Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the existence of a cherry-picking sequence

Janosch Döcker [a], Simone Linz [b,*]

[a] Department of Computer Science, University of Tübingen, Germany
[b] Department of Computer Science, University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 21 August 2017
Received in revised form 30 November 2017
Accepted 5 December 2017
Available online xxxx
Communicated by T. Calamoneri

Keywords:
2P2N-3-SAT
Cherry
Cherry-picking sequence
Intermezzo
Phylogenetic tree
Temporal phylogenetic network

ABSTRACT

Recently, the minimum number of reticulation events that is required to simultaneously embed a collection $\mathcal{P}$ of rooted binary phylogenetic trees into a so-called temporal network has been characterized in terms of cherry-picking sequences. Such a sequence is a particular ordering on the leaves of the trees in $\mathcal{P}$. However, it is well-known that not all collections of phylogenetic trees have a cherry-picking sequence. In this paper, we show that the problem of deciding whether or not $\mathcal{P}$ has a cherry-picking sequence is NP-complete for when $\mathcal{P}$ contains at least eight rooted binary phylogenetic trees. Moreover, we use automata theory to show that the problem can be solved in polynomial time if the number of trees in $\mathcal{P}$ and the number of cherries in each such tree are bounded by a constant.

1. Introduction

To represent evolutionary relationships among species, phylogenetic trees have long been a powerful tool. However, as we now not only acknowledge speciation but also non-tree-like processes such as hybridization and lateral gene transfer to be driving forces in the evolution of certain groups of organisms (e.g. bacteria, plants, and fish) [16,20], phylogenetic networks become more widely used to represent ancestral histories. A phylogenetic network is a generalization of a rooted phylogenetic tree. More precisely, such a network is a rooted directed acyclic graph whose leaves are labeled [14].

The following optimization problem, which is biologically relevant and mathematically challenging, motivates much of the theoretical work that has been done in reconstructing phylogenetic networks from phylogenetic trees. Given a collection $\mathcal{P}$ of rooted binary phylogenetic trees on a set of species such that $\mathcal{P}$ correctly represents the tree-like evolution of different parts of the species' genomes, what is the smallest number of reticulation events that is required to simultaneously embed the trees in $\mathcal{P}$ into a phylogenetic network? Here, reticulation events are collectively referring to all non-tree-like events and they are represented by vertices in a phylogenetic network whose in-degree is at least two. Without any structural constraints on a phylogenetic network, it is well-known that $\mathcal{P}$ can always be embedded into such a network [2,19] and, hence, the optimization problem is well-defined. Moreover, despite the problem being NP-hard [4], even for when $|\mathcal{P}| = 2$, several exact algorithms have been developed that, given two rooted phylogenetic trees, construct a phylogenetic network whose number of reticulation events is minimized over the space of all networks that embed both trees [1,7,18,22].

Motivated by the introduction of temporal networks [3,17], which are phylogenetic networks that satisfy several time constraints, Humphries et al. [12,13] recently investigated the special case of the aforementioned optimization problem for

* Corresponding author.
E-mail addresses: janosch.doecker@uni-tuebingen.de (J. Döcker), s.linz@auckland.ac.nz (S. Linz).

https://doi.org/10.1016/j.tcs.2017.12.005
0304-3975/

when one is interested in minimizing the number of reticulation events over the smaller space of all temporal networks that embed a given collection of rooted binary phylogenetic trees. More precisely, in the context of their two papers, the authors considered *temporal networks* to be phylogenetic networks that satisfy the following three constraints:

(1) speciation events occur successively,
(2) reticulation events occur instantaneously, and
(3) each non-leaf vertex has a child whose in-degree is one.

The second constraint implies that the three species that are involved in a reticulation event, i.e. the new species resulting from this event and its two distinct parents, must coexist in time. Moreover, a phylogenetic network that satisfies the third constraint (but not necessarily the first two constraints) is referred to as a *tree-child* network in the literature [6]. Intuitively, if a phylogenetic network $\mathcal{N}$ is temporal, then one can assign a time stamp to each of its vertices such that the following holds for each edge (u, v) in $\mathcal{N}$. If v is a reticulation, then the time stamp assigned to u is the same as the time stamp assigned to v. Otherwise, the time stamp assigned to v is strictly greater than that assigned to u. Baroni et al. [3] showed that it can be checked in polynomial time whether or not a given phylogenetic network satisfies the first two constraints.

Humphries et al. [12] have established a new characterization to compute the minimum number of reticulation events that is needed to simultaneously embed an arbitrarily large collection $\mathcal{P}$ of rooted binary phylogenetic trees into a temporal network. This characterization, which is formally defined in Section 2, is in terms of *cherries*, and the existence of a particular type of sequence on the leaves of the trees, called a *cherry-picking sequence*. It was shown that such a sequence for $\mathcal{P}$ exists if and only if the trees in $\mathcal{P}$ can simultaneously be embedded into a temporal network [12, Theorem 1]. Moreover, a cherry-picking sequence for $\mathcal{P}$ can be exploited further to compute the minimum number of reticulation events that is needed over all temporal networks. Importantly, not every collection $\mathcal{P}$ is guaranteed to have a solution, i.e. there may be no cherry-picking sequence for $\mathcal{P}$ and, hence no temporal network that embeds all trees in $\mathcal{P}$. It was left as an open problem by Humphries et al. [12] to analyze the computational complexity of deciding whether or not $\mathcal{P}$ has a cherry-picking sequence for when $|\mathcal{P}| = 2$.

In this paper, we make progress towards this question and show that it is NP-complete to decide if $\mathcal{P}$ has a cherry-picking sequence for when $|\mathcal{P}| \geq 8$. Translated into the language of phylogenetic networks, this result directly implies that it is computationally hard to decide if a collection of at least eight rooted binary phylogenetic trees can simultaneously be embedded into a temporal network. To establish our result, we use a reduction from a variant of the Intermezzo problem [9]. On a more positive note, we show that deciding if $\mathcal{P}$ has a cherry-picking sequence can be done in polynomial time if the number of trees and the number of cherries in each such tree are bounded by a constant. To this end, we explore connections between phylogenetic trees and automata theory and show how the problem at hand can be solved by using a deterministic finite automaton.

The remainder of the paper is organized as follows. The next section contains notation and terminology that is used throughout the paper. Section 3 establishes NP-completeness of a variant of the Intermezzo problem which is then, in turn, used in Section 4 to show that it is NP-complete to decide if $\mathcal{P}$ has a cherry-picking sequence for when $|\mathcal{P}| \geq 8$. In Section 5, we show that deciding if $\mathcal{P}$ has a cherry-picking sequence is polynomial-time solvable if the number of cherries in each tree and the size of $\mathcal{P}$ are bounded by a constant. We finish the paper with some concluding remarks in Section 6.

2. Preliminaries

This section provides notation and terminology that is used in the subsequent sections. Throughout this paper, X denotes a finite set.

Phylogenetic trees. A *rooted binary phylogenetic X-tree* $\mathcal{T}$ is a rooted tree with leaf set X and, apart from the root which has degree two, all interior vertices have degree three. Furthermore, a pair of leaves $\{a, b\}$ of $\mathcal{T}$ is called a *cherry* if a and b are leaves that are adjacent to a common vertex. Note that every rooted binary phylogenetic tree has at least one cherry. We denote by $c_{\mathcal{T}}$ the number of cherries in $\mathcal{T}$. We now turn to a rooted binary phylogenetic tree with exactly one cherry. More precisely, we call $\mathcal{T}$ a *caterpillar* if $|X| = n \geq 2$ and the elements in X can be ordered, say $x_1, x_2, \ldots, x_n$, so that $\{x_1, x_2\}$ is a cherry and, if p_i denotes the parent of x_i, then, for all $i \in \{3, 4, \ldots, n\}$, we have (p_i, p_{i-1}) as an edge in $\mathcal{T}$, in which case we denote the caterpillar by $(x_1, x_2, \ldots, x_n)$. To illustrate, Fig. 1 shows the caterpillar $(D_1, D_2, \ldots, D_{|A'|})$ with cherry $\{D_1, D_2\}$. Two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ are said to be *isomorphic* if the identity map on X induces a graph isomorphism on the underlying trees.

Subtrees. Now, let $\mathcal{T}$ be a rooted binary phylogenetic X-tree, and let $X' = \{x_1, x_2, \ldots, x_k\}$ be a subset of X. The minimal rooted subtree of $\mathcal{T}$ that connects all vertices in X' is denoted by $\mathcal{T}(X')$. Furthermore, the rooted binary phylogenetic tree obtained from $\mathcal{T}(X')$ by contracting all non-root degree-2 vertices is the *restriction of $\mathcal{T}$ to X'* and is denoted by $\mathcal{T}|X'$. We also write $\mathcal{T}[-x_1, x_2, \ldots, x_k]$ or $\mathcal{T}[-X']$ for short to denote $\mathcal{T}|(X - X')$. For a set $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m\}$ of rooted binary phylogenetic X-trees, we write $\mathcal{P}|X'$ (resp. $\mathcal{P}[-X']$) when referring to the set $\{\mathcal{T}_1|X', \mathcal{T}_2|X', \ldots, \mathcal{T}_m|X'\}$ (resp. $\{\mathcal{T}_1[-X'], \mathcal{T}_2[-X'], \ldots, \mathcal{T}_m[-X']\}$). Lastly, a rooted binary phylogenetic tree is *pendant* in $\mathcal{T}$ if it can be detached from $\mathcal{T}$ by deleting a single edge.

Please cite this article in press as: J. Döcker, S. Linz, On the existence of a cherry-picking sequence, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.12.005