Journal of Computer and System Sciences ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

A parameterized algorithm for the Maximum Agreement Forest problem on multiple rooted multifurcating trees ☆

Feng Shi [a], Jianer Chen [b,c], Qilong Feng [a], Jianxin Wang [a,*]

[a] School of Information Science and Engineering, Central South University, Changsha 410083, P.R. China
[b] School of Computer Science & Education Software, Guangzhou University, Guangzhou, Guangdong 510006, P.R. China
[c] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA

ARTICLE INFO

Article history:
Received 11 October 2015
Received in revised form 4 September 2016
Accepted 15 March 2018
Available online xxxx

Keywords:
Maximum agreement forest
Phylogenetic tree
Algorithm

ABSTRACT

The Maximum Agreement Forest problem has been extensively studied in phylogenetics. Most previous work is on two binary phylogenetic trees. In this paper, we study a generalized version of the problem: the Maximum Agreement Forest problem on multiple rooted multifurcating phylogenetic trees, from the perspective of parameterized algorithms. By taking advantage of a new branch-and-bound strategy, a parameterized algorithm with running time $O(2.42^k m^3 n^4)$ is presented for the problem, assuming that all polytomies in the multifurcating phylogenetic trees are hard.

1. Introduction

Phylogenetic trees (alternatively, evolutionary trees) are an invaluable tool in phylogenetics that are used to represent the evolutionary histories of homologous regions of genomes from a collection of extant species or, more generally, taxa. However, due to reticulation events, such as hybridization, recombination, or lateral gene transfer (LGT) in evolution, phylogenetic trees constructed by different homologous regions of genomes may have different structures. Since the reticulation events can be studied by examining these differences in structures, several metrics, such as Robinson–Foulds distance [1], Nearest Neighbor Interchange (NNI) distance [2], Hybridization number [3], Tree Bisection and Reconnection (TBR) distance, and Subtree Prune and Regraft (SPR) distance [4,5], have been proposed in the literature to compare these different phylogenetic trees. Among these metrics, the SPR distance has been studied extensively for investigating phylogenetic inference [6], lateral genetic transfer [7,8], and MCMC search [9].

Given two phylogenetic trees on the same collection of taxa, the SPR distance between the two trees is defined to be the minimum number of "Subtree Prune and Regraft" operations [10] needed to convert one tree to the other. Since the Subtree Prune and Regraft operation has been widely used as a method to model a reticulation event, the SPR distance provides a lower bound on the number of reticulation events needed to reconcile the two phylogenetic trees [11], which can give an indication how reticulation events influence the evolutionary history of the taxa under consideration.

For the study of SPR distance, Hein et al. [12] proposed the concept of *maximum agreement forest* (MAF) for two phylogenetic trees, which is a common subforest of the two trees with the minimum *order* over all common subforests of the two trees (the order of a forest is defined as the number of connected components in the forest). Bordewich and Semple [13]

☆ This work is supported by the National Natural Science Foundation of China under Grants (61420106009, 61672536, 61472449, 61572414).
* Corresponding author.
E-mail addresses: fengshi@csu.edu.cn (F. Shi), chen@cs.tamu.edu (J. Chen), csufeng@csu.edu.cn (Q. Feng), jxwang@mail.csu.edu.cn (J. Wang).

https://doi.org/10.1016/j.jcss.2018.03.002
0022-0000/

proved that the order of an MAF for two *rooted binary* phylogenetic trees minus 1 is equal to their rSPR distance. Since then, much work has been focused on studying the Maximum Agreement Forest problem on two rooted binary phylogenetic trees, which asks for an MAF for the two trees.

Biological researchers traditionally assumed that phylogenetic trees were bifurcating [14,15], which motivated most earlier work focused on the Maximum Agreement Forest problem for binary trees. However, more recent research in biology and phylogenetics has called a need to study the problem for general trees. For example, for many biological data sets in practice [16,17], the constructed phylogenetic trees always contain *polytomies* (alternatively called *multifurcations*). There are two different meanings to the polytomies in phylogenetic trees: (1) the polytomy refers to an event during which an ancestral species gave rise to more than two offspring species at the same time [18–21], i.e., a *hard* polytomy; and (2) the polytomy refers to ambiguous evolutionary relationships as a result of insufficient information, i.e., a *soft* polytomy. Note that the types of polytomies in the phylogenetic trees have a substantial impact on designing algorithms for comparing these trees. For example, a soft polytomy with three leaves (a, b, c) is not considered different from two resolved bifurcations of the same three leaves $((a, b), c)$, as the soft polytomy is ambiguous rather than conflicting, and the soft polytomy (a, b, c) can be *binary resolved* as $((a, b), c)$. On the other hand, if the polytomy (a, b, c) is hard, then (a, b, c) and $((a, b), c)$ are considered different as the hard polytomy is interpreted as simultaneous speciation. In this paper, we assume that all soft polytomies are correctly binary resolved, and all polytomies in the multifurcating phylogenetic trees we consider are hard. The rSPR distance between two rooted multifurcating phylogenetic trees also corresponds to their MAF.[1]

For the same collection of taxa, multiple (i.e., two or more) different phylogenetic trees may be constructed based on different data sets or different building methods. Thus, studying the Maximum Agreement Forest problem on multiple phylogenetic trees is biologically meaningful. For example, suppose that we have two phylogenetic trees that are constructed by two homologous regions of genomes from a collection of taxa. As mentioned above, studying the order of their MAF can indicate how reticulation events influence the evolutionary histories of the two homologous regions of the genomes. Note that these reticulation events that influenced the evolutionary histories of the two homologous regions may also influence the evolutionary histories of other homologous regions of the genomes. Thus, if we construct phylogenetic trees for each homologous region of the genomes, and study their MAF, then the order of their MAF can give a more comprehensive indication of the extent to which reticulation has influenced the evolutionary history of the collection of taxa. Moreover, consider an MAF F of order k for a set $\mathcal{C}$ of rooted phylogenetic trees. Since F is also an agreement forest (not necessarily an MAF) for any two trees T_i and T_j in $\mathcal{C}$, the rSPR distance between T_i and T_j would not be greater than $k - 1$. Thus, the order of an MAF for $\mathcal{C}$ provides an upper bound for the rSPR distance between any two trees in $\mathcal{C}$. Last but not least, constructing an MAF for multiple phylogenetic trees is a critical step in studying the reticulate networks with the minimum number of reticulation vertices for multiple phylogenetic trees [24], which is a hot topic in phylogenetics. The reason is that among all reticulate networks for the given multiple phylogenetic trees, the number of reticulation vertices in the reticulate network with the minimum number of reticulation vertices is equal to the order of an MAF for the given multiple phylogenetic trees minus one if the MAF is acyclic [25].

To summarize, it makes perfect sense to study the Maximum Agreement Forest problem on multiple rooted multifurcating phylogenetic trees. In this paper, we will focus on parameterized algorithms for the Maximum Agreement Forest problem on multiple rooted multifurcating phylogenetic trees. In the following, we first review previous related work on the Maximum Agreement Forest problem. Note that there are two kinds of phylogenetic trees, rooted or unrooted, depending on whether an ancestor-descendant relation is defined in the tree. Although in this paper we only study rooted phylogenetic trees, we also present previous related work on unrooted phylogenetic trees. In particular, Allen and Steel [10] proved that the TBR distance between two unrooted binary phylogenetic trees is equal to the order of their MAF minus 1.

In terms of the computational complexity of the problems, it has been proved that computing the order of an MAF is NP-hard and MAX SNP-hard for two unrooted binary phylogenetic trees [12], as well as for two rooted binary phylogenetic trees [13].

Approximation Algorithms. For the Maximum Agreement Forest problem on two rooted binary phylogenetic trees, Hein et al. [12] proposed an approximation algorithm of ratio 3. However, Rodrigues et al. [26] found a subtle error in [12], showed that the algorithm in [12] has ratio at least 4, and presented a new approximation algorithm which they claimed had ratio 3. Borchwich and Semple [13] corrected the definition of an MAF for the rSPR distance. Using this definition, Bonet et al. [27] provided a counterexample and showed that, with a slight modification, both the algorithms in [12] and [26] compute a 5-approximation of the rSPR distance between two rooted binary phylogenetic trees. A 3-approximation algorithm was achieved by Bordewich et al. [11], but the running time of the algorithm is increased to $O(n^5)$. A second 3-approximation algorithm was presented in [28] with running time $O(n^2)$. Whidden et al. [29] presented a third 3-approximation algorithm, which runs in linear-time. Shi et al. [30] improved the ratio to 2.5, but the algorithm has running time $O(n^2)$. Schalekamp et al. [31] presented a polynomial-time approximation algorithm of ratio 2 for the problem based on LP Duality. Chen et al. [32] gave an approximation algorithm of ratio 7/3 for the problem with running time $O(n^2)$, then improved the ratio to 2 with the same time complexity [33], which is the best known approximation algorithm for the Maximum Agreement Forest problem on two rooted binary trees. For the Maximum Agreement Forest problem on

[1] The relationship between MAF and the metric of rSPR distance on binary trees can be naturally extended to that on multifurcating trees [22,23].

Please cite this article in press as: F. Shi et al., A parameterized algorithm for the Maximum Agreement Forest problem on multiple rooted multifurcating trees, J. Comput. Syst. Sci. (2018), https://doi.org/10.1016/j.jcss.2018.03.002