JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 9, 2011

Pp. 1041–1053
DOI: 10.1089/cmb.2011.0108

A 2-Approximation for the Minimum Duplication Speciation Problem

AÏDA OUANGRAOUA,[1] KRISTER M. SWENSON,[2,3] and CEDRIC CHAUVE[3,4]

ABSTRACT

We consider the following problem: given a set of gene family trees, spanning a given set of species, find a first speciation which splits these species 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, we propose a polynomial time 2-approximation algorithm for the Minimum Duplication Bipartition Problem. We apply this algorithm to the inference of species trees on synthetic datasets and on two datasets of eukaryotic species.

Key words: computational molecular biology, dynamic programming, genomics rearrangements.

1. INTRODUCTION

Gene duplication is a fundamental evolutionary mechanism for important groups of eukaryotes such as vertebrates (Blomme et al., 2006), insects (Hahn et al., 2007), plants (Sanderson and McMahon, 2007) or fungi (Wapinski et al., 2007). Gene duplications, together with gene losses, result in *gene families*, which can contain several copies of a certain 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 (Li et al., 2006). Phylogenomics aims at reconstructing the evolution of *species (genomes)* by inferring a species tree from a set of gene trees. The *Minimum Duplication Problem* (MD Problem), also known as the *Gene Tree Parsimony Problem* (GTP 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 (Ma et al., 2000), and is neither fixed-parameter tractable (FPT) using the number of gene duplications as parameter, nor approximable with a constant ratio (Bansal and Shamir, 2011; Guillemot, 2008). Recent advances in local search heuristics proved to be useful (Bansal et al., 2007) and have been applied on several eukaryotic datasets with interesting results (Sanderson and McMahon, 2007; Wehe et al., 2008), but with no optimality guarantee.

Recently, Chauve and El-Mabrouk (2009) and Scornavacca et al. (2011) described a formal link between the Minimum Duplication Problem and a problem of supertrees (Bininda-Edmonds, 2004), 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

[1]INRIA Lille–Nord Europe, LIFL, Université Lille 1, Villeneuve d' Ascq, France.
[2]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada.
[3]Lacim, Université du Québec à Montréal, Montréal, Quebec, Canada.
[4]Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, Canada.

to reconstruct a species tree that disagrees with the minimum number of gene trees (Byrka et al., 2010). 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 (Bryant, 1997). 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 (Page, 2002; Semple and Steel, 2000). In such heuristics, each Minimum Edge-Cut splits the set of considered species into two subsets, corresponding to a speciation which 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 natural greedy heuristics to reconstruct a species tree from a set of gene trees. Our work is also motivated by understanding the combinatorial properties of speciations 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. Although our focus here is primarily theoretical, we also provide experimental results, on small simulated datasets, and two real eukaryotic large-scale datasets.

We first define, in Section 2, gene trees, species trees, duplications, and the optimization problems considered in this article. In Section 3, we show how the Minimum Duplication Bipartition Problem can be described as a variant of the classical Minimum Edge-Cut Problem. Then we use this description to reduce the Minimum Duplication Bipartition Problem to a Set Function Minimization Problem. We finally describe a 2-approximation algorithm through a Minimum Hypergraph Cut Problem. In Section 4, we describe results of experiments on simulated and real eukaryotic data, illustrating our algorithm in a realistic phylogenomics context.

2. PRELIMINARIES: OBJECTS, PROBLEMS, BACKGROUND

Consider, $\mathcal{G} = \{1, 2, \ldots, k\}$ as a set of integers representing k different species (genomes).

Gene and species trees, bipartitions. A *species tree* on $\mathcal{G}$ is a rooted tree with exactly k leaves, where each $i \in \mathcal{G}$ is the label of a single leaf. A tree is a complete binary tree if every internal vertex has exactly two children. A *gene tree* on $\mathcal{G}$ is a complete 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 $\mathcal{F}$, that is a set of genes that all originate from a single ancestral gene; a leaf labeled i in a gene tree represents a gene belonging to $\mathcal{F}$ and 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 (and then two internal vertices).

Given a tree T whose leaves are labeled by integers from $\mathcal{G}$ and a vertex x of T, 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 an internal vertex, we denote by x_ℓ and x_r the two children of x.

A *bipartition* B on a set S is a partition of S into two subsets S_1 and S_2. 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 $L(v_\ell) \cap L(v_r) = \emptyset$, $L(v_\ell) \cup L(v_r) = S$, $L(v_\ell) = S_1$ and $L(v_r) = S_2$.

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 v of S such that $L(x) \subseteq L(v)$ and, either v is 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* (Fig. 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.

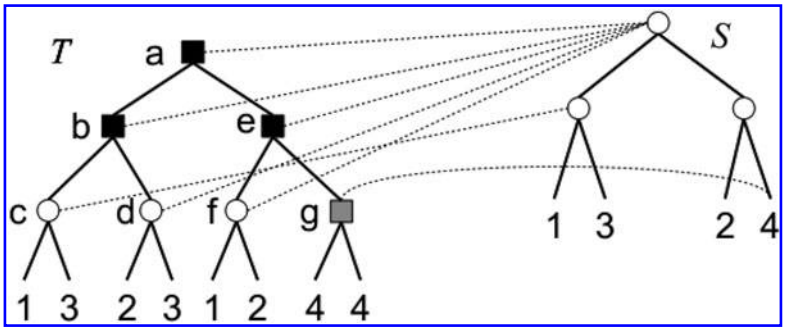

FIG. 1. A gene tree T and a species tree S on a set of genome $\mathcal{G} = \{1, 2, 3, 4\}$. The LCA mapping from vertices of T to vertices of S is indicated by dashed lines linking vertices. The vertices of T that are duplications with respect to S are represented by square vertices; the black colored 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.

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 duplications are called *non-apparent duplications*. Given a bipartition B on $\mathcal{G}$ with root v, a duplication vertex 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*. We denote by $d_1(F, B)$, the number of pre-duplications of F with respect to B.

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 (Chauve and El-Mabrouk, 2009; Górecki and Tiuryn, 2006), 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 (Ma et al., 2000), even in the case where every gene tree is a uniquely leaf-labeled and rooted triplet (Bryant, 1997), 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 (Bansal and Shamir, 2011; Guillemot, 2008) (contrary to what was claimed in Stege [1999]),
- second, it cannot be approximated within a constant ratio unless P = NP (Byrka et al., 2010).

Hence, for solving the MD Problem, one has to rely on exponential time algorithms (Hallett and Lagergren, 2000) or local-search methods with no optimality guarantee (Bansal et al., 2007; Wehe et al., 2008).

In Chauve and El-Mabrouk (2009), it was shown that the MD Problem is in fact a variant of a supertree problem (see also Scornavacca et al. [2011], which 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 (Semple and Steel, 2000), and in Chauve and El-Mabrouk (2009), an application of such a heuristic showed promising results on synthetic data. This motivates the introduction of the main problem we study in this article:

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 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 $\{s,t\} \subseteq L(x_r)$.

Previous work. As far as we know, the MDB problem was introduced in Stege (1999), where an exponential time algorithm was proposed. Unlike the MD Problem, it is obviously FPT, as there are $2^{|\mathcal{G}|}$ possible bipartitions and computing the number of pre-duplications associated to a bipartition can be done

in polynomial time. It was also shown in Chauve and El-Mabrouk (2009), 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. However, the hardness of the MDB Problem is still an open problem. In Chauve and Ouangraoua (2009), it was shown that if F contains a single gene tree, the MDB Problem is 3-approximable; an alternative combinatorial model, based on prefix of gene trees was also introduced. 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 Chauve and Ouangraoua (2009) computes a first speciation that can imply up to $2d + t$ duplications. In the present work, we improve on this result and we show that the MDB Problem can be approximated with a constant ratio of 2 in polynomial time.

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, $E_G(B) = \{(s, t) \in E \parallel s \in L(v_\ell), t \in L(v_r)\}$. Hence, a subset X of V induces a bipartition on V: if we denote its root by v, then $L(v_\ell) = X$ 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 (Fujishige, 2005), in particular the MEC Problem. Given a submodular function f, the following optimization problem, which is tractable (Iwata and Orlin, 2009), 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.

Remark 1. The Minimum Edge-Cut problem is a special case of the SFM Problem: given a graph $H = (V, E)$, if we denote by g the cut-set function defined from 2^V to $\mathbb{R}$ that associates to a subset X of V the number of edges in the edge-cut $E_H(X)$, then g is a submodular function. The problem of finding a subset of vertices $X \subset V$ that minimizes $g(X)$ is then equivalent to the Minimum Edge-Cut problem on H. However, the MLEC Problem is not an SFM Problem, and, due to its relationship with the Minimum Label-Cut problem (Chauve and Ouangraoua, 2009), is in fact NP-complete, even when each edge label appears at most two times in the graph (Zhang et al., 2011; Fellows, 2010).

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

We first describe how the MDB problem can naturally be translated into an edge-cut problem; however the tractability of this edge-cut problem is unknown, as it cannot be reduced to an SFM problem. Next, we describe our main result, that is an efficient 2-approximation algorithm for the MDB problem.

From now on, 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.

3.1. A set function minimization problem

Given a gene tree forest F, we define the *edge-labeled graph* $H(F) = (V, E)$ associated to F as follows (Fig. 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 shows how the MDB problem can be described as a Minimum Labeled-Edge-Cut (MLEC) Problem.

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$(E_{H(F)}(B))$.*

Proof. If a is the label of a pre-duplication x in F with respect to the bipartition B with root v, then, from Property 1, there is a pair $(s, t) \in L(v_\ell) \times L(V_r)$ such that $\{s, t\} \subseteq L(x_\ell)$ or $\{s, t\} \subseteq L(x_r)$, and thus (s, t) is an edge of $H(F)$ labeled with a; this edge is then contained in the edge-cut $E_{H(F)}(B)$. Conversely, if $E_H(B)$ contains an edge (s, t) labeled with $a \in \Sigma$, then, also from Property 1, there is a vertex of F labeled with a that is a pre-duplication with respect to B. ■

Remark 2. If a forest F of gene trees contains uniquely leaf-labeled triplets, then each label a appears on a single edge, and the MDB problem is equivalent to the Minimum Edge-Cut problem on $H(F)$, and then tractable.

The MLEC Problem (and then the MDB 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 $label(E_G(X))$. Solving the MLEC Problem on G can then 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.

Modeling the MDB problem as an edge-cut problem naturally leads to ask if this edge-cut problem is an instance of the SFM problem, in which case it would be tractable. As noted in Remark 1, the MLEC Problem is NP-hard. However, it is not difficult to see that the graphs used in Jha et al. (2002) to prove the hardness of the Minimum label Cut Problem can not be obtained from gene tree forests. Unfortunately, the cut-set function f_G associated to an edge-labeled graph G is not always a submodular function, as shown below.

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

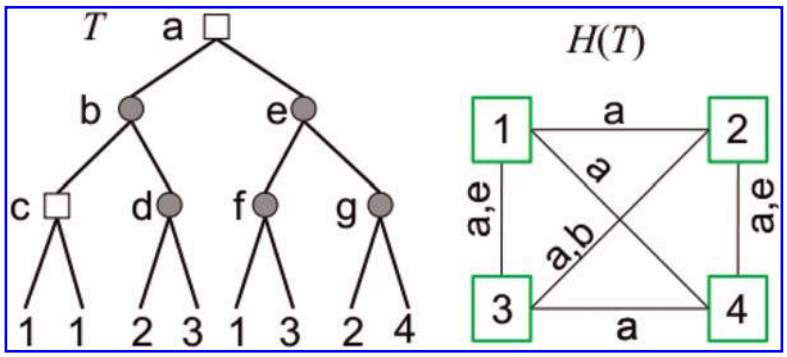

FIG. 2. A gene tree T on the set of genomes $\mathcal{G} = \{1, 2, 3, 4\}$ and the corresponding edge-labeled graph $H(T)$. Apparent duplication vertices of T appear as square vertices.

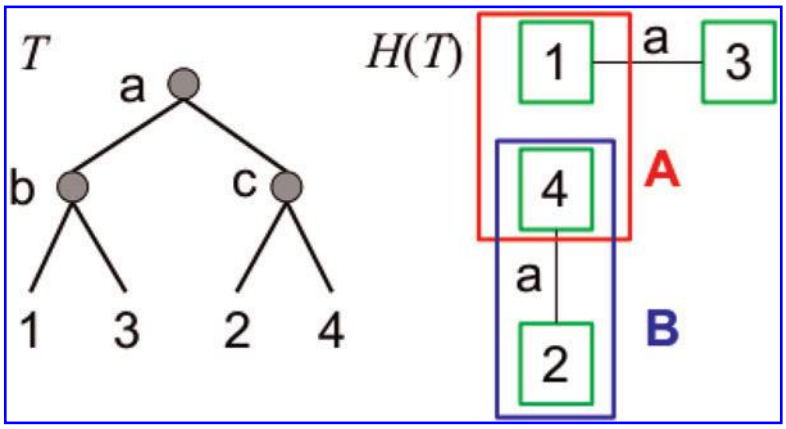

FIG. 3. Illustration of Property 2: a gene tree T (left) and the corresponding graph $H(T)$ (right), and two vertex sets A and B that contradict the submodularity of the cut-set function f_H.

For example (Fig. 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.

We described in Chauve and Ouangraoua (2009) and Ouangraoua et al. (2010) the necessary conditions for the cut-set function f_H to be non-submodular. Based on these conditions, we were able to show that the graph $H(F)$ can be augmented into a graph $I(F)$ in such a way that the (labeled) cut-set function f_I is submodular. This result, aside from its theoretical interest, lead to a 2-approximation algorithm for the MDB Problem, based on solving a SFM Problem on the graph $I(F)$. In Section 3.2 below, we describe another 2-approximation through a direct link with the Minimum Hypergraph Cut problem, whose exposition is simpler.

3.2. An approximation through the Minimum Hypergraph Cut problem

Given a gene tree forest F, we define the *hypergraph* $J(F) = (V, E)$ associated to F as follows (Fig. 4): $V = L(F)$ and, for each apparent duplication x in F, $L(x)$ is a hyperedge of $J(F)$ and, for each non-apparent duplication y in F, $L(y_\ell)$ and $L(y_r)$ are two hyperedges of $J(F)$.

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. $d(F, B) \leq 2d$ by solving the MHC Problem on the graph $J(F)$.*

Proof. For any bipartition B on $L(F)$, the cardinality of $label(E_H(B))$ (resp. $E_J(B)$) 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. A label $a \in \text{label}(E_H(B))$ corresponds to at least one but at most two hyperedges in $E_J(B)$: if a is the label of an apparent duplication x, then we have $a \in \text{label}(E_H(B)) \Leftrightarrow L(x) \in E_J(B)$; if a is the

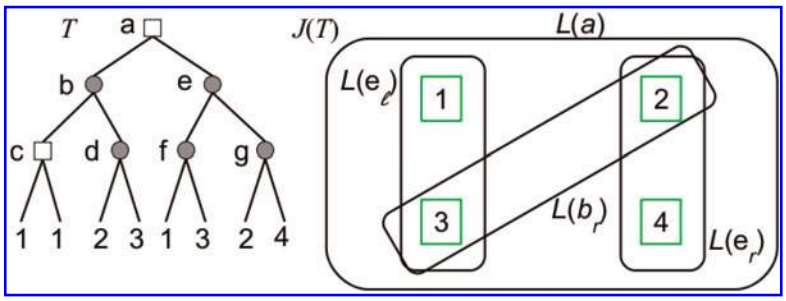

FIG. 4. A gene tree T on the set of genomes $\mathcal{G} = \{1, 2, 3, 4\}$ and the hypergraph $J(F)$, where hyperedges containing only one vertex are not displayed. Apparent duplication vertices of T appear as square vertices.

label of a non-apparent duplication y, then we have $a \in \text{label}(E_H(B)) \Leftrightarrow L(y_\ell) \in E_J(B)$ or $L(y_r) \in E_J(B)$. 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 hyperedge-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 $d_J(B) \geq d_H(B) > 2 * d_H(B') \geq d_J(B')$. Hence, for any bipartition B on $L(F)$ that induces an optimal hyperedge-cut 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 time complexity stated in Theorem 1 follows from the algorithm described in Mak, (2005) solving the MHC Problem on a hypergraph G in time $O(kn)$ where k (resp. n) is the number of vertices (resp. hyperedges) of G. ■

We showed in Ouangraoua et al. (2010) that the approximation ratio of 2 is tight. The approximation algorithm also has the following straightforward algorithmic property.

Property 3. If there exists at least one parsimonious first speciation B' such that all the corresponding pre-duplications are apparent duplications in F, then solving the MHC Problem on $J(F)$ gives a parsimonious bipartition.

To prove Property 3, we only need to note that if a first speciation B' induces only pre-duplications which are apparent, then B' defines 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.

4. EXPERIMENTAL RESULTS

We performed four different experiments.[1] 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 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). Second, 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. Third, we applied the approximation algorithm on a real dataset of 6808 gene families from 23 fungal genomes. Last, we performed the same experiment on a real dataset of 18584 gene families from 50 eukaryotic genomes.

4.1. Experiment on simulated datasets

Datasets. We analyzed the four synthetic datasets that were studied in Chauve and El-Mabrouk (2009); 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 (Bie et al., 2006) from real Drosophilia gene families (Hahn et al., 2007). The duplication/loss rates range from 0.02 event/million years to 0.2 event/million years, and the number of duplication ranges from roughly 1000 (with rate 0.02) to roughly 6000 (see Table 1 in Chauve and El-Mabrouk [2009]).

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 leads to the exact species tree (Fig. 5), 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

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

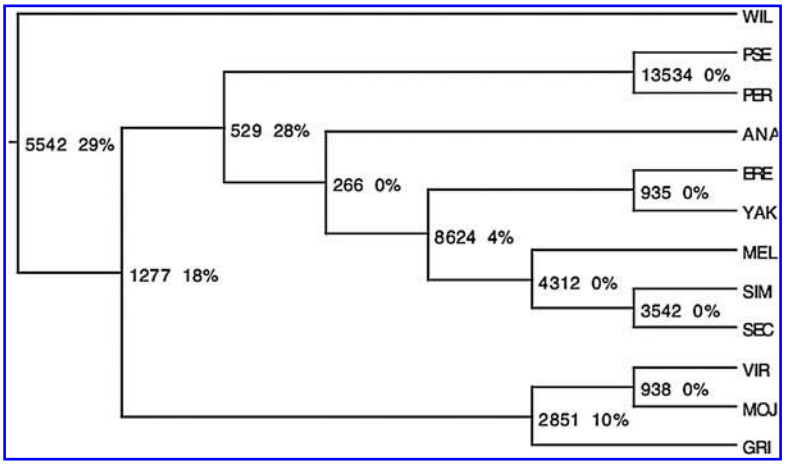

FIG. 5. The species tree computed for the fourteen synthetic datasets on 12 species. Each speciation (vertex) of the tree is labeled with the total number of duplications preceding the speciation in all datasets, and the proportion of non-apparent duplications.

context, especially as the heuristic described in Chauve and El-Mabrouk (2009) 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 (Fig. 5).

4.2. Experiment on a fungal dataset

Datasets. We uploaded from the Fungal Orthogroup Repository[2] the 6808 fungal gene trees containing genes belonging to at least three different species, among 23 ascomycete fungal species.

Results. 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[3] (Fig. 6).

We also noticed that a significant number of speciations corresponded to apparent duplications only, while some of them are associated with a large number of duplications. The apparent duplications correspond to the speciation vertices labeled with 0% non-apparent pre-duplications in Figure 6. Such speciations are then parsimonious (see the discussion at the end of the previous section). Moreover, aside from one branch (leading to the group containing Kwal, Agos, Klac, Sklu) of the species tree, where 103 pre-duplications out of 322 (31% of the pre-duplications) are non-apparent, all other branches are associated with very few non-apparent pre-duplications, which suggests that 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.

4.3. Experiment on a large-scale eukaryotic dataset

Datasets. We uploaded from the release 58 of the Ensembl database the 18584 eukaryotic gene trees (Vilella et al., 2009) containing genes belonging to 50 eukaryotic species, and obtained from CDS back-translated protein-based multiple alignments (Vilella et al., 2009). The reference phylogeny of the 50 species is represented in Figure 7.

Results. We applied the same method as for the previous dataset, and obtained a species tree (Fig. 8) that is close to the one widely accepted in fungi/metazoa phylogenomics (Fig. 7). Among the 38 reference clades (i.e. set of species descending from the same internal vertex present in the reference tree), the inferred species tree recovers 35 clades, and does not recover only 3 clades inside the eutherian group (placental mammals): Xenarthra, Afrotheria, and Insectivora. However, the relative position of the species

[2]Version 1.1, `www.broadinstitute.org/regev/orthogroups/.`
[3]The same tree was obtained from `www.broadinstitute.org/regev/orthogroups/.`

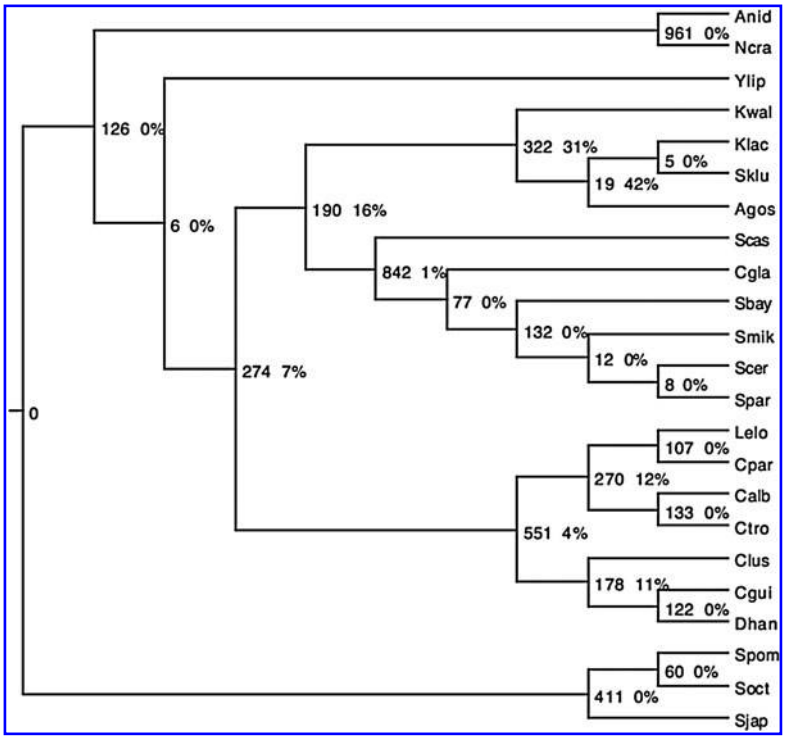

FIG. 6. The species tree computed for the fungal datasets on 23 species. Each speciation of the tree is labeled with the number of duplications preceding the speciation (pre-duplications), and the proportion of non-apparent duplications. Species: *A. nidulans* (Anid), *N. crassa* (Ncra), *Y. lipolytica* (Ylip), *K. waltii* (Kwalt), *K. lactis* (Klac), *S. kluyveri* (Sklu), *A. gossypii* (Agos), *S. castellii* (Scas), *C. glabrata* (Cgla), *S. bayanus* (Sbay), *S. mikatae* (Smik), *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *L. elongosporus* (Lelo), *C. parapsilosis* (Cpar), *C. albicans* (Calb), *C. tropicalis* (Ctro), *C. lusitaniae* (Clus), *C. guilliermondii* (Cgui), *D. hansenii* (Dhan), *S. pombe* (Spom), *S. octosporus* (Soct), *S. japonicus* (Sjap).

of these 3 clades compared to the others species remains congruent with recent studies on the phylogeny of placental mammals (Kjer and Honeycutt, 2007; Murphy et al., 2007; Nikolaev et al., 2008; Prasad et al., 2008).

First, we noticed that the two whole-genome duplications that occurred on the branch leading to the fish (*Danio rerio*, … , *Gasterosteus aculeatus*), and on the branch leading to vertebrates (*Danio rerio*, … , *Tursiops truncatus*), both left traces that are still visible today: a high number of pre-duplications (3846, 9375) with a low proportion of non-apparent duplications (2%, 11%).

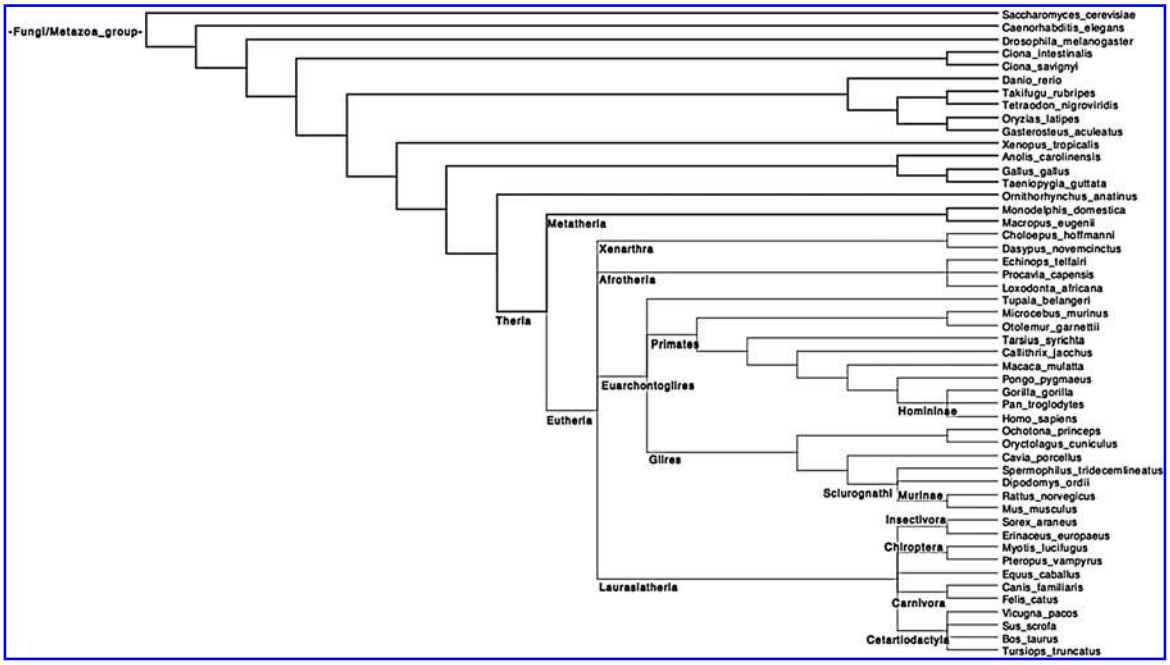

FIG. 7. The 50 considered eukaryotic species. Tree obtained from the NCBI taxonomy at www.ncbi.nlm.nih.gov.

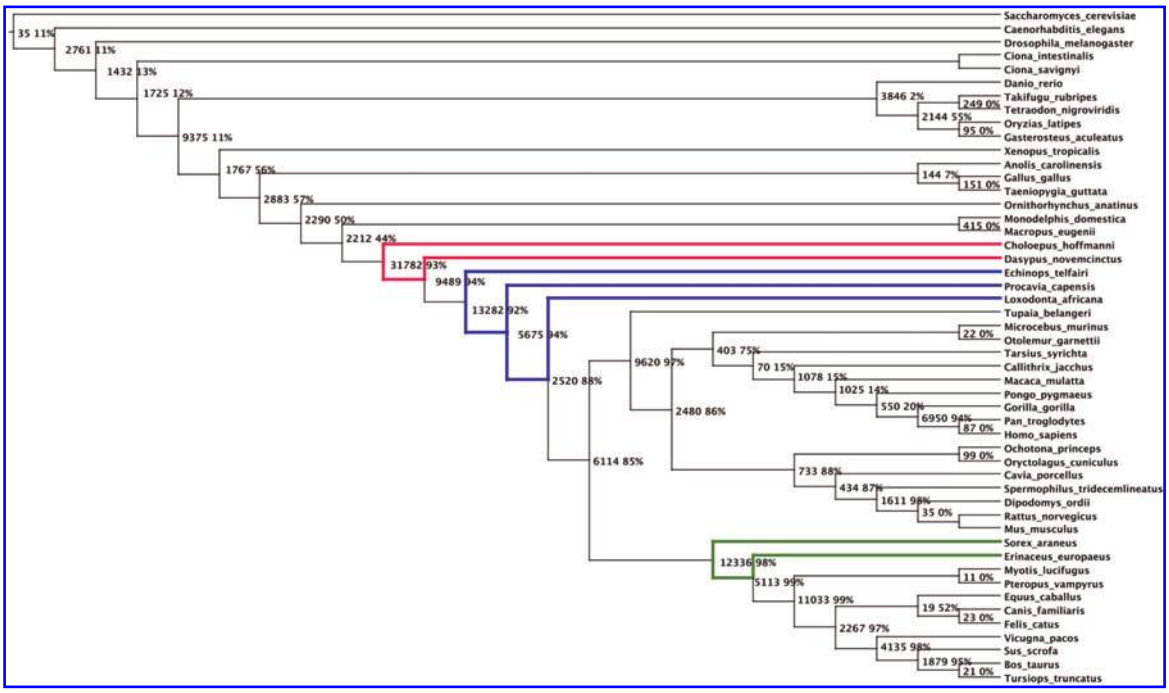

FIG. 8. The species tree computed for the eukaryotic datasets on 50 species. Each speciation of the tree is labeled with the number of duplications preceding the speciation, and the ratio of non-apparent duplications. The clades that are missing regarding the NCBI reference tree are represented in red (Xenarthra), blue (Afrotheria), and green (Insectivora) colors.

From a structural point of view, we notice that the 3 speciations which are responsible for missing the monophyletic clades Xenarthra, Afrotheria, and Insectivora are characterized by very high numbers of pre-duplications (31782, 13282, 12336), coupled with high proportions of non-apparent duplications (93%, 92%, 98%). Moreover, all the speciations obtained for the unresolved nodes (i.e. nodes having more than two descending branches) of the reference tree (Eutheria, Euarchontoglires, Homininae, Laurasiatheria, and Cetartiodactyla) alse have high numbers of pre-duplications (more than 5000), coupled again with high proportions of non-apparent duplications (more than 90%).

In Table 1, we summarize the numbers of duplications associated with all the possible binary configurations for the unresolved nodes of the reference tree. We can observe that our algorithm misses the most parsimonious configuration for the roots of 4 clades: Eutheria, Afrotheria, Euarchontoglires, and Laurasiatheria. This is caused by the choice of a first speciation that has the lowest local pre-duplication cost, but leads to a non globally optimal configuration. These observations clearly suggests that considering a parsimony criterion in terms of gene duplications might not always be helpful to resolve difficult speciations. One explanation is that issues in reconstructing accurate and consistent phylogenetic trees might also impact the gene duplications signal, that is detected by reconciling gene trees and species trees. In these cases, the consideration of parsimonious k-furcations (where a speciation is the particular case where $k = 2$) instead of parsimonious speciations might help resolving the difficult phylogenies.

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 has been proposed in relation to 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. This algorithm is also a natural generalization of the classical

TABLE 1. NUMBERS OF DUPLICATIONS ASSOCIATED TO THE CONCURRENT BINARY CONFIGURATIONS FOR THE UNRESOLVED NODES OF THE REFERENCE TREE

	n_1	n_2	n_3	n
Eutheria				
((Xenarthra,Afrotheria),Boreoeutheria)	46791	42		46833
(Afrotheria,(Xenarthra,Boreoeutheria))	42418	5724		48142
(Xenarthra,(Afrotheria,Boreoeutheria))	34351	14115		48466
Afrotheria				
(Echinops,(Procavia,Loxondonta))	1329	74		1403
(Loxondonta,(Procavia,Echinops))	6315	23		6338
(Procavia,(Echinops,Loxondonta))	6774	22		6796
Euarchontoglires				
(Primates,(Tupaia,Glires))	10418	423		10841
(Glires,(Tupaia,Primates))	11372	237		11609
(Tupaia,(Primates,Glires))	9620	2480		12100
Hominimae**				
(Gorilla,(Pan,Homo))	6950	87		7037
(Pan,(Gorilla,Homo))	12594	127		12721
(Homo,(Pan,Gorilla))	13070	40		13110
Sciurognathi**				
(Spermophilus,(Dipodomys,Murinae))	1611	35		1646
(Dipodomys,(Spermophilus,Murinae))	4493	15		4508
(Murinae,(Dipodomys,Spermophilus))	4510	4		4514
Laurasiatheria				
(Insectivora,(Chiroptera,Ferungulata))	13436	11033		24469
(Chiroptera,(Insectivora,Ferungulata))	23563	2806		26369
(Ferungulata,(Chiroptera,Insectivora))	26450	98		26548
Ferungulata**				
(Cetartiodactyla,(Equus,Carnivora))	2267	19		2286
(Carnivora,(Equus,Cetartiodactyla))	2303	61		2364
(Equus,(Cetartiodactyla,Carnivora))	2380	87		2467
Cetartiodactyla**				
(Vicugna,(Sus,(Bos,Tursiops)))	4135	1879	21	6035
((Vicugna,Sus),(Bos,Tursiops))	7981	11	21	8013
(Sus,(Vicugna,(Bos,Tursiops)))	7425	659	21	8105
(Vicugna,(Tursiops,(Bos,Sus))	4135	5199	65	9399
(Vicugna,(Bos,(Sus,Tursiops))	4135	5442	5	9582
(Sus,(Tursiops,(Bos,Vicugna)))	7425	2379	4	9808
(Sus,(Bos,(Vicugna,Tursiops)))	7425	2396	4	9825
(Tursiops,(Vicugna,(Bos,Sus)))	13182	446	65	13693
(Tursiops,(Bos,(Sus,Vicugna)))	13182	586	11	13779
(Tursiops,(Sus,(Bos,Vicugna)))	13182	614	4	13800
((Vicugna,Tursiops),(Bos,Sus))	13739	4	65	13808
(Bos,(Vicugna,(Sus,Tursiops)))	13431	452	5	13888
((Vicugna,Bos),(Sus,Tursiops))	13978	4	5	13987
(Bos,(Sus,(Vicugna,Tursiops)))	13431	555	4	13990
(Bos,(Tursiops,(Vicugna,Sus)))	13431	553	11	13995

n_i, number of duplications preceding the i^{th} speciation; n, total number of pre-duplications associated to each binary configuration. The cases where our algorithm recovers the most parsimonious configurations (i.e. inducing a minimum global number of duplications) are marked with**.

minimum edge-cut algorithm that is used in supertree consistency problems, which is highlighted by its link with the Submodular Function Minimization Problem.

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 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.

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. However, we can notice that, on a large-scale eukaryotic dataset, some speciations, that were already considered as difficult to resolve by traditional phylogenetic methods, are associated with an unrealisticly large number of duplications and, more importantly, a large number of non-apparent pre-duplications. We investigated alternative hypothesis for such speciations, that resulted in comparable results. This suggests a possible way to characterize such difficult speciations, that are not supported by a strong signal in gene trees.

ACKNOWLEDGMENTS

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

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bansal, M., Burleigh, J., Eulenstein, O., et al. 2007. Heuristics for the gene-duplication problem: a $\theta(n)$ speed-up for the local search, 238–252. *In* Speed, T.P., Huang, H., eds. *Research in Computational Molecular Biology*. Springer, New York.

Bansal, M., and Shamir, R. 2011. A note on the fixed parameter tractability of the gene-duplication problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 848–850.

Bie, T.D., Cristianini, N., Demuth, J., et al. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271.

Bininda-Edmonds, O., ed. 2004. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer, New York.

Blomme, T., Vandepoele, K., Bodt, S.D., et al. 2006. The gain and loss of genes during 600 millions years of vertebrate evolution. *Genome Biol.* 7, R43.

Bryant, D. 1997. Hunting for trees, building trees and comparing trees: theory and methods in phylogenetic analysis [Ph.D. dissertation]. Department of Mathematics, University of Canterbury, New Zealand.

Byrka, J., Guillemot, S., and Jansson, J. 2010. New results on optimizing rooted triplets consistency. *Discrete Appl. Math.* 158, 1136–1147.

Chauve, C., and El-Mabrouk, N. 2009. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees, 46–58. *In* Batzoglou, S., ed. *Research in Computational Molecular Biology*. Springer, New York.

Chauve, C., and Ouangraoua, A. 2009. A 3-approximation algorithm for computing a parsimonious first speciation in the gene duplication model. Available: arxiv.org/abs/0904.1645v2.

Fujishige, S. 2005. *Submodular Functions and Optimization. Annals of Discrete Mathematics. Volume 58*. Elsevier, New York.

Górecki, P., and Tiuryn, J. 2006. Dls-trees: a model of evolutionary scenarios. *Theoret. Comput. Sci.* 359, 378–399.

Guillemot, S. 2008. *Approches combinatoires pour le consensus d'arbres et de séquences* [Ph.D. dissertation]. Université Montpellier II, Montpellier, France.

Hahn, M., Han, M., and Han., S.-G. 2007. Gene family evolution across 12 *drosophilia* genomes. *PLoS Genet.* 3, e197.

Hallett, M., and Lagergren, J. 2000. New algorithms for the duplication-loss model, 138–146. *In* Shamir, R., Miyano, S., Istrail, S., et al., eds. *Research in Computational Molecular Biology*. ACM, New York.

Iwata, S., and Orlin, J. 2009. A simple combinatorial algorithm for submodular function minimization, 1230–1237. *In* Mathieu, C., ed. *ACM-SIAM Symposium on Discrete Algorithms*. SIAM, New York.

Jha, S., Sheyner, O., and Wing, J. 2002. Two formal analys s of attack graphs. *Proc. IEEE Comput. Security Found. Workshop* 49–63.

Kjer, K., and Honeycutt, R. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol. Biol.* 7, 8.

Li, H., Coghlan, A., Ruan, J., et al. 2006. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 34, 572–580.

Ma, B., Li, M., and Zhang, L. 2000. From gene trees to species trees. *SIAM J. Comput.* 30, 729–752.

Mak, W.-K. 2005. Faster min-cut computation in unweighted hypergraphs/circuit netlists. *Proc. Int. Symp. VLSI Design Automation Test* 67–70.

Fellows, M.R., Guo, J., and Kanj, I. 2010. The parameterized complexity of some minimum label problems. *J. Comput. Syst. Sci.* 76, 727–740.

Murphy, W., Pringle, T., Crider, T., et al. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* 17, 413–421.

Nikolaev, S., Montoya-Burgos, J., Margulies, E., et al. 2008. Early history of mammals is elucidated with the encode multiple species sequencing data. *PLoS Genet.* 3, e2.

Ouangraoua, A., Swenson, K.M., and Chauve, C. 2010. An approximation algorithm for computing a parsimonious first speciation in the gene duplication model, 290–301. *In* Tannier, E., ed. *Comparative Genomics*. Springer, New York.

Page, R. 2002. Modified mincut supertrees, 537–552. *In* Guigó, R., Gusfield, D., eds. *Algorithms in Bioinformatics*. Springer, New York.

Prasad, A., Allard, M., Program, N.C.S., et al. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol. Biol. Evol.* 25, 1795–1808.

Sanderson, M., and McMahon, M. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol. Biol.* 7.

Scornavacca, C., Berry, V., and Ranwez, V. 2011. Building species trees from larger parts of phylogenomic databases. *Inform. Comput.* 209, 590–605.

Semple, C., and Steel, M. 2000. A supertree method for rooted trees. *Discrete Appl. Math.* 105, 147–158.

Stege, U. 1999. Gene trees and species trees: the gene-duplication problem in fixed-parameter tractable, 288–293. *In* Dehne, F., Gupta, A., Sack, J.-R., et al., eds. *Algorithms and Data Structures*. Springer, New York.

Vilella, A., Severin, J., Ureta-Vidal, A., et al. 2009. Ensembl compara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335.

Wapinski, I., Pfeffer, A., Friedman, N., et al. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449, 54–61.

Wehe, A., Bansal, M., Burleigh, J., et al. 2008. Duptree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24, 1540–1541.

Zhang, P., Cai, J.-Y., Tang, L.-Q., et al. 2011. Approximation and hardness results for label cut and related problems. *J. Comb. Optim.* 21, 192–208.

Address correspondence to:
Dr. Aïda Ouangraoua
INRIA Lille–Nord Euorpe, LIFL
Université Lille 1
Villeneuve d'Ascq–Lille, 59650 France

E-mail: Aida.Ouangraoua@inria.fr