To appear in the proceedings of RECOMB-CG 2010, RECOMB Satellite Workshop on Comparative Genomics, Ottawa, October 9-11, 2010.

An approximation algorithm for computing a parsimonious first speciation in the gene duplication model

Aïda Ouangraoua[1,2,3], Krister M. Swenson[2,4], and Cedric Chauve[1,2]

[1] Department of Mathematics, SFU, Burnaby (BC), Canada cchauve@sfu.ca
[2] Lacim, Université du Québec à Montréal, Montréal (QC), Canada
[3] INRIA LNE, LIFL, Université Lille 1, France aida.ouangraoua@inria.fr
[4] Department of Mathematics and Statistics, University of Ottawa, Ottawa (ON), Canada thekswenson@gmail.com

Abstract. We consider the following problem: given a forest of gene family trees on a set of genomes, find a first speciation which splits these genomes into two subsets and minimizes the number of gene duplications that happened before this speciation. We call this problem the Minimum Duplication Bipartition Problem. Using a generalization of the Minimum Edge-Cut Problem, known as Submodular Function Minimization, we propose a polynomial time and space 2-approximation algorithm for the Minimum Duplication Bipartition Problem. We illustrate the potential of this algorithm on both synthetic and real data.

1 Introduction

Gene duplication is a fundamental evolutionary mechanism for important groups of eukaryotes such as vertebrates [3], insects [15], plants [19] or fungi [23]. Gene duplications, together with gene losses, result in *gene families*, which can contain several copies of a same gene in a given genome. Recent advances in methods for reconstructing phylogenetic trees for individual gene families, have resulted in large sets of accurate *gene trees* for eukaryote species, such as TreeFam [24]. Phylogenomics aims at reconstructing the evolution of *species (genomes)* by inferring a species tree for the set of genomes from a set of gene trees. The *Minimum Duplication Problem* (MD Problem), also known as the Gene Duplication Problem, is to infer, from a set of gene trees, a species tree that induces an evolutionary history with a minimum number of gene duplications. This problem is NP-hard, and is neither fixed-parameter tractable (FPT) nor approximable with a constant ratio [2, 12]. Recent advances in local search heuristics proved to be useful [1] and have been applied on several eukaryotic datasets with interesting results (see [19, 25]), but with no optimality guarantee.

Recently, Chauve and El-Mabrouk [7, 20] described a formal link between the Minimum Duplication Problem and a problem of supertrees, where, given a set of uniquely leaf-labeled gene trees (there is at most one copy of each gene in each genome), the goal is to reconstruct a species tree that disagrees with the minimum number of gene trees [6]. This problem — a version of the MD Problem restricted to uniquely leaf-labeled trees — is NP-hard too, even in the simple

case where each gene tree is a rooted triplet [4]. For supertree problems, greedy heuristics based on the principle of computing successive optimal speciations, modeled as edge-cuts in a graph whose vertices are the considered species, are now standard [18, 21]. In such heuristics, each Minimum Edge-Cut splits the set of considered species into two subsets that correspond to a speciation that results in two distinct lineages. Computing an optimal first speciation (a first speciation that disagrees with the least number of rooted triplets) is tractable, as the Minimum Edge-Cut problem is tractable. A complete species tree can then be obtained from a series of locally optimal speciations.

In the present work we consider the Minimum Duplication Bipartition (MDB) Problem: given a set of gene trees (where a gene occurs any number of times), find a bipartition of the considered genes (corresponding to a speciation) that minimizes the number of duplications that happened before this speciation. A first motivation for this problem is that it leads, as for supertree problems, to a natural greedy heuristics to reconstruct a species tree from a set of gene trees. Also, solving the Minimum Duplication Bipartition Problem can provide valuable information on the combinatorial nature of early speciations for large eukaryotic datasets with respect to gene duplications. Our main result is a polynomial time 2-approximation algorithm for the Minimum Duplication Bipartition Problem that generalizes the Minimum Edge-Cut approach used in supertrees and relies on Submodular Function Minimization [10]. Although our focus here is mostly theoretical, and explores the combinatorial structure of parsimonious first speciations, we also provide experimental results, on small datasets, which illustrate the potential of our approach.

We first define, in Section 2, gene trees, species trees, duplications, and the optimization problems considered in this paper as well as our motivation to introduce the MDB Problem. In Section 3 we describe our 2-approximation algorithm[1]. Our experimental results are described in Section 4.

2 Preliminaries: objects, problems, background

In this section and the sequel, $\mathcal{G} = \{1, 2, \ldots, k\}$ is a set of integers representing k different species (genomes).

Gene and species trees, bipartitions. A *species tree* on $\mathcal{G}$ is a tree with exactly k leaves, where each $i \in \mathcal{G}$ is the label of a single leaf. A tree is binary if every internal vertex has exactly two children. A *gene tree* on $\mathcal{G}$ is a binary tree where each leaf is labeled with an integer from $\mathcal{G}$. A gene tree is a formal representation of a phylogenetic tree of a gene family, where each leaf labeled i represents a gene which is a member of the gene family located on genome i. A gene tree is *uniquely leaf-labeled* if no two leaves have the same label. A gene tree is a *rooted triplet* if it has exactly three leaves.

Given a tree T and a vertex x of T whose leaves are labeled by integers from $\mathcal{G}$, we denote by $L(x)$ (resp. $L(T)$) the subset of $\mathcal{G}$ defined by the labels of the leaves of the subtree of T rooted at x (resp. the leaves of T). If x is not a leaf, we denote by x_ℓ and x_r the two children of x.

[1] Missing proofs are available at `http://www.cecm.sfu.ca/~cchauve/SUPP/RECOMBCG10`.

A *bipartition* B on a set S is a partition of S into two subsets. We represent a bipartition by a, possibly non-binary, species tree on S containing exactly three internal vertices — the root v and its two children v_ℓ and v_r — such that $L(v_\ell) \cap L(v_r) = \emptyset$.

Reconciliation between Gene Trees and Species Trees. The *Lowest Common Ancestor Mapping (LCA mapping)* is central in the problem of reconciling a gene tree and a species tree. Given a gene tree T and a species tree S on $\mathcal{G}$, the LCA mapping M maps vertices of T to vertices of S as follows: for a vertex x of T, $M(x) = v$ is the unique vertex of S such that $L(x) \subseteq L(v)$ and v is either a leaf of S or $L(x)$ is not included in the leaf set of any child of v. In other words, v is the deepest among all possible. A vertex x of T is then a *duplication with respect to* S if $M(x) = M(x_r)$ and/or $M(x) = M(x_\ell)$; otherwise, x is called a *speciation with respect to* S (see Figure 1). The same definitions apply to a forest F of gene trees on $\mathcal{G}$. The *duplication cost* of F given S denoted by $d(F, S)$ is the number of vertices of F that are duplications with respect to S. Note that the definitions of duplication and speciation apply to a species tree that is a bipartition on the set $\mathcal{G}$, as these definitions do not depend on the species tree being binary.

If $L(x_\ell) \cap L(x_r) \neq \emptyset$, then x is a duplication vertex with respect to any species tree S on $\mathcal{G}$. Such a vertex is called an *apparent duplication*. Vertices of F that are not apparent duplication are called *non-apparent duplication*.

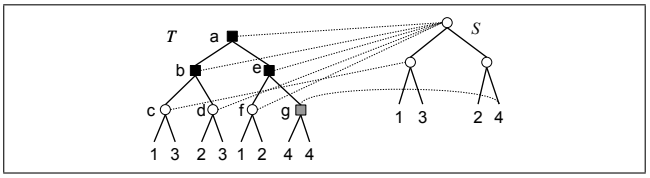

Fig. 1. A gene tree T and a species tree S on a set of genome $\mathcal{G} = \{1, 2, 3, 4\}$. The LCA mapping from vertices of T to vertices of S is indicated by dashed lines linking vertices. The vertices of T that are duplications with respect to S are represented by square vertices; the black colored square vertices correspond to pre-duplications while the grey colored square vertices are the duplications that are not pre-duplications. Here, the first speciation of S is the bipartition with root v such that $L(v_l) = \{1, 3\}$ and $L(v_r) = \{2, 4\}$. a, b and g are apparent duplications.

Inferring parsimonious species trees and speciations. It is well known that $d(F, S)$ is the minimum number of gene duplication events required in any evolutionary scenario that resulted in F (see [7, 11] and references there), which leads to the following optimization problem called the Minimum Duplication Problem (MD Problem): given a gene tree forest F find a species tree S such that $d(F, S)$ is minimum.

The MD Problem is NP-hard [16], even in the case where every gene tree is a uniquely leaf labeled and rooted triplet [4], in which case it is in fact equivalent to a supertree problem called the Minimum Rooted Triplet Inconsistency (MRTI)

Problem. This link with supertrees is important as recent hardness results on the MRTI Problem imply that first, the MD Problem is W[2]-hard, and thus not FPT [2, 12] — contrary to what was believed ten years ago [22] — but also that it cannot be approximated within a constant ratio unless P=NP [6]. Hence, for solving the MD Problem, one has to rely on exponential time algorithms such as [13] or local-search methods with no optimality guarantee [1, 25].

In [7], it was shown that the MD Problem is in fact a slight variant of a supertree problem (see also [20] that explores the link between gene duplications and supertrees). Greedy heuristics for hard supertree problems based on computing successive speciations events have proved to be effective [21], and in [7], an application of such a heuristic showed promising results on synthetic data. This motivates the introduction of the main problem we study in this paper.

Given a gene tree forest F on $\mathcal{G}$ and a bipartition B on $\mathcal{G}$ with root v, a duplication x of F with respect to B is said to *precede the first speciation with respect to* B if $M(x) = v$. Such vertices are called *pre-duplications* (e.g. a, b and e in Figure 1). We denote by $d_1(F, B)$, the number of pre-duplications of F with respect to B.

Minimum Duplication Bipartition Problem (MDB Problem):
Input: A gene tree forest F on $\mathcal{G}$;
Output: A bipartition B on $\mathcal{G}$ such that $d_1(F, B)$ is minimum.

Before discussing previous works, we state an obvious, but very useful, property related to duplication vertices of a forest of gene trees F on $\mathcal{G}$.

Property 1. Let x be a vertex of F. Given a bipartition B on $\mathcal{G}$ with root v, x is a pre-duplication with respect to B if and only if there exists a pair $\{s, t\} \subseteq L(v_\ell) \times L(v_r)$ such that $\{s, t\} \subseteq L(x_\ell)$ or $\{s, t\} \subseteq L(x_r)$.

As far as we know, the MDB problem was introduced in [22], where an exponential time algorithm was proposed. It was also shown in [7], although not formally stated, that if there exists a bipartition such that all pre-duplications are apparent duplications, then such a bipartition can be computed in polynomial time and space. The hardness of the MDB Problem is still open, but preliminary results showing the hardness of a slight variant using quadruplets as input gene trees (G. Blin and S. Vialette, personal communication), suggest it may be NP-complete. In [8], it was shown that if F contains a single gene tree, the MDB Problem is 3-approximable. However, in the more general case of a forest F with t gene trees, the approximation ratio is not constant: if a parsimonious first speciation implies d duplications, then the algorithm described in [8] computes a first speciation that can imply up to $2d + t$ duplications. In the present work, we show that the MDB Problem can be approximated with a constant ratio of 2 in polynomial time and space.

Related optimization problems. Given a connected graph $G = (V, E)$, an *edge-cut* of G is an edge set $E' \subseteq E$ whose removal disconnects the graph G. A bipartition B with root v on the set of vertices of G induces a unique edge-cut of G, denoted by $E_G(B)$, composed of the edges $(s, t) \in E$ such that $s \in L(v_\ell)$ and $t \in L(v_r)$. So $E_G(B) = \{(s, t) \in E \mid s \in L(v_\ell),\ t \in L(v_r)\}$. Hence, a subset X of V induces a bipartition on V with root v such that $L(v_\ell) = X$ and $L(v_r) = V - X$. We

denote this bipartition by $B_V(X)$ and the edge-cut of G induced by $B_V(X)$ is denoted by $E_G(X)$ ($E_G(X) = E_G(B_V(X))$).

The *Minimum Edge-Cut (MEC) Problem* asks for a bipartition on the vertices of G inducing an edge-cut of G of minimum cardinality. If the edges of G are labeled on a given set Σ of labels, given an edge-cut E' of G, the *label-set* of E' denoted by label(E') is the subset of Σ composed of the labels of the edges in E'. The following cut problem is a natural generalization of the MEC Problem and is essential in our algorithm:

Minimum Labeled-Edge-Cut (MLEC) Problem:
Input: A connected edge-labeled graph $G = (V, E)$;
Output: A bipartition B on V such that the cardinality of label($E_G(B)$) is minimum.

A *set function* is a function $f : 2^V \rightarrow \mathbb{R}$ defined from the set of the $2^{|V|}$ subsets of a finite set V onto the real numbers $\mathbb{R}$. The set V is called the *ground set* of f. The Set Function Minimization Problem asks to find a non-empty subset X of V such that $f(X)$ is minimum. A *submodular function* is a set function f with ground set V such that for any subsets A and B of V, $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$. Several combinatorial optimization problems have been linked to submodular functions [10], in particular the MEC Problem. Given a submodular function f, the following optimization problem, which is tractable [14], is a special case of the Set Function Minimization Problem:

Submodular Function Minimization (SFM) Problem:
Input: A submodular function $f : 2^V \rightarrow \mathbb{R}$ with ground set V;
Output: A non-empty subset X of V such that $f(X)$ is minimum.

A *hypergraph* is a pair (V, E) where V is a set of vertices and E is a set of non-empty subsets of V called *hyperedges*. Given a hypergraph $G = (V, E)$, a bipartition B on V with root v induces a *hyperedge-cut* of G defined by the following subset of E:
$E_G(B) = \{e \in E \mid e \cap L(v_\ell) \neq \emptyset \text{ and } e \cap L(v_r) \neq \emptyset\}$.

Minimum Hypergraph Cut (MHC) Problem:
Input: A hypergraph $G = (V, E)$;
Output: A bipartition B on V such that the cardinality of $E_G(B)$ is minimum.

3 A 2-approximation algorithm for the MDB Problem

3.1 A Set Function Minimization Problem

In the following, given a gene tree forest, we label arbitrarily its internal vertices with a set Σ of labels in such a way that no two internal vertices have the same label.

Given a gene tree forest F, we define the *edge-labeled graph* $H(F) = (V, E)$ associated to F as follows (see Figure 2): $V = L(F)$ and there is an edge labeled with $a \in \Sigma$ between two vertices s and t of $H(F)$ if and only if there exists an internal vertex x of F labeled with a such that $\{s, t\} \subseteq L(x_\ell)$ or $\{s, t\} \subseteq L(x_r)$.

Lemma 1. *Let F be a gene tree forest on $\mathcal{G}$. If B is a bipartition on $L(F)$, then the set of labels of the pre-duplications of F with respect to B is exactly the set* label($E_{H(F)}(B)$).

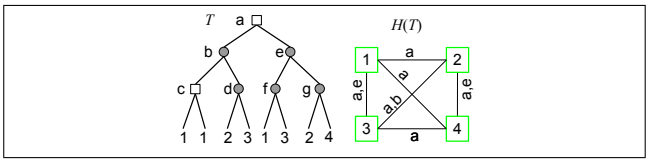

Fig. 2. A gene tree T on the set of genomes $\mathcal{G} = \{1, 2, 3, 4\}$ and the corresponding edge-labeled graph $H(T)$. Apparent duplication vertices of T appear as square vertices.

The MLEC Problem can be reduced to a Set Function Minimization Problem as follows: given an edge-labeled graph $G = (V, E)$, we define the *cut-set function* $f_G : 2^V \to \mathbb{R}$ as a function from the set of the subsets of V onto $\mathbb{R}$ such that, for any subset X of V, $f_G(X)$ is the cardinality of the set of labels $\text{label}(E_G(X))$. It is easy to see that solving the MLEC Problem on G can be achieved by minimizing f_G. In the following, given a gene tree forest F, we simply denote by f_G the cut-set function induced by an edge-labeled graph $G(F)$ associated to F.

Unfortunately, the cut-set function f_G associated to an edge-labeled graph G is not always a submodular function. For example (Figure 3), consider a single gene tree T with four leaves $\{1, 2, 3, 4\}$ and three internal vertices a, b and c whose sets of children are respectively $\{b, c\}$, $\{1, 3\}$ and $\{2, 4\}$. The edge-labeled graph $H(T)$ associated to T has four vertices $\{1, 2, 3, 4\}$ and only two edges $(1, 3)$ and $(2, 4)$ labeled with a. If we consider the subsets $A = \{1, 4\}$ and $B = \{2, 4\}$ of the set of $\{1, 2, 3, 4\}$, we see that $f_{H(T)}(A) = 1$, $f_{H(T)}(B) = 0$, $f_{H(T)}(A \cup B) = 1$ and $f_{H(T)}(A \cap B) = 1$. Then, $f_{H(T)}(A) + f_{H(T)}(B) = 1 < f_{H(T)}(A \cup B) + f_{H(T)}(A \cap B) = 2$, and $f_{H(T)}$ is not a submodular function which proves the following property:

Property 2. There exist gene trees forest F such that the cut-set function f_H, where $H = H(F)$, is not a submodular function.

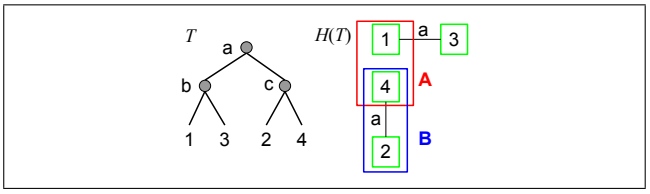

Fig. 3. Illustration of Property 2: a gene tree T (left) and the corresponding graph $H(T)$ (right), and two vertex sets A and B that contradict the submodularity of the cut-set function f_H.

3.2 Submodular Function Minimization

In this section, we prove our main result.

Theorem 1. *Let F be a gene tree forest with n vertices, on a set $\mathcal{G}$ of k genomes. If a most parsimonious bipartition B^* on $L(F)$ has cost $d_1(F, B^*) = d$, then it is possible to compute in time $O(kn)$ a bipartition B s.t. $d(F, B) \leq 2d$.*

The approximation results from transforming the graph $H(F)$ associated to a gene tree forest F in order to obtain a new edge-labeled graph $J(F)$ such that the set function f_J is a submodular function.

Characterization of non-submodularity. Given two subsets A and B of $L(F)$, we define the following four subsets of $L(F)$: $\mathrm{AB}_1 = A - B$, $\mathrm{AB}_2 = B - A$, $\mathrm{AB}_3 = L(F) - (A \cup B)$ and $\mathrm{AB}_4 = A \cap B$. Note that the intersection of any two of these subsets is empty, and the union of all of them is $L(F)$.

Given an edge-labeled graph G whose set of vertices is $L(F)$, we then define six sets of labels $\mathrm{AB}_{1,2}(G)$, $\mathrm{AB}_{1,3}(G)$ $\mathrm{AB}_{1,4}(G)$, $\mathrm{AB}_{2,3}(G)$, $\mathrm{AB}_{2,4}(G)$ and $\mathrm{AB}_{3,4}(G)$ as follows: given two integers i and j such that $1 \leq i < j \leq 4$, the set $\mathrm{AB}_{i,j}(G)$ is the set of labels of edges (s, t) in G such that $s \in \mathrm{AB}_i$ and $t \in \mathrm{AB}_j$ (see Figure 4).

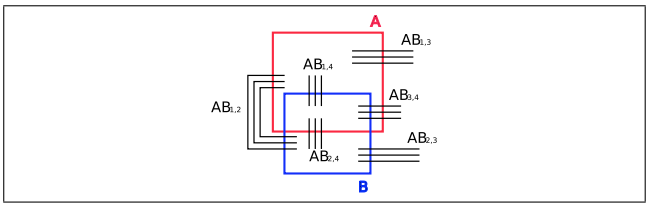

Fig. 4. Illustrations of the definition of the six sets of labels $\mathrm{AB}_{i,j}(G)$ $(1 \leq i < j \leq 4)$ associated to two subsets A and B of the set $L(F)$ of leaves of a gene tree forest F and an edge-labeled graph G whose set of vertices is $L(F)$: the A and B are represented by two squares and the solid black lines correspond to edges of G whose labels belong to one of the six sets according to membership of their extremities to the sets AB_i $(1 \leq i \leq 4)$.

In the following, given a gene tree forest F, two subsets A and B of $L(F)$, and the edge-labeled graph $H(F)$ associated to F, we simply denote any set $\mathrm{AB}_{i,j}(H(F))$ by $\mathrm{AB}_{i,j}$.

Lemma 2. *Let F be a gene tree forest and f_H be the cut-set function associated to $H(F)$. If f_H is not a submodular function then there exist two subsets A and B of $L(F)$ and an internal vertex x of F labeled with $a \in \Sigma$ that is a non-apparent duplication, such that at least one of the two following configurations holds:*

- *(1)* $a \in \mathrm{AB}_{1,3} \cap \mathrm{AB}_{2,4}$ *and* $a \notin \mathrm{AB}_{1,2} \cup \mathrm{AB}_{1,4} \cup \mathrm{AB}_{2,3} \cup \mathrm{AB}_{3,4}$ *or*
- *(2)* $a \in \mathrm{AB}_{2,3} \cap \mathrm{AB}_{1,4}$ *and* $a \notin \mathrm{AB}_{1,2} \cup \mathrm{AB}_{1,3} \cup \mathrm{AB}_{2,4} \cup \mathrm{AB}_{3,4}$.

Modification of the edge-labeled graph $H(F)$. We now present a modification of $H(F)$ that leads to our 2-approximation algorithm for the MDB Problem. The goal is to modify $H(F)$ into a new edge-labeled graph $J(F)$ such that the

two configurations of Lemma 2 never hold, leading to a cut-set function that is submodular. The transformation is as follows: for each vertex x of F (say labeled with $a \in \Sigma$) that is not an apparent duplication, reassign to edges (s,t) of $H(F)$ labeled with a such that $\{s,t\} \subseteq L(x_r)$, a new label that is different from a (see Figure 5).

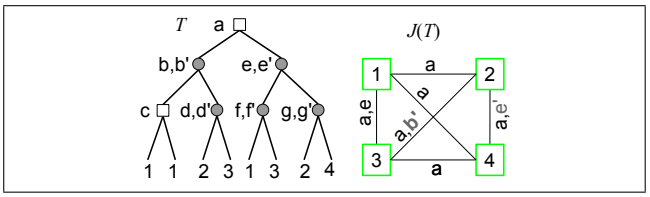

Fig. 5. A gene tree T on the set of genomes $\mathcal{G} = \{1,2,3,4\}$ and the corresponding modified edge-labeled graph $J(F)$, where modified labels are displayed in grey color. Apparent duplication vertices of T appear as square vertices.

More precisely, given a gene tree forest F we now label each non apparent duplication vertex of F with an ordered pair of labels in Σ, and define the *edge-labeled graph* $J(F) = (V,E)$ as follows (see Figure 5): the set V of vertices of $J(F)$ is $L(F)$, and there is an edge (s,t) labeled $a \in \Sigma$ in $J(F)$ if and only if there exists an internal vertex x of F such that:

- either $\{s,t\} \subseteq L(x_\ell)$ or $\{s,t\} \subseteq L(x_r)$ and x is an apparent duplication labeled with a,
- or $\{s,t\} \subseteq L(x_\ell)$ and x is a non-apparent duplication labeled with (a,a'),
- or $\{s,t\} \subseteq L(x_r)$ and x is a non-apparent duplication labeled with (a',a).

Lemma 3. *Given a gene tree forest F, the cut-set function f_J associated to $J(F)$ is a submodular function.*

Proof of Theorem 1. For any bipartition B on $L(F)$, the cardinality of label$(E_H(B))$ (resp. label$(E_J(B))$) is denoted by $d_H(B)$ (resp. $d_J(B)$).

We first prove that, for any bipartition B on $L(F)$ with root v, $d_H(B) \leq d_J(B) \leq 2 * d_H(B)$. It is relatively straightforward. First, note that $E_H(B) = E_J(B)$ since $J(F)$ differs from $H(F)$ only in the fact that the labels of some edges have been changed. Next, a label $a \in$ label$(E_H(B))$ corresponds to at least one but at most two labels in label$(E_J(B))$: if a is the label of an apparent duplication, then we have $a \in$ label$(E_H(B)) \Leftrightarrow a \in$ label$(E_J(B))$; if a is the label of a non-apparent duplication x, then the vertex x has two labels — a and a' — such that one or both belong to label$(E_J(B))$. This proves that $d_H(B) \leq d_J(B) \leq 2 * d_H(B)$.

Now, let B' be a bipartition on $L(F)$ inducing an optimal labeled edge-cut of $H(F)$ (i.e $d_H(B')$ is minimum). For any bipartition B on $L(F)$, if $d_H(B) > 2 * d_H(B')$ then B cannot induce an optimal labeled edge-cut of $J(F)$: indeed, if $d_H(B) > 2 * d_H(B')$, as $d_H(B) \leq d_J(B)$ and $d_J(B') \leq 2 * d_H(B')$, we have $d_J(B) > 2 * d_H(B') \geq d_J(B')$. Hence, for any bipartition B on $L(F)$ that is

optimal for $J(F)$, we have $d_H(B) \leq 2 * d_H(B')$. This completes the proof that computing an optimal labeled edge-cut for $J(F)$ achieves a ratio 2 approximation for the MDB Problem.

The complexity stated in Theorem 1 follows. The $O(kn)$ time complexity is derived from the reduction of the minimization of the function f_J to the MHC Problem as follows. Given a graph $G = (V, E)$ with edges labeled on a set Σ of labels, we define the hypergraph $G_h = (V_h, E_h)$ such that $V_h = V$ and, for each label $a \in \Sigma$, G_h contains an hyperedge composed of all vertices s in V that belong to an edge labeled by a. Given a gene tree forest F and a bipartition B on $L(F)$, if we consider the graph $J = J(F)$, it is obvious that the cardinality of label$(E_J(B))$ is equal to the cardinality of $E_{J_h}(B)$. Hence, the minimization of f_J can be reduced to the MHC Problem on J_h. The time complexity given in the theorem then follows from the algorithm described in [17] solving the MHC Problem on a hypergraph G in time and space $O(kn)$ where k (resp. n) is the number of vertices (resp. hyperedges) of G.

Additional remarks. If there exists at least one parsimonious first speciation B' such that all the corresponding pre-duplications are apparent duplications in F, then our algorithm computes a parsimonious bipartition, and not an approximation. Indeed, in such a case, B' induces an optimal cut for both $H(F)$ and $J(F)$. Note however that this does not imply that the cut-set function for $H(F)$ is submodular.

Conversely, the approximation ratio of 2 is tight, as illustrated in Figure 6. This example is easily expanded to any size by extending the top and bottom rows of the graph $J(F)$, adding pairs of labeled edges between the rows and the suitable number of unlabeled edges between the vertices of the top row and between the vertices of the bottom row.

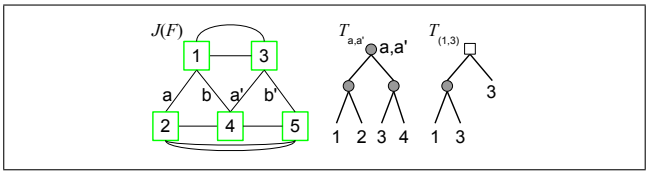

Fig. 6. The forest F is built from trees corresponding to the labeled and unlabeled edges in $J(F)$. For a pair of edges in $J(F)$ labeled with two labels a, a', say (p, q) and (u, v), we add to F the tree $((p, q), (u, v))$. For an unlabeled edge (u, v), we add to F the tree $((u, v), v)$. The tree $T_{a,a'}$ (resp. $T_{(1,3)}$) corresponds to the pair of edges labeled with a, a' (resp. the unlabeled edge $(1, 3)$).

4 Experimental results

We performed three different experiments[2]. First, on several datasets of simulated gene families on 12 species (genomes) we studied the ability of the greedy approach — that infers a species tree by computing successive parsimonious

[2] Data and results available at `http://www.cecm.sfu.ca/~cchauve/SUPP/RECOMBCG10`

speciations — to recover the exact species tree, using an exhaustive exploration of all possible speciations at each step (which was possible due to the fact we considered only 12 species). Next, on the same simulated datasets, we replaced the exhaustive exploration of all possible speciations at each step by our 2-approximation algorithm for computing a parsimonious speciation. Last, we performed the same experiment, using only the approximation algorithm on a large dataset of gene families on 23 fungal genomes.

Datasets. We uploaded from the Fungal Orthogroup Repository[3] the 6808 fungal gene trees containing genes belonging to at least three different species, among 23 fungal species.

We also analyzed the four synthetic datasets that were studied in [7]; each dataset contains 100 gene trees. Each gene tree was generated from a single ancestral gene with duplication (birth) and loss (death) rates computed using the software CAFE [9] from real Drosophilia gene families [15].

To balance the fact that each gene tree originates from a single ancestral gene (and then has no duplication before the first speciation), and to consider datasets including gene families generated with different duplication/loss rates, we created duplications that happened before the first speciation by clustering the 400 gene trees of the four datasets into 100 clusters of random size, and for every such cluster of a given size, say k, we generated a random binary tree with k leaves and replaced each leaf of this tree by a gene tree of the cluster, which amounts to creating around 4 duplications that precede the first speciation. We repeated this experiment ten times, generating ten different datasets.

Results. On each of the simulated datasets, we first observed that the greedy heuristic that computes successive parsimonious speciations using an exhaustive exploration lead to the exact species tree, i.e. the one that had been used to generate the synthetic gene trees. Despite the relatively modest size of our synthetic datasets (12 species), it illustrates the potential of this greedy heuristic in a phylogenomics context, especially as the heuristic described in [7] inferred a slightly incorrect species tree for the two datasets with the highest duplication/loss rates. Moreover, we observed that our approximation algorithm provided the exact species tree every time. We believe this result can be explained by the fact that most duplications that occurred during the generation of the gene trees are apparent duplications.

Due to the large number of species in the real data set — 6808 fungal gene trees from 23 species — we applied only our approximation algorithm. We observed that the inferred species tree is the one widely accepted in yeasts phylogenomics[4]. We also noticed that a significant number of speciations satisfied the property that all the associated pre-duplications were apparent duplications. Such speciations are then parsimonious (see the discussion at the end of the previous section). Moreover, aside from one branch (leading to node 28 of the species tree, where 103 pre-duplications are non-apparent) all other branches are associated with very few such non-apparent pre-duplications, which suggests that

[3] Version 1.1, `http://www.broadinstitute.org/regev/orthogroups/`

[4] The species tree can be seen at `http://www.cecm.sfu.ca/~cchauve/SUPP/RECOMBCG10`.

they might be parsimonious. Providing the gene trees we analyzed are correct, this clearly suggests that traces of most duplications that occurred during yeast evolution are still visible today.

5 Conclusion

We showed that computing a parsimonious first speciation in the gene duplication model can be approximated in polynomial time with a ratio of 2. As far as we know this is the first time a constant approximation algorithm is proposed in relation with the problem of inferring species trees using gene duplications. This result was obtained by describing the problem in terms of edge-cuts in particular graphs, which can be computed in polynomial time through submodular function minimization. This algorithm is also a natural generalization of the classical minimum edge-cut algorithm that is used in supertree consistency problems, which is highlighted by its link with the Submodular Function Minimization Problem. Our preliminary experiments showed that both the approach of inferring a species tree by computing successive parsimonious speciations and our approximation algorithm for computing such speciations are promising, and we plan to apply them on larger datasets, like those that will soon be available from [5, 26].

From a theoretical point of view, the hardness of the Minimum Duplication Bipartition Problem is still an open problem, but we conjecture the problem is NP-complete. It is interesting to note that, as for to the Gene Duplication Problem, when there is a parsimonious first speciation whose pre-duplications are all apparent duplications, it can be detected in polynomial time. Also when F contains only uniquely leaf-labeled rooted triplets, the graph $H(F)$ does not need to be augmented as every label appears only once, and computing a parsimonious first speciation can be done by computing a minimum edge-cut in $H(F)$. However, as we showed in the proof of Property 1, this tractability property no longer holds when quadruplets whose root is a non-apparent duplication are considered instead of triplets, as the cut-set function is no longer submodular. The role of non-apparent duplications, especially with respect to the non-submodularity of the cut-set function of $H(F)$, seems to be fundamental to the hardness of the problem, in particular to the understanding of which families of gene tree forests are tractable or fixed-parameter tractable.

Acknowledgments. Work supported by an NSERC Discovery grant to C.C. and a fellowship from the ANR (project ANR-06-BLAN-0045) for A.O. We thank Tamon Stephen, Mukul Bansal and Sylvain Guillemot for useful discussions.

References

1. M.S. Bansal *et al.*. Heuristics for the gene-duplication problem: A $\Theta(n)$ speed-up for the local search. In *RECOMB 2007*, *LNCS* 4453, pp.238–252. Springer.
2. M.S. Bansal and R. Shamir. 2010. A Note on the Fixed Parameter Tractability of the Gene-Duplication Problem. Submitted.
3. T. Blomme *et al.*.Y. van de Peer. 2006. The gain and loss of genes during 600 millions years of vertebrate evolution. *Genome Biol.* 7:R43.

4. D. Bryant. 1997. 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.
5. J. G. Burleig *et al.*. 2010. Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 discordant gene trees *Systematic Biology,* in press.
6. J. Byrka, S. Guillemot and J. Jansson. 2008. New Results on Optimizing Rooted Triplets Consistency. In *ISAAC 2008*, *LNCS* 5369, pp. 484-495. Springer.
7. C. Chauve and N. El-Mabrouk. 2009. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. In *RECOMB 2009*, *LNCS* 5541, pp. 46–58. Springer.
8. C. Chauve and A. Ouangraoua. 2009. A 3-approximation algorithm for computing a parsimonious first speciation in the gene duplication model. arXiv:0904.1645v2
9. T. De Bie *et al.*. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. 22, pp. 1269–1271.
10. S. Fujishige. 2005. Submodular Functions and Optimization. Second edition. *Annals of Discrete Math.* 58. Elsevier.
11. P. Górecki and J. Tiuryn. 2006. DLS-trees: a model of evolutionary scenarios. *Theoret. Comput. Sci.*, 359:378–399.
12. S. Guillemot. 2008. Approches combinatoires pour le consensus d'arbres et de séquences. Ph.D. thesis, Univ. Montpellier II, France.
13. M.T. Hallett and J. Lagergren. 2000. New algorithms for the duplication-loss model. In *RECOMB 2000*, pp. 138–146. ACM Press.
14. S. Iwata and J.B. Orlin. 2009. simple combinatorial algorithm for submodular function minimization. In *SODA 2009*, pp. 1230–1237. SIAM.
15. M.W. Hahn, M.V. Han and S.-G. Han. 2007. Gene family evolution across 12 *Drosophilia* genomes. *PLoS Genet.*, 3:e197.
16. B. Ma, M. Li and L. Zhang. 2000. From gene trees to species trees. *SIAM J. Comput.*, 30:729–752.
17. W.-K. Mak. 2005. Faster Min-Cut Computation in Unweighted Hypergraphs/Circuit Netlists. In *VLSI Design, Automation and Test, 2005. (VLSI-TSA-DAT)*, pp.67–70. IEEE.
18. R.D.M. Page. 2002. Modified mincut supertrees. In *WABI 2002*, *LNCS* 2452, pp.537–551. Springer.
19. M.J. Sanderson and M.M. McMahon. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol. Biol.*, 7:S3.
20. C. Scornavacca, V. Berry and V. Ranwez 2009. From gene trees to species trees through a supertree approach. In *LATA 2009*, *LNCS* 5457, pp. 702–714. Springer.
21. C. Semple and M. Steel. 2000. A supertree method for rooted trees. *Discrete Appl. Math.*, 105:147–158.
22. U. Stege. 1999. Gene Trees and Species Trees: The Gene-Duplication Problem in Fixed-Parameter Tractable In *WADS 1999*, *LNCS* 1663, pp. 288–293. Springer.
23. I. Wapinski *et al.*. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61.
24. H. Li *et al.*. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, 34:572–580. 2006.
25. A. Wehe *et al.*. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24:1540–1541.
26. X. Zhou, Z. Lin and H. Ma. 2010. Phylogenetic detection of numerous gene duplications shared by animals, fungi and plants *Genome Biol.*, 11:R38.