathcal{G}_M$ is irreconcilable, then there exists no binarization $B(M)$ of M that is reconcilable.*

3.1 Proof of Reconcilability

For each locus $l \in \mathbb{L}$, there exists a *locus tree*[4] which is the subtree of M whose leaves contain that locus. Let $r(M)$ denote the root of M, and for $u \in V(M), u \neq r(M)$, let the *parent edge* of u be the edge from u to its parent. Consider a node u and its parent edge e. If edge e is not used by any locus trees, then u is said to be *uncontained*. However, if one or more locus trees contain edge e, then, by definition, the set of those loci are in a single connected component C of $\mathcal{G}_M$, and we say that u is *contained* by C.

Given a non-binary tree M (Figure 3A), we want to efficiently determine whether or not there exists a binarization $B(M)$ of M that is reconcilable. We propose the following binarization algorithm:

1. For each multifurcation $v \in V(M)$, partition its children by the connected components in $\mathcal{G}_M$ that contain them, placing uncontained children arbitrarily (Figure 3B).
2. For each set in the partition, construct a *sub-expansion tree* by attaching all the children to the leaves of an arbitrary binary tree with the same number of leaves as children in the set (Figure 3C).
3. Join all sub-expansion trees together with another arbitrary binary tree of appropriate size, called the *connecting tree*, by attaching the roots of the sub-expansion trees to the leaves of the connecting tree. This results in our expansion tree for v (Figure 3C).

Constructing an expansion tree for each multifurcating node in M, in this way, results in our binarization $B(M)$. Note that since some aspects of the construction permit arbitrary decisions (e.g., placement of uncontained nodes, construction of the connecting tree), the resulting binarization is not unique.

We now relate $\mathcal{G}_M$ for M with $\mathcal{G}_{B(M)}$ for $B(M)$.

Lemma 10. *Let v be a multifurcating vertex in M, and let $B(M)$ be a binarization constructed by our algorithm. In $B(M)$, if a locus tree L for locus l contains an edge in the connecting tree of v, then it contains the parent edge of v.*

Proof. Suppose L contains an edge in the connecting tree of v but does not contain the parent edge of v. Then, by construction of $B(M)$, L has leaves g_1 and g_2 in two distinct sub-expansion trees of v. Therefore, in the original tree M, the path from g_1 to g_2 passes through v and thus passes through two children of v, denoted v_i and v_j. Since g_1 and g_2 are in distinct sub-expansion trees, it follows that v_i and v_j are each contained by a distinct connected component in

[4] Note that this locus tree is distinct from the locus tree of Rasmussen and Kellis [25].

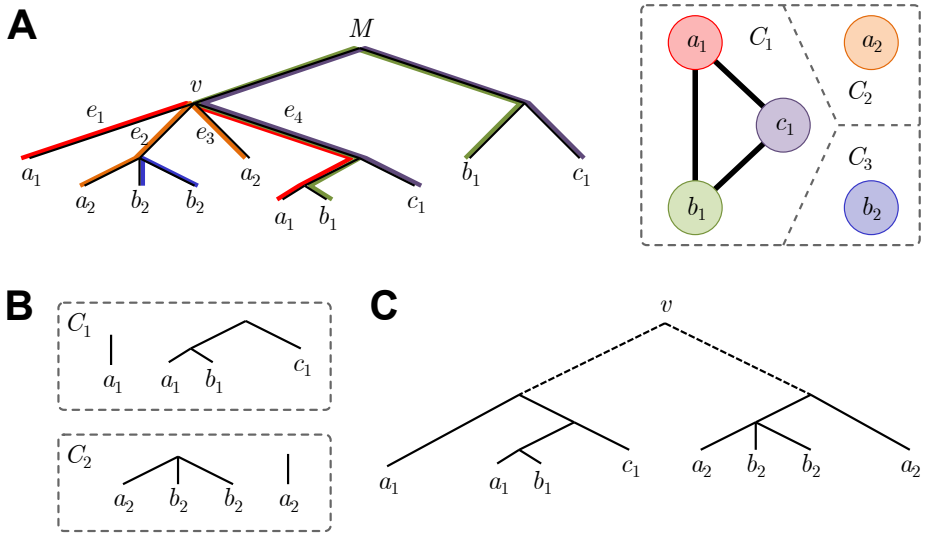

Fig. 3. Reconciliation feasibility for non-binary gene trees. (**A**) A multifurcating gene tree M and its locus equivalency graph $\mathcal{G}_M$. Superscripts indicating haploid samples have been omitted. Nodes with parent edges e_1 and e_4 are contained by connected component C_1, and nodes with parent edges e_2 and e_3 are contained by connected component C_2. (**B**) The partition of children from multifurcating node v. The first set includes the children contained by connected component C_1, and the second set includes the children contained by connected component C_2. (**C**) Sub-expansion trees (solid) joined through a connecting tree (dashed) to yield an expansion tree for v.

$\mathcal{G}_M$. But, by definition of $\mathcal{G}_M$, the path from g_1 to g_2 implies that v_i and v_j are covered by a single connected component in $\mathcal{G}_M$. □

Lemma 11. *Let T be any binarization of M. Let l be a locus and let L_M and L_T be the locus trees for l in M and T, respectively. The edge set of L_T is exactly the edge set of L_M, with the addition of a subset of edges from expansion trees.*

Proof. Note that L_M is the union of paths between all pairs of leaves with locus l in M, and similarly, L_T is the union of paths between all pairs of leaves with locus l in T. Every path in M corresponds to a unique path in T where all internal nodes are expanded into a path through the corresponding expansion tree. By the uniqueness of paths in trees, L_T is exactly L_M augmented with the edges traversed in the expansion trees. □

For graph $\mathcal{G}$ and two nodes $u, v \in V(\mathcal{G})$, we say that u and v are *connected* if they are in the same connected component and *disconnected* otherwise.

Lemma 12. *Let l and k be a pair of disconnected loci in $\mathcal{G}_M$. Then, there is no edge between l and k in $\mathcal{G}_{B(M)}$.*

Proof. Let $T = B(M)$. Consider any l and k in different connected components in $\mathcal{G}_M$ and any multifurcation v in M. Let L_M and K_M denote the locus trees for l and k, respectively, in M, and let L_T and K_T denote the corresponding locus trees in T.

Since l and k are in different connected components of $\mathcal{G}_M$, at least one of L_M or K_M does not use the parent edge of v. By Lemma 11, the edge set of L_M and K_M are subsets of L_T and K_T, respectively. Therefore, at least one of L_T or K_T does not use the parent edge of v. Therefore, by Lemma 10, at most one of L_T or K_T contains an edge in the connecting tree for v in T.

Next, we claim that if l and k are in different connected components C_l and C_k in $\mathcal{G}_M$, then L_T and K_T do not intersect in any sub-expansion tree in T. Suppose L_T and K_T intersect inside a sub-expansion tree of some vertex v. Since C_l and C_k are in different components in $\mathcal{G}_M$, this intersection must happen at an edge that was introduced when joining the children of v into sub-expansion trees; these edges correspond to edges from v to its children in M. Thus, in M, L_M and K_M must share an edge and are thus in the same component, which contradicts our assumption.

We have established that if l and k are in different connected components in $\mathcal{G}_M$, then they cannot share an edge in either a connection tree or a sub-expansion tree in T. Therefore, by Lemma 11, l and k cannot share any edge in T and thus there is no edge between them in $\mathcal{G}_T$. □

Finally, we prove Theorem 9a, which has been restated using $B(M)$ constructed by our algorithm.

Theorem 9a. *If $\mathcal{G}_M$ is reconcilable, then $\mathcal{G}_{B(M)}$ is reconcilable.*

Proof. Let $T = B(M)$. It suffices to show that if l and k are in different connected components in $\mathcal{G}_M$, then they are in different connected components in $\mathcal{G}_T$, implying that if $\mathcal{G}_M$ is reconcilable, then $\mathcal{G}_T$ is reconcilable.

Assume by way of contradiction that loci l and k are disconnected in M but connected in T. Then, there exists a path p from ℓ to k in $\mathcal{G}_T$. Let (u, v) be the first edge on p such that l and u are in the same connected component in $\mathcal{G}_M$ but u and v are in different connected components in $\mathcal{G}_M$. From Lemma 12, (u, v) cannot be an edge in $\mathcal{G}_T$, contradicting the assumption. □

Theorem 9a implies a polynomial-time algorithm for *both* determining if a non-binary gene tree is reconcilable *and*, if so, constructing one reconcilable binarization. Recall that, for $n = |L(G)|$ and $k = |\mathbb{L}|$, it takes $O(nk)+O(nk^2)+O(k^3)$ time to construct the PLCT and LEG and then test the LEG for reconcilability. For the binarization process, let c denote the maximum number of children over all multifurcations in the gene tree, and m denote the total number of multifurcations. The time required to build the expansion tree for each multifurcation is linear in c, and each multifurcation can be resolved independently. Thus, the total complexity of the binarization process is $O(cm)$.

For comparison, the number of distinct binary resolutions for multifurcation v with k_v children is $N_v = (2k_v - 3)!!$. A brute-force approach that enumerates each binarization and combines them would therefore result in $\prod_{v \in V(M): k_v > 2} N_v$ expansion trees, making it infeasible to enumerate and test each one for feasibility.

3.2 Proof of Irreconcilability

Theorem 9b. *If $\mathcal{G}_M$ is irreconcilable, then there exists no binarization $B(M)$ of M that is reconcilable.*

Proof. Let T be an arbitrary binarization of M. By Lemma 11, any pair of locus trees L_T and K_T in T contain all the edges of the corresponding locus tree, L_M and K_M, in M. Thus, any two loci that are connected by an edge in $\mathcal{G}_M$ must also have an edge in $\mathcal{G}_T$. □

It is not difficult to show that there exist non-binary gene trees that are not reconcilable.[5]

4 Discussion

We have presented an efficient algorithm that evaluates an arbitrary gene tree topology with an arbitrary number of samples per species under the DLC model and either (1) determines that there is a valid reconcilable binary resolution of that tree and constructs one such resolution or (2) determines that there exists no valid reconcilable binary resolution of that tree.

[5] For example, in Figure 3, swapping leaves labeled a_2 with leaves labeled c_1 would result in an irreconcilable LEG and thus a multifurcating gene tree for which there exists no reconcilable binarization.

In previous work [26], we reconstructed RAxML [28] gene trees, collapsed poorly-supported branches to yield non-binary gene trees, and analyzed the reconcilability of the associated LEG. Our work here allows us to directly relate LEG reconcilability to gene tree reconcilability. In particular, while gene tree reconcilability is affected by poorly-supported branches, even multifurcating gene trees with well-supported branches can be infeasible.

One limitation of our work is that given a non-binary gene tree, we are not guaranteed to construct an *optimal* binary resolution. That is, our binarization may not yield a gene tree with the lowest reconciliation cost under a parsimony framework. But our work suggests one possible approach. We propose to explore the space of reconcilable resolutions compared to the space of all resolutions. If, in-practice, most non-binary gene trees have a single or small number of reconcilable resolutions, it would imply that we could simply enumerate the resolutions, then apply existing reconciliation algorithms for binary trees.[6]

For irreconcilable gene trees, a possible research direction is to investigate error-correction algorithms. Such an algorithm could find the minimum number of topological rearrangements needed to yield a reconcilable gene tree. An alternative is to remove the minimum number of sampled individuals and explore possible patterns among the removed individuals. Such patterns could provide insight into whether certain populations are correlated with error and therefore more susceptible to problems elsewhere in a phylogenomic pipeline.

Acknowledgments

This work was supported by funds from the Department of Computer Science and the Dean of Faculty of Harvey Mudd College and by the U.S. National Science Foundation under grant IIS-1419739.

References

1. Åkerborg, Ö., Sennblad, B., Arvestad, L., Lagergren, J.: Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci USA **106**(14), 5714–5719 (2009)
2. Albalat, R., Cañestro, C.: Evolution by gene loss. Nat Rev Genet **17**, 379– (2016)
3. Bansal, M.S., Alm, E.J., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**(12), i283–i291 (2012)
4. Burleigh, J.G., Bansal, M.S., Eulenstein, O., Hartmann, S., Wehe, A., Vision, T.J.: Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees. Syst Biol **60**(2), 117–125 (2011)
5. Chang, W.C., Eulenstein, O.: Reconciling gene trees with apparent polytomies: Computing and combinatorics. In: Chen, D., Lee, D. (eds.) Lecture Notes in Comput Sci, vol. 4112, pp. 235–244. Springer, Berlin Heidelberg, Germany (2006)

[6] While most reconciliation algorithms do not support multiple samples per species nor non-binary gene trees, the former extension is fairly straightforward while the latter requires new algorithms.

6. Chauve, C., Doyon, J.P., El-Mabrouk, N.: Gene family evolution by duplication, speciation, and loss. J Comput Biol **15**(8), 1043–1062 (2008)
7. Chen, K., Durand, D., Farach-Colton, M.: NOTUNG: A program for dating gene duplications and optimizing gene family trees. J Comput Biol **7**(3-4), 429–447 (2000)
8. Chen, Z.Z., Deng, F., Wang, L.: Simultaneous identification of duplications, losses, and lateral gene transfers. IEEE/ACM Trans Comput Biol Bioinform **9**(5), 1515–1528 (2012)
9. Conow, C., Fielder, D., Ovadia, Y., Libeskind-Hadas, R.: Jane: a new tool for the cophylogeny reconstruction problem. Algorithm Mol Biol **5**(16) (2010)
10. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. Nature **469**(7328), 93–96 (2011)
11. Degnan, J.H., Rosenberg, N.A.: Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution **24**(6), 332–340 (2009)
12. Delsuc, F., Brinkmann, H., Philippe, H.: Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet **6**(5), 361–375 (2005)
13. Doyon, J.P., Scornavacca, C., Gorbunov, K., SzöllHosi, G.J., Ranwez, V., Berry, V.: An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: Tannier, E. (ed.) Comparative Genomics, Lecture Notes in Comput Sci, vol. 6398, pp. 93–108. Springer, Berlin Heidelberg, Germany (2011)
14. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool **28**(2), 132–163 (1979)
15. Górecki, P., Tiuryn, J.: Dls-trees: A model of evolutionary scenarios. Theoret Comput Sci **359**(1–3), 378–399 (2006)
16. Koonin, E.V.: Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet **39**(1), 309–338 (2005)
17. Kordi, M., Bansal, M.S.: Exact algorithms for duplication-transfer-loss reconciliation with non-binary gene trees. IEEE/ACM Trans Comput Biol Bioinform **PP**(99), 1–1 (2018)
18. Lafond, M., Swenson, K.M., El-Mabrouk, N.: An optimal reconciliation algorithm for gene trees with polytomies. In: Raphael, B., Tang, J. (eds.) Lecture Notes in Comput Sci, vol. 7534, pp. 106–122. Springer, Berlin Heidelberg, Germany (2012)
19. Maddison, W.P.: Gene trees in species trees. Syst Biol **46**(3), 523–536 (1997)
20. Ochman, H.: Lateral and oblique gene transfer. Curr Opin Genet Dev **11**(6), 616–619 (2001)
21. Ohno, S.: Evolution by Gene Duplication. Springer-Verlag New York, New York, NY, USA (1970)
22. Page, R.D.: Maps between trees and cladistic analysis of historical associations among genes,organisms, and areas. Syst Biol **43**(1), 58–77 (1994)
23. Rannala, B., Yang, Z.: Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics **164**(4), 1645–1656 (2003)
24. Rasmussen, M.D., Kellis, M.: A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol **28**(1), 273–290 (2011)
25. Rasmussen, M.D., Kellis, M.: Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res **22**, 755–765 (2012)
26. Rogers, J., Fishberg, A., Youngs, N., Wu, Y.C.: Reconciliation feasibility in the presence of gene duplication, loss, and coalescence with multiple individuals per species. BMC Bioinformatics **18**, 292– (2017)

27. Slowinski, J.B.: Molecular polytomies. Mol Phylogenet. Evol **19**(1), 114–120 (2001)
28. Stamatakis, A.: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**(21), 2688–2690 (2006)
29. Tofigh, A., Hallett, M., Lagergren, J.: Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans Comput Biol Bioinform **8**(2), 517–535 (2011)
30. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E.: EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res **19**(2), 327–335 (2009)
31. Wu, T., Zhang, L.: Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree. BMC Bioinf **12**(Suppl 9), S7– (2011)
32. Wu, Y.C., Rasmussen, M.D., Bansal, M.S., Kellis, M.: Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Research **24**(3), 475–486 (2014)
33. Zhang, L.: From gene trees to species trees ii: Species tree inference by minimizing deep coalescence events. IEEE/ACM Trans Comput Biol Bioinform **8**(6), 1685–1691 (2011)
34. Zheng, Y., Zhang, L.: Reconciliation with non-binary gene trees revisited. In: Proceedings of the 18th Annual International Conference on Research in Computational Moleculary Biology. pp. 418–432. RECOMB '14, Apr. 2–5, Springer International Publishing, Cham (2014)
35. Zmasek, C.M., Eddy, S.R.: A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics **17**(9), 821–828 (2001)