Theoretical Computer Science 530 (2014) 66–79

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Complexity insights of the Minimum Duplication problem

Guillaume Blin [a], Paola Bonizzoni [b], Riccardo Dondi [c], Romeo Rizzi [d], Florian Sikora [e,*]

[a] Université Paris-Est, LIGM, UMR 8049, France
[b] DISCo, Universitá degli Studi di Milano–Bicocca, Milano, Italy
[c] Dipartimento di Scienze Umane e Sociali, Universitá degli Studi di Bergamo, Bergamo, Italy
[d] Department of Computer Science, University of Verona, Verona, Italy
[e] PSL, Université Paris–Dauphine, LAMSADE, UMR 7243, Paris, France

ARTICLE INFO

Article history:
Received 11 March 2013
Received in revised form 29 January 2014
Accepted 18 February 2014
Communicated by M. Crochemore

Keywords:
Minimum Duplication problem
Comparative genomics
Computational complexity
APX-hardness
Randomized algorithm

ABSTRACT

The Minimum Duplication problem is a well-known problem in phylogenetics and comparative genomics. Given a set of gene trees, the Minimum Duplication problem asks for a species tree that induces the minimum number of gene duplications in the input gene trees. Recently, a variant of the Minimum Duplication problem, called Minimum Duplication Bipartite, has been introduced, where the goal is to find all *pre-duplications*, that is duplications that in the evolution precede the first speciation with respect to a species tree. In this paper, we investigate the complexity of both Minimum Duplication and Minimum Duplication Bipartite. First of all, we prove that the Minimum Duplication problem is APX-hard, even when the input consists of five uniquely leaf-labeled gene trees (improving upon known results on the complexity of the problem). Then, we show that the Minimum Duplication Bipartite problem can be solved efficiently with a randomized algorithm when the input gene trees have bounded depth. An extended abstract of this paper appeared in SOFSEM 2012 [1].

1. Introduction

The evolutionary history of the genomes of eukaryotes is the result of a series of evolutionary events, called *speciations*, that produce new species starting from a common ancestor. This evolutionary history has been deeply studied in computational biology, and it is usually represented using a phylogenetic tree called *species tree* [2]. A *species tree* is a rooted binary tree whose leaves are uniquely labeled by a set Λ representing the extant species, where the common ancestor of the contemporary species is associated with the root of the tree. The internal nodes represent hypothetical ancestral species (and the associated speciations). Speciations are not the only events that influence the evolution. Indeed, there are other events, such as gene duplications, gene losses and lateral gene transfers that, although not leading to new species, are fundamental in the evolution. In this paper we focus on gene duplications which are known to be essential for the evolution of many eukaryotes groups, such as vertebrates, insects and plants [3]. A gene duplication can be described as the genomic event that causes a gene inside a genome to be copied, resulting in two copies of the same gene that can evolve independently. Genes of extant species are called *homologous* if they evolved from a common ancestor through speciations and duplications events [4]. The evolution of homologous genes, with regards to the extant species, is usually represented

* Corresponding author.
E-mail addresses: gblin@univ-mlv.fr (G. Blin), bonizzoni@disco.unimib.it (P. Bonizzoni), riccardo.dondi@unibg.it (R. Dondi), Romeo.Rizzi@univr.it (R. Rizzi), florian.sikora@dauphine.fr (F. Sikora).

http://dx.doi.org/10.1016/j.tcs.2014.02.025
0304-3975/

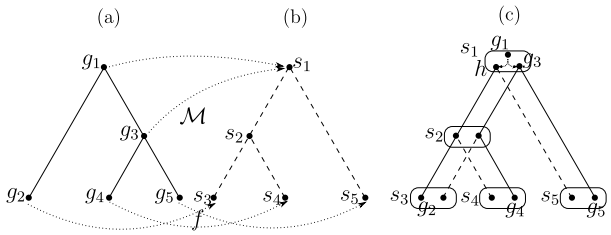

Fig. 1. (a) A gene tree T. (b) A species tree S where $\mathcal{M}$ is the `lca` mapping from T to S; each gene in $\{g_2, g_4, g_5\}$ is mapped by function f in the species that gene belongs to. Nodes of S are labeled with s's. (c) A reconciled tree for T and S based on the *a priori* duplication of gene g_1 into genes h and g_3.

using another special kind of phylogenetic tree, called *gene tree*. A *gene tree* is a rooted binary tree whose leaves are (not necessarily uniquely) labelled by elements of the set Λ. Despite the fact that biologically speaking leaves in the gene tree represent genes, for simplification, the gene tree is labelled according to the species from which the corresponding gene was sampled. Therefore, leaves similarly labelled represent duplicated genes that evolved independently and appear in a common extant species. As in the species tree, the root and the internal nodes respectively represent the common ancestor and ancestral genes explaining their evolution.

With regards to the set of labels Λ, gene and species trees are said to be *comparable*. Nevertheless, due to complex evolutionary processes, such as gene duplications and losses, comparable gene trees and species trees very often present incompatibilities. An interesting problem is then to reconcile the gene trees and species trees with hypothetical gene duplications. For example, in Fig. 1, given a comparable gene tree and species tree inducing incompatibilities, one can infer a reconciled tree based on the *a priori* duplication of gene g_1 into genes h and g_3 (h is a hypothetical ancestor of genes g_2, g_4), which afterwards both speciate according to the topology of the species tree.

Reconciliation is a widely-investigated problem, and different approaches have been proposed in the past based on the duplication-loss model [5–14] and also extended to consider later gene transfer [15–19]. Some approaches are based on a probabilistic model that aims to infer how a gene tree evolves within a given species tree [5,6,17].

Based on the principle of parsimony, one is interested in finding the minimum number of gene evolutionary events that can explain all the incompatibilities. Notice that, while we focus on minimizing duplications, other possible costs have been considered, for example the minimization of losses or the minimization of duplications and losses [12,9,20].

This last can be inferred by the so-called *lowest common ancestor mapping* (`lca` mapping), denoted by $\mathcal{M}$. $\mathcal{M}$ maps each ancestral gene g of the gene tree to the most recent common ancestor of the extant species from which all the descendants of g were sampled. Given $\mathcal{M}$, a gene in the gene tree is a gene duplication if it has a descendant with the same $\mathcal{M}$ mapping. Then, the reconciliation cost is defined as the number of gene duplications in the gene tree induced by the species tree. Computing a specie tree inducing the minimum cost for this distance has been widely investigated under the name of the Minimum Duplication problem [21,12,20,22] (defined formally afterwards).

1.1. Known results

The Minimum Duplication problem is known to be NP-hard [12]. Recently, the Minimum Duplication problem has been related to the Minimum Triplets Consistency problem [22], a problem known to be W[2]-hard [23] and not approximable within factor $O(\log n)$ [23]. These results coupled with the reduction provided in [22] implies that the Minimum Duplication problem is NP-hard, W[2]-hard (despite of [21]) and cannot be approximated within factor $O(2^{\log^{1-\varepsilon} n})$, even in the specific case of a forest composed of uniquely leaf-labelled gene trees with three leaves [22,24] (notice that if the forest consists of a constant number of uniquely leaf-labelled gene trees with three leaves, then the problem is trivially in P).

Therefore, different heuristics and Integer Linear Programs have been developed [25,26,9,27].

Recently, the Minimum Duplication Bipartite problem has been introduced to tackle the Minimum Duplication problem [28]. The Minimum Duplication Bipartite problem aims to find all the *pre-duplications*, that is duplications that in the evolution precede the first speciation with respect to a species tree (see Fig. 2 for an example). Roughly, this means that only the first level of the species tree is considered. Indeed, one is interested in knowing if a given species belongs to the subtree of S rooted at the left child of the root or at the right one. Therefore, one can view the species tree as a bipartition (Λ_1, Λ_2) of the set of species Λ. Solving the Minimum Duplication Bipartite problem recursively produces a natural greedy heuristic for the Minimum Duplication problem. The Minimum Duplication Bipartite problem was shown to be 2-approximable [28], but its complexity remains open.

In this contribution, we provide results for both the Minimum Duplication problem and the Minimum Duplication Bipartite problem. First of all, we prove that the Minimum Duplication problem is APX-hard, even when the input consists of five uniquely leaf-labelled gene trees (that is for a constant number of gene trees). Then, we show that the Minimum Duplication Bipartite problem can be solved efficiently with a randomized algorithm when the input gene trees have bounded depth.