Gene tree species tree reconciliation with gene conversion

Damir Hasic, Eric Tannier

manuscript No.
(will be inserted by the editor)

Gene tree species tree reconciliation with gene conversion

Damir Hasić · Eric Tannier

Abstract Gene tree/species tree reconciliation is a recent decisive progress in phylogenetic methods, accounting for the possible differences between gene histories and species histories. Reconciliation consists in explaining these differences by gene-scale events such as duplication, loss, transfer, which translates mathematically into a mapping between gene tree nodes and species tree nodes or branches. Gene conversion is a frequent and important biological event, which results in the replacement of a gene by a copy of another from the same species and in the same gene tree. Including this event in reconciliations has never been attempted because this changes as well the solutions as the methods to construct reconciliations. Standard algorithms based on dynamic programming become ineffective. We propose here a novel mathematical framework including gene conversion as an evolutionary event in gene tree/species tree reconciliation. We describe a randomized algorithm giving in polynomial running time a reconciliation minimizing the number of duplications, losses and conversions. We show that the space of reconciliations includes an analog of the Last Common Ancestor reconciliation, but is not limited to it. Our algorithm outputs any optimal reconciliation with non null probability. We argue that this study opens a research avenue on including gene conversion in reconciliation, which can be important for biology.

This work is funded by the Agence Nationale pour la Recherche, Ancestrome project ANR-10-BINF-01-01.

Damir Hasić
Department of Mathematics, Faculty of Science, University of Sarajevo, 71000 Sarajevo, Bosnia and Herzegovina
E-mail: damir.hasic@gmail.com, d.hasic@pmf.unsa.ba

Eric Tannier
Inria Grenoble Rhône-Alpes, F-38334 Montbonnot, France
Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France

Keywords phylogenetic reconciliation · gene conversion · gene duplication · gene loss · random algorithms · all optimal solutions

Mathematics Subject Classification (2000) 92D15 · 05C90 · 92-08 · 68W40

1 Introduction

1.1 Biological motivation

Due to various evolutionary events on a gene level, gene trees (trees used to describe the evolution of genes) and species trees (trees used to describe the evolution of species) are often not identical. Identifying these evolutionary events, such as speciation, duplication, transfer, conversion, transfer with replacement, and their positioning inside species tree is called *phylogenetic reconciliation*.

Tree reconciliation techniques become widely used in biology. For example they are used in testing hypotheses of horizontal transfer in some Bacterial and Archaeal species (Planet et al. 2003); studying parasites infecting tropheine cichlids (Vanhove et al. 2015); finding horizontal gene transfers of RH50 among prokaryotes (Matassi 2017). Reconciliation tools (Szöllősi et al. 2012, 2013a,b) are also used to explore the process of shaping gut microbiomes (Groussin et al. 2017). In Dufayard et al. (2005); Storm and Sonnhammer (2002); van der Heijden et al. (2007) reconciliations are used "for inferring orthology relationships" (Doyon et al. 2011), and in Bourgon et al. (2004); Searls (2003) "for identifying orthologs for use in function prediction, gene annotation, planning experiments in model organisms, and identifying drug targets" (Vernot et al. 2008). From Page and Charleston (1998); Brooks and Ferrao (2005) we can see that "reconciliation can also be used to study co-evolution between parasites and their hosts (parasitology), and between organisms and their living areas (biogeography)" (Doyon et al. 2011).

An evolutionary event of particular interest in this paper is *gene conversion*. It is a highly important genomic event for evolution and health (Chen et al. 2007). It results in the replacement of a gene in a genome by another homologous gene from the same genome, where homologous means that they have a common ancestor. It has largely contributed to shaping extant eukaryotic genomes and is involved in several known human genetic diseases (Ko et al. 2011).

However, gene conversion is nearly absent from the mathematical framework for phylogeny. Phylogenetic methods can handle base substitutions, indels (Felsenstein 2004), genome rearrangements (Hu et al. 2014), duplications, transfers and losses of genes (Szöllősi et al. 2015) or population scale events as incomplete lineage sorting (Mirarab et al. 2014). But the detection of gene conversion is still done with empirical examinations of gene trees combined with various genomic features (Hsu et al. 2010; Mansai and Innan 2010).

This absence of gene conversion can strongly bias evolutionary studies. Indeed, it introduces a discordance between the history of a gene and the history of a locus (Rasmussen and Kellis 2012) which stays unresolved. It makes the confusion between duplications and conversions (Boussau et al. 2013), whereas conversions are probably more frequent (Kejnovsky et al. 2007).

1.2 Mathematical and computational aspects of the problem

With $V(T)$ we denote the set of all nodes, and $L(T)$ is the set of all leaves of a tree T. We assume that a gene tree G and a species tree S are given, as well as a mapping $\phi : L(G) \to L(S)$ that places extant genes into extant species.

The problem is to find a mapping $\rho : V(G) \to V(T)$ that optimizes some objective function. How to determine ρ depends on a model that describes a problem of reconciliation. The model includes the set of allowed evolutionary events (speciation is usually always included) and the objective function, which is usually the likelihood of a reconciliation (maximization problem) or the weight of a reconciliation (minimization problem). The weight of a reconciliation, which is the sum of costs of all evolutionary events in a reconciliation, is a sort of measure of dissimilarity between G and S.

In this paper, the objective function is the weight of a reconciliation. Conversions are modeled as a pair of duplication and loss. Since we are pairing gene losses with gene duplications, there is a need to introduce *lost subtrees*, *i.e.* subtrees of the gene tree that were not given in the input. This means that, in order to obtain an optimal solution, we need to extend given gene tree G, and this extension we denote by G'. Because of pairing losses with duplications, we obtain that disjoint subtrees of G are not independent anymore. The loss of independence and the need to extend the given gene tree are things that make the problem harder than the usual duplication/loss reconciliation.

1.3 A review of some previous results

The first model of reconciliation to mention is the one with duplications, speciations and losses. A natural way to form a reconciliation, in this model, is to position every node from the gene tree as low as possible inside the species tree. This type of reconciliation is called the *Last Common Ancestor* (LCA). LCA minimizes the number of duplications and losses (Górecki and Tiuryn 2006), the number of duplications (Górecki and Tiuryn 2006), and the number of losses (Chauve and El-Mabrouk 2009; Chauve et al. 2008). LCA is the only reconciliation that minimizes duplications and losses (Górecki and Tiuryn 2006). These reconciliations can be found in linear time. There is a polynomial algorithm in Vernot et al. (2008) that finds the minimum number of duplications even when S is polytomous. The problem of reconciliation between a polytomous gene tree and a binary species tree minimizing the number of mutations (duplications + losses) is polynomial (Chang and Eulenstein 2006;

Lafond et al. 2012). In Zheng and Zhang (2017), $O(|G| + |S|)$ algorithms for reconciling a nonbinary gene tree and a binary species tree in the duplication, loss, mutation, and deep coalescence models are given.

A biologically important and mathematically much studied evolutionary event is *gene transfer*. Models that include duplications, losses, and transfer are called *DTL models*. When the transfers are included, then time constraints are introduced, because direct gene transfer can happen only between species that exist in the same moment. There are two ways of considering time constraints in reconciliations. One is to use an undated species tree but imposing a consistency between found transfers. This variant has been proved to be NP-hard in Tofigh et al. (2011) (while without time consistency it is solvable in time $O(m^2n)$, where m is the number of extant species and n is the number of extant genes). Another is to use a fully dated species tree as an input, that is, there is a total order on the internal nodes. In that case a reconciliation algorithm with duplications, transfers and losses is given in Doyon et al. (2010) with time complexity $\Theta(m^2n)$. In Chan et al. (2015) the space of all reconciliations is explored and formula for its size is given. Discrete and continuous cases for DTL model are equivalent (Ranwez et al. 2016). In Chan et al. (2017), duplications, transfers, losses, and incomplete lineage sorting are included in the model and the FPT (fixed-parameter-tractable) algorithm for the most parsimonious reconciliation is given. If a gene that is transfered replaces another gene, then we have *transfer with replacement*, which is to transfer what conversion is to duplication (see Hasić and Tannier (2017) for NP-hardness proof, and FPT algorithm) For a more detailed review on reconciliations see Szöllősi et al. (2015), Nakhleh (2013), and Doyon et al. (2011).

1.4 The contribution of this paper

Gene conversion can be modeled in the gene tree/species tree reconciliation framework. It consists in coupling a duplication (the donor sequence) and a loss (the receiver sequence). It is usually not included in reconciliation models because the usual algorithmic toolbox of gene tree/species tree reconciliation, based on dynamic programming assuming a statistical independence between lineages, does not allow to couple events from different lineages.

Our contribution is to explore the algorithmic possibilities of introducing conversion in reconciliations. We formally define a reconciliation with duplications, losses and conversions. We define the algorithmic problem of computing, given a gene tree and a species tree, a reconciliation minimizing a linear combination of the number of events of each type. We fully solve the problem in the particular case when all events are equally weighted. More precisely, we construct an algorithm which gives, in polynomial running time, an optimal solution, and we prove that any optimal solution can be output by the algorithm with a non null probability. The algorithm can be used as a polynomial delay enumeration of the whole space of solutions.

The space of solutions is non trivial. In contrast with the duplication and loss only reconciliations, solutions are not unique, they are not all given by the standard Last Common Ancestor (LCA) technique. Moreover, easy examples show that the LCA technique does not give the optimal solution if events are weighted differently. This opens a wide range of new open algorithmic problems related to gene tree/species tree reconciliations.

The paper is organized as follows. Section 2 introduces a gene tree/species tree reconciliation including gene conversion events, and states the relations with the classical duplication loss reconciliation. Section 3 is devoted to the presentation of an algorithm to find one optimal solution, which is called an LCA completion. In Section 4, we give an algorithm to find all optimal solutions, by the definition of a class of optimal solutions called zero-flow, containing but not limited to LCA completions. We prove that an algorithm finding all zero-flow reconciliations is sufficient to access the whole solution space, and we write such an algorithm. In Section 5 we complete the proof that the presented algorithm always gives an optimal solution, and that every optimal solution can be output with a non null probability.

2 Reconciliations with Duplication, Loss, Conversion

In this section we define the mathematical problem modeling the presence of gene conversion in gene tree species tree reconciliations. We start with the definition of the standard duplication and loss model, and then add the possibility of conversions.

2.1 Duplication-Loss reconciliations

Let us begin with some generalities about phylogenetic trees. All phylogenetic trees are binary rooted trees where the root node has degree 1, and its incident edge is called the *root edge*. The root edge of T is denoted by $root_E(T)$, and the root node by $root(T)$. If x is a node in a tree, then $L(x)$ denotes the set of leaves of the maximal subtree rooted at x. If $x \in V(T) \backslash L(T)$ then x_r, x_l denote the two children of x. Similarly, we can define the children e_r, e_l of an edge e. If x is a leaf or an edge incident to a leaf, then their children are NULL and $f(NULL) = 0$ for any function/procedure which returns some value. If x is a node/edge in a rooted tree T, then $p_T(x) = p(x)$ denotes its parent. Let $e = (x, p(x))$ be an edge, then $T(e)$ denotes the maximal rooted subtree with root edge e. If x is on the path from y to $root(T)$ then we say that x is an *ancestor* of y, or that y is a *descendant* of x, and we write $y \leq_T x$ or $y \leq x$, defining a partial order on the nodes. If x is neither ancestor nor descendant of y, we say that x and y are *incomparable*. Let x and y be comparable nodes in a rooted tree T, then with $d_T(x, y)$ or $d(x, y)$ we denote the distance, *i.e.* the number of edges in the path between x and y. For a partially ordered set A, we use *minimal* to denote an element m such that $x \leq m \implies x = m$,

$\forall x \in A$. We use this terminology for the partial order defined by rooted trees. For example, if V' is a subset of nodes of a tree, their *Last Common Ancestor* (LCA) is the minimal node which is an ancestor of all nodes in V'. We also use it for partial orders defined by inclusion on sets or by subtrees in trees. In particular we can use it for the partial order defined by the *extension* relation.

Definition 1 (Extension) A tree G' is said to be an *extension* of a gene tree G if G can be obtained from G' by pruning some subtrees and suppressing nodes of degree 2.

We define the gene tree species tree duplication loss (DL) reconciliation. We suppose we have two trees G and S, respectively called the *gene tree* and the *species tree*. Nodes of G (S) are called *genes* (*species*). A mapping $\phi : L(G) \to L(S)$ indicates the species in which genes are found in the data. Without loss of generality we suppose that ϕ verifies that the last common ancestor of all the leaves of S that are in the image of ϕ is the node adjacent to the root node (recall the root node has degree 1). The reconciliation is based on a function ρ, which is an extension of ϕ to all genes and species, including internal nodes.

Definition 2 (Consistency) A function $\rho : V(G') \to V(S)$ on the nodes of a tree G' is said to be *consistent* with a species tree S if $\rho(root(G')) = root(S)$ and for every $x \in V(G') \backslash L(G')$ one of the conditions holds (D) $\rho(x) = \rho(x_l) = \rho(x_r)$ or (S) $\rho(x)_l = \rho(x_l)$ and $\rho(x)_r = \rho(x_r)$. We also say that G' is ρ-consistent with S.

Obviously, both conditions (D) and (S) cannot hold for a single node.

Definition 3 (DL reconciliation) Let G and S be a gene and a species trees and $\phi : L(G) \to L(S)$. A *DL reconciliation* between G and S is a 5-tuple (G, G', S, ϕ, ρ) such that G' is an extension of G, G' is ρ-consistent with S, and $\rho / L(G) = \phi$.

Note that we allow some extant species not to have genes. The definition is equivalent to the standard ones Arvestad et al. (2004); Górecki and Tiuryn (2006); Chauve and El-Mabrouk (2009), although they can present some variations between them. For example we do no impose that losses are represented by subtrees extended to the leaves of S (which is the case for example in Chauve and El-Mabrouk (2009)), because of the particular use we make of loss subtrees in the sequel. An example of DL reconciliation is given in Figure 1 (a).

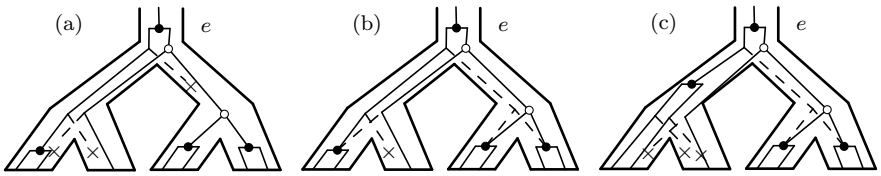

Fig. 1 Examples of reconciliations. The gene tree is depicted inside the species tree to signify the mapping ρ. Duplication nodes are black circles, speciation nodes are white circles, losses are leaves with crosses and conversions are duplication nodes which are also a leaf of the lost subtrees, which are dashed. (a) An LCA reconciliation. Total cost: $3l + 4d = 7$. (b) An LCA completion, obtained from LCA by extending losses and assigning them to duplications. Total cost: $l + d + 3c = 5$. (c) A non-optimal reconciliation. Total cost: $3l + 2d + 2c = 7$

Definition 4 (Duplication) Let $\mathfrak{R} = (G, G', S, \phi, \rho)$ be a DL reconciliation and $x \in V(G')\backslash L(G')$ satisfies condition (D). Then x is called a *duplication*. The set of all duplications is denoted by $\Delta = \Delta(\mathfrak{R})$.

Definition 5 (Speciation) Let $\mathfrak{R} = (G, G', S, \phi, \rho)$ be a DL reconciliation and $x \in V(G')\backslash L(G')$ satisfies condition (S). Then x is called a *speciation*. The set of all speciations is denoted by $\Sigma = \Sigma(\mathfrak{R})$.

Definition 6 (Loss) Let $\mathfrak{R} = (G, G', S, \phi, \rho)$ be a DL reconciliation and $x \in L(G')\backslash L(G)$. Then x is called a *loss*. The set of all losses is denoted by $\Lambda = \Lambda(\mathfrak{R})$.

We say that a duplication, loss or speciation x is *assigned* to s if $\rho(x) = s$. Let $\mathfrak{L}(s, \mathfrak{R}) = \mathfrak{L}(s) = |\rho^{-1}(s) \cap \Lambda(\mathfrak{R})|$ and $\mathfrak{D}(s, \mathfrak{R}) = \mathfrak{D}(s) = |\rho^{-1}(s) \cap \Delta(\mathfrak{R})|$ be the number of losses and the number of duplications assigned to $s \in V(S)$ in the reconciliation $\mathfrak{R}$. If $e = (s, p(s)) \in E(S)$, then $\mathfrak{L}(e, \mathfrak{R}) = \mathfrak{L}(e) = \mathfrak{L}(s, \mathfrak{R})$ and $\mathfrak{D}(e, \mathfrak{R}) = \mathfrak{D}(e) = \mathfrak{D}(s, \mathfrak{R})$.

The next definition extends the notion of loss.

Definition 7 (Lost subtree) Let $\mathfrak{R} = (G, G', S, \phi, \rho)$ be a DL reconciliation. A maximal subtree T of G' such that $V(T) \cap V(G) = \emptyset$ is called a *lost subtree*.

The next lemma introduces the standard Last Common Ancestor reconciliation, and its proof can be found in Chauve and El-Mabrouk (2009) or Chauve et al. (2008).

Lemma 1 *Let G and S be a gene and a species tree, and $\phi : L(G) \to L(S)$. There exists a DL reconciliation $\mathfrak{R} = (G, G', S, \phi, \rho)$ such that $\rho(x)$ is the root of the minimal subtree of S containing $L(\phi(x))$, $\forall x \in V(G)$.*

Definition 8 (LCA reconciliation) The DL reconciliation from Lemma 1 that minimizes $|\Lambda(\mathfrak{R})|$ is called the *Last Common Ancestor (LCA)* reconciliation and is noted $\mathfrak{R}_{lca} = (G, G'_{lca}, S, \phi, \rho_{lca})$.

Note that the LCA reconciliation is the unique reconciliation minimizing the number of duplications, or the number of losses, or any linear combination

of these two numbers Chauve and El-Mabrouk (2009). In Section 3 we will construct equivalents of the LCA reconciliation including conversions, called LCA completions, which will have the property of minimizing the sum of the number of duplications, losses and conversions. However in contrast it is not unique, it does not contain all optimal solutions (as we show it in Section 4) and does not optimize over any linear combinations of these numbers (see the conclusion for such an example).

2.2 Duplication-Loss-Conversion reconciliations

In the next definition we introduce an additional event, called *gene conversion*, which is a function δ pairing some losses and duplications. This models the replacement of a gene by a copy of another one from the same family.

Definition 9 (Conversion) Let (G, G', S, ϕ, ρ) be a DL reconciliation. Let $\delta : \Delta \to \Lambda$ be an injective partial function such that $\rho(x) = \rho(\delta(x))$ for all $x \in \delta^{-1}(\Lambda)$. If $x \in \delta^{-1}(\Lambda)$, then x is called a *conversion*, and $\delta(x)$ is its associate loss. The set of all conversions is denoted by Δ' and the set of associate losses by Λ'. The 6-tuple $(G, G', S, \phi, \rho, \delta)$ is called a *DLC reconciliation*.

We see that every DL reconciliation is also a DLC reconciliation with $\Delta' = \emptyset$. From now on, *reconciliation* stands for *DLC reconciliation*. Examples of DLC reconciliations are drawn on Figure 1.

The following properties are equivalents of standard properties of DL reconciliations Chauve et al. (2008), which have to be checked in the DLC case.

Lemma 2 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, $x, y \in V(G')$ and $x < y$. Then $\rho(x) \leq \rho(y)$.*

Proof If $x < y$, then we have $x_1, ..., x_k \in V(G')$ so that $x = x_0 < x_1 < x_2 < ... < x_k < x_{k+1} = y$, and x_i is a child of x_{i+1}. From Definition 2, we have that (D) or (S) holds, *i.e.* $\rho(x) \leq \rho(p(x))$, therefore $\rho(x) \leq \rho(x_1) \leq \rho(x_2) \leq ... \leq \rho(x_k) \leq \rho(y)$. □

Lemma 3 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, $s \in V(S) \backslash L(S)$, $x \in V(G') \backslash L(G')$ such that $\rho(x) = s$. Then $x \in \Sigma(\mathfrak{R})$ if and only if x is a minimal element of $\rho^{-1}(s)$.*

Proof Let x be a minimal element of $\rho^{-1}(s)$. Assume the opposite, then $x \in \Delta(\mathfrak{R})$. Let x_l, x_r be the children of x in G', hence $x_l < x$, $x_r < x$ and $\rho(x) = \rho(x_l) = \rho(x_l) = s$, which contradicts the minimality of x.

Let $x \in \Sigma(\mathfrak{R})$. Assume the opposite, that x is not a minimal element of $\rho^{-1}(s)$. Let $x' < x$, $\rho(x') = s$. Then $x' \leq x_l$ or $x' \leq x_r$. Let $x' \leq x_l$, hence $\rho(x') \leq \rho(x_l) \leq \rho(x)$. Therefore $\rho(x) = \rho(x_l)$, which contradicts $x \in \Sigma(\mathfrak{R})$. □

Next lemma states that we cannot have two comparable speciations assigned to the same node from $V(S)$.

Lemma 4 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation and $x, y \in V(G')$, $x < y$, $\rho(x) = \rho(y)$. Then $y \in \Delta(\mathfrak{R})$.*

Proof Follows directly from Lemma 3. □

Lemma 5 *Let $\mathfrak{R}_1 = (G, G'_1, S, \phi, \rho_1, \delta_1)$ and $\mathfrak{R}_2 = (G, G'_2, S, \phi, \rho_2, \delta_2)$ be reconciliations, and $x \in V(G)$. Then $\rho_1(x)$ and $\rho_2(x)$ are comparable.*

Proof Assume the opposite, *i.e.* $\rho_1(x)$ and $\rho_2(x)$ are incomparable. Then $T(\rho_1(x))$ and $T(\rho_2(x))$ are disjoint, and in particular $L(\rho_1(x)) \cap L(\rho_2(x)) = \emptyset$. Let $l \in L(x)$. Then $l \leq x$, therefore $\phi(l) = \rho_1(l) \leq \rho_1(x)$ and $\phi(l) = \rho_2(l) \leq \rho_2(x)$, hence $\phi(l) \in L(\rho_1(x))$ and $\phi(l) \in L(\rho_2(x))$, a contradiction. □

Definition 10 (The cost/weight of a reconciliation) Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, $d, l, c \in \mathbb{N}$ weights associated with duplication, loss and conversion. The cost (or weight) of $\mathfrak{R}$ is given by

$$\omega(\mathfrak{R}) = l \cdot |\Lambda \backslash \Lambda'| + d \cdot |\Delta \backslash \Delta'| + c \cdot |\Delta'|.$$

Examples of computations of this cost are given on Figure 1. As we can see, losses from Λ' are not counted as losses in the formula, so we call them *free losses*. If a lost subtree has only free losses then it is called a *free subtree*.

Definition 11 (Minimum/optimal reconciliation) Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation that minimizes $\omega(\mathfrak{R})$, for given G, S, and ϕ. Then it is called *minimum* (or *optimal*) reconciliation.

In the sequel we give an algorithm that is able to output all optimal reconciliations for $d = l = c$, so unless specified, we assume from now, and without loss of generality, that they are all equal to 1. We come back to the general case in the conclusion, stating open problems.

2.3 Completions and minimizations of reconciliations

Recall that any DL reconciliation is a DLC reconciliation by definition. However an optimal DL reconciliation is not an optimal DLC reconciliation. Completions and minimizations are operations on reconciliations that help constructing nonetheless a relation between optimal DL and DLC reconciliations.

Definition 12 (Loss extension) Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation. The reconciliation $\mathfrak{R}' = (G, G'', S, \phi, \rho', \delta')$ is said to be obtained from $\mathfrak{R}$ by *loss extension* if G'' is an extension of G', $\rho = \rho'/V(G')$, $\mathfrak{R}$ and $\mathfrak{R}'$ have the same number of lost subtrees.

Definition 13 (Completion) Let $\mathfrak{R}$ be a reconciliation, and $\mathfrak{R}'$ is a reconciliation with minimum weight among all reconciliations obtained from $\mathfrak{R}$ by extending some losses. Then $\mathfrak{R}'$ is called a *completion* of $\mathfrak{R}$.

It is obvious, by definition, that an optimal reconciliation is a completion, *i.e* a completion of a reconciliation $\mathfrak{R}$ has always a lower or equal cost than $\mathfrak{R}$ itself. The set of all completions of $\mathfrak{R}$ is denoted by $c(\mathfrak{R})$. When useful, $c(\mathfrak{R})$ can also be used to denote one arbitrary completion if it is clear that any completion works. For example the cost of a completion can be written $\omega(c(\mathfrak{R}))$ since by definition they all have the same cost.

The converse of a completion is a *minimization*. It is based on the following definition and lemma.

Definition 14 (Minimal reconciliation) A reconciliation $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ is called *minimal* if there does not exist G'' such that G' is a proper extension of G'', G'' is an extension of G, and G'' is ρ''-consistent, where $\rho'' = \rho/V(G'')$.

An example of minimal reconciliation is the LCA reconciliation. The next lemma shows how to construct a minimal reconciliation from any reconciliation.

Lemma 6 *Let G and S be a gene and a species tree, and $\rho' : V(G) \to V(S)$ such that*

- $\rho'(x) = \phi(x)$, $\forall x \in L(G)$,
- $x < y \implies \rho'(y) \leq \rho'(x)$,
- $\rho'(x)$ *belongs to the path from* $\rho_{lca}(x)$ *to root*(S).

Then there exists a unique (up to δ) minimal reconciliation $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ such that $\rho/V(G) = \rho'$.

Proof Assume that there exists a reconciliation $\mathfrak{R}_1 = (G, G'_1, S, \phi, \rho_1, \delta_1)$ such that $\rho_1/V(G) = \rho'$. Let $x \in V(G)$ with children x_l, x_r (in G). In the next three cases we show how to construct G'.

Case 1, $\rho_1(x_l) = \rho_1(x)$ and $\rho_1(x_r) < \rho_1(x)$. In that case $x \notin \Sigma(\mathfrak{R}_1)$, hence $x \in \Delta(\mathfrak{R}_1)$. Therefore $\exists x' \in V(G'_1)$ such that x' is the right child of x and $\rho_1(x') = \rho_1(x)$. Since $x_r < x' < x$, x' is not a leaf and it has the left subtree. Therefore $\exists x'' \in V(G'_1)$ such that x'' is a descendant of x' and $\rho_1(x'') = \rho_1(x')_l$. We have a similar situation for the case $\rho_1(x_r) = \rho_1(x)$ and $\rho_1(x_l) < \rho_1(x)$.

Case 2, $e = (s, p(s)) \in E(S)$, $s \in V(S)$ and $\rho_1(p_G(x)) > s$ and $\rho_1(x) < s$. We will prove that there exists a node $x_1 \in V(G'_1)$ such that $\rho(x_1) = s$ and $x < x_1 < p_G(x)$. Let x' be a minimal node of $V(G'_1)$ such that $x' \leq p_G(x)$ and $\rho(x') > s$. From Lemma 3, we have $x' \in \Sigma(\mathfrak{R}_1)$. Therefore it has children x'_l,x'_r (in G'_1) such that $\rho_1(x'_l) < \rho_1(x')$ and $\rho_1(x'_r) < \rho_1(x')$. From the properties of x', we get that one of the children maps to s. Let $\rho(x'_r) = s$, and we need to insert an additional child for x'_r, since x'_r cannot be a leaf.

Case 3, $\rho_1(x_l) \leq \rho_1(x)_l$ and $\rho_1(x_r) \leq \rho_1(x)_l$. Let x' be a child of x in G'_1. Therefore x' is comparable to x_l or x_r, and $\rho_1(x')$ is comparable to $\rho_1(x_l)$ or $\rho_1(x_r)$, hence $\rho_1(x')$ is comparable to $\rho_1(x)_l$. Next, $\rho_1(x')$ is incomparable to $\rho_1(x)_r$, hence $x \notin \Sigma(\mathfrak{R}_1)$ and $x \in \Delta(\mathfrak{R}_1)$. If x'_l, x'_r are the children of x in G'_1, then $\rho_1(x'_l) = \rho_1(x'_r) = \rho_1(x)$. This means that we need to insert x'_l, x'_r and additional children for x'_l, x'_r.

Insertions, described in the previous three cases, are for any reconciliation $\mathfrak{R}_1$. Let us prove that they are enough to form a reconciliation. From this will follow minimization and uniqueness.

Let us form G' and ρ in a way described in the previous three cases. We need to prove that G' is ρ-consistent. Let $x \in V(G')\backslash L(G')$ and x_l, x_r are the children of x in G'. We will prove that x satisfies condition (D) or (S) from Definition 2. If $\rho(x) = \rho(x_l) = \rho(x_r)$, then condition (D) is satisfied. Now assume that condition (D) is not satisfied, i.e. $\rho(x) \neq \rho(x_l)$ or $\rho(x) \neq \rho(x_r)$. Take $\rho(x_r) < \rho(x)$. From the Case 2, we get $\rho(x_r) = \rho(x)_r$. We are left to prove $\rho(x_l) = \rho(x)_l$. Assume the opposite, let $\rho(x_l) = \rho(x)_r$ or $\rho(x_l) = \rho(x)$. From Case 3 and the definition of duplication, we get that x is a duplication, this contradicts our assumption that $\rho(x_l) \neq \rho(x)_l$. □

The unique minimal reconciliation obtained from a reconciliation is called its *minimization*. In the next section we prove that minimization and completion are complementary operations, that is, an optimal reconciliation is always the completion of its minimization. This will lead to the important result that completions of the LCA reconciliations are optimal.

3 A family of optimal reconciliations: LCA reconciliations

In this section we provide a polynomial running time algorithm which finds an LCA completion, and prove that it is an optimal reconciliation. We present a more general algorithm, which finds a completion of any reconciliation. To this aim we present the important notion of *flow*, constantly used all along the paper. This settles the complexity of the defined problem when the weights d, l, c are all equal. However the algorithm described here does not find all LCA completions, and moreover the space of optimal reconciliations is not limited to LCA completions. Finding all solutions will be the subject of next section. Here we begin by stating general properties of reconciliations and optimal reconciliations, showing that they all share some important properties with LCA reconciliations.

3.1 Similarities of any reconciliation with the LCA reconciliation

Some properties of the LCA reconciliation are shared by all reconciliations.

Lemma 7 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, and $x \in V(G)$. Then $\rho(x)$ is not lower than $\rho_{lca}(x)$.*

Proof Follows directly from the definition of Last Common Ancestor. □

Lemma 8 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, and $x \in V(G)\backslash L(G)$. Then $\rho(x)$ is in the path in S from $\rho_{lca}(x)$ to root(S).*

Proof Follows directly from Lemmas 5 and 7. □

The next lemma states that if a node is a speciation in an arbitrary reconciliation then it is also a speciation in the LCA.

Lemma 9 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, and $x \in V(G)$. If $x \in \Sigma(\mathfrak{R})$, then $x \in \Sigma(\mathfrak{R}_{lca})$, and $\rho(x) = \rho_{lca}(x)$.*

Proof Let $x \in V(G) \cap \Sigma(\mathfrak{R})$. Let x''_l, x''_r be the children of x in $\mathfrak{R}$, x'_l, x'_r the children of x in $\mathfrak{R}_{lca}$, and x_l, x_r be the children of x in G. We have $\rho(x)_l = \rho(x''_l)$ and $\rho(x)_r = \rho(x''_r)$. From Lemma 8 we have $\rho_{lca}(x) \leq \rho(x)$.

Assume that $\rho_{lca}(x) < \rho(x)$. Hence $\rho(x)_l$ or $\rho(x)_r$ is incomparable to $\rho_{lca}(x)$. Assume that $\rho(x)_r = \rho(x''_r)$ is incomparable to $\rho_{lca}(x)$. Next, $x_r \leq x'_r < x$, $x_r \leq x''_r < x$, hence $\rho_{lca}(x_r) \leq \rho_{lca}(x'_r) \leq \rho_{lca}(x)$ and $\rho(x_r) \leq \rho(x''_r) \leq \rho(x)$. Therefore, $\rho(x_r)$ is incomparable to $\rho_{lca}(x)$, hence incomparable to $\rho_{lca}(x_r)$, which contradicts Lemma 5. Therefore $\rho_{lca}(x) = \rho(x)$.

Let us prove that $x \in \Sigma(\mathfrak{R}_{lca})$. Assume the opposite, $x \in \Delta(\mathfrak{R}_{lca})$. Thus $\rho_{lca}(x) = \rho_{lca}(x'_l) = \rho_{lca}(x'_r)$, and from LCA reconciliation, we have $\rho_{lca}(x) = \rho_{lca}(x_r)$ or $\rho_{lca}(x) = \rho_{lca}(x_l)$. Next, $\rho_{lca}(x_r) = \rho_{lca}(x) = \rho(x) > \rho(x_r)$ or $\rho_{lca}(x_l) = \rho_{lca}(x) = \rho(x) > \rho(x_l)$, which contradicts Lemma 7. □

Thanks to these properties we can define a distance from an arbitrary reconciliation to the LCA reconciliation. This distance will be used in the proofs of several properties, stating that there is always a way to lower the distance to the LCA without increasing the cost of a reconciliation.

Definition 15 Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be any reconciliation. Let $dist_{lca}(\mathfrak{R}) = \sum_{d \in V(G)} d_S(\rho(d), \rho_{lca}(d))$ be the *distance* from $\mathfrak{R}$ to the LCA reconciliation $\mathfrak{R}_{lca} = (G, G'_{lca}, S, \phi, \rho_{lca})$.

Lemma 10 *If for a reconciliation $\mathfrak{R}$ $dist_{lca}(\mathfrak{R}) > 0$, there exists a reconciliation $\mathfrak{R}'$ such that $dist_{lca}(\mathfrak{R}') < dist_{lca}(\mathfrak{R})$ and $\omega(\mathfrak{R}') \leq \omega(\mathfrak{R})$.*

Proof Take any $d' \in V(G)$ so that $\rho(d') > \rho_{lca}(d')$ and let d be a minimal element of $V(G)$ such that $\rho(d) = \rho(d')$ and $d \leq d'$. Since $d \leq d'$, we have $\rho_{lca}(d) \leq \rho_{lca}(d') < \rho(d') = \rho(d)$, therefore $\rho_{lca}(d) < \rho(d)$. By Lemma 9 $d \notin \Sigma(\mathfrak{R})$, so $d \in \Delta(\mathfrak{R})$.

Let d^1_l, d^1_r be the children of d in $\mathfrak{R}$. Since $d \in \Delta(\mathfrak{R})$, we have $\rho(d) = \rho(d^1_l) = \rho(d^1_r)$, and because of the minimality of d, we get $d^1_l, d^1_r \notin V(G)$. Similarly, all descendants of d in G', with the same ρ-value, are not in $V(G)$.

Let $d_1, ...d_k$ be these descendants and let $T_1, ..., T_k$ be lost subtrees such that $root(T_i) = d_i$, $(i = 1, \ldots, k)$. Prune all these subtrees, contract nodes of degree two (*i.e.* $d_1, \ldots, d_k$), and let G'' denotes the obtained extension of gene tree G. Let d^2_l, d^2_r be the children of d in G''.

If $\rho(d^2_l) \neq \rho(d^2_r)$, then G'' generates a new reconciliation $\mathfrak{R}'$, where d is a speciation, and $\rho'(d) = \rho(d)$. By Lemma 9, $\rho'(d) = \rho_{lca}(d)$, which contradicts $\rho(d) > \rho_{lca}(d)$.

Let $\rho(d^2_l) = \rho(d^2_r)$. Since $\rho(d^2_l) < \rho(d)$, we don't have consistency. Put $\rho'(d) = \rho(d^2_l)$ and insert x_1 into G'' so that $d < x_1 < p_{G'}(d)$, $\rho'(x_1) = \rho(d)$, and x_1 is the root of some of the pruned subtrees T_i (reinsert T_i). In this way we

get a new reconciliation $\mathfrak{R}''$, and d is a duplication in $\mathfrak{R}''$. Also $\omega(\mathfrak{R}'') \leq \omega(\mathfrak{R})$ and $dist_{lca}(\mathfrak{R}'') < dist_{lca}(\mathfrak{R})$.

If $d \in \Delta'(\mathfrak{R})$ and corresponding loss is l, then extend l so that one loss extensions follows d and the other can be some of the pruned subtrees T_i (reinsert T_i). □

The next lemma states that with LCA we get the smallest set of duplications.

Lemma 11 *Let $\mathfrak{R}_{lca}$ be the LCA reconciliation and $\mathfrak{R}$ be any reconciliation. Then $\Delta(\mathfrak{R}_{lca}) \subseteq \Delta(\mathfrak{R}) \cap V(G)$.*

Proof Let $x \in \Delta(\mathfrak{R}_{lca})$, then $x \notin \Sigma(\mathfrak{R}_{lca})$ and $x \in V(G)$. Assume the opposite, that $x \notin \Delta(\mathfrak{R}) \cap V(G)$, then $x \in \Sigma(\mathfrak{R})$. From Lemma 9 we get $x \in \Sigma(\mathfrak{R}_{lca})$, a contradiction. Therefore $x \in \Delta(\mathfrak{R}) \cap V(G)$. □

3.2 Properties of optimal reconciliations

We examine some properties of optimal reconciliations. Note that optimal reconciliations are not necessarily minimal, but we will state the relation between the two classes (see Lemma 15). The next lemma states that optimal reconciliations never contain duplication nodes in lost subtrees.

Lemma 12 *Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be an optimal reconciliation. Then $\Delta(\mathfrak{R}) \subseteq V(G)$,* i.e. *all duplications nodes are in G.*

Proof Assume the opposite. Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a reconciliation, and x is a minimal node of $\Delta(\mathfrak{R}) \backslash V(G)$. Let us prove that $\mathfrak{R}$ cannot be optimal. Let $x_l, x_r \in V(G')$ be the children of x. Since x is a duplication, we have $\rho(x) = \rho(x_l) = \rho(x_r)$. Observe two cases.

Case 1, $x_l, x_r \notin V(G)$

Case 1.1, x is a conversion, and l is the corresponding loss. Remove l and x, connect x_l with $p_{G'}(l)$, and x_r with $p_{G'}(x)$. In this way we get G''. Let $\rho' = \rho/G''$, and $\delta' = \delta/G''$. We get a reconciliation $\mathfrak{R}' = (G, G'', S, \phi, \rho', \delta')$ which has one duplication less, *i.e.* $\omega(\mathfrak{R}') = \omega(\mathfrak{R}) - 1$. Hence $\mathfrak{R}$ cannot be an optimal reconciliation.

Case 1.2, x is not a conversion. Remove $T(x_l)$ and x, then connect x_r with $p_{G'}(x)$. By a similar argument, we get a reconciliation with one duplication and all non-free losses from $T(x_l)$ less, *i.e.* we get a reconciliation with a strictly lower cost. Indeed, since x is a minimal duplication, subtree $T(x_l)$ cannot have any duplications, *i.e.* by removing $T(x_l)$ we cannot get to the situation where some free loss becomes non-free.

Case 2, $x_l \in V(G), x_r \notin V(G)$. Similarly, if x is not a conversion, remove $T(x_r)$ suppress x, and we get a reconciliation with strictly less cost. If x is a conversion and l is associate loss, then remove l, suppress x and connect x_r and $p_{G'}(l)$. We again obtain a cheaper reconciliation. □

The next lemma is a version of Lemma 10 for an optimal reconciliation.

Lemma 13 *Let $\mathfrak{R}_{lca}$ be the LCA reconciliation, and let $\mathfrak{R}$ be an optimal reconciliation. If $dist_{lca}(\mathfrak{R}) > 0$, there exists an optimal reconciliation $\mathfrak{R}'$ such that $\Delta(\mathfrak{R}') = \Delta(\mathfrak{R})$ and $dist_{lca}(\mathfrak{R}') < dist_{lca}(\mathfrak{R})$.*

Proof Follows directly from the proof of Lemma 10. We constructed $\mathfrak{R}'$ by pruning some of the lost subtrees and lowering duplication, which remained a duplication in $\mathfrak{R}'$. By Lemma 12 lost subtrees in optimal reconciliation cannot contain duplications, hence the set of duplications remained unchanged, *i.e.* $\Delta(\mathfrak{R}') = \Delta(\mathfrak{R})$. □

Next theorem states that all optimal reconciliations have the same sets of duplications.

Theorem 1 *Let $\mathfrak{R}_{lca} = (G, G'_{lca}, S, \phi, \rho_{lca})$ be the LCA reconciliation and $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be an optimal reconciliation. Then $\Delta(\mathfrak{R}_{lca}) = \Delta(\mathfrak{R})$.*

Proof Assume the opposite, there exist G, S and $\mathfrak{R}$ such that $\mathfrak{R}$ is an optimal reconciliation and $\Delta(\mathfrak{R}_{lca}) \neq \Delta(\mathfrak{R})$. By Lemma 11 and Lemma 12 we get $\Delta(\mathfrak{R}_{lca}) \subset \Delta(\mathfrak{R}) \cap V(G) = \Delta(\mathfrak{R})$. Assume that $\mathfrak{R}$ is an optimal reconciliation with $\Delta(\mathfrak{R}_{lca}) \subset \Delta(\mathfrak{R})$ and minimum $dist_{lca}(\mathfrak{R})$. We have $dist_{lca}(\mathfrak{R}) = 0$, otherwise we could get an optimal reconciliation $\mathfrak{R}'$ with $dist_{lca}(\mathfrak{R}') < dist_{lca}(\mathfrak{R})$ and $\Delta(\mathfrak{R}') = \Delta(\mathfrak{R})$ (Lemma 13). From $dist_{lca}(\mathfrak{R}) = 0$, we obtain $\rho(x) = \rho_{lca}(x)$, $\forall x \in V(G)$.

Let $x' \in \Delta(\mathfrak{R}) \backslash \Delta(\mathfrak{R}_{lca})$. By Lemma 12, we have $x' \in V(G)$. From $x' \notin \Delta(\mathfrak{R}_{lca})$ we get that $x' \in \Sigma(\mathfrak{R}_{lca})$. We will continue in a similar way as in the proof of Lemma 10. Let $x_1, ..., x_k$ be descendants of x' in $V(G')$ with the same ρ-value as x'.

Assume $x_1 \in V(G)$. Since $\rho(x) = \rho_{lca}(x)$, $\forall x \in V(G)$ and $\rho(x_1) = \rho(x')$ we get $\rho_{lca}(x_1) = \rho_{lca}(x')$, hence (Lemma 4) $x' \in \Delta(\mathfrak{R}_{lca})$, a contradiction. Therefore $x_1 \notin V(G)$.

By a similar argument, $x_1, \ldots, x_k \notin V(G)$. Let T_i be the lost subtrees rooted at x_i $(i = 1, \ldots, k)$. By pruning T_i and suppressing x_i $(i = 1, \ldots, k)$ we get G'', and a new reconciliation where node x' is a speciation. Hence we get a reconciliation with strictly lower cost, which contradicts the optimality of $\mathfrak{R}$. □

Next lemma states that, in an optimal reconciliation, we cannot have two comparable nodes $x, y \in V(G') \backslash V(G)$ such that $\rho(x) = \rho(y)$.

Lemma 14 *Let $\mathfrak{R}$ be an optimal reconciliation and $x, y \in V(G')$ such that $\rho(x) = \rho(y)$ and $x < y$. Then $y \in V(G) \cap \Delta(R_{lca}) = \Delta(\mathfrak{R}_{lca}) = \Delta(\mathfrak{R})$.*

Proof From Lemma 4 we have $y \in \Delta(\mathfrak{R})$. From Theorem 1, we obtain $\Delta(\mathfrak{R}) = \Delta(\mathfrak{R}_{lca})$. From Lemma 12, we have $y \in V(G) \supseteq \Delta(\mathfrak{R}) = \Delta(\mathfrak{R}_{lca})$. Therefore $y \in V(G) \cap \Delta(\mathfrak{R}_{lca})$. □

Next lemma states the relation between minimal and optimal reconciliations.

Lemma 15 *Let $\mathfrak{R}$ be an optimal reconciliation. Then there exists $\mathfrak{R}'$, a minimal reconciliation such that $\mathfrak{R}$ is a completion of $\mathfrak{R}'$.*

Proof Let $\mathfrak{R}'$ be the reconciliation obtained from $\mathfrak{R}$ by deleting all lost subtrees except their root edges. So $\mathfrak{R}$ is a completion of $\mathfrak{R}'$. We prove that $\mathfrak{R}'$ is minimal. Suppose the opposite. There is $e' = (x', p_{G'}(x')) \in E(G')\backslash E(G)$ such that by removing e' and suppressing $p_{G'}(x')$ we obtain again a reconciliation, denoted by $\mathfrak{R}''$. From the proof of Lemma 6, Case 2, we have that $\forall s \in V(S)$ and $x, y \in V(G')$, such that $x < y$, $\rho(x) < s < \rho(y)$, $\exists z \in V(G')$ such that $\rho(z) = s$ and $x < z < y$. Let x_1 be another child of $p_{G'}(x')$. Since there is no lost subtrees with more than one edge, we have $x_1 \in V(G)$.

Let $s = \rho(p_{G'}(x'))$. Since $\mathfrak{R}''$ is a reconciliation, $\exists x'' \in V(G'')$ such that $s = \rho(x'')$ and x'' comparable to x_1. Take minimal x'' with these properties, then (Lemma 3) $x'' \in \Sigma(\mathfrak{R}'')$. After bringing back e, we get that $p_{G'}(x')$ or x'' becomes a duplication (Lemma 4). Hence $\Delta(\mathfrak{R}'') \subset \Delta(\mathfrak{R}') = \Delta(\mathfrak{R})$, which contradicts the optimality of $\mathfrak{R}$ (Lemma 11 and Theorem 1). □

3.3 LCA completions are optimal

Theorem 2 *A completion of the LCA reconciliation is an optimal reconciliation.*

Proof Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be an optimal reconciliation with $dist_{lca}(\mathfrak{R})$ minimum. We prove that this reconciliation is a completion of the LCA. Since all completions of the LCA have the same weight by definition, this proves that all completions of the LCA are optimal reconciliations.

From Lemma 13 we get $dist_{lca}(\mathfrak{R}) = 0$ and therefore $\rho(x) = \rho_{lca}(x)$, $\forall x \in V(G)$. From Theorem 1 and Lemma 12, we have $\Delta(\mathfrak{R}) = \Delta(\mathfrak{R}_{lca}) \subseteq V(G)$.

Let t be a root of some lost subtree of G'. Let us prove that $t \in V(G'_{lca})$, and vice versa, if $t \in V(G'_{lca})\backslash V(G)$, then t is a root of some lost subtree of G'. This correspondence has to be bijective.

Let us prove that we can establish a bijection
$f : V(G) \cup \{t \mid \text{t is a root of some lost subtree of } V(G')\} \to V(G'_{lca})\backslash \Lambda(\mathfrak{R}_{lca})$ such that $f(x) = x$, $\forall x \in V(G)$, $x < y \implies f(x) < f(y)$, $\rho(x) = \rho(f(x))$.

First, put $f(x) = x$, $\forall x \in V(G)$.

Let $t \in V(G')\backslash V(G)$ be a root of some lost subtree of G', $\rho(t) = s$, $x < t < p_G(x)$. From Lemmas 12 and 3, we have $t \in \Sigma(\mathfrak{R})$ and t is a minimal element of $\rho^{-1}(s)$. Hence, there is no other element $t' \in V(G')$ such that $\rho(t') = s$, $x < t' < p_G(x)$. Since $t \in \Sigma(\mathfrak{R})$, we have $\rho(x) < \rho(t) \leq \rho(p_G(x))$. In $\mathfrak{R}_{lca}$ we also have $x' \in V(G'_{lca})$, such that $\rho(x') = s$, and $x < x' < p_G(x)$. Next, put $f(t) = x'$.

Above correspondence is obviously an injection. Let us prove that it is a surjection. In a similar way, let $x' \in V(G'_{lca})\backslash \Lambda(\mathfrak{R}_{lca})$, $\rho_{lca}(x') = s'$. If $x' \in V(G)$, then $x' = f(x')$. Now, assume $x' \notin V(G)$. Again from Lemmas 12 and 3 we have that $x' \in \Sigma(\mathfrak{R}_{lca})$ and x' is a minimal element of $\rho_{lca}^{-1}(s')$. Let $x < x' < p_G(x)$, $x \in V(G)$. Similarly, we have $\rho_{lca}(x) < \rho_{lca}(p_G(x))$ and x' is

the only element from $V(G')\backslash V(G)$ assigned to s' comparable to x. In order for $\mathfrak{R}$ to be ρ-consistent, there is a root of the lost subtree of G' (say t) such that: $\rho(t) = s'$, and $x < t < p_G(x)$ and it is unique. So, $f(t) = x'$.

We proved the existence of the described correspondence, therefore every lost subtree of $\mathfrak{R}$ is obtained as a loss extension in $\mathfrak{R}_{lca}$. □

The LCA reconciliation is easy to find, it is a well known result that there is a linear time algorithm to compute it Chauve and El-Mabrouk (2009). What remains in order to derive an algorithm to find an optimal reconciliation is to find a completion. Next section presents a method to find a completion of an arbitrary reconciliation.

3.4 Finding a completion and the flow of losses

Finding a completion is a kind of flow problem. We have demands, which are losses, that we supply by duplications, *i.e.* we associate them to duplications to form conversions. The amount and distribution of duplications in the phylogenetic tree tells how many losses can be supplied. The number of losses that can be supplied tells the value of a completion. We compute this number recursively along the tree. In consequence we have to define restriction of reconciliations to subtrees, which are *multiple reconciliations*.

Definition 16 (Multiple reconciliation) Let $\mathfrak{R}_i = (G_i, G_i', S, \phi_i, \rho_i)$ be DL reconciliations of gene trees G_i with species tree S, $(i = 1, \ldots, k)$. Let $T_1, \ldots, T_t$ be trees, $\rho_j' : V(T_j) \to V(S)$ verifying that $\rho_j'(root(T_j)) = root(S)$ and T_j is ρ_j'-consistent, $(j = 1, \ldots, t)$. Let $\mathfrak{R}_j' = (T_j, S, \rho_j')$, $(j = 1, \ldots, t)$. Next, let $\delta : \bigcup \Delta(\mathfrak{R}_i) \cup \bigcup \Delta(\mathfrak{R}_j') \to \bigcup \Lambda(\mathfrak{R}_i) \cup \bigcup \Lambda(\mathfrak{R}_j')$ be a partial injective function such that $\delta(d) = l$ implies that d and l are assigned to the same node in $V(S)$. Then the structure $\mathfrak{R}_m = (G, S, \mathfrak{R}_1, ..., \mathfrak{R}_k, \mathfrak{R}_1', ..., \mathfrak{R}_t', \delta)$ is called *multiple reconciliation*.

The crucial property of a multiple reconciliation is that a loss from one tree (G' or T_i) can be assigned by δ to a duplication from another gene tree. The cost of a multiple reconciliation is computed the same way as the cost of a reconciliation. The multiple reconciliation *induced* by a reconciliation $\mathfrak{R}$ and an edge e is composed of all parts of $\mathfrak{R}$ mapped to $S(e)$ by ρ. If it is evident from the context, instead of *multiple reconciliation*, we will write *reconciliation*, allowing additional lost subtrees. Let $\mathfrak{R}_m$ be a multiple reconciliation with $e \in E(S)$. Let $\mathfrak{R}_{m1}$ be the reconciliation obtained from $\mathfrak{R}_m$ by adding k new lost subtrees with only one root edge assigned to e. Obviously $\omega(\mathfrak{R}_m) + k = \omega(\mathfrak{R}_{m1})$, but it is possible that $\omega(c(\mathfrak{R}_m)) = \omega(c(\mathfrak{R}_{m1}))$ (see Figure 2).

Definition 17 (Flow) Let $\mathfrak{R}$ be a reconciliation, $e \in E(S)$, and $\mathfrak{R}(e)$ the multiple reconciliation induced with $\mathfrak{R}$ and e. Let $\mathfrak{R}'(e)$ be the reconciliation obtained from $\mathfrak{R}(e)$ by removing all $T_1, \ldots, T_l$ the lost trees containing only one loss assigned to e. With $\mathfrak{R}_k(e)$ denote multiple reconciliation obtained

from $\mathfrak{R}'(e)$ by adding k lost trees containing only one loss assigned to e (k may be lower or higher than l, if $k = l$ then $\mathfrak{R}_k = \mathfrak{R}$). Let k' be the maximum number such that $\omega(c(\mathfrak{R}_{k'}(e))) = \omega(c(\mathfrak{R}'(e)))$. With $F(e, \mathfrak{R}) = F(e) = k' - l$ is denoted the *flow* of the edge e.

Note that if $F(e) \geq 0$, then $F(e)$ is the maximum number of extra losses assigned to e that does not change the weight of the completion of $\mathfrak{R}(e)$. Opposite is also true, if $m \geq 0$ is the maximum number of extra losses assigned to e that does not change the weight of a completion of $\mathfrak{R}(e)$, then $m = F(e)$.

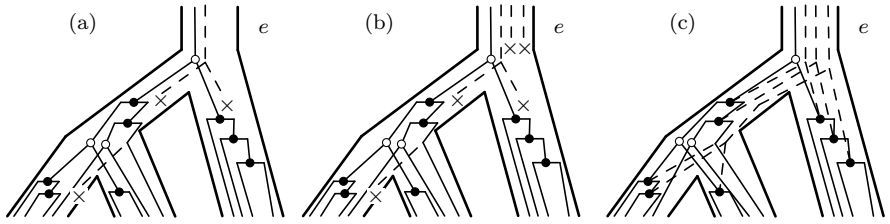

Fig. 2 Flow. (a) Multiple reconciliation $\mathfrak{R}_1$. (b) Multiple reconciliation $\mathfrak{R}_2$ obtained from $\mathfrak{R}_1$ by adding k extra ($k = 2$) losses to the edge e. We have $\omega(\mathfrak{R}_2) = \omega(\mathfrak{R}_1) + k$. (c) Completion of $\mathfrak{R}_2$. Completion of $\mathfrak{R}_1$ can be obtained by removing added lost subtrees. We see that $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1))$. Maximum number k for which the last relation holds is called flow of the edge e

We show how to efficiently compute the flow recursively with Lemma 16. Recall $\mathfrak{D}(e) = \mathfrak{D}(e, \mathfrak{R}_m)$, $\mathfrak{L}(e) = \mathfrak{L}(e, \mathfrak{R}_m)$ denote number of duplications and losses assigned to e in reconciliation $\mathfrak{R}_m$.

Lemma 16 *Let $\mathfrak{R}_m$ be a multiple reconciliation, $e \in E(S)$. Then*

$$F(e) = max\Big(min\Big(F(e_l), F(e_r)\Big), 0\Big) + \mathfrak{D}(e) - \mathfrak{L}(e).$$

Proof We will use mathematical induction on e. Let e be a leaf edge. Then $e_l = NULL$, $e_r = NULL$, and $F(e_l) = F(e_r) = 0$. The only way new losses, assigned to e, can be free is by pairing them with duplications in e. Therefore $k' = d$ and $F(e) = k' - l = d - l$.

Now, let e be a non-leaf edge, $m = max(min(F(e_l, \mathfrak{R}_m(e_l)), F(e_r, \mathfrak{R}_m(e_r))), 0)$, $d = \mathfrak{D}(e, \mathfrak{R}_m(e))$, and $l = \mathfrak{L}(e, \mathfrak{R}_m(e))$. By inductive hypothesis, we can extend m losses over e_r and e_l, so the weight of the completions of $\mathfrak{R}_m(e_l)$ and $\mathfrak{R}_m(e_r)$ is not changed. We can make d losses, assigned to e, free by pairing them with duplications in e. Hence $k' = m + d$ and $F(e) = k' - l = m + d - l$. □

Lemma 17 *Let $\mathfrak{R}_1$ be a multiple reconciliation with a root edge $e = (s, p(s))$, and $F(e, \mathfrak{R}_1) \leq 0$. By assigning an extra loss to e we obtain $\mathfrak{R}_2$. Then $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1)) + 1$.*

We postpone the proof of this Lemma to section 5 because it will use some notions introduced later.

The next lemma is a consequence of Lemma 17.

Lemma 18 *Let $\mathfrak{R}_1$ be a (multiple) reconciliation, e is the root edge of S, and $F(e, \mathfrak{R}_1) < 0$. Let $\mathfrak{R}_2$ be a reconciliation obtained from $\mathfrak{R}_1$ by removing a loss assigned to e. Then $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1)) - 1$*

Lemmas 17 and 18 are stated in a way of adding and removing a loss from the root edge e. Similar lemmas are in effect if we remove/add a duplication from/to the root edge e. Because of the obviousness we will not state them nor prove them.

Thanks to this flow computation we can find a completion of any reconciliation by a polynomial time algorithm, which pseudo-code is written in Algorithm 1 and 2.

Algorithm 1 Find a completion of a reconciliation.

```
1: procedure OneCompletion(ℜ)
2:     while there is a loss l ∈ Σ\Σ' assigned to edge e such that either there is a duplication that is not a conversion assigned to e or children of e have positive flow do
3:         ExtendLossIntoFreeTree(ℜ, l)
4:     end while
5: end procedure
```

Algorithm 2 Extends one loss into a free tree.

```
1: procedure ExtendLossIntoFreeTree(ℜ, l)
2:     l is assigned to e = (s, p(s)) and e1, e2 are children of e
3:     Δ''(e) is the set of all duplications that are not conversion assigned to e
4:     if Δ''(e) ≠ ∅ and F(e1) > 0 and F(e2) > 0 then
5:         Randomly choose between "assign" and "extend"
6:     end if
7:     if F(e1) ≤ 0 or F(e2) ≤ 0 or "assign" has been chosen then
8:         Assign l to a random d ∈ Δ''(e)
9:     else
10:        extend l over e1, e2
11:        l1, l2 are new losses assigned to e1, e2
12:        ExtendLossIntoFreeTree(ℜ, l1)
13:        ExtendLossIntoFreeTree(ℜ, l2)
14:    end if
15: end procedure
```

Let us introduce a convention. If we say that, *e.g.* $\mathfrak{R}'$ *is an output of* ExtendLosses($\mathfrak{R}$), then the procedure ExtendLosses(.) is observed as a standalone procedure with the input $\mathfrak{R}$. But if we say that $\mathfrak{R}'$ *is an output of* ExtendLosses (no input parameters), then we observe ExtendLosses as a part (sub procedure) of the main procedure, and ExtendLosses receives parameters as described.

Lemma 19 *Let $\mathfrak{R}$ be a reconciliation, l is a non-free loss assigned to $e \in E(S)$, e_1, e_2 are children of e. Next, $\Delta''(e) \neq \emptyset$ or ($F(e_1) > 0$ and $F(e_2) > 0$). Then the procedure* ExtendLossIntoFreeTree$(\mathfrak{R}, l)$ *extends l into a free tree.*

Proof Note that if $\Delta''(e) = \emptyset$ and $F(e_1) = F(e_2) = 0$, then $F(e) \leq 0$.

We will use mathematical induction on e. Let e be a leaf edge. Then $e_1 = NULL$, $e_2 = NULL$ and $F(e_1) = F(e_2) = 0$. Hence $\Delta''(e) \neq \emptyset$, and l is assigned to a random duplication from $\Delta''(e)$.

Assume that e is not a leaf edge. If $\Delta''(e) \neq \emptyset$ and *assign* is chosen, then l is assigned to a random element from $\Delta''(e)$, *i.e.* l is extended into a free tree with one edge. If $\Delta''(e) = \emptyset$ or *extend* is chosen, then $F(e_1) > 0$, $F(e_2) > 0$ and l is extended into l_1 and l_2. Since $F(e_i) > 0$, $(i = 1, 2)$ then e_i satisfies *if* condition in OneCompletion. Hence, by inductive hypothesis, ExtendLossIntoFreeTree$(\mathfrak{R}, l_i)$ extends l_i into a free tree, *i.e.* l is extended into free tree. □

Let us introduce a convention. Let $e = (x, p_{G'}(x)) \in E(G')$. If $\rho(p_{G'}(x)) = p(\rho(x))$, then we can write $\rho(e) = (\rho(x), \rho(p_{G'}(x))) \in E(S)$. This property does not hold for any edge of G', but it holds for any edge of a lost subtree, since we do not observe lost subtrees with duplications (an optimal reconciliation cannot have a lost subtree with a duplication). Let T be a subtree of G', then $\rho(t) = \{\rho(e) \mid e \in E(T)\}$. Sometimes we will identify lost trees with their root, *i.e.* v can denote both a root of a tree or a tree with root v. The reason for this is that lost subtrees are dynamical, they extend or switch (an operation introduced later), but their roots are not.

Lemma 20 *Let $\mathfrak{R}$ be a reconciliation with non-extended losses, t_i $(i = 1, \ldots, k)$ and t'_j $(j = 1, \ldots, m)$ are free and non-free lost subtrees of $c(\mathfrak{R})$ such that $t'_j \geq t_i$ whenever t_i and t'_j overlap. All non-free lost subtrees t'_j $(j = 1, \ldots, m)$ are non-extended,* i.e. *they have one edge each. Then $c(\mathfrak{R})$ is a possible output of* OneCompletion$(\mathfrak{R})$.

Proof Let $\mathfrak{R}_0 = \mathfrak{R}$, $\mathfrak{R}_i$ is obtained from $\mathfrak{R}_{i-1}$ by extending corresponding loss to the tree t_i $(i = 1, \ldots, k)$. Hence $\mathfrak{R}_k = c(\mathfrak{R})$.

Assume that trees $t_1, \ldots, t_{i-1}$ $(i \geq 1)$ are constructed by iterations of ExtendLossIntoFreeTree. Take t_i that has the minimal root among free lost subtrees that are not added. Let us prove that $F(e, \mathfrak{R}_{i-1}) > 0$, $\forall e \in E(\rho(t_i)) \backslash \{root_E(\rho(t_i))\}$. Assume the opposite, let $F(e_1, \mathfrak{R}_{i-1}) \leq 0$, and since free subtree t_i extends over e_1, we have that some loss in $S(e_1)$ becomes non-free. More precisely, $\omega(c(\mathfrak{R}_{i-1}(e_1))) < \omega(c(\mathfrak{R}_i(e_1)))$. This means that $|\Lambda \backslash \Lambda'(c(\mathfrak{R}_{i-1}(e_1)))| < |\Lambda \backslash \Lambda'(c(\mathfrak{R}_i(e_1)))|$. Since trees $t_1, \ldots, t_{i-1}$ (and t_i) are free and already present in $\mathfrak{R}_{i-1}$ (i.e. $\mathfrak{R}_i$), then we can assume that they are not changed in $c(\mathfrak{R}_{i-1})$ (i.e. $c(\mathfrak{R}_i)$), because we gain nothing by further extending free losses (although it is possible).

Observe $c(\mathfrak{R}_i(e_1))$. Let T_S be the maximal subtree of $S(e_1)$ (see Figure 7) such that if $v_0 \in V(T_S) \backslash L(T_S)$ is a lost subtree in $c(\mathfrak{R}_i(e_1))$, then there are

lost subtrees (in $c(\mathfrak{R}_i(e_1))$) $v_1, \ldots, v_s$, $\rho(v_0) < \rho(v_1) < \ldots < \rho(v_s)$, v_i overlaps with v_{i+1} $(i = 0, \ldots, s-1)$ and $v_s = t_i$.

Let $v \in V(T_S) \backslash L(T_S)$ be a lost subtree. Let us prove that v is a free tree (in $c(\mathfrak{R}_{i-1}(e_1))$, $c(\mathfrak{R}_i(e_1))$, and $c(\mathfrak{R})$). From $v \in V(T_S) \backslash L(T_S)$ we have $v = v_0 < v_1 < \ldots, v_{s-1} < v_s = t_i$ and v_{i-1} overlaps v_i. Since v_{s-1} overlaps t_i (in $c(\mathfrak{R}_i(e_1))$) and t_i is the same in both $c(\mathfrak{R}_i(e_1))$ and $c(\mathfrak{R})$, we have that v_{s-1} overlaps t_i in $c(\mathfrak{R})$, hence v_{s-1} is a free tree in $c(\mathfrak{R})$, *i.e.* $v_{s-1} \in \{t_1, \ldots, t_{i-1}\}$. Applying the same argument on v_{s-1}, we get $v_{s-2} \in \{t_1, \ldots, t_{i-1}\}$. Proceeding in this manner, we have $v \in \{t_1, \ldots, t_{i-1}\}$, hence v is a free tree.

Let $f_1, \ldots, f_r$ be the children of leaf edges of T_S. From the maximality of T_S, we have there is no lost subtree in $c(\mathfrak{R}_{i-1}(e_1))$ nor in $c(\mathfrak{R}_i(e_1))$ that expands over f_j, $(j = 1, \ldots, r)$. All non-free losses from $S(e_1)$ are contained in $S(f_j)$, $(j = 1, \ldots, r)$. This holds for both $c(\mathfrak{R}_{i-1})$ and $c(\mathfrak{R}_i)$. Therefore the structure of the lost subtrees in $\mathfrak{R}_{i-1}(f_j)$ can be identical to the structure of the lost subtrees in $\mathfrak{R}_i(f_j)$, $(j = 1, \ldots, r)$, and thus obtaining that a completion of $\mathfrak{R}_{i-1}(e_1)$ has the same weight as an extension of $\mathfrak{R}_i(e_1)$, a contradiction.

Hence the procedure ExtendLossIntoFreeTree can give us t_i, $(i = 1, \ldots, k)$. □

It is proved in Section 5, in a more general framework, that these procedures indeed compute a completion, and hence, if the input reconciliation is the LCA reconciliation, it computes an optimal reconciliation.

4 Zero-flow reconciliations and the space of all optimal reconciliations

Here we introduce zero-flow reconciliations and use them as a hinge to find all optimal reconciliations. Zero-flow (ZF) reconciliations are a subspace of optimal reconciliations and they contain LCA reconciliations, but these inclusions are strict: all sets are distinct. We first show how to find any ZF reconciliation, up to completion, from an LCA reconciliation. Then by a different procedure we show how to access the whole space of optimal reconciliations, up to completion, from a ZF reconciliation. Finally, as these reductions work up to completion, we show how to navigate in all completions for a given reconciliation.

Let $e = (s, p(s))$ be an edge of S and $\mathfrak{R}$ a reconciliation. We note $X(e, \mathfrak{R}) = \{d \in V(G) \mid \rho_{lca}(d) \leq s, \rho(d) \geq p(s)\}$ the set of nodes (duplications or conversions) which are assigned under s in the LCA reconciliation and above $p(s)$ in $\mathfrak{R}$.

Definition 18 An optimal reconciliation $\mathfrak{R}$ is said to be a *zero-flow* (ZF) reconciliation if for all s internal node of S with children edges e_1 and e_2, $F(e_1, \mathfrak{R}) < 0 \implies X(e_1, \mathfrak{R}) = X(e_2, \mathfrak{R}) = \emptyset$.

In other words, an optimal reconciliation is ZF if all duplications assigned to or above a node s, when strictly below in the LCA, verify that the flow

the children edges of s is non negative. By definition LCA reconciliations are ZF ($X(e, \mathfrak{R}_{lca}) = \emptyset$ for all e). But we will see that the converse is not true. Similarly ZF reconciliations are optimal by definition but some optimal reconciliations are not ZF.

4.1 Computing ZF reconciliations by duplication raising

Duplication raising consists in changing the position of a duplication from its position in a minimal reconciliation to an upper position in the species tree. It is a concept that was previously used to explore DL reconciliations Chauve et al. (2008).

Definition 19 (Node raising) Let $\mathfrak{R} = (G, G', S, \phi, \rho, \delta)$ be a minimal reconciliation and $x \in V(G)$. We say that reconciliation $\mathfrak{R}' = (G, G'', S, \phi, \rho', \delta')$ is obtained from $\mathfrak{R}$ by *raising* node x if $\mathfrak{R}'$ is a minimal reconciliation such that $\rho(x') = \rho'(x')$, $\forall x' \in V(G)\backslash\{x\}$ and $\rho'(x) = p(\rho(x))$.

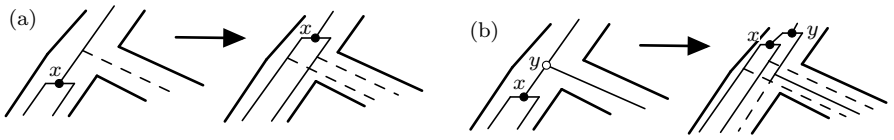

Fig. 3 Duplication raising. Given duplication x (a) No speciation. After raising x, a new loss is created. An optimal solution can be generated by this operation. (b) We have $y = p_G(x)$ and $\rho(y) = p(\rho(x))$. After raising x, y becomes duplication and three new losses are generated. This cannot be optimal

Depending on the assignment and event status of the parent node of x, raising x has different effects. If $p_G(x)$ is a speciation (see Figure 3) and $\rho(p_G(x)) = p(\rho(x))$, after raising x, $p_G(x)$ becomes a duplication and three new losses are generated. This cannot lead to an optimal solution because of the additional duplication (Theorem 1). If $\rho(p_G(x)) > p(\rho(x))$ or $p_G(x)$ is a duplication, after raising x, only one additional loss is generated. This condition, which is necessary to yield an optimal solution, is formalized as follows.

$$x \in \Delta(\mathfrak{R}) \wedge \Big(p(\rho(x)) < \rho(p_G(x)) \vee \Big(p(\rho(x)) = \rho(p_G(x)) \wedge p_G(x) \in \Delta(\mathfrak{R}) \Big) \Big) \tag{1}$$

The next lemma states that raising a duplication cannot decrease the weight of a completion. The proof of the lemma also describes how to lower a duplication. This procedure will be important later in some proofs.

Lemma 21 *Let $\mathfrak{R}$ be a minimal reconciliation, $\mathfrak{R}_1$ is a minimal reconciliation obtained from $\mathfrak{R}$ by raising a duplication. Then $\omega(c(\mathfrak{R})) \leq \omega(c(\mathfrak{R}_1))$.*

Proof Let x be the raised duplication, $e_1, e_2 \in E(S)$ are siblings, e is their parent, x is assigned to e_1 in $\mathfrak{R}$ and to e in $\mathfrak{R}_1$.

Let T be the lost subtree such that $root(T)$ is a child of x in $c(\mathfrak{R}_1)$ and T is expanded over e_2. Observe two cases (see Figure 4).

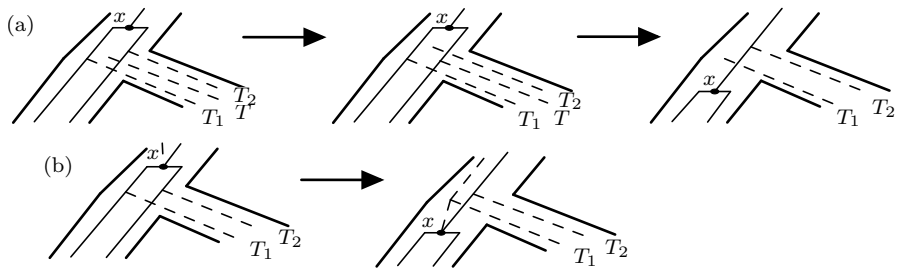

Fig. 4 Lowering duplication. (a) Lowering duplication (that is not a conversion). If there is non-free subtree T, not associated with x, we can make it associated by rearranging the roots. By lowering x we can delete one lost subtree (T), and if it is non-free, then we get a cheaper reconciliation. (b) Lowering a conversion. Loss assigned to x can be extended to one of the lost subtrees on the right side. We get a reconciliation of the same weight

Case 1, $x \notin \Delta'(c(\mathfrak{R}_1))$. Start with $c(\mathfrak{R}_1)$, place x back to e_1, and remove T. We get an extension of $\mathfrak{R}$ with a cost at most the one of $c(\mathfrak{R}_1)$, *i.e.* $\omega(c(\mathfrak{R})) \leq \omega(c(\mathfrak{R}_1))$.

Case 2, $x \in \Delta'(c(\mathfrak{R}_1))$. Let l be a loss assigned to x in $c(\mathfrak{R}_1)$. Start with $c(\mathfrak{R}_1)$, place x back to e_1, extend l, so that in e_1 is paired with x (staying free loss), and in e_2 is connected to T. In this way, we get an extension of $\mathfrak{R}$ of the same weight as $c(\mathfrak{R}_1)$, i.e. $\omega(c(\mathfrak{R})) \leq \omega(c(\mathfrak{R}_1))$. □

As a consequence of Lemma 21, no optimal reconciliation can be obtained by raising a duplication from a reconciliation that has no optimal completion. We will now see in which conditions a duplication raising of a reconciliation with an optimal completion can lead to another reconciliation with optimal completion.

The next lemma states when raising a duplication does not increase the weight of a reconciliation.

Lemma 22 *Let $\mathfrak{R}$ be a minimal reconciliation, and $e_1, e_2 \in E(S)$ the children of edge e. If $x \in \Delta(\mathfrak{R})$ assigned to e_1 satisfies condition* (1)*, $\mathfrak{R}_1$ is a minimal reconciliation obtained by raising x, $F(e_1, \mathfrak{R}) > 0$ and $F(e_2, \mathfrak{R}) > 0$, then $\omega(c(\mathfrak{R}_1)) = \omega(c(\mathfrak{R}))$.*

Proof First, construct an extension of $\mathfrak{R}_1$, by using $c(\mathfrak{R})$. By raising x, we generate one new loss in e_2. Since $F(e_2, \mathfrak{R}) > 0$, we have $\omega(c(\mathfrak{R}(e_2))) = \omega(c(\mathfrak{R}_1(e_2)))$, *i.e.* the loss generated by the duplication raising can become a free loss.

Let $x \in \Delta'(c(\mathfrak{R}))$ and assigned to $l \in \Lambda'(c(\mathfrak{R}))$. If l is non-extended (in $c(\mathfrak{R})$) and since $F(e_1, \mathfrak{R}) > 0$, we have that l can be assigned to some other

duplication in e_1 or extend over children of e_1 and become free. If l is part of a lost subtree T_l in $c(\mathfrak{R})$, then by raising x, we ca also raise l, remove subtree of T_l expanding over e_2, leave l assigned to x.

Thus we obtain an extension of $\mathfrak{R}_1$, not heavier than $c(\mathfrak{R})$, i.e. $\omega(c(\mathfrak{R}_1)) \leq \omega(c(\mathfrak{R}))$. From Lemma 21, we have $\omega(c(\mathfrak{R})) \leq \omega(c(\mathfrak{R}_1))$, hence $\omega(c(\mathfrak{R}_1)) = \omega(c(\mathfrak{R}))$. □

The next lemma follows directly from Lemma 22.

Lemma 23 *Under the hypotheses of Lemma 22, if completions of $\mathfrak{R}$ are optimal, then completions of $\mathfrak{R}_1$ are optimal.*

Algorithms 3, 4, and 5 describe how to generate a reconciliation which does not change the score of completions by raising duplications.

Algorithm 3 Raises duplications

```
1: procedure RaiseSeveralDuplications(d, ℜ)
2:     for d ∈ Δ(ℜ_lca) from top to bottom of V(G) do
3:         ℜ ← RaiseDuplication(d, ℜ_lca)
4:     end for
5: end procedure
```

Algorithm 4 Raises duplication (respecting $F > 0$)

```
1: procedure RaiseDuplication(d, ℜ)
2:     L ← PossiblePositions(d, ℜ)
3:     k = random(0,|L| − 1)
4:     ρ(d) = L[k]
5:     GenerateNewLosses()
6: end procedure
```

Procedure GenerateNewLosses adds lost subtrees to that the new ρ after raising a duplication is consistent with S.

The two next statements demonstrate that, up to completion, all the ZF reconciliations are reached by applying Algorithm RaiseDuplication on a LCA reconciliation.

Lemma 24 *Completions of $\mathfrak{R}$, an output of* RaiseDuplication *when the input is the LCA reconciliation, are optimal.*

Proof Completion of LCA reconciliation is an optimal (Theorem 2), raised duplications satisfy conditions of Lemma 23, and by this Lemma every time a duplication is raised we get that $c(\mathfrak{R})$ is an optimal reconciliation. □

Lemma 25 *Let $\mathfrak{R}'$ be a minimal reconciliation such that $c(\mathfrak{R}')$ is a ZF reconciliation. Then $\mathfrak{R}'$ is a possible output of* RaiseDuplication.

Algorithm 5 Possible new positions for a duplication.

```
1: procedure PossiblePositions(d, ℜ)
2:     s = ρ(d)
3:     e = (s, p(s))
4:     e' - sibling of e
5:     L ← {s}
6:     while F(e) > 0 and F(e') > 0 and (ρ(p_G(d)) > p(s) or (ρ(p_G(d)) == p(s) and
       p_G(d) ∈ Δ )) do
7:         s = p(s)
8:         e = (s, p(s))
9:         e' - sibling of e
10:        L ← L + {s}
11:    end while
12: end procedure
```

Proof Since $c(\mathfrak{R}')$ is an optimal reconciliation, $\mathfrak{R}'$ is obtained from LCA by raising duplications that satisfy condition (1). By raising a duplication, value of $F(e)$ cannot increase. Let $e_1, e_2 \in E(S)$ be siblings, e their parent, x a duplication assigned to e_1. Let us raise x to e. If before raising $F(e_1) \leq 0$ or $F(e_2) \leq 0$, then after raising $F(e_1) < 0$ or $F(e_2) < 0$, $X(e_1) \neq \emptyset$, and $X(e_2) \neq \emptyset$, a contradiction. Hence $F(e_1) > 0$ and $F(e_2) > 0$.

Thus all conditions, for raising a duplication, of the procedure RaiseDuplication are satisfied, hence $\mathfrak{R}'$ is a possible output. □

4.2 Reduction of optimal reconciliations to ZF reconciliations

Lemma 25 states that up to completion, we can generate all ZF reconciliation from LCA reconciliations. We now show how to generate all reconciliations from ZF reconciliations. This is done by conversion raising. Next lemma proves that only conversions are concerned by optimal non ZF reconciliations.

Lemma 26 *Let $\mathfrak{R}$ be an optimal reconciliation, $e_1 = (s_1, s), e_2 = (s_2, s) \in E(S)$. If $F(e_1, \mathfrak{R}) < 0$, then $X(e_1, \mathfrak{R})$ and $X(e_2, \mathfrak{R})$ are only conversions.*

Proof Assume the opposite, let $x \in X(e_1, \mathfrak{R})$ and x is not a conversion. Put back (lower) all elements of $X(e_1, \mathfrak{R})$ to e_1. The process is performed as in the proof of Lemma 21 (Figure 4). If we lower a conversion, the weight of a reconciliation is not changed, as well as $F(e_1)$. If we lower a duplication, then $F(e_1)$ is increased by 1 and the cost of a completion is decreased by one (Lemmas 17, 18 and the comment after), which is a contradiction with the optimality of $\mathfrak{R}$. Therefore, $X(e_1, \mathfrak{R})$ does not contain a duplication that is not a conversion.

Similar arguments apply to $X(e_2, \mathfrak{R})$. □

Lemma 27 *Procedure* RaiseConversions *does not change the weight of a reconciliation.*

Proof Let d be a raised conversion, and T_i is a lost subtree whose leaf is assigned to d. By raising d, we do not create an extra losses, but use existing subtree of T_i and reattach it under d (see Figure 4 (ii) in the opposite direction and Lemma 21, Case 2). The loss that was assigned to d is removed, and newly created loss is assigned to d at a new position. In this way we do not change the number of non-free losses, and the number of duplications/conversions, i.e. the weight of the reconciliation is not changed. □

Lemma 28 *Let $\mathfrak{R}$ be an optimal reconciliation. We can obtain a ZF reconciliation by lowering some conversions.*

Proof For all $e \in E(S)$, if $F(e) < 0$, take all elements from $X(e)$ and $X(e')$, where e' is the sibling of e, and lower them to e and e'. In this way we get $X(e) = X(e') = \emptyset$. Since these elements are conversions (Lemma 26) lower them as described in Lemma 21, Case 2.

In this way we obtain a ZF reconciliation of the same weight as $\mathfrak{R}$. □

In consequence it is possible to reach any optimal reconciliation by an algorithm which explores first ZF reconciliations and raises some conversions as in Algorithm 6.

Algorithm 6 raises some conversions

procedure RaiseConversions($\mathfrak{R}$)
  By convention let $e_1(d)$ denote the edge to which d is assigned, and $e_2(d)$ its sibling in S.
  Let $C = \{d \mid d \in \Delta', F(e_1(d)) < 0 \text{ or } F(e_2(d)) < 0\}$
  Let T_d be used to denote the lost subtree with a leaf paired with d by δ.
  while $C \neq \emptyset$ **do**
    $d \in C$ - **random**
    RaiseOneConversion($d, \mathfrak{R}, T_d$)
    $C = C \backslash \{d\}$
  end while
end procedure

Algorithm 7 raises one conversion

procedure RaiseOneConversion($d, \mathfrak{R}, T_d$)
  Let s be a random element of $V(S)$ satisfying
  (i) $s \geq p(\rho(d))$
  (ii) $s \leq min(\rho(root(T_d)), \rho(p_G(d)))$
  (iii) if $p_G(d) \in \Sigma$ then $s \neq \rho(p_G(d))$
  Note $\rho(d) = s_0 < s_1 < \ldots < s_k = s$
  T_d^j - subtree of T_d, $\rho(root(T_d^j)) = s_j$, $j = \overline{1,k}$
  assign d to random s_i
  node (leaf) of T_d, assigned to s_i, pair with d (and d stays conversion)
  root of every tree T_d^j position in G', under d, at an appropriate position
end procedure

4.3 Finding all completions

All previous results are valid up to completions. It means that we have an algorithm which is able to detect all duplications that can be conversions in one optimal solution for example. However we don't know all the possibilities by which it is converted. For that we need to enumerate all possible completions. The algorithm can be described by three procedures, as written in Algorithm 8.

Algorithm 8 finds a random completion

```
1: procedure AllCompletions(ℜ)
2:     OneCompletion(ℜ)
3:     ExtendLossesIntoNonFreeTrees(ℜ)
4:     Switch(ℜ_c)
5: end procedure
```

One procedure is to generate a completion by extending losses into free trees, which is described in Section 3.4. In order to generate the full diversity of possible reconciliations, there are two others described here, which consist in extending losses into non free lost subtrees, and switch between subtrees. The first one is described in Algorithms 9 and 10. In Algorithm 10 a loss is extended over two edges, one with positive F-value (say edge e_1), and the other with non-positive F-value (say edge e_2). The part (of the lost subtree) extended over e_1 is further extended as a free loss, while the part extended over e_2 is further (recursively) extended as a non-free loss.

Algorithm 9 randomly extends losses into non-free trees

```
1: procedure ExtendLossesIntoNonFreeTrees(ℜ)
2:     Σ_1 is the set of all non-free, non-extended losses in ℜ
3:     for all l ∈ Σ_1 do
4:         ExtendOneLossIntoNonFreeTree(ℜ, l)
5:     end for
6: end procedure
```

Lemma 29 *Let l be a non-free loss in a reconciliation $\mathfrak{R}$. Then procedure* ExtendOneLossIntoNonFreeTree($\mathfrak{R}, l$) *extends loss l into a non-free tree.*

Proof If l is not extended, since it is not assigned to a duplication (conversion) we will assume that it is extended into a non-free tree (with one edge).

Let l be assigned to the edge e, and e_1, e_2 are its children. We will use mathematical induction on e.

Let e be a leaf edge. Then $e_1 = NULL, e_2 = NULL$ and $F(e_1) = F(e_2) = 0$. In this case, the *if* condition is not satisfied, and therefore l is not extended.

Assume that e is not a leaf edge. If the *if* condition is not satisfied, then l is not extended, *i.e.* it is extended into a non-free tree with one edge. If the

Algorithm 10 randomly extends losses into non-free trees

```
1: procedure ExtendOneLossIntoNonFreeTree(ℜ, l)
2:     l is assigned to e = (s, p(s))
3:     e_1, e_2 are children of e and F(e_1) ≥ F(e_2)
4:     Randomly choose between "extend" or not.
5:     if F(e_1) > 0 and F(e_2) ≤ 0 and "extend" has been chosen then
6:         extend l over e_1, e_2
7:         l_1, l_2 are new losses assigned to e_1, e_2 and l is their parent
8:         ExtendOneLossIntoFreeTree(ℜ, l_1)
9:         ExtendOneLossIntoNonFreeTree(ℜ, l_2)
10:    end if
11: end procedure
```

if condition is satisfied, then $F(e_1) > 0$ and $F(e_2) \leq 0$, and l is extended into l_1, l_2. Then ExtendOneLossIntoFreeTree$(\mathfrak{R}, l_1)$ extends l_1 into a free tree (Lemma 19), and ExtendOneLossIntoNonFreeTree$(\mathfrak{R}, l_2)$ extends l_2 into a non-free tree (inductive hypothesis). Hence l is extended into a non-free tree. □

The next lemma is a consequence of Lemma 29

Lemma 30 *Procedure* ExtendLossesIntoNonFreeTrees *does not change the weight of a reconciliation.*

Lemma 31 *Let $\mathfrak{R}$ be a reconciliation with non-extended losses, t_i $(i = 1 \dots k)$ and t'_j $(j = 1 \dots m)$ are free and non-free lost subtrees of $c(\mathfrak{R})$ such that $t'_j \geq t_i$ whenever t_i and t'_j overlap. Then $c(\mathfrak{R})$ is a possible output of series of procedures* OneCompletion$(\mathfrak{R})$, ExtendLossesIntoNonFreeTrees$(\mathfrak{R})$.

Proof Let $\mathfrak{R}_0 = \mathfrak{R}$, $\mathfrak{R}_i$ is obtained from $\mathfrak{R}_{i-1}$ by extending corresponding loss to the tree t_i $(i = 1, \dots, k)$, $\mathfrak{R}'_0 = \mathfrak{R}_k$, $\mathfrak{R}'_j$ is obtained from $\mathfrak{R}'_{j-1}$ by extending corresponding loss to the tree t'_j $(j = 1, \dots, m)$. Hence $\mathfrak{R}'_m = c(\mathfrak{R})$.

The procedure OneCompletion can give us t_i, $(i = 1, \dots, k)$ (Lemma 20). Now we will prove that ExtendLossesIntoNonFreeTrees can give us t'_j, $(j = 1, \dots, m)$.

Assume that t_i, $(i = 1, \dots, k)$, $t'_1, \dots, t'_{j-1}$ $(j \geq 1)$ are added. Let us prove that ExtendLossesIntoNonFreeTrees can add t'_j. Let $e_1, e_2 \in E(S)$, $e = (s, p(s))$ is their parent, and $\rho(l'_j) = s$, where l'_j extends into t'_j. If $F(e, \mathfrak{R}'_{j-1}) > 0$, then l'_j can be free, thus obtaining a cheaper reconciliation than $c(\mathfrak{R})$, a contradiction, so $F(e, \mathfrak{R}'_{j-1}) \leq 0$.

Let $e'_1, e'_2 \in E(\rho(t'_j))$ be siblings, e' their parent, and $F(e'_1, \mathfrak{R}'_{i-1}) \geq F(e'_2, \mathfrak{R}'_{i-1})$. Subtree t'_j expands over e'_1, e'_2 and not necessarily originating at e'. Observe two cases.

Case 1, $F(e', \mathfrak{R}'_{j-1}) \leq 0$. If $F(e'_1, \mathfrak{R}'_{j-1}) \leq 0$ (and $F(e'_2, \mathfrak{R}'_{j-1}) \leq 0$), then by pruning t'_j both e'_1 and e'_2 don't gain a loss, so the cost of reconciliations $c(\mathfrak{R}'_{j-1}(e'_1))$ and $c(\mathfrak{R}'_{j-1})(e'_2)$ will not rise in $\mathfrak{R}'_j$, but $\mathfrak{R}'_j$ gain one non-free loss (pruned t'_j). Hence we gain a cheaper reconciliation, a contradiction.

Assume $F(e'_1, \mathfrak{R}'_{j-1}) > 0$ and $F(e'_2, \mathfrak{R}'_{j-1}) > 0$. Since $F(e', \mathfrak{R}'_{j-1}) \leq 0$, there is a loss l assigned to e' that is non-free (in $\mathfrak{R}'_{j-1}$). Then we can extend l

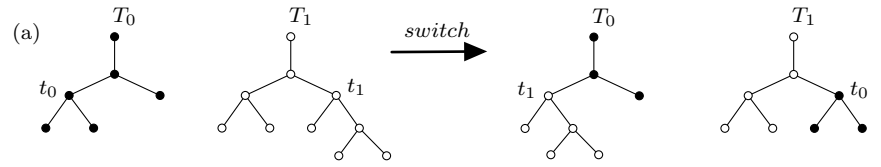

Fig. 5a Switch operation between binary trees. (a) Switch between T_0 and T_1 around t_0 and t_1

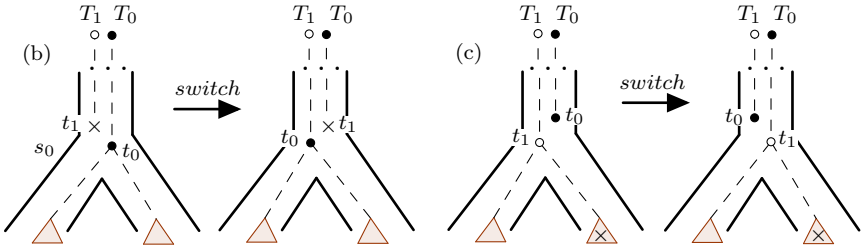

Fig. 5b Switch operation between reconciliations. (b-c) Switch on a reconciliation. Exactly one lost subtree receives a (nontrivial) subtree from the other lost subtree. A subtree with a non-free loss has to be involved in a switch operation. An empty triangle denotes a free subtree, while a triangle with x denotes a non-free subtree

over e'_1, e'_2 so it becomes free, and prune t'_j to a single edge (t'_j stays non-free). Hence obtaining a cheaper reconciliation than $c(\mathfrak{R})$, a contradiction.

Case 2, $F(e', \mathfrak{R}'_{j-1}) > 0$. If $F(e'_2, \mathfrak{R}'_{j-1}) \leq 0$, then e' has a duplication that is not a conversion. At least one of the subtrees of t'_j expanding over e'_1, e'_2 is a free tree. Assume that it is the one expanding over e'_1. Next, we can prune subtree of t'_j so that t'_j has a leaf assigned to e' and to the duplication, thus becoming a free loss. Since $F(e'_2, \mathfrak{R}'_{j-1}) \leq 0$ there is one non-free loss in $\mathfrak{R}'_{j-1}(e'_2)$ that can become free, thanks to the fact that t'_j does not expand over e'_1 anymore. Making this loss free enable us to obtain a cheaper reconciliation than $c(\mathfrak{R})$, a contradiction.

From the Cases 1 and 2, we have that if $F(e', \mathfrak{R}'_{j-1}) \leq 0$, then $F(e'_1, \mathfrak{R}'_{j-1}) > 0$, $F(e'_2, \mathfrak{R}'_{j-1}) \leq 0$, and if $F(e', \mathfrak{R}'_{j-1}) > 0$, then $F(e'_1, \mathfrak{R}'_{j-1}) > 0$, $F(e'_2, \mathfrak{R}'_{j-1}) > 0$. Hence conditions along $\rho(t'_j)$ of ExtendLossesIntoNonFreeTrees are satisfied, and therefore t'_j can be obtained by this procedure. □

To obtain all possible lost subtrees in an optimal reconciliation, we need to introduce an operation that exchanges parts of the lost subtrees. Notice that a lost subtree with more than one non-free leaf cannot appear in an optimal reconciliation.

Definition 20 (Switch operation on a binary rooted trees) Let T_0 and T_1 be binary rooted trees and $t_i \in V(T_i) \backslash \{root(T_i)\}$ $(i = 0, 1)$. A *switch* operation on T_0 and T_1 around t_0 and t_1 creates new trees by separating subtrees $T_i(t_i)$ from T_i and joining them with $p(t_{1-i}) \in T_{1-i}$ $(i = 0, 1)$.

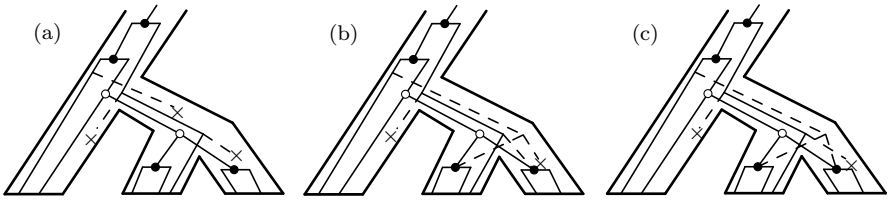

Fig. 6 Example for necessity of switch operation. (a) Minimal reconciliation. (b) The completion. We have one free and two non-free trees. (c) A completion obtained by switch operation. Note that this completion is not obtainable by standard extension into free and non-free trees

Definition 21 (Switch operation on a reconciliation) Let $\mathfrak{R}$ be a reconciliation, T_0 and T_1 free and non-free lost subtrees, $l \in L(T_1)$ is a non-free loss, p is a path in S from $\rho(l)$ to $\rho(root(T_1))$. Assume there exists a minimal element $s_0 \in \{s \mid s \in V(p) \cap V(\rho(T_0))\}\backslash\{\rho(root(T_0)), \rho(root(T_1))\}$, and $t_i \in V(T_i)$ such that $\rho(t_i) = s_0$ $(i = 0, 1)$. By *switch* operation on T_0 and T_1 we mean a switch operation on the binary trees T_0 and T_1 around t_0 and t_1.

Switch operation on a reconciliation is defined only for one free and one non-free lost subtree, and is possible only if trees T_0 and T_1 *overlap*, *i.e.* if $\rho(v_0) \in \rho(T_1)$ or $\rho(v_1) \in \rho(T_0)$, where $v_i = root(T_i)$, $(i = 0, 1)$. In the case $\rho(v_0) \in \rho(T_1)$, it must be $\rho(l) < \rho(v_0)$, where l is a non-free leaf of T_1. In these cases we say that T_0 and T_1 are *switchable*. We have that either T_0 gives a (non-trivial) subtree to T_1, or T_1 gives a (non-trivial) subtree to T_0, but both cannot happen.

When we apply a switch operation two times on the same trees, around the same nodes, we obtain starting trees, i.e. switching is self-inverse operation. After switch operation, involved trees still overlap.

For simplicity of notation, we introduce some conventions. We write *tree* instead of *lost subtree*. We will identify a tree with its root, *i.e.* instead of writing *a tree with the root* v, we will use *a tree* v. We do this because, when switching, trees are changed, but the roots are not. When we write $v_0 < v_1$, we mean $\rho(v_0) < \rho(v_1)$. Number of non-free leaves in a tree v is denoted by $\omega(v)$, thus $\omega(v) = 0$ means that v is a free lost subtree, and $\omega(v) = 1$ means that v is a non-free lost subtree.

If we will apply a switch operation on switchable trees v_0, v_1 such that $\omega(v_1) = 1$ and $\omega(v_0) = 0$, we say that v_1 carries over a (non-free) loss to v_0.

The next lemma is obvious.

Lemma 32 *Switch operation does not change the weight of a reconciliation.*

The next lemma tells us how to, from an arbitrary reconciliation, obtain a reconciliation with more convenient structure of lost subtrees.

Lemma 33 *Let $\mathfrak{R}$ be a reconciliation. Then there exists a reconciliation $\mathfrak{R}_1$ such that if v_0 and v_1 are free and non-free overlapping trees in $\mathfrak{R}_1$, then $v_0 \leq v_1$ and $\omega(\mathfrak{R}) = \omega(\mathfrak{R}_1)$.*

Proof Let $V_{lost} = \{v \mid v \text{ is a lost subtree}\}$. Take $v_0 \in V_{lost}$ such that $\omega(v_0) = 1$, and $v_1 \in V_{lost}$ such that $\omega(v_1) = 0$, $v_0 < v_1$, and v_0 is overlapping with v_1. By switching v_0 and v_1 we get $\omega(v_0) = 0$, $\omega(v_1) = 1$ and $v_0 < v_1$. Repeat the process as long as there are trees v_0, v_1 as described. We need to prove that this algorithm ends.

Let $d(V_{lost})$ be the total distance of all non-free $v \in V_{lost}$ from $root(S)$. Hence $d(V_{lost})$ is a non-negative integer. Every time, when switching is applied, $d(V_{lost})$ decreases, hence the algorithm must stop, because $d(V_{lost})$ cannot decrease indefinitely.

Switch operation does not change the weight of a reconciliation (Lemma 32). □

Algorithm 11 applies switch operation on lost subtrees

```
1: procedure SWITCH(𝔑)
2:     𝔗′ - the set of all non-free lost subtrees in 𝔑
3:     𝔗_v′ - the set of all free lost subtrees, less than v′, switchable with v′
4:     while 𝔗′ ≠ ∅ do
5:         v′ ∈ 𝔗′ - random
6:         v ∈ 𝔗_v′ ∪ {NULL}
7:         if v==NULL then
8:             𝔗′ = 𝔗′\{v′}
9:             continue while loop
10:        end if
11:        SWITCHSUBTREES(v, v′)
12:        𝔗′ = (𝔗′\{v′}) ∪ {v}
13:    end while
14: end procedure
```

Procedure SWITCHSUBTREES is described in Definition 21.

5 The algorithm

In this section, we prove that the algorithm returns an optimal reconciliation, and any optimal reconciliation can be an output of the algorithm. We also prove the remaining lemmas.

All elements are ready to write the main algorithm that generates a random optimal solution.

Algorithm 12 gives the main procedure.

Now we prove a lemma stated earlier.

Proof (Proof of Lemma 17) Let l be the number of assigned losses to e in $\mathfrak{R}_1$, $\mathfrak{R}$ is the (multiple) reconciliation obtained from $\mathfrak{R}_1$ by removing all (l) losses from e, k' is from the definition of flow. Then $F(e, \mathfrak{R}_1) = k' - l$. Therefore, the maximum number of extra losses that we can assign to e in $\mathfrak{R}$, without completion cost change, is k' and $k' \leq l$.

It is obvious that $\Delta(\mathfrak{R}) = \Delta(\mathfrak{R}_1) = \Delta(\mathfrak{R}_2)$. Also $\omega(c(\mathfrak{R})) < \omega(c(\mathfrak{R}_2))$.

Algorithm 12 Random reconciliation

```
1: procedure RandR(S, G, φ)
2:     Let ℜ_lca be the LCA reconciliation
3:     ℜ ← RaiseSeveralDuplications(d, ℜ_lca)
4:     ℜ_c ← AllCompletions(ℜ)
5:     Return RaiseConversions(ℜ_c)
6: end procedure
```

We have that $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1)) + 1$ or $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1))$. Assume that $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1))$.

Observe $c(\mathfrak{R}_2)$. Let $t_1, \ldots, t_l, t_{l+1}$ be the lost subtrees with the roots assigned to $p(s)$ (and expanding over e). If any of these subtrees are non-free in $c(\mathfrak{R}_2)$ then by removing it we get an extension of $\mathfrak{R}_1$ that has strictly less weight than $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1))$, a contradiction. Therefore all subtrees $t_1, \ldots, t_l, t_{l+1}$ are free in $c(\mathfrak{R}_2)$.

Let us prove that there is at least one non-free subtree in $c(\mathfrak{R}_2)$. Assume the opposite, *i.e.* all lost subtrees of $c(\mathfrak{R}_2)$ are free. Then we can have an extension of $\mathfrak{R}_1$ and $\mathfrak{R}$ with all free lost subtrees, by just removing one or all subtrees extending over e. Hence $\omega(c(\mathfrak{R})) = \omega(c(\mathfrak{R}_1)) = \omega(c(\mathfrak{R}_2)) = |\Delta(\mathfrak{R})|$. This means that we can assign at least $l+1$ losses to e in $\mathfrak{R}$ without completion cost change. This contradicts the fact that $k' < l + 1$. Therefore $c(\mathfrak{R}_2)$ has at least one non-free lost subtree.

Let us prove that there exists a chain of lost subtrees $v_1, \ldots, v_{m-1}, v_m$ (in $\mathfrak{R}_2$) such that $v_1 < \ldots < v_m$, v_i overlaps v_{i+1}, $(i = 1, \ldots, m-1)$, v_1 is a non-free tree, $v_2, \ldots, v_m$ are free trees and v_m is a tree assigned to $p(s)$ extending over e.

Assume the opposite. Let T_S be the maximum subtree with root edge e that contains only free lost subtrees (see Figure 7), and $f_1, \ldots, f_r$ edges of S that are children of leaf-edges of T_S. Because of the maximality of T_S and the assumption that there is no chain leading from non-free tree to one of the trees $t_1, \ldots, t_{l+1}$, we have that there is no tree expanding from inner node of T_S over one of the edges $f_1, \ldots, f_r$. Since $\mathfrak{R}_2$ has at least one non-free lost subtree, we have $r \geq 1$, *i.e.* edges $f_1, \ldots, f_r$ do exist.

Since $\omega(c(\mathfrak{R})) < \omega(c(\mathfrak{R}_2))$ and $c(\mathfrak{R}_2)$ has only free trees in T_S, then there is i such that $\omega(c(\mathfrak{R})(f_i)) < \omega(c(\mathfrak{R}_2)(f_i))$. Since no lost subtrees expands from inner node of T_S over f_i, we can take the lost subtrees with roots in $c(\mathfrak{R})(f_i)$ and use them in $c(\mathfrak{R}_2)$, instead of the lost subtrees in $c(\mathfrak{R}_2)(f_i)$. Thus we obtain an extension of $\mathfrak{R}_2$ with strictly less cost than $c(\mathfrak{R}_2)$, a contradiction. This means that there is a chain $v_1, \ldots, v_m$ with described properties (v_1 is non-free, *etc.*).

Now, apply switch operation on v_i, v_{i+1}, for every $i = 1, \ldots, m-1$. In this way v_m, which is one of the trees $t_1, \ldots, t_{l+1}$, becomes non-free. The weight of $c(\mathfrak{R}_2)$ is not changed with these switch operations. Now, by removing v_m, we obtain an extension of $\mathfrak{R}_1$ with strictly less cost than $c(\mathfrak{R}_2)$, which contradicts the assumption $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1))$. Therefore $\omega(c(\mathfrak{R}_2)) = \omega(c(\mathfrak{R}_1)) + 1$. □

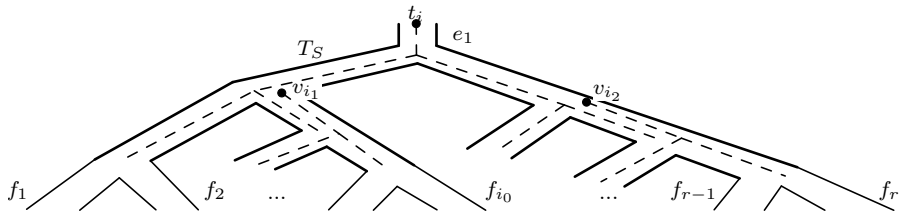

Fig. 7 Tree T_S is the maximum subtree of $S(e_1)$ rooted at e_1 that contains only free lost subtrees. There are no lost subtrees expanding from T_S over f_j. Black dots denote roots of lost subtrees (t_i, v_{i_1}, v_{i_2}). All non-free lost subtrees of $S(e_1)$ are in $S(f_j)$, $j = \overline{1,r}$

Let $e = (s, p(s)) \in E(S)$, and by $\mathfrak{L}'(e) = \mathfrak{L}'(e, \mathfrak{R}) = \mathfrak{L}'(s) = \mathfrak{L}'(s, \mathfrak{R})$ denote the number of non-free lost subtrees, in the reconciliation $\mathfrak{R}$, with a root assigned to $p(s) \in V(S)$, expanding over e.

Lemma 34 *Let $\mathfrak{R}$ be a reconciliation, $\mathfrak{R}_1$ output of* OneCompletion($\mathfrak{R}$)*, $e \in E(S)$, and e_l, e_r children of e. Then*

(a) $F(e, \mathfrak{R}) > 0 \implies \mathfrak{L}'(e, \mathfrak{R}_1) = 0$*;*
(b) $F(e, \mathfrak{R}) \leq 0 \implies \mathfrak{L}'(e, \mathfrak{R}_1) = \mathfrak{L}(e, \mathfrak{R}) - \mathfrak{D}(e, \mathfrak{R}) - max(min(F(e_l, \mathfrak{R}), F(e_r, \mathfrak{R})), 0)$*;*
(c) if $\mathfrak{R}_2$ is another output of OneCompletion($\mathfrak{R}$)*, then $\omega(\mathfrak{R}_1) = \omega(\mathfrak{R}_2)$.*

Proof Let $l = \mathfrak{L}(e, \mathfrak{R}), d = \mathfrak{D}(e, \mathfrak{R}), m = max(min(F(e_l, \mathfrak{R}), F(e_r, \mathfrak{R})), 0)$.

(a) Since OneCompletion extends losses only into edges e, if $F(e) > 0$, we have that the number of extra losses, expanded over e, is not greater than $F(e)$. Assume that $\mathfrak{R}_1$ generates extra f losses in e. Hence $f \leq F(e, \mathfrak{R}) = m + d - l$. Let f_m and l_m be the number of losses made free by extending over e_l, e_r, and f_d and l_d are number of losses made free by assigning them to the duplications in e. Hence $f_m + l_m \leq m$, $f_d + l_d \leq d$, $f_m + f_d \leq f$, $l_m + l_d \leq l$.

Assume the opposite, let $\mathfrak{L}'(e, \mathfrak{R}_1) > 0$. Then $f_m + f_d + l_m + l_d < f + l \leq d + m \implies f_d + l_d < d$ or $f_m + l_m < m$. Therefore one extra loss can be made free by assigning it to duplication in e, or extending it over e_r and e_l. This contradicts the procedure ExtendLossIntoFreeTree, which make loss free if $\Delta(e) \neq \emptyset$ or $F(e_1) > 0$ and $F(e_2) > 0$.

(b) Since $F(e, \mathfrak{R}) \leq 0$, $\mathfrak{R}_1$ does not extend any new losses over e, and $m + d \leq l$. At most m losses can be extended over e_r and e_l, and at most d losses can be assigned to the duplications in e. Therefore, number of losses that remained non-free is $l - d - m$.

(c) From (a) and (b), we have $\mathfrak{L}'(e, \mathfrak{R}_1) = \mathfrak{L}'(e, \mathfrak{R}_2)$, $\forall e \in E(S)$, hence $|\Lambda \backslash \Lambda'(\mathfrak{R}_1)| = |\Lambda \backslash \Lambda'(\mathfrak{R}_2)|$. Since ExtendLossIntoFreeTree does not create new duplications, we have $\Delta(\mathfrak{R}_1) = \Delta(\mathfrak{R}_2) = \Delta(\mathfrak{R})$. Therefore $\omega(\mathfrak{R}_1) = \omega(\mathfrak{R}_2)$. □

Lemma 35 *Let $\mathfrak{R}$ be a minimal reconciliation. Then* AllCompletions($\mathfrak{R}$) *returns a completion of $\mathfrak{R}$.*

Proof Let $\mathfrak{R}_1$ be a reconciliation from Lemma 33, obtained by applying switch operations on $c(\mathfrak{R})$. Then $\omega(\mathfrak{R}_1) = \omega(c(\mathfrak{R}))$ and $\mathfrak{R}_1$ satisfies the conditions

from Lemma 31. Hence $\mathfrak{R}_1$ is a possible output of the series of procedures OneCompletion($\mathfrak{R}$), ExtendLossesIntoNonFreeTrees($\mathfrak{R}$).

Let $\mathfrak{R}_2$ be another output of this series of procedures with the input $\mathfrak{R}$. From Lemmas 19 and 29 we have that $\mathfrak{R}_2$ is an extension of $\mathfrak{R}$. From Lemmas 34 (c) and 30 we have $\omega(\mathfrak{R}_1) = \omega(\mathfrak{R}_2)$. Since $\mathfrak{R}_1$ is a completion of $\mathfrak{R}$, we have $\mathfrak{R}_2$ is a completion of $\mathfrak{R}$.

Since Switch does not change the weight of a reconciliation (Lemma 32) and $\mathfrak{R}_2$ is a completion of $\mathfrak{R}$, we have that AllCompletions($\mathfrak{R}$) is also a completion of $\mathfrak{R}$. □

Theorem 3 *Algorithm* 12 *returns an optimal solution.*

Proof The algorithm starts with LCA reconciliation $\mathfrak{R}_1$. LCA's completion is an optimal reconciliation (Theorem 2), therefore completion of $\mathfrak{R}_1$ is an optimal reconciliation.

Let $\mathfrak{R}_2$ be an output of RaiseSeveralDuplications($\mathfrak{R}_1$). Then $c(\mathfrak{R}_2)$ is an optimal reconciliation (Lemma 24).

Let $\mathfrak{R}_3$ be an output of AllCompletions($\mathfrak{R}_2$). Then (Lemma 35) it is s completion of $\mathfrak{R}_2$, hence $\mathfrak{R}_3$ is an optimal reconciliation.

Assume that $\mathfrak{R}_4$ is an output of RaiseConversions($\mathfrak{R}_3$). From Lemma 27 we have $\omega(\mathfrak{R}_4) = \omega(\mathfrak{R}_3)$. Hence $\mathfrak{R}_4$ is an optimal reconciliation. Note that $\mathfrak{R}_4$ is an output of RandR(S, G, ϕ). □

Next lemma states that all duplications raised on a path going though a vertex with non positive flow on its children are conversions.

Lemma 36 *Let $\mathfrak{R}$ be a ZF reconciliation such that if v', v are non-free and free lost subtrees that overlap, then $v \leq v'$. Then $\mathfrak{R}$ is a possible output of* ExtendLossesIntoNonFreeTrees.

Proof From Lemma 25 we have that $\mathfrak{R}'$ is a possible output of RaiseDuplication, where $\mathfrak{R}'$ is the minimization of $\mathfrak{R}$. From Lemma 31 and and this Lemma condition, $\mathfrak{R}$ is a possible output of the series of procedures OneCompletion($\mathfrak{R}'$), ExtendLossesIntoNonFreeTrees($\mathfrak{R}'$). Hence $\mathfrak{R}$ is a possible output of ExtendLossesIntoNonFreeTrees. □

Lemma 37 *Let $\mathfrak{R}$ be a ZF reconciliation. Then $\mathfrak{R}$ is a possible output of* Switch.

Proof Let v' and v be non-free and free lost subtrees in $\mathfrak{R}$. If they overlap and $v' < v$, apply switch operation. Previous procedure repeat as long as there are such trees. Let us prove that the procedure will stop.

Let d be the sum of the distances of the roots of the non-free subtrees to $root(S)$. With every switch operation d decreases. Since $d \geq 0$, it cannot decrease indefinitely. Hence the procedure will stop.

The reconciliation, obtained in this way, denote by $\mathfrak{R}_1$. Now, $\mathfrak{R}_1$ satisfies the conditions in Lemma 36, hence it is a possible output of ExtendLossesIntoNonFreeTrees.

So, by ExtendLossesIntoNonFreeTrees we obtain $\mathfrak{R}_1$, and by Switch($\mathfrak{R}_1$), where switch operations are applied in the reversed order, we obtain $\mathfrak{R}$. □

Theorem 4 *Any optimal solution can be generated by Algorithm 12.*

Proof Let $\mathfrak{R}$ be an arbitrary optimal reconciliation. By lowering some conversions, we can obtain a ZF reconciliation $\mathfrak{R}_1$ such that $\omega(\mathfrak{R}_1) = \omega(\mathfrak{R})$ (see Lemma 28).

By Lemma 37, $\mathfrak{R}_1$ is obtainable by SWITCH.

So, $\mathfrak{R}_1$ is a possible output of SWITCH, and $\mathfrak{R}$ is a possible output of RAISECONVERSIONS($\mathfrak{R}_1$), if conversion raising is applied in the reversed order. □

Theorem 5 *Algorithm 12 has time complexity* $O(m^2 + m \cdot n)$.

Proof Let $n = |V(G)|$, $m = |V(S)|$, then $E(G) \in O(n), E(S) \in O(m)$. LCA reconciliation can be determined in linear time (see Chauve and El-Mabrouk (2009)), say $O(m + n)$.

Algorithm 2 forms a set $\Delta''(e)$ and it takes $O(m)$ time. It extends a loss into free tree. The maximum size of a (non-)free tree is $O(m)$. Algorithm 1 applies Algorithm 2 $|\Sigma \backslash \Sigma'| \leq |\Sigma|$ times, hence it has time complexity $O(|\Sigma| \cdot m)$.

Algorithm 5 determines possible new positions for a duplication d. Since the height of the tree S is $O(m)$, we have that the number of possible positions is also $O(m)$ and this is the complexity of Algorithm 5. Algorithm 4 calls Algorithm 5 and generates $k \in O(m)$ new losses. Hence the complexity of Algorithm 4 is $O(m)$. Algorithm 3 calls Algorithm 4 $|\Delta|$ times and its complexity is $O(|\Delta| \cdot m)$.

Algorithm 7 raises one conversion. Maximal raise height is $O(m)$ and this is the complexity of the algorithm. Algorithm 6 calls Algorithm 7 $|C|$ times (C - the set of all conversions). Therefore the complexity of Algorithm 6 is $O(|C| \cdot m)$.

Algorithm 10 extends a loss into a non-free tree. The size of non-free tree is $O(m)$ an this is the complexity of the algorithm. Algorithm 9 uses Algorithm 10 $|\Sigma_1|$ times, and its complexity is $O(|\Sigma_1| \cdot m)$.

Algorithm 11 applies a switch operation on lost subtrees. With every switch, a root of a subtree with non-free loss is further away from $root(S)$. Longest distance from $root(S)$ is $O(m)$. Switch operation always include one non-free loss. Therefore the complexity of this algorithm if $O(|\Sigma \backslash \Sigma'| \cdot m)$.

When we add corresponding complexities we get $O(m+n) + O(|\Sigma| \cdot m) + O(|\Delta| \cdot m) + O(|C| \cdot m) + O(|\Sigma_1| \cdot m) + O(|\Sigma \backslash \Sigma'| \cdot m)$. Since $|\Sigma|, |\Sigma_1|, |\Sigma \backslash \Sigma'| \in O(m+n)$, $|\Delta| \in O(n)$, we have that the complexity of the main algorithm is $O(m^2 + m \cdot n)$. □

6 Conclusion

In this paper we give a polynomial algorithm that returns an optimal reconciliation in duplication, loss, conversion model. The algorithm can return any optimal reconciliation with a non-zero probability, and can enumerate the whole space of solutions.

A natural extension would be a uniform sampling of all solutions in order to statistically assess properties of the solution space. Because of the switch operation, this could be achieved by an Markov chain Monte Carlo method. Future work is to define adequate transition probabilities to ascertain fast convergence.

An interesting problem that we leave open for further research is the weighted case. Unfortunately the approach, used in this paper, is not useful for this case. A completion of LCA reconciliation does not have to be an optimal reconciliation (see Figure 8). It might be necessary to raise some speciations from $V(G)$ in order to obtain an optimal solution.

Adding transfers and recombinations significantly increases the complexity of the problem.

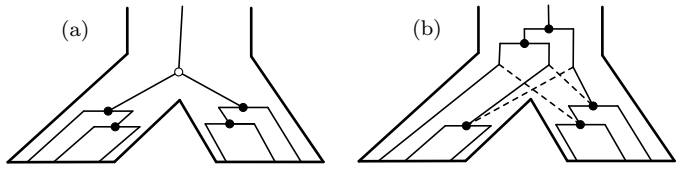

Fig. 8 Weighted case and $(d, l, c) = (2, 1, 1)$. (a) LCA reconciliation is equal to its completion (because there are no losses), and the weight is $4d = 8$. (b) The speciation and duplication are raised. Speciation is now duplication and three new losses are added. The weight is $2d + 3c = 7$

References

Arvestad L, Berglund AC, Lagergren J, Sennblad B (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In: Proc. eighth Annu. Int. Conf. Comput. Mol. Biol. - RECOMB '04. ACM Press, New York, New York, USA. pp 326–335. doi: 10.1145/974614.974657

Bourgon R, Delorenzi M, Sargeant T, Hodder AN, Crabb BS, Speed TP (2004) The serine repeat antigen (SERA) gene family phylogeny in Plasmodium: the impact of GC content and reconciliation of gene and species trees. Mol Biol Evol 21(11):2161–2171. doi: 10.1093/molbev/msh228

Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V (2013) Genome-scale coestimation of species and gene trees. Genome Res 23:323–330. doi: 10.1101/gr.141978.112

Brooks DR, Ferrao AL (2005) The historical biogeography of co-evolution: emerging infectious diseases are evolutionary accidents waiting to happen. J Biogeogr 32(8):1291–1299. doi: 10.1111/j.1365-2699.2005.01315.x

Chan Yb, Ranwez V, Scornavacca C (2015) Exploring the space of gene/species reconciliations with transfers. J Math Biol 71(5):1179–1209. doi: 10.1007/s00285-014-0851-2

Chan Yb, Ranwez V, Scornavacca C (2017) Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. J Theor Biol 432:1 – 13. doi: 10.1016/j.jtbi.2017.08.008

Chang WC, Eulenstein O (2006) Computing and Combinatorics. Lecture Notes in Computer Science, vol 4112. Springer Berlin Heidelberg, Berlin, Heidelberg. doi: 10.1007/11809678

Chauve C, El-Mabrouk N (2009) New perspectives on gene family evolution: Losses in reconciliation and a link with supertrees. In: Batzoglou S (ed) Res. Comput. Mol. Biol.. Springer Berlin Heidelberg, Berlin, Heidelberg. pp 46–58. doi: 10/dxfx65

Chauve C, Doyon JP, El-Mabrouk N (2008) Gene family evolution by duplication, speciation, and loss. J Comput Biol 15(8):1043–1062. doi: 10.1089/cmb.2008.0054

Chen JM, Cooper DN, Chuzhanova N, Frec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8:762–775. doi: 10.1038/nrg2193

Doyon JP, Scornavacca C, Gorbunov KY, Szöllősi GJ, Ranwez V, Berry V (2010) An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: Tannier E (ed) Comparative Genomics: International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg. pp 93–108. doi: 10.1007/978-3-642-16181-0_9

Doyon JP, Ranwez V, Daubin V, Berry V (2011) Models, algorithms and programs for phylogeny reconciliation. Brief Bioinformatics 12(5):392–400. doi: 10.1093/bib/bbr045

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21(11):2596–2603. doi: 10.1093/bioinformatics/bti325

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates

Górecki P, Tiuryn J (2006) DLS-trees: A model of evolutionary scenarios. Theor Comput Sci 359(1-3):378–399. doi: 10.1016/j.tcs.2006.05.019

Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ (2017) Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun 8:14,319. doi: 10.1038/ncomms14319

Hasić D, Tannier E (2017) Gene tree reconciliation including transfers with replacement is hard and FPT. submitted

van der Heijden RT, Snel B, van Noort V, Huynen MA (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics 8:83. doi: 10.1186/1471-2105-8-83

Hsu CH, Zhang Y, Hardison RC, Program NCS, Green ED, Miller W (2010) An effective method for detecting gene conversion events in whole genomes. J Comput Biol 17:1281–1297. doi: 10.1089/cmb.2010.0103

Hu F, Lin Y, Tang J (2014) MLGO: phylogeny reconstruction and ancestral inference from gene-order data. BMC Bioinformatics 15:354. doi: 10.1186/s12859-014-0354-6

Kejnovsky E, Hobza R, Kubat Z, Widmer A, Marais GAB, Vyskot B (2007) High intrachromosomal similarity of retrotransposon long terminal repeats: evidence for homogenization by gene conversion on plant sex chromosomes? Gene 390:92–97. doi: 10.1016/j.gene.2006.10.007

Ko WY, Kaercher KA, Giombini E, Marcatili P, Froment A, Ibrahim M, Lema G, Nyambo TB, Omar SA, Wambebe C, Ranciaro A, Hirbo JB, Tishkoff SA (2011) Effects of natural selection and gene conversion on the evolution of human glycophorins coding for mns blood polymorphisms in malaria-endemic african populations. Am J Hum Genet 88:741–754. doi: 10.1016/j.ajhg.2011.05.005

Lafond M, Swenson KM, El-Mabrouk N (2012) An optimal reconciliation algorithm for gene trees with polytomies. In: Algorithms in Bioinformatics: 12th International Workshop, WABI 2012, Ljubljana, Slovenia, September 10-12, 2012. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg. pp 106–122. doi: 10.1007/978-3-642-33122-0_9

Mansai SP, Innan H (2010) The power of the methods for detecting interlocus gene conversion. Genetics 184:517–527. doi: 10.1534/genetics.109.111161

Matassi G (2017) Horizontal gene transfer drives the evolution of Rh50 permeases in prokaryotes. BMC Evol Biol 17(1):2. doi: 10.1186/s12862-016-0850-6

Mirarab S, Bayzid MS, Boussau B, Warnow T (2014) Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science (New York, NY) 346:1250,463. doi: 10.1126/science.1250463

Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol (Amst) 28(12):719–728. doi: 10.1016/j.tree.2013.09.004

Page RD, Charleston MA (1998) Trees within trees: phylogeny and historical associations. Trends Ecol Evol (Amst) 13(9):356–359. doi: 10.1016/S0169-5347(98)01438-4

Planet PJ, Kachlany SC, Fine DH, DeSalle R, Figurski DH (2003) The widespread colonization island of actinobacillus actinomycetemcomitans. Nat Genet 34(2):193–198. doi: 10.1038/ng1154

Ranwez V, Scornavacca C, Doyon JP, Berry V (2016) Inferring gene duplications, transfers and losses can be done in a discrete framework. J Math Biol 72(7):1811–1844. doi: 10.1007/s00285-015-0930-z

Rasmussen MD, Kellis M (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res 22:755–765. doi: 10.1101/gr.123901.111

Searls DB (2003) Pharmacophylogenomics: genes, evolution and drug targets. Nat Rev Drug Discov 2(8):613–623. doi: 10.1038/nrd1152

Storm CE, Sonnhammer EL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18(1):92–99. doi: 10.1093/bioinformatics/18.1.92

Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci USA 109(43):17,513–17,518. doi: 10.1073/pnas.1202997109

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V (2013a) Efficient exploration of the space of reconciled gene trees. Syst Biol 62(6):901–912. doi: 10.1093/sysbio/syt054

Szöllősi GJ, Tannier E, Lartillot N, Daubin V (2013b) Lateral gene transfer from the dead. Syst Biol 62(3):386–397. doi: 10.1093/sysbio/syt003

Szöllősi GJ, Tannier E, Daubin V, Boussau B (2015) The inference of gene trees with species trees. Syst Biol 64(1):42–62. doi: 10.1093/sysbio/syu048

Tofigh A, Hallett M, Lagergren J (2011) Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans Comput Biol Bioinform 8(2):517–535. doi: 10.1109/TCBB.2010.14

Vanhove MPM, Pariselle A, Van Steenberge M, Raeymaekers JAM, Hablützel PI, Gillardin C, Hellemans B, Breman FC, Koblmüller S, Sturmbauer C, Snoeks J, Volckaert FAM, Huyse T (2015) Hidden biodiversity in an ancient lake: phylogenetic congruence between lake tanganyika tropheine cichlids and their monogenean flatworm parasites. Sci Rep 5:13,669. doi: 10.1038/srep13669

Vernot B, Stolzer M, Goldman A, Durand D (2008) Reconciliation with non-binary species trees. J Comput Biol 15(8):981–1006. doi: 10.1089/cmb.2008.0092

Zheng Y, Zhang L (2017) Reconciliation With Nonbinary Gene Trees Revisited. J ACM 64(4):1–28. doi: 10.1145/3088512