ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

On compatibility and incompatibility of collections of unrooted phylogenetic trees

David Fernández-Baca *, Sudheer R. Vakati [1]

Department of Computer Science, Iowa State University, Ames, IA 50010, USA

ARTICLE INFO

Article history:
Received 28 December 2015
Received in revised form 30 April 2017
Accepted 8 May 2017
Available online xxxx

Keywords:
Phylogenetics
Triangulations
Treewidth
Graph minors

ABSTRACT

Let $\mathcal{P} = \{T_1, \ldots, T_k\}$ be a collection of phylogenetic trees over various subsets of a set of species. For each $i \in \{1, \ldots, k\}$, let $\mathcal{L}(T_i)$ denote the set of species in tree T_i. A *supertree* for $\mathcal{P}$ is a phylogenetic tree with species set $\bigcup_{i=1}^{k} \mathcal{L}(T_i)$. The *tree compatibility problem* asks whether there exists a supertree T for $\mathcal{P}$ such that, for each $i \in \{1, \ldots, k\}$, T_i can be obtained from $T|\mathcal{L}(T_i)$ — the minimal subtree of T spanning $\mathcal{L}(T_i)$ — by zero or more contractions of internal edges. If the answer is "yes", then $\mathcal{P}$ is said to be *compatible*; otherwise, $\mathcal{P}$ is *incompatible*.

We characterize compatibility via graph triangulations and tree decompositions. We then study how to make an incompatible collection of trees compatible through edge contraction and tree deletion. Finally, we introduce the notion of a phylogenetic minor to study under which conditions edge contraction, tree removal, and species removal/renaming operations preserve compatibility.

1. Introduction

A *phylogenetic tree* T is an unrooted tree with no vertices of degree two, whose leaves are in one-to-one correspondence with a label set $\mathcal{L}(T)$. The labels are species – also known as *taxa*–, and T represents the evolutionary history of these species. The following problem arises when constructing phylogenetic trees for a collection of species. We are given a collection of phylogenetic trees with partially overlapping species sets. The question is to determine if there is a tree that exhibits the information from each of them — i.e., whether or not the information present in the input trees is compatible. To explain this precisely, we need some definitions.

Let $\mathcal{P} = \{T_1, T_2, \ldots, T_k\}$ denote a collection of k phylogenetic trees; we refer to $\mathcal{P}$ as a *profile*, and to each tree in $\mathcal{P}$ as an *input tree*.[2] We write $\mathcal{L}(\mathcal{P})$ to denote $\bigcup_{i=1}^{k} \mathcal{L}(T_i)$. A *supertree* for $\mathcal{P}$ is a phylogenetic tree T with label set $\mathcal{L}(T) = \mathcal{L}(\mathcal{P})$. The *tree compatibility problem* asks whether there exists a supertree T for $\mathcal{P}$ such that, for each $i \in \{1, \ldots, k\}$, T_i can be obtained from $T|\mathcal{L}(T_i)$ – the minimal subtree of T spanning $\mathcal{L}(T_i)$ – by zero or more contractions of internal edges. If the answer is "yes", then $\mathcal{P}$ is said to be *compatible*; otherwise, $\mathcal{P}$ is *incompatible*.

In this paper, we study the tree compatibility problem from the perspective offered by the *reduced display graph*, a graph that represents a profile solely by the topologies induced by the internal vertices of the input trees and the intersection patterns among sets of labels associated with internal vertices. By dispensing with species labels, reduced display graphs

* Corresponding author.
E-mail addresses: fernande@iastate.edu (D. Fernández-Baca), svakati@gmail.com (S.R. Vakati).
[1] Current address: amazon.com, Seattle, WA, USA.
[2] We allow $\mathcal{P}$ to contain subsets consisting of isomorphic input trees.

http://dx.doi.org/10.1016/j.dam.2017.05.002
0166-218X/

Please cite this article in press as: D. Fernández-Baca, S.R. Vakati, On compatibility and incompatibility of collections of unrooted phylogenetic trees, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.05.002.

enable us to define an equivalence relation among profiles, where two profiles – possibly with different label sets – are in the same equivalence class if they have isomorphic reduced display graphs. Using reduced display graphs, we

(i) characterize compatibility via triangulations and tree decompositions,
(ii) study the conditions under which incompatible profiles can be made compatible through edge contraction and tree deletion, and
(iii) define the notion of a *phylogenetic minor*, to study the conditions under which edge contraction, tree removal, and species removal/renaming preserve compatibility.

Before giving a detailed overview of our results, we review previous work on tree compatibility, in order to provide some context for our work.

Related work. The tree compatibility problem was proved to be NP-complete by Steel [26]. In the same paper, Steel showed that Aho et al.'s algorithm for inferring a tree from lowest common ancestor constraints [1] leads to a polynomial-time algorithm to test the compatibility of *rooted* trees. Note that a profile of rooted trees is equivalent to a profile of unrooted trees that share a common label, namely the root.

Steel's NP-completeness proof demonstrates that testing the compatibility of a collection of *quartets* – that is, phylogenetic trees with four leaves – is NP-complete. Thus, in his proof, both the total number of labels, n, and the number of trees, k, can be arbitrarily large. But what if we fix one of these two parameters? The case where n is fixed, is trivial, since the number of possible supertree topologies is also fixed, allowing us to test compatibility in time polynomial in k. The case where k is fixed is more interesting. In a seminal paper [7], Bryant and Lagergren showed that there exists an algorithm for the tree compatibility problem that runs in $O(n \cdot f(k))$ time, where f is a function of k; that is, tree compatibility is *fixed-parameter tractable* [12] with respect to k. In their proof, they introduced the notion of the *display graph* of a profile $\mathcal{P}$, the graph obtained from the disjoint union of the trees in $\mathcal{P}$ by identifying nodes with the same label[3] (see Section 2). Bryant and Lagergren showed that the display graph of a compatible profile must have tree-width at most k. They then showed that compatibility can be expressed via a monadic second-order logic (MSOL) formula on the display graph. The existence of a polynomial-time algorithm for compatibility follows from Courcelle's Theorem [2,9].

The connection between tree compatibility and tree-width is intuitive: the fact that the tree-width of the display graph of a compatible profile is at most k suggests that all input trees can be "packed" into a single tree. A further link with tree-width is suggested by Buneman's classic theorem on *character compatibility* [8], a problem that is closely related to tree compatibility (see Section 2, Theorem 2.1). Buneman's Theorem characterizes character compatibility via the existence of a certain kind of triangulation in a graph constructed from the given characters. Graph triangulations and chordal graphs are closely related to tree-width and tree decompositions.

The intuitive connections among tree compatibility, tree-width, and triangulations have motivated several efforts to elucidate the precise mathematical nature of these relationships. In a previous paper [28], we characterized tree compatibility in terms of triangulations of the display graph, in a manner analogous to Buneman's Theorem (we review this characterization in Section 4). Gysel and Gusfield [16] provided an alternative characterization, based on triangulations of a graph – the edge-label intersection graph – that is closely related to the display graph. In a subsequent paper [29], we used Gysel and Gusfield's ideas to obtain a characterization of compatibility based on the existence of certain sets of cuts in the display graph. We also provided a cut-based characterization of *agreement*, the version of compatibility where the supertree T must satisfy the stricter condition that $T|\mathcal{L}(T_i)$ is isomorphic to T_i. Note that testing agreement, like testing compatibility, is NP-complete; in fact, Steel's aforementioned proof of the NP-completeness of compatibility also implies the NP-completeness of agreement. Grigoriev et al. [15] studied the relationship between compatibility and tree-width for profiles of binary trees. They showed that if the tree-width of the display graph of such a profile is at most two, then the profile is compatible, and a supertree that displays the profile can be constructed in polynomial time. They also showed that there exist both compatible and incompatible profiles that have display graphs with tree-width three.

There has also been work exploring and extending the algorithmic implications of Bryant and Lagergren's result. Scornavacca et al. [24] showed that the agreement problem is fixed-parameter tractable in k. Their proof, like that of Bryant and Lagergren's result for compatibility, relies on bounded tree-width, MSOL, and Courcelle's Theorem. Although of great theoretical interest, results based on these techniques do not yield practical algorithms for compatibility or agreement. Some steps towards achieving more efficient solutions to both problems are reported by Baste et al. [4]. Display graphs, tree-width and MSOL have been used to study the existence of fixed-parameter algorithms to compute measures of incongruence between phylogenetic trees [19,20].

Recently, display graphs were used to obtain an algorithm to test the compatibility of a profile of rooted trees whose running time is within a polylogarithmic factor of optimality [11]. The algorithm was later extended to handle rooted trees where, as in the case of taxonomic trees, internal nodes may be labeled [10].

[3] We note that a graph similar to the display graph – the "tree alignment graph" – has been proposed independently by evolutionary biologists [18, especially Figure 3].

Please cite this article in press as: D. Fernández-Baca, S.R. Vakati, On compatibility and incompatibility of collections of unrooted phylogenetic trees, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.05.002.