On the Fixed Parameter Tractability of Agreement-based Phylogenetic Distances

Magnus Bordewich, Celine Scornavacca, Nihan Tokac and Mathias Weller

Preprint: 9th May 2016

Abstract Three important and related measures for summarizing the dissimilarity in phylogenetic trees are the minimum number of hybridization events required to fit two phylogenetic trees onto a single phylogenetic network (the *hybridization number*), the (rooted) subtree prune and regraft distance (the *rSPR distance*) and the tree bisection and reconnection distance (the *TBR distance*) between two phylogenetic trees. The respective problems of computing these measures are known to be NP-hard, but also fixed-parameter tractable in their respective natural parameters. This means that, while they are hard to compute in general, for cases in which a parameter (here the hybridization number and rSPR/TBR distance, respectively) is small, the problem can be solved efficiently even for large input trees. Here, we present new analyses showing that the use of the "cluster reduction" rule – already defined for the hybridization number and the rSPR distance and introduced here for the TBR distance – can transform any $O(f(p) \cdot n)$-time algorithm for any of these problems into an $O(f(k) \cdot n)$-time one, where n is the number of leaves of the phylogenetic trees, p is the natural parameter and k is a much stronger (that

Magnus Bordewich
School of Engineering and Computing Sciences, Durham University, Durham DH1 3LE, U.K., E-mail: m.j.r.bordewich@durham.ac.uk

Celine Scornavacca
Institut des Sciences de l'Evolution (Université de Montpellier, CNRS, IRD, EPHE), Place E. Bataillon CC 064 - 34095 Montpellier Cedex 5, France, E-mail: celine.scornavacca@umontpellier.fr

Nihan Tokac
School of Engineering and Computing Sciences, Durham University, Durham DH1 3LE, U.K., E-mail: nihan.tokac@durham.ac.uk

Mathias Weller
Institut de Biologie Computationnelle (IBC), Laboratory of Informatics, Robotics, and Microelectronics of Montpellier (LIRMM), Université de Montpellier, 161 rue Ada 34392 Montpellier Cedex 5, France, E-mail: mathias.weller@lirmm.fr

is, smaller) parameter: the minimum *level* of a phylogenetic network displaying both trees.

Keywords
Phylogenetic network, hybridization number, cluster reduction, SPR distance, TBR distance.

1 Introduction

Since Darwin first introduced the theory of evolution, one of the central goals of evolutionary biology has been to try to construct an accurate ancestral history of present day species. Reconstruction of *phylogenetic trees* has been the principal tool to study the relationships between taxa, however it has long been known that not all evolution can be represented by a tree. There are some groups (for example including some subgroups of plants and fish) for which the evolutionary history contains reticulation events, caused by processes including hybridization, lateral gene transfer and recombination. For such groups of species, it is appropriate to represent their ancestral history by *phylogenetic networks*: single-rooted acyclic digraphs, where arcs represent lines of genetic inheritance and vertices of in-degree at least two represent reticulation events.

One fundamental problem is to determine how much reticulation is required to explain the evolution of a given set of taxa: given a collection of rooted phylogenetic trees on a set of taxa that correctly represent the tree-like evolution of different parts of their genomes, what is the smallest number of reticulation events needed to display the trees within a single phylogenetic network (the Hybridization Number problem)?

This question, along with the closely related problems of determining the minimum number of subtree prune and regraft, respectively tree bisection and reconnection, operations required to transform one phylogenetic tree into another (the rSPR Distance and TBR Distance problem, respectively) has been considered in a number of papers [1, 2, 3, 5, 10, 12, 17, 18]. Key theoretical developments have shown that each of these three problems is NP-hard even in the restricted case that the input consists of two binary phylogenetic trees [4, 6, 12], but also that they are all fixed-parameter tractable in their respective natural parameters [1, 4, 5]. In essence, this means that there are efficient algorithms for computing the hybridization number and the rSPR/TBR distance on two trees of large size, as long as there have not been too many reticulations in the evolutionary history of the considered taxa. In these theoretical analyses, an operation known as *chain reduction* is used to prove fixed-parameter tractability, but this operation does not seem to help the algorithms much in practice. On the other hand another operation, the *cluster reduction* [3], which did not crop up in the theoretical analyses, greatly speeds up the algorithms in practice. The cluster reduction for Hybridization Number has been included in algorithms since the first parameterized algorithms appeared [7], and recent work has shown the applicability of an equivalent cluster reduction for rSPR Distance [16].

Here, we give a theoretical justification of why the cluster reduction for Hybridization Number is so useful in practice by showing that the divide-and-conquer approach that follows from it implies fixed-parameter tractability where the parameter is not the *total number of reticulations* in the optimal network displaying the two input trees, but instead the *maximum number of reticulations seen in any biconnected component* of such a network. This concept has been studied before as the *level* of the network (see for example [14, 21]). In essence, this means that for large input trees, even when there have been many reticulations, as long as not too many of the reticulations are entangled with each other, the problem may still be solved efficiently. This is what is expected to happen for real biological data, in part because reticulation events such as hybridization events are less likely to happen between genetically-distant species.

Actually, in this paper, we show something stronger: the use of the cluster reduction can transform any $O(f(p) \cdot n)$-time algorithm for any of the considered problems into an $O(f(k) \cdot n)$-time algorithm, where n is the number of leaves of the phylogenetic trees, p is the natural parameter and k is the minimum level of a phylogenetic network displaying both trees, which is a much stronger (that is, smaller) parameter than p.

The fact that the cluster reduction implies fixed-parameter tractability in the level for Hybridization Number was already implicitly present in [15, 20]. Still, we think that it is worth proving explicitly and formally, and extending the reasoning to rSPR Distance and TBR Distance, thus giving hard evidence for the importance of implementing the cluster reduction in available software.

In the next section, we give formal notation and definitions and we present the main theorems of the paper. We prove fixed-parameter tractability of Hybridization Number, rSPR Distance and TBR Distance with respect to the level in Sections 3, 4, and 5, respectively.

2 Definitions and Statement of Results

The notation and terminology in this paper follows Semple and Steel [19], unless explicitly stated otherwise. A directed graph (digraph) D is an ordered pair (V, A) consisting of a non-empty set V of vertices and a set $A \subseteq V \times V$ of arcs. A digraph is *acyclic* (a *DAG*) if it has no directed cycles. The *degree* of a vertex is the sum of its in- and out-degree. A vertex of degree zero is said to be isolated, and a vertex of in-degree one and out-degree zero is called a *leaf*. A vertex of in-degree zero is called a *root*.

A *rooted binary phylogenetic network* $N = (D, \phi)$ (on X) is an ordered pair consisting of a DAG D with a unique root ρ, and a map ϕ such that

1. ϕ bijectively maps X to the set of leaves of D,
2. ρ has in-degree zero and out-degree two, and
3. all other vertices have degree three.

The vertices of in-degree two (and out-degree one) are called *reticulation vertices*. We denote the set of leaf labels associated to a rooted binary phylogenetic network N by $\mathcal{L}(N)$ (note that $X = \mathcal{L}(N)$).

A *rooted binary phylogenetic X-tree* (or *rooted binary phylogenetic tree* on X) is a rooted binary phylogenetic network on X without reticulation vertices.

The number of arcs we need to remove from a rooted phylogenetic network N on X to obtain a rooted binary phylogenetic tree on X is denoted by $h(N)$ and referred to as the *hybridization number* of N. (Note that, for rooted *binary* phylogenetic networks, $h(N)$ coincides with the number of reticulation vertices in N). A *cut vertex (cut arc)* is a vertex (an arc) whose removal disconnects the graph. A *biconnected component* is a maximal connected subgraph that does not contain a cut vertex. The maximum $h(B)$ in any biconnected component B of N is called the *level* of the phylogenetic network N. For all vertices v of N, let $c(v)$ denote the subset of X consisting of the elements x for which there is a directed path in N from v to $\phi(x)$. We call $c(v)$ the *cluster* corresponding to v. A subset C of X is a *cluster* of N if there is some vertex v of N such that $C = c(v)$ and C is non-trivial if $C \neq X$ and $|C| > 1$.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree with root ρ. We define the size of the tree $\mathcal{T}$ to be $|\mathcal{T}| := |X|$ and abbreviate $n := |X|$. Let P be a set of leaves of $\mathcal{T}$. We denote the minimal rooted subtree of $\mathcal{T}$ that connects the leaves of P by $\mathcal{T}(P)$, and the root of $\mathcal{T}(P)$ is the unique degree-two vertex of $\mathcal{T}(P)$ that is closest to the root of $\mathcal{T}$ in $\mathcal{T}$. Furthermore, the *restriction* of $\mathcal{T}$ to P (denoted $T|P$) is the rooted binary phylogenetic tree that is obtained from $\mathcal{T}(P)$ by suppressing all non-root vertices of degree two. For a non-trivial cluster C corresponding to a vertex v of $\mathcal{T}$, we define the *contraction* of $\mathcal{T}$ with respect to C (denoted by $T{\downarrow_C}$) as the result of contracting the subgraph rooted at v in $\mathcal{T}$ onto v, removing all labels of C from X, and giving v a new label (we use the label a_C unless otherwise specified). *Cutting an arc* (u, v) of $\mathcal{T}$ means deleting the arc (u, v) from $\mathcal{T}$, producing disconnected subtrees $\mathcal{T}_u$ and $\mathcal{T}_v$, containing u and v, respectively, and then suppressing u if it has degree two in $\mathcal{T}_u$.

An *unrooted binary phylogenetic network* N on a set X is a graph G containing only vertices of degree three or one, with a bijection ϕ mapping the degree-one vertices of G to X. An *unrooted binary phylogenetic X-tree* (or *unrooted binary phylogenetic tree* on X) is an unrooted binary phylogenetic network on X that is connected and acyclic (a tree). All concepts, except that of a cluster, defined in this section for rooted binary phylogenetic networks/trees can be easily adapted to the unrooted framework by disregarding the root and considering the graph as undirected. In the unrooted framework, we will use the word *edge* instead of *arc*. To avoid confusion, we defer the definition of a cluster for unrooted trees to Section 5.

The Hybridization Number. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $N = (D, \phi)$ be a rooted phylogenetic network on X. We say that N *displays* $\mathcal{T}$ if $\mathcal{T}$ can be obtained from a rooted subtree of N by suppressing degree-two vertices. In other words, $\mathcal{T}$ can be obtained from N by first deleting a subset of

the arcs of D and then deleting isolated vertices and suppressing the non-root degree-two vertices. For two rooted binary phylogenetic X-trees, $\mathcal{T}$ and $\mathcal{T}'$, we define the *hybridization number* of $\mathcal{T}$ and $\mathcal{T}'$ as

$$h(\mathcal{T}, \mathcal{T}') := \min\{h(N) \mid N \text{ displays } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

We also define the *hybridization level* of $\mathcal{T}$ and $\mathcal{T}'$ as the minimum k such that there is a level-k rooted phylogenetic network, i.e. a rooted phylogenetic network with level k, that displays $\mathcal{T}$ and $\mathcal{T}'$. The decision problem, Hybridization Number, is formally stated as follows.

Hybridization Number
Input: Two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, and $l \in \mathbb{N}$.
Question: Is $h(\mathcal{T}, \mathcal{T}') \leq l$?

We can now state our first theorem, whose proof is deferred to Section 3.

Theorem 1 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees.* Hybridization Number *is fixed-parameter tractable with respect to the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$.*

Plugging in current results for Hybridization Number [22], Theorem 1 implies the following.

Corollary 1 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees.* Hybridization Number *can be solved in time $O(3.18^k \cdot n)$, where n is the size of the leaf set of $\mathcal{T}$ and k is the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$.*

It was already known that Hybridization Number is fixed-parameter tractable when parameterized by the hybridization number [5] but our result is stronger as the hybridization level can be small, even 1, for pairs of trees for which the hybridization number is arbitrarily large. On the other hand, it is clear that the hybridization level never exceeds the hybridization number.

The rSPR Problem. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. For the upcoming definition of a rooted subtree prune and regraft operation, we regard the root of $\mathcal{T}$ as a vertex labelled by a dummy taxon l_ρ at the end of a pendant arc adjoined to the original root (for details see [4]. This is done to be able to *regraft* above the original root). Now let $e = (u, v)$ be an arc of $\mathcal{T}$ not incident with the vertex labelled l_ρ. Let $\mathcal{T}'$ be the rooted binary phylogenetic X-tree obtained from $\mathcal{T}$ by deleting e and then reconnecting v to the component $\mathcal{T}_u$ by:

(i) creating a new vertex u' which subdivides an arc in $\mathcal{T}_u$,
(ii) adding the arc (u', v), and
(iii) contracting the degree-two vertex u.

We say that $\mathcal{T}'$ is obtained from $\mathcal{T}$ by one *rooted subtree prune and regraft (rSPR) operation*. We define the *rSPR distance* between two rooted binary

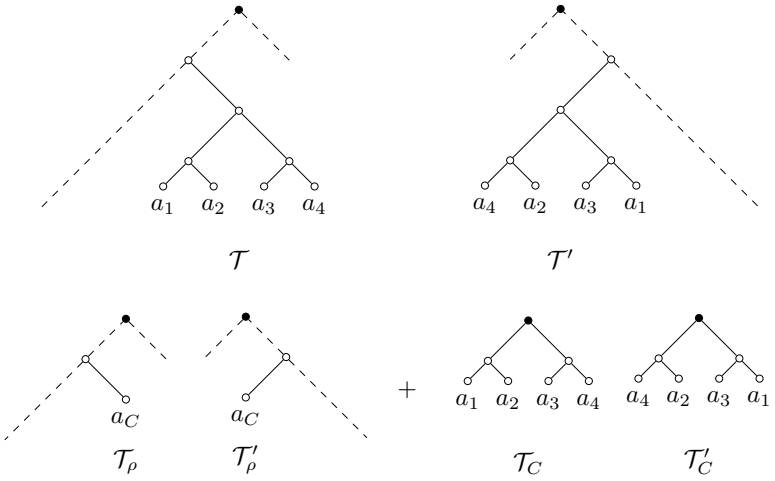

Fig. 1 An example of the rooted cluster reduction. Black vertices are the respective roots.

phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ to be the minimum number of rSPR operations that are required to transform $\mathcal{T}_1$ into $\mathcal{T}_2$. We denote this distance by $d_{\text{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)$. The associated decision problem is the following.

rSPR Distance
Input: Two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ and $l \in \mathbb{N}$.
Question: Is $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') \leq l$?

Our second theorem is an analogue of Theorem 1 for rSPR Distance instead of Hybridization Number. However, in order to define the required parameter, the *rSPR level* of two rooted binary phylogenetic X-trees, we need to define a cluster reduction, following [5].

Definition 1 (rooted cluster reduction) Let $\mathcal{T}$ and $\mathcal{T}'$ be rooted binary phylogenetic X-trees and let C be a non-trivial cluster common to both $\mathcal{T}$ and $\mathcal{T}'$. A *cluster reduction* is the operation of splitting $(\mathcal{T}, \mathcal{T}')$ into the two pairs of smaller trees $(\mathcal{T}_C, \mathcal{T}'_C), (\mathcal{T}_\rho, \mathcal{T}'_\rho) := (\mathcal{T}|C, \mathcal{T}'|C), (\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C})$. Note that $(\mathcal{T}_C, \mathcal{T}'_C)$ is a pair of phylogenetic C-trees, and $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ is a pair of phylogenetic $((X \setminus C) \cup \{a_C\})$-trees that contain the original roots of $\mathcal{T}$ and $\mathcal{T}'$ respectively. See Fig. 1 for an example.

We now define a cluster sequence, which is essentially the result of applying several cluster reductions to a pair of trees. Let $\mathcal{T}$ and $\mathcal{T}'$ be rooted binary phylogenetic X-trees. Set $\hat{\mathcal{T}}_0 = \mathcal{T}$ and $\hat{\mathcal{T}}'_0 = \mathcal{T}'$. For a cluster sequence consisting of t reductions, for $i = 1, \ldots, t$ let A_i be a non-trivial cluster common to both $\hat{\mathcal{T}}_{i-1}$ and $\hat{\mathcal{T}}'_{i-1}$, and define $\mathcal{T}_i := \hat{\mathcal{T}}_{i-1}|A_i$ and $\mathcal{T}'_i := \hat{\mathcal{T}}'|A_i$, and also $\hat{\mathcal{T}}_i := \hat{\mathcal{T}}_{i-1}{\downarrow_{A_i}}$ and $\hat{\mathcal{T}}'_i := \hat{\mathcal{T}}'_{i-1}{\downarrow_{A_i}}$, where the newly created leaf in $\hat{\mathcal{T}}_i$ and $\hat{\mathcal{T}}'_i$ is

labelled by a_i. Finally, we denote $(\hat{\mathcal{T}}_t, \hat{\mathcal{T}}'_t)$ as $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$, to emphasize that these two trees contain the original roots of $\mathcal{T}$ and $\mathcal{T}'$. The result is a sequence of pairs of trees $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ which we call a *cluster sequence*. Note that the leaf set of $\mathcal{T}_i$ and $\mathcal{T}'_i$ is A_i and the leaf set of $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$ is $(X \cup \bigcup_i \{a_i\}) \setminus \bigcup_i A_i$.

We say a cluster sequence is a *full* cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$ if at each step the cluster A_i is a minimal non-trivial common cluster and the trees $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$ contain no further non-trivial common clusters. Observe that the full cluster reduction is unique, up to the ordering of pairs, since any non-trivial common cluster of $\mathcal{T}$ and $\mathcal{T}'$ will at some point become minimal (once all common subclusters have been reduced), and it will then itself be reduced. In addition, no pair $(\mathcal{T}_i, \mathcal{T}'_i)$ in the full cluster reduction contains a non-trivial common cluster.

For two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, the *rSPR level* is the maximum rSPR distance between a pair of trees in a full cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$, i.e. the maximum of $d_{\mathrm{rSPR}}(\mathcal{T}_i, \mathcal{T}'_i)$ over $i \in \{1, \ldots, t, \rho\}$. We may now state the second theorem of the paper whose proof is deferred to Section 4.

Theorem 2 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees.* rSPR Distance *is fixed-parameter tractable with respect to the rSPR level of $\mathcal{T}$ and $\mathcal{T}'$.*

Note that, analogous to the hybridization number, the rSPR level of a pair of trees is at most the rSPR distance between the trees, and may be much smaller, even 1 for trees that have arbitrarily large rSPR distance. Plugging in current results for rSPR Distance [9], Theorem 2 implies the following.

Corollary 2 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees.* rSPR Distance *can be solved in time $O(2.344^k \cdot n)$, where n is the size of the leaf set of $\mathcal{T}$ and k is the rSPR level of $\mathcal{T}$ and $\mathcal{T}'$.*

The TBR Problem. Let $\mathcal{T}$ be an unrooted binary phylogenetic X-tree and $e = \{u, v\}$ be an edge of $\mathcal{T}$ such that neither u nor v is a leaf. Let $\mathcal{T}'$ be the unrooted binary phylogenetic X-tree obtained from $\mathcal{T}$ by deleting e and reconnecting the subtrees $\mathcal{T}_u$ and $\mathcal{T}_v$ by

(i) subdividing an edge of $\mathcal{T}_u$ with a new vertex w,
(ii) subdividing an edge of $\mathcal{T}_v$ with a new vertex z,
(iii) adding the edge $\{w, z\}$, and
(iv) suppressing any vertices of degree two.

The decision problem TBR Distance is formally stated as follows.

TBR Distance
Input: Two unrooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ and $l \in \mathbb{N}$.
Question: Is $d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') \leq l$?

Note that the notions of *displaying*, *hybridization number* and *hybridization level* of two unrooted trees are defined as in the rooted framework. Our third theorem is an analogue of Theorem 1 for TBR Distance instead of Hybridization Number.

Theorem 3 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two unrooted binary phylogenetic X-trees.* TBR Distance *is fixed-parameter tractable with respect to the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$.*

Plugging in current results for TBR Distance [8], Theorem 3 implies the following.

Corollary 3 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two unrooted binary phylogenetic X-trees.* TBR Distance *can be solved in time $O(3^k \cdot n)$, where n is the size of the leaf set of $\mathcal{T}$ and k is the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$.*

Note that the unrooted hybridization level is always smaller or equal to the TBR distance, since the unrooted hybridization number equals the TBR distance (see Theorem 6 in Section 5).

3 Proof of Theorem 1

The following lemma shows how the cluster reduction can be used as part of a divide-and-conquer approach to computing the hybridization number.

Lemma 1 ([3]) *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Suppose that $C \subset X$ is a cluster of both $\mathcal{T}$ and $\mathcal{T}'$, where $(\mathcal{T}_C, \mathcal{T}'_C)$ and $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ are the results of performing a cluster reduction of C on $(\mathcal{T}, \mathcal{T}')$. Then,*

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

A straightforward consequence of Lemma 1 is that if $(\mathcal{T}_1, \mathcal{T}'_1), \cdots, (\mathcal{T}_t, \mathcal{T}'_t)$, $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ is a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$, then

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

Next, we show that the hybridization level of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is equal to the maximum hybridization number between a pair of trees in a full cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$. Recall that, for a rooted phylogenetic network N, its level is the maximum number of reticulation vertices in any biconnected component of N.

Lemma 2 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees and let $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ be a full cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$. Then, the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$ equals*

$$\max_{i \in \{1,\ldots,t,\rho\}} h(\mathcal{T}_i, \mathcal{T}'_i).$$

Proof For each $i \in \{1, \ldots, t\}$, let N_i be a rooted phylogenetic network displaying $\mathcal{T}_i$ and $\mathcal{T}'_i$ with hybridization number $h(\mathcal{T}_i, \mathcal{T}'_i)$ and let A_i and a_i denote the set of leaves of $\mathcal{T}_i$ and the new leaf created to represent the cluster A_i in the i^{th} cluster reduction, respectively. We may now rebuild a rooted phylogenetic network N displaying $\mathcal{T}$ and $\mathcal{T}'$ from the smaller rooted phylogenetic networks N_i as follows. We start with $N = N_\rho$. While N contains a leaf v labelled a_i for some i, we replace v by a pendant copy of N_i in N. Since each arc incident with such a leaf is a cut arc of the resulting rooted phylogenetic network N, each biconnected component of N is a subnetwork of N_i for some $i \in \{1, \ldots, t, \rho\}$. Thus, N displays $\mathcal{T}$ and $\mathcal{T}'$ and the level of N is at most the maximum of $h(\mathcal{T}_i, \mathcal{T}'_i)$ over $i \in \{1, \ldots, t, \rho\}$, hence the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$ is at most the maximum of $h(\mathcal{T}_i, \mathcal{T}'_i)$ over $i \in \{1, \ldots, t, \rho\}$.

Conversely, let N be any rooted phylogenetic network displaying $\mathcal{T}$ and $\mathcal{T}'$ and let k denote its level. Let the vertex set of N be V and the root be ρ. We will construct a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$. Each cut arc (u, v) of N gives rise to a cluster $c(v)$ which is a common cluster to $\mathcal{T}$ and $\mathcal{T}'$. A cut arc (u, v) of N is trivial if v is a leaf of N, and it is a minimal non-trivial cut arc if there is no other non-trivial cut arc (w, x) of N such that there is a directed path from v to w in N. We obtain a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$ by iteratively:

- selecting v in V at the head of a minimal non-trivial cut arc of N, which gives rise to $c(v)$, a minimal non-trivial common cluster of $\mathcal{T}$ and $\mathcal{T}'$;
- performing the cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$ by $c(v)$ replacing the cluster with a new vertex c_v, and
- replacing the subnetwork below the cut edge with a single pendant leaf c_v in N

Note that the deleted subnetwork is either a subtree (in fact, due to minimality, just a pair of leaves with a common parent, which is known as a *cherry*) or a biconnected component of N with pendant leaves, since otherwise, we could choose a smaller common cluster. Since the level of the network is k, this subnetwork of N is a phylogenetic network on $c(v)$ containing at most k hybridization vertices and displaying $\mathcal{T}|c(v)$ and $\mathcal{T}'|c(v)$. Hence the cluster pair in the cluster reduction has hybridization number at most k. We repeat this process until N has no further cut arcs, obtaining a cluster sequence $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ for $\mathcal{T}$ and $\mathcal{T}'$. Every cluster pair $(\mathcal{T}_i, \mathcal{T}'_i)$ from the cluster sequence has hybridization number at most k. It remains to consider the final pair $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$. Since in the end N had no (non-trivial) cut arcs, either N was reduced to a cherry or N was a biconnected component with pendant leaves, and again we deduce that $h(\mathcal{T}_\rho, \mathcal{T}'_\rho) \leq k$. Thus if $\mathcal{T}$ and $\mathcal{T}'$ can be displayed on a level-k phylogenetic network, then there is a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$ such that the maximum hybridization number between a pair of trees in the cluster reduction is at most k.

It remains to show that the maximum hybridization number between a pair of trees in the *full* cluster reduction is therefore also at most k. We will make use of the fact that if a cluster reduction is not a reduction by a minimal non-trivial common cluster, then it can be broken down into a series of cluster

reductions each of which is by a minimal non-trivial common cluster. To see this consider a cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$ by a common cluster A and suppose it is not a minimal non-trivial common cluster. Then, there is a subset $A_1 \subset A$ such that A_1 is a minimal non-trivial common cluster. We first reduce by A_1, obtaining $(\mathcal{T}_{A_1}, \mathcal{T}'_{A_1}), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$, where there is a leaf a_1 in $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$ replacing the cluster A_1. We may then reduce by the common cluster $A \cup \{a_1\} \setminus A_1$ of $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$. This has broken the cluster reduction by A into a minimal cluster reduction by A_1 and a cluster reduction by a proper subset of A. By repeating this process until the remaining reduction is itself by a minimal non-trivial common cluster, we iteratively break down the cluster reduction by A into a sequence of cluster reductions, each of which is by a minimal non-trivial common cluster.

So we first form a full cluster reduction from $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ by following the same sequence of cluster reductions used to create the cluster sequence, but at each step where we would reduce $\mathcal{T}$ and $\mathcal{T}'$ by a common cluster A, we instead reduce by a sequence of minimal non-trivial common clusters, as described above, whose union contains all the elements of A. Finally, once we have finished breaking down the cluster reductions in the original cluster sequence, we continue to perform cluster reductions on $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$ by any remaining minimal common clusters until none remain. The result is a full cluster reduction $(\hat{\mathcal{T}}_1, \hat{\mathcal{T}}'_1), ..., (\hat{\mathcal{T}}_s, \hat{\mathcal{T}}'_s), (\hat{\mathcal{T}}_\rho, \hat{\mathcal{T}}'_\rho)$ such that each pair $(\mathcal{T}_i, \mathcal{T}'_i)$ of the original cluster sequence corresponds to a subsequence $(\hat{\mathcal{T}}_j, \hat{\mathcal{T}}'_j), \ldots, (\hat{\mathcal{T}}_q, \hat{\mathcal{T}}'_q)$ of the full cluster reduction, in the sense that $(\hat{\mathcal{T}}_j, \hat{\mathcal{T}}'_j), \ldots, (\hat{\mathcal{T}}_q, \hat{\mathcal{T}}'_q)$ is itself a cluster reduction of $(\mathcal{T}_i, \mathcal{T}'_i)$. Then, by Lemma 1,

$$h(\mathcal{T}_i, \mathcal{T}'_i) = \sum_{j \le l \le q} h(\hat{\mathcal{T}}_l, \hat{\mathcal{T}}'_l) \ge \max_{j \le l \le q} h(\hat{\mathcal{T}}_l, \hat{\mathcal{T}}'_l),$$

implying

$$k \ge \max_{i \in \{1,\ldots,t,\rho\}} h(\mathcal{T}_i, \mathcal{T}'_i) \ge \max_{j \in \{1,\ldots,s,\rho\}} h(\hat{\mathcal{T}}_j, \hat{\mathcal{T}}'_j),$$

and, since this holds for every phylogenetic network N displaying $\mathcal{T}$ and $\mathcal{T}'$, whatever the level of N, the lemma follows. □

From Lemmas 1 and 2 it follows that there is a network displaying $\mathcal{T}$ and $\mathcal{T}'$ minimizing the hybridization level that also minimizes the hybridization number.

Lemma 3 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. A full cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$ can be computed in time $O(n)$, where n is the size of the leaf set of $\mathcal{T}$.*

Proof We start by applying the algorithm in [11] to $\mathcal{T}$, which preprocesses $\mathcal{T}$ in time $O(n)$ and creates a data structure that returns the least common ancestor (LCA) of any two specific vertices of $\mathcal{T}$ in $O(1)$ time. Then, we compute, for each vertex x of $\mathcal{T}$, the number $l(x)$ of leaves below it in $O(n)$ total time. We do the same for $\mathcal{T}'$. Finally, for each vertex x of $\mathcal{T}$, we store the vertex x' of

$\mathcal{T}'$ with $x' := \mathrm{LCA}_{\mathcal{T}'}(c(x))$ as $m(x)$. Since, assuming the children of x are y and z, we have $m(x) = \mathrm{LCA}_{\mathcal{T}'}(m(y), m(z))$, this can be done in $O(n)$ time via a post-order traversal of $\mathcal{T}'$ using the precomputed data structure. Then, a cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$ can be found as follows:

1 $i \leftarrow 1$;
2 **for** *x in a post-order traversal of $\mathcal{T}$* **do**
3 **if** *$l(x) \geq 2$, $l(x) = l(m(x))$ and x is not the root of $\mathcal{T}$* **then**
4 $A_i \leftarrow c(x)$;
5 $(\mathcal{T}_i, \mathcal{T}'_i) \leftarrow (\mathcal{T}_{A_i}, \mathcal{T}'_{A_i})$;
6 reduce A_i to a single leaf a_i in both $\mathcal{T}$ and $\mathcal{T}'$;
7 $i \leftarrow i + 1$;
8 $(\mathcal{T}_\rho, \mathcal{T}'_\rho) \leftarrow (\mathcal{T}, \mathcal{T}')$;

The overall worst-case running time of this algorithm is $O(n)$; indeed, although there are $O(n)$ iterations of the outer loop, each one involving reducing a cluster A_i of size $O(n)$ in line 6, the sum of the sizes of the clusters is at most $O(n)$, and so the amortized running-time of this line is $O(1)$. □

We are now in a position to prove Theorem 1 and Corollary 1.

Proof of Theorem 1 and Corollary 1 Let the two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ and the integer l be an instance of Hybridization Number. Let $|X| = n$, and let k be the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$. We may first compute a full cluster reduction $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ of $\mathcal{T}$ and $\mathcal{T}'$ in time $O(n)$ by Lemma 3. We then apply the algorithm of [22] to each pair $(\mathcal{T}_i, \mathcal{T}'_i)$ to obtain $h(\mathcal{T}_i, \mathcal{T}'_i)$ in time $O(3.18^{h(\mathcal{T}_i, \mathcal{T}'_i)} \cdot |\mathcal{T}_i|)$. By Lemma 2, $h(\mathcal{T}_i, \mathcal{T}'_i) \leq k$, and clearly $\sum_i |\mathcal{T}_i| = O(n)$, hence we may compute $h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}_1, \mathcal{T}'_1) + ... + h(\mathcal{T}_t, \mathcal{T}'_t) + h(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ in time $O(3.18^k \cdot n)$. By a comparison of $h(\mathcal{T}, \mathcal{T}')$ and l we may answer the decision problem in the same time bound, and hence Hybridization Number is fixed parameter tractable when parameterized by the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$. □

4 Proof of Theorem 2

Recall that, for solving instances of rSPR Distance with two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, we add to each of them a vertex labelled by a dummy taxon l_ρ at the end of a pendant edge adjoined to the original root. Given such an "augmented" tree $\mathcal{T}$ and a label x, let $\mathcal{T}|_{l_\rho \to x}$ denote the result of removing the vertex labelled l_ρ and replacing the label x by l_ρ. In the following, we make use of the concept of *rooted agreement forests*: Given two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, a leaf-labelled forest $\mathcal{F}$ is called a *rooted agreement forest* of $\mathcal{T}$ and $\mathcal{T}'$ if $\mathcal{F}$ can be obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by a series of edge cuts as defined in Section 2. We say that a rooted agreement forest is *root-isolating* if it contains the singleton tree that consists of the leaf labelled l_ρ. A rooted agreement forest for a cluster sequence

$(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, is a leaf-labelled forest $\mathcal{F}$ on $X \cup \{a_1, \ldots, a_t\}$ which can be obtained from the forests $\{\mathcal{T}_1, ..., \mathcal{T}_t, \mathcal{T}_\rho\}$ and $\{\mathcal{T}'_1, ..., \mathcal{T}'_t, \mathcal{T}'_\rho\}$ by a series of edge cuts.

For the proof of Theorem 2, we need to define the concept of *cluster hierarchy*: the cluster hierarchy for a full cluster sequence $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is defined as the directed tree with a vertex for each component $(\mathcal{T}_i, \mathcal{T}'_i)$ of the cluster sequence, and a directed edge from vertex $(\mathcal{T}_i, \mathcal{T}'_i)$ to vertex $(\mathcal{T}_j, \mathcal{T}'_j)$ if a leaf labelled by a_j is present in $\mathcal{T}'_i$. Then, by starting with $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ as the root of the tree, and using a breadth-first search, since $t < n$ we have the following:

Observation 1 *The* cluster hierarchy *for a full cluster sequence can be computed in time $O(n)$, where n is the size of the leaf set of $\mathcal{T}$.*

For the proof of Theorem 2, we will also make use of the *Minimum-Weight Forest Algorithm* of Linz and Semple [16], which establishes the correctness of the use of a cluster reduction in a divide-and-conquer approach for computing the rSPR distance. In particular, they offer the following theorem and algorithm.

Theorem 4 (Theorem 2.2 of [16]) *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ be a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{G}$ be a rooted agreement forest for this sequence of minimum weight $w(\mathcal{G})$. Then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = w(\mathcal{G}) - 1$.*

Algorithm Minimum-Weight Forest [16]
Input: A cluster sequence $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, along with its cluster hierarchy.
Output: The minimum weight of a rooted agreement forest for this sequence.

Without needing to give a precise definition of a minimum-weight rooted agreement forest for a cluster sequence (for details see [16]), it suffices to note that if we start with two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, first compute a full cluster reduction and its cluster hierarchy, and then apply the *Minimum-Weight Forest* algorithm, our output is one more than the rSPR distance between $\mathcal{T}$ and $\mathcal{T}'$. It remains to bound the running time of this approach. To do so, we need the following lemma:

Lemma 4 *Let $\mathcal{T}$ and $\mathcal{T}'$ be rooted binary phylogenetic X-trees and let $x \in X$. Then, there is a root-isolating rooted maximum-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ if and only if $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{T}|_{l_\rho \to x}, \mathcal{T}'|_{l_\rho \to x}) + 1$.*

Proof Let $\mathcal{T}_* := \mathcal{T}|_{l_\rho \to x}$ and $\mathcal{T}'_* := \mathcal{T}'|_{l_\rho \to x}$.

"$\Rightarrow$": Let $\mathcal{F}$ be a root-isolating rooted maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ and let $\mathcal{T}_\rho$ be the tree in $\mathcal{F}$ that consists of the singleton labelled l_ρ. Then, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = |\mathcal{F}|$. Let $\mathcal{F}'$ be the result of removing $\mathcal{T}_\rho$ from $\mathcal{F}$ and relabelling the leaf labelled x by l_ρ. Clearly, $\mathcal{F}'$ is a rooted agreement forest for $\mathcal{T}_*$ and $\mathcal{T}'_*$ and, thus, $d_{\mathrm{rSPR}}(\mathcal{T}_*, \mathcal{T}'_*) \leq |\mathcal{F}'| = |\mathcal{F}| - 1$. To show that $\mathcal{F}'$

maximizes agreement, assume towards a contradiction that there is a rooted agreement forest $\mathcal{F}^*$ for $\mathcal{T}_*$ and $\mathcal{T}'_*$ with $|\mathcal{F}^*| < |\mathcal{F}'|$. Then, relabelling the leaf labelled l_ρ by x in $\mathcal{F}^*$ and adding $\mathcal{T}_\rho$ to $\mathcal{F}^*$ yields a rooted agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with $|\mathcal{F}^*| + 1 < |\mathcal{F}|$ components, contradicting optimality of $\mathcal{F}$.

"$\Leftarrow$": Let $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{T}_*, \mathcal{T}'_*) + 1$. We construct a root-isolating rooted maximum-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$. To this end, let $\mathcal{F}_*$ be a rooted maximum-agreement forest for $\mathcal{T}_*$ and $\mathcal{T}'_*$ and let $\mathcal{F}$ be the result of relabelling the leaf labelled l_ρ by x in $\mathcal{F}_*$ and adding a singleton tree whose only vertex is labelled l_ρ. Then, $|\mathcal{F}| = |\mathcal{F}_*| + 1 = d_{\mathrm{rSPR}}(\mathcal{T}_*, \mathcal{T}'_*) + 1 = d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. Thus, $\mathcal{F}$ is a root-isolating rooted maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. □

We have now all the building blocks to prove the main results of this section.

Proof of Theorem 2 and Corollary 2. Let the two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ and the integer l be an instance of rSPR Distance. Let $|X| = n$, and let k be the rSPR level of $\mathcal{T}$ and $\mathcal{T}'$. We may first compute a full cluster reduction $(\mathcal{T}_1, \mathcal{T}'_1), ..., (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ of $\mathcal{T}$ and $\mathcal{T}'$ and its cluster hierarchy in time $O(n)$ by Lemma 3 and Observation 1. We then apply the algorithm of [16] to obtain $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. The time-consuming step in this algorithm is finding a maximum-agreement forest for each pair $\mathcal{T}_i, \mathcal{T}'_i$ (if possible a root-isolating one). These may be found, using Lemma 4, in time $O(2.344^{d_{\mathrm{rSPR}}(\mathcal{T}_i, \mathcal{T}'_i)} \cdot |\mathcal{T}_i|)$ by the approach of [9]. By definition, $d_{\mathrm{rSPR}}(\mathcal{T}_i, \mathcal{T}'_i) \leq k$ and, clearly, $|\mathcal{T}_i| \in O(n)$. Hence, the whole algorithm runs in time $O(2.344^k \cdot n)$. By a comparison of $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ and l we may answer the decision problem in the same time bound, and hence rSPR Distance is fixed parameter tractable when parameterized by the rSPR level of $\mathcal{T}$ and $\mathcal{T}'$. □

Note that the hybridization number of two trees is always bigger than their rSPR distance [2], and so Lemma 2 and Corollary 2 imply the following:

Corollary 4 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees.* rSPR Distance *can be solved in time $O(2.344^k \cdot n)$, where n is the size of the leaf set of $\mathcal{T}$ and k is the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$.*

Note also that the authors of [22] claim to have an algorithm to solve rSPR Distance in $O(2^{d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')} \cdot n)$ [23]. If this is true, the running time in Corollaries 2 and 4 will reduce to $O(2^k \cdot n)$.

5 Proof of Theorem 3

In this section, we consider unrooted binary phylogenetic X-trees. Note that each edge e of any phylogenetic X-tree uniquely partitions X into nonempty sets C and $\overline{C} := X \setminus C$ such that all paths between a leaf labelled with an element of C and a leaf labelled with an element of $\overline{C}$ contain e. A set C for which such an edge exists in $\mathcal{T}$ is called a *cluster* of $\mathcal{T}$. A cluster is called *trivial* if $|C| = 1$ or $|\overline{C}| = 1$. Given an unrooted binary phylogenetic X-tree

$\mathcal{T}$ and a nontrivial cluster C of $\mathcal{T}$, let $\mathcal{T}|C$ denote the minimal subtree of $\mathcal{T}$ containing each leaf whose label is in C (analogous to the rooted case) and denote by $\mathcal{T}{\downarrow_C}$ the unrooted phylogenetic tree where $\mathcal{T}|C$ has been replaced by a leaf labelled by a_C. An *unrooted agreement forest* (uAF) for two unrooted phylogenetic X-trees is the unrooted version of a rooted agreement forest: it is a leaf-labelled forest $\mathcal{F}$ that can be obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by a series of edge deletions, deletions of unlabeled leaves, and suppressions of degree-two vertices. A uAF of minimal cardinality is called an unrooted *maximum*-agreement forest (uMAF). $\mathcal{F}$ is said to *isolate* some $x \in X$ if $\mathcal{F}$ contains a singleton tree consisting of the leaf labelled x (denoted by $\{x\} \in \mathcal{F}$). Finally, we denote the number of trees in $\mathcal{F}$ by $|\mathcal{F}|$.

In the following, we describe a cluster reduction for unrooted binary phylogenetic trees, slightly different from the rooted case.

Definition 2 (unrooted cluster reduction) Let $\mathcal{T}$ and $\mathcal{T}'$ be unrooted binary phylogenetic trees and let C be a non-trivial cluster common to both $\mathcal{T}$ and $\mathcal{T}'$ (note that $\overline{C}$ is also a common cluster of $\mathcal{T}$ and $\mathcal{T}'$). A *cluster reduction* is the operation of splitting $(\mathcal{T}, \mathcal{T}')$ into the two pairs of smaller trees $(\mathcal{T}_C, \mathcal{T}'_C), (\mathcal{T}_{\overline{C}}, \mathcal{T}'_{\overline{C}}) := (\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C}), (\mathcal{T}{\downarrow_{\overline{C}}}, \mathcal{T}'{\downarrow_{\overline{C}}})$. See Fig. 2 for an example.

Analogously to the rooted case, we call the result $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t)$ of repeatedly applying the cluster reduction to two unrooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ a *cluster sequence* for $\mathcal{T}$ and $\mathcal{T}'$ and such a sequence is called *full* if each cluster reduction leading to the sequence reduces a *minimal* non-trivial common cluster and the trees $\mathcal{T}_t$ and $\mathcal{T}'_t$ contain no further non-trivial common clusters. Again, the full cluster reduction is unique, up to the ordering of pairs and no pair $(\mathcal{T}_i, \mathcal{T}'_i)$ in the full cluster reduction contains a non-trivial common cluster.

Note that an unrooted cluster sequence can be computed as described in Lemma 3 by previously rooting the two trees on the same leaf.

The following results are fundamental for proving that TBR Distance is FPT in the hybridization level.

Theorem 5 ([1]) *Let $\mathcal{T}$ and $\mathcal{T}'$ be two unrooted binary phylogenetic X-trees. Let $\mathcal{F}$ be a uMAF for $\mathcal{T}$ and $\mathcal{T}'$. Then $d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = |\mathcal{F}| - 1$.*

Theorem 6 ([13]) *Let $\mathcal{T}$ and $\mathcal{T}'$ be unrooted binary phylogenetic X-trees. Then $h(\mathcal{T}, \mathcal{T}') = d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}')$.*

Note that the concepts of *hybridization number* and *level* refer to the *undirected* versions. The following observation is straightforward.

Observation 2 *A forest $\mathcal{F} = \{F_1, \ldots, F_k\}$ is a uAF of $\mathcal{T}$ and $\mathcal{T}'$ if and only if*

1. *each tree of $\mathcal{F}$ is displayed by both $\mathcal{T}$ and $\mathcal{T}'$,*
2. *all labels of $\mathcal{T}$ and $\mathcal{T}'$ occur in $\mathcal{F}$, and*
3. *the subtrees $\mathcal{T}(\mathcal{L}(F_1)), \ldots, \mathcal{T}(\mathcal{L}(F_k))$ and $\mathcal{T}'(\mathcal{L}(F_1)), \ldots, \mathcal{T}'(\mathcal{L}(F_k))$ are all vertex disjoint.*

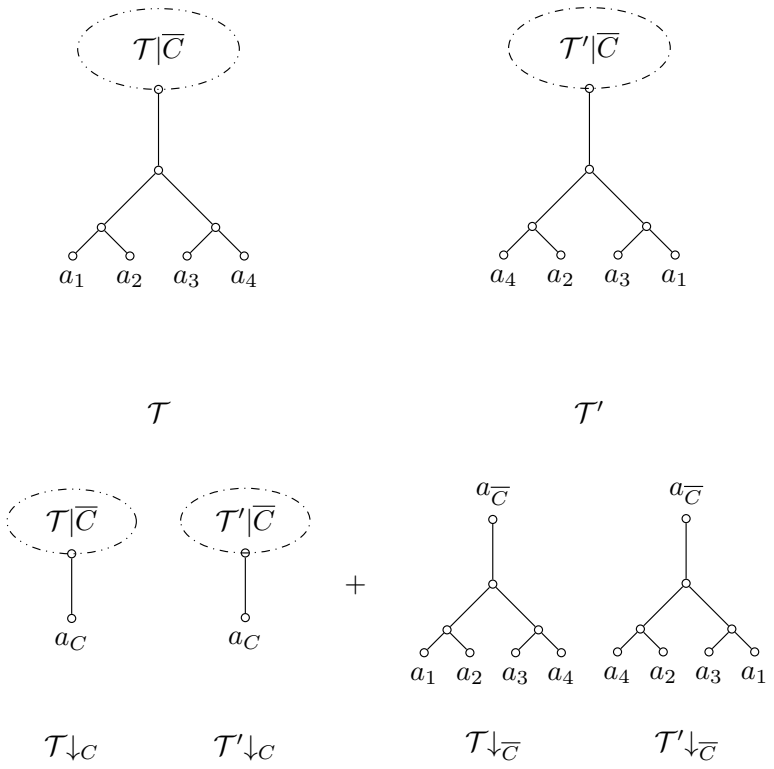

Fig. 2 An example of an unrooted cluster reduction. The common cluster is $C = \{a_1, a_2, a_3, a_4\}$.

The following two lemmas constitute a portation of Lemma 4 and Lemma 1 to unrooted binary phylogenetic trees.

Lemma 5 *Let $\mathcal{T}$ and $\mathcal{T}'$ be unrooted binary phylogenetic X-trees and let $x \in X$. If there is a uMAF $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ that isolates x, then*

$$d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{TBR}}(\mathcal{T}|(X-x), \mathcal{T}'|(X-x)) + 1$$

and, otherwise,

$$d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{TBR}}(\mathcal{T}|(X-x), \mathcal{T}'|(X-x)).$$

Proof Let $\mathcal{F}'$ be a uMAF for $\mathcal{T}|(X-x)$ and $\mathcal{T}'|(X-x)$.

First, suppose that there is a uMAF $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ that isolates x. Then, $\mathcal{F}$ can be turned into a uAF for $\mathcal{T}|(X-x)$ and $\mathcal{T}'|(X-x)$ by deleting the singleton tree containing x and $\mathcal{F}'$ can be turned into a uAF for $\mathcal{T}$ and $\mathcal{T}'$ by adding a singleton tree containing a vertex labelled x. Thus, $|\mathcal{F}| = |\mathcal{F}'| + 1$.

Next, suppose that there is no uMAF for $\mathcal{T}$ and $\mathcal{T}'$ that isolates x and let $\mathcal{F}$ be a uMAF for $\mathcal{T}$ and $\mathcal{T}'$. Since adding a singleton tree containing a

vertex labelled x to $\mathcal{F}'$ yields a uAF for $\mathcal{T}$ and $\mathcal{T}'$ that isolates x, we have $|\mathcal{F}| < |\mathcal{F}'| + 1$. However, since removing x from the tree of $\mathcal{F}$ that contains x yields a uAF for $\mathcal{T}|(X - x)$ and $\mathcal{T}'|(X - x)$, we also have $|\mathcal{F}| \geq |\mathcal{F}'|$. Thus, $|\mathcal{F}| = |\mathcal{F}'|$. The lemma follows by Theorem 5. □

Lemma 6 *Let $\mathcal{T}$ and $\mathcal{T}'$ be unrooted binary phylogenetic X-trees and let C be a nontrivial cluster of $\mathcal{T}$ and $\mathcal{T}'$. If there is a uMAF for $\mathcal{T}{\downarrow_C}$ and $\mathcal{T}'{\downarrow_C}$ that isolates the leaf labelled a_C, then*

$$d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{TBR}}(\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C}) + d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C),$$

and, otherwise,

$$d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{TBR}}(\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C}) + d_{\mathrm{TBR}}(\mathcal{T}{\downarrow_{\overline{C}}}, \mathcal{T}'{\downarrow_{\overline{C}}}).$$

Proof First off, suppose that there is a uMAF for $\mathcal{T}{\downarrow_C}$ and $\mathcal{T}'{\downarrow_C}$ that isolates the leaf labelled a_C.

"$\leq$": Let $\mathcal{F}_C$ be a uMAF for $\mathcal{T}|C$ and $\mathcal{T}'|C$. Let $\mathcal{F}_{\overline{C}}$ be analogous for $\overline{C}$. Let $\mathcal{F}' := \mathcal{F}_{\overline{C}} \uplus \mathcal{F}_C$. Then, all trees of $\mathcal{F}'$ are displayed by $\mathcal{T}$ and $\mathcal{T}'$ and by Observation 2, $\mathcal{F}'$ is a uAF for $\mathcal{T}$ and $\mathcal{T}'$. Thus,

$$\begin{aligned} d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') &\leq |\mathcal{F}'| - 1 \\ &= |\mathcal{F}_{\overline{C}}| + |\mathcal{F}_C| - 1 \\ &\overset{\text{Theorem 5}}{=} d_{\mathrm{TBR}}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C}) + d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) + 1 \\ &\overset{\text{Lemma 5}}{=} d_{\mathrm{TBR}}(\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C}) + d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) \end{aligned}$$

"$\geq$": Let $\mathcal{F}$ be a uMAF for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{F}(C)$ denote the set containing exactly the trees of $\mathcal{F}$ that contain only leaves labelled by elements of C. Let $\mathcal{F}(\overline{C})$ be defined analogously for $\overline{C}$.

Case 1: $\mathcal{F} = \mathcal{F}(C) \uplus \mathcal{F}(\overline{C})$. Then, $|\operatorname{uMAF}(\mathcal{T}|C, \mathcal{T}'|C)| = |\mathcal{F}(C)|$ since, otherwise, exchanging $\mathcal{F}(C)$ for a uMAF of $\mathcal{T}|C$ and $\mathcal{T}'|C$ in $\mathcal{F}$ yields a uAF that is smaller than $\mathcal{F}$, contradicting optimality of $\mathcal{F}$. Likewise, $|\operatorname{uMAF}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C})| = |\mathcal{F}(\overline{C})|$. Then,

$$\begin{aligned} d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = |\mathcal{F}| - 1 &= |\mathcal{F}(C)| + |\mathcal{F}(\overline{C})| - 1 \\ &\overset{\text{Theorem 5}}{=} d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) + d_{\mathrm{TBR}}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C}) + 1 \\ &\overset{\text{Lemma 5}}{=} d_{\mathrm{TBR}}(\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C}) + d_{\mathrm{TBR}}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C}) \end{aligned}$$

Case 2: There is a tree H in $\mathcal{F}$ containing a leaf labelled $x \in C$ and a leaf labelled $y \in \overline{C}$ (note that only one of such "mixed" trees can be present in $\mathcal{F}$; indeed, since C is a cluster of both trees, the existence of two such trees will contradict Condition 3 of Observation 2). Then, $\mathcal{F} = \mathcal{F}(C) \uplus \mathcal{F}(\overline{C}) \uplus \{H\}$. Let $H{\downarrow_C}$ denote the result of contracting all edges of H that are on a path between two leaves with labels of C in H and labelling the vertex on which they are all contracted with C. Let $H{\downarrow_{\overline{C}}}$ be analogous for $\overline{C}$. Then, all labels of C and the special label $a_{\overline{C}}$ occur in $\mathcal{F}_1 := \mathcal{F}(C) \uplus \{H{\downarrow_{\overline{C}}}\}$ and all its trees

are displayed by $\mathcal{T}\downarrow_{\overline{C}}$ and $\mathcal{T}'\downarrow_{\overline{C}}$. Thus, by Observation 2, $\mathcal{F}_1$ is a uAF for $\mathcal{T}\downarrow_{\overline{C}}$ and $\mathcal{T}'\downarrow_{\overline{C}}$. Likewise, $\mathcal{F}(\overline{C}) \uplus \{H\downarrow_C\}$ is a uAF for $\mathcal{T}\downarrow_C$ and $\mathcal{T}'\downarrow_C$. Thus,

$$\begin{aligned} d_{\mathrm{TBR}}(\mathcal{T},\mathcal{T}') = |\mathcal{F}| - 1 &= |\mathcal{F}(C) \uplus \mathcal{F}(\overline{C}) \uplus \{H\}| - 1 \\ &= |\mathcal{F}(C) \uplus \{H\downarrow_{\overline{C}}\}| + |\mathcal{F}(\overline{C}) \uplus \{H\downarrow_C\}| - 2 \\ &\geq d_{\mathrm{TBR}}(\mathcal{T}\downarrow_C, \mathcal{T}'\downarrow_C) + d_{\mathrm{TBR}}(\mathcal{T}\downarrow_{\overline{C}}, \mathcal{T}'\downarrow_{\overline{C}}) \\ &\geq d_{\mathrm{TBR}}(\mathcal{T}\downarrow_C, \mathcal{T}'\downarrow_C) + d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) \end{aligned}$$

Next, suppose that there is no uMAF for $\mathcal{T}\downarrow_C$ and $\mathcal{T}'\downarrow_C$ that isolates the leaf labelled a_C.

"$\leq$": First, note that if there is a uMAF for $\mathcal{T}\downarrow_{\overline{C}}$ and $\mathcal{T}'\downarrow_{\overline{C}}$ that isolates the leaf labelled $a_{\overline{C}}$, the first part of our proof implies that

$$\begin{aligned} d_{\mathrm{TBR}}(\mathcal{T},\mathcal{T}') &= d_{\mathrm{TBR}}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C}) + d_{\mathrm{TBR}}(\mathcal{T}\downarrow_{\overline{C}}, \mathcal{T}'\downarrow_{\overline{C}}) \\ &\leq d_{\mathrm{TBR}}(\mathcal{T}\downarrow_C, \mathcal{T}'\downarrow_C) + d_{\mathrm{TBR}}(\mathcal{T}\downarrow_{\overline{C}}, \mathcal{T}'\downarrow_{\overline{C}}). \end{aligned}$$

Now, let consider the case where there is no uMAF for $\mathcal{T}\downarrow_C$ and $\mathcal{T}'\downarrow_C$ (respectively $\mathcal{T}\downarrow_{\overline{C}}$ and $\mathcal{T}'\downarrow_{\overline{C}}$) that isolates the leaf labelled a_C (respectively labelled $a_{\overline{C}}$). Let $\mathcal{F}_C$ be a uMAF for $\mathcal{T}\downarrow_{\overline{C}}$ and $\mathcal{T}'\downarrow_{\overline{C}}$ and let $H_{\overline{C}}$ denote the tree of $\mathcal{F}_C$ containing the label $a_{\overline{C}}$. Let $\mathcal{F}_{\overline{C}}$ and H_C be analogous for C. Let H be the result of joining H_C and $H_{\overline{C}}$ by identifying the leaves labelled $a_{\overline{C}}$ and a_C, respectively and suppressing this degree-two vertex. Let $\mathcal{F}' := (\mathcal{F}_{\overline{C}} \setminus \{H_C\}) \uplus (\mathcal{F}_C \setminus \{H_{\overline{C}}\}) \uplus \{H\}$. Then, H is displayed by $\mathcal{T}$ and $\mathcal{T}'$ and, thus, all trees of $\mathcal{F}'$ are displayed by $\mathcal{T}$ and $\mathcal{T}'$. Moreover, it is easy to see that $\mathcal{T}(H)$ is vertex disjoint with the other trees in the forest, and the same holds for $\mathcal{T}'(H)$. Then, by Observation 2, $\mathcal{F}'$ is a uAF for $\mathcal{T}$ and $\mathcal{T}'$. Thus,

$$\begin{aligned} d_{\mathrm{TBR}}(\mathcal{T},\mathcal{T}') &\leq |\mathcal{F}'| - 1 \\ &= |\mathcal{F}_{\overline{C}} \setminus \{H_C\}| + |\mathcal{F}_C \setminus \{H_{\overline{C}}\}| + |\{H\}| - 1 \\ &= |\mathcal{F}_{\overline{C}}| + |\mathcal{F}_C| - 2 \\ &= d_{\mathrm{TBR}}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C}) + d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) \\ &\overset{\text{Lemma 5}}{=} d_{\mathrm{TBR}}(\mathcal{T}\downarrow_C, \mathcal{T}'\downarrow_C) + d_{\mathrm{TBR}}(\mathcal{T}\downarrow_C, \mathcal{T}'\downarrow_C) \end{aligned}$$

"$\geq$": Let $\mathcal{F}$ be a uMAF for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{F}(C)$ denote the set containing exactly the trees of $\mathcal{F}$ that contain only leaves labelled by elements of C. Let $\mathcal{F}(\overline{C})$ be defined analogously for $\overline{C}$.

Case 1: $\mathcal{F} = \mathcal{F}(C) \uplus \mathcal{F}(\overline{C})$. Then, $|\operatorname{uMAF}(\mathcal{T}|C, \mathcal{T}'|C)| = |\mathcal{F}(C)|$ since, otherwise, exchanging $\mathcal{F}(C)$ for a uMAF of $\mathcal{T}|C$ and $\mathcal{T}'|C$ in $\mathcal{F}$ yields a uAF that is smaller than $\mathcal{F}$, contradicting optimality of $\mathcal{F}$. Likewise, $|\operatorname{uMAF}(\mathcal{T}|\overline{C}, \mathcal{T}'|\overline{C})| = |\mathcal{F}(\overline{C})|$. Let $\mathcal{F}'(\overline{C})$ be a uMAF for $\mathcal{T}\downarrow_C$ and $\mathcal{T}'\downarrow_C$ and note that, by Lemma 5 $|\mathcal{F}'(\overline{C})| = |\mathcal{F}(\overline{C})|$. Further, let $\mathcal{F}'(C)$ be a uMAF for $\mathcal{T}\downarrow_{\overline{C}}$ and $\mathcal{T}'\downarrow_{\overline{C}}$ and note that $|\mathcal{F}'(C)| \leq |\mathcal{F}(C)| + 1$. Then,

$$\begin{aligned} d_{\mathrm{TBR}}(\mathcal{T},\mathcal{T}') = |\mathcal{F}| - 1 &= |\mathcal{F}(C)| + |\mathcal{F}(\overline{C})| - 1 \\ &\geq |\mathcal{F}'(C)| + |\mathcal{F}'(\overline{C})| - 2 \\ &= d_{\mathrm{TBR}}(\mathcal{T}\downarrow_C, \mathcal{T}'\downarrow_C) + d_{\mathrm{TBR}}(\mathcal{T}\downarrow_{\overline{C}}, \mathcal{T}'\downarrow_{\overline{C}}) \end{aligned}$$

Case 2: There is a tree H in $\mathcal{F}$ containing a leaf labelled $x \in C$ and a leaf labelled $y \in \overline{C}$. This is completely analogous to Case 2 above. □

It is worth mentioning that, in the two cases of Lemma 6, the TBR distances differ by exactly one, that is, $d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) \leq d_{\mathrm{TBR}}(\mathcal{T}{\downarrow_C}, \mathcal{T}'{\downarrow_C}) \leq d_{\mathrm{TBR}}(\mathcal{T}|C, \mathcal{T}'|C) + 1$, Lemma 6 implies that, if there is a uMAF for $\mathcal{T}{\downarrow_C}$ and $\mathcal{T}'{\downarrow_C}$ that isolates the leaf labelled a_C and a uMAF for $\mathcal{T}{\downarrow_{\overline{C}}}$ and $\mathcal{T}'{\downarrow_{\overline{C}}}$ that isolates the leaf labelled $a_{\overline{C}}$, then, when gluing the forests of the subtrees back together to form a uMAF $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$, then we have a tree that does not contain any labelled leaf. Thus, an optimal uMAF has size $|\mathcal{F}| - 1$. This means that, to minimize the size of a forest for $\mathcal{T}$ and $\mathcal{T}'$, we need to favor the forests isolating the dummy taxa. Then, we have the following:

Corollary 5 *Let $\mathcal{T}$ and $\mathcal{T}'$ be unrooted binary phylogenetic X-trees. Let $(\mathcal{T}_1, \mathcal{T}'_1)$, $\ldots, (\mathcal{T}_t, \mathcal{T}'_t)$ be a cluster sequence of $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{F}$ be a maximum-agreement forest of $\mathcal{T}$ and $\mathcal{T}'$. For $i \in \{1, \ldots, t\}$, let $\mathcal{F}_i$ be a maximum-agreement forest for $\mathcal{T}_i$ and $\mathcal{T}'_i$ such that $r := |\{C : \{a_C\}, \{a_{\overline{C}}\} \in \biguplus_i \mathcal{F}_i\}|$ is maximal. Then, $d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = (\sum_i |\mathcal{F}_i|) - t - r$.*

Corollary 5 is a drop-in replacement for Theorem 2.2 in [16] and lets us use the entire cluster-sequence-based machinery of [16] for unrooted phylogenetic trees. Thus, a slight modification of the Minimum-Weight Forest algorithm of [16] (solving the TBR Distance instead of the rSPR Distance and using the unrooted cluster reduction instead of the rooted one) leads right to the following theorem:

Theorem 7 *Let $\mathcal{T}$ and $\mathcal{T}'$ be two unrooted binary phylogenetic X-trees and let $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t)$ be a full cluster reduction of $\mathcal{T}$ and $\mathcal{T}'$. Then, the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$ equals*

$$\max_{i \in \{1,\ldots,t\}} d_{\mathrm{TBR}}(\mathcal{T}_i, \mathcal{T}'_i).$$

Proof First, from Lemma 6, we have that

$$\max_{i \in \{1,\ldots,t\}} d_{\mathrm{TBR}}(\mathcal{T}_i, \mathcal{T}'_i) = \max_{i \in \{1,\ldots,t\}} h(\mathcal{T}_i, \mathcal{T}'_i).$$

The fact that $\max_{i \in \{1,\ldots,t\}} h(\mathcal{T}_i, \mathcal{T}'_i)$ equals the hybridization level of $\mathcal{T}$ and $\mathcal{T}'$ can be proven similarly to Lemma 2, and we do not repeat the proof here. □

Thanks to Theorem 7, Theorem 3 and Corollary 3 can be proven similarly to Theorem 2 and Corollary 2, since TBR Distance can be solved in $O(3^k \cdot n)$, where k is the TBR distance of $\mathcal{T}$ and $\mathcal{T}'$ [8].

6 Conclusion

In this paper, we have shown better bounds for the running time of algorithms computing the hybridization number and the rSPR/TBR distance between two

phylogenetic trees using cluster reductions. We have thus given an explanation for the curious divergence between theoretical results and observed running time of algorithms using cluster reductions.

A deeper biological question that warrants further research is: why does real biological data partition so effectively under the cluster reduction? In other words, why are observed networks of low hybridization level?

Acknowledgment

The third author gratefully acknowledges the scholarship supplied to her from the Republic of Turkey, Ministry of National Education.

Part of the work has been conceived at the 7th workshop on Graph Classes, Optimization, and Width Parameters.

References

1. Allen BL, Steel M (2001) Subtree transfer operations and their induced metrics on evolutionary trees. Annals of combinatorics 5(1):1–15
2. Baroni M, Grünewald S, Moulton V, Semple C (2005) Bounding the number of hybridisation events for a consistent evolutionary history. Journal of mathematical biology 51(2):171–182
3. Baroni M, Semple C, Steel M (2006) Hybrids in real time. Systematic Biology 55(1):46–56
4. Bordewich M, Semple C (2005) On the computational complexity of the rooted subtree prune and regraft distance. Annals of combinatorics 8(4):409–423
5. Bordewich M, Semple C (2007) Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 4(3):458–466
6. Bordewich M, Semple C (2007) Computing the minimum number of hybridization events for a consistent evolutionary history. Discrete Applied Mathematics 155(8):914–928
7. Bordewich M, Linz S, John KS, Semple C (2007) A reduction algorithm for computing the hybridization number of two trees. Evolutionary bioinformatics online 3:86
8. Chen J, Fan JH, Sze SH (2013) Parameterized and approximation algorithms for the MAF problem in multifurcating trees. In: Graph-Theoretic Concepts in Computer Science, Springer, pp 152–164
9. Chen ZZ, Fan Y, Wang L (2013) Faster exact computation of rSPR distance. Journal of Combinatorial Optimization 29(3):605–635
10. Hallett MT, Lagergren J (2001) Efficient algorithms for lateral gene transfer problems. In: Proceedings of the fifth annual international conference on Computational biology, ACM, pp 149–156
11. Harel D, Tarjan RE (1984) Fast algorithms for finding nearest common ancestors. siam Journal on Computing 13(2):338–355

12. Hein J, Jiang T, Wang L, Zhang K (1996) On the complexity of comparing evolutionary trees. Discrete Applied Mathematics 71(1):153–169
13. van Iersel L, Kelk S, Stougie L, Boes O (In preparation) On unrooted and root-uncertain variants of several well-known phylogenetic network problems
14. Jansson J, Sung WK (2006) Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theoretical Computer Science 363(1):60–68
15. Kelk S, Scornavacca C, Van Iersel L (2012) On the elusiveness of clusters. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9(2):517–534
16. Linz S, Semple C (2011) A cluster reduction for computing the subtree distance between phylogenies. Annals of Combinatorics 15(3):465–484
17. Maddison WP (1997) Gene trees in species trees. Systematic biology 46(3):523–536
18. Nakhleh L, Warnow T, Linder CR (2004) Reconstructing reticulate evolution in species: theory and practice. In: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology, ACM, pp 337–346
19. Semple C, Steel MA (2003) Phylogenetics, vol 24. Oxford University Press
20. Van Iersel L, Kelk S (2011) When two trees go to war. Journal of Theoretical Biology 269(1):245–255
21. Van Iersel L, Kelk S, Mnich M (2009) Uniqueness, intractability and exact algorithms: reflections on level-k phylogenetic networks. Journal of Bioinformatics and Computational Biology 7(04):597–623
22. Whidden C, Beiko RG, Zeh N (2013) Fixed-parameter algorithms for maximum agreement forests. SIAM Journal on Computing 42(4):1431–1466
23. Whidden C, Beiko R, Zeh N (In preparation) Computing the SPR distance of binary rooted trees in $O(2^k n)$ time.