Reconciliation Revisited: Handling Multiple Optima when Reconciling with Duplication, Transfer, and Loss

Mukul S. Bansal[1], Eric J. Alm[2,3], and Manolis Kellis[1,3]

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA
[2] Dept. of Biological Engineering, Massachusetts Institute of Technology, Cambridge, USA
[3] Broad Institute of MIT and Harvard, Cambridge, USA
mukul@csail.mit.edu, ejalm@mit.edu, manoli@mit.edu

Abstract. Phylogenetic tree reconciliation is a powerful approach for inferring evolutionary events like gene duplication, horizontal gene transfer, and gene loss, which are fundamental to our understanding of molecular evolution. While Duplication-Loss (DL) reconciliation leads to a unique maximum-parsimony solution, Duplication-Transfer-Loss (DTL) reconciliation yields a multitude of optimal solutions, making it difficult the infer the true evolutionary history of the gene family.

Here, we present an effective, efficient, and scalable method for dealing with this fundamental problem in DTL reconciliation. Our approach works by sampling the space of optimal reconciliations uniformly at random and aggregating the results. We present an algorithm to efficiently sample the space of optimal reconciliations uniformly at random in $O(mn^2)$ time, where m and n denote the number of genes and species, respectively. We use these samples to understand how different optimal reconciliations vary in their node mapping and event assignments, and to investigate the impact of varying event costs.

Keywords: Gene family evolution, gene-tree/species-tree reconciliation, gene duplication, horizontal gene transfer, host-parasite cophylogeny, phylogenetics.

1 Introduction

The systematic comparison of a gene tree with its species tree under a reconciliation framework is a powerful technique for understanding gene family evolution. Specifically, gene tree/species tree reconciliation shows how the gene tree evolved inside the species tree while accounting for events like gene duplication, gene loss, and horizontal gene transfer, that drive gene family evolution. Thus, gene tree/species tree reconciliation is widely used and has many important applications; e.g., for inferring orthologs, paralogs and xenologs [1–6], reconstructing ancestral gene content and dating gene birth [7, 8], accurate gene tree reconstruction [5, 9], and whole genome species-tree reconstruction [10].

Duplication-Loss (DL) reconciliation, which accounts for only gene duplication and gene loss events, has been widely studied and extensively used [11–15]. However, since it does not account for horizontal gene transfer events, it only applies to multi-cellular eukaryotes, a very small part of the tree of life. An interesting and extremely useful

property of DL-reconciliation is that, assuming that loss events have a non-zero positive cost, the most parsimonious reconciliation is always unique [14]. In addition, the most parsimonious reconciliation remains the same irrespective of the chosen event costs for duplication and loss. Given these properties, there is no ambiguity in interpreting the results of DL-reconciliation, making it very easy to use in practice.

The limited applicability of DL reconciliation has led to the formulation of the Duplication-Transfer-Loss (DTL) reconciliation model, which can simultaneously account for duplication, transfer, and loss events and can be applied to species and gene families from across the entire tree of life. Indeed, the DTL-reconciliation model and its variants have been widely studied in the literature [8, 16–22]. In addition, DTL-reconciliation has also been indirectly studied in the context of the host-parasite cophylogeny problem [23–27].

The DTL-reconciliation problem is typically solved in a parsimony framework, where costs are assigned to duplication, transfer, and loss events, and the goal is to find a reconciliation with minimum total cost. DTL-reconciliations can sometimes be *time-inconsistent*; i.e, the inferred transfers may induce contradictory constraints on the dates for the internal nodes of the species tree. The problem of finding an optimal *time-consistent* reconciliation is known to be NP-hard [18, 27]. Thus, in practice, the goal is to find an optimal (but not necessarily time-consistent) DTL-reconciliation. The problem of finding an optimal time-consistent reconciliation does become efficiently solvable [17] if the species tree is fully dated. However, accurately dating the internal nodes of a species tree is a notoriously difficult problem [28], which severely restricts its applicability. Thus, for wider applicability and efficient solvability, in this work, unless otherwise stated, we assume the input species tree is undated and seek an optimal (not necessarily time-consistent) DTL-reconciliation [8, 18, 20, 21]. This problem can be solved very efficiently, with our own algorithm achieving the fastest known time complexity of $O(mn)$ [21], where m and n denote the number of nodes in the gene tree and species tree respectively.

Despite its extensive literature, the DTL-reconciliation problem remains difficult to use in practice for understanding gene family evolution. The first reason for this difficulty is that there are often multiple equally optimal reconciliations for a given gene tree and species tree and for a fixed assignment of event costs. The second reason is that event costs, which can be very difficult to assign confidently, play a much more important role than in DL reconciliation, as varying the costs can result in different optimal reconciliations.

Thus, when applying DTL-reconciliation in practice, it is unclear whether the evolutionary history implied by a particular given optimal solution is meaningful, as many other optimal reconciliations exist with the same minimal reconciliation cost. Moreover, it is unclear whether the properties of an optimal reconciliation are representative of the space of optimal reconciliations, and also how large and diverse this space is. Furthermore, the number of optimal reconciliations is often prohibitively large, as it can grow exponentially in the number of events required for the reconciliation, making even the basic task of enumerating all optimal reconciliations unfeasible for all but the smallest of gene trees [20]. Here, we directly address these problems and seek to make DTL-reconciliation as easy to use as the DL-reconciliation model.

Our contribution. In this work, we develop the first efficient and scalable approach to explore the space of optimal DTL-reconciliations and show how it can be used to infer the similarities and differences in the different optimal reconciliations for any given input instance. Our approach is based on uniformly random sampling of optimal reconciliations and we demonstrate the utility of our approach by applying it to a biological dataset of approximately 4700 gene trees from 100 (predominantly prokaryotic) taxa [8]. Specifically, our contributions are as follows:

1. We analyze the gene trees in the biological dataset and show that even gene trees with only a few dozen genes often have many millions of optimal reconciliations. This analysis provides the first detailed look into the prevalence of optimal reconciliations in biological datasets.
2. We show how to efficiently sample the space of optimal reconciliations uniformly at random. Our algorithm produces each random sample in $O(mn^2)$ time, where m and n denote the number of nodes in the gene tree and species tree, respectively. This algorithm is fast enough to be applied thousands of times to the same dataset and scalable enough to be applied to datasets with hundreds or thousands of taxa.
3. We use our algorithm for random sampling to explore the space of optimal reconciliations and investigate the similarities and differences between the different optimal reconciliations. We show how to distinguish between the parts of the reconciliation that have high support from those that are more variable across the different multiple optima.
4. We show that even in the presence of multiple optimal solutions, a large amount of shared information can be extracted from the different optimal reconciliations. For instance, we observed that, for fixed event costs, any internal node taken from a gene tree in the biological dataset had a 93.31% chance of having the same event assignment (speciation, duplication, or transfer) and a 73.15% chance of being mapped to the same species tree node, across all (sampled) optimal reconciliations.
5. Our method allows users to compare the space of optimal reconciliations for different event costs and extract the shared aspects of the reconciliation. This makes it possible to study the impact of using different event costs and to meaningfully apply DTL-reconciliation even if one is unsure of the exact event costs to use. We applied our method to the biological dataset using different event costs and observed that large parts of the reconciliation tend to be robust to event cost changes.

Thus, in this work, we introduce the first efficient and scalable method for exploring the space of optimal reconciliations. Our new method allows for the very first large-scale exploration of the space of optimal reconciliations in real biological datasets.

The remainder of the paper is organized as follows: The next section introduces basic definitions and preliminaries. In Section 3 we study the prevalence of multiple optimal reconciliations in biological data. We introduce our sampling based approach and algorithms in Section 4. Section 5 shows the results of our analysis of multiple optimal reconciliations for the biological dataset, and in Section 6 we show how our method can be applied to study the impact of using different reconciliation costs. Concluding remarks appear in Section 7.

2 Definitions and preliminaries

We follow the basic definitions and notation from [21]. Given a tree T, we denote its node, edge, and leaf sets by $V(T)$, $E(T)$, and $Le(T)$ respectively. If T is rooted, the root node of T is denoted by $rt(T)$, the parent of a node $v \in V(T)$ by $pa_T(v)$, its set of children by $Ch_T(v)$, and the (maximal) subtree of T rooted at v by $T(v)$. If two nodes in T have the same parent, they are called *siblings*. The set of *internal nodes* of T, denoted $I(T)$, is defined to be $V(T) \setminus Le(T)$. We define $\leq_T$ to be the partial order on $V(T)$ where $x \leq_T y$ if y is a node on the path between $rt(T)$ and x. The partial order $\geq_T$ is defined analogously, i.e., $x \geq_T y$ if x is a node on the path between $rt(T)$ and y. We say that v is an *ancestor* of u, or that u is a *descendant* of v, if $u \leq_T v$ (note that, under this definition, every node is a descendant as well as ancestor of itself). We say that x and y are *incomparable* if neither $u \leq_T v$ nor $v \leq_T u$. Given a non-empty subset $L \subseteq Le(T)$, we denote by $lca_T(L)$ the least common ancestor (LCA) of all the leaves in L in tree T; that is, $lca_T(L)$ is the unique smallest upper bound of L under $\leq_T$. Given $x, y \in V(T)$, $x \rightarrow_T y$ denotes the unique path from x to y in T. We denote by $d_T(x, y)$ the number of edges on the path $x \rightarrow_T y$. Throughout this work, unless otherwise stated, the term tree refers to a rooted binary tree.

We assume that each leaf of the gene trees is labeled with the species from which that gene was sampled. This labeling defines a *leaf-mapping* $\mathcal{L}_{G,S} \colon Le(G) \rightarrow Le(S)$ that maps a leaf node $g \in Le(G)$ to that unique leaf node $s \in Le(S)$ which has the same label as g. Note that gene trees may have more than one gene sampled from the same species. Throughout this work, we denote the gene tree and species tree under consideration by G and S respectively and will assume that $\mathcal{L}_{G,S}(g)$ is well defined.

2.1 Reconciliation and DTL-scenarios

Reconciling a gene tree with a species tree involves mapping the gene tree into the species tree. Next, we define what constitutes a valid reconciliation; specifically, we define a Duplication-Transfer-Loss scenario (DTL-scenario) [18, 21] for G and S that characterizes the mappings of G into S that constitute a biologically valid reconciliation. Essentially, DTL-scenarios map each gene tree node to a unique species tree node in a consistent way that respects the immediate temporal constraints implied by the species tree, and designate each gene tree node as representing either a speciation, duplication, or transfer event.

Definition 1 (DTL-scenario). *A DTL-scenario for G and S is a seven-tuple $\langle \mathcal{L}, \mathcal{M}, \Sigma, \Delta, \Theta, \Xi, \tau \rangle$, where $\mathcal{L} \colon Le(G) \rightarrow Le(S)$ represents the leaf-mapping from G to S, $\mathcal{M} \colon V(G) \rightarrow V(S)$ maps each node of G to a node of S, the sets Σ, Δ, and Θ partition $I(G)$ into speciation, duplication, and transfer nodes respectively, Ξ is a subset of gene tree edges that represent transfer edges, and $\tau \colon \Theta \rightarrow V(S)$ specifies the recipient species for each transfer event, subject to the following constraints:*

1. *If $g \in Le(G)$, then $\mathcal{M}(g) = \mathcal{L}(g)$.*
2. *If $g \in I(G)$ and g' and g'' denote the children of g, then,*
   (a) $\mathcal{M}(g) \not\leq_S \mathcal{M}(g')$ and $\mathcal{M}(g) \not\leq_S \mathcal{M}(g'')$,

(b) At least one of $\mathcal{M}(g')$ and $\mathcal{M}(g'')$ is a descendant of $\mathcal{M}(g)$.

3. *Given any edge $(g, g') \in E(G)$, $(g, g') \in \Xi$ if and only if $\mathcal{M}(g)$ and $\mathcal{M}(g')$ are incomparable.*
4. *If $g \in I(G)$ and g' and g'' denote the children of g, then,*
   (a) $g \in \Sigma$ only if $\mathcal{M}(g) = lca(\mathcal{M}(g'), \mathcal{M}(g''))$ and $\mathcal{M}(g')$ and $\mathcal{M}(g'')$ are incomparable,
   (b) $g \in \Delta$ only if $\mathcal{M}(g) \geq_S lca(\mathcal{M}(g'), \mathcal{M}(g''))$,
   (c) $g \in \Theta$ if and only if either $(g, g') \in \Xi$ or $(g, g'') \in \Xi$.
   (d) If $g \in \Theta$ and $(g, g') \in \Xi$, then $\mathcal{M}(g)$ and $\tau(g)$ must be incomparable, and $\mathcal{M}(g')$ must be a descendant of $\tau(g)$, i.e., $\mathcal{M}(g') \leq_S \tau(g)$.

Constraint 1 above ensures that the mapping $\mathcal{M}$ is consistent with the leaf-mapping $\mathcal{L}$. Constraint 2(a) imposes on $\mathcal{M}$ the temporal constraints implied by S. Constraint 2(b) implies that any internal node in G may represent at most one transfer event. Constraint 3 determines the edges of G that are transfer edges. Constraints 4(a), 4(b), and 4(c) state the conditions under which an internal node of G may represent a speciation, duplication, and transfer respectively. Constraint 4(d) specifies which species may be designated as the recipient species for any given transfer event.

In some cases, one may wish to restrict transfer events to only occur between co-existing species. This requires that divergence time information (either absolute or relative) be available for all the internal nodes of the species tree. In such cases, the definition of a DTL-scenario remains the same, except for the additional restriction on transfer events.

DTL-scenarios correspond naturally to reconciliations and it is straightforward to infer the reconciliation of G and S implied by any DTL-scenario. Figure 1 shows two simple DTL-scenarios. Given a DTL-scenario, one can directly count the minimum number of gene losses in the corresponding reconciliation. For brevity, we refer the reader to [21] for further details on how to count losses in DTL-scenarios.

Let P_Δ, P_Θ, and P_{loss} denote the costs associated with duplication, transfer, and loss events respectively. The reconciliation cost of a DTL-scenario is defined as follows.

Definition 2 (Reconciliation cost of a DTL-scenario). *Given a DTL-scenario $\alpha = \langle \mathcal{L}, \mathcal{M}, \Sigma, \Delta, \Theta, \Xi, \tau \rangle$ for G and S, the* reconciliation cost *associated with α is given by $\mathcal{R}_\alpha = P_\Delta \cdot |\Delta| + P_\Theta \cdot |\Theta| + P_{loss} \cdot Loss_\alpha$.*

Given G and S, along with event costs P_Δ, P_Θ, and P_{loss}, the goal is to find a most parsimonious reconciliation of G and S. More formally,

Problem 1 (Most Parsimonious Reconciliation (MPR)) *Given G and S, the* most parsimonious reconciliation (MPR) *problem is to find a DTL-scenario for G and S with minimum reconciliation cost.*

We distinguish two versions of the MPR problem: (i) The *Undated MPR (U-MPR)* problem where the species tree is undated, and (ii) the *Fully-dated MPR (D-MPR)* problem where every node of the species tree has an associated divergence time estimate (or there is a known total order on the internal nodes of the species tree) and transfer events are required to occur only between coexisting species.

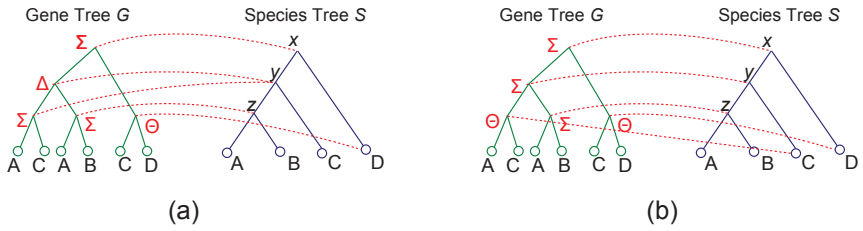

Fig. 1. Multiple optimal reconciliations. Parts (a) and (b) show two different reconciliations for the gene tree and species tree depicted in the figure. Both of the reconciliations are optimal for event costs $P_\Delta = 1$, $P_\Theta = 3$, and $P_{loss} = 1$. The reconciliation in part (a) invokes one duplication, one transfer, and two losses, while the reconciliation in part (b) invokes two transfers.

3 Multiple optimal solutions

In general, for any fixed values of P_Δ, P_Θ, and P_{loss}, there may be multiple equally optimal solutions to the MPR problem (both U-MPR and D-MPR). This is illustrated in Figure 1. The figure also illustrates the fundamental problem with having multiple optima: Given the different evolutionary histories implied by the different multiple optima, what is the true evolutionary history of the gene family? We address this problem in this paper. But first, in this section, we investigate the prevalence of optimal reconciliations in real datasets. For our study, we use a published biological dataset of 4735 gene trees and 100 (predominantly prokaryotic) species [8]. The gene trees in the dataset have median and average leaf-set sizes of 18 and 35.1, respectively. This dataset has been previously analyzed using DTL-reconciliation but without consideration of multiple optima. In our analysis of this dataset we used the same event costs as used in [8] (i.e., $P_\Delta = 2$, $P_\Theta = 3$, and $P_{loss} = 1$). Since the gene trees in the dataset are unrooted, we first rooted them optimally by choosing a root that minimized the reconciliation cost. In cases where there were multiple optimal rootings, we chose one of the optimal rootings at random. We computed the number of multiple optimal reconciliations for each of the rooted gene trees by augmenting the dynamic programming algorithm used to solve the MPR problem (e.g., [21]) to keep track of the number of optima for each sub-problem. Further algorithmic details appear in Section 4. Unless otherwise stated, all analyses in the manuscript were performed using the undated version of DTL-reconciliation.

Figure 2 shows the results of our analysis. As part (a) of the figure shows, only 17% of the approximately 4700 gene trees have a unique optimal reconciliation. Over half of the gene trees have over 100 optimal reconciliations and 15% have more than 10,000 optimal reconciliations. This illustrates the extent of the problem with multiple optimal reconciliations in biological datasets. As part (b) of the figure shows, the number of optimal reconciliations tends to increase exponentially with gene tree size. These results demonstrate the importance of considering multiple optima in DTL-reconciliation, and the impracticality of enumerating all optimal reconciliations for all but the smallest gene trees.

We also repeated the above analysis using the dated version of the DTL-reconciliation problem (i.e., the D-MPR problem), and observed no significant reduction in the num-

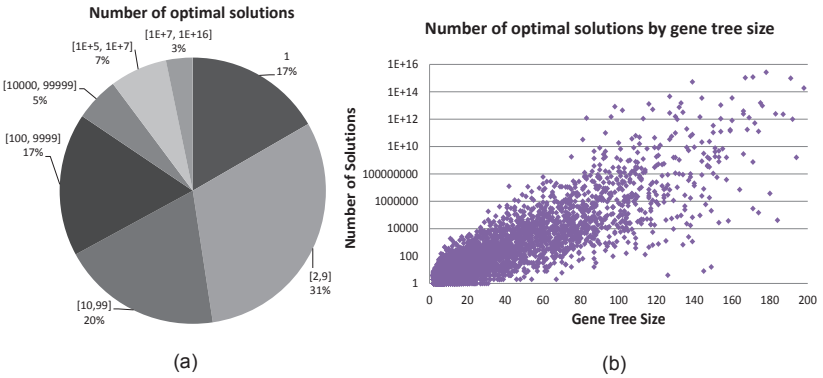

Fig. 2. Number of optimal reconciliations for the gene trees in the biological dataset. The pie chart in part (a) shows the distribution of the number of optimal reconciliations for the gene trees in the biological dataset. The dot plot in part (b) plots the size (number of internal nodes) and the number of optimal reconciliations for each gene tree. Due to arithmetic overflow concerns, results are only shown for the 4699 (out of 4735) gene trees that had fewer than 10^{16} optima.

ber of multiple optima. For instance, even for the dated version, 14% of the gene trees had more than 10,000 optimal reconciliations.

Recall that the gene trees in the dataset were originally unrooted. While the results above are for a fixed optimal rooting of these gene trees, we point out that about half the gene trees in the dataset have more than one optimal rooting. It may thus be necessary, in practice, to either consider all possible optimal rootings when studying multiple optimal reconciliations, or to use other information to assign a root uniquely.

4 Uniformly random sampling of optimal reconciliations

As Section 3 demonstrates, the exhaustive enumeration of all optimal reconciliations is only feasible for very small gene trees. In this section we show how to sample the space of reconciliations uniformly at random. Random sampling makes it possible to explore the space of optimal reconciliations without exhaustive enumeration, and makes it possible to understand the variability in the different reconciliations and to distinguish between the highly supported and weakly supported parts of a given optimal reconciliation. Our algorithm for random sampling is based on the dynamic programming algorithm for the MPR problem from [21]. The idea is to keep track of the number of optimal solutions for each subproblem considered in the dynamic programming algorithm. In the following, we show how to compute the number of optimal solutions at each step correctly and efficiently. First, we need a few definitions.

Given any $g \in I(G)$ and $s \in V(S)$, let $c_{\Sigma}(g, s)$ denote the cost of an optimal reconciliation of $G(g)$ with S such that g maps to s and $g \in \Sigma$. The terms $c_{\Delta}(g, s)$ and $c_{\Theta}(g, s)$ are defined similarly for $g \in \Delta$ and $g \in \Theta$ respectively. Given any $g \in V(G)$ and $s \in V(S)$, we define $c(g, s)$ to be the cost of an optimal reconciliation of $G(g)$ with

S such that g maps to s. The algorithm for the MPR problem performs a nested post-order traversal of the gene tree and species tree to compute the value of $c(g,s)$ for each g and s. The dynamic programming table is initialized as follows for each $g \in \mathit{Le}(G)$: $c(g,s) = 0$ if $s = \mathcal{M}(g)$, and $c(g,s) = \infty$ otherwise. For $g \in I(G)$, observe that $c(g,s) = \min\{c_{\Sigma}(g,s), c_{\Delta}(g,s), c_{\Theta}(g,s)\}$.

At each step, the values of $c_{\Sigma}(g,s)$, $c_{\Delta}(g,s)$, and $c_{\Theta}(g,s)$ for any $g \in I(G)$ and $s \in V(S)$, can be computed based on the previously computed values of $c(\cdot,\cdot)$. To show how $c_{\Sigma}(g,s)$, $c_{\Delta}(g,s)$, and $c_{\Theta}(g,s)$ are computed we need some additional notation. Let $\mathit{in}(g,s) = \min_{x \in V(S(s))}\{P_{loss} \cdot d_S(s,x) + c(g,x)\}$ and $\mathit{out}(g,s) = \min_{x \in V(S) \text{ incomparable to } s} c(g,x)$. In other words: $\mathit{out}(g,s)$ is the cost of an optimal reconciliation of $G(g)$ with S such that g may map to any node from $V(S)$ that is incomparable to s; and $\mathit{in}(g,s)$ is the cost of an optimal reconciliation of $G(g)$ with S such that g may map to any node, say x, in $V(S(s))$ but with an additional reconciliation cost of one loss event for each edge on the path from s to x. The values $c_{\Sigma}(g,s)$, $c_{\Delta}(g,s)$, and $c_{\Theta}(g,s)$ are computed as follows:

For any $g \in I(G)$ and $s \in I(S)$, let $\{g', g''\} = \mathit{Ch}_G(g)$ and $\{s', s''\} = \mathit{Ch}_S(s)$.

If $s \in \mathit{Le}(S)$ then,
$c_{\Sigma}(g,s) = \infty$,
$c_{\Delta}(g,s) = P_{\Delta} + c(g',s) + c(g'',s)$, and
If $s \neq \mathit{rt}(S)$, then $c_{\Theta}(g,s) = P_{\Theta} + \min\{\mathit{in}(g',s) + \mathit{out}(g'',s),\ \mathit{in}(g'',s) + \mathit{out}(g',s)\}$. Else, $c_{\Theta}(g,s) = \infty$.

If $s \in I(S)$ then,
$c_{\Sigma}(g,s) = \min\{\mathit{in}(g',s') + \mathit{in}(g'',s''),\ \mathit{in}(g'',s') + \mathit{in}(g',s'')\}$.

$$c_{\Delta}(g,s) = P_{\Delta} + \min \begin{cases} c(g',s) + \mathit{in}(g'',s'') + P_{loss},\ c(g',s) + \mathit{in}(g'',s') + P_{loss}, \\ c(g'',s) + \mathit{in}(g',s'') + P_{loss},\ c(g'',s) + \mathit{in}(g',s') + P_{loss}, \\ c(g',s) + c(g'',s),\ \mathit{in}(g',s') + \mathit{in}(g'',s'') + 2P_{loss}, \\ \mathit{in}(g',s'') + \mathit{in}(g'',s') + 2P_{loss},\ \mathit{in}(g',s') + \mathit{in}(g'',s') + 2P_{loss}, \\ \mathit{in}(g',s'') + \mathit{in}(g'',s'') + 2P_{loss}. \end{cases}$$

If $s \neq \mathit{rt}(S)$, then $c_{\Theta}(g,s) = P_{\Theta} + \min\{\mathit{in}(g',s) + \mathit{out}(g'',s),\ \mathit{in}(g'',s) + \mathit{out}(g',s)\}$. Else, $c_{\Theta}(g,s) = \infty$.

The optimal reconciliation cost of G and S is simply: $\min_{s \in V(S)} c(\mathit{rt}(G), s)$, and an optimal reconciliation with that cost can be reconstructed by backtracking in the dynamic programming table. We refer the reader to [21] for further algorithmic details.

To output optimal reconciliations uniformly at random we must keep track of the number of optimal reconciliations for each of the subproblems considered in the DP algorithm. We define the following: For any $g \in V(G)$ and $s \in V(S)$, let $N(g,s)$ denote the number of optimal solutions for reconciling $G(g)$ with S such that g maps to s. The idea is to compute $N(\cdot,\cdot)$ using the same nested post-order traversal used to compute the $c(\cdot,\cdot)$ values. The dynamic programming table for $N(\cdot,\cdot)$ is initialized as follows for each $g \in \mathit{Le}(G)$: $N(g,s) = 1$ if $s = \mathcal{M}(g)$, and $N(g,s) = 0$ otherwise. To compute $N(g,s)$, for $g \in I(G)$, we must consider all possible mappings of g' and g'' that yield a cost of $c(g,s)$. For the remainder of this discussion, in the interest of brevity and clarity, we will assume that $s \in I(S)$ and $s \neq \mathit{rt}(S)$; the cases when $s \in \mathit{Le}(S)$ or $s = \mathit{rt}(S)$ are easy to handle analogously.

Let a_1 through a_{13} denote the individual expressions in the $min\{\ \}$ blocks in the equations for $c_{\Sigma}(g,s)$, $c_{\Delta}(g,s)$, and $c_{\Theta}(g,s)$ above. Specifically, let a_1 denote $\mathit{in}(g',s') + \mathit{in}(g'',s'')$, a_2 denote $\mathit{in}(g'',s') + \mathit{in}(g',s'')$, a_3 through a_{11} denote the nine expressions in the $\min\{\ \}$ block for $c_{\Delta}(g,s)$, and a_{12} and a_{13} denote the two expressions in the $\min\{\ \}$ block for $c_{\Theta}(g,s)$. Each of these a_i's represents a certain cost, which we denote by $c(a_i)$, and a certain number of optimal reconciliations, which we denote by $N(a_i)$. Furthermore, let b_i, for $1 \leq i \leq 13$, be binary boolean variables associated with the a_i's such that $b_i = 1$ if a_i yields the minimum cost $c(g,s)$, and $b_i = 0$ otherwise. Specifically, for $i \in \{1,2\}$, $b_i = 1$ if and only if $c(a_i) = c(g,s)$; for $i \in \{3,\ldots,11\}$, $b_i = 1$ if and only if $c(a_i) + P_{\Delta} = c(g,s)$; and for $i \in \{12,13\}$, $b_i = 1$ if and only if $c(a_i) + P_{\Theta} = c(g,s)$. Then, we must have:

$$N(g,s) = \sum_{i=1}^{13} b_i \times N(a_i).$$

Next, we show how to compute $N(a_i)$ for any i for which $b_i = 1$. Observe that each a_i has one term involving g' and one term involving g''. These terms take one of the three forms: $c(\cdot,\cdot)$, $\mathit{in}(\cdot,\cdot)$, or $\mathit{out}(\cdot,\cdot)$. These terms, involving g' and g'', can be viewed as representing the choice of optimal mappings for g' and g'', respectively. For instance, $c(g',s)$ implies that g' must map to s, $\mathit{in}(g',s)$ implies that g' may map to any node $x \in V(S(s))$ for which $(P_{loss} \cdot d_S(s,x) + c(g',x))$ is minimized (recall the definition of $\mathit{in}(\cdot,\cdot)$), and $\mathit{out}(g',s)$ implies that g' may map to any node $x \in V(S)$ that is incomparable to s, for which $c(g',x)$ is minimized. Based on this observation, for any given a_i, we can compute a set of optimal mappings for g', which we will denote by X' and a set of optimal mappings for g'', which we will denote by X''. The value of $N(a_i)$ can then be computed as follows:

$$N(a_i) = \left(\sum_{x \in X'} N(g',x)\right) \times \left(\sum_{x \in X''} N(g'',x)\right).$$

The equations for $N(g,s)$ and $N(a_i)$ above make it possible to compute the value $N(g,s)$ for each $g \in I(G)$ and $s \in V(S)$ by using the same nested post-order traversal that is used for computing the values $c(\cdot,\cdot)$. Once all the $c(\cdot,\cdot)$ and $N(\cdot,\cdot)$ have been computed, an optimal reconciliation itself can be built by backtracking through the dynamic programming table. To ensure that reconciliations are generated uniformly at random the idea is to make the choice of mapping assignments based on the number of optimal solutions contained within each choice. For instance, if a node g has already been assigned a mapping, its two children g' and g'' must be assigned mappings jointly based on their joint probability mass. In the interest of brevity, further technical and algorithmic details, as well as a formal proof of correctness, are deferred to the full version of this paper.

It is not hard to implement this algorithm for uniformly random sampling in $O(mn^2)$ time, where m and n denote the size of the gene tree and species tree respectively. This is only a factor of n slower than the fastest known algorithm for the MPR problem [21]. Our implementation of this random sampling algorithm will be made available as part of the next version of the RANGER-DTL software package [21].

5 Exploring the space of optimal reconciliations

We applied our method to the biological dataset to understand the space of optimal reconciliations for the gene trees in this dataset. As before, we used event costs $P_{\Delta} = 2$, $P_{\Theta} = 3$, and $P_{loss} = 1$ for this analysis. For this study, we focused on understanding how similar the different optimal reconciliations are to each other. To that end, we used our algorithm to sample 500 optimal reconciliations for each gene tree, and wrote a program that reads in these samples and summarizes them as follows: For each internal node in the gene tree we (i) consider the fraction of times that node is mapped to the different nodes of the species tree, and (ii) consider the fraction of times that node is labeled as a speciation, duplication, and transfer event. We used this to investigate the stability of the embedding of the gene tree into the species tree (i.e., the stability of gene node mappings), and the stability of event assignments for the internal nodes of the gene tree.

We first checked to see how stable the gene node mappings were across the internal nodes in all the 4699 gene trees. Figure 3(a) shows the results of this analysis. Overall, we observed that mappings tended to be fairly well conserved across the different multiple optima. For instance, we observed that 73.15% of the internal gene tree nodes had the same mapping across all 500 samples. Recall that only 17% of the gene trees have a unique solution. We also repeated this analysis for event assignments and these results are also shown in Figure 3(a). Amazingly, we observed that 93.31% of the nodes had a consistent event assignment across all 500 samples. This suggests that event assignments tend to be highly conserved across the different multiple optima. Thus, even in those instances where there are many different optimal reconciliations it should be possible to confidently assign event types to most internal nodes of the gene tree (even though the mappings of the nodes themselves may not be consistent across the different multiple optima)). This has important implications for understanding gene family evolution, since the inference of orthologs, paralogs, and xenologs depends only on the event assignments for gene tree nodes.

In practice, users are often interested in analyzing the evolutionary history of a specific gene family. We thus asked the following question: Given a gene tree from the biological dataset, what fraction of its nodes can be expected to have (i) a consistent mapping, and (ii) a consistent event assignment, across all 500 samples. Figure 3(b) shows the results of this analysis. The results show that for most gene trees, event assignments are completely consistent across all samples for most of their internal nodes. For instance, we observed that 60.2% of the gene trees have a consistent event assignment for all of their internal nodes, and almost all gene trees had a consistent event assignment for at least half of their internal nodes. As we observed before, gene tree node mappings tend to be more variable, but still, over 91% of the gene trees had a consistent mapping for at least half of their internal nodes. We also tested to see if there was a correlation between the number of optimal reconciliations for a gene tree and fraction of its internal nodes with consistent mappings or consistent event assignments. To our surprise, we found no correlation (results not shown).

Our analyses above show that, even in the presence of multiple optimal reconciliations, most aspects of the reconciliation are highly conserved across the different multiple optima.

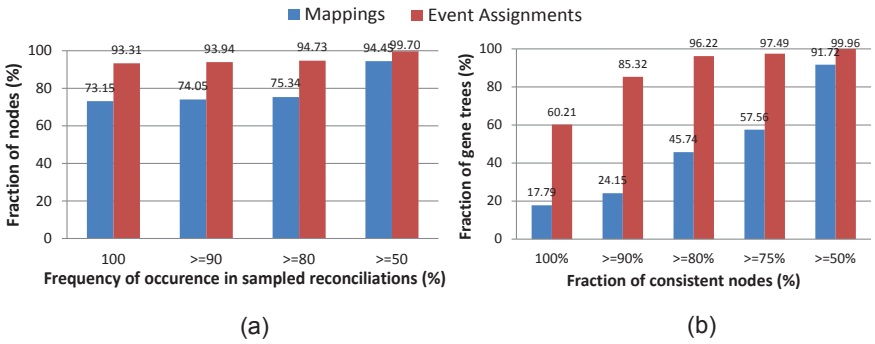

Fig. 3. Stability of mappings and event assignments. The plot in part (a) shows the fraction of internal nodes from the 4699 gene trees that have the same mapping or the same event assignment across at least a certain fraction of the 500 samples. The plot in part (b) plots the fraction of the 4699 gene trees that have at least a certain fraction of their nodes with a consistent mapping or a consistent event assignment across all 500 samples.

6 Application to understanding sensitivity to event costs

The ability to explore the space of multiple optimal reconciliations makes it possible to study the effect of using different event costs on the reconciliation. For instance, one can compare if the mapping or event assignments that are consistent across the multiple optima for a particular event cost assignment are also consistent across a different event cost assignment. Similarly, if one is unsure of which event cost assignment to use, one can try out all the different event costs, compute a set of random samples for each event cost assignment, and aggregate the samples from all event cost assignments into a single analysis to understand which aspects of the reconciliation are conserved across the different event cost assignments.

We performed a preliminary study of the effect of using different event costs on the analysis of the biological dataset. Recall that our default event costs are $P_{\Delta} = 2$, $P_{\Theta} = 3$, and $P_{loss} = 1$. For this study, we kept $P_{loss} = 1$, but considered the following combinations of the duplication and transfer costs: (i) $P_{\Delta} = 2$, $P_{\Theta} = 4$, (ii) $P_{\Delta} = 2$, $P_{\Theta} = 2$, (iii) $P_{\Delta} = 3$, $P_{\Theta} = 3$, and (iv) $P_{\Delta} = 1$, $P_{\Theta} = 1$. We computed 100 random samples for each setting of event costs. For our preliminary analysis, we asked the following question: What fraction of the gene tree nodes with consistent mappings (event assignments) under the default costs also have the same consistent mappings (resp. event assignments) under the alternative event costs? The results of this analysis for the four combinations of event costs listed aboveare as follows: For mappings, the fractions are 94%, 83.38%, 92.04%, and 63.97%, respectively. And, for event assignments, the fractions are 92.06%, 91.52%, 96.07%, and 80.37%, respectively. As the analysis indicates, consistent mappings and event assignments tend to be well conserved even when using different event costs. Even with the rather extreme event costs of $P_{\Delta} = P_{\Theta} = P_{loss} = 1$, almost 64% of the consistent mappings and over 80% of the event assignments are conserved. We defer a more detailed analysis of the differences

in the space of optimal reconciliations for the different event cost assignments to the full version of the paper.

7 Conclusion

In this work, we have presented an efficient and scalable approach for the problem of multiple optimal DTL-reconciliations. Our approach is based on random sampling and we show how to sample the space of optimal reconciliations uniformly at random efficiently in $O(mn^2)$ time per sample. The sampling based approach makes it possible for users to explore the space of optimal reconciliations and to distinguish between stable and unstable parts of the reconciliation. This approach also allows users to investigate the effect of using different event costs on the reconciliation. Our analysis of the biological dataset provides the first real insight into the space of multiple optima and reveals that many, if not most, aspects of the reconciliation remain consistent across the different multiple optima and that these can be efficiently inferred. We believe that this work represents an important step towards making DTL-reconciliation a practical method for understanding gene family evolution.

Many aspects of the space of optimal reconciliations remain to be explored. For instance, it would be interesting to investigate why so many of the input instances have millions (and more) of multiple optima. In this work we did not consider the effect of alternative optimal gene tree rootings on the reconciliation space and we would like to study this further. The ability to handle multiple optima also enables the systematic evaluation of the accuracy of DTL-reconciliation at inferring evolutionary history correctly and we plan to pursue this further. Similarly, we only performed a very preliminary study of the effect of different event costs and it would be instructive to study this more thoroughly.

Funding: This work was supported by a National Science Foundation CAREER award 0644282 to MK. National Institutes of Health grant RC2 HG005639 to M.K., and National Science Foundation AToL grant 0936234 to E.J.A. and M.K.

References

1. Storm, C.E.V., Sonnhammer, E.L.L.: Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics **18**(1) (2002) 92–99
2. Koonin, E.V.: Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics **39**(1) (2005) 309–338
3. Wapinski, I., Pferrer, A., Friedman, N., Regev, A.: Natural history and evolutionary principles of gene duplication in fungi. Nature **449** (2007) 54–61
4. van der Heijden, R., Snel, B., van Noort, V., Huynen, M.: Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics **8**(1) (2007) 83
5. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E.: Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research **19**(2) (2009) 327–335
6. Sennblad, B., Lagergren, J.: Probabilistic orthology analysis. Syst. Biol. **58**(4) (2009) 411–424

7. Chen, K., Durand, D., Farach-Colton, M.: Notung: dating gene duplications using gene family trees. In: RECOMB. (2000) 96–106
8. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. Nature **469** (2011) 93–96
9. Rasmussen, M.D., Kellis, M.: A bayesian approach for fast and accurate gene tree reconstruction. Molecular Biology and Evolution **28**(1) (2011) 273–290
10. 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) (2011) 117–125
11. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage. a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology **28** (1979) 132–163
12. Page, R.D.M.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. **43**(1) (1994) 58–77
13. Bonizzoni, P., Vedova, G.D., Dondi, R.: Reconciling a gene tree to a species tree under the duplication cost model. Theor. Comput. Sci. **347**(1-2) (2005) 36–53
14. Górecki, P., Tiuryn, J.: Dls-trees: A model of evolutionary scenarios. Theor. Comput. Sci. **359** (2006) 378–399
15. Chauve, C., Doyon, J.P., El-Mabrouk, N.: Gene family evolution by duplication, speciation, and loss. J. Comput. Biol.. **15**(8) (2008) 1043–1062
16. Gorbunov, K.Y., Liubetskii, V.A.: Reconstructing genes evolution along a species tree. Molekuliarnaia Biologiia **43**(5) (2009) 946–958
17. Doyon, J.P., Scornavacca, C., Gorbunov, K.Y., Szöllosi, G.J., Ranwez, V., Berry, V.: An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In Tannier, E., ed.: RECOMB-CG. Volume 6398 of Lecture Notes in Computer Science., Springer (2010) 93–108
18. Tofigh, A., Hallett, M.T., Lagergren, J.: Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans. Comput. Biology Bioinform. **8**(2) (2011) 517–535
19. Tofigh, A.: Using Trees to Capture Reticulate Evolution : Lateral Gene Transfers and Cancer Progression. PhD thesis, KTH Royal Institute of Technology (2009)
20. Chen, Z.Z., Deng, F., Wang, L.: Simultaneous identification of duplications, losses, and lateral gene transfers. IEEE/ACM Trans. Comput. Biology Bioinform. **9**(5) (2012) 1515–1528
21. Bansal, M.S., Alm, E.J., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**(12) (2012) 283–291
22. Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., Durand, D.: Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics **28**(18) (2012) 409–415
23. Charleston, M.: Jungles: A new solution to the host-parasite phylogeny reconciliation problem. Mathematical Biosciences **149** (1998) 191–223
24. Ronquist, F.: Parsimony analysis of coevolving species associations. In Page, R.D.M., ed.: Tangled Trees: Phylogeny, Cospeciation and Coevolution. The University of Chicago Press (2003) 22–64
25. Merkle, D., Middendorf, M., Wieseke, N.: A parameter-adaptive dynamic programming approach for inferring cophylogenies. BMC Bioinformatics **11**(Suppl 1) (2010) S60
26. Conow, C., Fielder, D., Ovadia, Y., Libeskind-Hadas, R.: Jane: a new tool for the cophylogeny reconstruction problem. Algorithm. Mol. Biol. **5**(1) (2010) 16
27. Ovadia, Y., Fielder, D., Conow, C., Libeskind-Hadas, R.: The cophylogeny reconstruction problem is np-complete. J. Comput. Biol. **18**(1) (2011) 59–65
28. Rutschmann, F.: Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. Divers. Distrib. **12**(1) (2006) 35–48