A CLUSTER REDUCTION FOR COMPUTING THE SUBTREE DISTANCE BETWEEN PHYLOGENIES

SIMONE LINZ AND CHARLES SEMPLE

Abstract. Calculating the rooted subtree prune and regraft (rSPR) distance between two rooted binary phylogenetic trees is a frequently applied process in various areas of molecular evolution. However, computing this distance is an NP-hard problem and practical algorithms for computing it exactly are rare. In this paper, a divide-and-conquer approach to calculating the rSPR distance is established. This approach breaks the problem instance into a number of smaller and more tractable subproblems. Two reduction rules which were previously used to show that computing the rSPR distance is fixed-parameter tractable can easily be used to complement this new theoretical result, and so a significant positive impact on the running time of calculating this distance in practice is likely.

1. Introduction

Since Charles Darwin's first sketch of a phylogenetic (evolutionary) tree in 1837, evolutionary biologists have been interested in the reconstruction of phylogenetic trees which correctly represent the ancestral history of a set of taxa. In such a tree, each leaf typically represents a present-day species and each interior vertex corresponds to a hypothetical (extinct) ancestor, while the edges indicate the relationship between distinct taxa. Due to the incompleteness of the fossil record, researchers often rely upon sequence data of contemporary species—such as DNA or protein sequences—to reconstruct phylogenetic trees. Depending on the data set and the tree reconstruction method under consideration, the resulting trees, even for the same set of present-day species, often reveal inconsistencies. Consequently, it is a particularly natural and important task to quantify the dissimilarity between two phylogenies.

A prominent tool for this quantification is that of the graph-theoretic operation of rooted subtree prune and regraft (rSPR) (see [11]). Loosely speaking, this operation cuts (prunes) a subtree and reattaches (regrafts) it to another part of the tree. The dissimilarity of two phylogenies is quantified by the minimum number of rSPR operations that transforms one tree into the other. This minimum number is referred to as the rSPR distance and, as well as a measure of dissimilarity, it is often used in the analysis of non-tree-like evolution (for example, see [3, 6, 12, 13, 15]).

Date: January 21, 2009.
1991 *Mathematics Subject Classification.* 05C05; 92D15.
Key words and phrases. Phylogenetic tree, subtree prune and regraft, cluster reduction.
We thank the New Zealand Marsden Fund for their support.

Such evolution is prevalently caused by evolutionary processes that include horizontal gene transfer, hybridization, and recombination.

Computing the rSPR distance exactly is a computationally hard problem [7]. However, kernalizing the problem by repeatedly applying two particular reduction rules that preserve the distance—subtree and chain rules—results in the problem being fixed-parameter tractable [7]. Recently, a closely-related computational problem in evolutionary biology was analyzed using three reduction rules [8]. The first two rules are analogues of the subtree and chain rules for computing the rSPR distance, while the third rule is a divide-and-conquer rule that allows the problem to be partitioned into a number of smaller problems. By applying the associated algorithm to a grass data set, the performance of these three rules was analyzed (see [8, Table 2]). It is clear from the investigations in [8] that it was the divide-and-conquer rule that greatly aided the computational process. For example, for one instance, the running time was 19 seconds using all three rules, while the running time increased to about 37.5 hours using just the first two rules. In this paper, we consider an analogous divide-and-conquer rule for rSPR and show how it can be applied in conjunction with the subtree and chain rules for computing the rSPR distance of two phylogenies. (Intuitively, the divide-and-conquer rule means that the inevitable exhaustive search part of any fixed-parameter algorithm for finding the rSPR distance can be applied to smaller instances than otherwise would have been possible without it.) This divide-and-conquer rule for rSPR was considered in [7] but, because of a potential difficulty, it appeared that it could not be used in practice. The main purpose of this paper is to show that this difficulty can be successfully overcome.

The paper is organized as follows. The next section contains some additional background and preliminaries as well as a formal statement of the key result of this paper. For the reader familiar with agreement forests, this result intuitively characterizes the rSPR distance of two phylogenies $\mathcal{T}$ and $\mathcal{T}'$ in terms of the sum of the sizes of agreement forests for pairs of subtrees of $\mathcal{T}$ and $\mathcal{T}'$. These pairs of subtrees are the result of repeated applications of the divide-and-conquer rule for rSPR. The proof of this result is shown in Section 3, while Section 4 describes a practical algorithm for computing the rSPR distance between two phylogenies based on this summation. The fact that the resulting algorithms works is given at the end of Section 4. The last section contains some final remarks on the fixed-parameter tractability of calculating the rSPR distance and shows how one can make use of all three reduction rules to compute this distance. Throughout the paper, notation and terminology follows [14].

2. Key Result

We begin this section with some preliminaries.

Phylogenetic trees. A *rooted phylogenetic X-tree* $\mathcal{T}$ is a rooted tree with no degree-two vertices, except for the root which has degree at least two, and whose leaf set is X. Furthermore, $\mathcal{T}$ is *binary* if its root has degree two and all other interior vertices have degree three. For example, ignoring the pendant edges with

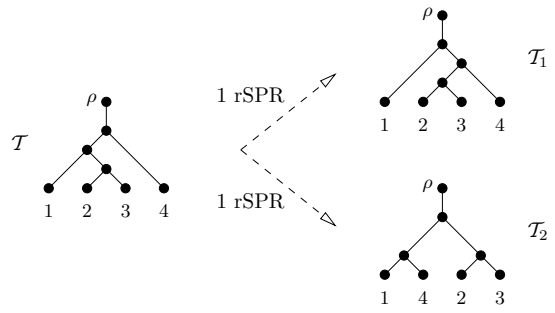


FIGURE 1. Each of $\mathcal{T}_1$ and $\mathcal{T}_2$ is obtained from $\mathcal{T}$ by a single rSPR operation.

end vertex ρ, each of the trees in Fig. 1 is a rooted binary phylogenetic tree. The leaf set X is the *label set* of $\mathcal{T}$ and is frequently denoted by $\mathcal{L}(\mathcal{T})$. A subset A of X is a *cluster* of $\mathcal{T}$ if there is an edge e, or equivalently a vertex v, that has precisely A as its set of descendant leaves. We denote this cluster by $C_{\mathcal{T}}(e)$.

For a rooted phylogenetic X-tree $\mathcal{T}$, several types of rooted subtrees will play an important role in this paper. Let X' be a subset of X. The *minimal rooted subtree* of $\mathcal{T}$ that connects all the leaves in X' is denoted by $\mathcal{T}(X')$. The *restriction of* $\mathcal{T}$ *to* X', denoted by $\mathcal{T}|X'$, is the rooted phylogenetic X'-tree obtained from $\mathcal{T}(X')$ by contracting all degree-two vertices apart from the root. Lastly, a rooted subtree of $\mathcal{T}$ is *pendant* if it can be detached from $\mathcal{T}$ by deleting a single edge.

rSPR and agreement forests. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. For the purposes of the upcoming definitions and indeed much of the paper, we view the roots of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ adjoined to the original root by a pendant edge. Furthermore, we regard ρ as part of the label sets of $\mathcal{T}$ and $\mathcal{T}'$, and so $\mathcal{L}(\mathcal{T}) = \mathcal{L}(\mathcal{T}') = X \cup \{\rho\}$.

Let $e = \{u, v\}$ be any edge of $\mathcal{T}$ not incident with ρ, where u is the vertex on the path from ρ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic X-tree obtained from $\mathcal{T}$ by deleting e and reattaching the resulting rooted subtree containing v via a new edge, f say, as follows. Subdivide an edge of the component that contains ρ with a new vertex u', join u' and v with f, and then contract u. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *rooted subtree prune and regraft* (rSPR) operation. As an example, in Fig. 1, each of $\mathcal{T}_1$ and $\mathcal{T}_2$ have been obtained from $\mathcal{T}$ by a single rSPR operation. The rSPR *distance* between two arbitrary rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is the minimum number of rSPR operations that transforms $\mathcal{T}$ into $\mathcal{T}'$. It is well known that one can always transform $\mathcal{T}$ into $\mathcal{T}'$ via a sequence of single rSPR operations, so this distance is well-defined. We denote this distance by $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two arbitrary rooted binary phylogenetic X-trees. An *agreement forest* $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \ldots, \mathcal{L}_k\}$ for $\mathcal{T}$ and $\mathcal{T}'$ is a partition of $X \cup \{\rho\}$ such that $\rho \in \mathcal{L}_\rho$ and the following properties are satisfied:

(i) for all $i \in \{\rho, 1, \ldots, k\}$, we have $\mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$, and

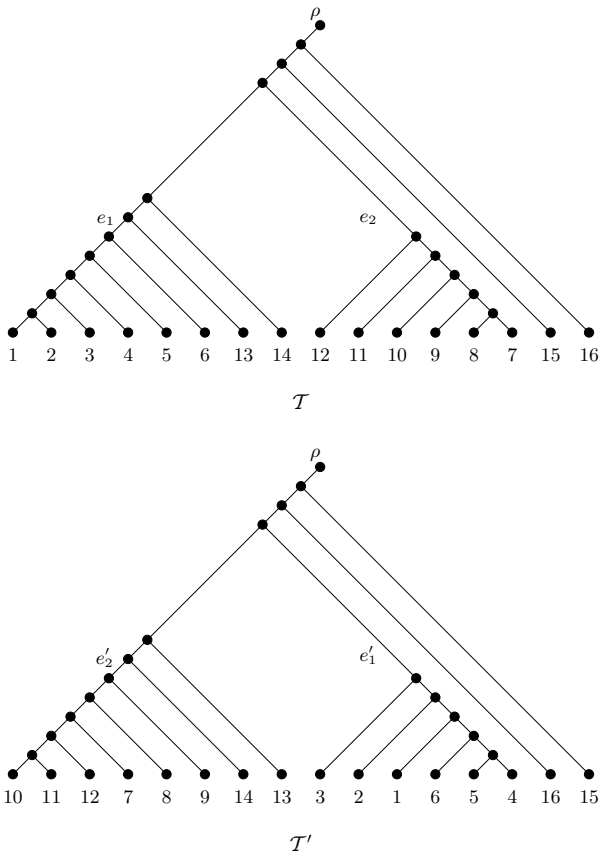


FIGURE 2. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ with their roots labeled ρ.

(ii) the trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, \ldots, k\}\}$ are vertex-disjoint subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

An agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is a *maximum-agreement forest* if, amongst all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, it has the smallest number of parts, in which case we denote this value of k by $m(\mathcal{T}, \mathcal{T}')$. To illustrate, consider the two binary phylogenetic trees shown in Fig. 2 where, for the moment, ignore the edge labels e_1, e_2, e'_1, and e'_2. An agreement forest for these two trees is

$$\big\{\{1,2,3\},\{4,5,6\},\{7,8,9\},\{10,11,12\},\{13,14\},\{15,16,\rho\}\big\}.$$

Bordewich and Semple [7] established the following characterization which expresses the rSPR distance in terms of agreement forests. This characterization is crucial to many of the computational results associated with computing the rSPR distance.

Theorem 2.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}').$$

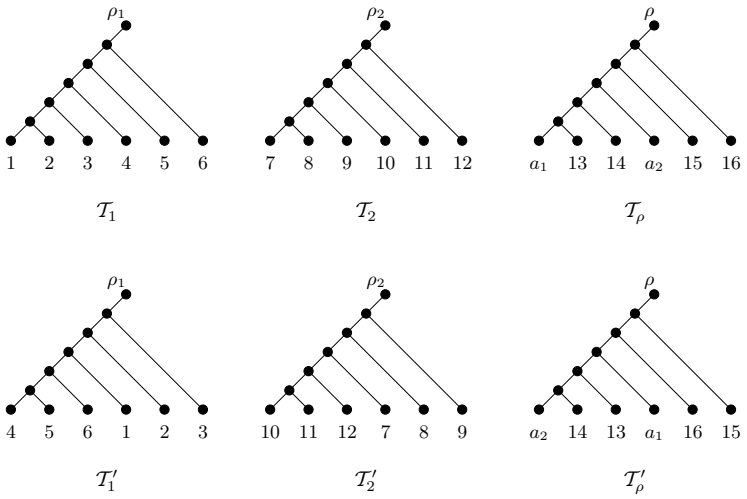


FIGURE 3. A cluster sequence $(\mathcal{T}_1, \mathcal{T}_1'), (\mathcal{T}_2, \mathcal{T}_2'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ for $\mathcal{T}$ and $\mathcal{T}'$ depicted in Fig. 2, where $A_1 = \{1, 2, \ldots, 6\}$ and $A_2 = \{7, 8, \ldots, 12\}$ are the common clusters whose corresponding minimal rooted subtrees have been replaced with a single vertex labeled a_1 and a_2, respectively.

We remark here that one may view the key result of this paper as a generalization of Theorem 2.1.

Cluster sequences. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. As previously, view the roots of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ adjoined to the original root by a pendant edge. We next describe algorithmically an associated sequence that is central to this paper. Setting $i = 1$, let A_i be a cluster of size at least two common to both $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{T}_i$ denote the rooted binary phylogenetic tree $\mathcal{T}|A_i$ (viewing the root of $\mathcal{T}_i$ as a vertex ρ_i adjoined to the original root by a pendant edge) and reset $\mathcal{T}$ to be the tree obtained from $\mathcal{T}$ by replacing $\mathcal{T}(A_i)$ with the new vertex a_i. Analogously, let $\mathcal{T}_i'$ denote the rooted binary phylogenetic tree $\mathcal{T}'|A_i$ (viewing the root of $\mathcal{T}_i'$ as a vertex ρ_i adjoined to the original root by a pendant edge) and reset $\mathcal{T}'$ to be the tree obtained from $\mathcal{T}'$ by replacing $\mathcal{T}'(A_i)$ with the vertex a_i. Increment i by 1 and repeat this process. Eventually, we obtain a sequence

$$(\mathcal{T}_1, \mathcal{T}_1'), (\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$$

of pairs of rooted binary phylogenetic trees, where $\mathcal{T}_\rho$ and $\mathcal{T}_\rho'$ denote the two trees after the replacement of $\mathcal{T}(A_t)$ and $\mathcal{T}'(A_t)$ with the vertex a_t. Observe that ρ is the root of $\mathcal{T}_\rho$ and $\mathcal{T}_\rho'$. We call this sequence a *cluster sequence* of $\mathcal{T}$ and $\mathcal{T}'$. An example of a cluster sequence for the two rooted binary phylogenetic trees shown in Fig. 2 is shown in Fig. 3.

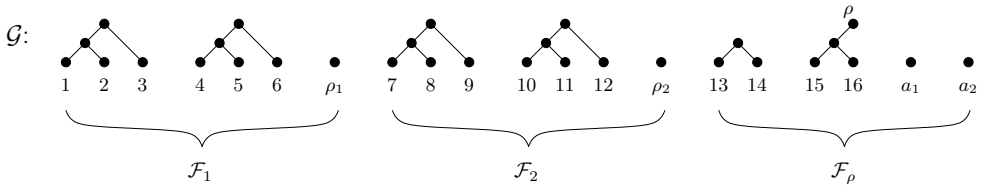


FIGURE 4. An agreement forest $\mathcal{G}$ for the cluster sequence shown in Fig. 3, where $\mathcal{F}_1$ is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$, and $\mathcal{F}_2$ and $\mathcal{F}_\rho$, respectively, are agreement forests for $\mathcal{T}_2$ and $\mathcal{T}_2'$, and $\mathcal{T}_\rho$ and $\mathcal{T}_\rho'$.

Extending the definition of an agreement forest to cluster sequences, an *agreement forest* for $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ is a partition $\mathcal{G}$ of

$$X \cup \{\rho\} \cup \{\rho_1, \rho_2, \ldots, \rho_t\} \cup \{a_1, a_2, \ldots, a_t\}$$

such that, for all $i \in \{1, \ldots, t, \rho\}$, there exists a subset $\mathcal{F}_i$ of $\mathcal{G}$ that is an agreement forest for $\mathcal{T}_i$ and $\mathcal{T}_i'$. The *weight* of $\mathcal{G}$, denoted $w(\mathcal{G})$, is defined to be

$$w(\mathcal{G}) = \sum_{i=1}^{t} |\mathcal{F}_i| + |\mathcal{F}_\rho| - |\{(\rho_i, a_i) : \{\rho_i\}, \{a_i\} \in \mathcal{G}\}| - t.$$

Note that $\sum_{i=1}^{t} |\mathcal{F}_i| + |\mathcal{F}_\rho|$ is simply $|\mathcal{G}|$ and that $|\{(\rho_i, a_i) : \{\rho_i\}, \{a_i\} \in \mathcal{G}\}|$ is the number of pairs (ρ_i, a_i) in which both ρ_i and a_i are singletons in $\mathcal{G}$. To illustrate, an agreement forest $\mathcal{G}$ (viewed as restricted subtrees) for the cluster sequence shown in Fig. 3 is shown in Fig. 4. The weight of this agreement forest is

$$(3 + 3 + 4) - 2 - 2 = 6.$$

For the reader familiar with maximum-agreement forests, it is interesting to note that $\mathcal{F}_\rho$ is not a maximum-agreement forest for $\mathcal{T}_\rho$ and $\mathcal{T}_\rho'$ since

$$\big\{\{a_1, 13\}, \{a_2, 14\}, \{15, 16, \rho\}\big\}$$

is an agreement forest for these two trees. However, it turns out that $\mathcal{G}$ is of minimum weight. Hence, to optimize the weighting, it is not sufficient to exclusively consider maximum-agreement forests for each pair of trees in a cluster sequence.

The point of the above weighting is because of the following theorem, the key result of the paper.

Theorem 2.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let*

$$(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$$

be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{F}$ be a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and let $\mathcal{G}$ be an agreement forest for this sequence of minimum weight. Then $|\mathcal{F}| = w(\mathcal{G})$. In particular,

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = w(\mathcal{G}) - 1.$$

The fact that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = w(\mathcal{G}) - 1$ in the statement of Theorem 2.2 is an immediate consequence of Theorem 2.1. The main part of the theorem will be establish in the next section. Furthermore, a divide-and-conquer algorithm for computing

the rSPR distance based upon this theorem will be described in Section 4. However, before doing this, we make three remarks.

First, a single application of the cluster reduction for computing the rSPR distance was considered in [7]. In the language of this paper, the following result was established.

Proposition 2.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and let $(\mathcal{T}_1, \mathcal{T}_1'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_1') + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}_\rho') \leq d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') + 1.$$

The potential difficulty of using Proposition 2.3 in practice is that either

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_1') + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}_\rho')$$

or

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_1') + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}_\rho') - 1$$

and both equalities are possible depending upon the pairs $(\mathcal{T}_1, \mathcal{T}_1')$ and $(\mathcal{T}_\rho, \mathcal{T}_\rho')$. Understanding and recognizing which of these equalities hold is the basis for this paper.

Second, as Theorem 2.2 is stated, it appears that we have to work globally to find an agreement forest for

$$(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$$

of minimum weight. Perhaps surprisingly given Proposition 2.3, this is not the case. It turns out that we can work locally using a bottom-up strategy starting at the leaves of both trees and working towards their respective roots, and only needing to find agreement forests for the smaller trees $\mathcal{T}_i$ and $\mathcal{T}_i'$ for all $i \in \{1, 2, \ldots, t, \rho\}$. Indeed, it turns out that it is sufficient to only consider maximum-agreement forests of subtrees of $\mathcal{T}_i$ and $\mathcal{T}_i'$ (see Section 4).

Third, a closely-related, and also computationally hard, problem to that of computing the rSPR distance is computing the so-called hybridization number $h(\mathcal{T}, \mathcal{T}')$ of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$. The value $h(\mathcal{T}, \mathcal{T}')$ is the minimum number of hybridization events that is required to simultaneously explain the evolutionary scenarios of $\mathcal{T}$ and $\mathcal{T}'$. For a formal definition, see [4]. The underlying reason for the closeness of the two computational problems is that $h(\mathcal{T}, \mathcal{T}')$ can also be characterized in terms of agreement forests. The only difference in the characterizations is that, for $h(\mathcal{T}, \mathcal{T}')$, we require the forest to have the additional property of being "acyclic". The reason for this property is so that the biologically well-motivated constraint that species cannot inherit genetic material from their own descendants is satisfied. However, it is this additional property that allows for a cleaner version of Proposition 2.3 for the hybridization number of $\mathcal{T}$ and $\mathcal{T}'$. In particular, in the language of this paper, Baroni *et al.* [5] established the following proposition.

Proposition 2.4. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and let $(\mathcal{T}_1, \mathcal{T}_1'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. Then*

$$h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}_1, \mathcal{T}_1') + h(\mathcal{T}_\rho, \mathcal{T}_\rho').$$

An immediate consequence of Proposition 2.4 is that if $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ is a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$, then

$$h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}_1, \mathcal{T}_1') + h(\mathcal{T}_2, \mathcal{T}_2') + \cdots + h(\mathcal{T}_\rho, \mathcal{T}_\rho').$$

This equality is the third rule mentioned in the introduction and it is this one that greatly aids the computational process in the study in [8], and because of this motivates the work done in this paper. For the reader familiar with acyclic-agreement forests, all maximum-acyclic-agreement forests have the property that ρ is never a singleton and it is this property that gives the equality in Proposition 2.4. This property on ρ does not necessarily hold in the context of maximum-agreement forests, and thus the reason for the inequalities in Proposition 2.3.

3. Proof of Theorem 2.2

This section contains the proof of Theorem 2.2. For convenience, we restate the theorem.

Theorem 2.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let*

$$(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$$

be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{F}$ be a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and let $\mathcal{G}$ be an agreement forest for this sequence of minimum weight. Then $|\mathcal{F}| = w(\mathcal{G})$. In particular,

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = w(\mathcal{G}) - 1.$$

Proof. We prove that $|\mathcal{F}| = w(\mathcal{G})$ by first showing that $|\mathcal{F}| \leq w(\mathcal{G})$ and then showing that $|\mathcal{F}| \geq w(\mathcal{G})$. Both parts are established by induction. For this purpose, let $\mathcal{S}$ be the rooted binary phylogenetic tree obtained from $\mathcal{T}$ by replacing the minimal rooted subtree $\mathcal{T}(\mathcal{L}(\mathcal{T}_1))$ with a single vertex labeled a_1 and, similarly, let $\mathcal{S}'$ be such a tree obtained from $\mathcal{T}'$ by replacing $\mathcal{T}'(\mathcal{L}(\mathcal{T}_1))$ with a single vertex labeled a_1. Note that $\mathcal{L}(\mathcal{T}_1) = \mathcal{L}(\mathcal{T}_1')$, and $(\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ is a cluster sequence for $\mathcal{S}$ and $\mathcal{S}'$ of length t.

We first show that $|\mathcal{F}| \leq w(\mathcal{G})$ by proving a slightly stronger result. Let $\mathcal{G}_T$ be an arbitrary agreement forest for the cluster sequence $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ of $\mathcal{T}$ and $\mathcal{T}'$. We will show that there exists an agreement forest $\mathcal{F}_T$ for $\mathcal{T}$ and $\mathcal{T}'$ such that $|\mathcal{F}_T| \leq w(\mathcal{G}_T)$ and, for each $x \in \mathcal{L}(\mathcal{T}) - \{\rho\}$ with $\{x\} \in \mathcal{G}_T$, we have $\{x\} \in \mathcal{F}_T$. For simplicity, we will refer to this last property in the following way: *$\mathcal{F}_T$ has the desired singleton property relative to $\mathcal{G}_T$*. Observe that, by choosing $\mathcal{G}_T$ to be $\mathcal{G}$ and noting that $|\mathcal{F}| \leq |\mathcal{F}_T|$, this stronger result establishes $|\mathcal{F}| \leq w(\mathcal{G})$.

The proof is by induction on t. If $t = 0$, then the definition of a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ coincides with the definition of an agreement forest of minimum weight for the cluster sequence $(\mathcal{T}, \mathcal{T}')$ of $\mathcal{T}$ and $\mathcal{T}'$. Therefore, we can choose $\mathcal{F}_T$ to be $\mathcal{G}_T$ and the result follows. Now suppose that the stronger result holds for all cluster sequences of two rooted binary phylogenetic trees with length at most t. As $\mathcal{G}_T$ is an agreement forest for

$$(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho'),$$

it is easily checked that

$$\mathcal{G}_S = \{\mathcal{L}_i \in \mathcal{G}_T : \mathcal{L}_i \cap \mathcal{L}(\mathcal{T}_1) = \emptyset\}$$

is an agreement forest for $(\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ and

$$\mathcal{F}_{T_1} = \{\mathcal{L}_i \in \mathcal{G}_T : \mathcal{L}_i \cap \mathcal{L}(\mathcal{T}_1) \neq \emptyset\}$$

is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$. Observe that $|\mathcal{G}_T| = |\mathcal{G}_S| + |\mathcal{F}_{T_1}|$. Since $(\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ is a cluster sequence of $\mathcal{S}$ and $\mathcal{S}'$ with length t, it follows by the induction assumption that there exists an agreement forest $\mathcal{F}_S$ for $\mathcal{S}$ and $\mathcal{S}'$ such that $|\mathcal{F}_S| \leq w(\mathcal{G}_S)$ and $\mathcal{F}_S$ has the desired singleton property relative to $\mathcal{G}_S$. Let $\mathcal{L}_{\rho_1}$ denote the label set of $\mathcal{F}_{T_1}$ containing ρ_1, and let $\mathcal{L}_{a_1}$ denote the label set of $\mathcal{F}_S$ containing a_1. Let

$$P = \{(\rho_i, a_i) : \{\rho_i\}, \{a_i\} \in \mathcal{G}_T \text{ and } i \in \{1, \ldots, t\}\},$$

and let

$$P_S = \{(\rho_i, a_i) : \{\rho_i\}, \{a_i\} \in \mathcal{G}_S \text{ and } i \in \{2, \ldots, t\}\}.$$

Noting that $|P| \in \{|P_S|, |P_S| + 1\}$, there are two cases to consider: (i) $|P| = |P_S|$ and (ii) $|P| = |P_S| + 1$.

If (i) holds, then $(\rho_1, a_1) \notin P$. Let

$$\mathcal{F}_T = (\mathcal{F}_S \cup \mathcal{F}_{T_1} - \{\mathcal{L}_{a_1}, \mathcal{L}_{\rho_1}\}) \cup \{(\mathcal{L}_{a_1} - \{a_1\}) \cup (\mathcal{L}_{\rho_1} - \{\rho_1\})\},$$

and note that $\{(\mathcal{L}_{a_1} - \{a_1\}) \cup (\mathcal{L}_{\rho_1} - \{\rho_1\})\}$ may consist of the empty set, in which case we set $\mathcal{F}_T = (\mathcal{F}_S \cup \mathcal{F}_{T_1}) - \{\mathcal{L}_{a_1}, \mathcal{L}_{\rho_1}\}$. Since $\mathcal{F}_S$ and $\mathcal{F}_{T_1}$ are agreement forests for $\mathcal{S}$ and $\mathcal{S}'$, and for $\mathcal{T}_1$ and $\mathcal{T}_1'$, respectively, $\mathcal{F}_T$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, by construction, $\mathcal{F}_T$ has the desired singleton property relative to $\mathcal{G}_T$. Now, since $|\mathcal{F}_S| \leq w(\mathcal{G}_S)$,

$$\begin{aligned}
|\mathcal{F}_T| &\leq |\mathcal{F}_S| - 1 + |\mathcal{F}_{T_1}| - 1 + 1 \\
&\leq w(\mathcal{G}_S) + |\mathcal{F}_{T_1}| - 1 \\
&= (|\mathcal{G}_T| - |\mathcal{F}_{T_1}| - |P_S| - (t-1)) + |\mathcal{F}_{T_1}| - 1 \\
&= |\mathcal{G}_T| - |P| - t = w(\mathcal{G}_T).
\end{aligned}$$

If (ii) holds, then $(\rho_1, a_1) \in P$. Therefore, $\{\rho_1\}, \{a_1\} \in \mathcal{G}_T$ and $\{\rho_1\} \in \mathcal{F}_{T_1}$. Since $\mathcal{F}_S$ has the desired singleton property relative to $\mathcal{G}_S$ and $\{a_1\} \in \mathcal{G}_S$, we also have $\{a_1\} \in \mathcal{F}_S$. In this case, let

$$\mathcal{F}_T = (\mathcal{F}_S - \{\mathcal{L}_{a_1}\}) \cup (\mathcal{F}_{T_1} - \{\mathcal{L}_{\rho_1}\})$$

Since $\mathcal{F}_S$ and $\mathcal{F}_{T_1}$ are agreement forests for $\mathcal{S}$ and $\mathcal{S}'$, and for $\mathcal{T}_1$ and $\mathcal{T}_1'$, respectively, and $\mathcal{L}_{a_1} = \{a_1\}$ and $\mathcal{L}_{\rho_1} = \{\rho_1\}$, it follows that $\mathcal{F}_T$ is an agreement forest for $\mathcal{T}$ and

$\mathcal{T}'$. Furthermore, by construction, $\mathcal{F}_T$ has the desired singleton property relative to $\mathcal{G}_T$. Now, as $|\mathcal{F}_S| \leq w(\mathcal{G}_S)$,

$$\begin{aligned}|\mathcal{F}_T| &= |\mathcal{F}_S| - 1 + |\mathcal{F}_{T_1}| - 1\\ &\leq w(\mathcal{G}_S) + |\mathcal{F}_{T_1}| - 2\\ &= (|\mathcal{G}_T| - |\mathcal{F}_{T_1}| - |P_S| - (t-1)) + |\mathcal{F}_{T_1}| - 2\\ &= |\mathcal{G}_T| - |P| - t = w(\mathcal{G}_T).\end{aligned}$$

Thus $|\mathcal{F}_T| \leq w(\mathcal{G}_T)$ as required.

For the other direction of the theorem, we again prove a slightly stronger result. In particular, let $\mathcal{F}_T$ be an arbitrary agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. We show that there exists an agreement forest $\mathcal{G}_T$ for the cluster sequence $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ of $\mathcal{T}$ and $\mathcal{T}'$ such that $|\mathcal{F}_T| \geq w(\mathcal{G}_T)$ and, for all $\mathcal{L}_j \in \mathcal{G}_T$, there exists a $\mathcal{L}_i \in \mathcal{F}_T$ with $\mathcal{L}_j \cap (X \cup \{\rho\})$ being a subset of $\mathcal{L}_i$. For simplicity, we will refer to this last property in the following way: $\mathcal{G}_T$ *has the desired subset property relative to* $\mathcal{F}_T$. Observe that, by choosing $\mathcal{F}_T$ to be $\mathcal{F}$ and noting that $w(\mathcal{G}_T) \geq w(\mathcal{G})$, this stronger result establishes the other direction.

The proof of this direction is by induction on t. If $t = 0$, then, as the definition of a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ coincides with the definition of an agreement forest of minimum weight for the cluster sequence $(\mathcal{T}, \mathcal{T}')$ of $\mathcal{T}$ and $\mathcal{T}'$, we can choose $\mathcal{G}_T$ to be $\mathcal{F}_T$ and this immediately establishes the result. Now suppose that the stronger result holds for all cluster sequences of two rooted binary phylogenetic trees with length at most t. We consider two cases, where A_1 denotes $\mathcal{L}(\mathcal{T}_1) - \{\rho_1\}$:

(i) There exists a label set $\mathcal{L}_m$ in $\mathcal{F}_T$ such that $\mathcal{L}_m \cap A_1 \neq \emptyset$ and $\mathcal{L}_m \cap ((X - A_1) \cup \{\rho\}) \neq \emptyset$.
(ii) For all label sets $\mathcal{L}_i \in \mathcal{F}_T$, either $\mathcal{L}_i \subseteq A_1$ or $\mathcal{L}_i \subseteq ((X - A_1) \cup \{\rho\})$.

Assume first that (i) holds and note that $\mathcal{L}_m$ is the unique label set in $\mathcal{F}_T$ with the described properties; otherwise $\mathcal{F}_T$ is not an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{L}_{m'} = \mathcal{L}_m \cap ((X - A_1) \cup \{\rho\})$, and let $\mathcal{L}_{m''} = \mathcal{L}_m \cap A_1$. Then, since $\mathcal{F}_T$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$,

$$\mathcal{F}_S = \{\mathcal{L}_i \in \mathcal{F}_T : \mathcal{L}_i \subseteq ((X - A_1) \cup \{\rho\})\} \cup \{\mathcal{L}_{m'} \cup \{a_1\}\}$$

is an agreement forest for $\mathcal{S}$ and $\mathcal{S}'$, and

$$\mathcal{F}_{T_1} = \{\mathcal{L}_i \in \mathcal{F}_T : \mathcal{L}_i \subseteq A_1\} \cup \{\mathcal{L}_{m''} \cup \{\rho_1\}\}$$

is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$. Since $(\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ is a cluster sequence of $\mathcal{S}$ and $\mathcal{S}'$ with length t, it follows by the induction assumption that there exists an agreement forest $\mathcal{G}_S$ for the cluster sequence $(\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ of $\mathcal{S}$ and $\mathcal{S}'$ such that $|\mathcal{F}_S| \geq w(\mathcal{G}_S)$ and, for all $\mathcal{L}_j \in \mathcal{G}_S$, there exists a $\mathcal{L}_i \in \mathcal{F}_S$ with $\mathcal{L}_j \cap ((X - A_1) \cup \{\rho, a_1\})$ being a subset of $\mathcal{L}_i$. As $\mathcal{F}_{T_1}$ is an agreement forest for $(\mathcal{T}_1, \mathcal{T}_1')$, it follows that

$$\mathcal{G}_T = \mathcal{G}_S \cup \mathcal{F}_{T_1}$$

is an agreement forest for $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$. By construction, $\mathcal{G}_T$ has the desired subset property relative to $\mathcal{F}_T$. Furthermore, as $\{\rho_1\} \notin \mathcal{F}_{T_1}$ and $w(\mathcal{G}_S) \leq$

$|\mathcal{F}_S|$,

$$\begin{aligned} w(\mathcal{G}_T) &= w(\mathcal{G}_S) + |\mathcal{F}_{T_1}| - 1 \\ &\leq |\mathcal{F}_S| + |\mathcal{F}_{T_1}| - 1 = |\mathcal{F}_T|. \end{aligned}$$

For (ii), choose

$$\mathcal{F}_S = \{\mathcal{L}_i \in \mathcal{F} : \mathcal{L}_i \subseteq ((X - A_1) \cup \{\rho\})\} \cup \{\{a_1\}\}$$

and

$$\mathcal{F}_{T_1} = \{\mathcal{L}_i \in \mathcal{F} : \mathcal{L}_i \subseteq A_1\} \cup \{\{\rho_1\}\}.$$

It is clear that $\mathcal{F}_S$ is an agreement forest for $\mathcal{S}$ and $\mathcal{S}'$, and $\mathcal{F}_{T_1}$ is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$. By the induction assumption, there exists an agreement forest $\mathcal{G}_S$ for the cluster sequence $(\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ of $\mathcal{S}$ and $\mathcal{S}'$ such that $|\mathcal{F}_S| \geq w(\mathcal{G}_S)$ and, for all $\mathcal{L}_j \in \mathcal{G}_S$, there exists a $\mathcal{L}_i \in \mathcal{F}_S$ with $\mathcal{L}_j \cap ((X - A_1) \cup \{\rho, a_1\})$ being a subset of $\mathcal{L}_i$. In particular, as $\{a_1\} \in \mathcal{F}_S$, we have $\{a_1\} \in \mathcal{G}_S$. Set

$$\mathcal{G}_T = \mathcal{G}_S \cup \mathcal{F}_{T_1}$$

and observe that $\mathcal{G}_T$ is an agreement forest for $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$. By construction, $\mathcal{G}_T$ has the desired subset property relative to $\mathcal{F}_T$. Moreover, as $\{\rho_1\}, \{a_1\} \in \mathcal{G}_T$, we have $w(\mathcal{G}_T) = w(\mathcal{G}_S) + |\mathcal{F}_{T_1}| - 2$, and so

$$\begin{aligned} w(\mathcal{G}_T) &= w(\mathcal{G}_S) + |\mathcal{F}_{T_1}| - 2 \\ &\leq |\mathcal{F}_S| + |\mathcal{F}_{T_1}| - 2 = |\mathcal{F}_T|. \end{aligned}$$

This completes the proof of the theorem. □

4. Computing the Minimum Weight

In this section, we present an algorithm for computing the minimum weight of an agreement forest for a cluster sequence $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ of $\mathcal{T}$ and $\mathcal{T}'$. Potentially, to find such a minimum-weight-agreement forest one may have to consider all agreement forests for each $(\mathcal{T}_i, \mathcal{T}_i')$ and compare over all such forests to minimize the weighting. However, in this section, we show that we can do significantly better than this by applying a 'bottom-up' approach. The fact that this approach works is shown in the second part of the section.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and let

$$(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$$

be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. For each $i \in \{1, 2, \ldots, t\}$, we denote by e_i the edge of $\mathcal{T}$ whose end vertex has been replaced by a_i at iteration i. Similarly, for each $i \in \{1, 2, \ldots, t\}$, we denote by e_i' the edge of $\mathcal{T}'$ whose end vertex has been replaced by a_i at iteration i. Note that e_i and e_i' are well-defined. We refer to e_i and e_i' as *reduction edges* of the sequence. Now let $\mathcal{J}$ denote the rooted tree with root ρ whose vertex set is $\{\rho, e_1, \ldots, e_t\}$ and where two vertices are joined by an edge precisely if the (unique) path connecting them in $\mathcal{T}$ does not traverse any other element in $\{\rho, e_1, \ldots, e_t\}$. Similarly, let $\mathcal{J}'$ denote the rooted tree with root ρ whose vertex set is $\{\rho, e_1', \ldots, e_t'\}$ and where two vertices are joined by an edge precisely if the (unique) path connecting them in $\mathcal{T}'$ does not traverse any other element in

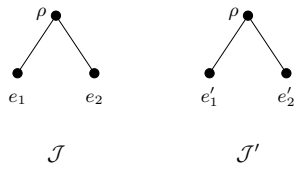


FIGURE 5. The two rooted trees $\mathcal{J}$ and $\mathcal{J}'$ for the cluster sequence $(\mathcal{T}_1, \mathcal{T}_1'), (\mathcal{T}_2, \mathcal{T}_2'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ shown in Fig. 3.

$\{\rho, e_1', \ldots, e_t'\}$. To illustrate, consider the cluster sequence $(\mathcal{T}_1, \mathcal{T}_1'), (\mathcal{T}_2, \mathcal{T}_2'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ shown in Fig. 3 of the two trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Fig. 2. The reduction edges e_1, e_2, e_1', and e_2' of this sequence are shown in Fig. 2. Furthermore, the rooted trees $\mathcal{J}$ and $\mathcal{J}'$ are shown in Fig. 5.

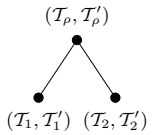
Lemma 4.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and let $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. Then*

(i) *$C_{\mathcal{T}}(e_i) = C_{\mathcal{T}'}(e_i')$ for all $i \in \{1, 2, \ldots, t\}$, and*
(ii) *$\mathcal{J}$ is isomorphic to $\mathcal{J}'$.*

Proof. Let Σ denote the sequence $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$. We prove both parts simultaneously by induction on the length of Σ. If Σ has length 1, that is $t = 0$, then (i) trivially holds and, as $\mathcal{J}$ and $\mathcal{J}'$ each consist of a single vertex ρ, (ii) also holds. Now suppose that (i) and (ii) holds for all cluster sequences of two rooted binary phylogenetic trees with length at most t. Let $\mathcal{S}$ be the rooted binary phylogenetic tree obtained from $\mathcal{T}$ by replacing the minimal rooted subtree $\mathcal{T}(\mathcal{L}(\mathcal{T}_1))$ with a single vertex labeled a_1 and, similarly, let $\mathcal{S}'$ be such a tree obtained from $\mathcal{T}'$ by replacing $\mathcal{T}'(\mathcal{L}(\mathcal{T}_1))$ with a single vertex labeled a_1. Since $\Sigma_S = (\mathcal{T}_2, \mathcal{T}_2'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ is a cluster sequence for $\mathcal{S}$ and $\mathcal{S}'$ of length t, it follows by the induction assumption, that $C_{\mathcal{S}}(e_i) = C_{\mathcal{S}'}(e_i')$ for all $i \in \{2, \ldots, t\}$ and that $\mathcal{J}_S$ and $\mathcal{J}_S'$ are isomorphic, where $\mathcal{J}_S$ and $\mathcal{J}_S'$ are the analogues of $\mathcal{J}$ and $\mathcal{J}'$ for $\mathcal{S}$ and $\mathcal{S}'$, respectively.

By the construction of $\mathcal{S}$ and $\mathcal{S}'$ from $\mathcal{T}$ and $\mathcal{T}'$, it is easily seen that $C_{\mathcal{T}}(e_i) = C_{\mathcal{T}'}(e_i')$ for all $i \in \{1, 2, \ldots, t\}$, so (i) holds. Furthermore, observe that $\mathcal{J}$ is obtained from $\mathcal{J}_S$ by adjoining e_1 to the vertex corresponding to the minimal cluster in $\{C_{\mathcal{T}}(e_i) : i \in \{2, \ldots, t\}\} \cup \{X\}$ that contains $C_{\mathcal{T}}(e_1)$. Similarly, $\mathcal{J}'$ is obtained from $\mathcal{J}_S'$ by adjoining e_1' to the vertex corresponding to the minimal cluster in $\{C_{\mathcal{T}}(e_i') : i \in \{2, \ldots, t\}\} \cup \{X\}$ that contains $C_{\mathcal{T}'}(e_1')$. It now follows by these observations and (i) that (ii) holds. This completes the proof of the lemma. □

Following Lemma 4.1(ii), let $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. We define the *cluster hierarchy* $\mathcal{H}$ of this sequence to be the rooted tree obtained from $\mathcal{J}$ (or equivalently $\mathcal{J}'$) by relabeling the vertex ρ with $(\mathcal{T}_\rho, \mathcal{T}_\rho')$ and, for all $i \in \{1, 2, \ldots, t\}$, relabeling e_i with $(\mathcal{T}_i, \mathcal{T}_i')$. We now present our algorithm.



$\mathcal{H}$

FIGURE 6. The cluster hierarchy $\mathcal{H}$ for the cluster sequence depicted in Fig. 3.

Algorithm: MINIMUM-WEIGHT FOREST
Input: A cluster sequence $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$.
Output: The minimum weight of an agreement forest for this sequence.

Step 1 Set $j = 1$, set $\mathcal{G}_j = \emptyset$, and set $\mathcal{H}_j$ to be the cluster hierarchy of

$$(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho).$$

Step 2 Select a leaf $(\mathcal{T}_i, \mathcal{T}'_i)$ of $\mathcal{H}_j$ and find an agreement forest $\mathcal{F}_i$ for $\mathcal{T}_i$ and $\mathcal{T}'_i$ that minimizes

$$|\mathcal{F}_i| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j \text{ and } \{a_{i'}\} \in \mathcal{F}_i\}|$$

and, provided $i \neq \rho$, amongst all such forests, choose one in which ρ_i is a singleton if possible. Note that $\rho_{i'} \in \{\rho_1, \rho_2, \ldots, \rho_t\}$ and $a_{i'} \in \{a_1, a_2, \ldots, a_t\}$.
Step 3 Set $\mathcal{G}_{j+1} = \mathcal{G}_j \cup \mathcal{F}_i$.
Step 4 If $j = t + 1$, STOP and return

$$|\mathcal{G}_{t+1}| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\}, \{a_{i'}\} \in \mathcal{G}_{t+1}\}| - t.$$

Step 5 Otherwise, increment j by 1, and set $\mathcal{H}_j$ to be the hierarchy obtained from $\mathcal{H}_{j-1}$ by deleting $(\mathcal{T}_i, \mathcal{T}'_i)$ and its incident edge. Return to Step 2.

To illustrate the algorithm, again consider the cluster sequence $(\mathcal{T}_1, \mathcal{T}'_1), (\mathcal{T}_2, \mathcal{T}'_2), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ shown in Fig. 3 of the two trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Fig. 2. The cluster hierarchy of this sequence is the rooted tree shown in Fig. 6. In the first iteration of MINIMUM-WEIGHT FOREST applied to $(\mathcal{T}_1, \mathcal{T}'_1), (\mathcal{T}_2, \mathcal{T}'_2), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$, either $(\mathcal{T}_1, \mathcal{T}'_1)$ or $(\mathcal{T}_2, \mathcal{T}'_2)$ is selected at Step 2. Say $(\mathcal{T}_1, \mathcal{T}'_1)$ is selected. The algorithm then finds an appropriate agreement forest of $\mathcal{T}_1$ and $\mathcal{T}'_1$. Such a forest $\mathcal{F}_1$ is shown in Fig. 4. The set $\mathcal{G}_1$ is initially empty, so $\mathcal{G}_2$ is set to be $\mathcal{F}_1$ at Step 3. In the second iteration, $(\mathcal{T}_2, \mathcal{T}'_2)$ is selected at Step 2 and an appropriate agreement forest $\mathcal{F}_2$ of $\mathcal{T}_2$ and $\mathcal{T}'_2$ is shown in Fig. 4. At Step 3, $\mathcal{G}_3$ is set to be $\mathcal{F}_1 \cup \mathcal{F}_2$. In the third and final iteration, $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ is considered. Using $\mathcal{F}_\rho$ shown in Fig. 4, a possible agreement forest of minimum weight for $\mathcal{T}$ and $\mathcal{T}'$ is

$$\{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}, \{10, 11, 12\}, \{13, 14\}, \{15, 16, \rho\}\}.$$

Thus, MINIMUM-WEIGHT FOREST returns 6, the weight of this forest.

Before establishing the correctness of MINIMUM-WEIGHT FOREST, we make three remarks. Firstly, observe that MINIMUM-WEIGHT FOREST is well-defined and $\mathcal{G}_{t+1}$ is an agreement forest for the initial sequence. Secondly, in practice,

finding such an appropriate forest in Step 2 comes down to finding a maximum-agreement forest. In particular, Step 2 can be restated as follows.

Step 2′ Select a leaf $(\mathcal{T}_i, \mathcal{T}_i')$ of $\mathcal{H}_j$ and find a maximum-agreement forest $\mathcal{F}_i'$ for $\mathcal{T}_i$ and $\mathcal{T}_i'$ restricted to $\mathcal{L}(\mathcal{T}_i) - \{a_{i'} : \{\rho_{i'}\} \in \mathcal{G}_j\}$ with the property that, amongst all such forests, choose one in which ρ_i is a singleton if possible. Set $\mathcal{F}_i$ to be $\mathcal{F}_i' \cup \{\{a_{i'}\} : \{\rho_{i'}\} \in \mathcal{G}_j\}$.

To see why Step 2′ is equivalent to Step 2 in Minimum-Weight Forest, let $\mathcal{F}_i''$ denote $\mathcal{F}_i' \cup \{\{a_{i'}\} : \{\rho_{i'}\} \in \mathcal{G}_j\}$ and recall that, in Step 2, $\mathcal{F}_i$ is an agreement forest for $\mathcal{T}_i$ and $\mathcal{T}_i'$ that minimizes

$$|\mathcal{F}_i| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j \text{ and } \{a_{i'}\} \in \mathcal{F}_i\}|$$

and, amongst all such forests, ρ_i is a singleton if possible. As $\mathcal{F}_i''$ is an agreement forest for $\mathcal{T}_i$ and $\mathcal{T}_i'$, it follows by the minimality of $\mathcal{F}_i$ that

$$|\mathcal{F}_i''| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j, \{a_{i'}\} \in \mathcal{F}_i''\}| \geq |\mathcal{F}_i| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j, \{a_{i'}\} \in \mathcal{F}_i\}|. \tag{1}$$

But the set obtained from $\mathcal{F}_i$ by removing any labels in $\{a_{i'} : \{\rho_{i'}\} \in \mathcal{G}_j\}$ and any resulting empty sets is an agreement forest for $\mathcal{T}_i$ and $\mathcal{T}_i'$ restricted to $\mathcal{L}(\mathcal{T}_i) - \{a_{i'} : \{\rho_{i'}\} \in \mathcal{G}_j\}$. Therefore, by the minimality of $\mathcal{F}_i'$,

$$\begin{aligned} |\mathcal{F}_i'| &\leq |\mathcal{F}_i| - |\{A' : A' \in \mathcal{F}_i,\ A' \subseteq \{a_{i'} : \{\rho_{i'}\} \in \mathcal{G}_j\}\}| \\ &\leq |\mathcal{F}_i| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j \text{ and } \{a_{i'}\} \in \mathcal{F}_i\}|. \end{aligned} \tag{2}$$

Since

$$|\mathcal{F}_i'| = |\mathcal{F}_i''| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j,\ \{a_{i'}\} \in \mathcal{F}_i''\}|,$$

it now follows that

$$|\mathcal{F}_i''| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j, \{a_{i'}\} \in \mathcal{F}_i''\}| = |\mathcal{F}_i| - |\{(\rho_{i'}, a_{i'}) : \{\rho_{i'}\} \in \mathcal{G}_j, \{a_{i'}\} \in \mathcal{F}_i\}|.$$

Furthermore, by the last equality, we now have equality in (1) and (2). Using these equalities, it is easily checked that ρ_i is a singleton in $\mathcal{F}_i' \cup \{\{a_{i'}\} : \{\rho_{i'}\} \in \mathcal{G}_j\}$ if and only if it is a singleton in $\mathcal{F}_i$. Thus Step 2′ is equivalent to Step 2.

Lastly, for computational reasons, it is useful to choose a cluster sequence that is as long as possible; thus breaking the problem instance into as many smaller subproblems as possible. Hence, in selecting clusters for this sequence, the best strategy is to choose minimal common clusters of size at least 2.

Theorem 4.2. *Let $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$ be a cluster sequence of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$. Then the* Minimum-Weight Forest *algorithm applied to this sequence returns the minimum weight of an agreement forest for it.*

Proof. Let Σ denote the sequence $(\mathcal{T}_1, \mathcal{T}_1'), \ldots, (\mathcal{T}_t, \mathcal{T}_t'), (\mathcal{T}_\rho, \mathcal{T}_\rho')$, and let $(\mathcal{T}_i, \mathcal{T}_i')$ be a pair in Σ. At some iteration of Minimum-Weight Forest, $(\mathcal{T}_i, \mathcal{T}_i')$ is selected at Step 2. Let e and e' be the reduction edges of $\mathcal{T}$ and $\mathcal{T}'$, respectively, corresponding to $\mathcal{T}_i$ and $\mathcal{T}_i'$ if $i \neq \rho$. Thus $e \in \{e_1, e_2, \ldots, e_t\}$ and $e' \in \{e_1', e_2', \ldots, e_t'\}$. Recalling

that $C_{\mathcal{T}}(e) = C_{\mathcal{T}'}(e')$ by Lemma 4.1(i), let $\mathcal{S}$ denote $\mathcal{T}|C_{\mathcal{T}}(e)$, and let $\mathcal{S}'$ denote $\mathcal{T}'|C_{\mathcal{T}}(e)$. If $i = \rho$, let $\mathcal{S}$ denote $\mathcal{T}$ and $\mathcal{S}'$ denote $\mathcal{T}'$. Let $\mathcal{H}$ denote the cluster hierarchy of Σ. Now observe that the subsequence Σ_S of Σ consisting of those pairs that are vertices in the pendant subtree of $\mathcal{H}$ whose root vertex is $(\mathcal{T}_i, \mathcal{T}'_i)$ is a cluster sequence for $\mathcal{S}$ and $\mathcal{S}'$. Recalling that $\mathcal{G}_{t+1}$ is the agreement forest for Σ found by Minimum-Weight Forest, let

$$\mathcal{G}_S = \{\mathcal{L}_i \in \mathcal{G}_{t+1} : \mathcal{L}_i \cap \mathcal{L}(\mathcal{T}_{i'}) \neq \emptyset, (\mathcal{T}_{i'}, \mathcal{T}'_{i'}) \in \Sigma_S\}.$$

Clearly, $\mathcal{G}_S$ is an agreement forest for Σ_S. To establish the theorem we show by induction on the length of Σ_S that $\mathcal{G}_S$ is an agreement forest of minimum weight for Σ_S where, amongst all such forests, the root label ρ_i is a singleton if possible.

If Σ_S has length 1, then $(\mathcal{T}_i, \mathcal{T}'_i)$ is a leaf of $\mathcal{H}$, and it immediately follows that $\mathcal{G}_S$ is an agreement forest of minimum weight for Σ_S. Now suppose that the result holds for all sequences of length less than Σ_S.

Let $f_1, f_2, \ldots, f_s$ denote the reduction edges of $\mathcal{T}$ corresponding to the child vertices of $(\mathcal{T}_i, \mathcal{T}'_i)$ in $\mathcal{H}$. Similarly, let $f'_1, f'_2, \ldots, f'_s$ denote the reduction edges of $\mathcal{T}'$ corresponding to the child vertices of $(\mathcal{T}_i, \mathcal{T}'_i)$ in $\mathcal{H}$. For each $r \in \{1, 2, \ldots, s\}$, the path in $\mathcal{T}$ from e to f_r does not contain any other element in $\{\rho, e_1, e_2, \ldots, e_t\}$ and the path in $\mathcal{T}'$ from e' to f'_r does not contain any other element in $\{\rho, e'_1, e'_2, \ldots, e'_t\}$. Without loss of generality, we may assume that $C_{\mathcal{T}}(f_r) = C_{\mathcal{T}'}(f'_r)$ for all r. In the construction of Σ, let b_r denote the replacement vertex at the ends of f_r and f'_r. Thus $b_r \in \{a_1, a_2, \ldots, a_t\}$. For each r, let $\mathcal{S}_r$ denote $\mathcal{T}|C_{\mathcal{T}}(f_r)$ and $\mathcal{S}'_r$ denote $\mathcal{T}'|C_{\mathcal{T}}(f_r)$. Furthermore, for each r, let Σ_r denote the subsequence of Σ consisting of those pairs that are vertices in the pendant subtree of $\mathcal{H}$ whose root vertex corresponds to the reduction edges f_r and f'_r. As above, note that Σ_r is a cluster sequence for $\mathcal{S}_r$ and $\mathcal{S}'_r$. By the induction assumption, for all r,

$$\mathcal{G}_r = \{\mathcal{L}_i \in \mathcal{G}_{t+1} : \mathcal{L}_i \cap \mathcal{L}(\mathcal{T}_{i'}) \neq \emptyset, (\mathcal{T}_{i'}, \mathcal{T}'_{i'}) \in \Sigma_r\}$$

is an agreement forest of minimum weight for Σ_r where, amongst all such forests, the root label, ζ_r say, of $\mathcal{S}_r$ and $\mathcal{S}'_r$ is a singleton if possible. Note that $\zeta_r \in \{\rho_1, \rho_2, \ldots, \rho_t\}$.

Let $\mathcal{G}^*_S$ be an agreement forest for Σ_S of minimum weight. For the purposes of obtaining a contradiction, suppose that either $w(\mathcal{G}^*_S) < w(\mathcal{G}_S)$, or $w(\mathcal{G}^*_S) = w(\mathcal{G}_S)$ and ρ_i is a singleton in $\mathcal{G}^*_S$ but it is not a singleton in $\mathcal{G}_S$. For all r, let

$$\mathcal{G}^*_r = \{\mathcal{L}_i \in \mathcal{G}^*_S : \mathcal{L}_i \cap \mathcal{L}(\mathcal{T}_{i'}) \neq \emptyset, (\mathcal{T}_{i'}, \mathcal{T}'_{i'}) \in \Sigma_r\}.$$

Since Σ_r is of smaller length than Σ_S, it follows by the induction assumption that $w(\mathcal{G}_r) \leq w(\mathcal{G}^*_r)$ for all r. Therefore, either

(i) $w(\mathcal{G}_r) = w(\mathcal{G}^*_r)$, in which case if ζ_r is a singleton in $\mathcal{G}^*_r$, then it is a singleton in $\mathcal{G}_r$, or
(ii) $w(\mathcal{G}_r) + 1 = w(\mathcal{G}^*_r)$, in which case ζ_r is not a singleton in $\mathcal{G}_r$, but it is a singleton in $\mathcal{G}^*_r$.

These are the only two possibilities, otherwise $w(\mathcal{G}_r) + 1 = w(\mathcal{G}^*_r)$, in which case ζ_r is a singleton in $\mathcal{G}_r$ or ζ_r is not a singleton in $\mathcal{G}^*_r$, or $w(\mathcal{G}_r) + 2 \leq w(\mathcal{G}^*_r)$. In both cases, $(\mathcal{G}^*_S - \mathcal{G}^*_r) \cup \mathcal{G}_r$ is an agreement forest of Σ_S with smaller weight than

$\mathcal{G}_S^*$; a contradiction to the minimality of $\mathcal{G}_S^*$. By reindexing if necessary, we may assume that $w(\mathcal{G}_r) = w(\mathcal{G}_r^*)$ for all $r \in \{1, \ldots, j\}$ and $w(\mathcal{G}_r) + 1 = w(\mathcal{G}_r^*)$ for all $r \in \{j+1, \ldots, s\}$.

Let $\mathcal{F}_i$ and $\mathcal{F}_i^*$ denote $\mathcal{G}_S - \bigcup_{r\in\{1,\ldots,s\}} \mathcal{G}_r$ and $\mathcal{G}_S^* - \bigcup_{r\in\{1,\ldots,s\}} \mathcal{G}_r^*$, respectively. Note that $\mathcal{F}_i$ is the set obtained in Step 2 in Minimum-Weight Forest and $\mathcal{F}_i^*$ is an agreement forest for $\mathcal{T}_i$ and $\mathcal{T}_i'$. Furthermore, let

$$p_1 = |\{(\zeta_r, b_r) : r \in \{1, \ldots, j\} \text{ and } \{\zeta_r\}, \{b_r\} \in \mathcal{G}_S\}|,$$

and let

$$p_1^* = |\{(\zeta_r, b_r) : r \in \{1, \ldots, j\} \text{ and } \{\zeta_r\}, \{b_r\} \in \mathcal{G}_S^*\}|$$

and

$$p_2^* = |\{(\zeta_r, b_r) : r \in \{j+1, \ldots, s\} \text{ and } \{\zeta_r\}, \{b_r\} \in \mathcal{G}_S^*\}|.$$

Observe that

$$s - j \geq p_2^*. \tag{3}$$

Since ζ_r is a singleton in $\mathcal{G}_r$ whenever ζ_r is a singleton in $\mathcal{G}_r^*$ for all $r \in \{1, \ldots, j\}$, it follows by the algorithm's choice of $\mathcal{F}_i$ in Step 2 that

$$|\mathcal{F}_i| - p_1 \leq |\mathcal{F}_i^*| - p_1^*. \tag{4}$$

Now

$$\begin{aligned} w(\mathcal{G}_S^*) &= \sum_{r=1}^{j} w(\mathcal{G}_r^*) + \sum_{r=j+1}^{s} w(\mathcal{G}_r^*) + |\mathcal{F}_i^*| - (p_1^* + p_2^*) - s \\ &= \sum_{r=1}^{s} w(\mathcal{G}_r) + (s-j) + |\mathcal{F}_i^*| - (p_1^* + p_2^*) - s \qquad (5) \end{aligned}$$

and

$$w(\mathcal{G}_S) = \sum_{r=1}^{s} w(\mathcal{G}_r) + |\mathcal{F}_i| - p_1 - s, \tag{6}$$

where (5) follows by the above reindexing. If $w(\mathcal{G}_S^*) < w(\mathcal{G}_S)$, then

$$s - j + |\mathcal{F}_i^*| - (p_1^* + p_2^*) < |\mathcal{F}_i| - p_1. \tag{7}$$

Combining (7) and (4),

$$s - j < p_2^*,$$

contradicting (3). Thus we may assume that $w(\mathcal{G}_S^*) = w(\mathcal{G}_S)$ and ρ_i is a singleton in $\mathcal{G}_S^*$ but it is not a singleton in $\mathcal{G}_S$. Therefore, by (4), (5), and (6),

$$s - j + |\mathcal{F}_i^*| - (p_1^* + p_2^*) = |\mathcal{F}_i| - p_1 \leq |\mathcal{F}_i^*| - p_1^*.$$

Since $(s-j) - p_2^* \geq 0$ by (3), it follows that $s - j = p_2^*$ and so, in particular, $|\mathcal{F}_i^*| - p_1^* = |\mathcal{F}_i| - p_1$. Furthermore, for all $r \in \{1, 2, \ldots, j\}$, whenever ζ_r is a singleton in $\mathcal{G}_r^*$, it is also a singleton in $\mathcal{G}_r$. Hence, as ρ_i is a singleton in $\mathcal{F}_i^*$, Minimum-Weight Forest would have not have chosen $\mathcal{F}_i$ in Step 2. We deduce that $\mathcal{G}_S$ is an agreement forest of Σ_S of minimum with the property that, amongst all such forests, ρ_i is a singleton if possible. This completes the proof of the theorem. □

5. Some Remarks on the Fixed-Parameter Tractability of Calculating the rSPR Distance

Fixed-parameter algorithms have recently attracted much attention in various areas of computational biology (e.g. see [2, 10] and references therein). The idea behind fixed-parameter algorithms is that although the general instance of a problem is NP-hard, many practical instances may be tractable in reasonable time. This may be the case if one can find an algorithm whose running time separates the size of the instance and the parameter of interest. In the case of computing the rSPR distance between two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, Bordewich and Semple [7] have shown that such an algorithm exists. In particular, they showed that this distance can be computed in time $O(f(k) + p(n))$, where k is the actual rSPR distance between $\mathcal{T}$ and $\mathcal{T}'$, $n = |X|$, f is a computable function, and p is a fixed polynomial. Thus, if k is small, the problem might be tractable even for a large n. For further information about fixed-parameter tractability, we refer the interested reader to [9].

To show that finding the rSPR distance between two arbitrary rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is fixed-parameter tractable, it suffices to kernalize the problem using two reduction rules [7]:

Rule 1. Replace a pendant subtree that occurs identically in both trees by a single leaf with a new label.

Rule 2. Replace a chain of at least three common pendant subtrees that occur identically and with the same orientation relative to the root in both trees by three new leaves, say a, b, c, correctly orientated to preserve the direction of the chain.

These reduction rules are the subtree and chain rules mentioned in the introduction. Both rules preserve the rSPR distance [7, Proposition 3.2]. That is, if $\mathcal{S}$ and $\mathcal{S}'$ are the two rooted binary phylogenetic trees resulting from a single application of either Rule 1 or Rule 2 to $\mathcal{T}$ and $\mathcal{T}'$, then

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{S}, \mathcal{S}')$$

or, equivalently, by Theorem 2.1, the size of a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is equal to the size of a maximum-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$. It is shown in [7] that repeated applications of both rules to $\mathcal{T}$ and $\mathcal{T}'$ until no further reductions are possible result in two rooted binary phylogenetic X'-trees, where $|X'|$ is linear in $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$; in particular, $|X'| \leq 28 d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. Applying an exhaustive search gives the aforementioned running time.

The obvious first way to make use of the subtree and chain rules in Minimum-Weight Forest is to preprocess the initial two trees by applying the these rules repeatedly before constructing any cluster sequence. This immediately implies that Minimum-Weight Forest is fixed-parameter tractable, and so one can think of the cluster sequence as a way of aiding the exhaustive search. Now suppose we have a cluster sequence of the resulting trees. In the following, we consider the use of the chain rule, but a similar analysis can be done for the simpler subtree rule.

Suppose that $\mathcal{S}$ and $\mathcal{S}'$ have been obtained from two arbitrary rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ by a single application of the chain rule. The reason that the chain rule preserves the rSPR distance is that there exists a maximum-agreement forest $\mathcal{F}_S$ for $\mathcal{S}$ and $\mathcal{S}'$ in which $\{a, b, c\}$ is a subset of a label set in $\mathcal{F}_S$ [7, Lemma 3.1]. By replacing a, b, and c in this label set with the original elements of the chain, we obtain an agreement forest $\mathcal{F}_T$ for $\mathcal{T}$ and $\mathcal{T}'$. By the optimality of $\mathcal{F}_S$, it follows that $\mathcal{F}_T$ is a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ (see [1, 7]). Using the equivalence of Steps 2 and 2′, we can incorporate the chain rule in Minimum-Weight Forest in Step 2 as follows:

(i) first, reduce $\mathcal{T}_i|(\mathcal{L}(\mathcal{T}_i)-\{a_{i'} : \{\rho_{i'}\} \in \mathcal{G}_j\})$ and $\mathcal{T}_i'|(\mathcal{L}(\mathcal{T}_i)-\{a_{i'} : \{\rho_{i'}\} \in \mathcal{G}_j\})$ with repeated applications of the chain rule;
(ii) second, find a maximum-agreement forest of the resulting trees such that each 3-element set $\{a, b, c\}$ resulting from the chain rule is a subset of a label set and, amongst all such forests, ρ_i a singleton if possible;
(iii) third, replace each 3-element set $\{a, b, c\}$ with the original elements of the associated chain to obtain $\mathcal{F}_i'$.

(Note that any element in $\{a_1, a_2, \ldots, a_t\}$ that is in the label sets of the restrictions in (i) does not effect the weighting as their counterpart in $\{\rho_1, \rho_2, \ldots, \rho_t\}$ is not a singleton in $\mathcal{G}_j$.) The two possible causes for concern are that

(I) we can find no maximum-agreement forest in (ii) such that each 3-element set $\{a, b, c\}$ is a subset of a label set, and
(II) there is a maximum-agreement forest for the restrictions in (i) with ρ_i a singleton and we no longer find it because no maximum-agreement forest found in (ii) has ρ_i as a singleton.

However, the proof of [7, Lemma 3.1] works by taking a maximum-agreement forest and making small modifications to get the desired outcome. An analysis of the proof shows that we can sequentially find a maximum-agreement forest so that if $\{a, b, c\}$ is a subset of a label set prior to the modifications, then it is also a subset of a label set after the modifications. This resolves (I). Furthermore, this analysis also shows that if ρ_i is a singleton prior to the modifications, then it is a singleton afterwards. It now follows that if there is a maximum-agreement forest for the trees resulting from repeated use of the chain rule with ρ_i a singleton, then the maximum-agreement forest found in (ii) also has ρ_i as a singleton, in which case, we obtain via (iii) a maximum-agreement forest for the restrictions in (i) with ρ_i a singleton. Similarly, the converse also holds by noting that common subtrees are never broken across different label sets in a maximum-agreement forest and that the analogous outcome of [7, Lemma 3.1] holds for all chains of size at least 3. This resolves (II).

References

[1] B.L. Allen and M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Comb. **5** (2001) 1-13.
[2] L.F. Ávila, G. García, M. Serna, D.M. Thilikos, A list of parameterized problems in bioinformatics, Technical report LSI-06-24-R (2006) Technical University of Catalonia.

[3] M. Baroni, C. Semple, and M. Steel, A framework for representing reticulate evolution, Ann. Comb. **8** (2004) 391-408.
[4] M. Baroni, S. Grünewald, V. Moulton, C. Semple, Bounding the number of hybridization events for a consistent evolutionary history, J. Math. Biol. **51** (2005) 171-182.
[5] M. Baroni, C. Semple, and M. Steel, Hybrids in real time, Syst. Biol. **55** (2006) 46-56.
[6] R. Beiko and N. Hamilton, Phylogenetic identification of lateral genetic transfer events, BMC Evol. Biol. **6**:15 (2006).
[7] M. Bordewich and C. Semple, C. (2004). On the computational complexity of the rooted subtree prune and regraft distance, Ann. Comb. **8** (2004) 409-423.
[8] M. Bordewich, S. Linz, K. St. John, and C. Semple, A reduction algorithm for computing the hybridization number of two trees, Evol. Bioinform. Online **3** (2007) 86-98.
[9] R. Downey and M. Fellows, Parameterized Complexity (Monographs in Computer Science), Springer Verlag, 1998.
[10] J. Gramm, A. Nickelsen, and T. Tantau, Fixed-parameter algorithms in phylogenetics, TCJ **51** (2008) 79-101.
[11] D.M. Hillis, B.K. Mable, and C. Moritz, Molecular Systematics, Sinauer Assoc., Sunderland, Mass., 1996.
[12] W. Maddison, Gene trees in species trees, Syst. Biol. **46** (1997) 523-536.
[13] L. Nakhleh, T. Warnow, C.R. Linder, and K. St. John, Reconstructing reticulate evolution in species—theory and practice, J. Comput. Biol. **12** (2005) 796-811.
[14] C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.
[15] Y. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events, In: Proceedings of the Workshop on Algorithms in Bioinformatics, Lecture Notes in Bioinformatics, vol. 2812, 2003, pp. 287-302.

Department of Computer Science, Heinrich-Heine-University, Düsseldorf, Germany

E-mail address: linz@cs.uni-duesseldorf.de

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: c.semple@math.canterbury.ac.nz