Abstract—The supertree problem asking for a tree displaying a set of consistent input trees has been largely considered for the reconstruction of species trees. Here, we rather explore this framework for the sake of reconstructing a gene tree from a set of input gene trees on partial data. In this perspective, the phylogenetic tree for the species containing the genes of interest can be used to choose among the many possible compatible “supergenetrees”, the most natural criteria being to minimize a reconciliation cost. We develop a variety of algorithmic solutions for the construction and correction of gene trees using the supertree framework. A dynamic programming supertree algorithm for constructing or correcting gene trees, exponential in the number of input trees, is first developed for the less constrained version of the problem. It is then adapted to gene trees with nodes labeled as duplication or speciation, the additional constraint being to preserve the orthology and paralogy relations between genes. Then, a quadratic time algorithm is developed for efficiently correcting an initial gene tree while preserving a set of “trusted” subtrees, as well as the relative phylogenetic distance between them, in both cases of labeled or unlabeled input trees. By applying these algorithms to the set of Ensembl gene trees, we show that this new correction framework is particularly useful to correct weakly-supported duplication nodes. The C++ source code for the algorithms and simulations described in the paper are available at https://github.com/UdeM-LBIT/SuGeT.

1. Introduction

The supertree problem consists in combining a set of input phylogenetic trees on possibly overlapping sets of data, into a single one for the whole set (see for example [1], [2], [3], [4], [5], [6], [7]). Ideally, the obtained tree should display each of the input trees, which is only possible if they are “consistent” i.e. if they do not contain conflicting phylogenetic information. The simplest formulation of the supertree problem is therefore to state whether an input set of trees is consistent, and if so, find a “compatible” tree, called a supertree, displaying them all. This problem is NP-complete for unrooted trees [8], [9], but solvable in polynomial time for rooted trees [10], [11], [12], [13]. However, even for rooted trees the set of all possible supertrees may be exponential in the number of genes.

Supertree methods have been mainly designed to reconstruct a species tree from gene trees obtained for various gene families. However, they can have applications for gene tree reconstruction as well. Indeed, they may be used to combine partial trees on overlapping gene sets available from various sources (various databases, various reconstruction tools, etc). Alternatively, in the case of large gene families, they may be used to combine gene trees for smaller sets of orthologs, usually obtained from clustering algorithms such as OrthoMCL [14], InParanoid [15] or Proteinortho [16]. In such a case, ideally, orthology relations should be preserved in the final tree. More generally, given a set of input “labeled gene trees”, i.e. gene trees with internal nodes labeled as duplication or speciation, we may be interested in a supertree preserving this labeling. As far as we know, no automated method accounting for labeling constraints has never been proposed. Here, we consider the problem of reconstructing a “supergenetree” in both cases of a labeled or unlabeled set of input gene trees.

In this paper, we also show that the supertree principle can be used for gene tree correction. For various reasons related to the considered model, method or data, gene trees can contain many errors (see for example [17] for a link with dubious high duplication nodes), and trees frequently exhibit branches with low statistical support. Two main approaches exist to correct gene trees, based on a local exploration principle to identify closely related trees that might have a better statistical support [18], a better reconciliation cost [19], [20], [21] or a combination of both [22], [23]. In the present work, we consider the second approach, based on the reconciliation cost with a given species tree. A way of correcting a gene tree is to remove weakly-supported branches, leading
to a set of subtrees, that should then be merged into a new one, according to some criterion. The most commonly considered criterion is a best fit with the species tree. A simple way is to consider the set of subtrees as the leaves of a polytomy (star-tree), and to resolve the polytomy in a way minimizing the reconciliation cost with the species tree (see NOTUNG [19], the Zheng and Zhang algorithm [20], PolytomySolver [24]). Such a correction method, not only preserves the input subtrees, but also the gene clusters inside the subtrees. In other words, the exhibited monophyly of input gene clusters is not challenged by a polytomy resolution method. However, it has been shown that genes under negative selection, while exhibiting the true topology, may be wrongly grouped into monophyletic groups (see for example [21], [25], [26], [27]). In this perspective, using a supertree method may be beneficial, as it preserves the topology of subtrees while allowing to group genes from different subtrees.

In [28], we introduced under the name of Minimum SuperGeneTree (MinSGT) the problem of finding, for a set of gene trees, a supertree that minimizes the reconciliation cost with a given species tree. Under the duplication cost, we have shown that this problem is NP-hard to approximate within a $n^{1-\epsilon}$ factor, for any $0 < \epsilon < 1$, even for instances in which there is only one gene per species in the input trees, and even if each gene appears in at most one input tree. In this paper, we carry out on MinSGT but for the more general reconciliation cost. Although NP-hardness proofs for the duplication cost do not apply to the duplication plus loss cost, the problem is conjectured NP-hard for this more general reconciliation cost, as accounting for losses in addition to duplications is unlikely to make the problem simpler. Given a set of consistent input gene trees, we provide various algorithmic results depending on the additional information we have on the trees.

In Section 3, we first exhibit a dynamic programming algorithm for the general case, exponential in the number of input trees. We show how this algorithm can be adapted to compute a supertree preserving the input trees labeling, as motivated above. In Section 4, we then consider the correction problem with as input a gene tree together with a set of subtrees which topology should be preserved in the final supertree. To avoid having a supertree grouping genes that are far apart in the original tree, the relative phylogenetic distance between gene clusters is considered as an additional constraint. Inspired by the Minimum Triplet Respecting History introduced in [21], we define the Minimum Triplet Respecting SuperGeneTree Problem asking for a supertree displaying all input subtrees, while preserving the topology of any triplet of genes taken from three different subtrees. We develop a quadratic-time algorithm for this problem. Finally, in Section 5, by applying these algorithms to a set of a few hundreds Ensembl vertebrate gene trees, we show that this new correction framework is particularly useful to correct weakly-supported upper duplication nodes, as we observe that the correction carried out by our algorithms often improves significantly the likelihood scores.

2. Preliminaries

All considered trees are rooted and binary. We denote by $r(T)$ the root, by $V(T)$ the set of nodes, and by $\mathcal{L}(T) \subseteq V(T)$ the leafset of a tree $T$. We say that $T$ is a tree for $L = \mathcal{L}(T)$. Given a node $x$ of $T$, the subtree of $T$ rooted at $x$ is denoted $T[x]$. When there is no ambiguity on the considered tree, we simply write $\mathcal{L}(x)$ instead of $\mathcal{L}(T[x])$.

We arbitrarily set one of the two children of an internal node $x$ as the left child $x_1$ and the other as the right child $x_r$, and denote by $(\mathcal{L}(x_1), \mathcal{L}(x_r))$ the bipartition induced by $x$. Also for the sake of simplicity, we just denote by $T_i$ and $T_r$ the left and right subtrees of the root of $T$. A node $x$ is an ancestor of a node $y$ if $x$ is on the path between $y$ and $r(T)$. If $x$ is an ancestor of $y$, $\text{inter}(x, y)$ is the number of nodes located on the path between $x$ and $y$, excluding $x$ and $y$. Two nodes $x$ and $y$ are separated in $T$ iff none is an ancestor of the other. In this case, we also say that the two subtrees $T[x], T[y]$ of $T$ are separated.

The lowest common ancestor (lca) of $L' \subset \mathcal{L}(T)$, denoted $\text{lca}_T(L')$, is the ancestor common to all leaves in $L'$ that is the most distant from the root. $T|_{L'}$ is the tree with leafset $L'$ obtained from the subtree of $T$ rooted at $\text{lca}_T(L')$ by removing all leaves that are not in $L'$, and then all internal nodes of degree 2, except the root. Let $T'$ be a tree such that $\mathcal{L}(T') = L' \subseteq \mathcal{L}(T)$. We say that $T$ displays $T'$ iff $T|_{L'}$ is isomorphic to $T'$ while preserving the same leaf-labeling.

Gene and species trees. A species tree $S$ for a set $\Sigma$ of species represents an ordered set of speciation events that have led to $\Sigma$. A gene family is a set of genes $\Gamma$ accompanied with a mapping function $s : \Gamma \to \Sigma$ mapping each gene to its corresponding species. Consider a gene family $\Gamma$ where each gene $x \in \Gamma$ belongs to a species $s(x)$ of $\Sigma$. The evolutionary history of $\Gamma$ can be represented as a gene tree $G$ for $\Gamma$. For example, in Figure 1, $G$ is a gene tree for $\Gamma = \{s_1, s_2, b_1, b_2, h_1, h_2, h_3, m_3, r_3\}$. Each internal node of $G$ refers to an ancestral gene at the moment of an event, either speciation ($\text{Spec}$) or duplication ($\text{Dup}$). The mapping function $s$ is generalized as follows: if $x$ is an internal node of $G$, then $s(x) = \text{lca}_S(s(x') : x' \in \mathcal{L}(x))$.

When the type of event is known for each internal node, the gene tree $G$ is said labeled. Formally, a labeled gene tree for $\Gamma$ is a pair $(G, \text{ev}_G)$, where $G$ is a tree for $\mathcal{L}(G) = \Gamma$, and $\text{ev}_G : V(G) \setminus \mathcal{L}(G) \to \{\text{Dup}, \text{Spec}\}$ is a function labeling each internal node of $G$ as a duplication or a speciation node.

According to the Fitch [29] terminology, given a labeled gene tree $(G, \text{ev}_G)$, we say that two genes $x, y$ are orthologs if $\text{ev}_G(\text{lca}_G(x, y)) = \text{Spec}$, and paralogs if $\text{ev}_G(\text{lca}_G(x, y)) = \text{Dup}$. For example, from the set of labeled gene trees in Figure 1, $s_1, h_3$ are orthologs while $s_1, h_2$ are paralogs.

While a history for $\Gamma$ can be represented as a labeled gene tree, the converse is not always true, as a labeled tree $(G, \text{ev}_G)$ for $\Gamma$ does not necessarily represent a valid history in agreement with a species tree $S$. For this to hold,
The number of duplications required to explain the evolution in the general case, and as its LCA-reconciliation cost underlined by a labeled gene tree as its reconciliation cost history. We refer to the number of duplications and losses underlined by such an evolutionary allows to further recover, in linear time, the minimum $G$ of $(\Gamma, G)$. The consistency problem of rooted trees for possibly overlapping subsets of $\Gamma$ is satisfiable, i.e. if there is a labeled tree $(G, ev_G)$ displaying the relations induced by all the input trees, and if there is such a tree which is $S$-consistent. Satisfiability is a well-studied problem. It reduces to verifying if a relation graph $R$ (vertices are genes and edges link orthologous genes) is $P_3$-free, i.e. no four vertices of $R$ induce a path of length 3. On the other hand, a cubic-time algorithm was developed in [30] for deciding whether a set of relations is $S$-consistent. Hereafter, we assume that the relations induced


We begin with the less constrained version of the problem. Given a set $G$ of consistent input gene trees, we ask for a compatible tree, also called supergartenetree $G$ for $G$, i.e. a tree displaying each tree of $G$. In addition, among all supergartenetrees for $G$, $G$ should be of minimum LCA-reconciliation cost (see $G$ in Figure 1).

MINIMUM SUPERGENETREE (MINSGT) PROBLEM:

Input: A species set $\Sigma$ and a species tree $S$ for $\Sigma$; a gene family $\Gamma$ of size $n$, a set $\Gamma_i$, $1 \leq i \leq k$ of subsets of $\Gamma$ such that $\bigcup_{i=1}^{k} \Gamma_i = \Gamma$, and a consistent set $G = \{G_1, G_2, \ldots, G_k\}$ of gene trees such that, for each $1 \leq i \leq k$, $G_i$ is a tree for $\Gamma_i$.

Output: Among all trees $G$ for $\Gamma$ compatible with $G$, one of minimum LCA-reconciliation cost.

Suppose now that the input trees are labeled, and consider this labeling as an additional constraint. The problem becomes one of finding a labeled supergartenetree preserving the input gene trees node labeling. As a labeled gene tree induces a full orthology and paralogy relation on the set of its leaves, this is possible only if the set of relations is satisfiable, i.e. if there is a labeled tree $(G, ev_G)$ displaying the relations induced by all the input trees, and if there is such a tree which is $S$-consistent. Satisfiability is a well-studied problem. It reduces to verifying if a relation graph $R$ (vertices are genes and edges link orthologous genes) is $P_3$-free, i.e. no four vertices of $R$ induce a path of length 3. On the other hand, a cubic-time algorithm was developed in [30] for deciding whether a set of relations is $S$-consistent. Hereafter, we assume that the relations induced
by the input trees are satisfiable and $S$-consistent.

Let $G$ and $G'$ be two trees with $\mathcal{L}(G') \subseteq \mathcal{L}(G)$ such that $G$ displays $G'$. Then $(G, ev_G)$ is said label-compatible with $(G', ev_{G'})$ iff, for any internal node $x$ of $G$ and $x'$ of $G'$ such that $x = lca_G(x')$, $ev_G(x) = ev_{G'}(x')$. A labeled supergenetree $G$ for a set $\mathcal{G}$ of trees is said label-compatible with $\mathcal{G}$ iff it is label-compatible with each of the labeled trees of $\mathcal{G}$. An illustration is provided by the supergenetree $G'$ in Figure 1. We are now ready to formulate our second problem.

**Minimum Labeled SuperGeneTree (MinLSGT) Problem:**

**Input:** A species set $\Sigma$ and a species tree $S$ for $\Sigma$; a gene family $\Gamma$ of size $n$, a set $\Gamma_i, 1 \leq i \leq k$ of subsets of $\Gamma$ such that $\bigcup_{i=1}^{k} \Gamma_i = \Gamma$, and a consistent set $\mathcal{G} = \{(G_1, ev_1), (G_2, ev_2), \ldots, (G_k, ev_k)\}$ of satisfiable and $S$-consistent labeled gene trees where, for each $1 \leq i \leq k$, $G_i$ is a tree for $\Gamma_i$.

**Output:** Among all labeled supergenetrees $(G, ev_G)$ for $\Gamma$ label-compatible with $\mathcal{G}$, one of minimum reconciliation cost.

The MinSGT and MinLSGT problems for the duplication cost were both shown NP-Hard in [28], even in the case where no two input trees have a gene in common and the trees only contain speciations.

3.1. The MinSGT problem

We describe a dynamic programming algorithm for the MinSGT problem leading to the following result.

**Theorem 1.** The MinSGT problem can be solved in $O((n + 1)^k \times 4^k \times k)$ time complexity.

The algorithm constructs the supergenetree $G$ from the root to the leaves. At each step, i.e. for each internal node $x$ being constructed in $G$, all possible reconciliations ($\mathcal{L}(x_l), \mathcal{L}(x_r)$) that could be induced by $x$ are tried, and the iteration continues on each of $\mathcal{L}(x_l)$ and $\mathcal{L}(x_r)$. For example, at the root, the goal is to find the best bipartition of $\Gamma$, i.e. the one leading to the minimum LCA-reconciliation cost. At each step, this cost is computed from a local reconciliation cost at $x$ (as defined in Lemma 1), and from the best reconciliation cost of the two created clusters. A key observation is that the constraint of being compatible with the input gene trees induces a strong constraint on the reconciliations, hence only a subset of the bipartition set has to be tested at each step.

First, a formulation of the reconciliation cost in terms of the sum of local reconciliation costs at each internal node $x$ is given. The next lemma is a reformulation of the reconciliation cost, as described in many papers [19], [32].

**Lemma 1.** The LCA-reconciliation cost of a gene tree $G$ is the sum of local LCA-reconciliation costs $cost(L_l, L_r)$ for all internal nodes $x$ of $G$, where $L = \mathcal{L}(x)$, and $(L_l, L_r) = (\mathcal{L}(x_l), \mathcal{L}(x_r))$, and $cost(L_l, L_r)$ equals to:

- $\text{inter}(s(L), s(L_l)) + \text{inter}(s(L), s(L_r))$ if $s(L) \neq s(L_l)$ and $s(L) \neq s(L_r)$;
- $1 + \text{inter}(s(L), s(L_l)) + \text{inter}(s(L), s(L_r))$ if $s(L) = s(L_l)$ and $s(L) = s(L_r)$;
- $2 + \text{inter}(s(L), s(L_l)) + \text{inter}(s(L), s(L_r))$ if $s(L) = s(L_l)$ and $s(L) \neq s(L_r)$ or conversely.

The node $x = (L_l, L_r)$ is a speciation node in the first case, and a duplication node in the two last cases (thus adding 1 duplication to the LCA-reconciliation cost, plus 1 loss in the third case). Note that $\text{inter}(s, t) = 0$ if $s = t$.

For example, the root of $G$ in Figure 1 fulfills the conditions of the first case, and thus it is a speciation node, whereas the root of $G'$ fulfills the condition of the third case.

Lemma 1 allows to recursively compute a minimum LCA-reconciliation cost supergenetree, by exploring, for each node $x$ from the root to the leaves, all “valid” bipartitions of $\mathcal{L}(x)$, remaining to be characterized formally. In the following, we define the properties of a bipartition $(L_l, L_r)$ induced by the root of a supergenetree $G$. It directly follows from the definition of a supergenetree that should display each individual gene tree.

**Property 1.** Let $\mathcal{G} = \{G_1, \ldots, G_k\}$ be a set of gene trees. The root of a supergenetree $G$ compatible with $\mathcal{G}$ subdivides $\bigcup_{i=1}^{k} \mathcal{L}(G_i)$ into a compatible bipartition $(L_l, L_r)$, i.e. a bipartition such that, for each $i$ s.t. $1 \leq i \leq k$, either: 1) $\mathcal{L}(G_i) \subseteq L_l$; or 2) $\mathcal{L}(G_i) \subseteq L_r$; or 3) $\mathcal{L}(G_i) \subseteq L_l$ and $\mathcal{L}(G_i) \subseteq L_r$; or 4) $\mathcal{L}(G_i) \subseteq L_l$ and $\mathcal{L}(G_i) \subseteq L_r$.

For example, the root of the supergenetree $G$ in Figure 1 satisfies the third condition for $G_1$, $G_2$ and $G_3$, and the second for $G_4$.

$B(G_1, \ldots, G_k)$ denotes the set of all bipartitions of $\bigcup_{i=1}^{k} \mathcal{L}(G_i)$ compatible with $\mathcal{G}$. For example, the two bipartitions defined by the roots of $G$ and $G'$ in Figure 1 are both compatible with the given set of gene trees. Figure 2 illustrates the set of all valid bipartitions compatible with two given trees.

**Lemma 2.** $|B(G_1, \ldots, G_k)| \leq \left(\frac{4^k}{2^k} - 1\right)$.

**Proof.** For each tree $G_i$, there are four possibilities for placing $\mathcal{L}(G_i)$ and $\mathcal{L}(G_i)$ in a bipartition $(L_l, L_r)$: either they are both in $L_l$, or both in $L_r$, or one in $L_l$ and the other in $L_r$. Therefore, $4^k$ distributions of left and right subtrees of the $k$ trees in $(L_l, L_r)$. However, as the left and right characterization of nodes is arbitrary, each distribution is counted twice, and thus the total number of different bipartitions is $\frac{4^k}{2^k}$. One of these bipartitions has a part that is empty. We discard it and the total number is then $\frac{4^k}{2^k} - 1$. However, a set $(L_l, L_r)$ obtained from such distribution of the $G_i$ subtrees is not necessarily a bipartition, as a same gene can be present in two different input trees, and end up placed in both $L_l$ and $L_r$. Therefore, $\left(\frac{4^k}{2^k} - 1\right)$ is only an upper bound of the number of compatible bipartitions. □

The constructive proof of Lemma 2 induces an algorithm for enumerating the members of $B(G_1, \ldots, G_k)$, which is
illustrated in Figure 2 for the case of two trees. Intuitively, to construct a bipartition \((L_i, L_r)\), each tree \(G_i\) of \(\mathcal{G}\) can choose to “send” in \(L_i\) either its left subtree \(G_{i,l}\), its right subtree \(G_{i,r}\), the whole tree \(G_i\) or nothing at all. What has not been sent in \(L_i\) is sent in \(L_r\). Then \(\mathcal{B}(G_1, \ldots, G_k)\) is the set of all possible combinations of choices. However, not every bipartition constructed in this manner yields a valid bipartition. For instance in Figure 2, the top-left bipartition cannot be valid if \(G_1\) and \(G_2\) share a leaf with the same label, as a gene cannot be sent both left and right. These cases, however, can be detected easily by verifying the sizes of \(L_i\) and \(L_r\).

We are now ready to give the main recurrence formula of our dynamic programming algorithm. Denote by \(\text{MinSGT}(G_1, \ldots, G_k)\) the minimum LCA-reconciliation cost of a supertree compatible with \(\mathcal{G} = \{G_1, \ldots, G_k\}\). The next lemma directly follows from Lemma 1 and Property 1.

**Lemma 3.** Let \(\mathcal{G} = \{G_1, \ldots, G_k\}\) be a set of gene trees.

1. \(\text{MinSGT}(G_1, \ldots, G_k) = 0\) if \(|\bigcup_{i=1}^k \mathcal{L}(G_i)| = 1\) (Stop condition);
2. Otherwise, \(\text{MinSGT}(G_1, \ldots, G_k) = \min_{(L_i, L_r) \in \mathcal{B}(G_1, \ldots, G_k)} \left\{ \frac{\text{cost}(L_i, L_r)}{|\mathcal{L}(G_i)|} + \frac{\text{MinSGT}(G_{1|L_i}, \ldots, G_{k|L_r})}{|\mathcal{L}(G_i)|} : \text{MinSGT}(G_{1|L_i}, \ldots, G_{k|L_r}) \right\} \)

Note that, given a bipartition \((L_i, L_r) \in \mathcal{B}(G_1, \ldots, G_k)\), for each \(i\) such that \(1 \leq i \leq k\), \(G_{i|L_i}\) and \(G_{i|L_r}\) are equal either to \(\emptyset\) or \(G_{i,l}\) or \(G_{i,r}\). Thus, \(G_{i|L_i}\) and \(G_{i|L_r}\) are either empty trees or complete subtrees of \(G_i\).

Note also that, at each step, the existence of a compatible bipartition follows from the fact that the input gene trees are assumed to be consistent, as stated in the formulation of the \(\text{MinSGT}\) problem. In the absence of this assumption, we have to add a third equation to Lemma 3: If \(|\bigcup_{i=1}^k \mathcal{L}(G_i)| > 1\) and \(|\mathcal{B}(G_1, \ldots, G_k)| = 0\), \(\text{MinSGT}(G_1, \ldots, G_k) = +\infty\).

**Complexity.** We now address the complexity of the dynamic programming algorithm defined by the recurrences of Lemma 3. Each call to the recursive procedure \(\text{MinSGT}\) receives as input at most one subtree from each tree \(G_i\). Let \(n\) be the maximum number of node in a tree \(G_i\). As each tree has at most \(n\) possible subtrees, there are at most \((n+1)^k\) possible calls to \(\text{MinSGT}\). Next, for any set of gene trees \(\{G_1, \ldots, G_k\}\), the number of distinct bipartitions \((L_i, L_r) \in \mathcal{B}(G_1, \ldots, G_k)\) to be tested is at most \(\frac{k^2}{2} - 1\) (Lemma 2). Finally, the value of \(\text{cost}(L_i, L_r)\) can be computed in time \(O(k)\) provided that the mapping \(s\) is precomputed for all nodes of the trees \(G_1, \ldots, G_k\), and \(\text{lea}(x, y)\) and \(\text{inter}(x, y)\) are precomputed for any pair \((x, y)\) of nodes in \(S\). The time complexity of the overall algorithm is therefore \(O((n+1)^k \times 4^k \times k)\), which completes the proof of Theorem 1.

### 3.2. The MinLSGT Problem

The algorithm for the \(\text{MinSGT}\) problem can be adapted to solve the MinLSGT problem, leading to the following result.

**Corollary 1.** The MinLSGT problem can be solved in \(O((n+1)^k \times 4^k \times k)\) time complexity.

The intuition behind the MinLSGT algorithm is quite simple. We proceed as in the \(\text{MinSGT}\) algorithm, but each time a bipartition \((L_i, L_r)\) is considered, we verify whether the root of a tree separating \(L_i\) and \(L_r\) should be a speciation or a duplication. If there are two genes \(g_l \in L_i\) and \(g_r \in L_r\) that disagree with this event, we treat \((L_i, L_r)\) as an invalid bipartition and do not consider it further.

Before describing this adaptation, we need few additional definitions and properties. Given a set of labeled gene trees \(\mathcal{G} = \{(G_1, ev(G_1)), \ldots, (G_k, ev(G_k))\}\) and a bipartition \((L_i, L_r) \in \mathcal{B}(G_1, \ldots, G_k)\), for any \(i\) s.t. \(1 \leq i \leq k\), we say that \(G_i\) is separated by \((L_i, L_r)\) if \(G_i\) satisfies the third or fourth condition of Property 1. We denote by \(\mathcal{G}(L_i, L_r)\)
the set of gene trees $G_i$, $1 \leq i \leq k$, that are separated by $(L_i, L_r)$.

**Lemma 4.** Let $G = \{(G_1, ev_{G_1}), \ldots, (G_k, ev_{G_k})\}$ be a set of labeled gene trees. Then, for any labeled supergenetree $(G, ev_G)$ label-compatible with $G$, the label $ev_G(x)$ of its root $x$ equals the label of the root of any gene tree $G_i$, $1 \leq i \leq k$, such that $G_i \in G(\mathcal{L}(x_l), \mathcal{L}(x_r))$.

**Proof.** Let $G_i$, $1 \leq i \leq k$ be a genetree such that $G_i \in G(\mathcal{L}(x_l), \mathcal{L}(x_r))$, and let $x_i$ be the root of $G_i$. Then $\text{lca}_G(\mathcal{L}(x_i)) = x$, and thus by definition of the label-compatibility of $G$ with $G_i$, we have $ev_G(x) = ev_{G_i}(x_i)$. □

From Lemma 4, we define a bipartition of $\bigcup_{i=1}^k S(G_i)$ label-compatible with $G$ as follows.

**Definition 1.** Let $G = \{(G_1, ev_{G_1}), \ldots, (G_k, ev_{G_k})\}$ be a set of labeled gene trees. A bipartition $(L_i, L_r)$ of $\bigcup_{i=1}^k S(G_i)$ is label-compatible with $G$ if it is compatible with $G$ and verifies:

1. if $| G(L_i, L_r) | > 0$, the roots of all gene trees in $G(L_i, L_r)$ have the same label denoted by $ev_G(L_i, L_r)$.
2. if $| G(L_i, L_r) | > 0$ and $ev_G(L_i, L_r) = \text{Spec}$, then $\text{lca}_G(\{|s(x) : x \in L_i \cup L_r\}) \neq \text{lca}_G(\{|s(x) : x \in L_i\})$ and $\text{lca}_G(\{|s(x) : x \in L_i \cup L_r\}) \neq \text{lca}_G(\{|s(x) : x \in L_i\})$.

For example, the bipartition determined by the root of the supergenetree $G$ ($\{s_1, s_2, b_1, b_2, h_1, h_2, h_3, m_3, r_3\}$) in Figure 1 is not label-compatible with the set of gene trees, as it separates both $G_1$ and $G_2$ which do not have the same root label.

The $\text{MinLSGT}$ algorithm for solving the $\text{MinSGT}$ problem is based on the same general dynamic programming framework as the $\text{MinSGT}$ algorithm: at each step, iterate over all possible bipartitions, and then proceed recursively for each partition. The two differences are: (1) given a set of labeled gene trees $G = \{(G_1, ev_{G_1}), \ldots, (G_k, ev_{G_k})\}$, we only test a subset of compatible bipartitions of $B(G_1, \ldots, G_k)$ that are label-compatible with $G$; (2) computing local reconciliation costs should not be done on the basis of the LCA-reconciliation, as some nodes that would be labeled as speciation nodes from the LCA-mapping should rather be duplication nodes in order to be label-compatible with some input gene trees. For example, in Figure 1, $\text{lca}_G(\{|s_2, b_2\})$ would be labeled $\text{Spec}$ by the LCA-mapping. However, it should be labeled $\text{Dup}$ to be label-compatible with $G_3$. The following Lemma is required, in place of Lemma 1.

**Lemma 5.** Let $x$ be an internal node of a labeled supergenetree $(G, ev_G)$. $L = \mathcal{L}(x)$ and $(L_i, L_r) = (\mathcal{L}(x_l), \mathcal{L}(x_r))$. The local reconciliation cost of $x$, $\text{cost}_G(L_i, L_r)$, is equal to:

- $3 + \text{inter}(s(L), s(L_i)) + \text{inter}(s(L), s(L_r))$ if $s(L) \neq s(L_i), s(L) \neq s(L_r), | G(L_i, L_r) | > 0$ and $ev_G(L_i, L_r) = \text{Dup}$;
- $\text{cost}(L_i, L_r)$ Otherwise.

In the first case, the node $x$ is a duplication node adding 1 duplication plus at least 2 losses to the reconciliation cost, and in the second case the local reconciliation cost is computed as for the LCA-reconciliation.

The complexity of the $\text{MinLSGT}$ algorithm remains in $O((n+1)^k \times 4^k \times k)$ provided that the sets of label-compatible bipartitions $(L_i, L_r)$ are constructed simultaneously with the sets $G(L_i, L_r)$.

### 3.3. Improved complexity from a core set of trees

Last, we show that the principles underlying the two algorithms described above can be improved to reduce the dependency in $k$. The key remark is that all bipartitions to consider can be identified by considering only a subset of the input trees provided they span the set of all genes of $\Gamma$.

Call $G' \subseteq G$ a core of $G$ if $\bigcup_{G \in G'} S(G) = \Gamma$. We introduce the following modified $\text{MinSGT}$ algorithm, that we call $\text{MinSGT-core}$:

1. Find a core $G' = \{G_1', \ldots, G_{k'}\}$ of $G$.
2. Apply the $\text{MinSGT}$ algorithm on $G'$, with the exception that, when considering a bipartition $B = (L_i, L_r)$ compatible with $G'$:
   - Verify that $B$ is also compatible with $G \setminus G'$. If not, then do not proceed recursively on $B$;
   - Compute $\text{cost}(L_i, L_r)$ on the whole set $G$.

**Theorem 2.** Let $G'$ be a core of $G$ composed of $k'$ trees. The $\text{MinSGT}$ problem can be solved in $O((n+1)^k \times 4^k \times k)$ time complexity.

**Proof.** The difference between the executions of a call of the $\text{MinSGT-core}$ algorithm on the input $G'$ and a call of the $\text{MinSGT}$ on $G$ lies in the set of bipartitions considered at each step of the recursion. At a given step of the recursion, let $B'$ and $B$ be the set of bipartitions compatible with $G'$ and $G$, respectively, and let $B^*$ be the set of bipartitions considered by $\text{MinSGT-core}$. We show that $B = B^*$.

Clearly $B \subseteq B'$, as a bipartition compatible with $G$ is also compatible with $G' \subseteq G$. Suppose that there is some $(L_i, L_r) \in B$ such that $(L_i, L_r) \notin B^*$. This implies that $(L_i, L_r)$ was filtered out of $B'$, meaning that it is not compatible with some tree $G \in G \setminus G'$. Therefore $(L_i, L_r)$ cannot be in $B$, a contradiction. We deduce that $B \subseteq B^*$. To see that $B^* \subseteq B$, observe that $B^*$ contains only bipartitions compatible with $G' \cup (G \setminus G') = G$, and that $B$ contains every such bipartition. So both algorithms consider the same set of bipartitions at each step, and lead to the same solution. □

It thus remain to describe how to find a core $G'$, as small as possible, as the size of the core is now the main complexity parameter. This problem is equivalent to the **MINIMUM SET COVER PROBLEM** known to be NP-hard. However, a natural heuristic is the following: choose a gene tree $G_i$ with the largest subset of $\Gamma$ as leafset, say of size $n-p$, and “complete” it with at most $p$ additional gene trees.
from $G$ each containing at least one of the missing genes. This obviously provide a core, leading to the following result.

**Corollary 2.** The MinSGT problem can be solved in $O((n + 1)p^4 \times 4^p + 1 \times k)$ time complexity, where $p$ is the smallest integer such that a gene tree of $G$ contains $n - p$ genes.

The same technique applies to MinLSGT and the same result could be stated for this problem.

### 4. Triplet Respecting Supergenetrees

We now consider a problem related to the correction of a gene tree. Assume that input gene trees $G_1, G_2, \ldots, G_k$ are separated subtrees (i.e. leaf-disjoint) of a given gene tree $G$. The MinSGT and MinLSGT problems can also be considered in this context to infer an alternative gene tree displaying them all and minimizing a reconciliation cost. However, this may lead to a new tree exhibiting a complete reorganization of the input subtrees and possibly grouping genes that were far apart in the initial tree. Therefore, assume in addition that we trust the hierarchy of upper branches. Then we ask for a supergenetree of minimum reconciliation cost which preserves the phylogenetic relation between subtrees, as given by $G^{\text{init}}$. Formally, we seek for a supertree respecting supergenetree, as defined below.

**Definition 2.** Let $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ be a set of separated subtrees of a gene tree $G^{\text{init}}$ for $\Gamma$ such that $\bigcup_{i=1}^k L(G_i) = \Gamma$. A tree $G^{\text{TR}}$ compatible with $\mathcal{G}$ is triplet respecting iff, for any triplet of leaves $L(G_{i_1}), L(G_{i_2})$ and $L(G_{i_3})$ in $\mathcal{G}$ and any triplet of genes $x \in G_{i_1}$, $y \in G_{i_2}$ and $z \in G_{i_3}$, $G^{\text{init}}$ and $G^{\text{TR}}$ display the same topology for the triplet $(x, y, z)$, i.e. $G^{\text{init}}|_{(x,y,z)} = G^{\text{TR}}|_{(x,y,z)}$.

For example in Figure 3, the supergenetree $G$ is not triplet respecting as for the triplet of genes $(h_3, m_1, m_2)$, $G$ does not display the same topology as the tree $G^{\text{init}}$.

**MINIMUM TRIPLET RESPECTING SUPERGENETREE (MinTRS) Problem:**

**Input:** A species set $\Sigma$ and a species tree $S$ for $\Sigma$; a gene family $\Gamma$ and a gene tree $G^{\text{init}}$ for $\Gamma$; a set $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ of separated subtrees of $G^{\text{init}}$ such that $\bigcup_{i=1}^k L(G_i) = \Gamma$.

**Output:** Among all triplet respecting gene trees for $\Gamma$ compatible with $\mathcal{G}$, one of minimum LCA-reconciliation cost.

A natural extension of the MinTRS Problem is the MINIMUM LABELED TRIPLET RESPECTING SUPERGENETREE (MinLTRS) Problem, where we are given a set of labeled separated subtrees of $G^{\text{init}}$ and we seek for a labeled triplet respecting supergenetree of minimum reconciliation cost. Here we focus on MinTRS, though all results extend naturally to MinLTRS, as briefly explained at the end of this section. Note that the MinTRS and MinLTRS problems can be reduced to the MinSGT and MinLSGT problems by considering as input of MinSGT and MinLSGT the set of subtrees $\mathcal{G}$ of $G^{\text{init}}$ augmented with the set of all rooted triplet trees that should be respected by the output supegenetree. However, the algorithms for MinTRS and MinLTRS problems induced by these reductions would remain exponential in the number of input subtrees.

We describe a more efficient recursive algorithm that solves the MinTRS problem by making use of the MinSGT solution. This algorithm leads to the following result.

**Theorem 3.** The MinTRS and MinLTRS problems can be solved in $O(n^2)$ time complexity.

The high-level description of the algorithm is as follows. The triplet-respecting property only allows a limited number of ways to combine the subtrees of $\mathcal{G}$ together. We distinguish two possible cases. First, if two subtrees $G_1, G_2$ of $\mathcal{G}$ form a “cherry” in $G^{\text{init}}$, meaning that $r(G_1)$ and $r(G_2)$ share the same parent in $G^{\text{init}}$, then $G_1$ and $G_2$ can be mixed together in any way without contradicting the triplet-respecting property. The optimal way of mixing the two trees is to compute $\text{MinSGT}(G_1, G_2)$, which gives a solution for the subtree of $G^{\text{init}}$ rooted at the parent of $r(G_1)$ and $r(G_2)$. For instance in Figure 3, the two children of the $\beta$ node in $G^{\text{init}}$ form a cherry of subtrees. Second, if instead a subtree $G_1$ of $\mathcal{G}$ is not part of such a cherry, then let $x$ be the sibling of $r(G_1)$ in $G^{\text{init}}$. Then we show that the following procedure can be performed: recursively compute $G_x$, an optimal solution for the subtree of $G^{\text{init}}$ rooted at $x$, then try grafting $G_x$ on $G_1$ in every possible way and keep the solution that minimizes the reconciliation cost. This gives a solution for the subtree of $G^{\text{init}}$ rooted at the parent of $r(G_1)$ and $x$. These two cases describe all the possible subtree mixings that can occur, and the rest of the $G^{\text{init}}$ topology must be conserved. For example in Figure 3, from a bottom-up point of view, the algorithm would compute $G_{3,4} = \text{MinSGT}(G_3, G_4)$, then obtain $G_{2,3,4}$ by finding the best place on which to graft $G_{3,4}$ on $G_2$ (in this case, above the parent of $m_1$ and $r_1$), then obtain a solution by grafting $G_{2,3,4}$ somewhere on $G_1$ (in this case above $h_1$). In the following we rather describe the algorithm in a top-down manner, i.e. we start at the root of $G^{\text{init}}$, obtain a solution recursively for its two child subtrees and combine them appropriately.

Before describing the algorithm in full detail, we give a few additional definitions and properties. Let $G$ and $G'$ be two gene trees for $\Gamma$. Define $s_{G' \rightarrow G}$ as the mapping from the nodes of $G'$ to the nodes of $G$ such that $s_{G' \rightarrow G}(x) = lca_G(L(x))$. For example in Figure 3, the image of the lowest $\delta$ node of $G^{\text{init}}$ by $s_{G^{\text{init}} \rightarrow G}$ is the highest $\delta$ node of $G$, since it is the $lca$ of $m_1, r_1, h_2$.

The algorithm for MinTRS constructs the triplet respecting supergenetree $G^{\text{TR}}$ by building recursively and independently the bi partitions $(L(y_l), L(y_r))$ induced by each internal node $y$ of $G^{\text{TR}}$ from the root to the leaves. The nodes of $G^{\text{TR}}$ can be considered independently in the algorithm because the constraint of being triplet respecting strongly predetermines the set of leaves $L(y)$ associated to some nodes $y$ of $G^{\text{TR}}$ as shown in Lemma 6.
Given a node \( x \) of \( G_{\text{Init}} \), we denote by \( G(x) \) the subset of \( G \) that are subtrees of \( G_{\text{Init}}[x] \). If there exists a node \( y \) in \( G^{TR} \) such that \( L(y) = L(x) \), then we also define \( G(y) = G(x) \). For example, call \( x \) the highest \( \delta \) node of \( G_{\text{Init}} \) in Figure 3. Then \( G(x) = \{G_2, G_3, G_4\} \). Now, for \( y \) being the lowest (non-loss) \( \delta \) node of \( G^{TR} \), \( L(y) = L(x) \) and so \( G(y) = \{G_2, G_3, G_4\} \).

**Lemma 6.** Let \( G^{TR} \) be a triplet respecting supegenetree for \( G_{\text{Init}} \) and \( G = \{G_1, \ldots, G_k\} \). For any node \( x \) of \( G_{\text{Init}} \) such that \( |G(x)| \geq 2 \), there exists a node \( y \) of \( G^{TR} \) such that \( L(y) = L(x) \).

**Proof.** Let \( x \) be a node of \( G_{\text{Init}} \) such that \( |G(x)| \geq 2 \). Each of the subtrees \( G_{\text{Init}}[x_l] \) and \( G_{\text{Init}}[x_r] \) then contains at least one tree of \( G \). Assume that (*) there exists no node \( y \) in \( G^{TR} \) such that \( L(y) = L(x) \). Let \( x' \) be the node of \( G^{TR} \) such that \( s_{G_{\text{Init}}}^{G^{TR}}(x) = x' \), and let \( x'' \) be the node of \( G_{\text{Init}} \) such that \( s_{G_{\text{Init}}}^{G^{TR}}(x') = x'' \). The assumption (*) implies that \( x'' \neq x \), so \( x'' \) is a strict ancestor of \( x \). Suppose w.l.o.g. that \( x \) belongs to the subtree \( G_{\text{Init}}[x''_l] \) and pick any gene \( g \in L(x') \cap L(x''_r) \). There exists a tree \( G_h \) of \( G \) such that \( c \in \mathcal{L}(G_h) \) and \( G_h \) is contained in \( G_{\text{Init}}[x''_r] \). Let \( a \) and \( b \) be two genes such that \( a \in \mathcal{L}(x_1) \cap \mathcal{L}(x_2) \) and \( b \in \mathcal{L}(x_3) \cap \mathcal{L}(x_4) \), or \( a \in \mathcal{L}(x_1) \cap \mathcal{L}(x_3) \) and \( b \in \mathcal{L}(x_2) \cap \mathcal{L}(x_4) \). Such two genes necessarily exist because \( s_{G_{\text{Init}}}^{G^{TR}}(x) = x' \). So, there exist two trees \( G_i \) and \( G_j \) of \( G \) such that \( (a, b) \in \mathcal{L}(G_i) \times \mathcal{L}(G_j) \). \( G_i \) is contained in \( G_{\text{Init}}[x_1] \) and \( G_j \) is contained in \( G_{\text{Init}}[x_2] \). So \( G_{\text{Init}} \) displays the topology \((a, b, c)\) for the triplet of genes \((a, b, c)\) while \( G^{TR} \) displays a different topology, either \((a, b, c)\) or \((ac, b)\). The assumption (*) is then impossible. \( \square \)

We denote by \( V_{\text{cons}}(G^{TR}) \) the subset of nodes \( y \) of \( G^{TR} \) such that there exists a node \( x \) of \( G_{\text{Init}} \) satisfying \( L(x) = L(y) \) and \( |G(y)| \geq 2 \). For example, in Figure 3, \( V_{\text{cons}}(G^{TR}) \) contains three nodes, the root, the lowest \( \delta \) node and the lowest duplication node of \( G^{TR} \). Lemma 6 allows to predetermine the sets of leaves \( L(y) \) associated to the nodes \( y \in V_{\text{cons}}(G^{TR}) \). We now describe how to find the best subtree \( G^{TR}[y] \) for each node \( y \in V_{\text{cons}}(G^{TR}) \), i.e. one leading to the minimum reconciliation cost.

Note that if \( y \in V_{\text{cons}}(G^{TR}) \) and \( |G(y)| = 2 \), say \( G(y) = \{G_i, G_j\} \), \( 1 \leq i < j \leq k \), then the MinSGT algorithm can be applied to build the subtree \( G^{TR}[y] \) as a minimum reconciliation cost supegenetree for \( G_i \) and \( G_j \). It then remains to describe a recursive procedure for finding the subtree \( G^{TR}[y] \) for a node \( y \in V_{\text{cons}}(G^{TR}) \) such that \( |G(y)| \geq 2 \).

In order to compute the reconciliation cost of the tree \( G^{TR} \), we need to account for the local reconciliation costs for the nodes \( y \in V_{\text{cons}}(G^{TR}) \), and also for the internal nodes \( z \) of \( G^{TR} \) such that \( z \notin V_{\text{cons}}(G^{TR}) \). To do so, given a node \( y \in V_{\text{cons}}(G^{TR}) \), we define \( \text{cost}_{TR}(y) \) as the local reconciliation cost for \( y \), plus the local reconciliation costs for all internal nodes \( z \in V(G^{TR}) \) such that \( z \notin V_{\text{cons}}(G^{TR}) \), \( y \) is an ancestor of \( z \) and there exists no node \( y' \in V_{\text{cons}}(G^{TR}) \) on the path between \( y \) and \( z \).

For example in Figure 3, call \( y \) the root of \( G^{TR} \). Then, \( y \in V_{\text{cons}}(G^{TR}) \) and \( \text{cost}_{TR}(y) = \text{cost}(y_r) = \text{cost}(L(y_r)) \). For all nodes \( y \notin V_{\text{cons}}(G^{TR}) \), \( y \) and \( y_r \notin V_{\text{cons}}(G^{TR}) \). We then obtain a formulation of the reconciliation cost of \( G^{TR} \) as the sum of \( \text{cost}_{TR}(y) \) for all nodes \( y \in V_{\text{cons}}(G^{TR}) \).

**Lemma 7.** Let \( G^{TR} \) be a triplet respecting supegenetree
for $G = \{G_1, \ldots, G_k\}$. Let $y$ be a node of $G^{TR}$ such that $y \in V_{cons}(G^{TR})$ and $|G(y)| > 2$. Let $x$ be the node of $G^{init}$ such that $L(y) = L(x)$.

1. If $|G(x_1)| = 1$ and $|G(x_r)| \geq 2$, let $y^* \in V_{cons}(G^{TR})$ be the node of $G^{TR}$ such that $L(y^*) = L(x_r)$. The subtree $G^{TR}[y]$ can be obtained by taking the tree $G^{init}[x_1]$ and grafting the tree $G^{TR}[y^*]$ onto it such that the root of $G^{TR}[y^*]$ appears as the sibling of a node $x^*$ of $G^{init}[x_1]$. The cost $cost_{TR}(y)$ is then given by the following formula:

$$cost_{TR}(y) = cost_{TR}(x_1, x_r, x^*) = \sum_{u \in A(x^*)} cost(L(u_i) \cup L(x_r), L(u_r)) + \sum_{u \in A_r(x^*)} cost(L(u_i), L(u_r) \cup L(x_r)) + \sum_{u \in \{V(G^{init}[x_1]) \backslash A(x^*)\}} cost(L(u_i), L(u_r)) + cost(L(x_1), L(x_r)) \quad (if \ x^* = x_1)$$

2. If $|G(x_1)| \geq 2$ and $|G(x_r)| = 1$, then this case is symmetric to the previous case.

3. If $|G(x_1)| \geq 2$ and $|G(x_r)| \geq 2$, then $G^{TR}[y]$ is such that $L(y_1) = L(x_1)$ and $L(y_r) = L(x_r)$, and $cost_{TR}(y) = cost(L(x_1), L(x_r))$.

Proof: In Case 1, we first deduce from Lemma 6 that there must exist a node $y^* \in V_{cons}(G^{TR})$ such that $L(y^*) = L(x_r)$. Next, $G^{init}[x_1]$ is one of the gene trees of the set $G$. So, it must be displayed by $G^{TR}[y]$ and then, $G^{TR}[y]$ can be obtained by taking $G^{init}[x_1]$ and grafting $G^{TR}[y^*]$ onto it. Finally, Case 2 is symmetric to Case 1 and Case 3 follows directly from Lemma 6. The formulas for $cost_{TR}(y)$ follows directly from the definition of $cost_{TR}$.

For example in Figure 3, the root and the highest $\delta$ node of $G^{TR}$ fulfills the conditions of the first case. There are no node $y \in V_{cons}(G^{TR})$ satisfying $|G(y)| > 2$ and fulfilling the conditions of the second or third case.

We are now ready to describe the recurrence formula of the recursive algorithm solving the $MinTRS$ problem. Given a node $x$ of $G^{init}$ such that $|G(x)| \geq 2$, we denote by $MinTRS(G^{init}[x])$ the minimum LCA-reconciliation cost of a triplet respecting supertree compatibility with $G(x)$.

Lemma 8. Let $G = \{G_1, \ldots, G_k\}$ be a set of separated subtrees of a gene tree $G^{init}$ for $\Gamma$ such that $\bigvee_{i=1}^{k} L(G_i) = \Gamma$. Let $x$ be a node of $G^{init}$ such that $|G(x)| \geq 2$.

1. (Stop condition) If $|G(x_1)| = 2$ and $G(x) = \{G_i, G_j\}$, 
   $$MinTRS(G^{init}[x]) = MinSGT(G_i, G_j).$$

2. Otherwise (i.e., if $|G(x)| > 2$),
   a. If $|G(x_1)| = 1$ and $|G(x_r)| \geq 2$,
      $$MinTRS(G^{init}[x]) = \min_{x^* \in V(G^{init}[x])} \{cost_{TR}(x_1, x_r, x^*)\} + MinTRS(G^{init}[x_1])$$
   b. If $|G(x_1)| \geq 2$ and $|G(x_r)| = 1$, this case is symmetric to the previous case.

   c. If $|G(x_1)| \geq 2$ and $|G(x_r)| \geq 2$,
      $$MinTRS(G^{init}[x]) = MinTRS(G^{init}[x_1]) + MinTRS(G^{init}[x_r])$$

Proof. The proof follows from Lemmas 6 and 7, and the fact that each call to the recursive procedure $MinTRS$ receives as input a subtree $G^{init}[x]$ such that $|G(x)| \geq 2$, starting with the whole tree rooted at $r(G^{init})$. Case 1 is trivial. For Case 2(a) (and symmetrically Case 2(b)), following Lemma 7, there are $V(G^{init}[x_1])$ possible configurations for the subtree $G^{TR}[y]$ rooted at $y = s_{G^{init} \rightarrow G^{TR}}(x)$, depending on which node $x^*$ of $G^{init}[x_1]$ is chosen to be the sibling of $y = s_{G^{init} \rightarrow G^{TR}}(x_r)$. Since $G^{TR}$ must be of minimum reconciliation cost, the configuration for $G^{TR}[y]$ must be one that locally minimizes the cost $cost_{TR}(y)$.

Finally, Case 3 follows directly from Lemma 7.

Complexity. We claim that Lemma 8 leads to a $O(n^2)$ algorithm for $MinTRS$, where $n = |V(G^{init})|$. Let $x$ be a node of $G^{init}$, let $n_x$ be the number of nodes in $G^{init}[x]$ and let $n_l$ and $n_r$ be the number of nodes in the left and right subtrees of $x$, respectively. As a base case, if $x$ falls into case 1 of Lemma 8, then running $MinSGT$ on the two child subtrees of $x$ takes time $O(max\{n_l, n_r\}^2) = O(n_x^2)$. Suppose instead that $x$ falls into case 2a, and thus $G(x_1) = 1$ and $G(x_r) \geq 2$. We may assume by induction that computing $MinTRS(G^{init}[x_1])$ requires time $O(n_l^2)$. Afterwards, grafting the resulting tree is done on each $O(n_l)$ branch of $G^{init}[x_1]$, and computing the cost can be done in time $O(n_l)$ for each grafting. Thus in total, case 2a can be handled in time $O(n_l^2 + n_r^2) = O(n_x^2)$. The case 2b of Lemma 8 is symmetric, and the case 2c can be handled in constant time. As the quadratic bound holds for every node, we get a bound of $O(n_x^2) = O(n_x^2)$ when $x$ is the root.

Algorithm for $MinLTRS$. The algorithm for the $MinTRS$ problem can be adapted to solve the $MinLTRS$ problem. The adaptation consists in (1) replacing the calls to $MinSGT(G_i, G_j)$ in the stop condition of Lemma 8 by calls to $MinLSGT((G_i, e_{\Gamma_1}), (G_j, e_{\Gamma_2}))$, and (2) replacing the use of $cost(l_i, l_r)$ in order to define $cost_{TR}(y)$ in Lemma 7 by the use of $cost(l_i, l_r)$. Moreover, Lemma 5 must be extended such that $cost_G(l_i, l_r) = +\infty$ if $(l_i, l_r)$ is not label-compatible with $G$. The complexity of the algorithm remains unchanged in $O(n^2)$.

5. Experiments

In the context of gene tree correction, we wanted to evaluate: (1) the benefit of the new supertree approach allowing to merge clades from different subtrees, compared with the more constrained polynomy resolution approach [24] which conserves input subtrees separated; (2) the benefit of the additional triplet preservation requirement of $MinTRS$ and $MinLTRS$. Both evaluations were made based on the conjecture of dubious highest duplication nodes in gene trees [17], [21].

<table>
<thead>
<tr>
<th>Method</th>
<th># of modified trees</th>
<th>Avg. running time</th>
<th>Avg. rec. cost reduction</th>
<th>Trees with better AU value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinSGT</td>
<td>211</td>
<td>205240 ms</td>
<td>22.5 (24.8%)</td>
<td>66.4%</td>
</tr>
<tr>
<td>MinLSGT</td>
<td>297</td>
<td>113 ms</td>
<td>19.3 (21.5%)</td>
<td>95.3%</td>
</tr>
<tr>
<td>MinTRS</td>
<td>211</td>
<td>3031 ms</td>
<td>15.5 (17.1%)</td>
<td>68.6%</td>
</tr>
<tr>
<td>MinLTRS</td>
<td>207</td>
<td>60 ms</td>
<td>13.5 (14.9%)</td>
<td>66.4%</td>
</tr>
<tr>
<td>PolytomySolver</td>
<td>20 ms</td>
<td>3 ms</td>
<td>3.05 (4.0%)</td>
<td>50.0%</td>
</tr>
</tbody>
</table>

For this purpose, we considered the gene trees of the Ensembl vertebrate database Release 84 rooted at a duplication node. For each tree $G^{Init}$, $G$ was defined as the set of all subtrees of $G^{Init}$ rooted at the “highest speciation nodes”, i.e. speciation nodes with only duplication nodes as ancestors. On average, $G^{Init}$ contains 121.2 leaves and is partitioned into 7.5 subtrees. Aiming at comparing all developed algorithms, including the exponential time MinSGT and MinLSGT, we restricted the sample to the 217 gene trees with at most 200 leaves and partitioned into at most 5 subtrees. We also applied, on these 217 trees, PolytomySolver [24] which, given a set of trees $G_1, \ldots, G_k$, finds a binary tree with leafset \{$G_1, \ldots, G_k$\} such that the reconciliation cost of the resulting tree is minimum. Results are given in Table 5.

While the four supertree algorithms correct more than 200 trees, corresponding to more than 95% of the 217 trees, PolytomySolver only corrects 20 trees corresponding to about 9% of the trees. Additionally, PolytomySolver reduces the reconciliation cost by only 4% on average on the 20 corrected trees, compared to more than 15% for supertree algorithms. Clearly, by exploring a larger solution space, MinSGT and MinLSGT allow to obtain the best solutions in terms of reconciliation cost.

As for MinTRS and MinLTRS, although more constrained than MinSGT and MinLSGT, they lead surprisingly to almost as much correction as these two algorithms, while the correction achieved by PolytomySolver is clearly less. The triplet preserving constraint appears to be less stringent than the conservation of the subtrees. In particular, for the trees leading to only two subtrees, PolytomySolver conserves the initial tree. Notice that introducing the labeling constraint (MinSGT versus MinLSGT and MinTRS versus MinLTRS) only leads to a slight decrease of the correction rates.

Finally, in order to assess the benefit of the triplet respecting constraint and the quality of the correction achieved, the trees corrected by MinTRS, MinLTRS and PolytomySolver were evaluated according to their statistical support. PhyML [34] was executed to obtain the log-likelihood values per site (note that 150 trees were included in this evaluation, as PhyML was very time-consuming on the larger trees). Consel [35] was then run to evaluate, using the AU (Approximately Unbiased) test, if the likelihood differences of pairs of Ensembl and corrected gene trees were significant enough to statistically reject one of them. A tree can be rejected if its AU value, interpreted as a p-value, is under 0.05. Otherwise, no significant evidence allows us to reject one of the two trees.

Interestingly, when compared to the tree output by MinTRS (respectively MinLTRS), 48.5% (resp. 46.5%) of the Ensembl gene trees are rejected compared to only 11.9% (resp. 11.0%) of the corrected gene trees. More than 68.6% (resp. 66.4%) of the corrected trees have better AU values than original trees. As for PolytomySolver, 5% of the Ensembl trees were rejected, as 25% of the corrected trees were rejected, with 50% of the corrected trees obtaining a better AU value. The performance of MinTRS and MinLTRS is rather surprising as our correction, based on the phylogenetic information of the species tree, is not expected to improve tree likelihood based on sequence similarity. This may be an indication that high duplications are actually dubious and that a correction specifically focusing on such duplications is able to significantly improve the accuracy of the tree. This observation is further supported by the fact that the number of highest duplications is lower for corrected trees than for initial trees (data not shown), showing that our correction algorithms have the general tendency of deleting high duplications.

6. Conclusion

This paper introduces a new methodology combining the supertree and reconciliation frameworks with the purpose of constructing a gene tree by combining a set of trees on partial, possibly overlapping data. We also show how this new paradigm is useful for gene tree correction. In particular, the artifact of duplications wrongly inferred close to the root of a gene tree has been reported in the literature. Here, we propose a new method for correcting a gene tree, by
first removing the higher duplication nodes and then finding the supertree best fitting the species tree, that preserves the remaining “trusted” subtrees, and possibly their hierarchical position in the initial gene tree. This supertree approach is shown to correct more trees than the approach based on resolving a polytomy, as the first correction allows the clustering of genes from different input subtrees. The corrected Ensembl gene trees are shown to exhibit less highest duplication nodes and a lower reconciliation cost. Corrected gene trees are also shown to have a better likelihood support.

This new gene tree construction and correction paradigm leads to many new open problems. In particular, no proof currently exists on the complexity of the problem of finding a supertree minimizing the reconciliation cost, although it is likely to be NP-hard, based on the fact that minimizing the duplication cost is hard. The two problems (reconciliation versus duplication costs) probably also share the same inapproximability properties. However, it is possible that the supertree problems presented here are fixed-parameter tractable with respect to parameters such as the number of trees, the minimum reconciliation cost or the size of the intersection between the leafset of the trees. This is an area that deserves a more in-depth investigation. In addition, while the extension to the labeled case has been done with the same exponential complexity, adding the label restriction strongly constrains the set of explored bipartitions, and we can expect a more efficient algorithm in this case.

The problems we consider are build upon strong underlying assumptions, such as the consistency of input gene trees, the compatibility and S-consistency of input gene relations. A natural extension is then to integrate the notion of a minimal correction of input trees to fit these preliminary conditions. Finally, from an application point of view, rather than removing higher duplication nodes, other types of gene tree pruning can be envisaged and used to select the initial “trusted” phylogenetic information that can then be combined using our supertree and reconciliation framework.

**Funding**

M.L. acknowledges the support of the Fonds de Recherche du Québec Nature et Technologie (FRQNT). C.C. acknowledges the support of the Natural Sciences and Engineering Research Council (NSERC) (Discovery Grant RGPIN-249834). N.E. acknowledges the support of NSERC. A.O. acknowledges the support of the Canada Research Chairs (CRC) (CRC Tier2 Grant 950-230577).

**References**


**Manuel Lafond** received the PhD degree from the University of Montreal, Canada in 2016. He is now a NSERC postdoctoral fellow in the Department of Mathematics and Statistics at the University of Ottawa, Canada, where he is developing new models and methods to distinguish orthologous genes from paralogous genes. His research interests mainly reside in the intersection of computational biology, algorithms and graph theory.

**Cedric Chauve** is a professor in mathematics at the Simon Fraser University. His research interests include algorithms design and analysis, genome evolution, genome rearrangements, and ancestral genome reconstruction.

**Nadia El-Mabrouk** is a full professor at the computer science department (DIRO) of the University of Montreal. She joined the department in 1998, after her Ph.D. in theoretical computer science from University “Paris VII” in France. Her research projects are related to the computational biology aspects of comparative genomics. She publishes in various computer science, bioinformatics and life science journals, such as “Bioinformatics”, “Journal of Computational Biology” or “Molecular Biology and Evolution”. She is involved, every year, in the program committees of several well-known bioinformatics conferences such as RECOMB, ISMB, WABI or APBC.

**Aïda Ouangraoua** received the master’s degree and PhD degree in Computer Science from Université de Bordeaux, France, in 2004 and 2007, respectively. She did her postdoctoral training at Simon Fraser University and Université du Québec à Montréal, Canada. She was a researcher at INRIA Lille, France from 2009 to 2014. She is currently a professor at the Department of Computer Science of Université de Sherbrooke, Québec, Canada. Her research interests include Algorithms and Computational Molecular Biology.