Hybridization in Non-Binary Trees

Simone LINZ and Charles SEMPLE

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2008/1 MARCH 2008

HYBRIDIZATION IN NON-BINARY TREES

SIMONE LINZ AND CHARLES SEMPLE

ABSTRACT. Reticulate evolution—the umbrella term for processes like hybridization, horizontal gene transfer, and recombination—plays an important role in the history of life of many species. Although the occurrence of such events is widely accepted, approaches to calculate the extent to which reticulation has influenced evolution are relatively rare. In this paper, we show that the NP-hard problem of calculating the minimum number of reticulation events for two (arbitrary) rooted phylogenetic trees parameterized by this minimum number is fixed-parameter tractable.

1. INTRODUCTION

Using mathematical models to reconstruct a tree of life from nucleotide or protein sequences is subject of many phylogenetic studies that aim at analyzing the complex evolutionary processes that have occurred during the development of the current diversity of species. Under the usual assumption that each species arises from its ancestor by a simple speciation event, tree-based methods have contributed significantly to approaching this task. However, due to non-tree-like events, not all groups of taxa are suited to this type of presentation. Such processes, collectively referred to as reticulation events, include hybridization, horizontal gene transfer, and recombination. Since reticulate evolution results in genomes that are mosaics of distinct ancestral genomes, there has been an increased interest in modeling evolutionary relationships using phylogenetic networks rather than phylogenetic trees.

In this paper, we focus our attention on hybridization and its impact on evolution. This has been an active and controversially discussed field of research for many years and even several definitions of the term hybridization have been suggested [8]. For the purposes of this article, we refer to the origin of a new species through a mating between two different species as a hybridization event. Hybridization is widely accepted to play an important role in the evolutionary history of certain groups of plants and fish. For a review of hybrid species, we refer the reader to [11].

Date: March 18, 2008.

1991 *Mathematics Subject Classification*. 05C05; 92D15.

Key words and phrases. Rooted phylogenetic tree, reticulate evolution, hybridization network.

The first and second author were supported by the New Zealand Marsden Fund. Part of this work was carried out while the authors were visiting the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

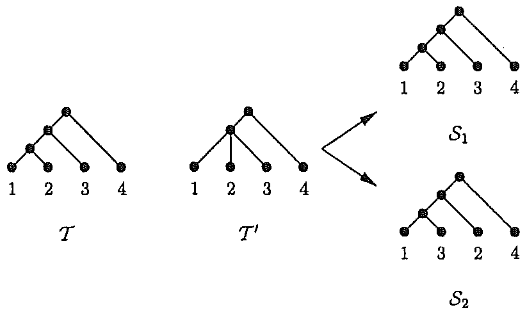

FIGURE 1. Two rooted phylogenetic tree $\mathcal{T}$ and $\mathcal{T}'$ and two binary refinements $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathcal{T}'$. The hybridization number for $\mathcal{S}_1$ and $\mathcal{T}$ is 0, while this number for $\mathcal{S}_2$ and $\mathcal{T}$ is 1.

To provide insight into the extent to which hybridization has influenced the evolution of a set of present-day species, this paper addresses the following fundamental problem: Given a collection of rooted phylogenetic trees that are correctly reconstructed for different genetic loci, what is the smallest number of hybridization events needed to simultaneously explain the evolutionary scenarios of the gene trees under consideration?

Bordewich and Semple [4] showed that the above problem is NP-hard even when the initial collection consists of two rooted binary phylogenetic trees. However, the same authors showed [5] that in the case of two binary trees the problem is fixed-parameter tractable. In particular, they showed that the minimum number of hybridization events can be computed in time $O(f(k)+p(|X|))$, where k is the actual minimum number, f is some computable function, and p is a fixed polynomial. Due to the NP-hardness of the problem, such a result is of importance, since for many practical instances, the minimum number of hybridization events is small and, therefore, the problem may be tractable, even for a large number of taxa. This can be seen by considering the separation of the variables k and $|X|$. For more details about fixed-parameter tractability, we refer the interested reader to [6].

Despite the above fixed-parameter tractable algorithm, for many biological data sets in practice (e.g. [7, 12]), the reconstructed phylogenetic trees are not fully resolved; that is, they contain *polytomies*. For example, this may be due to either the tree reconstruction method or the use of consensus trees for a certain analysis. Polytomies—alternatively called *multifurcations*—refer to vertices that have more than two direct descendants. A polytomy is said to be *hard* if it refers to an event during which an ancestral species gave rise to more than two offspring species at the same time, whereas a *soft* polytomy represents ambiguous evolutionary relationships as a result of insufficient information [10].

Since simultaneous speciation events only occur rarely, we typically assume that all polytomies in a phylogenetic tree are soft. The reconstruction of a strictly bifurcating (binary) tree may consequently force refinements that are not necessarily optimal in terms of the hybridization number. An example for that is depicted in

Fig. 1, where two binary refinements $\mathcal{S}_1$ and $\mathcal{S}_2$ of the tree $\mathcal{T}'$ are shown. While the hybridization number for $\mathcal{S}_1$ and $\mathcal{T}$ is 0, this number for $\mathcal{S}_2$ and $\mathcal{T}$ is 1.

In this paper, we show that the decision problem of asking whether the minimum number of hybridization events to explain two (arbitrary) rooted phylogenetic trees is at most k is fixed-parameter tractable. We now describe the above-mentioned problem formally beginning with several definitions.

A *rooted phylogenetic X-tree $\mathcal{T}$* is a rooted tree with no degree-2 vertices except possibly the root which has degree at least two, and with leaf set X. The set X is called the *label set* of $\mathcal{T}$ and is denoted by $\mathcal{L}(\mathcal{T})$. In addition, $\mathcal{T}$ is *binary* if, apart from the root which has degree two, all interior vertices have degree three.

Let Y be a subset of X. We call Y an *(edge) cluster* of $\mathcal{T}$ if there is an edge e, or equivalently a vertex v, whose set of descendants in X is precisely Y. We denote this cluster by $C_{\mathcal{T}}(v)$, or simply $C(v)$ if there is no ambiguity. The set of clusters of $\mathcal{T}$ is denoted by $\mathcal{C}(\mathcal{T})$. Furthermore, the *most recent common ancestor* of Y is the vertex v in $\mathcal{T}$ with $Y \subseteq C_{\mathcal{T}}(v)$ such that there exists no vertex v' with $Y \subseteq C_{\mathcal{T}}(v')$ and $C_{\mathcal{T}}(v') \subset C_{\mathcal{T}}(v)$. We denote v by $\text{mrca}_{\mathcal{T}}(Y)$.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees. We say that $\mathcal{T}'$ *refines* $\mathcal{T}$, or equivalently $\mathcal{T}'$ is a *refinement* of $\mathcal{T}$, if $\mathcal{C}(\mathcal{T}) \subseteq \mathcal{C}(\mathcal{T}')$. In addition, $\mathcal{T}'$ is a *binary refinement* if $\mathcal{T}'$ is binary. Note that $\mathcal{T}$ is a refinement of itself. Graphically speaking, it is straightforward to see that if $\mathcal{T}'$ refines $\mathcal{T}$, then $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by contracting interior edges.

Hybridization networks are a generalization of evolutionary trees that allow for a simultaneous visualization of several conflicting or alternating histories of life. Such a network embeds a collection of gene trees representing a set of present-day species, where each vertex whose in-degree is greater than 1 represents a hybrid species. Mathematically speaking, a *hybridization network* $\mathcal{H}$ (on X) is a rooted acyclic digraph with root ρ in which

(i) X is the set of vertices of out-degree zero,
(ii) the out-degree of ρ is at least 2, and
(iii) for each vertex with out-degree 1, its in-degree is at least 2.

To quantify the number of reticulation events, the *hybridization number* of a hybridization network $\mathcal{H}$ with root ρ is

$$h(\mathcal{H}) = \sum_{v \neq \rho} (d^{-}(v) - 1),$$

where v is a vertex of $\mathcal{H}$.

Let $\mathcal{T}$ be a rooted phylogenetic X-tree, and let $\mathcal{H}$ be a hybridization network. We say that $\mathcal{H}$ *displays* $\mathcal{T}$ if $\mathcal{L}(\mathcal{T}) \subseteq \mathcal{L}(\mathcal{H})$ and there is a rooted subtree of $\mathcal{H}$ that is a refinement of $\mathcal{T}$. In other words, $\mathcal{T}$ can be obtained from $\mathcal{H}$ by first deleting a subset of the edges of $\mathcal{H}$, deleting and suppressing any resulting degree-0 and degree-2 vertices, respectively, and then contracting edges. For a collection $\mathcal{P}$ of rooted phylogenetic trees, $\mathcal{H}$ *displays* $\mathcal{P}$ if each tree in $\mathcal{P}$ is displayed by $\mathcal{H}$.

Furthermore, extending the definition of the hybridization number of a network to $\mathcal{P}$, we set

$$h(\mathcal{P}) = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybridization network that displays } \mathcal{P}\}.$$

If $\mathcal{P}$ contains precisely two rooted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, then we denote the hybridization number $h(\mathcal{P})$ by $h(\mathcal{T}, \mathcal{T}')$ and remark that the beforehand given definition is equivalent to

$$h(\mathcal{T}, \mathcal{T}') = \min\{h(\mathcal{S}, \mathcal{S}') : \mathcal{S} \text{ and } \mathcal{S}' \text{ are binary refinements of } \mathcal{T} \text{ and } \mathcal{T}', \text{ respectively}\}.$$

Throughout the paper, both definitions are used interchangeably.

We can now formally state the decision problem for when $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$:

Hybridization Number
Instance: Two rooted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, and an integer k.
Question: Is $h(\mathcal{T}, \mathcal{T}') \leq k$?

Since computing $h(\mathcal{T}, \mathcal{T}')$ is NP-hard when $\mathcal{T}$ and $\mathcal{T}'$ are binary [4], calculating this value for when $\mathcal{T}$ and $\mathcal{T}'$ are arbitrary rooted phylogenetic X-trees is also NP-hard.

The main result of this paper is the following theorem.

Theorem 1.1. *The decision problem* Hybridization Number *is fixed-parameter tractable with $h(\mathcal{T}, \mathcal{T}')$ being the parameter.*

The overall approach in proving Theorem 1.1 is similar to that used to show that Hybridization Number is fixed-parameter tractable when the initial two trees are binary. Basically, we use three reductions to kernalize the problem instance before calculating exactly the minimum number of hybridization events using an exhaustive search. The reason that this is sufficient to prove Theorem 1.1 is that the size of the label set of the trees $\mathcal{S}$ and $\mathcal{S}'$ obtained from $\mathcal{T}$ and $\mathcal{T}'$ by repeatedly applying the three reductions is linear in $h(\mathcal{T}, \mathcal{T}')$.

The paper is organized as follows. The next section contains some additional preliminaries that are used throughout the paper. In Sections 3 and 4, we characterize Hybridization Number in terms of a particular type of agreement forest. This characterization is essential to getting the main result of the paper. Section 5 describes the three reductions that are used to kernalize the problem instance and also includes three key lemmas that are needed for the proof of Theorem 1.1. This proof is given in Section 6. The paper ends with some brief remarks in Section 7.

We end the introduction by remarking that despite the similarities between the approaches used to prove Theorem 1.1 and the analogous result for binary trees, we see no obvious way that this latter result can be used to directly establish Theorem 1.1. Part of the reason for this is that a number of additional and non-trivial complications arise in the non-binary case.

2. Preliminaries

In this section, we give some preliminary definitions that are used throughout the paper. Unless stated otherwise, the notation and terminology follows [13].

For a rooted phylogenetic X-tree $\mathcal{T}$, a subset Y of X is called a *vertex cluster* of $\mathcal{T}$ if there is a refinement of $\mathcal{T}$ in which Y is an edge cluster. For example, considering Fig. 1, the taxa set $\{1, 2\}$ is an edge cluster in $\mathcal{T}$, but a vertex cluster (and not an edge cluster) in $\mathcal{T}'$. Note that edge clusters are special types of vertex clusters.

Let $\mathcal{T}$ be a rooted phylogenetic X-tree. Several types of rooted subtrees of $\mathcal{T}$ play a central role in this paper. Let Y be a subset of X. The minimal rooted subtree of $\mathcal{T}$ that connects the leaves in Y is denoted by $\mathcal{T}(Y)$. Furthermore, the *restriction of* $\mathcal{T}$ *to* Y, denoted $\mathcal{T}|Y$, is the subtree obtained from $\mathcal{T}(Y)$ by contracting all non-root vertices of degree two. Furthermore, a subtree of $\mathcal{T}$ is *pendant* if it can be obtained from a refinement of $\mathcal{T}$ by deleting a single edge. Lastly, a subtree is *non-trivial* if it contains at least two leaves.

3. Agreement Forests

Various types of agreement forests have recently been used to analyze reticulate evolution for a set of gene trees and its impact on evolution [1, 3, 5, 14, 15]. All of these approaches are restricted to the case when the trees under consideration are binary. Here, we extend the definition of agreement forests to arbitrary rooted phylogenetic trees. For the reader familiar with agreement forests, we note that the following definitions coincide with those previously given for rooted binary phylogenetic trees.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees. For the purposes of the upcoming definitions, we regard the root of both $\mathcal{T}$ and $\mathcal{T}'$ as a vertex labeled ρ at the end of a pendant edge adjoined to the original root. Furthermore, we also regard ρ as part of the label set of $\mathcal{T}$ and $\mathcal{T}'$, thus we view their label sets as $X \cup \{\rho\}$.

A *forest* of $\mathcal{T}$ is a partition $\{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ of its label set $X \cup \{\rho\}$, where $\mathcal{L}_\rho$ contains ρ, no part is empty, and the trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are edge-disjoint rooted subtrees of $\mathcal{T}$. An *agreement forest* $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ is a forest $\{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ of $\mathcal{T}$ and $\mathcal{T}'$ such that, for all $i \in \{\rho, 1, 2, \ldots, k\}$, the trees $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement. To illustrate these concepts, two examples of agreement forests $\mathcal{F}_1$ and $\mathcal{F}_2$ are shown in Fig. 2 for the two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ also shown in that figure. Considering $\mathcal{F}_1$, it is easily checked that, for each label set $\mathcal{L}_i$, the restrictions of $\mathcal{T}$ and $\mathcal{T}'$, respectively, to $\mathcal{L}_i$ have a common binary refinement.

The subtree prune and regraft distance between two rooted binary phylogenetic X-trees can be characterized in terms of agreement forests. However, the corresponding characterization for the minimum number of hybridization events for the same pair of trees requires an additional condition. This condition excludes the

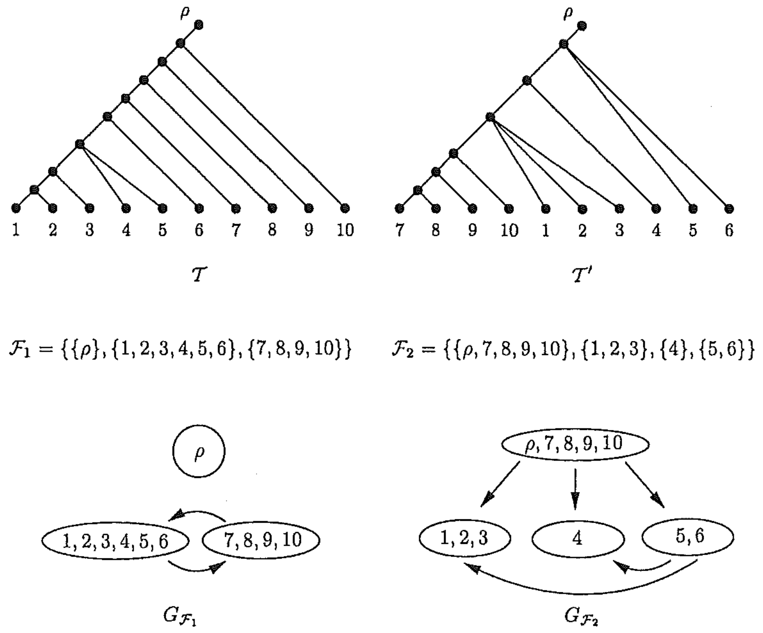

FIGURE 2. Two agreement forests $\mathcal{F}_1$ and $\mathcal{F}_2$ for the two rooted trees $\mathcal{T}$ and $\mathcal{T}'$ and their associated digraphs $G_{\mathcal{F}_1}$ and $G_{\mathcal{F}_2}$.

possibility that species inherit genetic material from their own descendants. Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be an agreement forest for two arbitrary rooted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$. Let $G_\mathcal{F}$ be the directed graph that has vertex set $\mathcal{F}$ and an arc $(\mathcal{L}_i, \mathcal{L}_j)$ from $\mathcal{L}_i$ to $\mathcal{L}_j$ precisely if $i \neq j$ and either

(I) the path from the root of $\mathcal{T}(\mathcal{L}_i)$ to the root of $\mathcal{T}(\mathcal{L}_j)$ contains an edge of $\mathcal{T}(\mathcal{L}_i)$, or

(II) the path from the root of $\mathcal{T}'(\mathcal{L}_i)$ to the root of $\mathcal{T}'(\mathcal{L}_j)$ contains an edge of $\mathcal{T}'(\mathcal{L}_i)$.

We say that $\mathcal{F}$ is an *acyclic-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ if $G_\mathcal{F}$ contains no directed cycles, that is, $G_\mathcal{F}$ is acyclic. For the example depicted in Fig. 2, $\mathcal{F}_2$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ since $G_{\mathcal{F}_2}$ is acyclic, whereas $\mathcal{F}_1$ is not an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. If $\mathcal{F}$ contains the smallest number of parts over all acyclic-agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, we say that $\mathcal{F}$ is a *maximum-acyclic-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$, in which case, we denote this value of k by $m_a(\mathcal{T}, \mathcal{T}')$. In the case that both $\mathcal{T}$ and $\mathcal{T}'$ are binary, these definitions again extend those typically given for two rooted binary phylogenetic trees. Baroni *et al.* [1] established the following characterization for binary trees.

Theorem 3.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then*

$$h(\mathcal{T}, \mathcal{T}') = m_a(\mathcal{T}, \mathcal{T}').$$

4. Characterizing $h(\mathcal{T},\mathcal{T}')$ in Terms of Agreement Forests

In this section, we prove the following analogue of Theorem 3.1 for arbitrary rooted phylogenetic trees. This analogue is crucial in proving the main result of the paper.

Theorem 4.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees. Then*

$$h(\mathcal{T},\mathcal{T}') = m_a(\mathcal{T},\mathcal{T}').$$

Essentially, all of the work in establishing this theorem is done in proving the next two lemmas.

Lemma 4.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees, and let $\mathcal{F}$ be an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Then there exist binary refinements $\mathcal{S}$ and $\mathcal{S}'$ of $\mathcal{T}$ and $\mathcal{T}'$, respectively, such that $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$.*

Proof. Suppose that $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and let $\mathcal{B}_i$ be a common binary refinement of $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ for all i. The proof of the lemma is by induction on k. Clearly, the result holds if $k = 0$. Now suppose that the result holds for all acyclic-agreement forests of $\mathcal{T}$ and $\mathcal{T}'$ of size at most k. Since $\mathcal{F}$ is acyclic, $G_\mathcal{F}$ contains a vertex, $\mathcal{L}_m$ say, with out-degree zero. Since $\mathcal{L}_m$ has out-degree zero, $\mathcal{T}(\mathcal{L}_m)$ is a pendant subtree of $\mathcal{T}$ and $\mathcal{T}'(\mathcal{L}_m)$ is a pendant subtree of $\mathcal{T}'$.

Let $\mathcal{T}_m$ and $\mathcal{T}'_m$ be the rooted phylogenetic trees $\mathcal{T}|((X\cup\{\rho\})-\mathcal{L}_m)$ and $\mathcal{T}'|((X\cup\{\rho\}) - \mathcal{L}_m)$, respectively, and let $\mathcal{F}_m = \mathcal{F} - \{\mathcal{L}_m\}$. Since $\mathcal{F}$ is an acyclic-agreement forest of $\mathcal{T}$ and $\mathcal{T}'$, it is easily checked that, as $\mathcal{T}(\mathcal{L}_m)$ is a pendant subtree of $\mathcal{T}$ and $\mathcal{T}'(\mathcal{L}_m)$ is a pendant subtree of $\mathcal{T}'$, the collection $\mathcal{F}_m$ is an acyclic-agreement forest of $\mathcal{T}_m$ and $\mathcal{T}'_m$. Therefore, by the induction assumption, there are binary refinements $\mathcal{S}_m$ and $\mathcal{S}'_m$ of $\mathcal{T}_m$ and $\mathcal{T}'_m$, respectively, such that $\mathcal{F}_m$ is an acyclic-agreement forest for $\mathcal{S}_m$ and $\mathcal{S}'_m$.

We now construct a binary refinement of $\mathcal{T}$ from $\mathcal{S}_m$. Let u be the vertex of $\mathcal{T}$ with the property that $\mathcal{C}(u)$ is the minimal cluster of $\mathcal{T}$ that properly contains $\mathcal{L}_m$. By construction, $\mathcal{C}(u) - \mathcal{L}_m$ is a cluster of $\mathcal{T}_m$. Furthermore, as $\mathcal{S}_m$ is a binary refinement of $\mathcal{T}_m$, the set $\mathcal{C}(u) - \mathcal{L}_m$ is a cluster of $\mathcal{S}_m$. Let u_m be the vertex of $\mathcal{S}_m$ such that $\mathcal{C}(u_m) = \mathcal{C}(u) - \mathcal{L}_m$. Let $\mathcal{S}$ be the rooted binary phylogenetic tree obtained from $\mathcal{S}_m$ by subdividing the edge coming into u_m with a new vertex v and adjoining the root of $\mathcal{B}_m$ to this new vertex v via a new edge. Observing that $\mathcal{C}(v) = \mathcal{C}(u)$, it is easily checked that $\mathcal{S}$ is a binary refinement of $\mathcal{T}$. Furthermore, by construction and because of the induction assumption, it follows that $\mathcal{F}$ is a forest of $\mathcal{S}$ and, for all i, we have $\mathcal{S}|\mathcal{L}_i = \mathcal{B}_i$.

By the same construction and argument, there is a binary refinement $\mathcal{S}'$ of $\mathcal{T}'$ such that $\mathcal{F}$ is a forest of $\mathcal{S}'$ and, for all i, we have $\mathcal{S}'|\mathcal{L}_i = \mathcal{B}_i$. It now follows that $\mathcal{F}$ is an agreement forest for $\mathcal{S}$ and $\mathcal{S}'$. Moreover, as $\mathcal{F}_m$ is an acyclic-agreement forest for $\mathcal{S}_m$ and $\mathcal{S}'_m$, it is easily seen that $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$. This completes the proof of the lemma. □

Lemma 4.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees, and let $\mathcal{S}$ and $\mathcal{S}'$ be binary refinements of $\mathcal{T}$ and $\mathcal{T}'$, respectively. If $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$, then $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.*

Proof. Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be an acyclic-agreement forest of $\mathcal{S}$ and $\mathcal{S}'$. Since $\mathcal{S}$ and $\mathcal{S}'$ are both binary, it is easily seen, for all i, that $\mathcal{S}|\mathcal{L}_i$ and $\mathcal{S}'|\mathcal{L}_i$ are binary. Therefore, as $\mathcal{S}$ and $\mathcal{S}'$ are binary refinements of $\mathcal{T}$ and $\mathcal{T}'$, respectively, $\mathcal{S}|\mathcal{L}_i$ is a common binary refinement of $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ for all i. To see that the trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are edge-disjoint rooted subtrees of $\mathcal{T}$, suppose that this is not the case. Then, for some $r \neq s$, the subtrees $\mathcal{T}(\mathcal{L}_r)$ and $\mathcal{T}(\mathcal{L}_s)$ are not edge-disjoint. That is, $\mathcal{T}(\mathcal{L}_r)$ and $\mathcal{T}(\mathcal{L}_s)$ have an edge $e = \{u, v\}$ in common. Let u be the end vertex of e closest to ρ. Since $\mathcal{S}$ is a binary refinement of $\mathcal{T}$, there are vertices u' and v' of $\mathcal{S}$ with $\mathcal{C}_\mathcal{S}(u') = \mathcal{C}_\mathcal{T}(u)$ and $\mathcal{C}_\mathcal{S}(v') = \mathcal{C}_\mathcal{T}(v)$. Now it is easily seen that $\mathcal{S}(\mathcal{L}_r)$ contains u' and v', and $\mathcal{S}(\mathcal{L}_s)$ contains u' and v'. In other words, $\mathcal{S}(\mathcal{L}_r)$ and $\mathcal{S}(\mathcal{L}_s)$ are not edge-disjoint in $\mathcal{S}$, contradicting that $\mathcal{F}$ is an agreement forest of $\mathcal{S}$ and $\mathcal{S}'$. Thus the trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are edge-disjoint rooted subtrees of $\mathcal{T}$ and, similarly, the trees in $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are edge-disjoint rooted subtrees of $\mathcal{T}'$. Hence, $\mathcal{F}$ is an agreement forest of $\mathcal{T}$ and $\mathcal{T}'$.

Now relative to $\mathcal{S}$ and $\mathcal{S}'$, the graph $G_\mathcal{F}$ is acyclic. With respect to $\mathcal{F}$, consider the analogous graph, $G'_\mathcal{F}$ say, for $\mathcal{T}$ and $\mathcal{T}'$. Noting that both graphs have the same vertex set, it is clear that if $(\mathcal{L}_r, \mathcal{L}_s)$ is an arc in $G'_\mathcal{F}$, then $(\mathcal{L}_r, \mathcal{L}_s)$ is an arc in $G_\mathcal{F}$. Thus the arc set of $G'_\mathcal{F}$ is a subset of the arc set of $G_\mathcal{F}$. Since $G_\mathcal{F}$ is acyclic, it follows that $G'_\mathcal{F}$ is acyclic. This completes the proof of the lemma. □

Proof of Theorem 4.1. Let $\mathcal{S}$ and $\mathcal{S}'$ be binary refinements of $\mathcal{T}$ and $\mathcal{T}'$ that satisfy the hypothesis of Lemma 4.2. Then, by that lemma, $m_a(\mathcal{T}, \mathcal{T}') \geq m_a(\mathcal{S}, \mathcal{S}')$. But, by Theorem 3.1, $m_a(\mathcal{S}, \mathcal{S}') = h(\mathcal{S}, \mathcal{S}')$. It now follows that, as $h(\mathcal{S}, \mathcal{S}') \geq h(\mathcal{T}, \mathcal{T}')$, we have $m_a(\mathcal{T}, \mathcal{T}') \geq h(\mathcal{T}, \mathcal{T}')$.

To establish the converse, now let $\mathcal{S}$ and $\mathcal{S}'$ be binary refinements of $\mathcal{T}$ and $\mathcal{T}'$ such that $h(\mathcal{S}, \mathcal{S}') = h(\mathcal{T}, \mathcal{T}')$. Then, by Theorem 3.1, there is an acyclic-agreement forest $\mathcal{F}$ of $\mathcal{S}$ and $\mathcal{S}'$ such that

$$|\mathcal{F}| - 1 = h(\mathcal{S}, \mathcal{S}') = h(\mathcal{T}, \mathcal{T}').$$

By Lemma 4.3, $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, so

$$m_a(\mathcal{T}, \mathcal{T}') \leq |\mathcal{F}| - 1 = h(\mathcal{T}, \mathcal{T}').$$

It now follows that $h(\mathcal{T}, \mathcal{T}') = m_a(\mathcal{T}, \mathcal{T}')$. This completes the proof of the theorem. □

5. Reducing the Size of the Problem Instance

In this section, we introduce three reductions which kernalize Hybridization Number. The *subtree* and *long-chain reductions* extend the subtree and chain reductions described in [5]. Additionally, we introduce the *short-chain reduction* which—in combination with the other two reductions—guarantees that all problem instances can be kernalized. We begin with some preliminaries.

Let $\mathcal{T}$ be a rooted phylogenetic X-tree, and let x be a leaf of $\mathcal{T}$. Viewing $\mathcal{T}$ as a directed graph with edges directed away from its root, the unique vertex, u say, of $\mathcal{T}$ such that (u, x) is an arc of $\mathcal{T}$ is called the *parent* of x and is denoted by $p_{\mathcal{T}}(x)$.

For all $n \geq 2$, an *n-chain* of $\mathcal{T}$ is an ordered tuple $(a_1, a_2, \ldots, a_n)$ of distinct elements of X that satisfies the following properties:

(i) for all $i \in \{1, 2, \ldots, n-1\}$, either $p_{\mathcal{T}}(a_i) = p_{\mathcal{T}}(a_{i+1})$ or $p_{\mathcal{T}}(a_i)$ is a child of $p_{\mathcal{T}}(a_{i+1})$, and
(ii) there is an ordering, $p_1, p_2, \ldots, p_m$ say, of the parents of $a_1, a_2, \ldots, a_n$ such that, for all $i \in \{1, 2, \ldots, m-1\}$, the vertex p_i is a child of p_{i+1} and, apart from p_1 and p_m, each of the vertices $p_2, p_3, \ldots, p_{m-1}$ has exactly one child not in $\{a_1, a_2, \ldots, a_m\}$.

If p is a parent of an element in $A = \{a_1, a_2, \ldots, a_n\}$, then p is called *internal* if it has exactly one child not in A; otherwise p is said to be external. An element of A is *internal* (resp. external) if its parent is internal (resp. external). Furthermore, the partition of $\{a_1, a_2, \ldots, a_n\}$ defined by putting a_i and a_j in the same part precisely if $p_{\mathcal{T}}(a_i) = p_{\mathcal{T}}(a_j)$ is called the *parent partition* of $(a_1, a_2, \ldots, a_n)$ induced by $\mathcal{T}$. Throughout the paper, we will assume that if $(a_1, a_2, \ldots, a_n)$ is an n-chain of both $\mathcal{T}$ and $\mathcal{T}'$, where $\mathcal{T}$ and $\mathcal{T}'$ are rooted phylogenetic X-trees, then $\mathcal{T}$ and $\mathcal{T}'$ have no common non-trivial pendant subtree whose label set is a subset of $\{a_1, a_2, \ldots, a_n\}$. As we will soon see, this assumption does not restrict the results in this paper; it is simply for convenience and to avoid repetition in the statements. As an illustration, $(a_1, a_2, \ldots, a_n)$ is an n-chain of the two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Fig. 3, where triangles represent subtrees outside of the chain.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees. Let P be a disjoint collection of subsets of X such that each set in P contains the elements of a chain $(a_1, a_2, \ldots, a_n)$ with $n \leq 5$ in both $\mathcal{T}$ and $\mathcal{T}'$ that is one of the following two types:

(i) For the first type, $3 \leq n \leq 5$ and there are exactly three elements that are internal in both trees. Furthermore, if $n = 4$, then one of the elements is external in one tree, but internal in the other, while if $n = 5$, then either a_1 or a_n is external in one tree, but internal in the other.
(ii) For the second type, $2 \leq n \leq 4$ and, in one of the trees, the chain has precisely two internal elements. Furthermore, if $n = 3$, then either a_1 or a_n is external in the same tree, while if $n = 4$, then both a_1 and a_n are external in this tree. On the other hand, regardless of size, the chain has exactly one parent in the other tree.

Depending on whether the subset is of the first or second types, we assign a triple of weights or a single weight from $\mathbb{Z}^+ \times \mathbb{Z}^+ \times \mathbb{Z}^+$ and $\mathbb{Z}^+$, respectively. We call such a pair of trees with associated weighted set P a *pair of weighted rooted phylogenetic X-trees.*

We now describe the three reductions. Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted rooted phylogenetic X-trees with an associated set P, and let A be a subset of X. We say that A does not *cross* P if, for each element S in P, the intersection $S \cap A$ is empty.

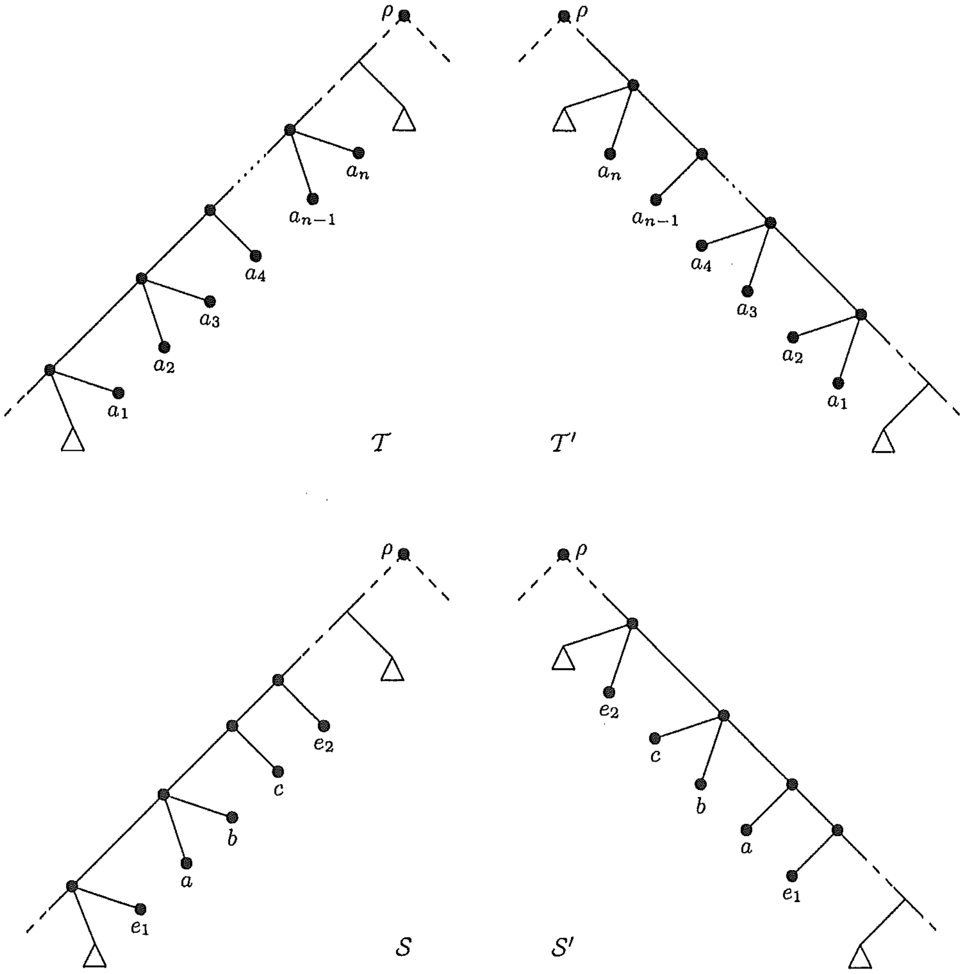

FIGURE 3. Two rooted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ reduced under the long-chain reduction, where $\mathcal{S}$ and $\mathcal{S}'$ are the resulting trees. Dotted lines indicate regions of the chain $(a_1, a_2, \ldots, a_n)$.

Subtree Reduction: For $|A| \geq 2$, if A is the label set of a pendant subtree in $\mathcal{T}$ and $\mathcal{T}'$ such that $\mathcal{T}|A$ and $\mathcal{T}'|A$ have a common binary refinement, A does not cross P, and A is maximal with these properties, then replace these subtrees with either a single new leaf labeled a or a pendant edge ending in a new leaf labeled a depending on whether the subtree can be obtained without or with refinement, respectively. In all cases, the new label is the same in both resulting trees.

Long-Chain Reduction: For $n \geq 4$, let $(a_1, a_2, \ldots, a_n)$ be an n-chain of both $\mathcal{T}$ and $\mathcal{T}'$ that does not cross an element of P and is maximal with the following properties:

(i) The chain has at least three internal parents in both $\mathcal{T}$ and $\mathcal{T}'$, and at least three elements that are internal in both $\mathcal{T}$ and $\mathcal{T}'$.
(ii) If a_1 is external in one of the trees, then a_2 is internal in the same tree and a_1 is internal in the other tree.
(iii) If a_n is external in one of the trees, then a_{n-1} is internal in the same tree while, in the other tree, a_n is internal and there are not exactly three internal parents one of which has a_n as its only child in $\{a_1, a_2, \ldots, a_n\}$.

Depending upon whether $\emptyset$, $\{a_1\}$, $\{a_n\}$, or $\{a_1, a_n\}$ is the subset of elements of $\{a_1, a_2, \ldots, a_n\}$ that are external in either $\mathcal{T}$ or $\mathcal{T}'$, respectively replace this chain in $\mathcal{T}$ and $\mathcal{T}'$ with the chain (a, b, c), (e_1, a, b, c), (a, b, c, e_2), or (e_1, a, b, c, e_2) as follows:

(i) In $\mathcal{T}$,
$$p_{\mathcal{T}}(e_1) \neq p_{\mathcal{T}}(a) = p_{\mathcal{T}}(b) \neq p_{\mathcal{T}}(c) \neq p_{\mathcal{T}}(e_2),$$
where e_1 (resp. e_2) is external if a_1 (resp. a_n) is external in $\mathcal{T}$, and e_1 (resp. e_2) is internal if a_1 (resp. a_n) is external in $\mathcal{T}'$.
(ii) In $\mathcal{T}'$,
$$p_{\mathcal{T}}(e_1) \neq p_{\mathcal{T}}(a) \neq p_{\mathcal{T}}(b) = p_{\mathcal{T}}(c) \neq p_{\mathcal{T}}(e_2),$$
where e_1 (resp. e_2) is external if a_1 (resp. a_n) is external in $\mathcal{T}'$, and e_1 (resp. e_2) is internal if a_1 (resp. a_n) is external in $\mathcal{T}$.

If m denotes the number of internal parents in $\mathcal{T}$ and m' denotes the number of internal parents in $\mathcal{T}'$, then respectively add the new set $\{a, b, c\}$, $\{e_1, a, b, c\}$, $\{a, b, c, e_2\}$, or $\{e_1, a, b, c, e_2\}$ to P and, calling this set S, assign it a tuple of weights in which the first coordinate w_1 is $n - |S|$, the second coordinate w_2 is m minus the number of internal parents of the resulting chain in $\mathcal{T}$, and the third coordinate w_3 is m' minus the number of internal parents of the resulting chain in $\mathcal{T}'$. Intuitively the reduction results in replacing a_1 and a_n with e_1 and e_2, respectively, if a_1 or a_n is external in either $\mathcal{T}$ or $\mathcal{T}'$, and replacing the elements of the chain that are internal in both trees with a, b, and c. Figure 3 depicts an example of the long-chain reduction, where $\mathcal{T}$ and $\mathcal{T}'$ are the trees before, and $\mathcal{S}$ and $\mathcal{S}'$ are the trees after applying the long-chain reduction. In this example, a_1 is external in $\mathcal{T}$, while a_n is external in $\mathcal{T}'$.

Short-Chain Reduction: For $n \geq 3$, let $(a_1, a_2, \ldots, a_n)$ be an n-chain of both $\mathcal{T}$ and $\mathcal{T}'$ that does not cross an element of P such that in one of the trees, say $\mathcal{T}$, this chain has exactly one parent, while in the other tree $\mathcal{T}'$ this chain has at least three internal parents. (Note that $p_{\mathcal{T}'}(a_1), \ldots, p_{\mathcal{T}'}(a_n)$ are pairwise distinct vertices in $\mathcal{T}'$ and so only a_1 or a_n may be external in $\mathcal{T}'$.) Suppose that the chain is maximal with these properties. Depending upon whether $\emptyset$, $\{a_1\}$, $\{a_n\}$, or $\{a_1, a_n\}$ is the subset of external elements of this chain in $\mathcal{T}'$, respectively replace this chain in $\mathcal{T}$ and $\mathcal{T}'$ with the chain (a, b), (e_1, a, b), (a, b, e_2), or (e_1, a, b, e_2) as follows:

(i) In $\mathcal{T}$,
$$p_{\mathcal{T}}(e_1) = p_{\mathcal{T}}(a) = p_{\mathcal{T}}(b) = p_{\mathcal{T}}(e_2).$$
(ii) In $\mathcal{T}'$,
$$p_{\mathcal{T}'}(e_1) \neq p_{\mathcal{T}'}(a) \neq p_{\mathcal{T}'}(b) \neq p_{\mathcal{T}'}(e_2),$$

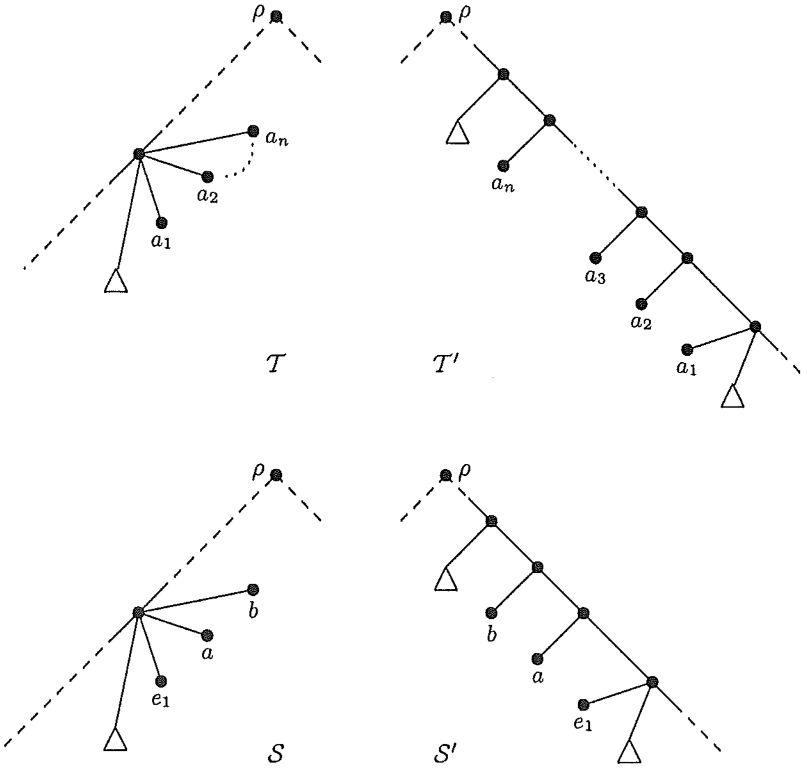

FIGURE 4. Two rooted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ reduced under the short-chain reduction, where $\mathcal{S}$ and $\mathcal{S}'$ are the resulting trees. Dotted lines indicate regions of the chain $(a_1, a_2, \ldots, a_n)$.

where e_1 (resp. e_2) is external if a_1 (resp. a_n) is external in $\mathcal{T}'$.

Furthermore, add the new set $\{a, b\}$, $\{e_1, a, b\}$, $\{a, b, e_2\}$, or $\{e_1, a, b, e_2\}$ to P and, calling this set S, assign it weight $n - |S|$. Intuitively, the reduction results in replacing a_1 and a_n with e_1 and e_2, respectively, if either a_1 or a_n is external in $\mathcal{T}'$ and, relative to $\mathcal{T}'$, replacing the internal elements with a and b. Figure 4 depicts an example of the short-chain reduction, where $\mathcal{T}$ and $\mathcal{T}'$ are the trees before, and $\mathcal{S}$ and $\mathcal{S}'$ are the trees after applying the short-chain reduction. Here a_1 is external in $\mathcal{T}'$, but a_n is internal in $\mathcal{T}'$, and so the chain $(a_1, a_2, \ldots, a_n)$ is replaced with the chain (e_1, a, b).

An agreement forest $\mathcal{F}$ for a pair of weighted rooted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is *legitimate* if $\mathcal{F}$ is acyclic and satisfies the following property, where, depending on the set in P, the elements e_1 and e_2 may or may not exist:

(P): If $\{e_1, a, b, c, e_2\} \in P$, then exactly one of the following holds:
(i) $\{e_1, a, b, c, e_2\}$ is a subset of a label set in $\mathcal{F}$,

(ii) $\{a\}$, $\{b\}$, and $\{c\}$ are label sets in $\mathcal{F}$, and e_1 and e_2 are in separate label sets in $\mathcal{F}$,

(iii) $\{a, b\}$ and $\{c\}$ are label sets in $\mathcal{F}$, and e_1 and e_2 are in separate label sets in $\mathcal{F}$,

(iv) $\{a\}$ and $\{b, c\}$ are label sets in $\mathcal{F}$, and e_1 and e_2 are in separate label sets in $\mathcal{F}$,

while if $\{e_1, a, b, e_2\} \in P$, then exactly one of the following holds:

(I) $\{e_1, a, b, e_2\}$ is a subset of a label set in $\mathcal{F}$,

(II) $\{a\}$ and $\{b\}$ are label sets in $\mathcal{F}$, and e_1 and e_2 are in separate label sets in $\mathcal{F}$.

Furthermore, referring to property (P), for an arbitrary agreement forest of $\mathcal{T}$ and $\mathcal{T}'$, we define the *weight* of $\mathcal{F}$, denoted by $w(\mathcal{F})$, to be

$$\begin{aligned} w(\mathcal{F}) = |\mathcal{F}| - 1 \quad &+ \sum_{S=\{e_1,a,b,c,e_2\}\in P; S \text{ satisfies (i) in } \mathcal{F}} w_1(S) \\ &+ \sum_{S=\{e_1,a,b,c,e_2\}\in P; S \text{ satisfies (ii) in } \mathcal{F}} w_2(S) \\ &+ \sum_{S=\{e_1,a,b,c,e_2\}\in P; S \text{ satisfies (iii) in } \mathcal{F}} w_3(S) \\ &+ \sum_{S=\{e_1,a,b,e_2\}\in P; S \text{ satisfies (I) in } \mathcal{F}} w(S). \end{aligned}$$

We denote the minimum weight of a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ by $f(\mathcal{T}, \mathcal{T}')$. Observe that $f(\mathcal{T}, \mathcal{T}') \geq h(\mathcal{T}, \mathcal{T}')$ as the weightings are non-negative, and $f(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}')$ whenever P is empty.

The next three results are key lemmas in proving that HYBRIDIZATION NUMBER is fixed-parameter tractable. Each lemma describes how particular common configurations in $\mathcal{T}$ and $\mathcal{T}'$ behave in a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight. For convenience in the proofs of these lemmas, we will frequently refer to the property of a forest $\mathcal{F}$ that the trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are edge-disjoint rooted subtrees of $\mathcal{T}$ as *no two label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$*.

Lemma 5.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted rooted phylogenetic X-trees. Let A be the label set of a pendant subtree common to $\mathcal{T}$ and $\mathcal{T}'$ that satisfies the properties of its namesake in the description of the subtree reduction. Then, for every legitimate-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight, A is a subset of a label set in $\mathcal{F}$.*

Proof. Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be a legitimate-agreement forest of $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight. Suppose that two subsets, $\mathcal{L}_i$ and $\mathcal{L}_j$ say, have the property that $\mathcal{L}_i \cap A$ and $\mathcal{L}_j \cap A$ are both non-empty. If $\mathcal{L}_i \subseteq A$ or $\mathcal{L}_j \subseteq A$, then it is easily checked that the partition of $X \cup \{\rho\}$ obtained from $\mathcal{F}$ by replacing $\mathcal{L}_i$ and $\mathcal{L}_j$ with $\mathcal{L}_i \cup \mathcal{L}_j$ is a legitimate-agreement forest of $\mathcal{T}$ and $\mathcal{T}'$ but with smaller weight than $\mathcal{F}$; a contradiction. Therefore we may assume that $\mathcal{L}_i \cap (X \cup \{\rho\})$ and $\mathcal{L}_j \cap (X \cup \{\rho\})$ are both non-empty. Because of this assumption, the pendant subtree with label set A cannot be obtained from either $\mathcal{T}$ or $\mathcal{T}'$ by deleting a single edge. Let e (resp. e') denote the edge of $\mathcal{T}$ (resp. $\mathcal{T}'$) that is directed into the vertex corresponding to the root of $\mathcal{T}|A$ (resp. $\mathcal{T}'|A$). Since no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$ and $\mathcal{T}'$, at most one of $\mathcal{T}(\mathcal{L}_i)$ and $\mathcal{T}(\mathcal{L}_j)$ includes e and at most one of $\mathcal{T}'(\mathcal{L}_i)$ and $\mathcal{T}'(\mathcal{L}_j)$ includes e'. Also, since $G_{\mathcal{F}}$ is acyclic, if $\mathcal{T}(\mathcal{L}_i)$ includes e, then $\mathcal{T}'(\mathcal{L}_j)$ does not include e'. Similar, conclusions hold for the other combinations including e or e'. Let $\mathcal{F}'$ be the partition of $X \cup \{\rho\}$ obtained from $\mathcal{F}$ by replacing $\mathcal{L}_i$ and $\mathcal{L}_j$ with $\mathcal{L}_i \cup \mathcal{L}_j$. Because of these last conclusions, it is clear that $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, it easily checked that $G_{\mathcal{F}'}$ is acyclic as $G_{\mathcal{F}}$ is acyclic and that $\mathcal{F}'$ satisfies (P) as $\mathcal{F}$ satisfies (P). Therefore $\mathcal{F}'$ is a legitimate-agreement forest of $\mathcal{T}$ and $\mathcal{T}'$. But $w(\mathcal{F}') < w(\mathcal{F})$; a contradiction. It now follows that A is a subset of a label set in $\mathcal{F}$, completing the proof of the lemma. □

Lemma 5.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted rooted phylogenetic X-trees. Let $(a_1, a_2, \ldots, a_n)$ be a chain that satisfies the properties of its namesake in the description of the long-chain reduction. Then, for every legitimate-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight, exactly one of the following holds:*

(i) *$\{a_1, a_2, \ldots, a_n\}$ is a subset of a label set in $\mathcal{F}$,*
(ii) *no label set in $\mathcal{F}$ contains at least two elements of the chain and, if a_i is an internal element of both $\mathcal{T}$ and $\mathcal{T}'$, then $\{a_i\}$ is a singleton in $\mathcal{F}$, or*
(iii) *for either $\mathcal{T}$ or $\mathcal{T}'$, say $\mathcal{T}$, two elements of the chain are in the same label set precisely if they have the same parent and, moreover, if that parent is internal in $\mathcal{T}$, then the corresponding set contains no other elements of $X \cup \{\rho\}$.*

Proof. Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight. Let $A = \{a_1, a_2, \ldots, a_n\}$. The proof is partitioned into two cases depending on which of the following properties, up to symmetry, is satisfied by $\mathcal{F}$:

(A) Whenever an element $a_i \in A$ is in a label set, $\mathcal{L}_i$ say, and $p_{\mathcal{T}}(a_1)$ is an ancestor of all elements in $\mathcal{L}_i - A$ in $\mathcal{T}$, then $p_{\mathcal{T}'}(a_1)$ is an ancestor of all elements in $\mathcal{L}_i - A$ in $\mathcal{T}'$.
(B) There is a label set, $\mathcal{L}_i$ say, in $\mathcal{F}$ with both $\mathcal{L}_i \cap A$ and $\mathcal{L}_i - A$ non-empty and such that, in $\mathcal{T}$, the vertex $p_{\mathcal{T}}(a_1)$ is an ancestor of all elements in $\mathcal{L}_i - A$, but, in $\mathcal{T}'$, the vertex $p_{\mathcal{T}'}(a_1)$ is not an ancestor of all elements in $\mathcal{L}_i - A$.

First consider (A). Let J index the label sets of $\mathcal{F}$ that contain elements of the chain. More precisely,

$$J = \{j \in \{\rho, 1, 2, \ldots, k\} : \mathcal{L}_j \cap \{a_1, a_2, \ldots, a_n\} \neq \emptyset\}.$$

Relative to the chain $(a_1, a_2, \ldots, a_n)$, we will call an edge of $\mathcal{T}$ or $\mathcal{T}'$ a *non-pendant chain edge* if the edge is non-pendant and incident with an internal parent in $\mathcal{T}$ or $\mathcal{T}'$, respectively. The analysis of (A) is partitioned into two subcases:

(I) There exists (not necessarily distinct) label sets $\mathcal{L}_i$ and $\mathcal{L}_{i'}$ in $\mathcal{F}$ such that $\mathcal{T}(\mathcal{L}_i)$ and $\mathcal{T}'(\mathcal{L}_{i'})$ contain a non-pendant edge of the chain $(a_1, a_2, \ldots, a_n)$ in $\mathcal{T}$ and $\mathcal{T}'$, respectively.
(II) $\mathcal{F}$ contains no such label sets $\mathcal{L}_i$ and $\mathcal{L}_{i'}$.

For (I), we may assume without loss of generality that $\mathcal{L}_i$ and $\mathcal{L}_{i'}$ are chosen so that the roots of $\mathcal{T}(\mathcal{L}_i)$ and $\mathcal{T}'(\mathcal{L}_{i'})$ are as close to ρ as possible in $\mathcal{T}$ and $\mathcal{T}'$. If neither $\mathcal{L}_i$ nor $\mathcal{L}_{i'}$ contains an element of A, then it is easily seen that $\mathcal{F}$ satisfies (ii) in the statement of the lemma. Thus, we may assume that either $\mathcal{L}_i$ or $\mathcal{L}_{i'}$, say $\mathcal{L}_i$, contains an element of A. If $\mathcal{L}_{i'}$ does not contain an element of A, then one of the following holds: (a) for some $a_j, a_{j'} \in (\mathcal{L}_i \cap A)$, we have $p_{\mathcal{T}}(a_j) \neq p_{\mathcal{T}}(a_{j'})$ but $p_{\mathcal{T}'}(a_j) = p_{\mathcal{T}'}(a_{j'})$; (b) $a_1 \in \mathcal{L}_i$, $a_n \notin \mathcal{L}_i$, and a_1 is an external element of the chain in $\mathcal{T}'$; or (c) $a_n \in \mathcal{L}_i$, $a_1 \notin \mathcal{L}_i$, and a_n is an external element of the chain in $\mathcal{T}'$. Since $\mathcal{L}_{i'}$ does not contain an element of A, it follows that if a label set in $\mathcal{F}$ contains an element in A and an element in $(X \cup \{\rho\}) - A$, then that label set contains either a_1 or a_n, in which case a_1 or a_n are external in $\mathcal{T}'$, respectively, but no other elements from A. Furthermore, no label set in $\mathcal{F}$ contains two elements of A that have different parents in $\mathcal{T}'$. It is now easily checked that, as $\mathcal{F}$ is a legitimate-agreement forest of minimum weight, $\mathcal{F}$ satisfies (iii) if (a) or (b) holds and $\mathcal{F}$ satisfies either (ii) or (iii) if (c) holds. In all cases, if (iii) holds, then $\mathcal{T}'$ is the distinguished tree.

Now assume that $\mathcal{L}_{i'}$ contains an element of A. The rest of the analysis for (I) is in two parts. Let $\mathcal{L}'_i$ (resp. $\mathcal{L}'_{i'}$) denote the subset of elements in $\mathcal{L}_i - A$ (resp. $\mathcal{L}_{i'} - A$) that are descendants of $p_{\mathcal{T}}(a_1)$ (resp. $p_{\mathcal{T}'}(a_1)$), and let X'_1 (resp. X_1) denote the subset of elements in $\mathcal{L}_i - A$ (resp. $\mathcal{L}_{i'} - A$) that are descendants of $p_{\mathcal{T}'}(a_1)$ in $\mathcal{T}'$ (resp. $p_{\mathcal{T}}(a_1)$ in $\mathcal{T}$).

For the first part, suppose that $\mathcal{L}'_i = X'_1$ and $\mathcal{L}'_{i'} = X_1$. Let $\mathcal{F}'$ be the forest obtained from $\mathcal{F}$ by removing each label set $\mathcal{L}_j$ with $j \in J$ and inserting the new label set $\mathcal{L}_a = \bigcup_{j \in J} \mathcal{L}_j$. Since we are in case (A), $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. To see that $\mathcal{F}'$ is acyclic. Consider the directed graphs $G_{\mathcal{F}}$ and $G_{\mathcal{F}'}$ associated with $\mathcal{F}$ and $\mathcal{F}'$, respectively. The vertex set of $G_{\mathcal{F}'}$ is obtained from $G_{\mathcal{F}}$ by deleting the vertices $\mathcal{L}_j$ for all $j \in J$, and adding the new vertex $\mathcal{L}_a$. Also, if $\mathcal{L}_r, \mathcal{L}_s \in \mathcal{F}' - \{\mathcal{L}_a\}$, then $(\mathcal{L}_r, \mathcal{L}_s)$ is an arc in $G_{\mathcal{F}'}$ if and only if $(\mathcal{L}_r, \mathcal{L}_s)$ is an arc in $G_{\mathcal{F}}$. Without loss of generality, we may assume that a_n is internal in $\mathcal{T}$. Regarding the arcs in $G_{\mathcal{F}'}$ incident with $\mathcal{L}_a$, there are two instances to consider. First assume that $\mathcal{L}_i - A$ is non-empty and contains an element that is not a descendant of $p_{\mathcal{T}}(a_1)$ in $\mathcal{T}$. Then $\mathcal{L}_i - A$ contains an element that is not a descendant of $p_{\mathcal{T}'}(a_1)$ in $\mathcal{T}'$. Since $G_{\mathcal{F}}$ is acyclic, there is no arc from $\mathcal{L}_{i'}$ to $\mathcal{L}_i$ in $G_{\mathcal{F}}$; otherwise, $G_{\mathcal{F}}$ contains a directed 2-cycle. Therefore either the roots of $\mathcal{T}'(\mathcal{L}_i)$ and $\mathcal{T}'(\mathcal{L}_{i'})$ coincide in $\mathcal{T}'$ or the root of $\mathcal{T}'(\mathcal{L}_{i'})$ is a descendant of the root of $\mathcal{T}'(\mathcal{L}_i)$. Since the root of $\mathcal{T}(\mathcal{L}_a)$ is the same as the root of $\mathcal{T}(\mathcal{L}_i)$ in $\mathcal{T}$, it follows that if $(\mathcal{L}_r, \mathcal{L}_a)$ is an arc in $G_{\mathcal{F}'}$, then $(\mathcal{L}_r, \mathcal{L}_i)$ and $(\mathcal{L}_r, \mathcal{L}_{i'})$ are arcs in $G_{\mathcal{F}}$. Moreover, if $(\mathcal{L}_a, \mathcal{L}_r)$ is an arc in $G_{\mathcal{F}'}$, then either $(\mathcal{L}_a, \mathcal{L}_i)$ or $(\mathcal{L}_a, \mathcal{L}_{i'})$ is an arc in $G_{\mathcal{F}}$. Thus, as $G_{\mathcal{F}}$ is acyclic, $G_{\mathcal{F}'}$ is also acyclic.

Second assume that either $\mathcal{L}_i - A$ is empty or if $\mathcal{L}_i - A$ is non-empty, then it only contains elements that are descendants of $p_{\mathcal{T}}(a_1)$. Because of the first instance, we may assume that the analogous property holds for $\mathcal{L}_{i'}$ and $\mathcal{T}'$. Then the root of $\mathcal{T}(\mathcal{L}_a)$ is $p_{\mathcal{T}}(a_n)$ in $\mathcal{T}$ and the root of $\mathcal{T}'(\mathcal{L}_a)$ is $p_{\mathcal{T}'}(a_n)$ in $\mathcal{T}'$. Suppose that $G_{\mathcal{F}'}$ contains the directed cycle C. Then, as $G_{\mathcal{F}}$ is acyclic, C must contain $\mathcal{L}_a$. Let $\mathcal{L}_l$ and $\mathcal{L}_m$ denote the vertices in C that immediately precede and succeed $\mathcal{L}_a$, respectively, in this directed cycle. Observe that, except for $\mathcal{L}_a$, all other vertices in C are also vertices in $G_{\mathcal{F}}$. Thus $(\mathcal{L}_i, \mathcal{L}_m)$ is an arc in $G_{\mathcal{F}}$. But $(\mathcal{L}_l, \mathcal{L}_i)$ is also an arc in $G_{\mathcal{F}}$, implying that $G_{\mathcal{F}}$ contains a directed cycle; a contradiction. Thus $G_{\mathcal{F}'}$ is acyclic. Hence $\mathcal{F}'$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, as $(a_1, a_2, \ldots, a_n)$ does not cross P and $\mathcal{F}$ satisfies (P), it is straightforward to check that $\mathcal{F}'$ satisfies (P). Thus if $|J| \geq 2$, then $w(\mathcal{F}') < w(\mathcal{F})$, contradicting the minimality of $\mathcal{F}$. Therefore A is a subset of a label set in $\mathcal{F}$ and so $\mathcal{F}$ satisfies (i) in the statement of the lemma.

For the second part, suppose that either $\mathcal{L}'_i \neq X'_1$ or $\mathcal{L}'_{i'} \neq X_1$. Without loss of generality, we may assume that $\mathcal{L}'_i \neq X'_1$ and $a_i \in \mathcal{L}_i \cap A$. Since we are in case (A), this implies that $p_{\mathcal{T}}(a_1)$ is an ancestor of at least one element in $\mathcal{L}_i - A$, but it is not an ancestor of all elements in $\mathcal{L}_i - A$. Let $\mathcal{L}''_i$ denote the subset of elements in $\mathcal{L}_i - A$ that are not descendants of $p_{\mathcal{T}}(a_1)$. Because we are in case (A), $X'_1 \neq \mathcal{L}_i - A$, and so there is an element in $\mathcal{L}_i - A$ that is not a descendant of $p_{\mathcal{T}'}(a_1)$ in $\mathcal{T}'$.

First assume that either a_i is internal in both $\mathcal{T}$ or $\mathcal{T}'$, or $a_i = a_1$. If $(\mathcal{L}_i - A) \cap X'_1$ is non-empty, then, as $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, it is easily seen that $\mathcal{L}'_i \subseteq X'_1$. Furthermore, if $a_i \neq a_1$ or $a_i = a_1$ and a_1 is internal in $\mathcal{T}'$, then the same reasoning implies that $X'_1 \cap \mathcal{L}''_i$ is empty. But then $X'_1 = \mathcal{L}'_i$; a contradiction. Therefore assume that $a_i = a_1$ and a_1 is external in $\mathcal{T}'$. If $a_n \notin \mathcal{L}_i$, then, as $\mathcal{F}$ is a legitimate-agreement forest of minimum weight, $\mathcal{F}$ satisfies (ii) in the statement of the lemma. So assume that $a_n \in \mathcal{L}_i$. If a_n is internal in $\mathcal{T}$, then, as $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, another check shows that $X'_1 \cap \mathcal{L}''_i$ is empty and so $X'_1 = \mathcal{L}'_i$. So now assume that a_n is external in $\mathcal{T}$, and therefore internal in $\mathcal{T}'$. Again as $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, it is straightforward to check that, for any two elements in $\mathcal{L}''_i \cap X'_1$ the path in $\mathcal{T}$ from each of these elements to ρ meets the path from a_n to ρ in exactly one place. With this in hand, let $\mathcal{F}'$ be the partition of $X \cup \{\rho\}$ obtained from $\mathcal{F}$ by removing each label set $\mathcal{L}_j$ with $j \in J$ and inserting the new label sets $\bigcup_{j \in J} \mathcal{L}_j - (\mathcal{L}''_i \cap X'_1)$ and $\mathcal{L}''_i \cap X'_1$. Clearly, $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and it is easily checked that, as $\mathcal{F}$ is acyclic, $\mathcal{F}'$ is acyclic. Furthermore, as $(a_1, a_2, \ldots, a_n)$ does not cross P and $\mathcal{F}$ satisfies (P), $\mathcal{F}'$ satisfies (P). Thus $\mathcal{F}$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. But, in $\mathcal{F}$, each of the elements of the chain that are internal in both $\mathcal{T}$ and $\mathcal{T}'$ are singletons. Since there are at three such elements, $w(\mathcal{F}') < w(\mathcal{F})$; a contradiction.

We may now assume $(\mathcal{L}_i - A) \cap X'_1$ is empty. As $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, it is easily seen that, for any two elements in $\mathcal{L}'_i$ the path in $\mathcal{T}'$ from each of these elements to ρ meets the path from a_n to ρ in exactly one place. If a_1 is external in $\mathcal{T}$, not in $\mathcal{L}_i$ and its label set contains elements in $(X \cup \{\rho\}) - A$, then, as we are in case (A), $p_{\mathcal{T}}(a_1)$ and $p_{\mathcal{T}'}(a_1)$ are ancestors of each of the elements in this label set. The same reasoning also shows that if a_n is external in $\mathcal{T}$ and not in $\mathcal{L}_i$, then its label set contains no elements in $(X \cup \{\rho\}) - A$. Furthermore, if a_j and a_k are internal elements of both $\mathcal{T}$ and $\mathcal{T}'$, then, as $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, the label set containing a_j is a subset of A if $p_{\mathcal{T}}(a_j) \neq p_{\mathcal{T}}(a_i)$. Also, as no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$, the elements a_j and a_k are in separate label sets in $\mathcal{F}$ if $p_{\mathcal{T}}(a_j) \neq p_{\mathcal{T}}(a_k)$. Thus there are two such subsets of A in $\mathcal{F}$. Now let $\mathcal{F}'$ be the partition of $X \cup \{\rho\}$ obtained from $\mathcal{F}$ by removing each label set $\mathcal{L}_j$ with $j \in J$ and inserting the new label sets $\bigcup_{j \in J} \mathcal{L}_j - \mathcal{L}'_i$

and $\mathcal{L}'_i$. It is clear that $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Moreover, it is easily seen that, as $\mathcal{F}$ is acyclic, $\mathcal{F}'$ is acyclic and that, as $(a_1, a_2, \ldots, a_n)$ does not cross P and $\mathcal{F}$ satisfies (P), $\mathcal{F}'$ satisfies (P). But $w(\mathcal{F}') < w(\mathcal{F})$; a contradiction.

It now follows that we may assume $\mathcal{L}_i \cap A = \{a_n\}$, where a_n is external in either $\mathcal{T}$ or $\mathcal{T}'$. By considering $\mathcal{T}$ it is easily seen that if a_j and a_k are internal elements in $\mathcal{T}$, then the label set in $\mathcal{F}$ containing a_j is a subset of A, and a_j and a_k can only be in the same label set in $\mathcal{F}$ if they have the same parent in $\mathcal{T}$. Now consider $\mathcal{T}'$. If $p_{\mathcal{T}'}(a_1)$ is an ancestor of an element in $\mathcal{L}_i$, then $\mathcal{F}$ satisfies (ii) in the statement of the lemma. Therefore assume that $p_{\mathcal{T}'}(a_1)$ is not an ancestor of any element in $\mathcal{L}_i$. Now $\mathcal{L}_{i'}$ contains an element of A and $\mathcal{T}'(\mathcal{L}_{i'})$ contains a non-pendant edge of $(a_1, a_2, \ldots, a_n)$. If $a_1 \in \mathcal{L}_{i'}$ and $\mathcal{L}_{i'}$ contains an element in $(X \cup \{\rho\}) - A$ that is not a descendant of $p_{\mathcal{T}'}(a_1)$ in $\mathcal{T}'$, then again $\mathcal{F}$ satisfies (ii) in the lemma. Noting that the label set containing a_1 can only contain another element of A if a_1 is internal in $\mathcal{T}$, it is now easily seen that, as $\mathcal{F}$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight, then $\mathcal{F}$ satisfies (iii) in the statement of the lemma with $\mathcal{T}$ as the distinguished tree. This completes the analysis of the second part, and therefore (I).

Now consider (II). We may assume that for one of the trees, say $\mathcal{T}$, whenever a label set $\mathcal{L}_r$ in $\mathcal{F}$ contains an element in A, then, unless this element is external, $\mathcal{L}_r \subseteq A$ and all elements in $\mathcal{L}_r$ have the same parent in $\mathcal{T}$. If $\mathcal{F}$ satisfies (ii) in the statement of the lemma, then we are done; so assume that this is not the case. Then there is a label set, $\mathcal{L}_i$ say, in $\mathcal{F}$ that contains at least two elements in A. In $\mathcal{T}'$, these elements have different parents. Since $\mathcal{F}$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight, it is now easily checked that $\mathcal{F}$ satisfies (iii) in the statement of the lemma. This completes the analysis of (II) and, therefore, (A)

Now suppose that $\mathcal{F}$ satisfies (B). First note that, since $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, $p_{\mathcal{T}'}(a_1)$ is not an ancestor of any element in $\mathcal{L}_i - A$ in $\mathcal{T}'$ unless $\mathcal{L}_i \cap A = \{a_1\}$ and a_1 is external in $\mathcal{T}$ or $\mathcal{L}_i \cap A = \{a_n\}$ and a_n is external in $\mathcal{T}'$. The analysis of this case is separated into two subcases:

(I) $\mathcal{L}_i \cap A$ contains an element that is internal in both $\mathcal{T}$ and $\mathcal{T}'$.
(II) $\mathcal{L}_i \cap A$ contains no element that is internal in both $\mathcal{T}$ and $\mathcal{T}'$.

For (I), let a_i be an element of $\mathcal{L}_i \cap A$ that is internal in both $\mathcal{T}$ and $\mathcal{T}'$. Let a_j be an element of A that is internal in both $\mathcal{T}$ and $\mathcal{T}'$. If $p_{\mathcal{T}}(a_j) \neq p_{\mathcal{T}}(a_i)$, then using the facts that no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$ or $\mathcal{T}'$, that $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, and that $\mathcal{F}$ is acyclic, it is easily checked that a_j is in a label set of $\mathcal{F}$ containing only elements of A and all of the elements in this set have the same parent in $\mathcal{T}$. Because of the requirement on internal parents in (iii) in the definition of the long-chain reduction, there are at least two such label sets. Also, if $p_{\mathcal{T}}(a_j) = p_{\mathcal{T}}(a_i)$ for some $j \neq i$ and $a_j \notin \mathcal{L}_i$, then, because $\mathcal{F}$ is acyclic and no label sets in $\mathcal{F}$ edge-overlap, a_j is in a label set of $\mathcal{F}$ containing only elements of A and all of the elements in this set have the same parent. Furthermore, since $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, any two distinct elements in $\mathcal{L}_i - A$ intersect the path from a_n to ρ in $\mathcal{T}'$ in exactly one place.

We next consider a_1 (resp. a_n) if a_1 (resp. a_n) is external in either $\mathcal{T}$ or $\mathcal{T}'$. If a_1 is external in $\mathcal{T}$, then, as $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, $a_1 \notin \mathcal{L}_i$. Furthermore, a_1 is in a label set of $\mathcal{F}$ that contains no other elements of A and, moreover, both $p_{\mathcal{T}}(a_1)$ and $p_{\mathcal{T}'}(a_1)$ are ancestors of all elements in this label set. If a_1 is external in $\mathcal{T}'$, then it easily checked that a_1 behaves in the same way as elements in A that are internal in both $\mathcal{T}$ and $\mathcal{T}'$. Now consider a_n. If a_n is external in $\mathcal{T}$, then, as $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, $a_n \notin \mathcal{L}_i$. Also, as $\mathcal{F}$ is acyclic, a_n is in a label set of $\mathcal{F}$ that contains no other elements of A and, moreover, $p_{\mathcal{T}}(a_n)$ is an ancestor of all elements in this label set, but $p_{\mathcal{T}}(a_1)$ is an ancestor of none. Furthermore, except for a_n, the vertex $p_{\mathcal{T}'}(a_1)$ is an ancestor of all elements in this set. Now assume that a_n is external in $\mathcal{T}'$. If $a_n \notin \mathcal{L}_i$, then, as no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}'$, the element a_n is the only element of A in its label set and, if this label set contains elements in $(X \cup \{\rho\}) - A$, then $p_{\mathcal{T}'}(a_1)$ is not an ancestor of any of these elements and all elements in $\mathcal{L}_i$ are descendants of $p_{\mathcal{T}'}(a_n)$.

With the above conclusions in hand and noting that it is possible for a_n to be external in $\mathcal{T}'$ and $a_n \in \mathcal{L}_i$, let J index the label sets of $\mathcal{F}$ that contain elements of the chain. Let $\mathcal{F}'$ be the forest obtained from $\mathcal{F}$ by removing each label set $\mathcal{L}_j$ with $j \in J$ and inserting the new label sets

$$\bigcup_{j\in J} \mathcal{L}_j - (\mathcal{L}_i - A) - (\mathcal{L}_n - \{a_n\}),$$

$\mathcal{L}'_i = \mathcal{L}_i - A$, and $\mathcal{L}'_n = \mathcal{L}_n - A$ if a_n is external in $\mathcal{T}$, where $\mathcal{L}_n$ is the label set in $\mathcal{F}$ containing a_n, and

$$\bigcup_{j\in J} \mathcal{L}_j - (\mathcal{L}_i - A)$$

and $\mathcal{L}'_i = \mathcal{L}_i - A$ if a_n is external in $\mathcal{T}'$. Note that $\mathcal{F}'$ is a partition of $X \cup \{\rho\}$. By considering the possibilities for a_1 and a_n, and noting that $p_{\mathcal{T}'}(a_1)$ is not an ancestor of any element in $\mathcal{L}_i - A$, it is clear that $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Using arguments similar to that used in (A), a straightforward check shows that, as $\mathcal{F}$ is acyclic, $\mathcal{F}'$ is acyclic. Since $(a_1, a_2, \ldots, a_n)$ does not cross P and $\mathcal{F}$ satisfies (P), $\mathcal{F}'$ satisfies (P). Therefore $\mathcal{F}'$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. But, as there are at least two label sets in $\mathcal{F}$ containing just elements of A, we have $w(\mathcal{F}') < w(\mathcal{F})$; contradicting the minimality of $\mathcal{F}$. Thus subcase (I) does not arise.

For the analysis of (II), first observe that $\mathcal{L}_i \cap A$ is a non-empty subset of $\{a_1, a_n\}$ and each of the elements in $\mathcal{L}_i \cap A$ is external in either $\mathcal{T}$ or $\mathcal{T}'$. Let $a_j, a_k \in A$ such that neither a_j nor a_k is a_1 if a_1 is external in either $\mathcal{T}$ or $\mathcal{T}'$ and neither a_j nor a_k is a_n if a_n is external in either $\mathcal{T}$ or $\mathcal{T}'$. Assume first that $a_1 \in \mathcal{L}_i$. Since $\mathcal{F}$ is acyclic and no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$ or $\mathcal{T}'$, it is easily checked that a_j and a_k are in separate label sets in $\mathcal{F}$ and none of these label sets contain elements in $(X \cup \{\rho\}) - A$. Arguing similarly, if a_n is external in $\mathcal{T}$, and therefore internal in $\mathcal{T}'$, then $\{a_n\}$ is a label set in $\mathcal{F}$. It now follows that if a_n is not external in $\mathcal{T}'$, then $\mathcal{F}$ satisfies (ii) in the statement of the lemma. Therefore, assume that a_n is external in $\mathcal{T}'$. If $a_n \notin \mathcal{L}_i$, then, as no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}'$, the elements a_j and a_n are not in the same label set in $\mathcal{F}$ for all j. Thus $\mathcal{F}$ again satisfies (ii) in the statement of the lemma, so assume that $a_n \in \mathcal{L}_i$. Since $\mathcal{T}|\mathcal{L}_i$

and $\mathcal{T}'|\mathcal{L}_i$ have a common binary refinement, $p_{\mathcal{T}'}(a_n)$ is an ancestor of all elements in $\mathcal{L}_i$. Let $\mathcal{F}'$ be the partition of $X \cup \{\rho\}$ that is obtained from $\mathcal{F}$ by replacing $\mathcal{L}_i$ and all other label sets containing elements of A with the three sets $\mathcal{L}_i'$, $\mathcal{L}_i''$, and A, where $\mathcal{L}_i''$ contains precisely the elements in $\mathcal{L}_i - A$ that are descendants of $p_{\mathcal{T}'}(a_1)$ in $\mathcal{T}'$ and $\mathcal{L}_i' = \mathcal{L}_i - (A \cup \mathcal{L}_i'')$. Clearly, $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, using arguments similar to that used in (A), it is easily checked that, as $\mathcal{F}$ is acyclic, $\mathcal{F}'$ is acyclic. Since $(a_1, a_2, \ldots, a_n)$ does cross P and $\mathcal{F}$ satisfies (P), $\mathcal{F}'$ satisfies (P) and so $\mathcal{F}'$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. But $\mathcal{F}$ has the property that $\{a_j\} \in \mathcal{F}$ for all $a_j \in A - \{a_1, a_n\}$. Since $|A| \geq 5$, this implies that $w(\mathcal{F}) < w(\mathcal{F}')$; a contradiction.

We may now assume that $a_n \in \mathcal{L}_i$ and $a_1 \notin \mathcal{L}_i$. First note that if $p_{\mathcal{T}'}(a_1)$ is an ancestor of an element in $\mathcal{L}_i$, then, as the label sets in $\mathcal{F}$ are edge-disjoint, $\mathcal{F}$ satisfies (ii) in the statement of the lemma. Thus we may also assume that $p_{\mathcal{T}'}(a_1)$ is not an ancestor of any element in $\mathcal{L}_i$. Since no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$, it follows that if $p_{\mathcal{T}}(a_j) \neq p_{\mathcal{T}}(a_k)$ or $p_{\mathcal{T}}(a_1) \neq p_{\mathcal{T}}(a_j)$, then a_j and a_k, and a_1 and a_j are in separate label sets in $\mathcal{F}$, respectively. Furthermore, unless $p_{\mathcal{T}}(a_j) = p_{\mathcal{T}}(a_n)$ and a_n is external in $\mathcal{T}'$, the label set containing a_j does not contain an element of $(X \cup \{\rho\}) - A$. Also, if a_1 is internal in $\mathcal{T}$, then its label set does not contain an element of $(X \cup \{\rho\}) - A$. It is now easily seen that if a_n is external in $\mathcal{T}$, then, as a_n is internal in $\mathcal{T}'$ and $\mathcal{F}$ is a legitimate-agreement forest of minimum weight, $\mathcal{F}$ satisfies (iii) in the statement of the lemma with $\mathcal{T}$ as the distinguished tree. Therefore, assume that a_n is external in $\mathcal{T}'$.

If a_1 is external in $\mathcal{T}$ and its label set contains an element in $(X \cup \{\rho\}) - A$ that is not an ancestor of $p_{\mathcal{T}'}(a_1)$, then $\mathcal{F}$ satisfies (ii) in the lemma. Thus if the label set containing a_1 contains an element in $(X \cup \{\rho\}) - A$, we may assume that it is a descendant of $p_{\mathcal{T}'}(a_1)$.

Now, apart from $\mathcal{L}_i$ and the label set containing a_1 if a_1 is external in $\mathcal{T}$, the only other possible label set, $\mathcal{L}_k$ say, in $\mathcal{F}$ that has a non-empty intersection with A and $(X \cup \{\rho\}) - A$ has the property that if $a_k \in \mathcal{L}_k \cap A$, then $p_{\mathcal{T}}(a_k) = p_{\mathcal{T}}(a_n)$. If no label set in $\mathcal{F}$ contains at least two elements of A each having a different parent in $\mathcal{T}'$ and there exists no such label set $\mathcal{L}_k$, then $\mathcal{F}$ satisfies (ii) in the statement of the lemma. Therefore, suppose that one of these two possibilities occur. Let $\mathcal{F}'$ be the partition of $X \cup \{\rho\}$ obtained from $\mathcal{F}$ by replacing $\mathcal{L}_i$, $\mathcal{L}_k$ if such a label set exists, and all other label sets containing elements in A with the sets $\mathcal{L}_i'$, $A \cup \mathcal{L}_1 \cup \mathcal{L}_k'$ and $\mathcal{L}_k''$, where $\mathcal{L}_i' = \mathcal{L}_i - \{a_n\}$, $\mathcal{L}_1$ is the label set of $\mathcal{F}$ containing a_1 if a_1 is external in $\mathcal{T}$, $\mathcal{L}_k''$ contains precisely the elements in $\mathcal{L}_k - A$ that are descendants of $p_{\mathcal{T}'}(a_1)$, and $\mathcal{L}_k' = \mathcal{L}_k - \mathcal{L}_k''$. Note that, as no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}$ or $\mathcal{T}'$, either $\mathcal{L}_1 - \{a_1\}$ or $\mathcal{L}_k''$ is empty. Clearly, $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, using the fact that one of the two above possibilities occur, it is easily checked that, as $\mathcal{F}$ is acyclic, $\mathcal{F}'$ is acyclic. Moreover, as $(a_1, a_2, \ldots, a_n)$ does not cross P and $\mathcal{F}$ satisfies (P), $\mathcal{F}'$ satisfies (P) and so $\mathcal{F}'$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. But $w(\mathcal{F}') < w(\mathcal{F})$ as $\mathcal{T}$ has at least three internal parents. This contradiction completes the proof of (B) and hence the lemma.

Lemma 5.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted rooted phylogenetic X-trees. Let $(a_1, a_2, \ldots, a_n)$ be a chain that satisfies the properties of its namesake in the description of the short-chain reduction. Then, for every legitimate-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight, exactly one of the following holds:*

(i) *$\{a_1, a_2, \ldots, a_n\}$ is a subset of a label set in $\mathcal{F}$, or*
(ii) *no label set in $\mathcal{F}$ contains at least two elements of the chain and, if a_i is an internal element of $(a_1, a_2, \ldots, a_n)$ in $\mathcal{T}'$, then $\{a_i\}$ is a singleton in $\mathcal{F}$.*

Proof. Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight, and let $A = \{a_1, a_2, \ldots, a_n\}$. Let J index the label sets of $\mathcal{F}$ that contain elements of A and let $\mathcal{L}_a = \bigcup_{j \in J} \mathcal{L}_j$. Suppose that neither (i) nor (ii) hold for $\mathcal{F}$. If no label set in $\mathcal{F}$ contains at least two elements of A, then, relative to $\mathcal{T}'$, there is a label set in $\mathcal{F}$ that contains an internal element of the chain as well as an element of $(X \cup \{\rho\}) - A$. By considering the structure of $(a_1, a_2, \ldots, a_n)$ in $\mathcal{T}'$, it is easily seen that, as $(a_1, a_2, \ldots, a_n)$ has at least three internal elements relative to $\mathcal{T}'$, at least one of these internal elements is a singleton in $\mathcal{F}$. A routine check shows that, apart from one exceptional case, we can replace such a singleton and a label set in $\mathcal{F}$ that contains an internal element of the chain in $\mathcal{T}'$ as well as an element of $(X \cup \{\rho\}) - A$ with the union of these two sets to obtain a legitimate-agreement forest of $\mathcal{T}$ and $\mathcal{T}'$ that has smaller weight then $\mathcal{F}$; a contradiction. In the exceptional case, there is exactly one label set, $\mathcal{L}_i$ say, in $\mathcal{F}$ that contains an internal element of the chain in $\mathcal{T}'$ and an element in $(X \cup \{\rho\}) - A$, and this set has the properties that $|\mathcal{L}_i \cap A| = 1$, and $p_{\mathcal{T}'}(a_1)$ is an ancestor of all the elements in $\mathcal{L}_i - A$, but $p_{\mathcal{T}}(a_1)$ is not an ancestor of all the elements in $\mathcal{L}_i$. Since $\mathcal{F}$ is acyclic, it follows that each of the remaining internal elements of the chain in $\mathcal{T}'$ are singletons in $\mathcal{F}$. A straightforward check now shows that

$$\{\mathcal{L} - A : \mathcal{L} \in \mathcal{F}\} \cup A$$

is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, but with smaller weight than $\mathcal{F}$. This contradiction implies that there is a label set in $\mathcal{F}$ containing at least two elements of A. Without loss of generality, we may assume that this set is $\mathcal{L}_i$ and that $a_i \in \mathcal{L}_i \cap A$, where $i > i'$ for all $a_{i'} \in \mathcal{L}_i \cap A$.

Suppose that there exists an $\mathcal{L}_h \in \mathcal{F} - \{\mathcal{L}_i\}$ such that $|\mathcal{L}_h \cap A| \geq 1$, $|\mathcal{L}_h \cap ((X \cup \{\rho\}) - A)| \geq 1$, and let $a_h \in (\mathcal{L}_h \cap A)$. If $p_{\mathcal{T}'}(a_h)$ is a descendant of $p_{\mathcal{T}'}(a_i)$, then, as $|\mathcal{L}_i| \geq 2$ and no label sets in $\mathcal{F}$ edge-overlap in $\mathcal{T}'$, the vertex $p_{\mathcal{T}'}(a_h)$ in $\mathcal{T}'$ is an ancestor of all elements in $\mathcal{L}_h \cap ((X \cup \{\rho\}) - A)$. Because $\mathcal{F}$ is acyclic, it follows that the vertex $p_{\mathcal{T}}(a_h)$ in $\mathcal{T}$ is an ancestor of all elements in $\mathcal{L}_h \cap ((X \cup \{\rho\}) - A)$; otherwise $G_{\mathcal{F}}$ contains a directed 2-cycle. Now assume that $p_{\mathcal{T}'}(a_h)$ is an ancestor of $p_{\mathcal{T}}(a_i)$. If $\mathcal{L}_i$ (resp. $\mathcal{L}_h$) contains an element z that is not a descendant of $p_{\mathcal{T}'}(a_n)$ in $\mathcal{T}'$, then, as $G_{\mathcal{F}}$ is acyclic, $p_{\mathcal{T}}(a_n)$ is an ancestor of all elements in $\mathcal{L}_h$ (resp. $\mathcal{L}_i$) in $\mathcal{T}$. Now let $\mathcal{F}'$ be the forest obtained from $\mathcal{F}$ by removing each label set $\mathcal{L}_j$ with $j \in J$ and inserting the new label set $\mathcal{L}_a$. Using the outcomes of the above two possibilities, it is easily seen that $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, as $(a_1, a_2, \ldots, a_n)$ does not cross P and $\mathcal{F}$ satisfies (P), $\mathcal{F}'$ satisfies (P). Using the facts that $\mathcal{F}$ is acyclic and at least one of the label sets in $\mathcal{F}$ contains at least two elements of A, arguments similar to that used in the proof of Lemma 5.2

show that $\mathcal{F}'$ is acyclic. But then $w(\mathcal{F}') < w(\mathcal{F})$; a contradiction to the minimality of $\mathcal{F}$. Thus $\mathcal{F}$ satisfies either (i) or (ii).

6. Hybridization Number is Fixed-Parameter Tractable

In this section, we prove Theorem 1.1. We begin by showing that each of the three reductions described in the last section preserves the minimum weight of a legitimate-agreement forest.

Proposition 6.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted rooted phylogenetic X-trees. Let $\mathcal{S}$ and $\mathcal{S}'$ be the pair of weighted rooted phylogenetic X'-trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by applying the subtree, long-chain, or short-chain reduction. Then $f(\mathcal{T},\mathcal{T}') = f(\mathcal{S},\mathcal{S}')$.*

Proof. It is an immediate consequence of Lemma 5.1 that if $\mathcal{S}$ and $\mathcal{S}'$ have been obtained from $\mathcal{T}$ and $\mathcal{T}'$ by an application of the subtree reduction, then the proposition holds. We next prove the result for when $\mathcal{S}$ and $\mathcal{S}'$ have been obtained from $\mathcal{T}$ and $\mathcal{T}'$ by applying the long-chain reduction. The proof of the result for the short-chain reduction is similar and omitted.

Suppose that $(a_1, a_2, \ldots, a_n)$ is the common chain of $\mathcal{T}$ and $\mathcal{T}'$ used in this application of the long-chain reduction. Now let $\mathcal{F}_\mathcal{T}$ be a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight. Then, by Lemma 5.2 one of the following holds:

(i) $\{a_1, a_2, \ldots, a_n\}$ is a subset of a label set of $\mathcal{F}_\mathcal{T}$,
(ii) no label set in $\mathcal{F}_\mathcal{T}$ contains at least two elements of the chain and, if a_i is an internal element of both $\mathcal{T}$ and $\mathcal{T}'$, then $\{a_i\}$ is a singleton in $\mathcal{F}_\mathcal{T}$, or
(iii) for either $\mathcal{T}$ or $\mathcal{T}'$, say $\mathcal{T}$, two elements of the chain are in the same label set precisely if they have the same parent and, moreover, if that parent is internal in $\mathcal{T}$, then the corresponding set contains no other elements of $X \cup \{\rho\}$.

Let $\mathcal{F}_\mathcal{S}$ be the forest obtained from $\mathcal{F}_\mathcal{T}$ by replacing a_1 and a_n with e_1 and e_2, respectively, if a_1 or a_n is external in either $\mathcal{T}$ or $\mathcal{T}'$, and then, depending on which of (i), (ii), or (iii) holds, respectively replace the remaining elements of A as follows: replace $a_1, a_2, \ldots, a_n$ with a, b, and c; collectively replace the label sets of the form $\{a_i\}$ with $\{a\}$, $\{b\}$, and $\{c\}$; or collectively replace the label sets of the form $\{a_i, a_{i+1}, \ldots, a_j\}$ with $\{a, b\}$ and $\{c\}$ and, if there is a label set of the form $\{e_1, a_2, \ldots, a_{i'}\}$ or $\{a_{j'}, a_{j'+1}, \ldots, e_2\}$, replace it with $\{e_1\}$ or $\{e_2\}$, respectively. Since $\mathcal{F}_\mathcal{T}$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, it is easily checked that $\mathcal{F}_\mathcal{S}$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$. In the case that (ii) holds, the contribution of the singletons containing elements that are internal in both $\mathcal{T}$ and $\mathcal{T}'$ to $w(\mathcal{F}_\mathcal{T})$ is exactly the same as the contribution of $\{a\}$, $\{b\}$, and $\{c\}$ to $w(\mathcal{F}_\mathcal{S})$. Furthermore, in the case that (iii) holds, the contribution of the label sets containing just internal elements of A in $\mathcal{T}$ to $\mathcal{F}_\mathcal{T}$ is equal to the contribution of $\{a, b\}$, $\{c\}$, and $\{e_1\}$ and $\{e_2\}$ if either e_1 or e_2 are internal elements of the reduced chain in $\mathcal{S}$ respectively, to $\mathcal{F}_\mathcal{S}$. Thus $w(\mathcal{F}_\mathcal{S}) = w(\mathcal{F}_\mathcal{T})$, and so $f(\mathcal{S},\mathcal{S}') \leq f(\mathcal{T},\mathcal{T}')$.

Now suppose that $\mathcal{F}_S$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ of minimum weight. As $\mathcal{F}_S$ is legitimate, one of the following holds, where e_1 and e_2 may or may not exist depending on whether a_1 or a_n is external in either $\mathcal{T}$ or $\mathcal{T}'$:

(i) $\{e_1, a, b, c, e_2\}$ is contained in a label set, $\mathcal{L}$ say, in $\mathcal{F}_S$,
(ii) $\{a\}$, $\{b\}$, and $\{c\}$ are label sets in $\mathcal{F}_S$, and e_1 and e_2 are in separate label sets in $\mathcal{F}_S$,
(iii) $\{a, b\}$ and $\{c\}$ are label sets in $\mathcal{F}_S$, and e_1 and e_2 are in separate label sets in $\mathcal{F}_S$, or
(iv) $\{a\}$ and $\{b, c\}$ are label sets in $\mathcal{F}_S$, and e_1 and e_2 are in separate label sets in $\mathcal{F}_S$.

Let $\mathcal{F}_T$ be the forest obtained from $\mathcal{F}_S$ by replacing e_1 and e_2 with a_1 and a_n, respectively, if a_1 or a_n is external in either $\mathcal{T}$ or $\mathcal{T}'$, and then, depending on which of (i) to (iv) holds, make one of the following replacements for a, b, and c:

(i) $\mathcal{L}$ with $(\mathcal{L} - \{a, b, c\}) \cup A$,
(ii) $\{a\}$, $\{b\}$, and $\{c\}$ with the sets $\{a_i\}$, where a_i is an internal element in both $\mathcal{T}$ and $\mathcal{T}'$,
(iii) $\{a, b\}$ and $\{c\}$ with the parts of the parent partition of $(a_1, a_2, \ldots, a_n)$ induced by $\mathcal{T}$ whose corresponding parents are internal in $\mathcal{T}$, and deleting $\{a_1\}$ or $\{a_n\}$ if e_1 or e_2 is internal in $\mathcal{S}$, or
(iv) $\{a\}$ and $\{b, c\}$ with the parts of the parent partition of $(a_1, a_2, \ldots, a_n)$ induced by $\mathcal{T}'$ whose corresponding parents are internal in $\mathcal{T}'$, and deleting $\{a_1\}$ or $\{a_n\}$ if e_1 or e_2 is internal in $\mathcal{S}'$.

A routine check shows that, as $\mathcal{F}_S$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$, the collection $\mathcal{F}_T$ of sets is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. In (ii), the contribution of the singletons $\{a\}$, $\{b\}$, and $\{c\}$ to $w(\mathcal{F}_S)$ is the same as the contribution of the sets $\{a_i\}$ to $w(\mathcal{F}_T)$, where a_i is an internal element of both $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, in (iii) and analogously in (iv), the contribution of $\{a, b\}$ and $\{c\}$, and $\{e_1\}$ and $\{e_2\}$ if e_1 or e_2, respectively, are internal in $\mathcal{S}$ to $\mathcal{F}_S$ is equal to the contribution of the label sets in $\mathcal{F}$ which exclusively contain internal elements of A in $\mathcal{T}$ to $\mathcal{F}_T$. Thus $w(\mathcal{F}_T) = w(\mathcal{F}_S)$, and so $f(\mathcal{T}, \mathcal{T}') \leq f(\mathcal{S}, \mathcal{S}')$. Hence $f(\mathcal{T}, \mathcal{T}') = f(\mathcal{S}, \mathcal{S}')$, completing the proof of the proposition. □

Lemma 6.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted rooted phylogenetic X-trees, and let $(a_1, a_2, \ldots, a_n)$ be a chain of both $\mathcal{T}$ and $\mathcal{T}'$ that does not cross P and is maximal with these properties. Then, by a sequence of long- and short-chain reductions applied to this chain, the length of the resulting chain is at most* 11.

Proof. We begin by partitioning the chain into at most four smaller ordered tuples. Suppose first that there is an element of the chain that is internal in both $\mathcal{T}$ and $\mathcal{T}'$. With $i \leq j$, choose a_i and a_j as follows:

(i) If a_1 (resp. a_n) is internal in both $\mathcal{T}$ and $\mathcal{T}'$, choose a_i (resp. a_j) to be a_1 (resp. a_n).

(ii) If a_1 (resp. a_n) is external in both $\mathcal{T}$ and $\mathcal{T}'$, but a_2 (resp. a_{n-1}) is internal in both $\mathcal{T}$ and $\mathcal{T}'$, choose a_i (resp. a_j) to be a_2 (resp. a_{n-1}).

(iii) If neither (i) nor (ii) holds, then, for some $\mathcal{R} \in \{\mathcal{T}, \mathcal{T}'\}$ (resp. $\mathcal{S} \in \{\mathcal{T}, \mathcal{T}'\}$), a_1 (resp. a_n) and a_2 (resp. a_{n-1}) are external in $\mathcal{R}$ (resp. $\mathcal{S}$). In this case, choose a_i (resp. a_j) to be the element of the chain that is external in $\mathcal{R}$ and has maximum (resp. minimum) index with $a_1, a_2, \ldots, a_i$ (resp. $a_j, a_{j+1}, \ldots, a_n$) all external in $\mathcal{R}$ (resp. $\mathcal{S}$).

Having picked a_i and a_j, consider the chain $(a_i, a_{i+1}, \ldots, a_j)$. If this chain satisfies (i) and the condition on internal parents at the end of (iii) in the description of the long-chain reduction, then we can apply this reduction to get a chain with at most 5 elements. Furthermore, if $(a_1, a_2, \ldots, a_{i-1})$ is a chain with at least three internal elements in the tree in $\{\mathcal{T}, \mathcal{T}'\}$ that is not $\mathcal{R}$, then we can apply the short-chain reduction to get a chain with at most 3 elements. Lastly, if $(a_{j+1}, a_{j+2}, \ldots, a_n)$ is a chain with at least three internal elements in the tree in $\{\mathcal{T}, \mathcal{T}'\}$ that is not $\mathcal{S}$, then we can again apply the short-chain reduction to get a chain with at most 3 elements. Note that if we cannot apply the first or the second of these short-chain reductions, then $i - 1 \leq 3$ and $n - j \leq 3$, respectively. It now follows that after these three reductions, the resulting chain has length at most 11.

Now assume that $(a_i, a_{i+1}, \ldots, a_j)$ does not satisfy (i) and the condition on internal parents at the end of (iii) in the description of the long-chain reduction. Then, up to the possibility of an additional internal parent which only has a_j as its only child in $\{a_i, a_{i+1}, \ldots, a_j\}$, this chain has at most two internal parents p_1 and p_2 in either $\mathcal{T}$ or $\mathcal{T}'$. Except for the children of these two parents that are in A, all of the remaining elements of A are external in either $\mathcal{T}$ or $\mathcal{T}'$. In particular, $a_1, \ldots, a_{i-1}$ share the same parent in $\mathcal{R}$, and $a_{j+1}, \ldots, a_n$ share the same parent in $\mathcal{S}$. As $(a_1, a_2, \ldots, a_n)$ has an internal element in both $\mathcal{T}$ and $\mathcal{T}'$, these two shared parents are distinct. Applying at most four short-chain reductions, it is easily checked that the resulting chain has length at most 10.

Now suppose that no element of the chain is internal in both $\mathcal{T}$ and $\mathcal{T}'$, then each element of the chain is external in either $\mathcal{T}$ or $\mathcal{T}'$. In this case, either we apply a single application of the short-chain reduction to get a chain of length at most 4 or we apply two applications of the short-chain reduction to get a chain of length at most 6. This completes the proof of the lemma. □

Proposition 6.1 showed that the weight function is preserved under each of the three reductions. Part (i) of the next lemma shows that these reductions can be applied so that the size of the label set of the resulting rooted phylogenetic trees is a linear function of the value of this function.

Lemma 6.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees, and let P be an empty collection of subsets of X. Let $\mathcal{S}$ and $\mathcal{S}'$ be two weighted rooted phylogenetic X'-trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by repeatedly applying the subtree reduction until no further reduction is possible, and then, for each maximal chain common to both resulting trees, repeatedly applying the long-chain and short-chain reductions. Then*

(i) *$\mathcal{S}$ and $\mathcal{S}'$ have no pendant subtrees with common label set A such that $\mathcal{S}|A$ and $\mathcal{S}'|A$ have a common binary refinement and $|A| \geq 2$,*

(ii) *the length of any chain common to both $\mathcal{S}$ and $\mathcal{S}'$ is at most* 11, *and*

(iii) $|X'| < 59h(\mathcal{T}, \mathcal{T}')$.

Proof. For the proof of (i) and (ii), let $\mathcal{T}_1$ and $\mathcal{T}_1'$ be the rooted phylogenetic trees obtained from $\mathcal{T}$ and $\mathcal{T}'$ after repeatedly applying the subtree reduction until no further reduction is possible. Furthermore, observe that if $P_1, P_2 \in P$, then $\mathcal{S}(P_1)$ and $\mathcal{S}(P_2)$ are edge-disjoint, and $\mathcal{S}'(P_1)$ and $\mathcal{S}'(P_2)$ are edge-disjoint. Consider (i), and let A be such a label set. Without loss of generality, we may assume that A is maximal. Then, because of maximality, if A intersects a set in P, then that set is a subset of A. Now let A' be the set obtained from A by replacing the elements belonging to a set in P with their original counterparts. Using the above observation, it is easily seen that A' is a pendant subtree of $\mathcal{T}_1$ and $\mathcal{T}_1'$. But, as $\mathcal{S}|A$ and $\mathcal{S}'|A$ have a common binary refinement, $\mathcal{T}_1|A'$ and $\mathcal{T}_1'|A'$ have a common binary refinement; a contradiction. Thus (i) holds,

For (ii), suppose that there exists a chain common to both $\mathcal{S}$ and $\mathcal{S}'$ that has at least 12 elements. Without loss of generality, we may assume that this chain is maximal. Let A denote the label set of this common chain. Analogous to (i), because of maximality, if A intersects a set in P, then that set is a subset of A. Moreover, if this intersection involves a set that was part of a sequence of reductions to reduce a common chain in $\mathcal{T}_1$ and $\mathcal{T}_1'$, then all of the associated sets in P are subsets of A. Using Lemma 6.2 to get a contradiction, a similar argument used to establish (i) can now be used to establish (ii).

Now consider (iii). Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ of minimum weight. Let $\mathcal{B}$ and $\mathcal{B}'$ be two binary refinements of $\mathcal{S}$ and $\mathcal{S}'$, respectively, so that $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{B}$ and $\mathcal{B}'$. By Lemma 4.2, such binary refinements exist. If $\mathcal{B}$ and $\mathcal{B}'$ have a common pendant subtree with label set A and $|A| \geq 2$, then this subtree is a common binary refinement of $\mathcal{S}|A$ and $\mathcal{S}'|A$, contradicting (i). Thus $\mathcal{B}$ and $\mathcal{B}'$ have no such pendant subtree. Furthermore, if $\mathcal{B}$ and $\mathcal{B}'$ have a common chain with label set A and $|A| \geq 12$, then this implies that $\mathcal{S}$ and $\mathcal{S}'$ have such a chain, contradicting (ii). Hence any chain common to both $\mathcal{B}$ and $\mathcal{B}'$ has at most 11 elements. With these restrictions on $\mathcal{B}$ and $\mathcal{B}'$, we can now use the argument for the analogous result for binary trees in [5] to complete the proof of (iii). The only modification necessary is to replace chains of size 2 with chains of size at most 11. Making this change and working through the straightforward algebra gives the desired result. □

Proof of Theorem 1.1. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees, and let P be an empty collection of subsets of X. Let k be an integer. Let $\mathcal{S}$ and $\mathcal{S}'$ be the weighted rooted phylogenetic X'-trees obtained from $\mathcal{T}$ and $\mathcal{T}'$ by repeatedly applying the subtree reduction until no further reduction is possible, and then, for each maximal chain common to both resulting trees, repeatedly applying the long-chain and short-chain reductions. As P is empty, $h(\mathcal{T}, \mathcal{T}') = f(\mathcal{T}, \mathcal{T}')$ and so, by Proposition 6.1,

$$h(\mathcal{T}, \mathcal{T}') = f(\mathcal{T}, \mathcal{T}') = f(\mathcal{S}, \mathcal{S}').$$

It is clear that $\mathcal{S}$ and $\mathcal{S}'$ can be found in time polynomial in $|X|$, say $p(|X|)$. By Lemma 6.3(iii), $|X'| \leq 59h(\mathcal{T},\mathcal{T}')$ and so, if $|X'| > 59k$, we declare that $h(\mathcal{T},\mathcal{T}') > k$.

Now suppose that $|X'| \leq 59k$. The time taken to check whether a partition of $X' \cup \{\rho\}$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ takes time polynomial in k. Note that for deciding if two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_1'$ have a common binary refinement, one simply needs to check whether or not $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_1')$ is a hierarchy, that is, for all (edge) clusters $C_1, C_2 \in \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_1')$, the set $C_1 \cap C_2 \in \{\emptyset, C_1, C_2\}$. Furthermore, as $|X'| \leq 59k$, the number of forests with at most $k+1$ parts is bounded by a computable function in k, say $f(k)$. If one of these forests is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ with weight at most k, then we declare $h(\mathcal{T},\mathcal{T}') \leq k$; otherwise, we declare $h(\mathcal{T},\mathcal{T}') > k$. Hence we can answer the HYBRIDIZATION NUMBER decision problem for $\mathcal{T}$ and $\mathcal{T}'$ in time $O(f(k) + p(|X|))$. Thus HYBRIDIZATION NUMBER is fixed-parameter tractable. □

Remark. While one could explicitly give a function in k that bounds the number of partitions to consider in the proof of Theorem 1.1, it is unlikely to be the best theoretically and we expect in practice much better methods.

7. Concluding Remarks

We end the paper with some remarks.

1. In this paper, we reduced a chain using two types of chain reductions. However, we believe that it is possible to do this with a single type of chain reduction. The drawback of such a reduction is that the number of possibilities for a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ increase. Since the goal of the paper is to show that HYBRIDIZATION NUMBER is fixed-parameter tractable, we decided to use the two types of reductions, thereby reducing the complexity and lengths of the proofs.
2. The subtree, long-chain, and short-chain reductions are enough to kernalize HYBRIDIZATION NUMBER and yield an algorithm that is fixed-parameter tractable. These reductions extend the two reductions used to kernalize HYBRIDIZATION NUMBER when the initial two trees are both binary [5]. However, there is another type of reduction for binary trees that turns out to particularly useful. This additional reduction, called the *cluster reduction* [2], allows for an attractive divide-and-conquer approach that breaks the problem into a number of smaller and, therefore, more tractable subproblems. Details on how this reduction can easily be fitted into the framework of (arbitrary) rooted phylogenetic trees can be found in [9].

8. Acknowledgments

We thank Magnus Bordewich and Dominic Welsh for useful discussions.

References

[1] M. Baroni, S. Grünewald, V. Moulton, and C. Semple (2005). Bounding the number of hybridisation events for a consistent evolutionary history, *Journal of Mathematical Biology*, **51**:171-182.

[2] M. Baroni, C. Semple, M. Steel (2006). Hybrids in real time. *Systematic Biology*, **55**:46-56.

[3] M. Bordewich and C. Semple (2004). On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, **8**:409-423.

[4] M. Bordewich and C. Semple (2007). Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics*, **155**:914-928.

[5] M. Bordewich and C. Semple (2007). Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4**:458-466.

[6] R. Downey and M. Fellows (1998). *Parameterized Complexity.* Springer.

[7] J. Fehrer, B. Gemeinholzer, J. Chrtek Jr, S. Bräutigam (2007). Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (*Hieracium*, Cichorieae, Asteraceae). *Molecular Phylogenetics and Evolution*, **42**:347-361.

[8] R. G. Harrison (1993). Hybrids and hybrid zones: historical perspectives. In *Hybrid zones and the evolutionary process.* Oxford University Press.

[9] S. Linz (2008). Reticulation in Evolution. PhD thesis, Heinrich-Heine-Universität, Düsseldorf.

[10] W. P. Maddison (1989). Reconstructing character evolution on polytomous cladograms. *Cladistics*, **5**:365-377.

[11] J. Mallet (2005). Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, **20**:229-237.

[12] O. Paun, C. Lehnebach, J. T. Johansson, P. Lockhart, E. Hörandl (2005). Phylogenetic relationships and biogeography of *Ranunculus* and allied genera (Ranunculaceae) in the Mediterranean region and in the European Alpine System. *Taxon*, **54**:911-930.

[13] C. Semple and M. Steel (2003). *Phylogenetics.* Oxford University Press.

[14] Y. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotyde blocks: finding the minimum number of recombination events. In *Algorithms in Bioinformatics (WABI2003)*, Lecture Notes in Bioinformatics, vol. 2812, 2003, pp. 287-302.

[15] Y. Song and J. Hein (2005). Constructing minimal ancestral recombination graphs. *Journal of Computational Biology*, **12**:147-16.

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`