Discrete Applied Mathematics 157 (2009) 2361–2369

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Phylogenetic graph models beyond trees

Ulrik Brandes, Sabine Cornelsen *

Department of Computer & Information Science, University of Konstanz, Box D 67, 78457 Konstanz, Germany

ARTICLE INFO

Article history:
Received 5 January 2007
Received in revised form 16 June 2007
Accepted 11 June 2008
Available online 11 September 2008

Keywords:
Phylogenetic trees
Graph models
Splits
Compatibility
Cactus model

ABSTRACT

A graph model for a set $\mathfrak{S}$ of splits of a set X consists of a graph and a map from X to the vertices of the graph such that the inclusion-minimal cuts of the graph represent $\mathfrak{S}$. Phylogenetic trees are graph models in which the graph is a tree. We show that the model can be generalized to a cactus (i.e. a tree of edges and cycles) without losing computational efficiency. A cactus can represent a quadratic rather than linear number of splits in linear space. We show how to decide in linear time in the size of a succinct representation of $\mathfrak{S}$ whether a set of splits has a cactus model, and if so construct it within the same time bounds. As a byproduct, we show how to construct the subset of all compatible splits and a maximal compatible set of splits in linear time. Note that it is $\mathcal{NP}$-complete to find a compatible subset of maximum size. Finally, we briefly discuss further generalizations of tree models.

1. Introduction

The goal of phylogenetic analysis is to determine and describe the evolutionary relationship between species (taxa). A phylogenetic tree describes in particular the branching process when during time a species is divided into two separate species. One method of obtaining such an evolutionary tree is to consider a bunch of properties (binary characters) that the actual species may or may not have. Then the goal is to construct a tree such that in particular the leaves are labeled with the different species and the properties correspond to the edges: The sets of species mapped to the two connected components of the tree deleting one edge corresponds to the set of species having or not having the corresponding property. Hence, a binary character induces a split, i.e., a partition of the set of taxa into two non-empty parts. In the following, we assume that each split is given by the smaller of its two subsets.

Not every set of splits can be represented in a phylogenetic tree. The splits have to be pairwise compatible, i.e., the intersection of two splits S and T has to be S, T, or the empty set. Given a set $\mathfrak{S}$ of m pairwise compatible splits of a set X of n taxa, Gusfield [19] showed how to construct an evolutionary tree in $\mathcal{O}(mn)$ time. Although he also gives a matching lower bound for the worst case, Agarwala et al. [1] improved the running time for constructing a phylogenetic tree to $\mathcal{O}(n+m+f)$ time, where $f \leq mn/2$ is the sum of the sizes of all splits. The Buneman graph [8] or the tree-popping algorithm of Meacham [27,28] are other approaches for constructing the phylogenetic tree of a set of compatible splits. One way of handling incompatible sets of splits is to compute a significant compatible subset of splits. It was shown by Day and Sankoff [10] that the problem of finding a maximum compatible subset of splits is $\mathcal{NP}$-complete. On the other hand, it is well known, that a maximal compatible subset of splits can be found greedily. We sketch how the greedy algorithm can be implemented to run in $\mathcal{O}(n+m+f)$ time. Both, a maximum and a maximal compatible subset of splits have the disadvantage that they are not unique. Thus, we consider the subset of splits of $\mathfrak{S}$ that are pairwise compatible with all other splits in $\mathfrak{S}$. This subset of all compatible splits is also known as the splits of the loose consensus tree [6] or the kernel splits [3]. A comparison of this set of splits to other compatible subsets of splits can be found in [7, Chapter 6].

* Corresponding author. Tel.: +49 7531 88 4431; fax +49 7531 88 3577.
E-mail addresses: Ulrik.Brandes@uni-konstanz.de (U. Brandes), Sabine.Cornelsen@uni-konstanz.de (S. Cornelsen).

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.06.031

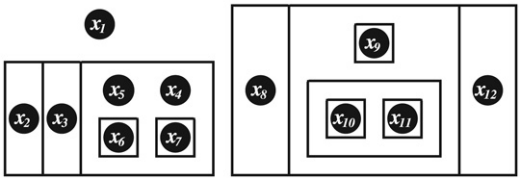

Fig. 1. A drawing of the set of splits in Fig. 2 with axis-parallel rectangles.

Recently, McConnell [25] gave a linear time algorithm for constructing a generalized PQ-tree from which the set of all compatible splits can be deduced. The algorithm is based on the overlap components of Dahlhaus [9]. We give a different algorithm that computes the subset of all compatible splits in $\mathcal{O}(n+m+f)$ time. Both, the algorithm of Dahlhaus [9] and our algorithm use lexicographical sorting as a basic construction step, however, for completely different purposes. We believe that our algorithm is easier to understand and to implement. It also leads to a complete characterization of incompatible splits in terms of prefix trees.

Another way of handling incompatible sets of splits is to extend phylogenetic trees to more complex networks. An overview on the different kinds of phylogenetic networks and their construction can be found, e.g., in [24,23,30]. Two basic types of networks representing splits are: Recombination networks (see e.g. [21]) and networks like splits graphs [2,17] that represent incompatible splits by some minimal cuts of a graph. We use a representation that is similar to the latter case.

A graph model for a set of splits is a graph in which some of the vertices are labeled by the taxa such that there is a one-to-one correspondence between the minimal cuts of the graph and the splits. Note that a phylogenetic tree is a tree model. A cactus is a tree of edges and cycles. Both, a tree model and a cactus model require only $\mathcal{O}(n)$ space. While a tree model can represent linearly many, cactus models can represent up to a quadratic number of splits. A cactus model for the set of splits in Fig. 2(a) is given in Fig. 5. Note that galled trees [21] are also trees of edges and cycles. However, they belong to the category of recombination networks and represent sets of splits differently.

Originally, the cactus model was introduced by Dinitz et al. [13] to represent the set of all minimum cuts of a connected graph. Dinitz and Nutov [14] later characterized all sets of splits that have a cactus model. While the proof is constructive [15], it appears to be difficult to translate it into an algorithm. So far, efficient algorithms are known for constructing the cactus of all minimum cuts of a graph [12,18,31,33]. A cactus model can also be deduced from the generalized PQ-tree of McConnell [25].

Sets of splits that have a cactus model also have a characterization in terms of graph drawing. They are exactly the sets of splits that have a drawing with axis-parallel rectangles [5]. See Fig. 1 for such a drawing. In this paper, we give an algorithm that decides in $\mathcal{O}(n+m+f)$ time whether a set of splits can be represented in a cactus, and if so constructs the model in the same asymptotic running time. The construction is based on the tree model of the subset of all compatible splits. In addition it uses only some easy counting and sorting arguments. In the conclusion, we discuss that our algorithm extends to graph models consisting of trees of edges and cliques or trees of edges, cycles, and cliques, respectively. Sets of splits having such graph models have been characterized in [29,15], respectively.

The paper is organized as follows. In Section 2, we give basic definitions and introduce the necessary concepts. We define the graph model of a set of splits in Section 3. In Section 4 we recall how to construct a tree model for a compatible set of splits utilizing so-called tries. In Sections 5 and 6 we show how to construct a maximal compatible subset and the subset of all compatible splits, respectively. Finally, in Section 7 we examine the existence and construction of a cactus model. We conclude in Section 8.

2. Preliminaries

Throughout this paper, let $X = \{x_1, \ldots, x_n\}$ denote a finite set of n taxa. We will denote the *size* n of the set X by $|X|$. By $S \subset X$ we denote that S is a subset of X including that S might be equal to X. A *split* of X is the unordered pair $\{S, X \setminus S\}$ such that $\emptyset \subsetneq S \subsetneq X$. We say that S induces $\{S, X \setminus S\}$. We will assume that the splits are given by the smaller subset. So, throughout this paper, let $\mathfrak{S}$ denote a set of m non-empty subsets S of X such that $|S| \leq |X \setminus S|$ and such that $\{S, X \setminus S\} \neq \{T, X \setminus T\}$ for two elements $S, T \in \mathfrak{S}$. We will refer to the elements of $\mathfrak{S}$ also as splits. Further, let $f = \sum_{S \in \mathfrak{S}} |S|$. Fixing an ordering $x_1, \ldots, x_n$ of X, a split can be represented by a *characteristic vector*. The characteristic vector of the split induced by S is the vector $(v_1, \ldots, v_n) \in \{0, 1\}^n$ such that for all $i = 1, \ldots, n$ we have $v_i = 1$ if and only if $x_i \in S$. Hence, a set of splits can be represented by a matrix where the columns are the characteristic vectors of the splits (provided an ordering of the splits is fixed). An example is given in Fig. 2(a). A more succinct way of representing a split $\{S, X \setminus S\}$ is to represent the set S that induces it by a *member list*, i.e., the list of elements of S. An example is given in Fig. 2(b). Throughout this paper we assume that splits are given in this succinct representation.

Let $A = (a_{ij})_{i=1,\ldots,n, j=1,\ldots,m}$ be the matrix of characteristic vectors of the ordered set $\mathfrak{S} = \{S_1, \ldots, S_m\}$ of splits with respect to an ordering $x_1, \ldots, x_n$ of the taxa. We say that the splits are *lexicographically sorted* with respect to the fixed ordering of the taxa if the columns of A are sorted lexicographically, i.e., if for $1 \leq j < k \leq m$ it holds that there is an $1 \leq \ell \leq n$ such that $a_{\ell j} = 1$, $a_{\ell k} = 0$ and $a_{ij} = a_{ik}$, $i < \ell$. Analogously, we say that the taxa are lexicographically sorted with respect to an ordering of the splits if the columns of A are lexicographically sorted. A lexicographical sorting of splits or taxa can be constructed in $\mathcal{O}(n + m + f)$ time using partition refinement [34].