Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 322 (2016) 35–50

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Correcting Gene Trees by Leaf Insertions: Complexity and Approximation

Stefano Beretta[1]

DISCo
Università di Milano-Bicocca
Milano, Italy

Riccardo Dondi[2]

Dipartimento di Scienze Umane e Sociali
Università di Bergamo
Bergamo, Italy

Abstract

Gene tree correction has recently gained interest in phylogenomics, as it gives insights in understanding the evolution of gene families. Following some recent approaches based on leaf edit operations, we consider a variant of the problem where a gene tree is corrected by inserting leaves with labels in a multiset M. We show that the problem of deciding whether a gene tree can be corrected by inserting leaves with labels in M is NP-complete. Then, we consider an optimization variant of the problem that asks for the correction of a gene tree with leaves labeled by a multiset M', with $M' \supseteq M$, having minimum size. For this optimization variant of the problem, we present a factor 2 approximation algorithm.

Keywords: Computational biology, Phylogenomics, Gene tree corrections, Gene Tree-Species Tree Reconciliation, Algorithms, Computational Complexity.

1 Introduction

The understanding of genome evolution is related to the identification of which evolutionary events (mainly speciations, duplications and losses, in some models lateral gene transfers) lead to the evolution of a genome [23,16]. The evolution of a *gene family* (a set of genes that originate through duplications from an ancestral gene) for a given set of species is usually represented by a *gene tree*. Once a gene tree is computed, usually via methods that rely on sequence similarity, it is compared with a species tree (a tree that represents the evolution of the set of species analysed)

[1] Email: stefano.beretta@disco.unimib.it

[2] Email: riccardo.dondi@unibg.it

http://dx.doi.org/10.1016/j.entcs.2016.03.004
1571-0661/© 2016 The Authors. Published by Elsevier B.V.

in order to identify which evolutionary events occurred in the evolution of the considered gene family [24,26]. Species trees are phylogenetic trees that are based only on speciation events, thus the evolutionary histories represented by a gene tree and a species tree can be different. The *reconciliation* of a gene tree and a species tree [7,8,9,18,25,27,31,14,3] compares the two trees, in order to infer the evolutionary events represented in the gene tree.

A related problem is the inference of the species tree, starting from a set of potentially discordant gene trees. This problem has been intensively studied under different models (see for example [10,22,2,30]) and it is known to be intractable [6,22,12,4,20].

One of the main drawbacks of reconciliation is that gene trees usually contain errors that alter the resulting evolutionary scenario [19,28]. Thus several approaches [11,15,17,29,13,21,5] have been proposed to correct gene trees before the reconciliation.

In this paper, we consider an approach that aims to remove a special kind of duplications, called *Non-Apparent Duplications* (NAD). NAD nodes can be related to errors in the gene trees [10,29], since they represent a disagreement between a gene tree and a species tree that is not directly related to a gene duplication. Thus, some recent approaches to gene tree correction aim to modify the structure of a given gene tree so that it does not contain NAD nodes, via polytomy refinement [21] or by edit operations (removal and modification) on misplaced leaves/labels [29,13,5]. More precisely, the approaches considered in [29,5] introduced two edit operations on leaves (leaf deletion and leaf modification). Here, following a similar approach, we introduce a third edit operation on leaves, *leaf insertion*, and we consider a combinatorial problem, called LeafIns, that aims to remove NAD nodes by inserting leaves associated with a given multiset M of labels. The multiset M represents a set of candidate missing leaves in the gene tree, due to errors in the reconstruction process. We consider the computational complexity of the LeafIns problem, and we show in Section 3 that it is NP-complete. Then, we consider a natural optimization version of this problem, called MinLeafIns, that aims at correcting a given gene tree by inserting the minimum number of leaves labeled by a multiset $M' \supseteq M$. For this optimization problem (which is NP-hard by the previous result), we give in Section 4 a polynomial time approximation algorithm of factor 2.

The paper is organized as follows. In Section 2, we give some preliminary definitions and properties of gene trees and species trees, and we formally introduce the two combinatorial problems we are interested in. In Section 3, we show that the LeafIns problem is NP-complete, while in Section 4 we give an approximation algorithm of factor 2 for MinLeafIns. Finally, we conclude the paper with some open problems.

2 Preliminaries

In this section, first we introduce some preliminary concepts, and we give the formal definitions of the two combinatorial problems we are interested in.

Let $\Lambda = \{l_1, l_2, \ldots, l_m\}$ be a set of labels, where each label represents a different