

## A Practical Fixed-Parameter Algorithm for Constructing Tree-Child Networks from Multiple Binary Trees

van Iersel, Leo; Janssen, Remie; Jones, Mark; Murakami, Yukihiro; Zeh, Norbert

**DOI**

[10.1007/s00453-021-00914-8](https://doi.org/10.1007/s00453-021-00914-8)

**Publication date**

2022

**Document Version**

Final published version

**Published in**

Algorithmica

**Citation (APA)**

van Iersel, L., Janssen, R., Jones, M., Murakami, Y., & Zeh, N. (2022). A Practical Fixed-Parameter Algorithm for Constructing Tree-Child Networks from Multiple Binary Trees. *Algorithmica*, 84(4), 917-960. <https://doi.org/10.1007/s00453-021-00914-8>

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

***Green Open Access added to TU Delft Institutional Repository***


***'You share, we take care!' - Taverne project***

**<https://www.openaccess.nl/en/you-share-we-take-care>**

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



# A Practical Fixed-Parameter Algorithm for Constructing Tree-Child Networks from Multiple Binary Trees

Leo van Iersel<sup>1</sup> · Remie Janssen<sup>1</sup> · Mark Jones<sup>1</sup> · Yukihiro Murakami<sup>1</sup> · Norbert Zeh<sup>2</sup> 

Received: 19 July 2019 / Accepted: 25 September 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

We present the first fixed-parameter algorithm for constructing a tree-child phylogenetic network that displays an arbitrary number of binary input trees and has the minimum number of reticulations among all such networks. The algorithm uses the recently introduced framework of cherry picking sequences and runs in  $O((8k)^k \text{poly}(n, t))$  time, where  $n$  is the number of leaves of every tree,  $t$  is the number of trees, and  $k$  is the reticulation number of the constructed network. Moreover, we provide an efficient parallel implementation of the algorithm and show that it can deal with up to 100 input trees on a standard desktop computer, thereby providing a major improvement over previous phylogenetic network construction methods.

---

Leo van Iersel, Remie Janssen, Mark Jones and Yukihiro Murakami were supported by the Netherlands Organization for Scientific Research (NWO), including Vidi grant 639.072.602, and van Iersel also by the 4TU Applied Mathematics Institute. Norbert Zeh was supported by the Natural Sciences and Engineering Research Council of Canada.

---

✉ Norbert Zeh  
nzeh@cs.dal.ca

Leo van Iersel  
L.J.J.vanIersel@tudelft.nl

Remie Janssen  
R.Janssen-2@tudelft.nl

Mark Jones  
M.E.L.Jones@tudelft.nl

Yukihiro Murakami  
Y.Murakami@tudelft.nl

<sup>1</sup> Delft Institute of Applied Mathematics, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

<sup>2</sup> Faculty of Computer Science, Dalhousie University, 6050 University Ave, Halifax, NS B3H 1W5, Canada

**Keywords** Fixed-parameter algorithms · Phylogenetic networks · Hybridization number

## 1 Introduction

Evolutionary histories are usually described by phylogenetic trees or networks. A phylogenetic tree describes how a collection of studied taxa (e.g., species, strains or languages) have evolved over time by divergence events, often also called *speciation* events. A phylogenetic network can additionally describe events where lineages merge, such as hybridization or lateral gene transfer, which are called *reticulation* events. A central goal of computational phylogenetics is to develop methods for reconstructing phylogenetic networks from various types of inputs.

One of the most fundamental problems in this area, HYBRIDIZATION NUMBER, is to find a phylogenetic network with the minimum number of reticulation events among all networks that contain a given collection of phylogenetic trees. The network is said to *display* each of the input trees and is also referred to as a *hybridization network* of the set of input trees. Each of these trees represents the evolution, through speciation events and mutation, of a particular gene. Accordingly, we refer to it as a “gene tree”. Reticulation events such as hybridization or lateral gene transfer can lead to discordance between gene trees. The requirement that each gene tree should be contained in the constructed network ensures that the network provides the required paths along which each gene could be passed from ancestors to descendants in a manner consistent with its gene tree. Following the parsimony principle, a network with the minimum number of reticulations that displays all input trees offers a simplest possible model of the evolution of a set of taxa consistent with the given gene trees. Hence the goal to compute a phylogenetic network with as few reticulations as possible. Since not all discordance between gene trees is due to reticulation events, such a network provides only an estimate of the actual number of reticulation events. Nevertheless, hybridization networks have proven to be a valuable tool in the study of the evolution of different sets of taxa. Computing hybridization networks with the minimum number of reticulations, however, has proven to be a major challenge.

Initial research focused on the special case when the input consists of only two trees, in which case there exists a nice mathematical characterization of the problem in terms of maximum acyclic agreement forests (MAAFs) [3]. This characterization has shown to be extremely useful for the development of fixed-parameter algorithms for phylogenetic network construction problems on two trees [6,9,19], with the currently fastest algorithm for HYBRIDIZATION NUMBER running in  $O(3.18^k n)$  time [19].

When the input consists of more than two trees, the problem becomes significantly harder. Kernelization is still possible [16,18]. However, existing algorithms for solving kernelized instances, TREETISTIC [11], PIRN [21], PIRNs [15] and HYBROSCALE [1,2], are limited to (very) small numbers of input trees and/or (very) small numbers of reticulation events. None of these algorithms is fixed-parameter tractable (FPT) unless combined with kernelization. A bounded-search FPT algorithm with running time  $O(c^k \text{poly}(n))$  for the special case of three input trees was proposed in [17] ( $n$  is the

number of taxa,  $k$  the number of reticulations), but the constant  $c$  is much too big for the algorithm to be useful in practice.

The main bottleneck hindering the development of practical algorithms seemed to be the missing mathematical characterization for the problem on more than two trees, analogous to the MAAF characterization for two trees. Such a characterization, in terms of cherry picking sequences, was developed recently and is very different from the MAAF characterization for two trees. The first characterization in terms of cherry picking sequences was developed for the restricted class of temporal networks [10]. Subsequently, it was generalized to the larger class of tree-child networks [14], in which each non-leaf vertex is required to have at least one non-reticulate child. However, Humphries, Linz, and Semple [10] provide only a theoretical FPT result based on kernelization for temporal networks, and Linz and Semple [14] do not present any algorithmic results. Hence, the fixed-parameter tractability of the tree-child version of HYBRIDIZATION NUMBER remained open, as well as the development of practical FPT algorithms based on the new characterization.

Our contribution is to fill this algorithmic gap. We show that there exists an FPT algorithm for HYBRIDIZATION NUMBER restricted to tree-child networks on an arbitrary collection of binary input trees. Its running time is  $O((8k)^k \cdot \text{poly}(n, t))$ , where  $n$  is the number of taxa,  $t$  is the number of trees, and  $k$  is the number of reticulations in the computed network. We verify experimentally that, combined with two heuristic improvements that both preserve the correctness of the algorithm, it can solve fairly complex instances of tree-child HYBRIDIZATION NUMBER. These two heuristics are cluster reduction [7] and a redundant branch elimination technique introduced in this paper. The implementation used in our experiments is available from [https://github.com/nzeh/tree\\_child\\_code](https://github.com/nzeh/tree_child_code).

The main practical benefit of our algorithm is that it can handle many more input trees than existing methods. Indeed, in experiments on synthetic inputs, the running time grows roughly linearly in the number of trees and taxa. On the other hand, the running time still has a large exponential dependency on the number of reticulation events  $k$ . Nevertheless, as long as  $k$  is small (at most 7–12), our algorithm can solve inputs with up to 100 input trees and 200 taxa. In our experiments on real-world data, we observed that these data sets have substantially more structure than random synthetic data sets, which makes cluster reduction and redundant branch elimination more effective and allowed our algorithm to solve inputs with up to 8 trees and 50 reticulations. As the number of trees increases, however, the inputs become less “clusterable”, which reduces the number of reticulations our algorithm can handle.

We also compared our algorithm directly to HYBROSCALE. For instances consisting of two input trees, HYBROSCALE is much faster because it exploits the MAAF characterization for this case. When the number of input trees is at least three, our algorithm turns out to be much faster than HYBROSCALE. HYBROSCALE was able to handle only very few instances with more than five trees.

We restrict our attention to tree-child networks for two reasons. First, although Linz and Semple [14] also provided a characterization of unrestricted hybridization networks in terms of cherry picking sequences, this characterization is based on adding leaves; since it is not known where to add these leaves, this characterization does not seem to be directly useful for developing FPT algorithms. Furthermore, we observed

in our experiments that the optimal tree-child network for a set of trees often has the same number of reticulations as an optimal unrestricted hybridization network. Hence, the restriction to tree-child networks allows us to deal with larger numbers of input trees without changing the problem substantially.

The remainder of this paper is organized as follows: Sect. 2 formally defines the key concepts including the HYBRIDIZATION NUMBER and TREE-CHILD HYBRIDIZATION problems. Section 3 presents our FPT algorithm for TREE-CHILD HYBRIDIZATION. Section 4 presents our redundant branch elimination heuristic for speeding up the algorithm in practice. This section also shows that redundant branch elimination preserves the correctness of the computed cherry picking sequence. Section 5 presents some details of our implementation of the algorithm and discusses our experimental results. We present some concluding remarks in Sect. 6.

## 2 Preliminaries and Definitions

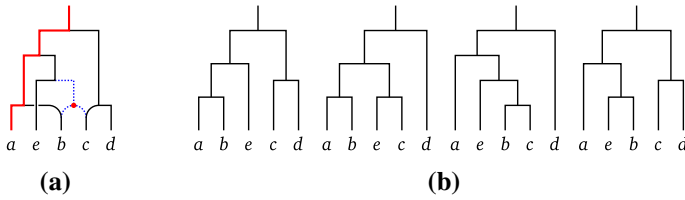
### 2.1 Phylogenetic Trees and Networks

Throughout this paper, we denote by  $X$  a finite non-empty set of taxa. A *phylogenetic network on a subset  $X' \subseteq X$*  is a directed acyclic graph  $N$  whose nodes satisfy the following properties: There is a single node of in-degree 0 and out-degree 2, called the *root*; the nodes of in-degree 1 and out-degree 0 are bijectively labelled with elements from  $X'$  (the *leaves*); all other nodes either have in-degree 1 and out-degree 2 (the *tree nodes*) or have out-degree 1 and in-degree at least 2 (the *reticulations*). This is illustrated in Fig. 1a. A *phylogenetic tree on  $X'$*  is a phylogenetic network on  $X'$  without reticulations; see Fig. 1b. Given a directed edge  $uv$  in a phylogenetic network or tree, we say that  $u$  is a *parent* of  $v$  and  $v$  is a *child* of  $u$ . Unless stated otherwise, edges and paths are always directed in this paper.

For brevity, we usually refer to phylogenetic networks and phylogenetic trees as *networks* and *trees*, respectively. When we feel the need to state the label set  $X'$  of a phylogenetic tree explicitly, especially when we want to emphasize that a set of trees all share the same leaf set, we do refer to this tree as an  $X'$ -*tree*.

Given an edge  $uv$  in a network  $N$ , we call  $uv$  a *reticulation edge* if  $v$  is a reticulation; otherwise,  $uv$  is a *tree edge*. A *tree path* in  $N$  is a path composed of only tree edges. A tree path is shown in red in Fig. 1a. The *reticulation number* of  $N$  is the number of reticulation edges in  $N$  minus the number of reticulations. Alternatively, the reticulation number is the number of edges that need to be deleted from the network to obtain a tree.

Let  $X' \subseteq X$  be a subset of the label set of an  $X$ -tree  $T$ , and let  $T''$  be the smallest subtree of  $T$  that contains all edges on undirected paths between leaves in  $X'$ . The *restriction* of  $T$  to  $X'$  is the tree obtained from  $T''$  by suppressing all vertices with in-degree 1 and out-degree 1. To suppress a node  $v$  with parent  $u$  and child  $w$  is to delete  $v$  and its incident edges and add an edge  $uw$  connecting  $u$  to  $w$ . If  $T$  is an  $X$ -tree and  $T'$  is the restriction of  $T$  to some subset  $X' \subseteq X$ , we write  $T' \subseteq T$ . We also write  $T \setminus T'$  to denote the difference  $X \setminus X'$  of the label sets of the two trees.



**Fig. 1** **a** A phylogenetic network that is not tree-child because both children of the red node are reticulations. Its reticulation number is 2. A tree path from the root to the leaf labelled *a* is shown in red. **b** The four phylogenetic trees displayed by the network in **(a)**. For example, the first tree can be obtained by deleting the dotted edges in **(a)**. The red and black edges constitute an embedding of this tree into the network (Colour figure online)

Let  $N'$  be a subgraph (e.g., a path) of the network  $N$ . Any edge  $uv \in N$  such that  $u \in N'$  and  $v \notin N'$  is called a *pendant edge* of  $N'$ ;  $v$  is a *pendant node* of  $N'$ . When  $N$  is a tree, we say the subtree rooted at  $v$  is a *pendant subtree* of  $N'$ .

**Remark** We note that all nodes of a phylogenetic network as defined in this paper have out-degree at most 2. This is consistent with the definitions used by Linz and Semple [14]. As noted by Linz and Semple, restricting network nodes to have out-degree at most 2 does not result in any loss of generality. In particular, for the problems discussed in this paper, any instance that has a network with out-degree greater than 2 as a solution also has a network with out-degree at most 2 as a solution.

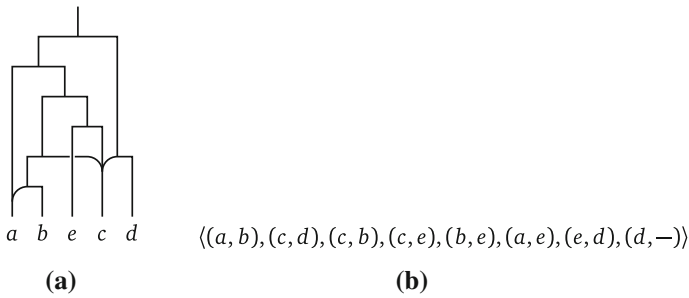
While phylogenetic trees may in general have unbounded out-degree, we require phylogenetic trees to have maximum out-degree 2 in this paper, that is, we restrict our attention to binary trees. It is an open question whether our algorithm can be extended to input trees of unbounded out-degree. We note that Linz and Semple’s result relating tree-child networks to tree-child cherry picking sequences imposes no restriction on the out-degree of phylogenetic trees but does not offer any algorithm to find an optimal tree-child cherry picking sequence or network even for binary trees.

### 2.2 Minimum Tree-Child Hybridization

Given a network  $N$  on a set of taxa  $X$  and a tree  $T$  on a subset  $X' \subseteq X$ , we say that  $N$  displays  $T$  if  $T$  can be obtained from a subgraph of  $N$  by suppressing nodes of out-degree and in-degree 1. Equivalently,  $N$  displays  $T$  if there exists a function  $f$ , called an *embedding* of  $T$  into  $N$ , that maps nodes of  $T$  to nodes of  $N$ , and edges of  $T$  to paths in  $N$ , such that

- Every leaf of  $T$  is mapped to the leaf of  $N$  with the same label;
- For each edge  $uv$  in  $T$ , the path  $f(uv)$  is a path in  $N$  from  $f(u)$  to  $f(v)$ ; and
- For any two distinct edges  $e$  and  $e'$  of  $T$ , the paths  $f(e)$  and  $f(e')$  are edge-disjoint.

For any embedding  $f$  and any node or edge  $x$ , we call  $f(x)$  the *image* of  $x$  (under  $f$ ). This definition extends naturally to arbitrary subgraphs  $T' \subseteq T$  by defining the image  $f(T')$  of  $T'$  to be the union of the images of all nodes and edges in  $T'$ . For a set of trees  $\mathcal{T} = \{T_1, \dots, T_t\}$ , we say that  $N$  displays  $\mathcal{T}$  if  $N$  displays every tree  $T_i \in \mathcal{T}$ .



**Fig. 2** **a** An optimal tree-child network for the four trees in Fig. 1b. Note that this network has reticulation number 3, one more than the non-tree-child hybridization network for these trees in Fig. 1a. The tree-child cherry picking sequences corresponding to this network is shown in **(b)**

For example, the network in Fig. 1a displays all trees in Fig. 1b. An embedding of the first tree into the network is shown.

The **MINIMUM HYBRIDIZATION** problem takes as input a set  $\mathcal{T}$  of phylogenetic trees and an integer  $k$ , and asks for a network displaying  $\mathcal{T}$  and with reticulation number at most  $k$ , if such a network exists. In this paper, we focus on a restricted version of **MINIMUM HYBRIDIZATION**, described below.

A network  $N$  is *tree-child* if every non-leaf node of  $N$  has at least one child that is a tree node or leaf. Note that this is equivalent to requiring that every node in  $N$  has a tree path to a leaf. The network in Fig. 1a is not tree-child because the children of the red node are both reticulations. A tree-child network displaying the trees in Fig. 1b is shown in Fig. 2a.

**MINIMUM TREE-CHILD HYBRIDIZATION**

**Input:** A set  $\mathcal{T} = \{T_1, \dots, T_t\}$  of phylogenetic trees on  $X$  and an integer  $k$ .  
**Output:** A tree-child phylogenetic network  $N$  on  $X$  that displays  $\mathcal{T}$  and has at most  $k$  reticulations, if such a network exists; **NONE** otherwise.

For a set  $\mathcal{T} = \{T_1, T_2, \dots, T_t\}$  of  $X$ -trees, let  $h(\mathcal{T})$  denote the *hybridization number* of  $\mathcal{T}$ , that is, the minimum reticulation number of all networks that display  $\mathcal{T}$ . Similarly, let  $h_{tc}(\mathcal{T})$  denote the *tree-child hybridization number* of  $\mathcal{T}$ , that is, the minimum reticulation number of all tree-child networks that display  $\mathcal{T}$ .

**2.3 Cherry Picking Sequences**

For any tree  $T$  on  $X' \subseteq X$  and any two taxa  $x, y \in X'$ , we say that  $\{x, y\}$  is a *cherry* of  $T$  if the leaves labelled with  $x$  and  $y$  are siblings in  $T$ . Observe that any tree with two or more leaves contains at least one cherry. A pair  $\{x, y\}$  is a cherry of a set of trees  $\mathcal{T}$  if it is a cherry of at least one tree in  $\mathcal{T}$ . It is a *trivial cherry* of  $\mathcal{T}$  if  $\{x, y\}$  is a cherry of every tree in  $\mathcal{T}$  that contains both  $x$  and  $y$ .

Linz and Semple [14] gave a characterization of tree-child hybridization number in terms of *cherry picking sequences*, which we define next. Informally, a cherry picking sequence is a sequence of pairs of leaves, describing a sequence of operations on a set of trees  $\mathcal{T}$ . In particular a "proper" pair of the form  $(x, y)$  denotes the operation



of removing the leaf  $x$  from any tree in  $\mathcal{T}$  that has  $\{x, y\}$  as a cherry, while a pair of the form  $(x, -)$  is added to the end of the sequence if the proper pairs in the sequence reduce at least one tree in  $\mathcal{T}$  to the single leaf  $x$ .

Formally, a *cherry picking sequence* is a sequence

$$S = \langle (x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_{r+1}, -), (x_{r+2}, -), \dots, (x_s, -) \rangle$$

with  $\{x_1, x_2, \dots, x_s, y_1, y_2, \dots, y_r\} \subseteq X$ . We write  $|S|$  to denote the length  $s$  of  $S$ . It may be that  $s = r$ , in which case the last element is  $(x_r, y_r)$ , that is, there are no pairs of the form  $(x_j, -)$ . We call such a sequence a *partial cherry picking sequence*. A sequence is *full* if  $s > r$  and  $\{x_1, \dots, x_s\} = X$ . For any  $1 \leq i \leq j \leq s$ , we denote by  $S_{i,j}$  the subsequence  $\langle (x_i, y_i), \dots, (x_j, y_j) \rangle$  (where  $y_h$  is replaced with  $-$  for  $h > r$ ). Given two sequences  $S = \langle (x_1, y_1), \dots, (x_r, y_r) \rangle$  and  $S' = \langle (x'_1, y'_1), \dots, (x'_{r'}, y'_{r'}) \rangle$ , we denote by  $S \circ S'$  the sequence  $\langle (x_1, y_1), \dots, (x_r, y_r), (x'_1, y'_1), \dots, (x'_{r'}, y'_{r'}) \rangle$ . We say that  $S \circ S'$  is an *extension* of  $S$ , and that  $S$  is a *prefix* of  $S \circ S'$ . If  $S' \neq \langle \rangle$ , then we call  $S$  a *proper prefix* of  $S \circ S'$ .

For a tree  $T$  on  $X' \subseteq X$ , the sequence  $S$  defines a sequence of trees  $\langle T^{(0)}, T^{(1)}, \dots, T^{(r)} \rangle$  as follows:

- $T^{(0)} = T$ ;
- If  $\{x_j, y_j\}$  is a cherry of  $T^{(j-1)}$ , then  $T^{(j)}$  is obtained from  $T^{(j-1)}$  by removing  $x_j$  and suppressing  $y_j$ 's parent. Otherwise,  $T^{(j)} = T^{(j-1)}$ .

For notational convenience, we refer to  $T^{(r)}$  as  $T/S$ , the tree obtained by *applying* the sequence  $S$  to  $T$ . In addition, for a set of trees  $\mathcal{T} = \{T_1, \dots, T_t\}$ , we write  $\mathcal{T}^{(j)}$  to denote the set  $\{T_1^{(j)}, \dots, T_t^{(j)}\}$ , and  $\mathcal{T}/S$  to denote the set  $\{T_1/S, \dots, T_r/S\}$ .

A full cherry picking sequence  $S = \langle (x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_{r+1}, -), (x_{r+2}, -), \dots, (x_s, -) \rangle$  is a cherry picking sequence *for a set of trees*  $\mathcal{T}$  if every tree in  $\mathcal{T}/S$  has a single leaf and that leaf is in  $\{x_{r+1}, \dots, x_s\}$ . The *weight*  $w(S)$  of  $S$  is defined to be  $|S| - |X|$ .

A cherry picking sequence  $S$  is *tree-child* if  $s \leq r + 1$  and  $y_j \neq x_i$  for all  $1 \leq i < j \leq s$ . (Thus, if  $S$  is a tree-child cherry picking sequence for  $\mathcal{T}$ , then  $T/S$  consists of the single leaf  $x_s$  for every tree  $T \in \mathcal{T}$ .) The tree-child cherry picking sequence for the set of trees in Fig. 1b corresponding to the tree-child network in Fig. 2a is shown in Fig. 2b. Given a partial tree-child cherry picking sequence  $S = \langle (x_1, y_1), \dots, (x_r, y_r) \rangle$ , we call the leaves  $\{x_1, \dots, x_r\}$  *forbidden leaves with respect to*  $S$  because every tree-child extension  $S \circ S' = \langle (x_1, y_1), \dots, (x_r, y_r), (x_{r+1}, y_{r+1}), \dots, (x_{r'}, y_{r'}) \rangle$  of  $S$  satisfies the condition  $\{x_1, \dots, x_r\} \cap \{y_{r+1}, \dots, y_{r'}\} = \emptyset$ , that is, the leaves  $\{x_1, \dots, x_r\}$  are forbidden to appear as the second element of any pair  $(x_j, y_j)$  with  $r < j \leq r'$  in  $S'$ . We say that  $S \circ S'$  is an *optimal tree-child extension* of  $S$  if  $S \circ S'$  is a tree-child cherry picking sequence for  $\mathcal{T}$  and every extension  $S \circ S''$  of  $S$  that is a tree-child cherry picking sequence for  $\mathcal{T}$  satisfies  $w(S \circ S'') \geq w(S \circ S')$ . For the purpose of the algorithmic construction of sequences, we adopt the convention that  $S \circ \text{NONE} = \text{NONE}$  for any sequence  $S$  and that  $w(\text{NONE}) = \infty$ .

Let  $s_{tc}(\mathcal{T})$  be the minimum weight of all tree-child cherry picking sequences for  $\mathcal{T}$ . Linz and Semple showed that the problem of finding the tree-child hybridization

number of a set  $\mathcal{T}$  of  $X$ -trees is equivalent to finding the minimum weight of a tree-child cherry picking sequence for  $\mathcal{T}$ :

**Theorem 1** (Linz and Semple [14]) *Let  $X$  be a set of taxa, and  $\mathcal{T} = \{T_1, T_2, \dots, T_t\}$  a collection of phylogenetic  $X$ -trees. Then*

$$s_{\text{tc}}(\mathcal{T}) = h_{\text{tc}}(\mathcal{T}).$$

### 3 Finding an Optimal Tree-Child Sequence

In this section, we show that MINIMUM TREE-CHILD HYBRIDIZATION is fixed-parameter tractable with respect to  $k$ . Our proof is based on Linz and Semple's characterization of tree-child hybridization number in terms of tree-child cherry picking sequences (see Theorem 1). As such, our main technical contribution is to give a fixed-parameter algorithm, TCS, for the problem of finding a tree-child cherry picking sequence of weight at most  $k$ , if such a sequence exists. By the following proposition, a corresponding tree-child network can then be found in polynomial time.

**Proposition 2** (Linz and Semple [14]) *There exists a linear-time algorithm that, given a set  $\mathcal{T}$  of  $X$ -trees and a tree-child cherry picking sequence  $S$  for  $\mathcal{T}$ , computes a tree-child network  $N$  displaying  $\mathcal{T}$  with  $h(N) \leq w(S)$ .*

For completeness, the pseudocode of this algorithm, TREECHILDNETWORKFROMSEQUENCE, is given in the appendix. (Linz and Semple do not state a running time for this algorithm, but it is easy to observe that their algorithm takes linear time in the length of the given cherry picking sequence.)

Our algorithm for computing a tree-child cherry picking sequence of length at most  $k$  has the following structure: Starting with the set of trees  $\mathcal{T}$  and the empty sequence  $S = \langle \rangle$ , the algorithm repeats the following as long as  $\mathcal{T}/S$  still has a cherry. If  $\mathcal{T}/S$  has a trivial cherry  $\{x, y\}$  such that  $y$  is not forbidden with respect to  $S$ , it adds  $(x, y)$  to the end of  $S$ . If  $\mathcal{T}/S$  has no trivial cherry, we show that  $\mathcal{T}/S$  has at most  $4k$  unique cherries or  $h_{\text{tc}}(\mathcal{T}) > k$ . The algorithm makes one recursive call for each pair  $(x, y)$  such that  $\{x, y\}$  is a cherry of  $\mathcal{T}/S$ , starting each recursive call by adding  $(x, y)$  to the end of  $S$ . (Note that every cherry  $\{x, y\}$  of  $\mathcal{T}/S$  gives rise to up to two recursive calls, one for the pair  $(x, y)$  and one for the pair  $(y, x)$ .) As this kind of branching step cannot occur more than  $k$  times in a sequence of weight at most  $k$ , this gives a search tree for our algorithm of depth  $k$  and branching number at most  $8k$ .

In the remainder of this section, we prove the correctness of procedure TCS and analyze its running time. This is summarized in the following theorem (we denote by  $\lg$  the logarithmic function with base 2).

**Theorem 3** *Given a collection  $\mathcal{T}$  of  $t$   $X$ -trees with  $|X| = n$ , it takes  $O((8k)^k nt \lg t + nt \lg nt)$  time to decide whether  $\mathcal{T}$  has tree-child hybridization number at most  $k$  and, if so, compute a corresponding tree-child cherry picking sequence.*

Combined with Proposition 2, this proves the following corollary.

**Procedure TCS( $\mathcal{T}, S, k$ )**

---

```

Input: A collection of phylogenetic trees  $\mathcal{T}$ , a partial tree-child cherry picking sequence  $S$ , and an integer  $k$ 
Output: An optimal solution of  $(\mathcal{T}, S)$  if  $(\mathcal{T}, S)$  has a solution of weight at most  $k$ ; NONE otherwise
1 while there exists a trivial cherry  $\{x, y\}$  of  $\mathcal{T}/S$  with  $y$  not forbidden with respect to  $S$  do
2   |  $S \leftarrow S \circ \langle(x, y)\rangle$ ;
3  $\mathcal{T}' \leftarrow \mathcal{T}/S$ ;
4 if  $\mathcal{T}'$  contains a cherry  $\{x, y\}$  with  $x, y$  both forbidden with respect to  $S$  then
5   | return NONE;
6 else
7   |  $n' \leftarrow |\{x \in X : x \text{ is a leaf of a tree in } \mathcal{T}'\}|$ ;
8   |  $k' \leftarrow |S| - |X| + n'$ ;
9   |  $C \leftarrow \{(x, y) \mid \{x, y\} \text{ is a cherry of some tree in } \mathcal{T}'\}$ ;
10  | if  $|C| = 0$  then
11  |   | return  $S \circ \langle(x, -)\rangle$ , where  $x$  is the last remaining leaf in all trees;
12  | else if  $|C| > 8k$  or  $k' \geq k$  then
13  |   | return NONE;
14  | else
15  |   |  $S_{opt} \leftarrow \text{NONE}$ ;
16  |   | foreach  $(x, y) \in C$  with  $y$  not forbidden with respect to  $S$  do
17  |     |  $S_{temp} \leftarrow \text{TCS}(\mathcal{T}, S \circ \langle(x, y)\rangle, k)$ ;
18  |     | if  $w(S_{temp}) < w(S_{opt})$  then
19  |       |   |  $S_{opt} \leftarrow S_{temp}$ ;
20  |   | return  $S_{opt}$ ;

```

---

**Corollary 4** Given a collection  $\mathcal{T}$  of  $t$   $X$ -trees with  $|X| = n$ , it takes  $O((8k)^k nt \lg t + nt \lg nt)$  time to decide whether  $\mathcal{T}$  has tree-child hybridization number at most  $k$  and, if so, compute a corresponding tree-child hybridization network that displays  $\mathcal{T}$ .

It is easy to see that procedure TCS returns a sequence  $S$  only if it is a valid tree-child cherry picking sequence for  $\mathcal{T}$ . Thus, it suffices to show that if a partial tree-child cherry picking sequence  $S$  has an extension  $S \circ S'$  of weight at most  $k$  that is a cherry picking sequence for  $\mathcal{T}$ , then the invocation  $\text{TCS}(\mathcal{T}, S, k)$  finds a shortest such extension. In the remainder of this section, we call an extension  $S \circ S'$  of a partial tree-child cherry picking sequence  $S$  a *solution* of  $(\mathcal{T}, S)$  if  $S \circ S'$  is a cherry picking sequence for  $\mathcal{T}$ ;  $S \circ S'$  is an *optimal* solution of  $(\mathcal{T}, S)$  if there is no solution of  $(\mathcal{T}, S)$  that is shorter than  $S \circ S'$ .

We split the proof of Theorem 3 into two parts: First, we show that we deal with trivial cherries correctly: if  $(\mathcal{T}, S)$  has a solution of weight at most  $k$  and  $\mathcal{T}' = \mathcal{T}/S$  has a trivial cherry  $\{x, y\}$  such that  $y$  is not forbidden with respect to  $S$ , then  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  has a solution of weight at most  $k$  and any optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  is also an optimal solution of  $(\mathcal{T}, S)$ . Thus, adding trivial cherries to  $S$  as TCS does in lines 1–2 is safe. Section 3.1 presents this first part of our proof. Second, we show that if  $\mathcal{T}'$  has no trivial cherries, then either the trees in  $\mathcal{T}'$  have at most  $4k$  unique cherries or  $(\mathcal{T}, S)$  has no solution of weight at most  $k$ . Thus, aborting the search if  $|C| > 8k$  (since  $C$  contains two pairs for each cherry of  $\mathcal{T}'$ ), as we do in line 13, is correct. The proof of this bound on the number of unique cherries is divided into two parts. In Sect. 3.2, we show that this bound holds if  $S = \langle \rangle$ , that is, if all trees in  $\mathcal{T}'$  are

$X$ -trees. In Sect. 3.3, we extend this result to arbitrary partial tree-child cherry picking sequences  $S$ . Sect. 3.4 then completes the proof of Theorem 3.

### 3.1 Pruning Trivial Cherries

Our algorithm begins by repeatedly pruning trivial cherries in lines (i)–(ii); that is, as long as there exists a trivial cherry  $\{x, y\}$  in  $\mathcal{T}/S$  with  $y$  not forbidden with respect to  $S$ , the algorithm extends  $S$  by adding the pair  $(x, y)$  to  $S$ . In this section, we show that this is safe: if  $(\mathcal{T}, S)$  has a solution of weight at most  $k$ , then so does  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$ , and any optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  is an optimal solution of  $(\mathcal{T}, S)$ . We begin with some simple observations.

**Proposition 5** *Let  $S = \langle(x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_{r+1}, -)\rangle$  be a tree-child cherry picking sequence for a set of  $X$ -trees  $\mathcal{T}$ . Then the following properties hold for all  $j \in [r]$ :*<sup>1</sup>

- (i) *If  $y \in X$  is not forbidden with respect to  $S_{1,j}$ , then  $y$  is a leaf in every tree in  $\mathcal{T}^{(j)}$ .*
- (ii) *If  $\{x, y\}$  is a cherry of  $\mathcal{T}^{(j)}$ , then either  $(x, y)$  or  $(y, x)$  is a pair in  $S_{j+1,r}$ .*
- (iii) *If  $\{x_j, y_j\}$  is a trivial cherry of  $\mathcal{T}^{(j-1)}$ , then  $x_j$  is not in any tree in  $\mathcal{T}^{(j)}$ .*

**Proof** Property (i) holds because  $y$  is not forbidden with respect to  $S_{1,j}$  and, thus,  $y \neq x_i$  for all  $i \in [j]$ . Property (ii) follows because  $S_{j+1,r}$  must delete at least one of  $x, y$  from the tree containing  $\{x, y\}$  as a cherry and only the pair  $(x, y)$  or  $(y, x)$  achieves this. To see why Property (iii) holds, observe that  $y_j$  is not forbidden with respect to  $S_{1,j-1}$ . Thus, by Property (i), every tree in  $\mathcal{T}^{(j-1)}$  contains  $y_j$  as a leaf. In particular, every tree in  $\mathcal{T}^{(j-1)}$  containing  $x_j$  also contains  $y_j$ . Thus, by the definition of a trivial cherry, every tree in  $\mathcal{T}^{(j-1)}$  containing  $x_j$  contains the cherry  $\{x_j, y_j\}$ . Thus, applying the pair  $(x_j, y_j)$  to  $\mathcal{T}^{(j-1)}$  deletes  $x_j$  from any tree containing  $x_j$  and no tree in  $\mathcal{T}^{(j)}$  contains  $x_j$ . □

**Lemma 6** *Let  $S = \langle(x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_{r+1}, -)\rangle$  be a tree-child cherry picking sequence for a set of  $X$ -trees  $\mathcal{T}$  and suppose that  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}^{(j)}$  and  $y$  is not forbidden with respect to  $S_{1,j}$ . Then there exists a tree-child cherry picking sequence  $S'$  for  $\mathcal{T}$  such that  $|S'| = |S|$ ,  $S'_{1,j} = S_{1,j}$ , and  $(x, y)$  is a pair in  $S'_{j+1,r}$ .*

**Proof** We start with the following trivial observation: Let  $\mathcal{T}$  be a set of trees and let  $S$  be a tree-child cherry picking sequence for  $\mathcal{T}$ . For an arbitrary permutation  $\pi$  of  $X$  and any  $X$ -tree  $T$ , let  $T_{|\pi}$  be the tree obtained from  $T$  by changing the label of each leaf from its label  $z$  in  $T$  to the label  $\pi(z)$  in  $T_{|\pi}$ . Let  $\mathcal{T}_{|\pi} = \{T_{|\pi} \mid T \in \mathcal{T}\}$ . Similarly, let  $S_{|\pi}$  be the sequence obtained from  $S$  by replacing every occurrence of an element  $z \in X$  in  $S$  with  $\pi(z)$ . Then  $S_{|\pi}$  is a tree-child cherry picking sequence for  $\mathcal{T}_{|\pi}$ . Here, we consider the permutation  $\pi$  such that  $\pi(x) = y$ ,  $\pi(y) = x$ , and  $\pi(z) = z$  for all  $z \in X \setminus \{x, y\}$ , where  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}^{(j)}$ .

By Proposition 5(ii), either  $(x, y)$  or  $(y, x)$  is a pair in  $S_{j+1,r}$ . In the former case, the sequence  $S' = S$  satisfies the lemma. In the latter case, neither  $x$  nor  $y$  is forbidden

<sup>1</sup> We use  $[m]$  to denote the set of integers  $\{1, \dots, m\}$  and  $[m]_0$  to denote the set of integers  $\{0, \dots, m\}$ .

with respect to  $S_{1,j}$ . It follows from Proposition 5(i) and the fact that  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}^{(j)}$  that every tree in  $\mathcal{T}^{(j)}$  has  $\{x, y\}$  as a cherry. In particular, neither  $x$  nor  $y$  is part of a pair in  $S_{1,j}$ . Thus, since  $S$  is a tree-child cherry picking sequence, the sequence  $S' = S_{1,j} \circ (S_{j+1,r+1})|_{\pi}$  is a tree-child cherry picking sequence such that  $S'_{1,j} = S_{1,j}$  and  $(x, y) \in S'_{j+1,r}$ . To see that  $S'$  is a tree-child cherry picking sequence for  $\mathcal{T}$ , observe that  $S_{j+1,r+1}$  is a tree-child cherry picking sequence for  $\mathcal{T}^{(j)}$ . Thus, as just observed,  $(S_{j+1,r+1})|_{\pi}$  is a tree-child cherry picking sequence for  $\mathcal{T}^{(j)}$ . However, since  $\{x, y\}$  is a cherry of every tree in  $\mathcal{T}^{(j)}$ , we have  $\mathcal{T}^{(j)}|_{\pi} = \mathcal{T}^{(j)}$ , that is,  $(S_{j+1,r+1})|_{\pi}$  is a tree-child cherry picking sequence for  $\mathcal{T}^{(j)}$  and  $S' = S_{1,j} \circ (S_{j+1,r+1})|_{\pi}$  is a tree-child cherry picking sequence for  $\mathcal{T}$ .  $\square$

**Lemma 7** *Let  $T$  be an  $X$ -tree, let  $T' \subseteq T$ , and let  $S = \langle (x_1, y_1), (x_2, y_2), \dots, (x_r, y_r) \rangle$  be a partial tree-child cherry picking sequence such that  $(T \setminus T') \cap \{y_1, y_2, \dots, y_r\} = \emptyset$ . Then  $T'/S \subseteq T/S$ .*

**Proof** We prove the claim by induction on  $|S|$ . If  $|S| = 0$ , then  $T'/S = T' \subseteq T = T/S$ , so the claim holds in this case. If  $|S| > 0$ , then let  $R' = T'/S_{1,1}$  and  $R = T/S_{1,1}$ . Note that  $R \supseteq T - x_1$ . If  $x_1 \notin T'$ , then  $R' = T' \subseteq T - x_1 \subseteq R$ . If  $y_1 \notin T'$ , then  $y_1 \notin T$  because  $y_1 \notin T \setminus T'$ . Thus,  $R' = T' \subseteq T = R$ .

So assume that  $x_1, y_1 \in T'$ . If  $\{x_1, y_1\}$  is a cherry of  $T'$ , then  $R' = T' - x_1 \subseteq T - x_1 \subseteq R$ . If  $\{x_1, y_1\}$  is not a cherry of  $T'$ , then  $x_1, y_1 \in T'$  implies that the path from  $x_1$  to  $y_1$  in  $T'$  has at least one pendant subtree. Since  $T' \subseteq T$ , this implies that the path from  $x_1$  to  $y_1$  in  $T$  also has at least one pendant subtree, that is,  $\{x_1, y_1\}$  is not a cherry of  $T$  either. Therefore,  $R' = T' \subseteq T = R$ .

We have shown that in all possible cases,  $R' \subseteq R$ . Now observe that  $R \setminus R' \subseteq (T \setminus T') \cup \{x_1\}$ . Since  $S$  is a partial tree-child cherry picking sequence,  $S_{2,r}$  is a partial tree-child cherry picking sequence and  $x_1 \notin \{y_2, y_3, \dots, y_r\}$ . Since  $(T \setminus T') \cap \{y_2, y_3, \dots, y_r\} = \emptyset$ , this implies that  $(R \setminus R') \cap \{y_2, y_3, \dots, y_r\} = \emptyset$ . Thus, by the induction hypothesis,  $T'/S = R'/S_{2,r} \subseteq R/S_{2,r} = T/S$ .  $\square$

We are now ready to prove a stronger version of Lemma 6, which establishes that pruning trivial cherries is safe.

**Proposition 8** *Let  $S = \langle (x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_{r+1}, -) \rangle$  be a tree-child cherry picking sequence for a set of  $X$ -trees  $\mathcal{T}$  and suppose that  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}^{(j)}$  and  $y$  is not forbidden with respect to  $S_{1,j}$ . Then there exists a tree-child cherry picking sequence  $S' = \langle (x'_1, y'_1), (x'_2, y'_2), \dots, (x'_{r'}, y'_{r'}), (x'_{r'+1}, -) \rangle$  for  $\mathcal{T}$  such that  $|S'| \leq |S|$ ,  $S'_{1,j} = S_{1,j}$ , and  $(x'_{j+1}, y'_{j+1}) = (x, y)$ .*

**Proof** By Lemma 6, there exists a tree-child cherry picking sequence  $S' = \langle (x'_1, y'_1), (x'_2, y'_2), \dots, (x'_{r'}, y'_{r'}), (x'_{r'+1}, -) \rangle$  for  $\mathcal{T}$  such that  $r' \leq r$ ,  $S'_{1,j} = S_{1,j}$  and  $(x, y) \in S'_{j+1,r'}$ . We choose  $S'$  from the set of all such cherry picking sequences so that the index  $j' > j$  with  $(x_{j'}, y_{j'}) = (x, y)$  is minimized. If  $j' = j + 1$ , the lemma holds. If  $j' > j + 1$ , we obtain a contradiction to the choice of  $S'$  by transforming  $S'$  into another tree-child cherry picking sequence  $S'' = \langle (x''_1, y''_1), \dots, (x''_{r''}, y''_{r''}), (x''_{r''+1}, -) \rangle$  for  $\mathcal{T}$  such that  $|S''| \leq |S'| \leq |S|$ ,  $S''_{1,j} = S'_{1,j} = S_{1,j}$ , and  $(x''_{j'-1}, y''_{j'-1}) = (x, y)$ .

So assume that  $j' > j + 1$  and let  $(x'_{j'-1}, y'_{j'-1}) = (v, w)$ . We distinguish two cases:

**$w = x$ :** In this case, we set  $r'' = r' - 1$ ,  $(x''_h, y''_h) = (x'_h, y'_h)$  for all  $1 \leq h \leq j' - 2$ , and  $(x''_h, y''_h) = (x'_{h+1}, y'_{h+1})$  for all  $j' - 1 \leq h \leq r'' + 1$ ; that is, we obtain  $S''$  by deleting the pair  $(x'_{j'-1}, y'_{j'-1})$  from  $S'$ . Thus,  $S'' \subset S'$ ,  $|S''| < |S'|$ , and  $(x''_{j'-1}, y''_{j'-1}) = (x'_{j'}, y'_{j'}) = (x, y)$ . Since  $S'$  is a tree-child cherry picking sequence, this implies that  $S''$  also is a tree-child cherry picking sequence. To see that  $S''$  is a tree-child cherry picking sequence for  $\mathcal{T}$ , it suffices to prove that  $T/S''_{1,j'-2} = T/S'_{1,j-1}$  and, thus,  $T/S'' = (T/S''_{1,j'-2})/S'_{j',r'} = (T/S'_{1,j'-2})/S'_{j',r'} = T/S'$  for every tree  $T \in \mathcal{T}$ .

To prove this, observe that  $v \neq y$  and  $y \in T/S'_{1,h}$  for all  $T \in \mathcal{T}$  and all  $1 \leq h < j'$  because  $y_{j'} = y$ , that is,  $y$  is not forbidden with respect to  $S'_{1,j'-1}$ . Thus, since  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}/S_{1,j} = \mathcal{T}/S'_{1,j}$  and  $j < j'$ ,  $\{x, y\}$  is a cherry of every tree  $T/S'_{1,j}$  in  $\mathcal{T}/S'_{1,j}$  that contains  $x$ . Since  $y$  is also a leaf of every tree  $T/S'_{1,j'-2}$  in  $\mathcal{T}/S'_{1,j'-2}$  (again, because  $y$  is not forbidden with respect to  $S'_{1,j'-1}$ ), this implies that  $\{x, y\}$  is also a cherry of every tree in  $\mathcal{T}/S'_{1,j'-2}$  that contains  $x$ . In particular, since  $v \neq y$ ,  $\{v, w\} = \{v, x\}$  is not a cherry of any tree  $T/S'_{1,j'-2}$  in  $\mathcal{T}/S'_{1,j'-2}$  and  $T/S'_{1,j'-2} = T/S'_{1,j'-1}$  for all  $T \in \mathcal{T}$ .

**$w \neq x$ :** In this case, we set  $(x''_{j'-1}, y''_{j'-1}) = (x'_{j'}, y'_{j'})$ ,  $(x''_h, y''_h) = (x'_{j'-1}, y'_{j'-1})$ , and  $(x''_h, y''_h) = (x'_h, y'_h)$  for all  $h \notin \{j' - 1, j'\}$ , that is, we obtain  $S''$  by swapping  $(x'_{j'-1}, y'_{j'-1}) = (v, w)$  and  $(x'_{j'}, y'_{j'}) = (x, y)$  in  $S$ . This clearly implies that  $|S''| = |S'|$  and  $(x''_{j'-1}, y''_{j'-1}) = (x'_{j'}, y'_{j'}) = (x, y)$ . To see that  $S''$  is a tree-child cherry picking sequence, observe that every pair  $(x''_h, y''_h)$  in  $S''$  with  $h \neq j'$  is preceded by a subset of the pairs that precede it in  $S'$ . Thus, since  $S'$  is a tree-child cherry picking sequence,  $y''_h$  is not forbidden with respect to  $S''_{1,h-1}$ . For the pair  $(x''_{j'}, y''_{j'})$ ,  $y''_{j'}$  is not forbidden with respect to  $S''_{1,j'-2}$  because  $S''_{1,j'-2} = S'_{1,j'-2}$  and  $(x'_{j'}, y'_{j'}) = (x'_{j'-1}, y'_{j'-1})$ . This implies that  $y''_{j'}$  is not forbidden with respect to  $S''_{1,j'-1}$  because  $y''_{j'} = y'_{j'-1} = w \neq x = x'_{j'} = x'_{j'-1}$ .

It remains to show that  $T/S'' = T/S'$  for all  $T \in \mathcal{T}$ . To this end, it suffices to show that  $T/S'' \subseteq T/S'$  because  $T/S'$  has only one leaf,  $x_{r+1}$ , and  $T/S'' \neq \emptyset$ , that is,  $T/S'' \subseteq T/S'$  implies that  $T/S'' = T/S'$ .

To see that  $T/S'' \subseteq T/S'$ , let  $T' = T/S'_{1,j'-2}$ . Then  $T'/\langle(x, y)\rangle \subseteq T'$ ,  $T' \setminus (T'/\langle(x, y)\rangle) \subseteq \{x\}$ , and  $x \notin \{w, y\}$ . By Lemma 7, this implies that  $T'/\langle(x, y), (v, w), (x, y)\rangle \subseteq T'/\langle(v, w), (x, y)\rangle$ . However, as argued above,  $\{x, y\}$  is a cherry of  $T'$ , so  $x \notin T'/\langle(x, y)\rangle$  and, thus,  $x \notin T'/\langle(x, y), (v, w)\rangle$ . This implies that  $T'/\langle(x, y), (v, w), (x, y)\rangle = T'/\langle(x, y), (v, w)\rangle$  and, therefore,  $T'/\langle(x, y), (v, w)\rangle \subseteq T'/\langle(v, w), (x, y)\rangle$ . Since  $T' = T/S'_{1,j'-2} = T/S''_{1,j'-2}$ ,  $S'_{1,j'} = S'_{1,j'-2} \circ \langle(v, w), (x, y)\rangle$ , and  $S''_{1,j'} = S''_{1,j'-2} \circ \langle(x, y), (v, w)\rangle$ , this shows that  $T/S''_{1,j'} \subseteq T/S'_{1,j'}$ .

Using Lemma 7 again, this shows that  $T/S'' = (T/S''_{1,j'})/S''_{j'+1,r'} = (T/S''_{1,j'})/S'_{j'+1,r'} \subseteq (T/S'_{1,j'})/S'_{j'+1,r'} = T/S'$ .  $\square$

### 3.2 Bounding the Number of Cherries in Irreducible X-Trees

Once the algorithm has eliminated all trivial cherries from a set of input trees, each of the remaining (non-trivial) cherries of  $\mathcal{T}/S$  is a candidate for being the next pair to be added to  $S$ . Our algorithm makes one recursive call for each possible choice of this next pair (lines 15–20). In order to limit the number of recursive calls it makes, the algorithm aborts and reports failure if there are more than  $8k$  choices to branch on. To prove that this does not prevent us from finding a tree-child cherry picking sequence of weight at most  $k$ , if such a sequence exists, we need to prove the following claim:

**Proposition 9** *If  $(\mathcal{T}, S)$  has a solution of weight at most  $k$  and  $\mathcal{T}/S$  has no trivial cherries, then the number of unique cherries in  $\mathcal{T}/S$  is at most  $4k$ .*

Note that this claim refers to the weight  $k$  of the whole sequence  $S \circ S'$ , not the weight of  $S'$ . This is because the proof uses the structure of  $S$  and  $S'$  to bound the number of unique cherries in  $\mathcal{T}/S$ .

Our proof has two parts: In this subsection, we consider the case when  $S = \langle \rangle$ , that is, when we have a set of  $X$ -trees  $\mathcal{T}$  with tree-child hybridization number at most  $k$  and no trivial cherries. In the next subsection, we prove the claim for  $S \neq \langle \rangle$ , via a reduction to the case when  $S = \langle \rangle$ .

**Lemma 10** *If  $\mathcal{T}$  is a set of  $X$ -trees without trivial cherries and with tree-child hybridization number  $k$ , then the total number of cherries of the trees in  $\mathcal{T}$  is at most  $4k$ .*

**Proof** Let  $N$  be a tree-child network with  $k$  reticulations that displays  $\mathcal{T}$  and, for each tree  $T_i \in \mathcal{T}$ , let  $f_i$  be an embedding of  $T_i$  into  $N$ . Our strategy is to charge each cherry  $\{x, y\}$  of  $\mathcal{T}$  to some reticulation edge in a manner that charges every reticulation edge for at most two cherries. Since  $N$  has at most  $k$  reticulations and, therefore, at most  $2k$  reticulation edges, this proves the lemma.

We start by proving a number of auxiliary claims about how the images of cherries interact with reticulation edges and with each other. The first three claims consider a fixed cherry  $\{x, y\}$  of some tree  $T_i \in \mathcal{T}$  and a fixed tree  $T_j$  that does not have  $\{x, y\}$  as a cherry. Since  $\{x, y\}$  is non-trivial, such a tree  $T_j$  exists. Let  $p$  be the common parent of  $x$  and  $y$  in  $T_i$  and let  $e_x = px$  and  $e_y = py$  be the parent edges of  $x$  and  $y$  in  $T_i$ , respectively. Since  $T_j$  is an  $X$ -tree, we have  $x, y \in T_j$ . Let  $u$  be the lowest common ancestor (LCA) of  $x$  and  $y$  in  $T_j$ , that is, the node in  $T_j$  farthest from the root that is an ancestor of both  $x$  and  $y$ , and let  $P_x$  and  $P_y$  be the paths from  $u$  to  $x$  and from  $u$  to  $y$  in  $T_j$ , respectively. Since  $\{x, y\}$  is not a cherry of  $T_j$ , the undirected path  $P_x \cup P_y$  has at least one pendant edge.

**Claim 1** *All pendant nodes of  $f_i(e_x) \cup f_i(e_y)$  are reticulations.*

**Proof** Consider any pendant node  $w$  of  $f_i(e_x) \cup f_i(e_y)$  and let  $e$  be the edge connecting  $w$  to a node  $v$  in  $f_i(e_x) \cup f_i(e_y)$ . Neither endpoint of  $e$  is the root of  $N$ . Since  $N$  is

a tree-child network, there exists a tree path  $Q$  from  $w$  to a leaf  $f_i(\ell_w)$ . Consider the path  $P$  from the root to  $\ell_w$  in  $T_i$ . Since  $e_x$  and  $e_y$  are not in  $P$ ,  $f_i(e_x) \cup f_i(e_y)$  and  $f_i(P)$  are edge-disjoint. On the other hand, since  $Q$  is a tree path,  $Q \subseteq f_i(P)$ . Since  $w$  is not the root of  $N$  and  $f_i(P)$ 's top endpoint is the root of  $N$ ,  $Q$  is a proper subpath of  $f_i(P)$ , that is,  $f_i(P)$  contains a parent edge of  $w$ . If  $f_i(P)$  contained  $e$ , then  $f_i(P)$  would be a proper superpath of  $Q \cup e$  because  $e$ 's top endpoint also is not the root of  $N$ . Thus,  $f_i(P)$  would contain the parent edge of  $v$ , that is,  $f_i(P)$  and  $f_i(e_x) \cup f_i(e_y)$  would not be edge-disjoint, a contradiction. Therefore,  $e \notin f_i(P)$  and  $w$  has another parent edge, that is,  $w$  is a reticulation.  $\square$

**Claim 2** *The undirected path  $f_i(e_x) \cup f_i(e_y)$  contains at most one reticulation. This reticulation is a child of  $f_i(p)$ .*

**Proof** We prove that only the top edge of  $f_i(e_x)$  can be a reticulation edge. An analogous argument shows that only the top edge of  $f_i(e_y)$  can be a reticulation edge. Thus, all reticulation edges in  $f_i(e_x) \cup f_i(e_y)$  are incident to  $f_i(p)$ . If the top edges of  $f_i(e_x)$  and  $f_i(e_y)$  are both reticulation edges, then both children of  $f_i(p)$  are reticulations, a contradiction because  $N$  is a tree-child network. Thus,  $f_i(e_x) \cup f_i(e_y)$  contains at most one reticulation.

So assume that  $f_i(e_x)$  contains a reticulation edge and choose such an edge  $e = vw$  that is farthest from  $f_i(p)$ . If  $v = p$ , then our claim holds. So assume that  $v \neq p$ . If  $v$  is a reticulation node, then  $w$  is its only child. Since  $e$  is a reticulation edge,  $w$  is also a reticulation node, a contradiction because  $N$  is a tree-child networks. Thus,  $v$  must be a tree node. By Claim 1, this implies that both of  $v$ 's children are reticulations, a contradiction again because  $N$  is a tree-child network.  $\square$

**Claim 3** *If the path  $f_i(e_x) \cup f_i(e_y)$  contains no reticulation, then it has at least one pendant node.*

**Proof** If  $f_i(e_x) \cup f_i(e_y)$  contains no reticulation and has no pendant nodes, then  $f_i(x)$  and  $f_i(y)$  are children of  $f_i(p)$  in  $N$ . Thus, both  $f_j(P_x)$  and  $f_j(P_y)$  include  $f_i(p)$ . Since  $f_j(P_x)$  and  $f_j(P_y)$  share only their top endpoint  $f_j(u)$ , we have  $f_j(u) = f_i(p)$  and thus  $f_j(P_x) = f_i(e_x)$  and  $f_j(P_y) = f_i(e_y)$ . This, however, is a contradiction because  $f_i(e_x) \cup f_i(e_y)$  has no pendant nodes but  $P_x \cup P_y$  has a pendant node in  $T_j$ , that is,  $f_j(P_x) \cup f_j(P_y)$  must also have a pendant node in  $N$ .  $\square$

For the next two claims, fix two distinct cherries  $\{x, y\}$  and  $\{w, z\}$  of two trees  $T_i \in \mathcal{T}$  and  $T_j \in \mathcal{T}$ , respectively. Let  $p$  be the common parent of  $x$  and  $y$  in  $T_i$ , and let  $q$  be the common parent of  $w$  and  $z$  in  $T_j$ .

**Claim 4**  *$f_i(e_x) \cup f_i(e_y)$  and  $f_j(e_w) \cup f_j(e_z)$  do not share any reticulation edge.*

**Proof** Assume the contrary. Then let  $e$  be a reticulation edge in  $(f_i(e_x) \cup f_i(e_y)) \cap (f_j(e_w) \cup f_j(e_z))$  and assume w.l.o.g. that  $e \in f_i(e_x) \cap f_j(e_w)$ . By Claim 2,  $f_i(p) = f_j(q)$ ;  $e$  is the first edge in both  $f_i(e_x)$  and in  $f_j(e_w)$ ;  $f_i(e_y)$  and  $f_j(e_z)$  are both tree paths from  $f_i(p)$  to  $f_i(y)$  and  $f_j(z)$ , respectively; and the subpaths of  $f_i(e_x)$  to  $f_i(e_w)$  from  $e$ 's bottom endpoint to  $f_i(x)$  and  $f_j(w)$ , respectively, are also tree paths.

Since every pendant node of  $f_i(e_y)$  is a reticulation, by Claim 1, none of these pendant nodes can belong to  $f_j(e_z)$ . Thus,  $f_j(z) = f_i(y)$ , that is,  $z = y$ . Similarly,



none of the pendant nodes of the subpath of  $f_i(x)$  from  $e$ 's bottom endpoint to  $f_i(x)$  can belong to  $f_j(w)$ . Thus,  $f_j(w) = f_i(x)$ , that is,  $w = x$ . This shows that  $\{x, y\} = \{w, z\}$ , a contradiction.  $\square$

**Claim 5** *If neither  $f_i(e_x) \cup f_i(e_y)$  nor  $f_j(e_w) \cup f_j(e_z)$  contains a reticulation edge, then these two undirected paths are vertex-disjoint.*

**Proof** Assume that neither  $f_i(e_x) \cup f_i(e_y)$  nor  $f_j(e_w) \cup f_j(e_z)$  contains a reticulation edge and assume first that  $f_i(e_x) \cup f_i(e_y)$  and  $f_j(e_w) \cup f_j(e_z)$  are not *edge-disjoint*. Then, w.l.o.g.,  $f_i(e_x)$  and  $f_j(e_w)$  share an edge  $e$ . Since  $f_i(e_x)$  and  $f_j(e_w)$  are tree paths, the same argument as in the proof of Claim 4 shows that  $x = w$ . If  $f_i(e_y)$  and  $f_j(e_z)$  also share an edge, then the same argument shows that  $y = z$ . Otherwise, w.l.o.g.  $f_j(q)$  is an internal node of  $f_i(e_x)$  and the first node after  $f_j(q)$  in  $f_j(e_z)$  is a pendant node of  $f_i(e_x)$ . By Claim 1, this node is a reticulation, a contradiction. This shows that  $f_i(e_x) \cup f_i(e_y)$  and  $f_j(e_w) \cup f_j(e_z)$  are edge-disjoint.

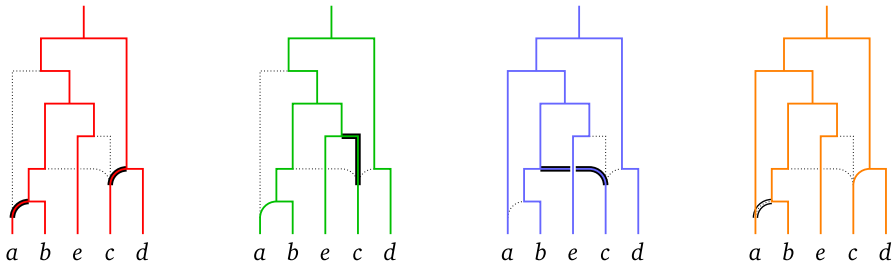
If  $f_i(e_x) \cup f_i(e_y)$  and  $f_j(e_w) \cup f_j(e_z)$  are edge-disjoint but not vertex-disjoint, then their shared vertex  $v$  satisfies either  $v \neq f_i(p)$  and  $v \neq f_j(q)$  or w.l.o.g.  $v = f_i(p)$ . In the former case, the parent edge of  $v$  belongs to both  $f_i(e_x) \cup f_i(e_y)$  and  $f_j(e_w) \cup f_j(e_z)$ , a contradiction. In the latter case, both child edges of  $v$  belong to  $f_i(e_x) \cup f_i(e_y)$  and  $f_j(e_w) \cup f_j(e_z)$  has to contain at least one of them, again a contradiction.  $\square$

Now we call a cherry  $\{x, y\}$  of some tree  $T_i$  a *type-I cherry* if the undirected path  $f_i(e_x) \cup f_i(e_y)$  contains a reticulation edge; otherwise, it is a *type-II cherry*. We charge each type-I cherry  $\{x, y\}$  to the reticulation edge in  $f_i(e_x) \cup f_i(e_y)$ . By Claim 4, every reticulation edge is charged for at most one type-I cherry. For every type-II cherry  $\{x, y\}$ , Claim 3 shows that w.l.o.g.,  $f(x)$ 's sibling  $v$  in  $N$  is a pendant node of  $f(e_x)$ . By Claim 1,  $v$  is a reticulation. Thus, the edge  $e$  between  $v$  and  $f(x)$ 's parent is a reticulation edge. We charge the cherry  $\{x, y\}$  to  $e$ . Since  $e$  has an endpoint in  $f(e_x)$ , Claim 5 implies that  $e$  is charged for only one type-II cherry. This proves that every reticulation edge is charged for at most two cherries, one of type I and one of type II. Figure 3 illustrates this. This finishes the proof.  $\square$

### 3.3 Bounding the Number of Cherries in General Irreducible Trees

Having shown, in Lemma 10, that Proposition 9 holds when  $S = \langle \rangle$ , we extend the proof to arbitrary partial tree-child cherry picking sequences in this section, thereby completing the proof of Proposition 9. The main idea is to construct a set of  $X$ -trees  $\hat{\mathcal{T}}$  that has the same set of cherries as  $\mathcal{T}/S$  (and in particular has no trivial cherries) and then show that  $\hat{\mathcal{T}}$  has reticulation number at most  $k$ . By Lemma 10, this implies that  $\hat{\mathcal{T}}$ , and thus  $\mathcal{T}/S$ , has at most  $4k$  cherries.

**Lemma 11** *Let  $\mathcal{T}$  be a set of  $X$ -trees and let  $S = \langle (x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_{r+1}, -) \rangle$  be a tree-child cherry picking sequence for  $\mathcal{T}$  of weight at most  $k$ . For any  $j \in [r]_0$ , either there exists a trivial cherry of  $\mathcal{T}^{(j)}$ , or  $\mathcal{T}^{(j)}$  has at most  $4k$  unique cherries.*



**Fig. 3** The embeddings of the four trees in Fig. 1b into the network in Fig. 2a. The set of cherries of these trees is  $\{\{a, b\}, \{c, d\}, \{c, e\}, \{b, c\}, \{b, e\}\}$ . We use the embedding of the first tree to charge the cherries  $\{a, b\}$  and  $\{c, d\}$  to reticulation edges, the embedding of the second tree to charge the cherry  $\{c, e\}$  to a reticulation edge, the embedding of the third tree to charge the cherry  $\{b, c\}$  to a reticulation edge, and the embedding of the fourth tree to charge the cherry  $\{b, e\}$  to a reticulation edge. The cherries  $\{a, b\}$  and  $\{c, d\}$  are type-I cherries because the two undirected paths between  $a$  and  $b$  and between  $c$  and  $d$  in the embedding of the first tree contain the two highlighted reticulation edges, which are the edges we charge for these cherries. Similarly, the cherry  $\{c, e\}$  is a type-I cherry because the undirected path between  $c$  and  $e$  in the embedding of the second tree contains the highlighted reticulation edge, which is charged for this cherry. The cherry  $\{b, c\}$  is also a type-I cherry because the undirected path between  $b$  and  $c$  in the embedding of the third tree contains the highlighted reticulation edge, which is charged for this cherry. Finally, the cherry  $\{b, e\}$  is a type-II cherry because the undirected path between  $b$  and  $e$  in the embedding of the fourth tree contains only tree edges. The parent of  $b$  on this path has a reticulation node as its other child, and the edge between  $b$ 's parent and this reticulation is the edge we charge for the cherry  $\{b, e\}$

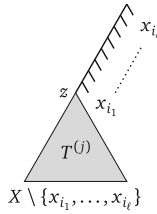
**Proof** For  $j = 0$ , the claim holds by Lemma 10. For  $j > 0$ , we cannot apply Lemma 10 directly because the trees in  $\mathcal{T}^{(j)}$  may have different leaf sets. Assume that  $\mathcal{T}^{(j)}$  has no trivial cherry, because otherwise the lemma holds. In order to use Lemma 10 to bound the number of unique cherries in  $\mathcal{T}^{(j)}$ , we transform  $\mathcal{T}^{(j)}$  into a set of  $X$ -trees  $\hat{\mathcal{T}}^{(j)}$  with the following properties:

1.  $\hat{\mathcal{T}}^{(j)}$  has the same unique cherries as  $\mathcal{T}^{(j)}$ ;
2.  $\hat{\mathcal{T}}^{(j)}$  has no trivial cherries; and
3.  $\hat{\mathcal{T}}^{(j)}$  has tree-child hybridization number at most  $k$ .

By Properties 2 and 3 and Lemma 10,  $\hat{\mathcal{T}}^{(j)}$  has at most  $4k$  unique cherries. Thus, by Property 1,  $\mathcal{T}^{(j)}$  has at most  $4k$  unique cherries.

To obtain  $\hat{\mathcal{T}}^{(j)}$  from  $\mathcal{T}^{(j)}$ , let  $\mathcal{T}' \subseteq \mathcal{T}$  be the subset of trees  $T \in \mathcal{T}$  such that  $T^{(j)}$  has at least two leaves. We can assume that  $\mathcal{T}' \neq \emptyset$  because otherwise,  $\mathcal{T}^{(j)}$  has no cherries and the claim holds. Also note that every cherry of  $\mathcal{T}^{(j)}$  is a cherry of some tree  $T^{(j)}$  with  $T \in \mathcal{T}'$ . Now consider any tree  $T \in \mathcal{T}'$  and let  $i_1 < \dots < i_\ell$  be the indices in  $[j]$  such that  $(x_{i_h}, y_{i_h})$  is a cherry of  $T^{(i_h-1)}$  for all  $1 \leq h \leq \ell$ . In other words,  $T^{(i)} \neq T^{(i-1)}$  if and only if  $i \in \{i_1, \dots, i_\ell\}$ . Observe that  $T^{(j)}$  has label set  $X \setminus \{x_{i_1}, \dots, x_{i_\ell}\}$ . Let  $C$  be a caterpillar with leaf set  $\{z, x_{i_1}, \dots, x_{i_\ell}\}$ , from bottom to top; that is,  $C$  is a tree such that  $z$  and  $x_{i_1}$  are siblings and, for  $1 \leq j < \ell$ , the parent of  $x_{i_j}$  is a sibling of  $x_{i_{j+1}}$ . We construct a tree  $\hat{T}^{(j)}$  from  $T^{(j)}$  and  $C$  by identifying  $z$  with the root of  $T^{(j)}$ . This is illustrated in Fig. 4.  $\hat{\mathcal{T}}^{(j)}$  is the set of all such trees  $\hat{T}^{(j)}$ :  $\hat{\mathcal{T}}^{(j)} = \{\hat{T}^{(j)} \mid T \in \mathcal{T}'\}$ .

Property 1 holds because the trees in  $\mathcal{T}^{(j)} \setminus (\mathcal{T}')^{(j)}$  have no cherries and, for every tree  $T \in \mathcal{T}'$ ,  $\hat{T}^{(j)}$  has the same cherries as  $T^{(j)}$ :  $T^{(j)}$  is a pendant subtree of  $\hat{T}^{(j)}$ , so



**Fig. 4** The construction of the tree  $\hat{T}^{(j)}$  from  $T^{(j)}$  and a caterpillar  $C$  with leaf set  $\{z, x_{i_1}, \dots, x_{i_\ell}\}$

every cherry of  $T^{(j)}$  is a cherry of  $\hat{T}^{(j)}$ . Every cherry of  $\hat{T}^{(j)}$  that is not a cherry of  $T^{(j)}$  would have to involve some leaf  $x_{i_h}$ , but none of these leaves is part of a cherry because  $T^{(j)}$  has at least two leaves.

To see that Property 2 holds, observe that every trivial cherry  $\{x, y\}$  would have to be a cherry of every tree in  $\hat{\mathcal{T}}^{(j)}$  because all trees in  $\hat{\mathcal{T}}^{(j)}$  have the same label set. Thus, by Property 1,  $\{x, y\}$  would be a cherry of every tree  $T \in \mathcal{T}'$ . By the definition of  $\mathcal{T}'$ ,  $\{x, y\}$  would therefore be a trivial cherry of  $\mathcal{T}^{(j)}$ , but  $\mathcal{T}^{(j)}$  has no trivial cherries. Thus,  $\hat{\mathcal{T}}^{(j)}$  has no trivial cherries.

To prove that  $\hat{\mathcal{T}}^{(j)}$  has tree-child hybridization number at most  $k$  (Property 3), we construct a tree-child cherry picking sequence  $\hat{S}^{(j)}$  of weight at most  $k$  for  $\hat{\mathcal{T}}^{(j)}$ . This sequence is defined as

$$\hat{S}^{(j)} = \langle (x_{j+1}, y_{j+1}), \dots, (x_r, y_r), (x_1, x_{r+1}), \dots, (x_j, x_{r+1}), (x_{r+1}, -) \rangle,$$

that is, we swap the subsequences  $\langle (x_1, y_1), \dots, (x_j, y_j) \rangle$  and  $\langle (x_{j+1}, y_{j+1}), \dots, (x_r, y_r) \rangle$  of  $S$  and then replace  $y_i$  with  $x_{r+1}$  in every pair  $(x_i, y_i)$  with  $1 \leq i \leq j$ . By construction,  $\hat{S}^{(j)}$  has the same weight as  $S$ , that is, its weight is at most  $k$ .

To see that  $\hat{S}^{(j)}$  is a tree-child cherry picking sequence, observe that  $\langle (x_{j+1}, y_{j+1}), \dots, (x_r, y_r) \rangle$  is a subsequence of a tree-child cherry picking sequence, namely  $S$ , and is thus a partial tree-child cherry picking sequence. Since  $S$  reduces each tree in  $\mathcal{T}$  to the single leaf  $x_{r+1}$ , we have  $x_{r+1} \notin \{x_1, \dots, x_r\}$ , so  $x_{r+1}$  is not forbidden with respect to  $\langle (x_{j+1}, y_{j+1}), \dots, (x_r, y_r), (x_1, x_{r+1}), \dots, (x_i, x_{r+1}) \rangle$ , for any  $i \in [j]_0$ . Thus,  $\hat{S}^{(j)}$  is a tree-child cherry picking sequence.

It remains to prove that  $\hat{S}^{(j)}$  is a cherry picking sequence for every tree  $\hat{T}^{(j)} \in \hat{\mathcal{T}}^{(j)}$ . Observe that the sequence  $S' = \langle (x_{j+1}, y_{j+1}), \dots, (x_r, y_r) \rangle$  reduces  $T^{(j)}$  to the single leaf  $x_{r+1}$ . Thus, after applying  $S'$  to  $\hat{T}^{(j)}$ , we obtain a subtree  $C'$  of the caterpillar  $C$  with  $z$  replaced with  $x_{r+1}$ . ( $S'$  may also delete some leaves of  $C$ .) Since the leaves  $x_{i_1}, \dots, x_{i_\ell}$  of  $C$  appear in this order from bottom to top in  $C$ , the sequence  $\langle (x_1, x_{r+1}), \dots, (x_j, x_{r+1}) \rangle$  reduces  $C'$  to the single leaf  $x_{r+1}$ . Thus,  $\hat{S}^{(j)}$  is a cherry picking sequence for  $\hat{T}^{(j)}$ .  $\square$

### 3.4 Proof of Theorem 3

Using the results from the previous three subsections, we are now ready to prove Theorem 3. While our algorithm computes  $\mathcal{T}'$  only in line 3, and  $n', k',$  and  $C$

only in lines 7–9, it is convenient for the sake of this proof to view them as quantities that evolve over time, as functions of  $S$ . We define  $n'(\mathcal{T}, S) = |\{x \in X \mid x \text{ is a leaf of a tree in } \mathcal{T}/S\}|$  and  $k'(\mathcal{T}, S) = |S| - |X| + n'(\mathcal{T}, S)$  for any partial tree-child cherry picking sequence  $S$ .

We divide the proof of Theorem 3 into three parts. First, we prove that  $k'(\mathcal{T}, S)$  is invariant over the course of any invocation  $\text{TCS}(\mathcal{T}, S, k)$  and that  $0 \leq k'(\mathcal{T}, S) \leq k$  in every invocation the algorithm makes. This will be used in the analysis of the running time of the algorithm and in proving the correctness of the algorithm in the case when it returns a sequence in line 11. Then, we bound the running time of the algorithm by  $O((8k)^k nt \lg t + nt \lg nt)$ , where  $n = |X|$  and  $t = |\mathcal{T}|$ . This implies in particular that the number of recursive calls the algorithm makes is finite, a fact that will be used in the correctness proof. Finally, we consider the tree of recursive calls the algorithm makes and use induction on the number of descendant invocations of any invocation  $\text{TCS}(\mathcal{T}, S, k)$  to prove the correctness of this invocation.

**Lemma 12** *For a collection of  $X$ -trees  $\mathcal{T}$ , any partial cherry picking sequence  $S$ , and any non-trivial cherry  $\{x, y\}$  of  $\mathcal{T}/S$ ,  $k'(\mathcal{T}, S \circ \langle(x, y)\rangle) = k'(\mathcal{T}, S) + 1$ .*

**Proof** Since  $\{x, y\}$  is a non-trivial cherry of  $\mathcal{T}/S$ , there exists a tree  $T/S \in \mathcal{T}/S$  that contains both  $x$  and  $y$  but not the cherry  $\{x, y\}$ . Thus, applying the pair  $(x, y)$  to  $\mathcal{T}/S$  does not remove  $x$  from all trees in  $\mathcal{T}/S$ . In particular,  $n'(\mathcal{T}, S \circ \langle(x, y)\rangle) = n'(\mathcal{T}, S)$  and, therefore,  $k'(\mathcal{T}, S \circ \langle(x, y)\rangle) = |S \circ \langle(x, y)\rangle| - |X| + n'(\mathcal{T}, S \circ \langle(x, y)\rangle) = |S| + 1 - |X| + n'(\mathcal{T}, S) = k'(\mathcal{T}, S) + 1. \quad \square$

**Lemma 13** *The value of  $k'(\mathcal{T}, S)$  is invariant over the course of any invocation  $\text{TCS}(\mathcal{T}, S, k)$  and satisfies  $0 \leq k'(\mathcal{T}, S) \leq k$ . Moreover, an invocation  $\text{TCS}(\mathcal{T}, S, k)$  satisfies  $k'(\mathcal{T}, S) = 0$  if and only if  $S = \langle \rangle$ .*

**Proof** First we prove that  $k'(\mathcal{T}, S)$  does not change over the course of any invocation  $\text{TCS}(\mathcal{T}, S, k)$ . Note that in a given invocation  $\text{TCS}(\mathcal{T}, S, k)$ ,  $S$  changes only in line 2. Each execution of line 2 adds a pair  $(x, y)$  to  $S$ , thereby increasing  $|S|$  by one. Since  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}$  and  $y$  is not forbidden with respect to  $S$  in this case, this also removes  $x$  from all trees in  $\mathcal{T}/S$ , so  $n'(\mathcal{T}, S)$  decreases by one and  $k'(\mathcal{T}, S) = |S| - |X| + n'(\mathcal{T}, S)$  remains unchanged.

We prove the bounds on  $k'(\mathcal{T}, S)$  for each invocation  $\text{TCS}(\mathcal{T}, S, k)$  by induction on  $|S|$ .

If  $|S| = 0$ , then  $S = \langle \rangle$ . In this case,  $\mathcal{T}/S = \mathcal{T}$ , so  $n'(\mathcal{T}, S) = |X|$ , that is,  $k'(\mathcal{T}, S) = |S| - |X| + n'(\mathcal{T}, S) = |S| - |X| + |X| = 0$ .

If  $|S| > 0$ , then  $\text{TCS}(\mathcal{T}, S, k)$  is called by another invocation  $\text{TCS}(\mathcal{T}, S', k)$  with  $|S'| < |S|$ . By the induction hypothesis, we have  $k'(\mathcal{T}, S') \geq 0$ . Let  $S''$  be a snapshot of  $S'$  in line 12 of the invocation  $\text{TCS}(\mathcal{T}, S', k)$ . Then  $S = S'' \circ \langle(x, y)\rangle$ , where  $\{x, y\}$  is a non-trivial cherry of  $\mathcal{T}/S''$ . Thus, by Lemma 12,  $k'(\mathcal{T}, S) = k'(\mathcal{T}, S'') + 1$ . Since  $k'(\mathcal{T}, S'') = k'(\mathcal{T}, S')$ , this implies that  $k'(\mathcal{T}, S) > k'(\mathcal{T}, S') \geq 0$ . By the second condition in line 12, we have  $k'(\mathcal{T}, S'') < k$  (because  $\text{TCS}(\mathcal{T}, S', k)$  makes the recursive call  $\text{TCS}(\mathcal{T}, S, k)$ ), so  $k'(\mathcal{T}, S) = k'(\mathcal{T}, S'') + 1 \leq k. \quad \square$

The following proposition now establishes the running time bound stated in Theorem 3.

**Proposition 14** *The total running time of the invocation  $\text{TCS}(\mathcal{T}, \langle \rangle, k)$  and all its descendant invocations is  $O((8k)^k nt \lg t + nt \lg nt)$ , where  $n = |X|$  and  $t = |\mathcal{T}|$ .*

**Proof** We only provide a sketch of the argument that the algorithm’s state can be initialized in  $O(nt \lg nt)$  time and that each invocation of procedure TCS, excluding the recursive calls it makes, has cost  $O(nt \lg t)$ . A careful proof is straightforward but tedious. To prove the proposition, it then suffices to prove that the algorithm makes  $O((8k)^k)$  recursive calls.

Instead of computing  $\mathcal{T}'$  from scratch as in the pseudo-code of procedure TCS, we first construct the state of the top-level invocation  $\text{TCS}(\mathcal{T}, \langle \rangle, k)$  consisting of  $\mathcal{T}'$  and the lists of trivial and non-trivial cherries. Whenever an invocation makes a recursive call, it makes a copy of its state to be modified by the recursive call.

Identifying the cherries in  $\mathcal{T}' = \mathcal{T}$  for the top-level invocation  $\text{TCS}(\mathcal{T}, \langle \rangle, k)$  takes  $O(nt \lg nt)$  time using appropriate dictionaries (e.g., balanced binary search trees) to identify leaves with the same labels in different trees and to collect all occurrences of the same cherry in different trees.

Copying the state of the current invocation for each recursive call the algorithm makes takes  $O(nt)$  time because the state is easily seen to have size  $O(nt)$ . We charge this cost to the recursive call. Each pair added to  $S$  eliminates the corresponding cherry from up to  $t$  trees and thereby creates up to  $t$  new cherries. Updating  $\mathcal{T}'$  and the lists of trivial and non-trivial cherries for each such cherry takes  $O(\lg t)$  time,  $O(t \lg t)$  time in total for each pair added to  $S$ . Each invocation adds at most  $n$  pairs corresponding to trivial cherries to  $S$ , in line 2. Each pair  $(x, y)$  added to  $S$  in line 17 can be charged to the recursive call  $\text{TCS}(\mathcal{T}, S \circ \langle (x, y) \rangle, k)$  made in line 17. Thus, each invocation adds at most one pair corresponding to a non-trivial cherry to  $S$ . The cost of updating  $\mathcal{T}'$  and the list of trivial and non-trivial cherries in each invocation is thus  $O(nt \lg t)$ . Adding the cost of making a copy of the parent invocation’s state at the beginning of each invocation, the cost per invocation is thus  $O(nt \lg t)$ . To obtain the time bound stated in the proposition, it remains to bound the number of recursive calls the algorithm makes by  $O((8k)^k)$ .

Let  $m_{k'}$  be the number of invocations  $\text{TCS}(\mathcal{T}, S, k)$  with  $k'(\mathcal{T}, S) = k'$ . By Lemma 13, every invocation  $\text{TCS}(\mathcal{T}, S, k)$  the algorithm makes satisfies  $0 \leq k'(\mathcal{T}, S) \leq k$  and the total number of invocations is therefore  $\sum_{k'=0}^k m_{k'}$ . Also by Lemma 13, there is exactly one invocation  $\text{TCS}(\mathcal{T}, S, k)$  with  $k'(\mathcal{T}, S) = 0$ , namely the top-level invocation  $\text{TCS}(\mathcal{T}, \langle \rangle, k)$ . Finally, by Lemma 12, every child invocation  $\text{TCS}(\mathcal{T}, S_2, k)$  of an invocation  $\text{TCS}(\mathcal{T}, S_1, k)$  satisfies  $k'(\mathcal{T}, S_2, k) = k'(\mathcal{T}, S_1, k) + 1$ . Thus, since each invocation makes at most  $8k$  recursive calls in line 17, we obtain  $m_{k'+1} \leq 8k \cdot m_{k'}$ . A simple inductive argument now shows that  $m_{k'} \leq (8k)^{k'}$  for all  $0 \leq k' \leq k$ . Thus, the total number of recursive calls the algorithm makes is at most  $\sum_{k'=0}^k (8k)^{k'} = \frac{(8k)^{k+1} - 1}{8k - 1} = O((8k)^k)$ .  $\square$

To establish the correctness of procedure TCS, we need a few simple auxiliary lemmas.

**Lemma 15** *Let  $S$  be a partial cherry picking sequence  $S$  without any pairs of the form  $(x, -)$ . Any solution of  $(\mathcal{T}, S)$  has weight at least  $k'(\mathcal{T}, S)$ .*

**Proof** Consider any cherry picking sequence  $S \circ S'$  for  $\mathcal{T}$ . Let  $X_1$  be the set of leaf labels of the trees in  $\mathcal{T}/(S \circ S')$ , and let  $X_2$  be the subset of leaf labels of the trees in  $\mathcal{T}/S$  that are not in  $X_1$ . Then  $n'(\mathcal{T}, S) = |X_1| + |X_2|$ .

Every leaf  $x \in X_2$  must be removed from the trees in  $\mathcal{T}/S$  by at least one pair  $(x, y) \in S'$ . For every leaf  $x \in X_1$ ,  $S'$  must contain a pair  $(x, -)$ . Thus,  $|S'| \geq |X_1| + |X_2| = n'(\mathcal{T}, S)$ . Therefore,  $|S \circ S'| - |X| = |S| + |S'| - |X| \geq |S| - |X| + n'(\mathcal{T}, S) = k'(\mathcal{T}, S)$ .  $\square$

**Lemma 16** *Let  $\mathcal{T}$  be a collection of  $X$ -trees, and  $S$  a partial tree-child cherry picking sequence such that at least one tree in  $\mathcal{T}/S$  has more than one leaf. Then any optimal solution of  $(\mathcal{T}, S)$  is an extension of some sequence  $S \circ \langle(x, y)\rangle$ , where  $\{x, y\}$  is a cherry of  $\mathcal{T}/S$ .*

**Proof** Consider any optimal solution  $S \circ S'$  of  $(\mathcal{T}, S)$ . Since there exists a tree  $T \in \mathcal{T}$  such that  $T/S$  has at least two leaves, the first pair in  $S'$  is a pair  $(x, y)$  with  $x, y \in X$ . Let  $S' = \langle(x, y)\rangle \circ S''$  and assume for the sake of contradiction that  $\{x, y\}$  is not a cherry of any tree in  $\mathcal{T}/S$ . Then  $S \circ S'' \subset S \circ S'$ , so  $S \circ S''$  is a tree-child cherry picking sequence and  $|S \circ S''| < |S \circ S'|$ . Since  $\{x, y\}$  is not a cherry of any tree in  $\mathcal{T}/S$ , we have  $T/(S \circ \langle(x, y)\rangle) = \mathcal{T}/S$  for all  $T \in \mathcal{T}$ . Thus,  $T/(S \circ S'') = T/(S \circ \langle(x, y)\rangle \circ S'') = T/(S \circ S')$  for all  $T \in \mathcal{T}$ . Since  $S \circ S'$  is a cherry picking sequence for  $\mathcal{T}$ , this shows that  $S \circ S''$  is a cherry picking sequence for  $\mathcal{T}$ , a contradiction.  $\square$

The following proposition now finishes the proof of Theorem 3 by proving that the invocation  $\text{TCS}(\mathcal{T}, \langle\rangle, k)$  returns a shortest tree-child cherry picking sequence for  $\mathcal{T}$  if and only if  $\mathcal{T}$  has a tree-child cherry picking sequence of weight at most  $k$ .

**Proposition 17** *Given a set  $\mathcal{T}$  of  $X$ -trees, a partial tree-child cherry picking sequence  $S$ , and an integer  $k$ ,  $\text{TCS}(\mathcal{T}, S, k)$  returns an optimal solution of  $(\mathcal{T}, S)$  if and only if  $(\mathcal{T}, S)$  has a solution of weight at most  $k$ . Otherwise, it returns NONE.*

**Proof** Consider the tree of recursive calls made by the algorithm and let  $|\text{TCS}(\mathcal{T}, S, k)|$  be the number of descendant invocations of the invocation  $\text{TCS}(\mathcal{T}, S, k)$ , including the invocation  $\text{TCS}(\mathcal{T}, S, k)$  itself. By Proposition 14,  $|\text{TCS}(\mathcal{T}, S, k)|$  is finite. Thus, we can use induction on  $|\text{TCS}(\mathcal{T}, S, k)|$  to prove the proposition.

If  $|\text{TCS}(\mathcal{T}, S, k)| = 1$ , then  $\text{TCS}(\mathcal{T}, S, k)$  makes no recursive calls. Thus, it returns a sequence in line 11 or NONE in line 5 or 13. (Note that  $\text{TCS}(\mathcal{T}, S, k)$  cannot reach line 20 without making a recursive call, as this is only possible if  $|C| = 0$  or every cherry  $\{x, y\}$  of some tree in  $\mathcal{T}$  has  $x, y$  both forbidden, and these cases are covered by lines 11 and 5, respectively.) By Proposition 8, if  $S_1$  is a snapshot of  $S$  at the start of the invocation  $\text{TCS}(\mathcal{T}, S, k)$  and  $S_2$  is a snapshot of  $S$  in line 3, then  $(\mathcal{T}, S_1)$  has a solution of weight at most  $k$  if and only if  $(\mathcal{T}, S_2)$  has a solution of weight at most  $k$ , and any optimal solution of  $(\mathcal{T}, S_2)$  also is an optimal solution of  $(\mathcal{T}, S_1)$ .

If  $\text{TCS}(\mathcal{T}, S, k)$  returns NONE in line 5, then  $\mathcal{T}/S_2$  has a cherry  $\{x, y\}$  with both  $x$  and  $y$  forbidden with respect to  $S_2$ . Any solution  $S_2 \circ S'$  of  $(\mathcal{T}, S_2)$  must include the pair  $(x, y)$  or  $(y, x)$  in  $S'$  because otherwise the tree in  $\mathcal{T}/S_2$  that has  $\{x, y\}$  as a cherry is not reduced to a single leaf by  $S'$ . Since both  $x$  and  $y$  are forbidden with respect to  $S_2$ , there is no such extension  $S_2 \circ S'$  of  $S_2$  that is tree-child. Thus,  $(\mathcal{T}, S_2)$  has no solution, and neither does  $(\mathcal{T}, S_1)$ . It is therefore correct to return NONE.

If  $TCS(\mathcal{T}, S, k)$  returns  $S_2 \circ \langle(x, -)\rangle$  in line 11, then observe that  $S_2$  is a partial tree-child cherry picking sequence. Indeed, by the assumption of the proposition,  $S_1$  is a partial tree-child cherry picking sequence. For every pair  $(x, y)$  added to  $S$  in line 2,  $y$  is not forbidden with respect to  $S$ , so  $S \circ \langle(x, y)\rangle$  is also tree-child. By applying this argument inductively, we conclude that  $S_2$  is tree-child.

Since  $TCS(\mathcal{T}, S, k)$  returns  $S_2 \circ \langle(x, -)\rangle$  in line 11 only if  $|C| = 0$ ,  $S_2$  reduces each tree in  $\mathcal{T}$  to a single leaf. Since  $S_2$  is tree-child, this is the same leaf  $x$  for every tree  $T \in \mathcal{T}$ . Indeed, assume that  $S_2$  reduces some tree  $T \in \mathcal{T}$  to some leaf  $x$  and another tree  $T' \in \mathcal{T}$  to some leaf  $y \neq x$ . Let  $x'$  be the last leaf pruned from  $T$  by  $S_2$ , and let  $y'$  be the last leaf pruned from  $T'$  by  $S_2$ . Then  $S_2$  contains the two pairs  $(x', x)$  and  $(y', y)$ . W.l.o.g., assume that  $(x', x)$  occurs before  $(y', y)$  in  $S_2$ . Since  $S_2$  reduces  $T$  to the single leaf  $x$ , it prunes  $y$  from  $T$ . Since  $x'$  is the last leaf pruned from  $T$ , the pair  $(y, z)$  in  $S_2$  used to prune  $y$  from  $T$  cannot occur after the pair  $(x', x)$  in  $S_2$ . Thus, this pair  $(y, z)$  occurs before the pair  $(y', y)$  and  $S_2$  is not a tree-child cherry picking sequence, a contradiction.

Since  $S_2$  reduces every tree  $T \in \mathcal{T}$  to the same leaf  $x$ , the sequence  $S_2 \circ \langle(x, -)\rangle$  is a solution of  $(\mathcal{T}, S_2)$ . Since every solution  $S_2 \circ S'$  of  $(\mathcal{T}, S_2)$  must include at least one pair  $(z, -)$  in  $S'$ ,  $S_2 \circ \langle(x, -)\rangle$  is an optimal solution of  $(\mathcal{T}, S_2)$  and, therefore, also of  $(\mathcal{T}, S_1)$ . Finally, by Lemma 13,  $|S_2| - |X| + n'(\mathcal{T}, S_2) = k'(\mathcal{T}, S_2) \leq k$ ;  $n'(\mathcal{T}, S_2) = 1$  because, as just observed, each tree in  $\mathcal{T}/S_2$  has  $x$  as its only leaf. Thus,  $|S_2| - |X| < k$  and  $|S_2 \circ \langle(x, -)\rangle| - |X| \leq k$ , that is,  $(\mathcal{T}, S_2)$  and  $(\mathcal{T}, S_1)$  both have solutions of weight at most  $k$  and returning  $S_2 \circ \langle(x, -)\rangle$  is correct.

Finally, if  $TCS(\mathcal{T}, S, k)$  returns NONE in line 13, then  $|C| > 8k$  or  $C \neq \emptyset$  and  $k'(\mathcal{T}, S_2, k) \geq k$ .

If  $|C| > 8k$ , then  $\mathcal{T}/S_2$  has more than  $4k$  unique cherries. Since  $\mathcal{T}/S_2$  has no trivial cherries, Proposition 9 shows that  $(\mathcal{T}, S_2)$  has no solution of weight at most  $k$ , and neither does  $(\mathcal{T}, S_1)$ . Thus, returning NONE is correct.

If  $C \neq \emptyset$  and  $k'(\mathcal{T}, S_2) \geq k$ , then observe that  $\{x, y\}$  is a non- $k$ -trivial cherry of  $\mathcal{T}/S_2$  for every pair  $(x, y) \in C$ . Lemma 12 shows that  $k'(\mathcal{T}, S_2 \circ \langle(x, y)\rangle) = k'(\mathcal{T}, S_2) + 1 > k$  for all  $(x, y) \in C$ . By Lemma 15, this shows that  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$  has no solution of weight at most  $k$  for any  $(x, y) \in C$ . By Lemma 16, this implies that  $(\mathcal{T}, S_2)$  has no solution of weight at most  $k$ , and neither does  $(\mathcal{T}, S_1)$ . Thus, returning NONE is correct. This finishes the proof that every invocation  $TCS(\mathcal{T}, S, k)$  that makes no recursive calls gives a correct answer.

Next consider an invocation  $TCS(\mathcal{T}, S, k)$  that does make recursive calls. Then  $C \neq \emptyset$ . By Lemma 16,  $(\mathcal{T}, S_2)$  (and thus  $(\mathcal{T}, S_1)$ ) has a solution of weight at most  $k$  if and only if there exists a pair  $(x, y) \in C$  such that  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$  has a solution of weight at most  $k$ . Moreover, if such a pair exists, then one such pair has the property that any optimal solution of  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$  also is an optimal solution of  $(\mathcal{T}, S_2)$  and, thus, of  $(\mathcal{T}, S_1)$ .

If there exists a pair  $(x, y)$  such that  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$  has a solution of weight at most  $k$ , then choose  $(x, y)$  so that any optimal solution of  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$  also is an optimal solution of  $(\mathcal{T}, S_1)$ . By the induction hypothesis, the invocation  $TCS(\mathcal{T}, S_2 \circ \langle(x, y)\rangle, k)$  in line 17 returns an optimal solution  $S'$  of  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$ . The solution  $S_{\text{opt}}$  of  $(\mathcal{T}, S_1)$  returned in line 20 is no longer than  $S'$ . Since  $S_{\text{opt}}$  is a solution of some instance  $(\mathcal{T}, S_2 \circ \langle(x', y')\rangle)$  with  $(x', y') \in C$ , it is a solution of  $(\mathcal{T}, S_2)$  and is thus

an optimal solution of  $(\mathcal{T}, S_2)$  and  $(\mathcal{T}, S_1)$ . Thus, the algorithm produces the correct answer.

If there is no pair  $(x, y) \in C$  such that  $(\mathcal{T}, S_2 \circ \langle(x, y)\rangle)$  has a solution of weight at most  $k$ , then all recursive calls made in line 17 of the invocation  $\text{TCS}(\mathcal{T}, S, k)$  return NONE. Thus,  $\text{TCS}(\mathcal{T}, S, k)$  also returns NONE. Since Lemma 16 shows that  $(\mathcal{T}, S_1)$  has no solution of weight at most  $k$  in this case, this is correct.  $\square$

## 4 Redundant Branch Elimination: A Heuristic Improvement

In this section, we discuss a method used in our implementation of procedure TCS to improve its running time. We prove that it preserves the correctness of the algorithm, but we do not know whether it provably improves the algorithm's running time. In this sense, it is a heuristic.

The intuition behind redundant branch elimination is the following: Suppose that  $\mathcal{T}/\langle(x, y), (z, w)\rangle$  and  $\mathcal{T}/\langle(z, w), (x, y)\rangle$  result in the same set of trees. (This can easily happen, for example, if  $x, y, z, w$  are all distinct.) Then the branch of the algorithm that starts by applying the sequence  $\langle(x, y), (z, w)\rangle$  finds a solution if and only if the branch that starts by applying the sequence  $\langle(z, w), (x, y)\rangle$  does. So the algorithm does not need to explore this second branch; it is redundant, and redundant branch elimination ensures that the algorithm does not make this recursive call.

Procedure TCS2 below is a modified version of procedure TCS that uses redundant branch elimination. The only difference between procedures TCS and TCS2 is that TCS2 maintains a set  $R$  of redundant pairs (with  $R$  set to  $\emptyset$  in the top-level invocation  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$ ) and ignores extensions  $S \circ \langle(x, y)\rangle$  of the current sequence  $S$  such that  $(x, y) \in R$ . If  $\{x, y\}$  is a trivial cherry, this means that the invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  returns NONE. If  $\{x, y\}$  is a non-trivial cherry, then  $\text{TCS2}(\mathcal{T}, S, k, R)$  does not make the recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R)$ . Note that  $R$  does not contain *all* redundant pairs for  $S$ , only a subset for which we prove below that they can safely be ignored based on the recursive calls the algorithm has made so far.

Procedure TCS2 calls a procedure UpdateR in lines 3 and 22. Given a partial tree-child cherry picking sequence  $S$ , a set of pairs  $R$  that are redundant for  $S$ , and a pair  $(x, y)$ ,  $\text{UpdateR}(\mathcal{T}, S, (x, y), R)$  returns the subset  $R' \subseteq R$  containing all pairs that are redundant also for  $S \circ \langle(x, y)\rangle$ .



The following definition formalizes the concept of a redundant pair.

---

**Procedure** TCS2( $\mathcal{T}, S, k, R$ )

---

**Input:** A collection of phylogenetic trees  $\mathcal{T}$ , a partial tree-child cherry picking sequence  $S$ , an integer  $k$ , and a set  $R$  of redundant pairs for  $S$

**Output:** An optimal solution of  $(\mathcal{T}, S)$  if  $(\mathcal{T}, S)$  has a solution of weight at most  $k$  and there do not exist a proper prefix  $S_p \subset S$  and a pair  $(x, y) \in R$  such that  $S_p \circ \langle (x, y) \rangle$  dominates some optimal solution of  $(\mathcal{T}, S)$ . NONE if  $(\mathcal{T}, S)$  has no solution of weight at most  $k$ . In any other case, the output may be NONE or a (possibly suboptimal) solution of  $(\mathcal{T}, S)$ .

```

1 while there exists a trivial cherry  $\{x, y\}$  of  $\mathcal{T}/S$  with  $y$  not forbidden with respect to  $S$  do
2   if  $(x, y) \notin R$  then
3      $R \leftarrow \text{UpdateR}(\mathcal{T}, S, (x, y), R)$ ;
4      $S \leftarrow S \circ \langle (x, y) \rangle$ ;
5   else
6     Return NONE;
7  $\mathcal{T}' \leftarrow \mathcal{T}/S$ ;
8 if  $\mathcal{T}'$  contains a cherry  $\{x, y\}$  with  $x, y$  both forbidden with respect to  $S$  then
9   return NONE;
10 else
11    $n' \leftarrow |\{x \in X : x \text{ is a leaf of a tree in } \mathcal{T}'\}|$ ;
12    $k' \leftarrow |S| - |X| + n'$ ;
13    $C \leftarrow \{(x, y) \mid \{x, y\} \text{ is a cherry of some tree in } \mathcal{T}'\}$ ;
14   if  $|C| = 0$  then
15     return  $S \circ \langle (x, -) \rangle$ , where  $x$  is the last remaining leaf in all trees;
16   else if  $|C| > 8k$  or  $k' \geq k$  then
17     return NONE;
18   else
19      $S_{\text{opt}} \leftarrow \text{NONE}$ ;
20      $R' \leftarrow R$ ;
21     foreach  $(x, y) \in C \setminus R$  with  $y$  not forbidden with respect to  $S$  do
22        $R'' \leftarrow \text{UpdateR}(\mathcal{T}, S, (x, y), R')$ ;
23        $S_{\text{temp}} \leftarrow \text{TCS2}(\mathcal{T}, S \circ \langle (x, y) \rangle, k, R'')$ ;
24       if  $w(S_{\text{temp}}) < w(S_{\text{opt}})$  then
25          $S_{\text{opt}} \leftarrow S_{\text{temp}}$ ;
26        $R' \leftarrow R' \cup \{(x, y)\}$ ;
27   return  $S_{\text{opt}}$ ;

```

---



---

**Procedure** UpdateR( $\mathcal{T}, S, (x, y), R$ )

---

**Input:** A collection of phylogenetic  $X$ -trees  $\mathcal{T}$ , a partial tree-child cherry picking sequence  $S$ , a pair  $(x, y) \in X \times X$ , a set  $R$  of redundant pairs for  $S$ ;

**Output:** A subset  $R' \subseteq R$  of redundant pairs for  $S \circ \langle (x, y) \rangle$ ;

```

1 return  $\{(x', y') \in R \mid x' \neq y \text{ and } \text{count}(x', y', \mathcal{T}/(S \circ \langle (x, y) \rangle)) = \text{count}(x', y', \mathcal{T}/S)\}$ ;

```

---

**Definition 1** Let  $\mathcal{T}$  be a set of  $X$ -trees,  $S$  a tree-child cherry picking sequence, and  $(x, y) \in X \times X$ . Let  $\text{count}(x, y, \mathcal{T}/S)$  be the number of trees in  $\mathcal{T}/S$  that have  $\{x, y\}$

as a cherry. An extension  $S \circ S'$  of  $S$  is *dominated* by  $S \circ \langle(x, y)\rangle$  if there exists an index  $j > 1$  that satisfies the following conditions:

- $(x, y)$  is the  $j$ th element of  $S'$ ;
- $\text{count}(x, y, \mathcal{T}/S) = \text{count}(x, y, \mathcal{T}/(S \circ S'_{1,j-1}))$ ; and
- for all  $(x', y') \in S'_{1,j-1}$ ,  $y' \neq x$  and  $\{x', y'\} \neq \{x, y\}$ .

If a sequence  $S \circ S' \circ \langle(x, y)\rangle$  is dominated by  $S \circ \langle(x, y)\rangle$ , we say that  $(x, y)$  is a *redundant pair* for  $S \circ S'$ .

The next observation follows immediately from Definition 1.

**Observation 18** *If a sequence  $S \circ S'$  is dominated by  $S \circ \langle(x, y)\rangle$ , then so is any extension of  $S \circ S'$  and any prefix  $S \circ S'' \subseteq S \circ S'$  such that  $(x, y) \in S''$ .*

**Lemma 19** *If a sequence  $S \circ S'$  is dominated by  $S \circ \langle(x, y)\rangle$  and  $(x, y)$  is the  $j$ th pair in  $S'$ , then  $\text{count}(x, y, \mathcal{T}/(S \circ S'_{1,i})) = \text{count}(x, y, \mathcal{T}/(S \circ S'_{1,i-1}))$  for all  $i \in [j - 1]$ .*

**Proof** Let  $\mathcal{T}' = \mathcal{T}/S$  and let  $(x'_i, y'_i)$  be the  $i$ th pair in  $S'$ , for any index  $i \in [j - 1]$ . Since  $\{x'_i, y'_i\} \neq \{x, y\}$ , the pair  $(x'_i, y'_i)$  does not eliminate the cherry  $\{x, y\}$  from any tree in  $\mathcal{T}'/S'_{1,i-1}$  that contains this cherry, so  $\text{count}(x, y, \mathcal{T}'/S'_{1,i}) \geq \text{count}(x, y, \mathcal{T}'/S'_{1,i-1})$ . By Definition 1,  $\text{count}(x, y, \mathcal{T}') = \text{count}(x, y, \mathcal{T}'/S'_{1,j-1})$ . Thus, if  $\text{count}(x, y, \mathcal{T}'/S'_{1,i}) > \text{count}(x, y, \mathcal{T}'/S'_{1,i-1})$ , then there also exists an index  $i' \in [j - 1]$  such that  $\text{count}(x, y, \mathcal{T}'/S'_{1,i'}) < \text{count}(x, y, \mathcal{T}'/S'_{1,i'-1})$ , a contradiction because we just argued that  $\text{count}(x, y, \mathcal{T}'/S'_{1,i'}) \geq \text{count}(x, y, \mathcal{T}'/S'_{1,i'-1})$  for all  $i' \in [j - 1]$ . This proves that  $\text{count}(x, y, \mathcal{T}'/S'_{1,i}) = \text{count}(x, y, \mathcal{T}'/S'_{1,i-1})$  for all  $i \in [j - 1]$ . □

**Lemma 20** *Let  $(x, y) \in X \times X$ , and  $S \circ S_1 \circ S_2 \circ S_3$  a cherry picking sequence. If  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ \langle(x, y)\rangle$  and  $S \circ S_1 \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S'$ , for some sequence  $S'$ , then  $S \circ \langle(x, y)\rangle$  also dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ , for any sequence  $S''$ .*

**Proof** First assume that  $|S_1| > 0$ ,  $|S_3| > 0$ , and  $(x, y) \notin S_1 \circ S_2 \circ S_3$ . Then  $(x, y)$  is the  $j$ th element of  $S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ , for  $j = |S_1 \circ S_2 \circ S_3| + 1 > 1$ . Since  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ \langle(x, y)\rangle$  and  $(x, y) \notin S_1 \circ S_2$ , we have  $y' \neq x$  and  $\{x, y\} \neq \{x', y'\}$  for every pair  $(x', y') \in S_1 \circ S_2$  and Lemma 19 shows that  $\text{count}(x, y, \mathcal{T}/S) = \text{count}(x, y, \mathcal{T}/(S \circ S_1)) = \text{count}(x, y, \mathcal{T}/(S \circ S_1 \circ S_2))$ . Similarly, since  $S \circ S_1 \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S'$  and  $(x, y) \notin S_2 \circ S_3$ , we have  $y' \neq x$  and  $\{x', y'\} \neq \{x, y\}$  for every pair  $(x', y') \in S_2 \circ S_3$  and  $\text{count}(x, y, \mathcal{T}/(S \circ S_1)) = \text{count}(x, y, \mathcal{T}/(S \circ S_1 \circ S_2 \circ S_3))$ . Together, these two observations imply that  $\text{count}(x, y, \mathcal{T}/S) = \text{count}(x, y, \mathcal{T}/(S \circ S_1 \circ S_2 \circ S_3))$  and  $y' \neq x$  and  $\{x', y'\} \neq \{x, y\}$  for every pair  $(x', y') \in S_1 \circ S_2 \circ S_3$ . Thus,  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ .

If  $|S_1| = 0$ , then  $S \circ \langle(x, y)\rangle = S \circ S_1 \circ \langle(x, y)\rangle$  and it follows immediately that  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S'$ . By Observations 18, this implies that  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle$  and thus also  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ .

If  $|S_3| = 0$ , then  $S \circ S_1 \circ S_2 \circ \langle(x, y)\rangle = S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle$ , so it follows immediately that  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle$ . By Observation 18, this implies that  $S \circ \langle(x, y)\rangle$  also dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ .

If  $(x, y) \in S_1 \circ S_2$ , then the fact that  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ \langle(x, y)\rangle$  and Observation 18 imply that it also dominates  $S \circ S_1 \circ S_2$  and thus also  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ .

Finally, if  $(x, y) \in S_3$ , then consider the longest prefix  $S'_3 \subseteq S_3$  such that  $(x, y) \notin S'_3$ . Then, by Observation 18,  $S \circ S_1 \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S'_3 \circ \langle(x, y)\rangle$ . As shown so far, this implies that  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S'_3 \circ \langle(x, y)\rangle$ . Since  $S'_3 \circ \langle(x, y)\rangle$  is a prefix of  $S_3$  and, thus, of  $S_3 \circ \langle(x, y)\rangle \circ S''$ , Observation 18 now shows that  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$ .  $\square$

The significance of redundant pairs stems from the following proposition.

**Proposition 21** *Let  $\mathcal{T}$  be a set of  $X$ -trees, and  $S \circ S'$  a tree-child cherry picking sequence for  $\mathcal{T}$ . Suppose that  $S \circ S'$  is dominated by  $S \circ \langle(x, y)\rangle$ , for some pair  $(x, y) \in X \times X$ . Then there exists a tree-child cherry picking sequence  $S \circ \langle(x, y)\rangle \circ S''$  for  $\mathcal{T}$  with  $w(S \circ \langle(x, y)\rangle \circ S'') \leq w(S \circ S')$ .*

In other words: If some branch of the algorithm already looks for an optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$ , then there is no need to also look for an optimal solution of  $(\mathcal{T}, S \circ S'')$ , for any sequence  $S \circ S''$  that is dominated by  $S \circ \langle(x, y)\rangle$ .

**Proof** We can write  $S' = S'' \circ \langle(x, y)\rangle \circ S'''$  such that  $(x, y) \notin S''$ . Let  $|S''| = h$ . For  $0 \leq i \leq h$ , let  $S'_i = S''_{1,i} \circ \langle(x, y)\rangle \circ S'''_{i+1,h} \circ S'''$ . We prove by induction on  $h - i$  that  $S \circ S'_i$  is a tree-child cherry picking sequence for  $\mathcal{T}$ , for all  $0 \leq i \leq h$ . Since  $S'_0 = \langle(x, y)\rangle \circ S'' \circ S'''$  and  $w(S \circ S'_0) = w(S \circ S')$ , this proves the proposition.

$S \circ S'_h$  is clearly a tree-child cherry picking sequence for  $\mathcal{T}$  because  $S'_h = S'$ . So assume that  $i < h$  and that  $S \circ S'_{i+1}$  is a tree-child cherry picking sequence for  $\mathcal{T}$ .

Let  $(x', y')$  be the  $(i + 1)$ st pair in  $S''$ , that is,  $(x', y')$  is the predecessor pair of  $(x, y)$  in  $S'_{i+1}$ . Since  $S \circ \langle(x, y)\rangle$  dominates  $S \circ S'$ , the choice of  $S''$  implies that  $y' \neq x$  and, by Lemma 19,  $count(x, y, \mathcal{T}/(S \circ S''_{1,i})) = count(x, y, \mathcal{T}/(S \circ S''_{1,i+1}))$ . Since  $S \circ S'_{i+1}$  is tree-child, the former implies that  $S \circ S'_i$  is tree-child. We use the latter in the following proof that  $S \circ S'_i$  is a cherry picking sequence for  $\mathcal{T}$ .

Let  $T \in \mathcal{T}$  be an arbitrary tree, let  $T' = T/(S \circ S''_{1,i})$ , let  $T_a = T'/\langle(x', y'), (x, y)\rangle$ , and let  $T_b = T'/\langle(x, y), (x', y')\rangle$ . We show that  $T_b \subseteq T_a$  and that  $T_a \setminus T_b \subseteq \{x'\}$ . Thus, since  $S \circ S'_{i+1}$  is a tree-child cherry picking sequence and, therefore,  $x' \neq y''$  for all  $(x'', y'') \in S''_{i+2,h} \circ S'''$ , Lemma 7 shows that  $T/(S \circ S'_i) = T_b/(S''_{i+2,h} \circ S''') \subseteq T_a/(S''_{i+2,h} \circ S''') = T/(S \circ S'_{i+1})$ . Since  $T/(S \circ S'_{i+1})$  has a single leaf and  $T/(S \circ S'_i)$  has at least one leaf, this shows that  $T/(S \circ S'_i) = T/(S \circ S'_{i+1})$ , that is,  $S \circ S'_i$  is a cherry picking sequence for  $T$ . Since this is true for every tree  $T \in \mathcal{T}$ ,  $S \circ S'_i$  is a cherry picking sequence for  $\mathcal{T}$ .

It remains to show that  $T_b \subseteq T_a$  and  $T_a \setminus T_b \subseteq \{x'\}$ . Since  $count(x, y, \mathcal{T}/(S \circ S''_{1,i})) = count(x, y, \mathcal{T}/(S \circ S''_{1,i+1}))$ , either both  $T' = T/(S \circ S''_{1,i})$  and  $T'/\langle(x', y')\rangle = T/(S \circ S''_{1,i+1})$  contain  $\{x, y\}$  as a cherry or neither of them does.

If neither  $T'$  nor  $T'/\langle(x', y')\rangle$  contains  $\{x, y\}$  as a cherry, then  $T_a = T'/\langle(x', y'), (x, y)\rangle = T'/\langle(x', y')\rangle = T'/\langle(x, y), (x', y')\rangle = T_b$ , so  $T_b \subseteq T_a$  and  $T_a \setminus T_b = \emptyset \subseteq \{x'\}$ .

If both  $T'$  and  $T'/\langle(x', y')\rangle$  contain  $\{x, y\}$  as a cherry, then observe that  $T'/\langle(x', y')\rangle$  does not contain  $\{x', y'\}$  as a cherry. If  $T'$  also does not contain  $\{x', y'\}$  as a cherry, then we have that  $T_a = T'/\langle(x', y'), (x, y)\rangle = T'/\langle(x, y)\rangle$  and  $T_b = T'/\langle(x, y), (x', y')\rangle = T_a/\langle(x', y')\rangle$ . Since applying the pair  $(x', y')$  to  $T_a$  can only remove the leaf  $x'$ , this shows that  $T_a \subseteq T_b$  and  $T_a \setminus T_b \subseteq \{x'\}$ .

The final case is when  $T'$  contains both  $\{x, y\}$  and  $\{x', y'\}$  as cherries. Since  $\{x', y'\} \neq \{x, y\}$ ,  $T'$  must contain distinct vertices  $p$  and  $q$  such that  $p$  is the common parent of  $x$  and  $y$ , and  $q$  is the common parent of  $x'$  and  $y'$ . It follows that  $T_b$  and  $T_a$  can both be derived from  $T'$  by deleting  $x$  and  $x'$  and suppressing  $p$  and  $q$ . Thus,  $T_a = T_b$ , that is, once again,  $T_b \subseteq T_a$  and  $T_a \setminus T_b = \emptyset \subseteq \{x'\}$ .  $\square$

While our algorithm uses redundant pairs to ignore some dominated sequences in its search for a shortest tree-child cherry picking sequence, it cannot ignore *all* dominated sequences. Indeed, in many cases, every possible tree-child cherry picking sequence for  $\mathcal{T}$  is dominated by another sequence. Consider, for example, a binary tree on  $X = \{a, b, c, d\}$  with cherries  $\{a, b\}$  and  $\{c, d\}$ . Any sequence for this tree must begin with  $(a, b)$ ,  $(b, a)$ ,  $(c, d)$  or  $(d, c)$ . If the first pair is  $(a, b)$ , then the second pair must be either  $(c, d)$  or  $(d, c)$ . But the sequence  $\langle(a, b), (c, d)\rangle$  is dominated by  $\langle(c, d)\rangle$ , and similarly  $\langle(a, b), (d, c)\rangle$  is dominated by  $\langle(d, c)\rangle$ . A similar argument applies to any other sequence we might try. Thus, if we did ignore *all* redundant pairs for *every* sequence, the algorithm would not find any cherry picking sequence for  $\mathcal{T}$ . This is the reason why procedure TCS2 explicitly keeps a set  $R$  of redundant pairs that are safe to ignore; it ignores a sequence  $S \circ \langle(x, y)\rangle$  *only* if  $(x, y) \in R$ .

Following the terminology of Linz and Semple [14], we call a pair  $(x_j, y_j)$  in a partial cherry picking sequence  $S = \langle(x_1, y_1), \dots, (x_r, y_r)\rangle$  *essential* if  $\mathcal{T}/S_{1,j} \neq \mathcal{T}/S_{1,j-1}$ , that is,  $\{x_j, y_j\}$  is a cherry of at least one tree in  $\mathcal{T}/S_{1,j-1}$  and, therefore, applying the pair  $(x_j, y_j)$  to  $\mathcal{T}/S_{1,j-1}$  removes  $x_j$  from at least one tree in  $\mathcal{T}/S_{1,j-1}$ .

Our correctness proof of procedure TCS2 is divided into two parts: First we prove that if, for a given invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$ , every pair in  $S$  is essential and every pair in  $R$  is redundant for  $S$ , then

- (i) This is true at any time during the execution of this invocation (even though the invocation may modify  $S$  and  $R$ ) and
- (ii) For every recursive call  $\text{TCS2}(\mathcal{T}, S'', k, R'')$  this invocation makes, every pair in  $S''$  is essential and every pair in  $R''$  is redundant for  $S''$ .

Since the top-level invocation  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$  satisfies  $S = \langle\rangle$  and  $R = \emptyset$ , that is, all pairs in  $S$  are trivially essential and all pairs in  $R$  are trivially redundant for  $S$ , an inductive argument then implies that every pair in  $S$  is essential and every pair in  $R$  is redundant for  $S$  at any time during the execution of any invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$ . The second part of the proof shows that, under this condition, the invocation  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$  returns a shortest tree-child cherry picking sequence for  $\mathcal{T}$  if this sequence has weight at most  $k$ ; otherwise, it returns NONE.

The following lemma shows that replacing  $R$  with the set returned by  $\text{UpdateR}(\mathcal{T}, S, (x, y), R)$  whenever we append a pair  $(x, y)$  to a sequence  $S$  maintains the property that every pair in  $R$  is redundant for  $S$ .

**Lemma 22** *Let  $S \circ \langle (x, y) \rangle$  be a partial tree-child cherry picking sequence whose pairs are all essential, and let  $R \subseteq X \times X$ . For every pair  $(x', y')$  in the subset  $R' \subseteq R$  returned by  $\text{UpdateR}(\mathcal{T}, S, (x, y), R)$ , the sequence  $S \circ \langle (x, y), (x', y') \rangle$  is dominated by  $S \circ \langle (x', y') \rangle$ .*

**Proof** By the definition of  $R'$  in line 1 of procedure  $\text{UpdateR}$ , we have  $x' \neq y$  and  $\text{count}(x', y', \mathcal{T}/S) = \text{count}(x', y', \mathcal{T}/(S \circ \langle (x, y) \rangle))$  for all  $(x', y') \in R'$ . Observe also that  $\{x, y\} \neq \{x', y'\}$ . Indeed, since every pair in  $S \circ \langle (x, y) \rangle$  is essential, there exists a tree in  $\mathcal{T}/S$  that has  $\{x, y\}$  as a cherry, while there is no tree in  $\mathcal{T}/(S \circ \langle (x, y) \rangle)$  that has  $\{x, y\}$  as a cherry. Thus, if  $\{x, y\} = \{x', y'\}$ , we would have  $\text{count}(x', y', \mathcal{T}/S) \neq \text{count}(x', y', \mathcal{T}/(S \circ \langle (x, y) \rangle))$ , so  $(x', y') \notin R'$ . Since  $(x', y')$  is not the first pair in  $\langle (x, y), (x', y') \rangle$ , the sequence  $S \circ \langle (x, y), (x', y') \rangle$  is therefore dominated by  $S \circ \langle (x', y') \rangle$ .  $\square$

We are now ready to prove Claims (i) and (ii) above. Since each invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  may modify  $S$  and  $R$ , we use  $S_0$  and  $R_0$  to refer to the values of  $S$  and  $R$  passed as arguments to this invocation, and  $S$  and  $R$  to refer to the current values of  $S$  and  $R$  at any point during the execution of  $\text{TCS2}(\mathcal{T}, S, k, R)$ .

**Lemma 23** *Consider any invocation  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  such that every pair in  $S_0$  is essential and every pair in  $R_0$  is redundant for  $S_0$ . Then*

- (i) *At any time during the execution of this invocation, every pair in  $S$  is essential and there exists a proper prefix  $S_p \subset S_0$  for each pair  $(x', y') \in R$  such that  $S_p \circ \langle (x', y') \rangle$  dominates  $S \circ \langle (x', y') \rangle$ ; and*
- (ii) *For every recursive call  $\text{TCS2}(\mathcal{T}, S'', k, R'')$  this invocation makes, every pair in  $S''$  is essential and every pair in  $R''$  is redundant for  $S''$ .*

**Proof** (i) Initially, we have  $S = S_0$  and  $R = R_0$ . Thus, since every pair in  $S_0$  is essential and every pair in  $R_0$  is redundant for  $S_0$ , (i) holds for this choice of  $S$  and  $R$ . Next we prove that any modification the invocation makes to  $S$  and  $R$  maintains (i). Observe that  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  modifies  $S$  and  $R$  only in lines 3 and 4. Consider one iteration of the loop in lines 1–6 and let  $(x, y)$  be the pair added to  $S$  in this iteration. Since  $\{x, y\}$  is a trivial cherry of  $\mathcal{T}/S$  in this case and every pair in  $S$  essential, every pair in  $S \circ \langle (x, y) \rangle$  is essential. By Lemma 22, every pair  $(x', y')$  in the set  $R'$  returned by  $\text{UpdateR}(\mathcal{T}, S, (x, y), R)$  in line 3 has the property that  $S \circ \langle (x', y') \rangle$  dominates  $S \circ \langle (x, y), (x', y') \rangle$ . Since  $R' \subseteq R$ , there exists a proper prefix  $S_p \subset S_0$  such that  $S_p \circ \langle (x', y') \rangle$  dominates  $S \circ \langle (x', y') \rangle$ . Thus, by Lemma 20,  $S_p \circ \langle (x', y') \rangle$  also dominates  $S \circ \langle (x, y), (x', y') \rangle$  (where  $S$  and  $S \circ S_1$  in Lemma 20 correspond to  $S_p$  and  $S$  respectively,  $S_2 = \langle \rangle$ , and  $S_3 = \langle (x, y) \rangle$ ). Therefore, replacing  $S$  with  $S \circ \langle (x, y) \rangle$ , and  $R$  with the set returned by  $\text{UpdateR}(\mathcal{T}, S, (x, y), R)$  maintains that every pair in  $S$  is essential and, for every every pair  $(x', y') \in R$ , there exists a proper prefix  $S_p \subset S_0$  such that  $S_p \circ \langle (x', y') \rangle$  dominates  $S \circ \langle (x', y') \rangle$ .

(ii) Consider any recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle (x, y) \rangle, k, R'')$  the invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  makes in line 23. By (i), all pairs in  $S$  are essential. Since  $(x, y) \in C$ ,

$\{x, y\}$  is a cherry of  $\mathcal{T}/S$ . Thus, every pair in  $S \circ \langle(x, y)\rangle$  is essential. By Lemma 22, the set  $R''$  returned by  $\text{UpdateR}(\mathcal{T}, S, (x, y), R')$  in line 22 contains only pairs that are redundant for  $S \circ \langle(x, y)\rangle$ . Thus, (ii) holds.  $\square$

The following corollary follows by applying Lemma 23 inductively after observing that  $S_0 = \langle \rangle$  and  $R_0 = \emptyset$  for the top-level invocation  $\text{TCS2}(\mathcal{T}, \langle \rangle, k, \emptyset)$ .

**Corollary 24** *At any point during the execution of an invocation  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$ , there exists a proper prefix  $S_p \subset S_0$  for each pair  $(x', y') \in R$  such that  $S_p \circ \langle(x', y')\rangle$  dominates  $S \circ \langle(x', y')\rangle$ .*

The next lemma states the fairly weak correctness guarantee that each invocation  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  provides. As we show below, in Corollary 26, this lemma implies that the invocation  $\text{TCS2}(\mathcal{T}, \langle \rangle, k, \emptyset)$  returns a shortest tree-child cherry picking sequence for  $\mathcal{T}$  if there is such a sequence of weight at most  $k$ .

**Lemma 25** *Consider any invocation  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  the algorithm makes. If  $(\mathcal{T}, S_0)$  has a solution of weight at most  $k$ , then either  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  returns an optimal solution of  $(\mathcal{T}, S_0)$  or there exist an extension  $S_0 \circ S'$  of  $S_0$ , a pair  $(x, y) \in R_0$ , and a proper prefix  $S_p \subset S_0$  such that  $S_p \circ \langle(x, y)\rangle$  dominates  $S_0 \circ S'$ .*

**Proof** Since no invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  makes more recursive calls than the corresponding invocation  $\text{TCS}(\mathcal{T}, S, k)$ , Proposition 14 shows that each invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  has a finite number of descendant invocations, which we denote by  $|\text{TCS2}(\mathcal{T}, S, k, R)|$ . Thus, if the lemma does not hold, we can choose an invocation  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  that violates the lemma and has the minimum number of descendant invocations  $|\text{TCS2}(\mathcal{T}, S_0, k, R_0)|$  among all such invocations.

Since  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  fails to find an optimal solution of  $(\mathcal{T}, S_0)$ ,  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  returns NONE in line 6, 9, 17 or 27, or it returns a suboptimal solution of  $(\mathcal{T}, S_0)$  in line 15 or 27. Next we consider these different cases:

**TCS2( $\mathcal{T}, S_0, k, R_0$ ) returns NONE in line 9 or 17:** In this case,  $\text{TCS}(\mathcal{T}, S_0, k)$  would have returned NONE in line 5 or 13. Thus, by Proposition 17,  $(\mathcal{T}, S_0)$  has no solution of weight at most  $k$ , a contradiction.

**TCS2( $\mathcal{T}, S_0, k, R_0$ ) returns a sequence  $S_0 \circ S'$  in line 15:** In this case,  $\text{TCS}(\mathcal{T}, S_0, k)$  would have returned the same sequence in line 11. Thus, by Proposition 17,  $S_0 \circ S'$  is an optimal solution of  $(\mathcal{T}, S_0)$ , a contradiction.

**TCS2( $\mathcal{T}, S_0, k, R_0$ ) returns NONE in line 6:** In this case, consider the contents of  $S$  and  $R$  immediately before  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  returns. There exists a trivial cherry  $\{x, y\}$  of  $\mathcal{T}/S$  such that  $y$  is not forbidden with respect to  $S$  and  $(x, y) \in R$ . Since  $(\mathcal{T}, S_0)$  has a solution of weight at most  $k$ , Proposition 8 shows that  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  also has a solution of weight at most  $k$  and any optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  is also an optimal solution of  $(\mathcal{T}, S)$ . By Corollary 24, there exists a proper prefix  $S_p \subseteq S_0$  such that  $S_p \circ \langle(x, y)\rangle$  dominates  $S \circ \langle(x, y)\rangle$  and, thus, by Observation 18,  $S_p \circ \langle(x, y)\rangle \circ S'$ , for any optimal solution  $S \circ \langle(x, y)\rangle \circ S'$  of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$ , a contradiction.

**TCS2( $\mathcal{T}, S_0, k, R_0$ ) returns NONE or a suboptimal solution in line 27:** In this case, the corresponding invocation  $\text{TCS}(\mathcal{T}, S_0, k)$  would have reached line 20.

Since  $(\mathcal{T}, S_0)$  has a solution of weight at most  $k$ , Proposition 17 shows that  $\text{TCS}(\mathcal{T}, S_0, k)$  would have returned an optimal solution  $S_0 \circ S'$  of  $(\mathcal{T}, S_0)$ . This solution satisfies  $S_0 \circ S' = S \circ \langle(x, y)\rangle \circ S''$ , for some pair  $(x, y) \in C$ , referring to the state of  $S$  in line 3 of  $\text{TCS}(\mathcal{T}, S, k)$ . This shows that there exists a pair  $(x, y) \in C$  such that  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  has a solution of weight at most  $k$  and any optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  is also an optimal solution of  $(\mathcal{T}, S_0)$ .

Now consider the subset  $C_{\text{opt}} \subseteq C$  of all pairs  $(x, y)$  such that  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  has a solution of weight at most  $k$  and any optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$  is an optimal solution of  $(\mathcal{T}, S_0)$ . Order the pairs in  $C_{\text{opt}}$  so that the pairs in  $C_{\text{opt}} \setminus R$  precede the pairs in  $C_{\text{opt}} \cap R$ , and the pairs in  $C_{\text{opt}} \setminus R$  are arranged in the order in which  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  makes the corresponding recursive calls  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, R'')$ . If for a pair  $(x, y) \in C_{\text{opt}}$ ,  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  makes the recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, R'')$  and this recursive call returns an optimal solution  $S \circ \langle(x, y)\rangle \circ S''$  of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$ , then  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  returns a solution  $S_0 \circ S'$  of  $(\mathcal{T}, S_0)$  that is no longer than  $S \circ \langle(x, y)\rangle \circ S''$ . By the choice of  $C_{\text{opt}}$ ,  $S_0 \circ S'$  is thus an optimal solution of  $(\mathcal{T}, S_0)$ . Since we assume that  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  does not return an optimal solution of  $(\mathcal{T}, S_0)$ , it follows that for each pair  $(x, y) \in C_{\text{opt}}$ , either  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  does not make the recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R'')$  (that is,  $(x, y) \in C_{\text{opt}} \cap R$ ) or it makes this recursive call (that is,  $(x, y) \in C_{\text{opt}} \setminus R$ ) but the recursive call returns NONE or a suboptimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$ .

Now let  $(x, y)$  be the first pair in  $C_{\text{opt}}$  according to the ordering defined above.

- If  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  does not make the recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R'')$ , then  $(x, y) \in R$ . Thus, by Corollary 24, there exists a proper prefix  $S_p \subset S_0$  such that  $S_p \circ \langle(x, y)\rangle$  dominates  $S \circ \langle(x, y)\rangle$ . Since  $S \circ \langle(x, y)\rangle$  is an extension of  $S_0$ , this is a contradiction.
- If  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  does make the recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R'')$ , then  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R'')$  does not return an optimal solution of  $(\mathcal{T}, S \circ \langle(x, y)\rangle)$ . Thus, since  $|\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R'')| < |\text{TCS2}(\mathcal{T}, S_0, k, R_0)|$ , the choice of  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  implies that there exist an extension  $S \circ \langle(x, y)\rangle \circ S'$  of  $S \circ \langle(x, y)\rangle$ , a prefix  $S_p \subseteq S$ , and a pair  $(x', y') \in R''$  such that  $S_p \circ \langle(x', y')\rangle$  dominates  $S \circ \langle(x, y)\rangle \circ S'$ . Now we distinguish two cases.
  - If  $(x', y') \in R$ , we prove that there exists a proper prefix  $S'_p \subset S_0$  such that  $S'_p \circ \langle(x', y')\rangle$  dominates  $S \circ \langle(x, y)\rangle \circ S'$ . Since  $S_0 \subseteq S \circ \langle(x, y)\rangle \circ S'$  and  $R \subseteq R_0$ , this implies that  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  does not violate the lemma, a contradiction. If  $S_p \subset S_0$ , we can set  $S'_p = S_p$ . So assume that  $S_0 \subseteq S_p \subseteq S$ . Since  $(x', y') \in R$ , Corollary 24 shows that there exists a proper prefix  $S'_p \subset S_0$  such that  $S'_p \circ \langle(x', y')\rangle$  dominates  $S \circ \langle(x', y')\rangle$ . By Lemma 20,  $S'_p \circ \langle(x', y')\rangle$  also dominates  $S \circ \langle(x, y)\rangle \circ S'$  (where  $(x, y)$  in Lemma 20 corresponds to  $(x', y')$ , and  $S, S \circ S_1, S \circ S_1 \circ S_2, S \circ S_1 \circ S_2 \circ S_3 \circ \langle(x, y)\rangle \circ S''$  correspond to  $S'_p, S_0, S_p, S \circ \langle(x, y)\rangle \circ S'$  respectively).
  - If  $(x', y') \notin R$ , then  $(x', y') \in R' \setminus R$ , which implies that  $(x', y') \in C \setminus R$  and, therefore,  $\text{TCS2}(\mathcal{T}, S_0, k, R_0)$  makes a recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x', y')\rangle, k, R'')$  before the recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle(x, y)\rangle, k, R'')$ . Since  $S \circ \langle(x', y')\rangle$  dominates  $S \circ \langle(x, y)\rangle \circ S'$ , Proposition 21 shows that there exists a solution  $S \circ \langle(x', y')\rangle \circ S''$  of  $(\mathcal{T}, S \circ \langle(x', y')\rangle)$  that satisfies  $w(S \circ \langle(x', y')\rangle \circ S'' < w(S \circ \langle(x, y)\rangle \circ S'))$ .

$S'' \leq w(S \circ \langle(x, y)\rangle \circ S')$ . Since  $S \circ \langle(x, y)\rangle \circ S'$  is an optimal solution of  $(\mathcal{T}, S_0)$ , this implies that  $S \circ \langle(x', y')\rangle \circ S''$  is also an optimal solution of  $(\mathcal{T}, S_0)$ . Thus,  $(x', y') \in C_{\text{opt}}$ , a contradiction because  $(x, y)$  is the first pair in  $C_{\text{opt}}$ .

□

**Corollary 26** *The invocation  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$  returns a shortest tree-child cherry picking sequence for  $\mathcal{T}$  if there exists such a sequence of weight at most  $k$ . Otherwise,  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$  returns NONE.*

**Proof** If there is no tree-child cherry picking sequence for  $\mathcal{T}$  of weight at most  $k$ , then Proposition 17 shows that the invocation  $\text{TCS}(\mathcal{T}, \langle\rangle, k)$  returns NONE. Since each invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  is easily seen to return a sequence only if  $\text{TCS}(\mathcal{T}, S, k)$  returns a sequence, this implies that  $\text{TCS}(\mathcal{T}, \langle\rangle, k, \emptyset)$  returns NONE if there is no tree-child cherry picking sequence of weight at most  $k$ .

So assume that there exists a tree-child cherry picking sequence for  $\mathcal{T}$  of weight at most  $k$ . If  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$  does not return a shortest tree-child cherry picking sequence for  $\mathcal{T}$ , then Lemma 25 states that there exists an extension  $S$  of  $\langle\rangle$ , proper prefix  $S_p \subseteq \langle\rangle$ , and a pair  $(x, y) \in \emptyset$  such that  $S_p \circ \langle(x, y)\rangle$  dominates  $S$ . However, neither  $S_p$  nor the pair  $(x, y)$  can exist. Thus,  $\text{TCS2}(\mathcal{T}, \langle\rangle, k, \emptyset)$  returns a shortest tree-child cherry picking sequence for  $\mathcal{T}$ . □

As already observed in the proof of Lemma 25, each invocation  $\text{TCS2}(\mathcal{T}, S, k, R)$  makes at most as many recursive calls as its corresponding invocation  $\text{TCS}(\mathcal{T}, S, k)$ , so the total number of recursive calls made by the algorithm is still bounded by  $O((8k)^k)$ . Using standard techniques, including binary search trees and integer sorting, and a careful implementation of lines 1–6 that avoids calling UpdateR in each iteration, it is possible to show that the cost per recursive call remains  $O(nt \lg t)$ , including the cost to query and maintain  $R$ . Thus, the worst-case running time of the algorithm remains  $O((8k)^k nt \lg t + nt \lg nt)$ . Since we are interested in using redundant branch elimination mainly as a heuristic improvement of the running time of the algorithm in practice, we do not prove this here. Note that redundant branch elimination is a heuristic only as far as improving the running time is concerned; Corollary 26 above shows that it preserves the algorithm’s correctness.

## 5 Implementation and Experiments

In order to evaluate the usefulness of the algorithm presented in this paper, we implemented it and ran experiments on synthetic and realistic inputs to answer the following questions:

- How difficult inputs can our algorithm handle, both in terms of the number of reticulations in the computed network and the number of trees in the input?
- How does the running time of our algorithm compare to that of its closest competitor, HYBROSCALE?

The answer to this second question is that, for inputs with at least 3 trees, our algorithm ran significantly faster than HYBROSCALE. Since HYBROSCALE computes optimal



hybridization networks, without any restrictions on their structure, while our algorithm computes optimal *tree-child* networks, we effectively buy this faster running time at the price of restricting the types of outputs we can compute and, consequently, possibly missing some optimal networks that are not tree-child. This raises the following natural question:

- For inputs for which both our algorithm and HYBROSCALE were able to compute a network, by how much did the reticulation numbers of the computed networks differ?

The discussion of our experimental results is divided into the following subsections: Sect. 5.1 discusses the hardware and software environment on which we ran our experiments, as well as some high-level characteristics of our implementation. The complete source code, test data, and the programs we used to prepare the test data are available from [https://github.com/nzeh/tree\\_child\\_code](https://github.com/nzeh/tree_child_code), including detailed documentation. Section 5.2 describes the data sets used in our experiments. Section 5.3 briefly discusses the tuning parameters of our implementation used throughout our experiments. Section 5.4 discusses our experimental results.

## 5.1 Evaluation Environment and Some Implementation Details

Our evaluation platform was a Linux system with a quad-core Intel Xeon W3570 running at 1.7GHz and 24GB of DDR3 RAM clocked at 1333MHz. The operating system was Debian GNU/Linux 9 with a 4.19.46-64 Linux kernel. Our code for computing a tree-child network was implemented in Rust version 1.27.0. HYBROSCALE was implemented in Java, and we used Java version 1.8.0\_161 to run it.

Our code implements procedure TCS2, that is, it uses redundant branch optimization. It also uses a number of additional optimizations:

**Check for redundant pairs using occurrence counts:** The check for redundant pairs (pairs in  $R$ ) was implemented by recording two counts  $c_{(x,y)}$  and  $c_{(y,x)}$  for each cherry  $\{x, y\}$  of  $\mathcal{T}/S$ ;  $c_{(x,y)}$  is the number of trees that contained the cherry  $\{x, y\}$  the last time an ancestor invocation made a recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle (x, y) \rangle, k, R)$ ;  $c_{(y,x)}$  is the number of trees that contained this cherry the last time an ancestor invocation made a recursive call  $\text{TCS2}(\mathcal{T}, S \circ \langle (y, x) \rangle, k, R)$ . Whenever we append a pair  $(v, w)$  to  $S$ , we set  $c_{(x,y)} = 0$  if  $w = x$ , and  $c_{(y,x)} = 0$  if  $w = y$ . It is easy to verify that this ensures that  $(x, y)$  is redundant for the current sequence  $S$  if and only if the number of trees in  $\mathcal{T}/S$  that contain the cherry  $\{x, y\}$  equals  $c_{(x,y)}$ . Similarly,  $(y, x)$  is redundant for  $S$  if and only if the number of trees in  $\mathcal{T}/S$  that contain the cherry  $\{x, y\}$  equals  $c_{(y,x)}$ .

**No copying of an invocation's state for each recursive call:** The state of each invocation (current set of trees, set of trivial cherries, set of non-trivial cherries, partial tree-child cherry picking sequence, and information about the cherries and trees containing each leaf) is fairly large. To avoid the overhead of copying this state for each recursive call, each recursive call instead modifies its parent invocation's state without making a copy. These modifications are recorded

in a log and are undone when the recursive call returns, thereby restoring the parent invocation's state.

**Search for the optimal  $k$ :** The search for an optimal tree-child cherry picking sequence calls the procedure  $\text{TCS2}(\mathcal{T}, \langle \rangle, k, \emptyset)$  with increasing values of  $k$  until it reports success. This guarantees that the parameter  $k$  is no larger than the tree-child hybridization number of each input.

**Parallelization:** The different branches of the recursive search for an optimal tree-child cherry picking sequence are clearly independent and can thus be assigned to different threads of a parallel implementation of procedure  $\text{TCS2}$ . One challenge was that, especially in the presence of redundant branch elimination, the computational costs of different branches can differ substantially.

To balance the load between threads, we implemented a work sharing scheduler that allows idle threads to send messages to busy threads to request part of their workload. In response to such a request, the busy thread sends a branch on its recursion stack that is yet to be explored to the requesting thread. In the interest of minimizing the number of messages exchanged between threads, the busy thread always shares the next branch from the *bottom* of its recursion stack, hopefully corresponding to a large subtree in the algorithm's recursion.

The communication protocol was implemented using light-weight spinlocks to minimize the amount of time busy threads spend on communicating with other threads.

**Cluster reduction** Cluster reduction [4,13] has been observed to be the most important optimization in phylogenetic network construction methods for pairs of trees [12]. While we expect cluster reduction to be less effective for more than two trees, our implementation still applies cluster reduction because it is relatively cheap and should still have a significant impact on the algorithm's running time for real-world inputs.

In order to complete all our experiments in a reasonable amount of time, we limited every run of our algorithm or of  $\text{HYBROSCALE}$  to 60 minutes. If the algorithm did not produce a result within this time limit, we consider this input to be unsolvable by the algorithm in the context of this evaluation.

## 5.2 Test Data

We used synthetic and real-world data for the performance evaluation of our algorithm.

### 5.2.1 Synthetic Data

To generate a test instance with  $t$  trees over a set of  $n$  leaves and with tree-child hybridization number close to  $k$ , we generated a random tree-child network  $N$  on  $n$  leaves and with  $k$  reticulations. Then we extracted a random set of  $t$  trees displayed by  $N$ .

*Network generation* To generate the network  $N$ , we initialized  $N$  to be a tree with two leaves. A network with  $n$  leaves and  $k$  reticulations could then be obtained by adding  $s_r = n + k - 2$  tree nodes and  $k_r = k$  reticulations to  $N$ . The total number of non-leaf

nodes to be added was  $s_r + k_r$ . Thus, as long as  $s_r > 0$  and  $k_r > 0$ , we added either a tree node or a reticulation.

To add a tree node, we chose an existing leaf  $u$  and added two new leaves  $v$  and  $w$  with parent  $u$ . This turned  $u$  into a tree node while not affecting any existing reticulations or tree nodes. Thus,  $s_r$  decreased by one while  $k_r$  remained unchanged.

To add a reticulation, we chose two leaves  $u$  and  $v$ ; merged  $v$  into  $u$ , making  $u$  and  $v$  the same node; and then added a new leaf  $w$  with parent  $u$ . This turned  $u$  into a reticulation while not affecting any existing reticulations or tree nodes. Thus,  $k_r$  decreased by one while  $s_r$  remained unchanged.

In order to ensure that the network was tree-child, the two nodes  $u$  and  $v$  to be merged were chosen from the set  $M$  of all nodes whose parents and siblings were tree nodes or leaves. We also ensured that the network had no parallel edges by picking  $u$  and  $v$  so that they had different parents. Thus, if  $|M| = 1$  or  $|M| = 2$  and the two nodes in  $M$  had the same parent, then there were no two nodes  $u$  and  $v$  that could be added while keeping the network tree-child and not introducing any parallel edges. In this case, we added a new tree node. If it was possible to add a reticulation node, then we added a tree node with probability  $\frac{s_r}{s_r + k_r}$  and a reticulation with probability  $\frac{k_r}{s_r + k_r}$ .

If we added a tree node, we chose the leaf  $u$  to be turned into a tree node uniformly at random from the current set of leaves.

If we added a reticulation, we chose  $u$  and  $v$  uniformly at random from the set  $M$ . If the two chosen nodes  $u$  and  $v$  had the same parent, we repeated this selection process until they did not.

This random addition of tree nodes and reticulations continued until  $s_r = 0$  or  $k_r = 0$ . If  $k_r = 0$  and  $s_r > 0$ , we kept adding tree nodes using the procedure above until  $s_r = 0$ . If  $s_r = 0$  and  $k_r > 0$ , we kept adding reticulations using the procedure above until either  $k_r = 0$  or it was impossible to add more reticulations because either  $|M| = 1$  or  $|M| = 2$  and the two leaves in  $M$  had the same parent.

*Tree generation* We selected  $t$  (or fewer) trees displayed by  $N$  by repeating the following process: We deleted one of the parent edges of each reticulation in  $N$  uniformly at random and suppressed every node with only one child in the resulting tree. If the newly generated tree already existed within the list of trees (with the same Newick representation), then we did not add it to the list. We maintained a count of the number of times this occurred. Once this count reached 100 or we had  $t$  trees in our list, we terminated the process and returned the trees.

Note that the set of trees generated using this process was not guaranteed to have tree-child hybridization number  $k$ . First, the network generation did not guarantee that we obtained a network with  $k$  reticulations if we stopped the network generation with a value of  $k_r > 0$  and without any pairs of leaves that could still be merged. Second, even if  $N$  did have  $k$  reticulations, there may exist a tree-child network with fewer than  $k$  reticulations that also displays the obtained set of trees.

### 5.2.2 Real-World Data

The real-world data we used in our experiments was derived from a collection of gene trees for 159,905 distinct homologous gene sets found in a set of 1,173 bacterial and archaeal genomes. These gene trees were constructed by Beiko and are described in

more detail in [5]. They were also used as a test data set, for example, in the evaluation of a method for constructing SPR supertrees [20]. Beiko's data set (as almost every real-world data set) poses two challenges for our algorithm. First, bipartitions with low support in this data set were collapsed, so the input trees are multifurcating. Second, since not all genes are present in all taxa, the label sets of the input trees differ.

To obtain a collection of binary trees over the same label set, we used a two-step process: First, given the desired number of leaves  $n$  as a parameter, we selected a subset of  $n$  taxa  $X$  and all trees that contained all of these taxa. Then we restricted the selected trees to the chosen label set  $X$ , thereby obtaining a collection of multifurcating trees over this set of  $n$  taxa. Second, we resolved multifurcations in these trees to obtain a collection of binary trees. If we had resolved multifurcations randomly, it would have been very likely that any network displaying the constructed trees contains many reticulations that result only from inconsistent resolutions of the input trees. To avoid this, we introduced inconsistent resolutions into different input trees only if the input trees forced us to do so. This procedure is described in more detail below and at [https://github.com/nzeh/tree\\_child\\_code](https://github.com/nzeh/tree_child_code).

We did not evaluate whether the resulting trees are biologically plausible (beyond the degree to which every binary resolution of a well supported multifurcating tree is plausible). Our only goal was to construct a test data set whose characteristics, in terms of number of reticulations and existence of clusters that allow the input to be decomposed into easier inputs, resemble those of typical real-world inputs, in order to evaluate the usefulness of our algorithm to construct phylogenetic networks for non-trivial real-world inputs.

*Selection of leaf set and trees* To extract as many trees with a given number of common leaves  $n$ , we used the following strategy: we started with an empty set of leaves  $X = \emptyset$  and the entire set of 159,905 input trees  $\mathcal{T}$ . Then we repeated the following process  $n$  times: Let  $Y$  be the set of all unique taxa of the trees in  $\mathcal{T}$  and let  $x \in Y \setminus X$  be a taxon that occurs in the maximum number of trees in  $\mathcal{T}$ . Then we added  $x$  to  $X$  and discarded all trees from  $\mathcal{T}$  that did not contain  $x$ . At the end of this iterative process, we obtained a set of trees  $\mathcal{T}$  that contained all taxa in  $X$ . As already mentioned, the next step was to restrict every tree in  $\mathcal{T}$  to the label set  $X$ .

*Binary resolution* Binary resolutions were obtained by repeating the following process until all trees were binary: Inspect the trees in  $\mathcal{T}$  in an arbitrary order. For each tree, inspect its multifurcations in an arbitrary order. For each multifurcation  $u$ , consider all pairs  $\{v, w\}$  such that  $v$  and  $w$  are children of  $u$ . For each such pair, count the number of resolved triplets (triplets of the form  $ab|c$  as opposed to  $a|b|c$ ) that would be introduced by resolving  $\{v, w\}$  (that is, by making  $v$  and  $w$  children of a new node  $u'$  and making  $u'$  a child of  $u$ ) and which are also present in at least one other tree in  $\mathcal{T}$ .

If there exists such a pair  $\{v, w\}$  that introduces at least one introduced resolved triplet that exists also in some other tree in  $\mathcal{T}$ , then resolve the pair that maximizes the number of introduced resolved triplets that exist in other trees. If no such pair is found, then move on to the next multifurcation in the current tree or to the next tree if there are no more multifurcations left to inspect in the current tree.

If the above steps resolve at least one multifurcation, then start another iteration. Otherwise, pick an arbitrary multifurcation in one of the trees and a random pair of children of this multifurcation and resolve it. Then start another iteration. (This random resolution will be matched by all other trees in the next iteration, thus forcing consistency between the trees.)

*Test instances* By running the above procedure with parameter  $n \in \{10, 20, 30, 40, 50, 60, 80, 100, 150\}$ , we generated tree sets with this number of leaves and with between 21 and 1,684 trees for  $n = 150$  and  $n = 20$ , respectively. To obtain an input with a given number of leaves  $n$  and a given number of trees  $t$ , we selected  $t$  of the trees with  $n$  leaves uniformly at random.

### 5.3 Parameter Tuning

Our implementation of procedure TCS2 accepts a number of command-line arguments, mainly to facilitate the type of performance evaluation we conducted. The most important options are turning cluster reduction on or off, turning redundant branch elimination on or off, configuring the number of threads across which to distribute the algorithm's work, and controlling how frequently busy threads check for work requests from idle threads. More threads allow the operating system to help with load balancing but too many threads result in scheduling overhead. Similarly, frequent checks for work requests from idle threads help with load balancing by ensuring that idle threads never remain idle for too long but increase the overhead that slows down busy threads.

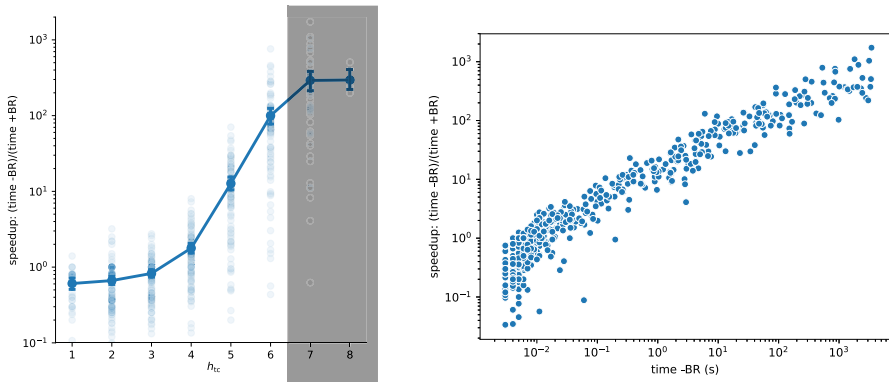
In preliminary experiments, we determined that we obtained the best performance using eight threads (`-p 8`) on our system. The frequency of checks for work requests had negligible impact on the algorithm's performance as long as idle threads did not wait for work for too long. Throughout the experiments discussed here, we made a busy thread check for work requests from idle threads every 100 iterations through its main loop (`-w 100`). Cluster reduction never hurt performance but helped substantially on most real-world inputs, so we never turned it off. Since redundant branch elimination is a potentially important optimization of our algorithm discussed in Sect. 4, we dedicate a separate section to discussing its impact on the algorithm's performance.

### 5.4 Results

#### 5.4.1 Does Redundant Branch Elimination Help?

Our first experiments concerned whether redundant branch elimination helps to reduce the running time of the algorithm in practice. To evaluate this, we ran the algorithm with redundant branch elimination on a synthetic data set. For the runs with redundant branch elimination, we used three test inputs for every possible combination of the following parameters:

- Number of trees:  $t \in \{2, 5, 10, 15, 20, 50, 100\}$
- Number of reticulations in the original network:  $k \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$



**Fig. 5** The speed-up (running time without redundant branch elimination (–BR) divided by the running time with redundant branch elimination (+BR)) achieved by redundant branch elimination on 658 instances solvable with and without redundant branch elimination. **a** as a function of the number of reticulations and **b** as a function of the running time without redundant branch elimination. The shading of reticulation numbers 7 and 8 indicate that not all inputs with 7 or 8 reticulations were solved by the algorithm, so particularly the flattening of the curve may be the result of limiting the running time of the algorithm and testing only a restricted set of inputs. We would expect that the effect of redundant branch elimination keeps increasing as the number of reticulations increases, given that there seems to be no plateauing of the speed-up as a function of running time in (**b**)

- Number of leaves:  $n \in \{20, 50, 100, 150, 200\}$

resulting in a set of 1155 inputs. The algorithm was able to solve 1016 of these inputs within the 1-h time limit. Without redundant branch elimination, the algorithm was not able to solve any synthetic inputs with  $k > 8$  within the time limit. Of the 735 inputs with  $k \leq 8$ , it was able to solve 658 inputs within the time limit.

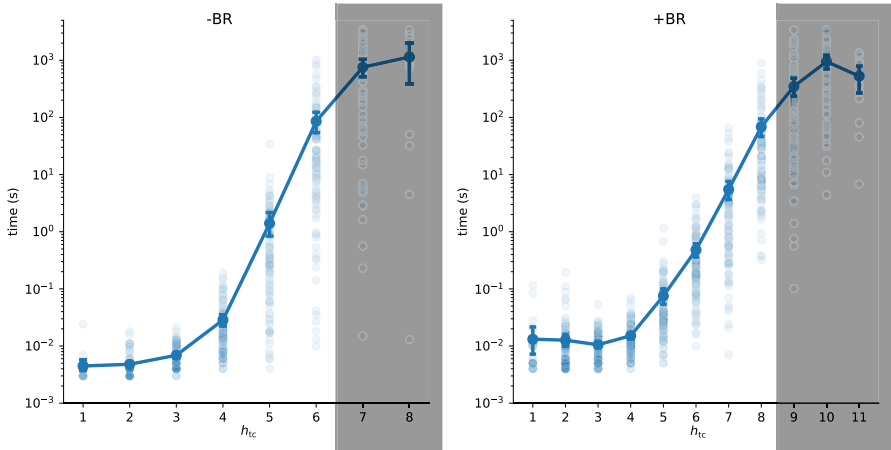
Figure 5 shows the speed-up achieved by using redundant branch elimination on the 658 inputs the algorithm was able to solve without it. As can be seen, the effect of redundant branch elimination increased with increasing reticulation number and, correspondingly, with increasing running time of the algorithm, reaching a speed-up of up to 1000 on some instances with six and seven reticulations.

Figure 6 shows that redundant branch elimination increased the difficulty of inputs our algorithm was able to solve within the 1-h time limit. Without branch reduction, the algorithm was able to solve all instances with reticulation numbers up to six and some instances with up to eight reticulations. With redundant branch reduction, the algorithm was able to solve all instances with reticulation numbers up to eight and some instances with up to 11 reticulations.

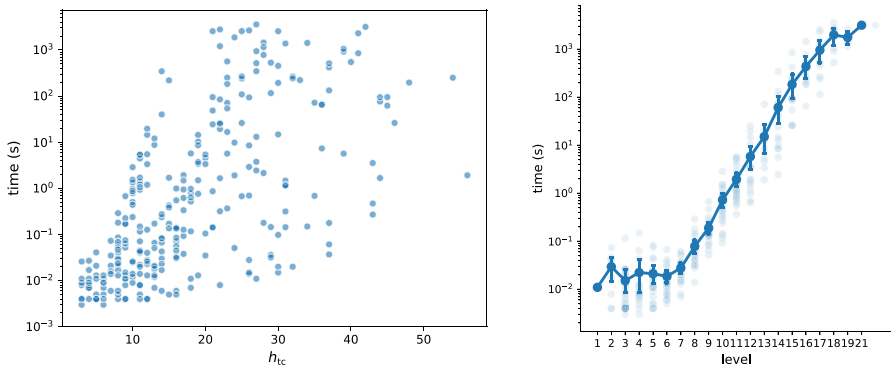
### 5.4.2 Real-World Inputs That Can Be Solved

Our next experiment tested whether we can solve real-world instances with non-trivial numbers of reticulations efficiently using our algorithm. For this experiment, we extracted ten test instances from the real-world data set for every possible combination of the following parameters:

- Number of trees:  $t \in \{2, 3, 4, 5, 6, 7, 8\}$



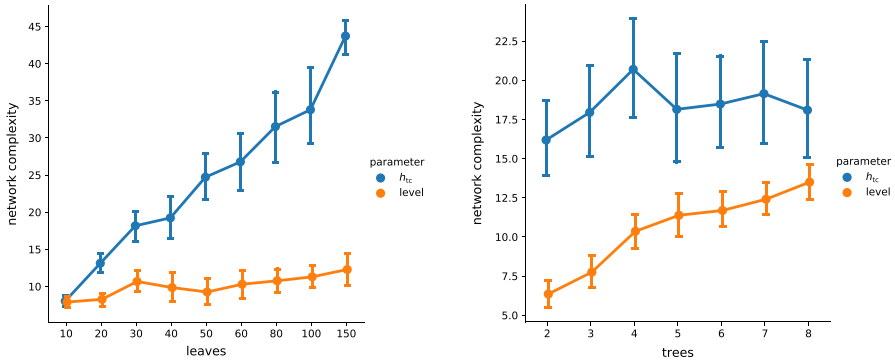
**Fig. 6** Running times of our algorithm with and without redundant branch elimination, as functions of the number of reticulations. As in Fig. 5, the shaded regions indicate reticulation numbers for which not all input instances were solved within the 1-h time limit. Transparent dots are data points, opaque dots indicate the average together with the 95% confidence intervals



**Fig. 7** Running times of our algorithm on real-world data as a function of the reticulation number (left) or the level (right)

- Number of leaves:  $n \in \{10, 20, 30, 40, 50, 60, 80, 100, 150\}$

The algorithm was run with redundant branch elimination and cluster reduction. Of the 630 test inputs, our algorithm was able to solve 306 within the 1-hour time limit. The left graph in Fig. 7 shows the running time of our algorithm on the instances it was able to solve as a function of the number of reticulations. We make two important observations: First, even though our algorithm was not able to solve any synthetic inputs with more than 11 reticulation even with redundant branch elimination turned on, it was able to solve real-world inputs with up to 50 reticulations. Second, the running time varied greatly across instances with the same number of reticulations. Both observations can be explained by the fact that the real-world data has much more structure and can be decomposed into non-trivial clusters. The time needed to find a (tree-child) hybridization network for such an input then depends primarily on



**Fig. 8** The reticulation number and the level as a function of the number of leaves and trees in the real-world inputs

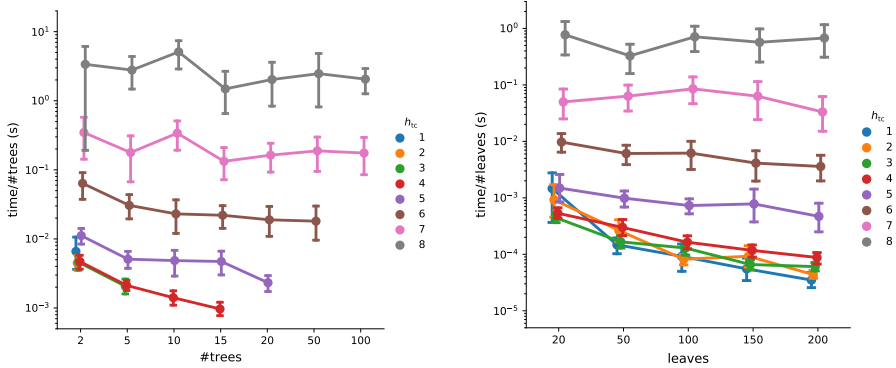
the maximum (tree-child) hybridization number of these clusters (as opposed to the hybridization number of the whole input, which can be much higher). This maximum (tree-child) hybridization number of all clusters is commonly referred to as the (tree-child) *level* of the input. Figure 8 shows the number of reticulations and the level of the real-world inputs as a function of the number of trees. These figures demonstrate that the network levels were significantly lower than the number of reticulations, something that had also been observed for inputs consisting of two trees and which is the key to the fast running times of MAAF-based algorithms for pairs of trees. It came as a bit of a surprise that the same was true also for more than two trees. However, the right graph in Fig. 8 demonstrates that the gap between level and reticulation number narrowed as the number of trees increased.

Using cluster reduction, the running time of the algorithm is determined by the level of the computed network rather than the reticulation number. Thus, the right graph in Fig. 7 shows the running time as a function of the level of the computed network. This figure highlights another important fact: We were able to solve real-world instances with level up to 21 whereas level 11 was the limit for synthetic inputs. This suggests that even the clusters seemed to have significantly more structure than random instances, which allowed the algorithm to branch on fewer non-trivial cherries in each recursive call than on synthetic instances.

#### 5.4.3 Dependence of the Running Time on the Number of Trees and Number of Leaves

The theoretical analysis of our algorithm predicts an exponential dependence of its running time only on the number of reticulations  $h_{tc}$ , whereas the running time should depend only nearly linearly on both  $n$  and  $t$ . To verify this, we divided the observed running times, for each value of  $h_{tc}$  between 1 and 8, by  $n$  and then by  $t$ . Figure 9 shows the results. The negative slopes of these curves confirms that the running time in practice depends at most linearly on each of  $n$  and  $t$ .





**Fig. 9** Running times of the algorithm with redundant branch elimination on all synthetic test inputs divided by the number of trees (left) and the number of leaves (right). Error bars denote a 95% confidence interval

### 5.4.4 Comparison with HYBROSCALE

The most interesting question is whether optimal tree-child networks can be computed significantly faster than unrestricted hybridization networks. To answer this question, we compared the running time of our algorithm against that of its closest competitor HYBROSCALE, which computes unrestricted hybridization networks. For this comparison, we used synthetic data and real-world data. In order to test a wide range of test inputs, we limited the time per run to 20 min for synthetic inputs and to 60 min for real-world inputs. Since we ran our algorithm with eight threads, we did the same for HYBROSCALE.

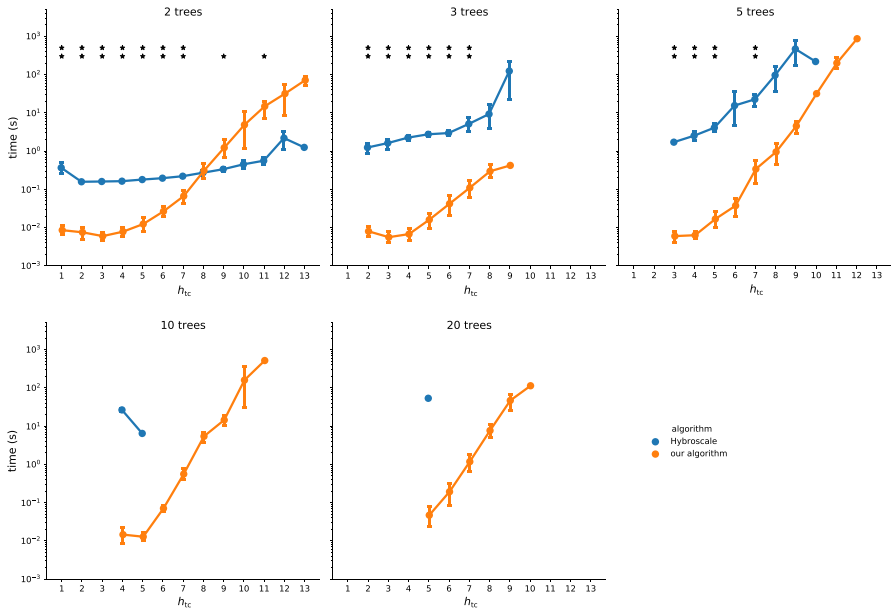
*Synthetic data* We tested both our algorithm and HYBROSCALE on six test inputs for every possible combination of the following parameters:

- Number of trees:  $t \in \{3, 5, 10, 20\}$
- number of reticulations in the original network:  $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

and on six inputs with 2 trees and  $k \in \{2, 4, \dots, 28, 30\}$ . All instances had 20 leaves. We used a wider range of reticulation numbers (and compensated for this by using only three instances for each value of  $k$ ) for inputs with only two trees because we expected HYBROSCALE to run very fast on such inputs (because MAAF-based algorithms are very fast for pairs of trees).

As can be seen in Fig. 10 and as expected, HYBROSCALE outperformed our algorithm on inputs consisting of two trees and for more than seven reticulations. For more than two trees, our algorithm ran faster than HYBROSCALE due to the near-linear dependence of our algorithm on the number of trees and the exponential dependence of HYBROSCALE on the number of trees. The difference became very pronounced for 10 and 20 trees, where HYBROSCALE was unable to solve most instances whereas our algorithm solved all test instances within the 20-min time limit. Additionally, HYBROSCALE ran out of memory on certain occasions.

*Real-world data* For this experiment, we used the same data set as in Sect. 5.4.2. As mentioned before, our algorithm solved 306 of the 630 inputs in the 1-hour time limit;



**Fig. 10** Running times of our Algorithm and HYBROSCALE on synthetic inputs. Since our algorithm solved all test instances and HYBROSCALE did not, we chose the tree-child hybridization number as the x-axis. Bars indicate a 95% confidence interval. Stars indicate significant differences between the running times of the two algorithms using an independent *t*-test with unequal variances (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ )

HYBROSCALE solved 152 inputs, which were a subset of the 306 inputs solved by our algorithm. On 5 of the 2-tree inputs, HYBROSCALE outperformed our algorithm. On all other inputs, including all other 2-tree inputs, our algorithm was faster. Figure 11 shows the detailed results.

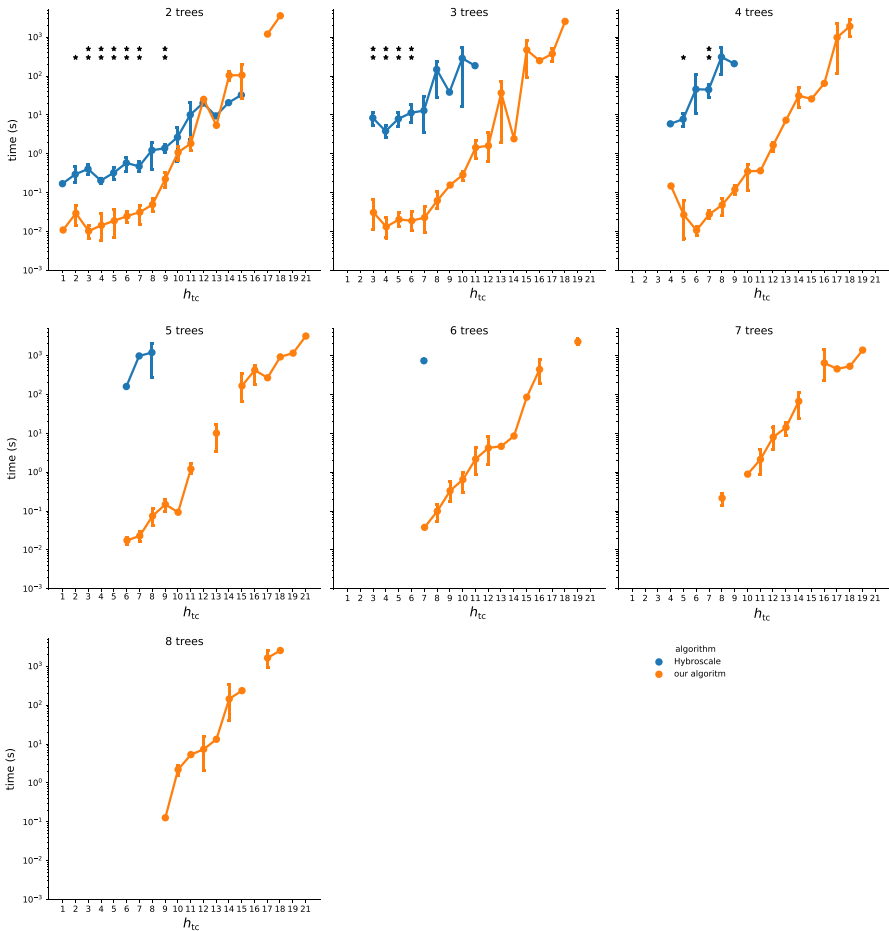
### 5.4.5 Hybridization Versus Tree-Child Hybridization

The final question we were interested in was whether optimal tree-child hybridization networks have significantly more reticulations than the optimal unrestricted hybridization networks for the same sets of trees or whether tree-child hybridization networks are often also optimal hybridization networks.

Of the 268 synthetic inputs that both our algorithm and HYBROSCALE were able to solve, only 3 had a greater tree-child hybridization number than their hybridization number. For all three inputs, the difference was 1.

Of the 142 real-world inputs solved by both our algorithm and HYBROSCALE, 21 had a greater tree-child hybridization number than their hybridization number. For 20 of these inputs, the difference was 1; for 1 input, the difference was 2.

This indicates that very often, tree-child hybridization networks achieve the optimal hybridization number and, even when they do not, they offer a reasonable approximation of optimal hybridization networks. Given that they are substantially easier to compute, as our results in the previous subsection demonstrate, tree-child networks



**Fig. 11** Running times of our algorithm and HYBROSCALE on real-world inputs. Since our algorithm solved all test instances that HYBROSCALE was able to solve, we chose the tree-child level as the x-axis. Bars indicate a 95% confidence interval. Stars indicate significant differences between the running times of the two algorithms using an independent *t*-test with unequal variances (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ )

therefore offer a useful analysis tool that can be used in place of hybridization networks in many instances.

## 6 Conclusion

We have presented the first fixed-parameter algorithm for computing optimal tree-child networks for many binary trees on the same label set, based on the recently introduced concept of tree-child cherry picking sequences. While the theoretical running time of our algorithm is substantially greater than that of MAAF-based network construction methods for two trees, our experimental results confirm that our algorithm can be used

to solve non-trivial real-world inputs efficiently. Similarly to MAAF-based algorithms for two trees, a key factor determining whether an instance can be solved efficiently is whether it can be decomposed into non-trivial clusters. While it comes as no surprise that randomly generated inputs consisting of more than two trees (almost) cannot be decomposed into clusters and thus cannot be solved efficiently, except for fairly small numbers of reticulations, the real-world inputs in our experiments contained sufficiently many non-trivial clusters, which allowed us to solve some inputs with up to 50 reticulations within one hour or less.

The closest competitor of our algorithm, HYBROSCALE, which computes unrestricted hybridization networks, outperformed our algorithm on inputs consisting of two trees, which was to be expected because MAAF-based methods are very efficient for computing optimal hybridization networks for pairs of trees. Already for three trees, our algorithm outperformed HYBROSCALE and, for more than six trees, HYBROSCALE was not able to solve any of the inputs our algorithm was able to solve, due to its exponential dependence on the number of trees.

While our results are promising, they should only be considered to be a first important step towards efficient algorithms for computing (tree-child) hybridization networks from many input trees. Here are two natural and important open questions to be addressed by future work:

Can tree-child hybridization networks be computed faster than in  $O((ck)^k \cdot \text{poly}(n, t))$  time, ideally in  $O(c^k \cdot \text{poly}(n, t))$  time? For temporal networks, a recent result [8] shows that this is indeed the case. An interesting open question is whether the techniques used in that algorithm can also be used to obtain faster algorithms for computing general tree-child networks.

Most real-world inputs are multifurcating, as a result of suppressing branches in gene trees with low support. Thus, it would be of great importance to obtain efficient methods for constructing (tree-child) hybridization networks from multifurcating trees. Our algorithm is able to do this but only if we sacrifice the FPT bound on its running time: the bound on the number of non-trivial cherries in Proposition 9, which is the key to bounding the branching number of our algorithm, holds only if the input trees are binary. It remains an open question whether there exists a fixed-parameter algorithm for computing optimal tree-child hybridization networks for multifurcating trees.

**Acknowledgements** We would like to thank the anonymous reviewers for valuable comments that helped us improve this manuscript.

## A Construction of a Tree-Child Network from a Tree-Child Cherry Picking Sequence

---

### Procedure TreeChildNetworkFromSequence( $\mathcal{T}$ , $S$ )

---

**Input:** A set of  $X$ -trees  $\mathcal{T}$  and a tree-child cherry picking sequence

$S = ((x_1, y_1), \dots, (x_r, y_r), (x_{r+1}, -))$  for  $\mathcal{T}$

**Output:** A tree-child phylogenetic network  $N$  on  $X$  that displays  $\mathcal{T}$  and with reticulation number at most  $w(S)$

```

1 if  $|X| = 1$  then
2   return the unique network consisting of a single node labelled with the element of  $X$ ;
3 else
4    $N \leftarrow$  the directed graph with nodes  $\rho$  and  $x_{r+1}$  and a single edge  $\rho x_{r+1}$ ;
5   for  $j \leftarrow r$  downto 1 do
6     Split the parent edge of  $y_j$  in  $N$  by adding a node  $p$ ;
7     if  $x_j$  is a leaf of  $N$  then
8       if  $x_j$ 's parent in  $N$  is a reticulation  $r$  then
9          $q \leftarrow r$ ;
10      else
11        Split the parent edge of  $x_j$  in  $N$  by adding a node  $q$ ;
12      else
13        Add  $x_j$  to  $N$ ;
14         $q \leftarrow x_j$ ;
15      Add the edge  $pq$  to  $N$ ;
16  return  $N$ ;
```

---

## References

- Albrecht, B.: Computing hybridization networks for multiple rooted binary phylogenetic trees by maximum acyclic agreement forests. [arXiv:1408.3044](https://arxiv.org/abs/1408.3044) (2014)
- Albrecht, B.: Computing all hybridization networks for multiple binary phylogenetic input trees. *BMC Bioinf* **16**(1), 236 (2015)
- Baroni, M., Grünewald, S., Moulton, V., Semple, C.: Bounding the number of hybridisation events for a consistent evolutionary history. *J. Math. Biol.* **51**(2), 171–182 (2005)
- Baroni, M., Semple, C., Steel, M.: Hybrids in real time. *Syst. Biol.* **55**, 46–56 (2006)
- Beiko, R.G.: Telling the whole story in a 10,000-genome world. *Biol. Direct* **6**(1), 34 (2011)
- Bordewich, M., Semple, C.: Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **4**(3), 458–466 (2007)
- Bordewich, M., Linz, S., John, K.S., Semple, C.: A reduction algorithm for computing the hybridization number of two trees. *Evol. Bioinf. Online* **3**, 86–98 (2007)
- Borst, S.: New FPT algorithms for finding the temporal hybridization number for sets of phylogenetic trees. Master's thesis, TU Delft, the Netherlands (2020)
- Chen, Z.-Z., Wang, L.: Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **9**(2), 372–384 (2012)
- Humphries, P.J., Linz, S., Semple, C.: Cherry picking: a characterization of the temporal hybridization number for a set of phylogenies. *Bull. Math. Biol.* **75**(10), 1879–1890 (2013)
- Kelk, S.: Treetistic. <http://skelk.sdf-eu.org/clustistic/>, (2012)
- Li, Z., Zeh, N.: Computing maximum agreement forests without cluster partitioning is folly. In: *Proceedings of the 25th Annual European Symposium on Algorithms*, pp. 56:1–56:14 (2017)
- Linz, S., Semple, C.: A cluster reduction for computing the subtree distance between phylogenies. *Ann. Comb.* **15**(3), 465–484 (2011)

14. Linz, S., Semple, C.: Attaching leaves and picking cherries to characterise the hybridisation number for a set of phylogenies. *Adv. Appl. Math.* **105**, 102–129 (2019)
15. Mirzaei, S., Yufeng, W.: Fast construction of near parsimonious hybridization networks for multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **13**(3), 565–570 (2016)
16. van Iersel, L., Linz, S.: A quadratic kernel for computing the hybridization number of multiple trees. *Inf. Process. Lett.* **113**(9), 318–323 (2013)
17. van Iersel, L., Kelk, S., Lekic, N., Whidden, C., Zeh, N.: Hybridization number on three rooted binary trees is EPT. *SIAM J. Discret. Math.* **30**(3), 1607–1631 (2016)
18. van Iersel, L., Kelk, S., Scornavacca, C.: Kernelizations for the hybridization number problem on multiple nonbinary trees. *J. Comput. Syst. Sci.* **82**(6), 1075–1089 (2016)
19. Whidden, C., Beiko, R.G., Zeh, N.: Fixed-parameter algorithms for maximum agreement forests. *SIAM J. Comput.* **42**(4), 1431–1466 (2013)
20. Whidden, C., Zeh, N., Beiko, R.G.: Supertrees based on the subtree prune-and-regraft distance. *Syst. Biol.* **63**(4), 566–581 (2014)
21. Yufeng, W.: Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics* **26**(12), i140–i148 (2010)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.