New Perspectives on Gene Family Evolution: Losses in Reconciliation and a Link with Supertrees*

Cedric Chauve[1] and Nadia El-Mabrouk[2]

[1] Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada
[2] Département Informatique et Recherche Opérationnelle, Université de Montréal, Montréal (QC), Canada
cedric.chauve@sfu.ca, mabrouk@iro.umontreal.ca

Abstract. Reconciliation between a set of gene trees and a species tree is the most commonly used approach to infer the duplication and loss events in the evolution of gene families, given a species tree. When a species tree is not known, a natural algorithmic problem is to infer a species tree such that the corresponding reconciliation minimizes the number of duplications and/or losses. In this paper, we clarify several theoretical questions and study various algorithmic issues related to these two problems. (1) For a given gene tree T and species tree S, we show that there is a single history explaining T and consistent with S that minimizes gene losses, and that this history also minimizes the number of duplications. We describe a simple linear-time and space algorithm to compute this parsimonious history, that is not based on the Lowest Common Ancestor (LCA) mapping approach; (2) We show that the problem of computing a species tree that minimizes the number of gene duplications, given a set of gene trees, is in fact a slight variant of a supertree problem; (3) We show that deciding if a set of gene trees can be explained using only apparent duplications can be done efficiently, as well as computing a parsimonious species tree for such gene trees. We also characterize gene trees that can be explained using only apparent duplications in terms of compatible triplets of leaves.

1 Introduction

Applying local similarity search tools to genomes of closely related species usually reveal large clusters of homologous genes, also called *gene families*. Such grouping by sequence similarity is not sufficient to infer a common function for genes. Indeed, in addition to orthologs which are copies in different species related through speciation, gene families are likely to contain paralogs, which are copies that have evolved by duplication. Paralogs are more likely to have acquired new functions. In addition to gene duplication, gene losses, arising through the pseudogenization of previously functional genes, also play a key role in the

* Both authors are supported by grants from NSERC.

S. Batzoglou (Ed.): RECOMB 2009, LNCS 5541, pp. 46–58, 2009.

evolution of gene families [23,21,14,10,4,11,19]. Understanding the evolution of gene families is thus a fundamental question in functional genomics, but also in evolutionary biology and phylogenomics [28,32].

The most commonly used methods to infer evolutionary scenarios for gene families are based on the *reconciliation* approach that compares the species tree (describing the relationships among taxa) to the gene trees, and implicitly infers a set of gene duplications and losses. Given a species tree and a set of gene trees, there can be several reconciliations, and a natural approach is then to select one optimizing a given criterion, either combinatorial [22] or probabilistic [3]. Natural combinatorial criteria are the number of duplications (duplication cost), losses (loss cost) or both combined (mutation cost). The so called Lowest Common Ancestor (LCA) mapping between a gene tree and a species tree, introduced in [17] and is widely used studies [24,15,25,33,22,26,5,16,13], defines a reconciliation that minimizes both the duplication and mutation costs [16]. Although losses appear to be an important phenomenon in the evolution of a gene family, they have only recently been explicitly used as a parsimony criterion [7]. It can be computed efficiently, in linear time [33] or using a simple quadratic time algorithm [34]. When no preliminary knowledge on the species tree is given, a natural problem is to infer, from a set of gene trees, a species tree leading to a parsimonious evolution scenario, for a chosen cost. Similarly to the case of a known species tree, methods have been developed for the duplication and mutation costs [22,18,8]. For both criteria, the problem of inferring an optimal species tree given a set of gene trees is hard [22].

In this paper, we present various theoretical results related to the optimization problems of inferring, for a given gene tree (or a forest of gene trees), an evolution scenario minimizing a given cost, in both cases of a known and an unknown species tree.

In Section 3, we clarify the link between the duplication and loss cost criteria for reconciliation. Given a gene tree T and a species tree S, we show that there is a single history explaining T and consistent with S minimizing losses, and that this history also minimizes duplications. This refines recent results showing that there is a unique reconciliation minimizing the mutation cost [16]. We describe a simple linear-time reconciliation method, not based on the LCA mapping, computing this most parsimonious history. Although our new reconciliation algorithm is not the only one running in linear time [33,15], its implementation is simpler, and it highlights the important combinatorial role of gene losses regarding parsimonious evolution scenarios for gene families.

In Section 4, we describe the problem of computing, from a set of gene trees, a most parsimonious species tree for the duplication cost (the Minimum Duplication Problem), as an instance of the following restricted supertree problem: given a set of uniquely-leaf labeled gene trees where only the first speciation is resolved, compute a species tree that agrees with the largest number of such gene trees. Clearly, these two problems share some common ground in terms of goal – inferring a species tree from a collection of gene trees –, but differ in terms of data – duplicated leaves versus uniquely leaf-labeled trees – and

considered evolutionary mechanisms: duplication are ignored, at least explicitly, in supertree problems. The link between these two problems suggests that heuristics for the supertree problem, such as the min-cut greedy heuristic [29,27], are natural candidate heuristics for the Minimum Duplication Problem. The parallel with supertrees implies also an efficient algorithm to decide if a set of gene trees can be explained using only apparent duplications, as well as an efficient algorithm to compute all most parsimonious species trees for a set of such *MD-trees* (Minimum-Duplication trees). We also provide a combinatorial characterization of MD-trees as trees not containing triplets of leaves leading to contradictory phylogenetic information. The latter characterization of gene trees may be useful to detect ambiguous phylogenetic relationships or possible errors in a set of gene trees, as we illustrate in Section 5 on a simulated dataset.

2 Preliminaries

Trees. Let $\mathcal{G} = \{1, 2, \cdots, g\}$ be a set of integers representing g different species (genomes). A *species tree* on $\mathcal{G}$ is a binary tree with exactly g leaves, where each $i \in \mathcal{G}$ is the label of a single leaf. A *gene tree* on $\mathcal{G}$ is a binary tree where each leaf is labeled by an integer from $\mathcal{G}$ (each tree represents a gene family, where each leaf labeled i represents a gene copy located on genome i).

For a given vertex x of a tree T, we denote by T_x the subtree of T rooted at x and by $L(x)$ the subset of $\mathcal{G}$ defined by the labels of the leaves of T_x. $L(x)$ is called the *genome set of* x. We denote by x_ℓ and x_r the two children of x, if x is not a leaf, and by x_p its parent if x not the root. An *expanded leaf* of T is a vertex x such that $|L(x)| = 1$ and $L(x) \neq L(x_p)$, or x is the root of T. A *cherry* of a tree is an internal vertex x for which both children are expanded leaves.

Reconciliation. There are several definitions of reconciliation between a gene tree and a species tree. Here we define reconciliation in terms of subtree insertions, following an approach used in [16,7]. A *subtree insertion* in a tree T consists in grafting a new subtree onto an existing branch of T. A tree T' is said to be an *extension* of T if it can be obtained from T by a sequence subtree insertions in T.

Given a gene tree T on $\mathcal{G}$ and a species tree S on $\mathcal{G}$, T is said to be *DS-consistent with* S (following the terminology used in [7]) if, for every vertex x of T such that $|L(x)| \geq 2$, there exists a vertex u of S such that $L(x) = L(u)$ and one of the two following conditions (D) or (S) holds: (D) either $L(x_r) = L(x_\ell)$, or (S) $L(x_r) = L(u_r)$ and $L(x_\ell) = L(u_\ell)$.

A *reconciliation* between a gene tree T and a species tree S is an extension R of T that is DS-consistent with S (this definition is easily shown to be equivalent to other definitions of reconciliation [3,12]). Such a reconciliation between T and S implies an unambiguous evolution scenario for the gene family T where a vertex of R that satisfies property (D) represents a duplication (the number of duplications induced by R is denoted by $d(R, S)$), and an inserted subtree represents a gene loss (the number of gene losses induced by R is denoted by $\ell(R, S)$). Vertices of R that satisfy property (S) represent speciation events (see Fig. 1).

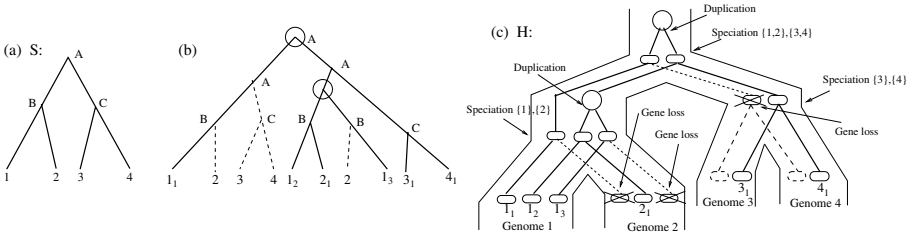

Fig. 1. (a) A species tree S; (b) The reconciliation R of S with the gene tree T represented by plain lines. Dotted lines represent subtree insertions (3 insertions). The correspondence between vertices of R and S is indicated by vertices labels. Circles represent duplications. All other internal vertices of R are speciation vertices; (c) Evolution scenario resulting from R. Each oval is a gene copy.

Given a gene tree T, it is immediate to see that every vertex x of T such that $L(x_\ell) \cap L(x_r) \neq \emptyset$ will always be a duplication vertex in any reconciliation R between T and S. Such a vertex is called an *apparent duplication vertex* (or just apparent duplication for short). For example, in the gene tree T represented by plain lines in Fig. 1.b., both duplication vertices are apparent duplications.

The notion of reconciliation can naturally be extended to the case of a set, or *forest*, of gene trees $\mathcal{F} = \{T_1, \ldots, T_m\}$: a reconciliation between $\mathcal{F}$ and S is a set $\mathcal{R} = \{R_1, \ldots, R_m\}$ of reconciliations, respectively for $T_1, \ldots, T_m$, such that each R_i is DS-consistent with S. We denote by $\mathcal{R}(\mathcal{F}, S)$ the set of all reconciliations between $\mathcal{F}$ and S.

Optimization problems: We consider three cost measures for a reconciliation $\mathcal{R}(\mathcal{F}, S)$ between a gene tree forest $\mathcal{F} = \{T_1, \ldots, T_m\}$ and a species trees S. The *duplication cost* is given by $d(\mathcal{R}, S) = \sum_{i=1}^{m} d(R_i, S)$, the *loss cost* by $\ell(\mathcal{R}, S) = \sum_{i=1}^{m} \ell(R_i, S)$ and the *mutation cost* by $m(\mathcal{R}, S) = \sum_{i=1}^{m} d(R_i, S) + \ell(R_i, S)$. For a given cost measure C (here d, ℓ or m), there are two natural combinatorial optimization problems, depending on whether a species tree is known or not.

Minimum Reconciliation C Problem:
Input: A gene tree forest $\mathcal{F}$ on $\mathcal{G}$ and a species tree S for $\mathcal{G}$;
Output: A reconciliation $\mathcal{R}$ with minimum cost $C(\mathcal{R}, S)$.

Minimum C Problem:
Input: A gene tree forest $\mathcal{F}$ on $\mathcal{G}$;
Output: A species tree S such that $\min_{\mathcal{R} \in \mathcal{R}(\mathcal{F}, S)} C(\mathcal{R}, S)$ is minimum.

The Minimum Duplication Problem and Minimum Mutation Problem with multiple gene trees are NP-complete [22]. The complexity status of the Minimum Loss Problem is still unknown.

3 Reconciled Trees

Let T be a gene tree on $\mathcal{G}$. We assume that a species tree S is already known for $\mathcal{G}$. The LCA mapping between T and S, denoted by M, maps every vertex

x of a gene tree T towards the Lowest Common Ancestor (LCA) of $L(x)$ in S. This mapping induces a reconciliation between T and S (see [12] for example) where an internal vertex x of T leads to a duplication vertex if $M(x_\ell) = M(x)$ and/or $M(x_r) = M(x)$. We denote by $M(T, S)$ the reconciliation between T and S defined by the LCA mapping. It has been shown recently [16] that $M(T, S)$ is the only reconciliation that minimizes the mutation cost, while there can be several reconciliations that minimize the duplication cost. The following theorem refines this result.

Theorem 1. *Given a gene tree T and a species tree S, $M(T, S)$ minimizes the duplication, loss and mutation costs. Moreover, $M(T, S)$ is the only reconciliation between T and S that minimizes the loss cost and minimizes the mutation cost.*

In [7, Prop. 1], it was shown that $M(T, S)$ is optimal for the loss cost. On the other hand, the fact that $M(T, S)$ is optimal for the duplication cost is a well known result (see [16] for a recent reference for example). It follows that $M(T, S)$ is optimal for each of the three costs. It then remains to show that $M(T, S)$ is the unique reconciliation between T and S that is optimal for the loss cost. This would imply that $M(T, S)$ is also the unique reconciliation that minimizes the mutation cost (a result proved in [16] although in a more complicated way) and complete our proof. To do so, we rely on a new simple linear-time algorithm that computes a reconciliation between T and S and minimizes the loss cost.

Algorithm Minimum-Reconciliation described below takes a gene tree T and a species tree S as input, and returns a reconciled tree R. Roughly speaking, the algorithm proceeds as follows: it traverses the gene tree T from the leaves to the root, and completes by subtree insertions the subtrees of T corresponding to the successive speciation events of S, from the latest ones to the earliest one. An example is given in Figure 2.

Algorithm Minimum-Reconciliation (T,S):

1. Set $R = T$, $R' = R$ and $S' = S$;
2. While S' is not reduced to a single vertex.
   (a) Visit the expanded leaves of R' given S'
      i. Let x be the current expanded leaf of R' and y its sibling;
      ii. Let u be the leaf of S' with $L(u) = L(x)$ and v its sibling;
      iii. If $L(y) \neq L(v)$ then insert in R on the branch between x and x_p a leaf labeled by $L(v)$;
   (b) Reduce each subtree of S' and R' corresponding to a cherry of S' to a single leaf labeled with the cherry genome set;
3. Return (R);

Theorem 2. *Given a gene tree T on $\mathcal{G}$ and a species tree S for $\mathcal{G}$,* Algorithm Minimum-Reconciliation *reconstructs the unique reconciliation between T and S that minimizes the number of gene losses. It can be implemented to run in $O(n)$ time and space.*

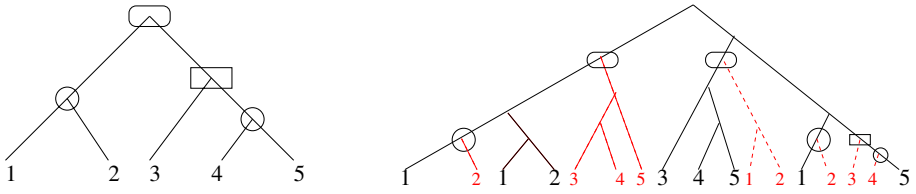

Fig. 2. Left: speciation tree S. Right: reconciliation tree R constructed over the gene tree T represented by solid lines. Three executions of Algorithm Minimum-Reconciliation loop of step 2 are required, and the successive lists of considered cherries of S' are represented by circles for the first iteration, a rectangle for the second iteration and ovals for the third iteration.

Proof. At the end of this algorithm, the resulting tree R is DS-consistent with S as it recursively completes cherries in R according to S. Therefore, R is a reconciliation of T and S. Moreover, as the algorithm considers each vertex of T exactly once, it can be implemented in a single post-order traversal of T. Handling vertices labeled by genome sets can be done efficiently by replacing such sets by integers, as was done in the algorithm DS-Recognition described in [7] that decides if a gene tree can be explained without gene loss by comparing the cherries of the gene tree to the cherries of the species tree. In fact Algorithm Minimum-Reconciliation can be implemented as a direct extension of algorithm DS-Recognition. Combined with the fact that subtree insertions can be implemented in constant times using pointers on vertices of S, this leads to a linear time and space complexity with low constants and using only simple data structures.

From now on, we denote by $MinR(S,T)$ the reconciliation tree obtained by Algorithm Minimum-Reconciliation, and we show that $MinR(S,T)$ is the only reconciliation between T and S that minimizes the number of losses. To prove this, we use the fact that each gene loss is represented by a subtree insertion. Given a species tree S, a gene tree T, and a cherry u of S such that $L(u) = \{a, b\}$, it follows from the definition of DS-consistency that every expanded leaf x of T with $L(x) = \{a\}$ or $L(x) = \{b\}$ has to be completed, if required, by inserting a sibling in order to form a cherry labeled $\{a, b\}$. Hence, all subtree insertions performed by Algorithm Minimum-Reconciliation when visiting the set of expanded leaves of T (first iteration of step 2) are required in order to extend T into a tree that is DS-consistent with S. The same property holds recursively to the following iterations of step 2, which implies that all subtree insertions performed by Algorithm Minimum-Reconciliation are necessary in order to extend T into a tree that is DS-consistent with S. Combined with the fact that, at the end of the algorithm, $MinR(S,T)$ is DS-consistent with S and the fact that the number of subtree insertions is the number of gene losses induced by $MinR(S,T)$, this completes the proof. □

Remark 1. It follows from Theorem 1 that minimizing losses results in minimizing duplications, as the unique solution to the Minimum Reconciliation

Loss Problem is a solution to the Minimum Reconciliation Duplication Problem. The converse is not true, as more than one solution may exist, in general, for the Minimum Reconciliation Duplication Problem. Stated differently, the loss cost criterion is more constraining than the duplication cost criterion for reconciliation. This property does not hold anymore in the case of inferring a species tree from a set of gene trees: the species tree that minimizes losses does not always minimizes duplications, and conversely. Even a weaker property does not hold in this case: there is not always a common solution to the Minimum Duplication Problem and the Minimum Loss Problem.

4 Gene Duplication and Supertrees

The general supertree problem can be stated as follows: given a set of uniquely leaf-labeled gene trees (i.e. in each gene tree no two leaves have the same label), compute a species tree optimizing some combinatorial criterion. A natural criterion is to maximize the number of input gene trees that are DS-consistent with the species tree (called a supertree).

The Minimum Duplication Problem is a supertree problem. We follow [31] to introduce terminology on supertrees. Given two uniquely leaf-labeled trees T and T', possibly non-binary, we say that T' *refines* T (denoted by $T' \rightarrow T$) if T can be obtained from T' by a sequence of contraction of internal edges of T'. Given a species tree S on $\mathcal{G}$ and a subset $\mathcal{H}$ of $\mathcal{G}$, we denote by $S_{|\mathcal{H}}$ the induced species tree on $\mathcal{H}$, obtained by first removing all vertices x of S such that $L(x) \cap \mathcal{H} = \emptyset$ and next removing all vertices of degree two. A, possibly non-binary, gene tree T is *consistent* with a species tree S if $S_{|L(T)} \rightarrow T$, and *inconsistent* otherwise. Finally, a uniquely leaf-labeled tree T on $\mathcal{G}$ is said to be a *bipartition* of $\mathcal{G}$ if T contains only three internal vertices x (the root), x_r and x_ℓ, and $L(x_r) \cap L(x_\ell) = \emptyset$ (x_r and x_ℓ are possibly non-binary vertices). For a given set $\mathcal{B} = \{B_1, \ldots, B_m\}$ of bipartitions of $\mathcal{G}$ and a species tree S on $\mathcal{G}$, we denote by $c(\mathcal{B}, S)$ the number of B_i's that are inconsistent with S.

We now introduce a simple variant of the general supertree problem, where each gene tree indicates a single speciation.

Minimum Bipartition Inconsistency Supertree (MBIS) problem
Input: A set of bipartitions $\mathcal{B}$ of $\mathcal{G}$;
Output: A species tree S such that $c(\mathcal{B}, S)$ is minimum.

Given a binary gene tree T and a vertex x of T that is not an apparent duplication, we define the bipartition associated to x, denoted $B(T, x)$, as the bipartition with root y and internal vertices y_r (resp. y_ℓ), such that $L(y_r) = L(x_r)$ and $L(y_\ell) = L(x_\ell)$. Given a forest $\mathcal{F} = \{T_1, \ldots, T_m\}$ of binary gene trees, we denote by $\mathcal{B}(\mathcal{F})$ the set of bipartitions associated to all vertices of the trees of $\mathcal{F}$ that are not apparent duplications.

Theorem 3. *Let $\mathcal{F}$ be a forest of gene trees on $\mathcal{G}$ and k be the number of apparent duplications present in the trees of $\mathcal{F}$. Then, for any species tree S on $\mathcal{G}$, $d(\mathcal{F}, S) = k + c(\mathcal{B}(\mathcal{F}), S)$.*

Proof. Apparent duplications are associated to duplications for every species tree S. Hence the remaining duplications (there are $d(\mathcal{F}, S) - k$ such duplications, for a given species tree S) are not apparent duplications. Let x be such a vertex, belonging to a tree T of $\mathcal{F}$. As the reconciliation with S implies that x is a duplication, without loss of generality, we can assume that $M(x_r) = M(x)$ (M is the LCA mapping between T and S). Hence there are two elements $a, b \in \mathcal{G}$ such that $a, b \in L(x_r)$ and, if $M(x) = u$, $a \in L(u_r)$ and $b \in L(u_\ell)$. This implies that $S_{|L(x)}$ does not refine $B(T, x)$, and thus $B(T, x)$ is not consistent with S.

Conversely, let $B = B(T, x)$ be a bipartition of $\mathcal{B}(\mathcal{F})$ that is not consistent with S. It is clear that if $S_{|L(B)}$ refines B, then it can be transformed into B by contracting all internal edges that are not incident to its root. Thus, if $\mathcal{B}(\mathcal{F})$ is not consistent with S, there should be two elements $a, b \in \mathcal{G}$ that do not belong to a proper subtree of $S_{|L(B)}$, but belong to a proper subtree of B (say the subtree rooted at x_r). This implies that $M(x) = M(x_r)$ and then that x is no an apparent duplication but counts for a duplication when reconciled with S. □

This result shows that inferring a most parsimonious species tree for the duplication cost is equivalent to a restricted supertree problem that considers only very pathological input gene trees (bipartitions). Note however that despite the very restricted nature of its input trees, the MBIS Problem is NP-complete, which is deduced from the NP-completeness of the Minimum Dup. Problem. Another simple variant of the supertree problem, where input gene trees are rooted triplets (the Max. Triplet Consistency Supertree prob.), has been shown to be NP-complete [6].

The link between the two problems has the interesting consequence that heuristics for the supertree problem are then natural candidate heuristics for the Minimum Duplication Problem. In particular, min-cut based heuristics such as those developed in [29,27] can be directly applied to bipartitions (see Section 5). Such heuristics can then be seen as greedy approaches to the Minimum Duplication Problem, that, as far as we know, has never been used for the Minimum Dup. Problem, while it follows very naturally from its description as a supertree problem. The resulting species tree can then be used as a starting point for local-search algorithms such as the one presented in [2].

Minimum Duplication trees and compatible trees. A gene tree forest $\mathcal{F}$ is said to be a *Minimum Duplication forest* (from now an *MD-forest*) if there exists a species tree S such that $d(\mathcal{F}, S)$ is exactly the number of apparent duplications present in the trees of $\mathcal{F}$. In such case, $\mathcal{F}$ is said to be *MD-consistent* with S.

Theorem 4. *Deciding whether a forest of gene trees $\mathcal{F}$ is an MD-forest and computing the set of all species trees S such that $\mathcal{F}$ is MD-consistent with S can be done in polynomial time and space.*

Proof. Assume that $\mathcal{F}$ contains p vertices that are non-apparent duplications, we first note that $\mathcal{B}(\mathcal{F})$ contains $O(p)$ bipartitions. Following Theorem 3, the problem of deciding if $\mathcal{F}$ is an MD-forest reduces to deciding if $c(\mathcal{B}(\mathcal{F}), S) = 0$. Several algorithms exist that answer this question in polynomial when the input consists of rooted triplets [1] or unconstrained rooted binary trees [9,20]. However, these algorithms can be easily adapted to our situation. For example, the algorithm of [1], as described in [30], can be used if we simply define the edges of connectivity graph as follows: for two elements $i, j \in \mathcal{G}$, there is an edge between i and j if and only if there is a bipartition B of $\mathcal{F}$ such that i and j belong to the same proper subtree of B.

To compute the set of all species trees such that $\mathcal{F}$ is MD-consistent with S, we can use the polynomial time and space algorithm of [9] by replacing each subtree rooted at a non-binary vertex x of the bipartitions $\mathcal{B}(\mathcal{F})$, with leaf set $L(x) = \{i_1, \ldots, i_k\}$ such that $i_1 < i_2 < \ldots < i_k$ by the caterpillar tree $(i_1, (i_2, \ldots (i_{k-1}, i_k), \ldots))$. □

We now provide a simple combinatorial characterization of MD-trees and MD-forests in terms of triplets of species. A vertex of T is said to *split* three species $\{a, b, c\}$, into $\{a, b; c\}$ if the genome set of one of its children contains a and b but not c, and the genome set of its other child contains c but neither a nor b. Let x and y be two vertices of a gene tree T, that are non-apparent duplications. They *disagree* on a triplet $\{a, b, c\}$ of species if they split $\{a, b, c\}$ in different ways (say $\{a, b; c\}$ and $\{a, c; b\}$ for example). A gene tree T on $\mathcal{G}$ is *compatible* if no pair of non-apparent duplication vertices disagrees on any triplet of species. For a given species tree S on $\mathcal{G}$, T is said to be a *compatible gene tree consistent with* S if it is compatible, and every triplet of species $\{a, b, c\}$ is split in the same way by the LCA of these species in S and by any non-apparent duplication vertex of T that split them. These definitions extend naturally to a forest of gene trees.

Theorem 5. *Let $\mathcal{F}$ be a gene tree forest on $\mathcal{G}$, and S be a species tree for $\mathcal{G}$. Then $\mathcal{F}$ is a compatible gene tree forest consistent with S if and only if $\mathcal{F}$ is an MD-forest consistent with S.*

Proof. We consider a compatible gene tree T, as the proof generalizes in a straightforward way to forests.

Suppose first that T is not a compatible gene tree. Then there are two non-apparent duplication vertices v and w that split a triplet of species $\{a, b, c\}$ into two different ways, say $\{a, b; c\}$ for v and $\{a, c; b\}$ for w. If $\{a, b, c\}$ are split into $\{a, b; c\}$ in a species tree S, then w is a duplication vertex as it maps to the same vertex of S than its child that contains leaves labeled by a and c, and then T is not an MD-tree. Similarly, v is a duplication vertex if $\{a, b, c\}$ are split into $\{a, c; b\}$ in S and both v and w are duplication vertices if $\{a, b, c\}$ are split into $\{b, c; a\}$ in S.

Suppose that T is a compatible gene tree that is not consistent with S. Then there is a triplet $\{a, b, c\}$ of elements of $\mathcal{G}$ and a non-apparent duplication vertex v of T that splits $\{a, b, c\}$ in a different way than they are in S. W.l.o.g, let assume that v splits them into $\{a, b; c\}$ while in S they are split into $\{b, c; a\}$.

Then v is a duplication vertex, as it maps to the same vertex of S than its child that contains leaves labeled by a and b. Therefore, as T contains a vertex that is a duplication vertex but not an apparent duplication vertex, T is not an MD-tree consistent with S.

Conversely, suppose that T is not an MD-tree consistent with S. Then there is a vertex v in T that is a duplication vertex but not an apparent duplication vertex. As v is a duplication vertex, v maps to the same vertex of S than one of its child v_ℓ or v_r, let say its left vertex v_ℓ. Moreover, as v is not an apparent duplication, there are two leaves x_a and x_b of T_{v_ℓ} labeled respectively a and b, and a leaf x_c in T_{v_r} labeled c that imply that $\{a, b, c\}$ is split into $\{a, b; c\}$ by v, while $\{a, b, c\}$ is split into $\{b, c; a\}$ in S. Therefore, T is not a compatible gene tree consistent with S. □

Corollary 1. *A gene tree forest $\mathcal{F}$ on $\mathcal{G}$ is a compatible gene tree forest if and only if $\mathcal{F}$ is an MD-forest.*

Proof. We will prove the result on a single gene tree T. The generalization to a forest F is straightforward.
"⇐" This case follows directly from the previous proof.
"⇒" Suppose that T is a compatible tree. Then for any triplet $\{a, b, c\}$ of distinct elements of $\mathcal{G}$, any non-apparent duplication vertex of T splits them in the same way. Then there is a DLS-history H for T leading to a species tree S such that, for any triplet $\{a, b, c\}$ of distinct elements of $\mathcal{G}$, S splits $\{a, b, c\}$ in the same way than any non-apparent duplication vertex of T. It follows that any vertex v of T that is not an apparent duplication vertex is not a duplication vertex for H. □

From a theoretical point of view, the above results are interesting as they can be seen to be the MD-trees counterpart of a well known result about supertrees stating that deciding if, given a set of gene trees, there is a species tree that agrees with all of them, is equivalent to checking the same property on all triplets induced by these gene trees. From a practical point of view, triplets of species can be used to point at possibly ambiguous phylogenetic relationships and possibly misplaced genes in the gene tree, as we illustrate in the next section.

5 Experimental Results

We generated gene families, as in [7], using the species tree of 12 *Drosophilia* species given in [19] (including branch length) and a birth-and-death process, starting from a single ancestral gene, with four different gene gain/loss rates (expected number of events by million years): 0.02 (the highest rate identified in [19]), 0.05, 0.1 and 0.2. For each rate, we generated 250 gene trees, described in Table 1. Note that more than 95% of gene duplications lead to an apparent duplication vertex. Note also that the number of informative bipartitions (i.e. bipartitions with at least two leaves) induced by non-apparent duplication vertices decreases dramatically as the rate of gene gain/loss increases [1].

[1] All the material is available at: `http://www.cecm.sfu.ca/~cchauve/SUPP/RECOMB09`

Table 1. Characteristics of simulated gene trees. Considered bipartitions are those containing more than two species.

Rate	Nb. of Duplications	Nb. of Losses	Nb. of Genes	Nb. of Int. vertices	Nb. of Apparent duplications	Nb. of Bipartitions
0.02	1080	976	3014	2752	1057	831
0.05	2018	1366	3622	3360	1948	593
0.1	3126	1603	4376	4114	3007	358
0.2	6123	2552	7709	7447	5875	429

For each of the four datasets, we extracted the informative bipartitions induced by the non-apparent duplication vertices. We then used the Modified Min-Cut algorithm described in [27] to compute a species tree from these bipartitions. With rates 0.02 and 0.04, this species tree is the correct species tree, while with rate 0.1, it differs from the correct one by a single branch swap, and with rate 0.2, it differs from the correct one by the fact that two consecutive binary nodes have been replaced by a single quaternary node. The fit statistic associated to the inferred species tree, that measures how well it agrees with the bipartitions, is very high, ranging from 0.98 to 0.855 (maximum fit is 1). This shows the effectiveness of the supertree approach using bipartitions, at least on a dataset of relatively close species where few vertices indicating a speciation are false positive.

We also studied the phylogenetic signal given by triplets of species that were split by non-apparent duplication vertices. With rates 0.02 and 0.05, for each triplet of species, there is a phylogeny that appears in most cases. However, with rates 0.1 and 0.2, among the triplets that appear a significant number of times (at least 50 times), the ones where the dominant phylogeny appears in less than 90% of the bipartitions splitting this triplet, contain the two species involved in the branch swap or species involved in the unresolved node that differs from the correct species tree. This illustrates the interest in using triplets of species that are split by non-apparent duplication vertices to point at possible locations of an inferred species tree that are associated with a weaker phylogenetic signal.

6 Conclusion

In this paper, we show that minimizing losses is a more constraining criterion than minimizing duplications for reconciliation. This highlights the importance of the former criterion from a combinatorial point of view, although it has been rarely considered alone in reconciliation approaches. Our second main result relates the problem of inferring a species tree minimizing duplications (given a set of gene trees), to a supertree problem. This link has important implications, as it allows, for example, to use min-cut based algorithms to infer a species tree from a set of gene trees. Moreover, this link with supertree problems allowed us to highlight properties of gene trees that could be exploited for gene tree correction. Indeed, a major problem with reconciliation, and its generalization

to an unknown species tree, is that errors in gene trees usually lead to erroneous duplication/loss histories, and potentially to a wrong species tree. Therefore, eliminating a number of potentially misleading gene copies is an important preliminary step to any reconciliation approach. In this context, non-apparent duplications, or triplets leading to contradictory phylogenetic informations, may point at gene copies that are possibly erroneously placed in the gene tree. Our preliminary experimental results tend to support this strategy for pruning gene trees.

References

1. Aho, A.V., Sagiv, Y., Szymanski, T.G., Ullman, J.D.: Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J. Comput. 10, 405–421 (1981)
2. Bansal, M.S., Burleigh, J.G., Eulenstein, O., Wehe, A.: Heuristics for the gene-duplication problem: A $\Theta(n)$ speed-up for the local search. In: Speed, T., Huang, H. (eds.) RECOMB 2007. LNCS (LNBI), vol. 4453, pp. 238–252. Springer, Heidelberg (2007)
3. Arvestad, L., Berglung, A.-C., Lagergren, J., Sennblad, B.: Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In: RECOMB 2004, pp. 326–335 (2004)
4. Blomme, T., Vandepoele, K., De Bodt, S., Silmillion, C., Maere, S., van de Peer, Y.: The gain and loss of genes during 600 millions years of vertebrate evolution. Genome Biol. 7, R43 (2006)
5. Bonizzoni, P., Della Vedova, G., Dondi, R.: Reconciling a gene tree to a species tree under the duplication cost model. Theoret. Comput. Sci. 347, 36–53 (2005)
6. Bryant, D.: Hunting for trees, building trees and comparing trees: theory and methods in phylogenetic analysis. Ph.D. thesis, Dept. of Math., Univ. of Canterbury, New Zealand (1997)
7. Chauve, C., Doyon, J.-P., El-Mabrouk, N.: Gene family evolution by duplication, speciation and loss. J. Comput. Biol. 15, 1043–1062 (2008)
8. Chen, K., Durand, D., Farach-Colton, M.: NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. 7, 429–444 (2000)
9. Constantinescu, M., Sankoff, D.: An efficient algorithm for supertrees. J. Classification 12, 101–112 (1995)
10. Cotton, J.A., Page, R.D.M.: Rates and patterns of gene duplication and loss in the human genome. Proc. R. Soc. Lond. B 272, 277–283 (2005)
11. Demuth, J.P., De Bie, T., Stajich, J., Cristianini, N., Hahn, M.W.: The evolution of mammalian gene families. PLoS ONE 1, e85 (2006)
12. Doyon, J.-P., Chauve, C., Hamel, S.: Algorithms for exploring the space of gene tree/species tree reconciliations. In: Nelson, C.E., Vialette, S. (eds.) RECOMB-CG 2008. LNCS (LNBI), vol. 5267, pp. 1–13. Springer, Heidelberg (2008)
13. Durand, D., Haldórsson, B.V., Vernot, B.: A hybrid micro-macroevolutionary approach to gene tree reconstruction. J. Comput. Biol. 13, 320–3354 (2006)
14. Eichler, E.E., Sankoff, D.: Structural dynamics of eukaryotic chromosome evolution. Science 301, 793–797 (2003)
15. Eulenstein, O., Mirkin, B., Vingron, M.: Comparison of annotating duplication, tree mapping, and copying as methods to compare gene trees with species trees. In: Mathematical hierarchies and biology. DIMACS Series Discrete Math. Theoret. Comput. Sci., vol. 37, pp. 71–93 (1997)

16. Gorecki, P., Tiutyn, J.: DLS-trees: a model of evolutionary scenarios. Theoret. Comput. Sci. 359, 378–399 (2006)
17. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28, 132–163 (1979)
18. Hallett, M.T., Lagergren, J.: New algorithms for the duplication-loss model. In: RECOMB 2000, pp. 138–146 (2000)
19. Hahn, M.W., Han, M.V., Han, S.-G.: Gene family evolution across 12 Drosophilia genomes. PLoS Genet. 3, e197 (2007)
20. Henzinger, M.R., King, V., Warnow, T.: Constructing a Tree from Homeomorphic Subtrees, with Applications to Computational Evolutionary Biology. Algorithmica 24, 1–13 (1999)
21. Lynch, M., Conery, J.S.: The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155 (2000)
22. Ma, B., Li, M., Zhang, L.: From gene trees to species trees. SIAM J. Comput. 30, 729–752 (2000)
23. Ohno, S.: Evolution by gene duplication. Springer, Heidelberg (1970)
24. Page, R.D.M.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43, 58–77 (1994)
25. Page, R.D.M., Charleston, M.A.: Reconciled trees and incongruent gene and species trees. Mathematical hierarchies and biology. DIMACS Series Discrete Math. Theoret. Comput. Sci. 37, 57–70 (1997)
26. Page, R.D.M.: GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14, 819–820 (1998)
27. Page, R.D.M.: Modified mincut supertrees. In: Guigó, R., Gusfield, D. (eds.) WABI 2002. LNCS, vol. 2452, pp. 537–551. Springer, Heidelberg (2002)
28. Sanderson, M.J., McMahon, M.M.: Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7, S3 (2007)
29. Semple, C., Steel, M.: A supertree method for rooted trees. Discrete Appl. Math. 105, 147–158 (2000)
30. Snir, S., Rao, S.: Using max cut to enhance rooted trees consistency. IEEE/ACM Trans. Comput. Biol. and Bioinform. 3, 323–333 (2006)
31. Steel, M.: The complexity of reconstructing trees from qualitative characters and subtrees. J. Classification 9, 91–116 (1992)
32. Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Natural history and evolutionary principles of gene duplication in fungi. Nature 449, 54–61 (2007)
33. Zhang, L.X.: On Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. J. Comput. Biol. 4, 177–188 (1997)
34. Zmasek, C.M., Eddy, S.R.: A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17, 821–828 (2001)