Journal of Discrete Algorithms 8 (2010) 65–75

Contents lists available at ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Worst-case optimal approximation algorithms for maximizing triplet consistency within phylogenetic networks

Jaroslaw Byrka [a,b], Pawel Gawrychowski [c], Katharina T. Huber [d], Steven Kelk [a,*]

[a] Centrum voor Wiskunde en Informatica, Kruislaan 413, NL-1098 SJ Amsterdam, The Netherlands
[b] Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[c] Institute of Computer Science, University of Wroclaw, ul. Joliot-Curie 15, 50-383 Wroclaw, Poland
[d] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

ARTICLE INFO

Article history:
Received 7 July 2008
Accepted 18 January 2009
Available online 23 February 2009

Keywords:
Triplet
Phylogenetic network
Level-k network

ABSTRACT

The study of phylogenetic networks is of great interest to computational evolutionary biology and numerous different types of such structures are known. This article addresses the following question concerning rooted versions of phylogenetic networks. What is the maximum value of $p \in [0, 1]$ such that for every input set T of rooted triplets, there exists some network $\mathcal{N}$ such that at least $p|T|$ of the triplets are consistent with $\mathcal{N}$? We call an algorithm that computes such a network (where p is maximum) worst-case optimal. Here we prove that the set containing all triplets (the full triplet set) in some sense defines p. Moreover, given a network $\mathcal{N}$ that obtains a fraction p' for the full triplet set (for any p'), we show how to efficiently modify $\mathcal{N}$ to obtain a fraction $\geqslant p'$ for any given triplet set T. We demonstrate the power of this insight by presenting a worst-case optimal result for level-1 phylogenetic networks improving considerably upon the 5/12 fraction obtained recently by Jansson, Nguyen and Sung. For level-2 phylogenetic networks we show that $p \geqslant 0.61$. We emphasize that, because we are taking $|T|$ as a (trivial) upper bound on the size of an optimal solution for each specific input T, the results in this article do not exclude the existence of approximation algorithms that achieve approximation ratio better than p. Finally, we note that all the results in this article also apply to weighted triplet sets.

1. Introduction

Capturing the idea that evolution proceeds in a tree-like manner, *phylogenetic trees* – rooted trees whose root vertex is of degree 2 and whose leaf set is bijectively labeled by a set X of taxa – have been used for many decades to model evolution. Advances in DNA sequencing have not only resulted in large amounts of data on which such trees may be based but also provided evidence that, at least for some plant or microbial species, a tree-like evolutionary model might be too simplistic to capture their evolutionary past due to hybridization and lateral gene transfer [8,18–20,25]. The development of structures capable of accommodating such *reticulation events* has received much attention in the literature (see e.g. [2,6,7,22]) and lies at the heart of the thriving area of phylogenetic network construction.

Numerous different models of (phylogenetic) networks have been introduced over the years, see e.g., [11,12] for an overview. Amongst these models *level-k networks*, k some non-negative integer, have recently received a considerable amount of attention (see Fig. 1(a) for an example for $k = 1$) [13,14,16].

* Corresponding author.
E-mail addresses: J.Byrka@cwi.nl (J. Byrka), gawry@cs.uni.wroc.pl (P. Gawrychowski), katharina.huber@cmp.uea.ac.uk (K.T. Huber), S.M.Kelk@cwi.nl (S. Kelk).

1570-8667/$ – see front matter
doi:10.1016/j.jda.2009.01.004

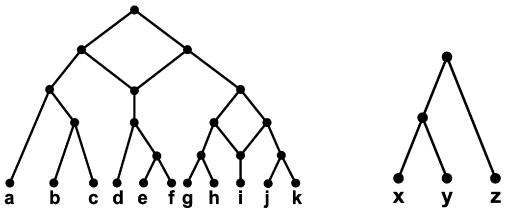

Fig. 1. (a) An example of a level-1 network on the set $\{a, \ldots, k\}$. (b) A rooted triplet $xy|z$ on the set $\{x, y, z\}$. In all figures in this article the root is the topmost vertex and arcs are assumed to be directed downwards, away from it.

Level-k networks are underpinned by a rooted directed acyclic graph whose root (which has degree 2) is the unique source and all of whose sinks are bijectively labeled by the species under consideration. Representing a reticulation event in terms of a reticulation vertex (i.e., a vertex in the network which has outdegree 1 and indegree 2), then the parameter k is the maximum number of reticulation vertices that each biconnected component of the network may have (see Section 2 for a precise definition). Level-1 networks are also known as *galled networks* in the literature.

(*Rooted*) *triplets*, that is, phylogenetic trees on just 3 species (see Fig. 1(b) for an example) are sometimes called the fundamental building blocks of phylogenetics. One of their main attractions lies in the fact that they can be estimated without too much difficulty [13]. Thus it is not surprising that the problem of reconstructing a phylogenetic tree or network from a set of triplets such that the tree or network is *consistent* (formal definition follows later) with that set has received a considerable amount of attention (see e.g. [3]). For example, Aho et al. [1] showed a simple polynomial-time algorithm which, given a set of triplets, finds a phylogenetic tree consistent with all the triplets, or shows that no such tree exists. For the equivalent problem in level-1 and level-2 networks the problem becomes NP-hard [14,16], although the problem becomes polynomial-time solvable if the input triplets are *dense* i.e., if there is at least one triplet in the input for each subset of three species [14,17].

Several authors have considered algorithmic strategies of use when the algorithms from [1] and [17] fail to find a phylogenetic tree or network. Gąsieniec et al. [9] gave a polynomial-time algorithm which always finds a phylogenetic tree consistent with at least 1/3 of the (weighted) input triplets, and furthermore showed that 1/3 is best possible when all possible triplets on n species (the *full triplet set*) are given as input. On the negative side, [4,15,27] showed that it is NP-hard to find a tree consistent with a maximum number of input triplets. In the context of level-1 networks, in [16] a polynomial-time algorithm is presented which produces a level-1 network consistent with at least $5/12 \approx 0.4166$ of the input triplets. The authors of that paper also described an upper-bound, which is a function of the number of distinct species n in the input, on the percentage of input triplets that can be consistent with a level-1 network. As in [9] this upper bound is tight in the sense that it is the best possible for the full triplet set on n species. They computed a value of n for which their upper bound equals approximately 0.4883, showing that in general a fraction better than this is not possible. The apparent convergence of this bound from above to $0.4880\ldots$ begs the question, however, whether a fraction better than 5/12 is possible for level-1 networks, and whether the full triplet set is in general always the worst-case scenario for such fractions.

In this paper we answer these questions in the affirmative, and in fact we give a much stronger result. In particular, we develop a probabilistic argument that (as far as such fractions are concerned) the full triplet set is indeed always the worst possible case, irrespective of the type of network being studied (Proposition 1). Furthermore, our main result (Theorem 1) establishes that, for any network $\mathcal{N}$ which achieves a fraction p for the full triplet set, its leaves (i.e., the outdegree zero vertices) can be relabeled so that the resulting network $\mathcal{N}'$ achieves a fraction $\geqslant p$ for a given triplet set T. It relies on the (re)labeling procedure LEAFLABEL. This procedure is not only fully general but also enjoys an optimized running time thanks to fast, novel algorithms for checking triplet consistency.

In addition to shedding more light into the results established in [9], Theorem 1 also has consequences for level-1 and level-2 networks. More precisely, it is used to establish Theorem 2 which states that a level-1 network can be constructed in polynomial-time which is consistent with a fraction *exactly equal* to the level-1 upper-bound identified in [16]. This is worst case optimal for all level-1 networks in the sense that we are optimizing with respect to $|T|$, the number of triplets in the input, not $Opt(T)$, the size of the optimal solution for that specific T. In addition, it is also crucial for establishing Theorem 3 which states that for any triplet set T, a level-2 network can be constructed in polynomial time that is consistent with at least a fraction 0.61 of the triplets in T.

The paper is organized as follows. In the next section we present some preliminaries. In Section 3, we present the derandomization procedure which underpins our main result. Sections 3.1–3.4 are devoted to its proof. Theorems 2 and 3 are established in Sections 4.1 and 4.2, respectively. We discuss in Section 5 the complexity of optimization with regards to $Opt(T)$ (as opposed to $|T|$). The underlying rationale being that since $Opt(T)$ is always bounded above by $|T|$, an algorithm that obtains a fraction p of the input T is trivially also a p-approximation for the corresponding problem in terms of $Opt(T)$.

Throughout the paper, X is a finite set of taxa (e.g., species) and $V(G)$ and $A(G)$ denote the vertex set and arc set of a graph G, respectively. In addition, in all figures, the (unique) root of a rooted directed graph is always on the top and all arcs are assumed to be directed downwards, away from the root.