Computing the Hybridization Number of Two Phylogenetic Trees Is Fixed-Parameter Tractable

Magnus Bordewich and Charles Semple

Abstract—Reticulation processes in evolution mean that the ancestral history of certain groups of present-day species is non-tree-like. These processes include hybridization, lateral gene transfer, and recombination. Despite the existence of reticulation, such events are relatively rare and, so, a fundamental problem for biologists is the following: Given a collection of rooted binary phylogenetic trees on sets of species that correctly represent the tree-like evolution of different parts of their genomes, what is the smallest number of "reticulation" vertices in any network that explains the evolution of the species under consideration? It has been previously shown that this problem is NP-hard even when the collection consists of only two rooted binary phylogenetic trees. However, in this paper, we show that the problem is fixed-parameter tractable in the two-tree instance when parameterized by this smallest number of reticulation vertices.

Index Terms—Rooted phylogenetic tree, reticulate evolution, hybridization network, agreement forest, subtree prune and regraft.

1 Introduction

Evolutionary (phylogenetic) trees are used in biology to represent the ancestral history of a collection of present-day species. While this is appropriate for many groups of species, there are some groups (including certain plant and fish species) for which the ancestral history is non-tree-like. This is caused by processes that include hybridization, lateral gene transfer, and recombination. Collectively, these processes are referred to as reticulation events. For such species, it is more appropriate to represent their ancestral history using rooted acyclic digraphs, where vertices of in-degree at least two represent reticulation events

Although reticulation events do occur, they are still relatively rare and, so, a fundamental problem for biologists studying the evolution of species is the following: Given a collection of rooted phylogenetic trees on sets of species that correctly represents the tree-like evolution of different parts of their genomes, what is the smallest number of reticulation events needed to explain the evolution of the species under consideration? This smallest number sets a lower bound on the number of such events.

This question has been considered in a number of papers including [2], [3], [6], [10], [14], [15]. Furthermore, variants of it (particularly when the input is a collection of binary sequences) have also been considered, for example see [8], [9], [11], [12], [13], [18]. In an earlier paper [6], we showed that, computationally, the above problem is NP-hard even when the initial collection consists of two rooted binary phylogenetic trees. However, the main result of this paper shows that, in the case where the input consists of two such trees, there is a fixed-parameter algorithm for finding the optimal solution.

The idea behind fixed-parameter complexity is that, while the general case of computing the minimum number of reticulation events is NP-hard, many biologically relevant cases have a relatively small number of hybridization events and, so, may be tractable. In particular, we show that this minimum number can be computed in time $O(f(k) + p(n))$, where n is the number of species, k is the actual minimum number, f is some computable function, and p is a fixed polynomial. The importance of this result is in the separation of the variables n and k; it shows that, for a reasonable range of k, the problem may be tractable even for a very large n.

To formally describe the above problem and, in particular, the main result, we need several definitions. A *rooted binary phylogenetic X-tree* is a rooted tree whose root has degree two and all other interior vertices have degree three and whose leaf set is X. The set X is called the *label set* of $\mathcal{T}$ and is often denoted $\mathcal{L}(\mathcal{T})$. Two rooted binary phylogenetic trees are shown in Fig. 1a.

A *hybridization network* (on X) is a rooted acyclic digraph with root ρ in which

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

For completeness, if $|X| = 1$, then the digraph consisting of an isolated vertex labeled by the element in X is also defined to be a hybridization network on X. The set X represents a set of present-day species and vertices of in-degree at least two

- *M. Bordewich is with the Department of Computer Science, Durham University, Durham DH1 3LE, UK. E-mail: m.j.r.bordewich@durham.ac.uk.*
- *C. Semple is with the Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand. E-mail: c.semple@math.canterbury.ac.nz.*

Manuscript received 30 May 2006; revised 20 Aug. 2006; accepted 24 Aug. 2006; published online 10 Jan. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0119-0506.
Digital Object Identifier no. 10.1109/TCBB.2007.1019.

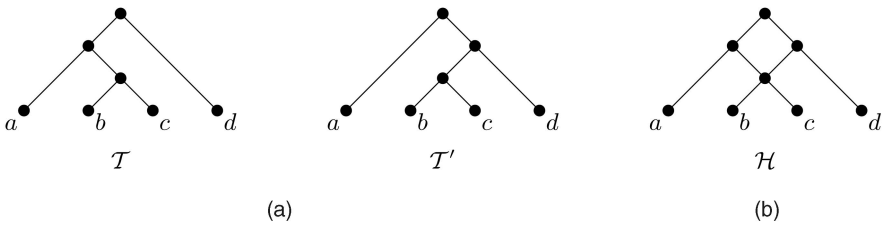


Fig. 1. (a) Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$. (b) A hybridization network $\mathcal{H}$ that displays them.

represent an inheritance of genetic information from their parents. Generically, we call such vertices *hybridization vertices*. A hybridization network is shown in Fig. 1b. For convenience, throughout the paper, we adopt the convention that hybridization networks are always drawn with their arcs directed downward and, so, we omit the arrowheads. Note that hybridization networks are referred to as "hybrid phylogenies" in [2], [3].

To quantify the number of hybridization events of a hybridization network $\mathcal{H}$, we define the *hybridization number* of $\mathcal{H}$, denoted $h(\mathcal{H})$, to be

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

where ρ denotes the root of $\mathcal{H}$ and $d^-(v)$ denotes the indegree of v. Apart from the root, every vertex has at least one parent and, so, "$(d^-(v) - 1)$" represents the number of "additional" parents of v. In Fig. 1b, $h(\mathcal{H}) = 1$.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $\mathcal{H}$ be a hybridization network. We say that $\mathcal{H}$ *displays* $\mathcal{T}$ if $\mathcal{T}$ can be obtained from a rooted subtree of $\mathcal{H}$ by suppressing degree-two vertices. In other words, $\mathcal{T}$ can be obtained from $\mathcal{H}$ by first deleting a subset of the edges of $\mathcal{H}$ and then deleting the isolated vertices and suppressing nonroot degree-two vertices. The hybridization network in Fig. 1b displays the two trees in Fig. 1a. For two rooted binary phylogenetic X-trees, $\mathcal{T}$ and $\mathcal{T}'$, we set

$$h(\mathcal{T}, \mathcal{T}') = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybridization network that displays } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

The decision problem HYBRIDIZATION NUMBER is formally stated as follows:

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

The main result of this paper is the following theorem:

Theorem 1.1. *The decision problem* HYBRIDIZATION NUMBER, *parameterized by* $h(\mathcal{T}, \mathcal{T}')$, *is fixed-parameter tractable.*

We note here that, while Theorem 1.1 provides the first fixed-parameter algorithm for HYBRIDIZATION NUMBER, Hallet and Lagergren [10] give a fixed-parameter algorithm in a slightly different setting which may be interpreted as a restricted version of this problem.

Informally, the overall approach we use in proving Theorem 1.1 is as follows: We start by taking the input to HYBRIDIZATION NUMBER and reducing its size using two reduction rules in a regulated way. We show that, once fully reduced, the resulting input size is linear in our parameter: the hybridization number of the original pair of input trees. We then apply brute force to compute the hybridization number on the smaller input, which may take exponential time but is only ever performed on the bounded size input. The resulting solution immediately provides the hybridization number of the original pair of input trees.

This approach is similar to that used in showing that "rooted subtree prune and regraft (rSPR) distance" is fixed-parameter tractable [5]; in particular, we kernalize the problem by using two rules that reduce the size of the input trees sufficiently. Loosely speaking, for two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, the rSPR distance is the minimum number of subtrees that must be "moved" to transform $\mathcal{T}$ into $\mathcal{T}'$. Denoting this distance by $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$, the decision problem rSPR DISTANCE is to decide whether $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq k$ for some given k. Like HYBRIDIZATION NUMBER, this problem is also NP-hard [5]. In the last section, Section 4, we compare the two approaches and highlight an interesting observation with regard to finding a polynomial-time approximation algorithm for HYBRIDIZATION NUMBER.

The paper is organized as follows: In the next section, we describe two notions of an "agreement forest." Both of these notions have proved fruitful in the study of rSPR DISTANCE and HYBRIDIZATION NUMBER. A third notion, which extends the other two and will be central to the results in this paper, will be described in Section 3, where the proof of Theorem 1.1 is established. Unless otherwise stated, the notation and terminology follow [17]. For an authoritative reference on fixed-parameter tractability, we refer the reader to [7].

2 Agreement Forests

Agreement forests have become an essential tool in understanding the decision problem HYBRIDIZATION NUMBER and the closely related problem rSPR DISTANCE. In this section, we describe two notions of agreement forests. The second notion provides a characterization of HYBRIDIZATION NUMBER that underpins many of the results in this area.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let X' be a subset of X. The minimal rooted subtree of $\mathcal{T}$ that connects the vertices of $\mathcal{T}$ labeled by the elements of X' is denoted by $\mathcal{T}(X')$. Furthermore, the *restriction* of $\mathcal{T}$ to X', denoted by $\mathcal{T}|X'$, is the rooted binary phylogenetic tree that is obtained from $\mathcal{T}(X')$ by suppressing any nonroot vertices of degree two.

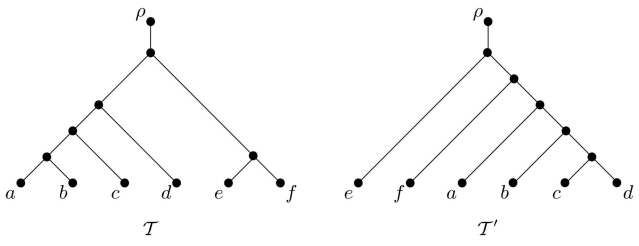


Fig. 2. Two rooted binary phylogenetic trees with their roots labeled.

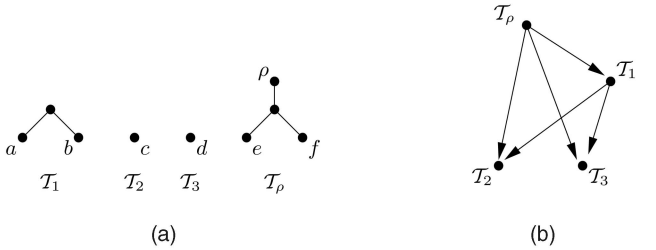


Fig. 3. (a) A maximum-acyclic-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 2 and (b) the graph $G_{\mathcal{F}}$.

Now, let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. For the purposes of the definition of an agreement forest, we regard the root of both $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ at the end of a pendant edge adjoined to the original root. Furthermore, we also regard ρ as part of the label sets of both $\mathcal{T}$ and $\mathcal{T}'$; thus, we view their label sets as $X \cup \{\rho\}$. For example, in Fig. 2, we have adjoined the vertex ρ to each of the original roots of $\mathcal{T}$ and $\mathcal{T}'$. An *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is a collection $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ of restricted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, where $\mathcal{T}_\rho$ is a rooted tree whose (leaf) label set $\mathcal{L}_\rho$ includes ρ and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ are rooted binary phylogenetic trees with label sets $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k$, respectively, such that the following properties are satisfied:

1. The label sets $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k$ partition $X \cup \{\rho\}$.
2. For all $i \in \{\rho, 1, 2, \ldots, k\}$, $\mathcal{T}_i \cong \mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$.
3. The trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are vertex disjoint subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

It is easily seen that, if $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, then $\mathcal{F}$ can be obtained from each of $\mathcal{T}$ and $\mathcal{T}'$ by deleting $|\mathcal{F}| - 1$ edges and suppressing nonroot vertices of degree two. An agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is a *maximum-agreement forest* if it has the smallest number of components among all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, in which case, we denote the value of k by $m(\mathcal{T}, \mathcal{T}')$.

While rSPR DISTANCE can be characterized in terms of agreement forests [5] (see Section 4), such a characterization for HYBRIDIZATION NUMBER requires an additional condition. This condition excludes the possibility of circular inheritance, that is, inheriting genetic information from your own descendants. Suppose that $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $G_{\mathcal{F}}$ be the directed graph whose vertex set is $\mathcal{F}$ and, for distinct vertices $\mathcal{T}_i$ and $\mathcal{T}_j$, the ordered pair $(\mathcal{T}_i, \mathcal{T}_j)$ is an arc precisely if either

1. the root of $\mathcal{T}(\mathcal{L}_i)$ in $\mathcal{T}$ is an ancestor of the root of $\mathcal{T}(\mathcal{L}_j)$ in $\mathcal{T}$ or
2. the root of $\mathcal{T}'(\mathcal{L}_i)$ in $\mathcal{T}'$ is an ancestor of the root of $\mathcal{T}'(\mathcal{L}_j)$ in $\mathcal{T}'$.

We say that $\mathcal{F}$ is an *acyclic-agreement forest* if $G_{\mathcal{F}}$ is acyclic, that is, $G_{\mathcal{F}}$ contains no directed cycles. Furthermore, if $\mathcal{F}$ contains the smallest number of components 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 $m_a(\mathcal{T}, \mathcal{T}')$. To illustrate these definitions, Fig. 3a shows a maximum-acyclic-agreement forest $\mathcal{F}$ for the two rooted binary phylogenetic trees shown in Fig. 2, while Fig. 3b shows the graph $G_{\mathcal{F}}$.

The following result is established in [2]:

Theorem 2.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then, $h(\mathcal{T}, \mathcal{T}') = m_a(\mathcal{T}, \mathcal{T}')$.*

To provide some intuition for Theorem 2.1, suppose that $\mathcal{H}$ is a hybridization network that displays $\mathcal{T}$ and $\mathcal{T}'$ such that $h(\mathcal{H}) = h(\mathcal{T}, \mathcal{T}')$. Then, it is easy to see that the in-degree of every hybridization vertex is two. Furthermore, up to suppressing degree-two vertices, an acyclic-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ can be obtained by deleting each of the edges coming into every hybridization vertex. In this case, $|\mathcal{F}| - 1 = h(\mathcal{T}, \mathcal{T}')$ and, so, we have one direction of the statement (in particular, $m_a(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}')$). Biologically, the deleted edges correspond to different paths of genetic inheritance. Consequently, the fewer edges deleted, the smaller the number of hybridization events required to explain $\mathcal{T}$ and $\mathcal{T}'$. On the other hand, if we have an acyclic-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$, then the acyclicity of $G_{\mathcal{F}}$ allows one to construct a hybridization network $\mathcal{H}$ that displays $\mathcal{T}$ and $\mathcal{T}'$ in which $h(\mathcal{H}) \leq |\mathcal{F}| - 1$. This gives the other direction of Theorem 2.1.

3 Fixed-Parameter Tractability

In this section, we prove the main result of this paper, Theorem 1.1. As mentioned in the introduction, we use two reduction rules to kernalize the problem. We begin this section by describing these two rules.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. For $n \geq 2$, an *n-chain* of $\mathcal{T}$ is an ordered tuple $(a_1, a_2, \ldots, a_n)$ of leaves of $\mathcal{T}$ such that the parent of a_1 is either the same as the parent of a_2 or a child of the parent of a_2 and, for all $i \geq 2$, the parent of a_i is a child of the parent of a_{i+1}. To illustrate, the tree $\mathcal{T}$ in Fig. 5 has an n-chain $(a_1, a_2, \ldots, a_n)$. Furthermore, a *pendant* subtree of $\mathcal{T}$ is one that can be detached by deleting a single edge.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let P be a disjoint collection of 2-element subsets of X such that each pair $\{a, b\} \in P$ is a 2-chain in both $\mathcal{T}$ and $\mathcal{T}'$. Let $w : P \to \mathbb{Z}^+$ be a weight function on the elements of P, that is, each pair in P is assigned a positive integer weight. In the remainder of the paper, we refer to such a pair of trees with associated set P and weight function w as *a pair of weighted phylogenetic trees on X*.

The above-mentioned reduction rules are as follows: Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted phylogenetic trees on X.

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

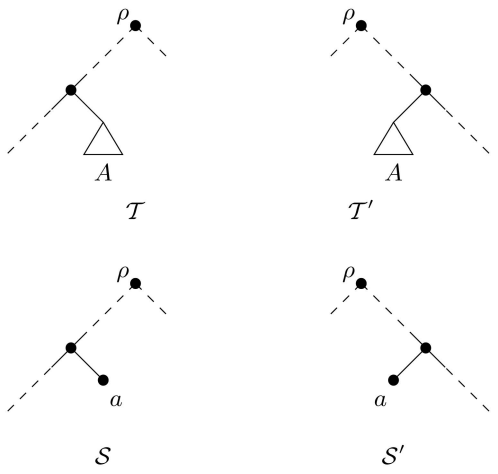


Fig. 4. Two weighted phylogenetic trees reduced under Rule 1, where $\mathcal{S}$ and $\mathcal{S}'$ are the resulting trees.

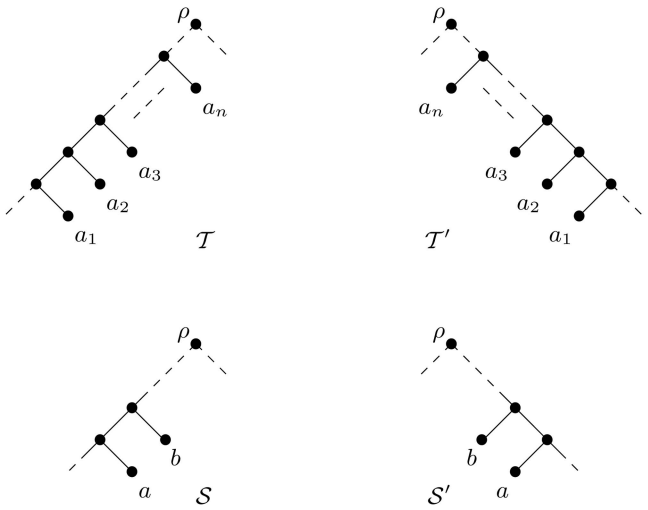


Fig. 5. Two weighted phylogenetic trees reduced under Rule 2, where $\mathcal{S}$ and $\mathcal{S}'$ are the resulting trees.

Furthermore, delete all members of P whose elements label leaves of the pendant subtree.

Rule 2. For $n \geq 3$, replace any maximal n-chain $(a_1, a_2, \ldots, a_n)$ that occurs identically in both $\mathcal{T}$ and $\mathcal{T}'$ by a 2-chain with new labels a, b. Furthermore, add the new 2-element set $\{a, b\}$ to P with weight

$$w(\{a,b\}) = n - 2 + \sum_{\{a_i,a_j\} \in P; a_i, a_j \in \{a_1,\ldots,a_n\}} w(\{a_i, a_j\})$$

and then delete all pairs in P whose elements are in $\{a_1, a_2, \ldots, a_n\}$.

Rules 1 and 2 are illustrated in Figs. 4 and 5, respectively.

Remark. The label set of any *maximal* pendant subtree or *maximal* chain which appears in both $\mathcal{T}$ and $\mathcal{T}'$ must intersect each pair in P in either both elements or neither. Hence, the rules above are well-defined. We freely use this fact in the rest of the paper.

We next introduce a third notion of agreement forests. This notion extends the previous two and is central to this paper. For a pair of weighted phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, an agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ is *legitimate* if it is acyclic and the following pairwise property holds:

(P) If $\{a, b\} \in P$, then either a and b are both contained in the label set of some component of $\mathcal{F}$ or a and b label isolated vertices in $\mathcal{F}$.

Furthermore, let $\mathcal{F}$ be an (ordinary) agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. We define the *weight* of $\mathcal{F}$, denoted $w(\mathcal{F})$, to be

$$w(\mathcal{F}) = (|\mathcal{F}| - 1) + \sum_{\{a,b\} \in P; a \text{ and } b \text{ isolated in } \mathcal{F}} w(\{a,b\})$$

and set $f(\mathcal{T}, \mathcal{T}')$ to be the minimum weight of a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Note that we always have $f(\mathcal{T}, \mathcal{T}') \geq h(\mathcal{T}, \mathcal{T}')$ since the weight function is nonnegative, and $f(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}')$ whenever the set P is empty.

The next lemma is a key result in establishing Theorem 1.1. For a vertex v of a rooted binary phylogenetic X-tree $\mathcal{T}$, the subset of X whose elements are precisely the descendants of v is a *cluster* of $\mathcal{T}$, while the *most recent common ancestor* of a subset A of X, denoted $\mathrm{mrca}_{\mathcal{T}}(A)$, is the vertex of $\mathcal{T}$ whose associated cluster is the minimal cluster of $\mathcal{T}$ containing A.

Lemma 3.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted phylogenetic trees on X. Let A be the leaf set of a maximal pendant subtree common to $\mathcal{T}$ and $\mathcal{T}'$ and let $(a_1, a_2, \ldots, a_n)$ be a maximal n-chain common to both $\mathcal{T}$ and $\mathcal{T}'$, where $n \geq 3$. Then, every legitimate-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight has the following properties:*

1. *$\mathcal{F}$ contains a tree whose label set contains every element of A and*
2. *either $\mathcal{F}$ contains a tree whose label set contains $\{a_1, a_2, \ldots, a_n\}$ or each of $a_1, a_2, \ldots, a_n$ labels an isolated vertex in $\mathcal{F}$.*

Proof. We start with the proof of 1. Let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight. Assume for a contradiction that no single component contains every element of A in its label set. We form a new legitimate-agreement forest $\mathcal{F}'$ which satisfies 1 and has smaller weight than $\mathcal{F}$. Let J index the components of $\mathcal{F}$ which include members of A in their label sets. To be precise, $J = \{j \in \{\rho, 1, \ldots, k\} : \mathcal{L}_j \cap A \neq \emptyset\}$. Let $\mathcal{F}'$ be the forest that is obtained from $\mathcal{F}$ by deleting each tree $\mathcal{T}_j$ such that $j \in J$ and inserting the new tree $\mathcal{T}_A = \mathcal{T}|(\cup_{j \in J} \mathcal{L}_j)$ with label set $\mathcal{L}_A$, say. Observe that $\mathcal{L}_j - A \neq \emptyset$ for at most one member of J since the corresponding subtrees in $\mathcal{T}$ (and $\mathcal{T}'$) must be vertex disjoint. Hence, $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, it is acyclic since the elements of A labeled a pendant subtree and legitimate since A was maximal. It remains to observe that $w(\mathcal{F}) > w(\mathcal{F}')$, since $\mathcal{F}'$ has fewer components and no additional pairs in P whose elements are isolated, which gives a contradiction.

We now turn to the proof of 2. Let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight and assume that some a_i does not label an isolated vertex. Then, without loss of

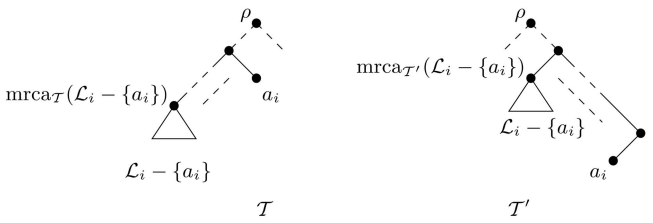


Fig. 6. Under the assumption that this configuration appears in $\mathcal{F}$, then the rest of the members of $\{a_1, a_2, \ldots, a_n\} - \{a_i\}$ must label isolated vertices in $\mathcal{F}$.

generality, the label a_i is contained in the label set $\mathcal{L}_i$ of $\mathcal{T}_i$, where $\mathcal{L}_i - \{a_i\}$ is nonempty. First, we eliminate a particular way in which a_i may be related to $\mathcal{L}_i - \{a_i\}$ in $\mathcal{T}$ and $\mathcal{T}'$.

Suppose a_i is adjoined to the root of $\mathcal{T}_i$ such that the parent of a_i in one of the original trees, $\mathcal{T}$ say, is an ancestor of $\mathrm{mrca}_{\mathcal{T}}(\mathcal{L}_i - \{a_i\})$ while the parent of a_i in $\mathcal{T}'$ is not an ancestor of $\mathrm{mrca}_{\mathcal{T}'}(\mathcal{L}_i - \{a_i\})$ (see Fig. 6). Then, each of the elements in $\{a_1, a_2, \ldots, a_n\} - \{a_i\}$ must label an isolated vertex in $\mathcal{F}$; otherwise, the corresponding subtrees of two components of $\mathcal{F}$ in either $\mathcal{T}$ or $\mathcal{T}'$ overlap. By deleting a_i from $\mathcal{T}_i$ and replacing these isolated vertices with a single tree that is isomorphic to $\mathcal{T}|\{a_1, a_2, \ldots, a_n\}$, it is easily seen that the resulting agreement forest $\mathcal{F}'$ is acyclic. Since $(a_1, a_2, \ldots, a_n)$ is a maximal n-chain and $\mathcal{F}$ is legitimate, it follows that $\mathcal{F}'$ satisfies (P). But, $w(\mathcal{F}') < w(\mathcal{F})$, contradicting the minimality of $\mathcal{F}$. Thus, we may assume that, if a_i is adjoined to the root of $\mathcal{T}_i$ and the parent of a_i in $\mathcal{T}$ is an ancestor of $\mathrm{mrca}_{\mathcal{T}}(\mathcal{L}_i - \{a_i\})$, then the parent of a_i in $\mathcal{T}'$ is also an ancestor of $\mathrm{mrca}_{\mathcal{T}'}(\mathcal{L}_i - \{a_i\})$.

Now, let J index the components of $\mathcal{F}$ which contain elements of the chain. To be precise, $J = \{j \in \{\rho, 1, \ldots, k\} : \mathcal{L}_j \cap \{a_1, a_2, \ldots, a_n\} \neq \emptyset\}$. Observe that $\mathcal{L}_j - \{a_1, \ldots, a_n\} \neq \emptyset$ for at most two members of J since the corresponding subtrees in $\mathcal{T}$ (and $\mathcal{T}'$) are vertex disjoint. Let $\mathcal{F}'$ be the forest that is obtained from $\mathcal{F}$ by deleting each tree $\mathcal{T}_j$ such that $j \in J$ and inserting the new tree $\mathcal{T}_a = \mathcal{T}|(\cup_{j \in J} \mathcal{L}_j)$ with label set $\mathcal{L}_a$, say. Essentially, we have joined the components in $\mathcal{F}$ involving elements of $\{a_1, a_2, \ldots, a_n\}$ together, along the chain. An illustration of this is shown in Fig. 7, where the left-hand side of the figure shows the components of $\mathcal{F}$ containing elements in $\{a_1, a_2, \ldots, a_n\}$, while the right-hand side shows $\mathcal{T}_a$ in $\mathcal{F}'$. It follows from the assumption at the end of the previous paragraph that $\mathcal{F}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ since the chain is common to both trees. Furthermore, as $(a_1, a_2, \ldots, a_n)$ is maximal, $\mathcal{F}'$ satisfies (P).

We next show that $\mathcal{F}'$ is acyclic. Consider the directed graphs $G_{\mathcal{F}'}$ and $G_{\mathcal{F}}$ associated with $\mathcal{F}'$ and $\mathcal{F}$, respectively. First, the vertex set of $G_{\mathcal{F}'}$ is obtained from $G_{\mathcal{F}}$ by deleting the vertices $\mathcal{T}_j$ for all $j \in J$ and introducing the new vertex $\mathcal{T}_a$. Furthermore, if $\mathcal{T}_l, \mathcal{T}_m \in \mathcal{F}' - \{\mathcal{T}_a\}$, then $(\mathcal{T}_l, \mathcal{T}_m)$ is an arc in $G_{\mathcal{F}'}$ if and only if $(\mathcal{T}_l, \mathcal{T}_m)$ is an arc in $G_{\mathcal{F}}$. Regarding the arcs incident with $\mathcal{T}_a$, there are two cases to consider. First, suppose there is some $j_1 \in J$ such that the root of $\mathcal{T}(\mathcal{L}_{j_1})$ in $\mathcal{T}$ is above a_n (i.e., on the path from a_n to ρ). Then, the root of $\mathcal{T}(\mathcal{L}_a)$ is the same as the root of $\mathcal{T}(\mathcal{L}_{j_1})$ and, under our assumptions, the respective roots must also coincide in $\mathcal{T}'$. This occurs in the example given in Fig. 7, where, in both $\mathcal{T}$ and $\mathcal{T}'$, the root of $\mathcal{T}(\mathcal{L}_2 \cup \{a_6, a_7\})$ is the same as the root of $\mathcal{T}(\mathcal{L}_a)$. So, $(\mathcal{T}_a, \mathcal{T}_l)$ and $(\mathcal{T}_l, \mathcal{T}_a)$ are arcs in $G_{\mathcal{F}'}$ if and only if $(\mathcal{T}_{j_1}, \mathcal{T}_l)$ and $(\mathcal{T}_l, \mathcal{T}_{j_1})$ are arcs in $G_{\mathcal{F}}$, respectively. Since $G_{\mathcal{F}}$ is acyclic, $G_{\mathcal{F}'}$ must be also. Second, suppose there is no such $j_1 \in J$. Then, the root of $\mathcal{T}(\mathcal{L}_a)$ is the parent of a_n in $\mathcal{T}$ and, likewise, the root of $\mathcal{T}'(\mathcal{L}_a)$ is the parent of a_n in $\mathcal{T}'$. Since not all of the elements in $\{a_1, \ldots, a_n\}$ are isolated in $\mathcal{F}$, there is some $j_2 \in J$ such that the root of $\mathcal{T}(\mathcal{L}_{j_2})$ in $\mathcal{T}$ is above a_1. It again follows that $(\mathcal{T}_a, \mathcal{T}_l)$ and $(\mathcal{T}_l, \mathcal{T}_a)$ are arcs in $G_{\mathcal{F}'}$ if and only if $(\mathcal{T}_{j_2}, \mathcal{T}_l)$ and $(\mathcal{T}_l, \mathcal{T}_{j_2})$ are arcs in $G_{\mathcal{F}}$, respectively, and, so, $G_{\mathcal{F}'}$ is acyclic. Hence, $\mathcal{F}'$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. If $a_1, \ldots, a_n$ are not all in the same component of $\mathcal{F}$ (i.e., if $|J| > 1$), then we have reduced the number of components and, so, $w(\mathcal{F}') < w(\mathcal{F})$. This contradicts the minimality of $\mathcal{F}$. Hence, under the original assumption that some a_i does not label an isolated vertex, we conclude that the chain is entirely contained in a single component of $\mathcal{F}$. This concludes the proof of the lemma. □

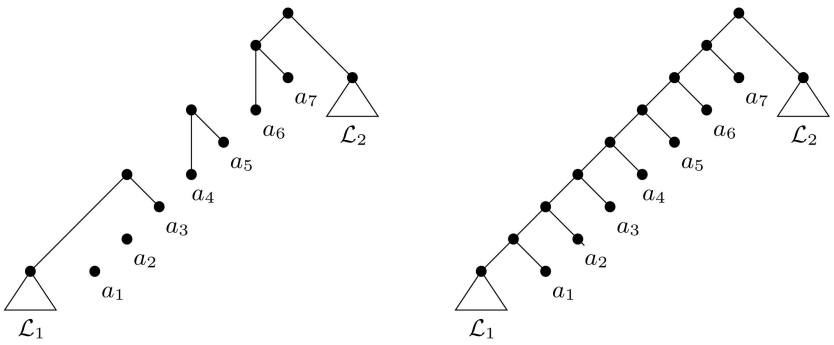
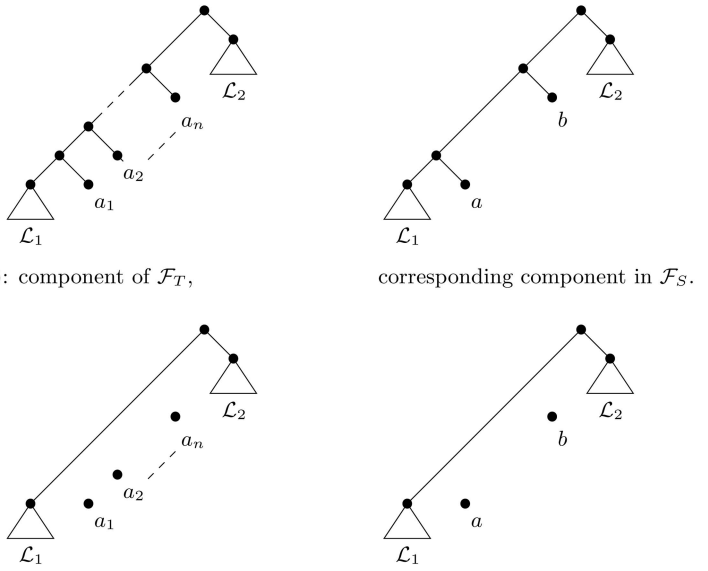


Fig. 7. Joining the components of $\mathcal{F}$ containing elements in $\{a_1, a_2, \ldots, a_n\}$ to form a new component $\mathcal{T}_a$ in $\mathcal{F}'$.



Case (1): component of $\mathcal{F}_T$, corresponding component in $\mathcal{F}_S$.

Case (2): components of $\mathcal{F}_T$, corresponding components in $\mathcal{F}_S$.

Fig. 8. Corresponding components of $\mathcal{F}_T$ and $\mathcal{F}_S$ for Cases 1 and 2.

Proposition 3.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be a pair of weighted phylogenetic X-trees on X. Let $\mathcal{S}$ and $\mathcal{S}'$ be the pair of weighted phylogenetic X'-trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by applying either Rule 1 or Rule 2. Then, $f(\mathcal{T},\mathcal{T}') = f(\mathcal{S},\mathcal{S}')$.*

Proof. It is an immediate consequence of Lemma 3.1.1 that the proposition holds if $\mathcal{S}$ and $\mathcal{S}'$ have been obtained from $\mathcal{T}$ and $\mathcal{T}'$ by applying Rule 1. Therefore, consider a single application of Rule 2 to $\mathcal{T}$ and $\mathcal{T}'$, where the common n-chain of $\mathcal{T}$ and $\mathcal{T}'$ that is used is $(a_1, a_2, \ldots, a_n)$ and the resulting 2-chain is (a, b).

Let $\mathcal{F}_T$ be a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of minimum weight. Then, by Lemma 3.1.2, either

1. $\{a_1, a_2, \ldots, a_n\}$ is contained in the label set of a tree in $\mathcal{F}_T$ or
2. each of $a_1, a_2, \ldots, a_n$ label isolated vertices in $\mathcal{F}_T$.

Let $\mathcal{F}_S$ be the forest obtained from $\mathcal{F}_T$ by either replacing the n-chain $(a_1, a_2, \ldots, a_n)$ with the 2-chain (a, b) or replacing the isolated vertices labeled with the elements of this n-chain with two isolated vertices labeled a and b depending upon whether 1 or 2 holds, respectively. Illustrations of $\mathcal{F}_T$ and $\mathcal{F}_S$ for 1 and 2 are shown in Fig. 8. Since $\mathcal{F}_T$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, a routine check shows that $\mathcal{F}_S$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$. Moreover, in the case where 2 holds, the contribution of the isolated vertices $a_1, a_2, \ldots, a_n$ to $w(\mathcal{F}_T)$ is exactly the same as the contribution of the isolated vertices a, b to $w(\mathcal{F}_S)$. It now follows that $f(\mathcal{S},\mathcal{S}') \leq f(\mathcal{T},\mathcal{T}')$.

Now, suppose that $\mathcal{F}_S$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ with minimum weight. Since $\mathcal{F}_S$ is legitimate, either

1. there is a tree, $\mathcal{S}_i$, say, in $\mathcal{F}_S$ whose label set contains a and b or
2. a and b label isolated vertices in $\mathcal{F}_S$.

Depending on which holds, let $\mathcal{F}_T$ be the forest obtained from $\mathcal{F}_S$ by either replacing $\mathcal{S}_i$ with the restriction of $\mathcal{T}$ to $(\mathcal{L}(\mathcal{S}_i) - \{a, b\}) \cup \{a_1, a_2, \ldots, a_n\}$ or replacing the isolated vertices labeled a and b with n isolated vertices labeled $a_1, a_2, \ldots, a_n$, respectively. Since $\mathcal{F}_S$ is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$, a routine check shows that $\mathcal{F}_T$ is a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, as the contribution of the isolated vertices labeled a, b to $w(\mathcal{F}_S)$ is the same as the contribution of the isolated vertices labeled $a_1, a_2, \ldots, a_n$ to $w(\mathcal{F}_T)$ in case 2, we have that $f(\mathcal{T},\mathcal{T}') \leq f(\mathcal{S},\mathcal{S}')$. This completes the proof of the proposition. □

Proposition 3.2 says that the tree reduction rules, Rules 1 and 2, preserve the function f. We now show that Rules 1 and 2 can be applied until the label set of the resulting rooted binary phylogenetic trees has size bounded by a linear function of the value of f.

Lemma 3.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees and let P be an empty collection of 2-element subsets of X. Let $\mathcal{S}$ and $\mathcal{S}'$ be two weighted phylogenetic X'-trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by repeatedly applying Rules 1 and 2 until no further reduction is possible. Then, $|X'| < 14h(\mathcal{T},\mathcal{T}')$.*

Proof. As in [5, Lemma 3.3], we follow the approach in [1, Lemma 3.7]. Let $\{\mathcal{S}_\rho, \mathcal{S}_1, \ldots, \mathcal{S}_k\}$ be a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ with minimum weight. For $i = \rho, 1, 2, \ldots, k$, set $\mathcal{L}_i = \mathcal{L}(\mathcal{S}_i)$ and let n_i denote the number of edges in $E(\mathcal{S}) - E(\mathcal{S}(\mathcal{L}_i))$ which are incident with the subtree $\mathcal{S}(\mathcal{L}_i)$ and let n'_i denote the number of edges in $E(\mathcal{S}') - E(\mathcal{S}'(\mathcal{L}_i))$ which are incident with the subtree $\mathcal{S}'(\mathcal{L}_i)$. The proof essentially consists of two claims.

Claim 1. *$\sum_i n_i \leq 2k$ and $\sum_i n'_i \leq 2k$.*

By symmetry, it suffices to show that $\sum_i n_i \leq 2k$. Consider the tree (V, E) obtained from $\mathcal{S}$ by contracting

each subtree $\mathcal{S}(\mathcal{L}_i)$ to a single vertex. In this tree, V consists of the vertices corresponding to the trees $\mathcal{S}_i$, each of which has degree n_i, and the additional vertices of degree 3. Hence, by the Handshaking Lemma, $\sum_i n_i + 3(|V| - (k+1)) = 2|E|$. Therefore, as (V, E) is a tree and so $|V| = |E| + 1$, it follows that

$$\sum_i n_i = 2(|V| - 1) - 3(|V| - (k+1)) = 3k - |V| + 1 \leq 2k.$$

Thus, Claim 1 holds.

Claim 2. *For each i, the number of leaves in $\mathcal{S}_i$ is at most $5(n_i + n_i') - 6$.*

Let I be the set of edges e of $\mathcal{S}_i$ such that, in the path of edges corresponding to e in either $\mathcal{S}(\mathcal{L}_i)$ or $\mathcal{S}'(\mathcal{L}_i)$, one of the vertices in this path is incident with an edge in $E(\mathcal{S}) - E(\mathcal{S}(\mathcal{L}_i))$ or $E(\mathcal{S}') - E(\mathcal{S}(\mathcal{L}_i))$, respectively. Note that $|I| \leq n_i + n_i'$. Let $\mathcal{S}_i'$ denote the tree obtained from the minimal subtree of $\mathcal{S}_i$ that contains the edges in I by suppressing nonroot degree-2 vertices not incident with an edge in I. Let J denote the set consisting of these new edges, $E(\mathcal{S}_i') - I$, and let I_{pend} denote the set of pendant edges of $\mathcal{S}_i'$. Note that $I_{pend} \subseteq I$. Observe that every subtree of $\mathcal{S}_i$ below an edge in I_{pend} will have been replaced by a single vertex using Rule 1, as these pendant subtrees are clearly common to both trees since they are in the agreement forest and they are maximal by Lemma 3.1. Similarly, each chain of subtrees in $\mathcal{S}_i$ corresponding to an edge in J will have been replaced by a 2-chain using Rules 1 and 2. Furthermore, the only other place a subtree, again reduced to a leaf under Rule 1, could attach itself to $\mathcal{S}_i$ is at a degree-2 vertex that is incident with two edges in I. If we identify each such vertex by the edge in I above it, it is clear that there are at most $|I| - |I_{pend}|$ such leaves. Hence, the number of leaves in $\mathcal{S}_i$ is at most $|I_{pend}| + 2|J| + (|I| - |I_{pend}|) = |I| + 2|J|$.

Let m_2 and m_3 denote the number of vertices of $\mathcal{S}_i'$ of degree 2 and 3, respectively. Then, as $|I_{pend}|$ is the number of vertices of degree 1, it follows by the Handshaking Lemma that $2|E(\mathcal{S}_i')| = |I_{pend}| + 2m_2 + 3m_3$. Therefore, as $|E(\mathcal{S}_i')| = |V(\mathcal{S}_i')| - 1$,

$$2\big(|I_{pend}| + m_2 + m_3 - 1\big) = |I_{pend}| + 2m_2 + 3m_3.$$

This last equality implies that $m_3 = |I_{pend}| - 2$. Furthermore,

$$|J| + |I| = (|I_{pend}| + m_2 + m_3) - 1$$

and, by construction, any degree-2 vertex in $\mathcal{S}_i'$ must be adjacent to at least one edge in I, so $m_2 \leq 2|I| - |I_{pend}|$. Therefore, the number of leaves in $\mathcal{S}_i$ is at most

$$\begin{aligned}
|I| + 2|J| &= |I| + 2(|I_{pend}| + m_2 + m_3 - 1 - |I|) \\
&\leq |I| + 2(|I_{pend}| + (2|I| - |I_{pend}|) + (|I_{pend}| - 2) \\
&\quad - 1 - |I|) \\
&= 2|I_{pend}| + 3|I| - 6 \\
&\leq 5|I| - 6 \\
&\leq 5(n_i + n_i') - 6.
\end{aligned}$$

This proves Claim 2.

Now, by Claim 1, we have $\sum_i (n_i + n_i') \leq 4k$ and, so,

$$\sum_i |\mathcal{L}_i| \leq 5 \sum_i (n_i + n_i') - 6(k+1) \leq 14k - 6.$$

By the definition of f and Proposition 3.2, $k \leq f(\mathcal{S}, \mathcal{S}') = f(\mathcal{T}, \mathcal{T}')$. Since P is initially empty, we also have $f(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}')$ and the result follows. □

We are now in a position to show that the decision problem HYBRIDIZATION NUMBER is fixed-parameter tractable.

Proof of Theorem 1.1. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees and let P be an empty collection of 2-element subsets of X. Let k be an integer. Let $\mathcal{S}$ and $\mathcal{S}'$ be the weighted phylogenetic X'-trees obtained from $\mathcal{T}$ and $\mathcal{T}'$ by repeatedly applying Rules 1 and 2 until no further reduction is possible. Then, as P is empty, $h(\mathcal{T}, \mathcal{T}') = f(\mathcal{T}, \mathcal{T}')$ and, by Proposition 3.2, $f(\mathcal{T}, \mathcal{T}') = f(\mathcal{S}, \mathcal{S}')$, thus $h(\mathcal{T}, \mathcal{T}') = f(\mathcal{S}, \mathcal{S}')$. As in [1] and [5], $\mathcal{S}$ and $\mathcal{S}'$ can be found in time polynomial in $|X|$ ($p(|X|)$, say). By Lemma 3.3, $|X'| \leq 14h(\mathcal{T}, \mathcal{T}')$. Thus, if $|X'| > 14k$, we declare that $h(\mathcal{T}, \mathcal{T}') > k$.

Now, suppose that $|X'| \leq 14k$. We next consider the time taken to check whether there is a legitimate-agreement forest for $\mathcal{S}$ and $\mathcal{S}'$ of weight at most k by deleting up to k edges of $\mathcal{S}$ and then seeing if the resulting forest is such a legitimate-agreement forest. Note that checking for legitimacy takes polynomial time. For a given rooted binary phylogenetic X'-tree, there are $2|X'| - 1$ possible edges to delete, including the edge incident with ρ. Thus, there are at most $\sum_{i=0}^{k} \binom{2|X'|-1}{i} \leq \sum_{i=0}^{k} (2|X'| - 1)^i \leq 2(2|X'| - 1)^k$ forests to examine, which can be done in time $O((2|X'|)^k) = O((28k)^k)$. 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 $O(f(k) + p(|X|))$, where $f(k)$ is the computable function $(28k)^k$ and $p(|X|)$ is the polynomial bound for reducing the trees using Rules 1 and 2. This satisfies the conditions for HYBRIDIZATION NUMBER to be fixed-parameter tractable. □

Remark. By making an organized comparison of the set of clusters of $\mathcal{T}$ and the set of clusters of $\mathcal{T}'$, a naive approach for fully reducing $\mathcal{T}$ and $\mathcal{T}'$ using Rules 1 and 2 results in an $O(n^3)$ algorithm, where $n = |X|$. While a further such approach for deciding if a particular set of k edge cuts produces a legitimate-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ gives an $O(k^2 + |P|)$ algorithm. We omit the details of these algorithms as they are not necessarily the best theoretically and we expect, in practice, much quicker methods. An implementation of the associated fixed-parameter algorithm and an analysis of its running time is the subject of ongoing research.

4 Some Remarks on rSPR Distance and Hybridization Number

In this section, we compare the approach used to prove Theorem 1.1 with that used in [5] for showing that rSPR DISTANCE is fixed-parameter tractable. We begin by formally defining the subtree prune and regraft operation.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and, as in the definition of an agreement forest, view the root of $\mathcal{T}$ as a vertex ρ adjoined to the original root by a new pendant edge. Let $e = \{u, v\}$ be an edge of $\mathcal{T}$ that is not incident with ρ, where u is in the path from ρ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic X-tree obtained from $\mathcal{T}$ by deleting e and then adjoining a new edge f between v and the component C_u that contains u as follows: Create a new vertex u' which subdivides an edge in C_u, adjoin f between u' and v, and then suppress the degree-two vertex u. The tree $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a single *rooted subtree prune and regraft* (rSPR) operation. The rSPR *distance* (d_{rSPR}) between two arbitrary rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is the minimum number of rSPR operations required to transform $\mathcal{T}$ into $\mathcal{T}'$.

Historically, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ has been used as a replacement for $h(\mathcal{T}, \mathcal{T}')$. The reason for this is that individual hybridization events correspond to individual rSPR operations and, indeed, a collection of hybridization events can be modeled by a sequence of rSPR operations. However, the converse does not hold since an arbitrary sequence of rSPR operations may include circular inheritance. It is shown in [2] that the difference between rSPR DISTANCE and HYBRIDIZATION NUMBER can be arbitrarily large. Nevertheless, the two values are closely related. Recall Theorem 2.1, which says that, for two rooted binary phylogenetic X-trees, $\mathcal{T}$ and $\mathcal{T}'$, the value $h(\mathcal{T}, \mathcal{T}')$ is one less than the number of components in a maximum-*acyclic*-agreement forest, $m_a(\mathcal{T}, \mathcal{T}')$. In comparison, we have the following result from [5]:

Theorem 4.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}')$, where $m(\mathcal{T}, \mathcal{T}')$ denotes the size of a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ minus one.*

The overall approach we have used to prove Theorem 1.1 is similar to that used in [5] to show that rSPR DISTANCE is fixed-parameter tractable (parameterized by $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$), but there are some crucial differences. In both papers, the problems are kernalized using two reduction rules which bound the size of the leaf sets of the resulting pairs of trees in terms of the parameter. The first rule in [5] is essentially identical to Rule 1 here, but the second rule differs from Rule 2 here. The lack of the acyclicity constraint means that there is a maximum-agreement forest in which every common n-chain $(n \geq 3)$ is a connected subtree of a component [5, Lemma 3.1] and, so, each such chain can be replaced by an unweighted 3-chain.

The implication of this is that there is no need for weighted forests, so, if $\mathcal{S}$ and $\mathcal{S}'$ are the rooted binary phylogenetic X'-trees resulting from applying the appropriate two rules, then the size of a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is bounded above by $|X'|$, the number of leaves in $\mathcal{S}$ (or $\mathcal{S}'$). The consequence is that the fixed-parameter algorithm for rSPR DISTANCE in [5] also provides a polynomial-time approximation algorithm for this problem. The analogue of Lemma 3.3 in [5] (with the upper bound on $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ included) is that

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq |X'| \leq 28 d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}').$$

Therefore, the size of the label sets of the reduced trees $\mathcal{S}$ and $\mathcal{S}'$ gives a 28-approximation for $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. With some modifications along the lines of legitimate-agreement forests, this approach can be made to yield a 9-approximation. However, no such approximation algorithm for HYBRIDIZATION NUMBER follows in an analogous way from the results in this paper since $|X'|$ does not bound the hybridization number due to the presence of weights. Indeed, there is currently no polynomial-time approximation algorithm for HYBRIDIZATION NUMBER.

Using a different approach, based upon ideas in [13], [16], the current best polynomial-time approximation algorithm for rSPR DISTANCE is a 5-approximation algorithm by Bonet et al. [4]. Intuitively, this algorithm builds an agreement forest by looking only at local structures. One might hope that this algorithm extends to HYBRIDIZATION NUMBER (using Theorem 2.1), but, due to the additional global condition on an *acyclic*-agreement forest, it seems unlikely that such an approach will work.

Acknowledgments

Magnus Bordewich was supported by the EPSRC and Charles Semple was supported by the New Zealand Marsden Fund.

References

[1] B.L. Allen and M. Steel, "Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees," *Annals of Combinatorics,* vol. 5, pp. 1-13, 2001.

[2] M. Baroni, S. Grünewald, V. Moulton, and C. Semple, "Bounding the Number of Hybridization Events for a Consistent Evolutionary History," *Math. Biology,* vol. 51, pp. 171-182, 2005.

[3] M. Baroni, C. Semple, and M. Steel, "Hybrids in Real Time," *Systematic Biology,* vol. 55, pp. 46-56, 2006.

[4] M.K. Bonet, K. St. John, R. Mahindru, and N. Amenta, "Approximating Subtree Distances between Phylogenies," *J. Computational Biology,* vol. 13, pp. 1419-1434, 2006.

[5] M. Bordewich and C. Semple, "On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance," *Annals of Combinatorics,* vol. 8, pp. 409-423, 2004.

[6] M. Bordewich and C. Semple, "Computing the Minimum Number of Hybridisation Events for a Consistent Evolutionary History," *Discrete Applied Math.,* vol. 155, pp. 914-928, 2007.

[7] R. Downey and M. Fellows, *Parameterized Complexity.* Springer, 1998.

[8] D. Gusfield and V. Bansal, "A Fundamental Decomposition Theory for Phylogenetic Networks and Incompatible Characters," *Proc. Ninth Ann. Int'l Conf. Research in Computational Molecular Biology (RECOMB '05),* S. Miyano et al., eds. pp. 217-232, 2005.

[9] D. Gusfield, S. Eddhu, and C. Langley, "Optimal, Efficient Reconstruction of Phylogenetic Networks with Constrained Recombination," *J. Bioinformatics and Computational Biology,* vol. 2, pp. 173-213, 2004.

[10] M. Hallett and J. Lagergren, "Efficient Algorithms for Lateral Gene Transfer Problems," *Proc. Fifth Ann. Int'l Conf. Research in Computational Molecular Biology (RECOMB '01),* pp. 149-156, 2001.

[11] J. Hein, "Reconstructing Evolution of Sequences Subject to Recombination Using Parsimony," *Math. Biosciences,* vol. 98, pp. 185-200, 1990.

[12] J. Hein, "A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination," *J. Molecular Evolution,* vol. 36, pp. 396-405, 1993.
[13] J. Hein, T. Jing, L. Wang, and K. Zhang, "On the Complexity of Comparing Evolutionary Trees," *Discrete Applied Math.,* vol. 71, pp. 153-169, 1996.
[14] W. Maddison, "Gene Trees in Species Trees," *Systematic Biology,* vol. 46, pp. 523-536, 1997.
[15] L. Nakhleh, T. Warnow, C.R. Linder, and K. St. John, "Reconstructing Reticulate Evolution in Species—Theory and Practice," *J. Computational Biology,* vol. 12, pp. 796-811, 2005.
[16] E.M. Rodrigues, M.-F. Sagot, and Y. Wakabayashi, "Some Approximation Results for the Maximum Agreement Forest Problem," *Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques,* M. Goemans et al., eds., pp. 159-169, Springer, 2001.
[17] C. Semple and M. Steel, *Phylogenetics.* Oxford Univ. Press, 2003.
[18] Y.S. Song and J. Hein, "Parsimonious Reconstruction of Sequence Evolution and Haplotyde Blocks: Finding the Minimum Number of Recombination Events," *Algorithms in Bioinformatics, Proc. WABI '03,* G. Benson and R. Page, eds., pp. 287-302, 2003.
[19] L. Wang, K. Zhang, and L. Zhang, "Perfect Phylogenetic Networks with Recombination," *J. Computational Biology,* vol. 8, pp. 69-78, 2001.

Magnus Bordewich received the MMath (1998) and DPhil degrees in mathematics (2003) from New College, Oxford University. He is a lecturer in the Department of Computer Science at the University of Durham. Subsequent to receiving his degrees, he was a postdoctoral research fellow in the Department of Mathematics and Statistics at the University of Canterbury, New Zealand, and then in the School of Computer Science at Leeds University, United Kingdom. He joined Durham University as a lecturer in 2006 and was recently awarded an EPSRC postdoctoral fellowship in theoretical computer science, researching randomized algorithms and approximation in phylogenetics.

Charles Semple received the BSc (Hons) degree in mathematics from Massey University and the MSc and PhD degrees in mathematics from Victoria University of Wellington. He is a senior lecturer in the Department of Mathematics and Statistics at the University of Canterbury, New Zealand. Initially a postdoctoral fellow, he has been a permanent staff member at the University of Canterbury since 2001. Other academic positions include visiting research fellow at Merton College, University of Oxford (2003) and visiting professor at the University of Montpellier II (2005). His main research interests are combinatorics, computational complexity, and computational biology.