

Deep kernelization for the Tree Bisection and Reconnect (TBR) distance in phylogenetics

Steven Kelk¹, Simone Linz², Ruben Meuwese¹

¹ Department of Data Science and Knowledge Engineering (DKE),
Maastricht University, The Netherlands,
steven.kelk@maastrichtuniversity.nl, ruben.meuwese@maastrichtuniversity.nl

² School of Computer Science, University of Auckland, New Zealand,
s.linz@auckland.ac.nz

Abstract. We describe a kernel of size $9k - 8$ for the NP-hard problem of computing the Tree Bisection and Reconnect (TBR) distance k between two unrooted binary phylogenetic trees. We achieve this by extending the existing portfolio of reduction rules with three novel new reduction rules. Two of the rules are based on the idea of topologically transforming the trees in a distance-preserving way in order to guarantee execution of earlier reduction rules. The third rule extends the local neighbourhood approach introduced in [15] to more global structures, allowing new situations to be identified when deletion of a leaf definitely reduces the TBR distance by one. The bound on the kernel size is tight up to an additive term. Our results also apply to the equivalent problem of computing a Maximum Agreement Forest (MAF) between two unrooted binary phylogenetic trees. We anticipate that our results will be more widely applicable for computing agreement-forest based dissimilarity measures.

Keywords: phylogenetics, agreement forest, TBR distance, kernelization, fixed parameter tractability.

1 Introduction

A phylogenetic tree is essentially a tree in the usual graph-theoretical sense whose leaves are bijectively labeled by a set of labels X [20]. Such trees have a central role in the study of evolution. The labels X represents a set of contemporary species, the unlabelled interior nodes of the tree represent hypothetical (extinct) ancestors of X and the topology of the tree encodes the history of branching events, such as speciation, which caused those ancestors to diversify into the set of species X . A central challenge in the field of phylogenetics is to accurately infer such trees from data obtained solely from X , such as DNA data [7]. However, it is not uncommon to obtain different trees for the same set X ; this can be methodological (e.g. different objective functions or multiple optima) or due to the fact that some species have multiple distinct tree signals woven into their genome [21]. This motivates the use of distance measures in phylogenetics, which rigorously quantify the dissimilarity of two phylogenetic trees. Such distance measures can communicate important information about the biological significance of the observed differences [25] and can help us to understand the behaviour of tree-construction algorithms that traverse the space of phylogenetic trees by applying local rearrangement operations [13,18]. Distances can also be used as part of the toolkit for constructing non-treelike hypotheses of evolution, known as phylogenetic networks [12].

In this article we are concerned with one such distance, Tree Bisection and Reconnect (TBR) distance, which is a metric on the space of unrooted (i.e. undirected) binary phylogenetic trees (see Figure 1). This distance represents the minimum number of times a subtree

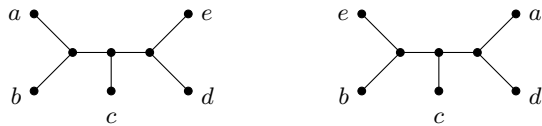


Fig. 1: Two unrooted binary phylogenetic trees on $X = \{a, b, c, d, e\}$.

of one tree has to be detached, and reattached elsewhere, in order to transform it into the other tree (see Figure 2). It is NP-hard to compute [1,11]. The problem has an equivalent, alternative formulation using *agreement forests*. An agreement forest is a partition of X such that the spanning trees induced by the blocks of the partition are disjoint in both trees *and* the induced spanning trees have the same topology in both trees, up to suppression of degree 2 vertices. An agreement forest with a minimum number of blocks is called a *maximum agreement forest* (MAF); it is well-known that the TBR distance (d_{TBR}) is equal to the number of blocks in a MAF (d_{MAF}) minus 1 [1]. In the last decade maximum agreement forests have received sustained attention from the mathematics, computer science and bioinformatics communities, see e.g. [2,4,6,17,19,24]. One response to the NP-hardness of computing d_{TBR} is *kernelization*. Here the goal is to apply polynomial-time preprocessing rules such that d_{TBR} is preserved, or decreased in a controlled fashion, such that the reduced trees have at most $f(d_{\text{TBR}})$ leaves for some function f that depends only on d_{TBR} . For further background on kernelization we refer to the book [10]. The core idea is that, if d_{TBR} is small, then the reduced trees (known as the *kernel*) will be small even if $|X|$ is very large. d_{TBR} can then be computed on these small trees using optimized exponential-time algorithms. The use of kernelization in this context is not coincidental: phylogenetics continues to be a rich source of open problems in, and application opportunities for, parameterized complexity [3].

In 2001 it was shown in [1] that the *subtree* and *chain* reduction rules suffice to obtain a kernel of size at most $28k$, where k is d_{TBR} . These function by reducing common pendant subtrees and common chains (i.e. caterpillar-like regions), respectively. Almost 20 years later the present authors proved that the same reduction rules actually yield a kernel of size at most $15k - 9$, and in fact that this is tight [14]. A critical insight in [14] was that computation of d_{TBR} (or d_{MAF}) can equivalently be viewed as the problem of adding the labels X , and a set of *breakpoints* (essentially: edge cuts), to an (unknown) cubic multigraph, known as a *generator*, such that the original two trees can be retrieved. This insight was subsequently leveraged in [15] to design five new reduction rules which, when added to the subtree and chain reduction rules, yield a tight kernel of size $11k - 9$. An empirical follow-up showed that the new rules in [15] have added reductive power in practice [23], and recently similar techniques have been used to design new reduction rules for distances and agreement forests on *rooted* trees [16].

The natural question is: how far can the $11k - 9$ bound for d_{TBR} be improved? Here we give a kernel of size $9k - 8$, which is tight up to an additive term. We use the analytical and counting bottlenecks identified in [15] as a starting point, and use these to guide the design of three new, novel reduction rules. The reduction rules have a rather different flavour to what has come before. The first new reduction rule addresses the following bottleneck: some of the topological structures that contribute heavily to the $11k - 9$ bound, and which we thus wish to target for reduction, could potentially be leveraged by a depth-bounded branching algorithm that recursively cuts edges in the input trees to obtain an agreement forest. However, the cuts

applied by such a direct, non-preprocessing algorithm yield a different, more general problem, on forests rather than trees, which is analytically far harder to deal with from a kernelization perspective. The first new reduction rule, Reduction 8, circumvents this by applying a d_{TBR} -preserving transformation to one of the trees, such that the classical subtree reduction rule can be applied and the number of leaves can be reduced; in this way we stay in the world of trees. The transformation itself requires a very careful analysis of the way common chains behave when one of the chains is ‘interrupted’ in the other tree. Essentially, the transformation works by deleting an edge in one of the trees and replacing it with an edge that is ‘buried’ inside an artificially lengthened common chain, which ensures that d_{TBR} does not change.

The second new reduction rule, Reduction 9, works by identifying other topological structures which contribute heavily to the $11k - 9$ bound, and transforming them into structures that can be attacked by Reduction 8. Reduction 9 only applies when the region surrounding the topological structure contains many leaves; conversely, if Reduction 9 does not apply, the region is sparse. Reduction 10 is similar in spirit to Reduction 9, but is more direct: if it triggers, it is parameter reducing i.e. d_{TBR} is definitely reduced by 1. Once Reductions 8–10 no longer apply (or the earlier reduction rules), there is extensive sparsity in the underlying generator, which we use to obtain the new bound of $9k - 8$. We show that this bound is (essentially) tight by describing irreducible pairs of trees with TBR distance k that have $9k - 9$ leaves.

We anticipate that the new reduction rules will yield new advances for other agreement-forest based distances in phylogenetics, contribute to a deeper understanding of the combinatorics of agreement forests, and facilitate the ongoing advancement of kernelization within phylogenetics.

2 Preliminaries

2.1 Notation and terminology

Our notation closely follows [15]. Throughout this paper, X denotes a non-empty finite set of *taxa*.

Phylogenetic trees. An *unrooted binary phylogenetic tree* T on X is a simple, connected, and undirected tree whose leaves are bijectively labeled with X and whose other vertices all have degree 3. The set X is often referred to as the *leaf set* of T . See Figure 1 for an example of two unrooted binary phylogenetic trees on $X = \{a, b, c, d, e\}$. For simplicity and since most phylogenetic trees in this paper are unrooted and binary, we refer to an unrooted binary phylogenetic trees as a *phylogenetic tree*. If a definition or statement applies to all unrooted phylogenetic trees, regardless of whether they are binary or not, we make this explicit. Two leaves, say a and b , of T are called a *cherry* $\{a, b\}$ of T if they are adjacent to a common vertex. Moreover, for each $x \in X$, we use p_x to denote the unique neighbor of x in T and refer to p_x as the *parent* of x .

For $X' \subseteq X$, we write $T[X']$ to denote the unique, minimal subtree of T that connects all elements in X' . For brevity we call $T[X']$ the *embedding* of X' in T . For an edge e of T , we say that $T[X']$ *uses* e , if e is an edge of $T[X']$. Furthermore, we refer to the phylogenetic tree on X' obtained from $T[X']$ by suppressing degree-2 vertices as the *restriction of T to X'* and we denote this by $T|X'$.

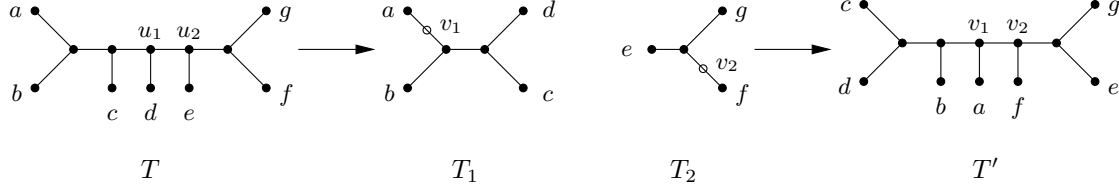


Fig. 2: A single TBR operation that transforms T into T' . First, T_1 and T_2 are obtained from T by deleting the edge $\{u_1, u_2\}$ in T . Second, T' is obtained from T_1 and T_2 by subdividing an edge in both trees as indicated by the open circles v_1 and v_2 and adding a new edge $\{v_1, v_2\}$.

Subtrees and chains. Let T be a phylogenetic tree on X . We say that a subtree of T is *pendant* if it can be detached from T by deleting a single edge. For $n \geq 2$, let $C = (\ell_1, \ell_2, \dots, \ell_n)$ be a sequence of distinct taxa in X . We call C an n -chain of T if there exists a walk p_1, p_2, \dots, p_n in T and the elements in p_2, p_3, \dots, p_{n-1} are all pairwise distinct. Note that ℓ_1 and ℓ_2 may have a common parent or ℓ_{n-1} and ℓ_n may have a common parent. Furthermore, if $p_1 = p_2$ or $p_{n-1} = p_n$ holds, then C is said to be *pendant* in T . To ease reading, we sometimes write C to denote the set $\{\ell_1, \ell_2, \dots, \ell_n\}$. It will always be clear from the context whether C refers to the associated sequence or set of taxa. If a pendant subtree S (resp. an n -chain C) exists in two phylogenetic trees T and T' on X , we say that S (resp. C) is a *common* subtree (resp. chain) of T and T' .

Tree bisection and reconnection. Let T be a phylogenetic tree on X . Apply the following three-step operation to T :

1. Delete an edge in T and suppress any resulting degree-2 vertex. Let T_1 and T_2 be the two resulting phylogenetic trees.
2. If T_1 (resp. T_2) has at least one edge, subdivide an edge in T_1 (resp. T_2) with a new vertex v_1 (resp. v_2) and otherwise set v_1 (resp. v_2) to be the single isolated vertex of T_1 (resp. T_2).
3. Add a new edge $\{v_1, v_2\}$ to obtain a new phylogenetic tree T' on X .

We say that T' has been obtained from T by a single *tree bisection and reconnection (TBR) operation* (or, *TBR move*). Furthermore, we define the TBR *distance* between two phylogenetic trees T and T' on X , denoted by $d_{\text{TBR}}(T, T')$, to be the minimum number of TBR operations that are required to transform T into T' . To illustrate, the trees T and T' in Figure 2 have a TBR distance of 1. It is well known that d_{TBR} is a metric [1]. By building on an earlier result by Hein et al. [11, Theorem 8], Allen and Steel [1] showed that computing the TBR distance is an NP-hard problem.

Agreement forests. Let T and T' be two phylogenetic trees on X . Furthermore, let $F = \{B_0, B_1, B_2, \dots, B_k\}$ be a partition of X , where each block B_i with $i \in \{0, 1, 2, \dots, k\}$ is referred to as a *component* of F . We say that F is an *agreement forest* for T and T' if the following conditions hold.

- (1) For each $i \in \{0, 1, 2, \dots, k\}$, we have $T|B_i = T'|B_i$.
- (2) For each pair $i, j \in \{0, 1, 2, \dots, k\}$ with $i \neq j$, we have that $T[B_i]$ and $T[B_j]$ are vertex-disjoint in T , and $T'[B_i]$ and $T'[B_j]$ are vertex-disjoint in T' .

Let $F = \{B_0, B_1, B_2, \dots, B_k\}$ be an agreement forest for T and T' . The *size* of F is simply its number of components; i.e. $k + 1$. Moreover, an agreement forest with the minimum number of components (over all agreement forests for T and T') is called a *maximum agreement forest* (MAF) for T and T' . The number of components of a maximum agreement forest for T and T' is denoted by $d_{\text{MAF}}(T, T')$. The following theorem is well known.

Theorem 1. [1, Theorem 2.13] *Let T and T' be two phylogenetic trees on X . Then*

$$d_{\text{TBR}}(T, T') = d_{\text{MAF}}(T, T') - 1.$$

A maximum agreement forest for the trees T and T' shown in Figure 2, which have TBR distance 1, therefore contains two components. $F = \{\{a, b, c, d\}, \{e, f, g\}\}$ is an example of such a forest (in fact, here it is the only maximum agreement forest).

Phylogenetic networks. An *unrooted binary phylogenetic network* N on X is a simple, connected, and undirected graph whose leaves are bijectively labeled with X and whose other vertices all have degree 3. Let E and V be the edge and vertex set of N , respectively. As with phylogenetic trees, we refer to an unrooted binary phylogenetic network simply as a *phylogenetic network*. Furthermore, we define the *reticulation number* of a phylogenetic network N as the number of edges in E that need to be deleted from N to obtain a spanning tree. More formally, we have $r(N) = |E| - (|V| - 1)$. If $r(N) = 0$, then N is simply a phylogenetic tree on X .

Let N be a phylogenetic network on X , and let T be a phylogenetic tree on X . We say that N displays T if, up to suppressing degree-two vertices, T can be obtained from N by deleting edges and vertices, in which case, the resulting subgraph of N is an *image* of T in N . Observe that an image of T in N is a subdivision of T .

Generators. Let k be a positive integer. For $k \geq 2$, a k -*generator* (or short *generator* when k is clear from the context) is a connected cubic multigraph with edge set E and vertex set V such that $k = |E| - (|V| - 1)$. The edges of a generator are called its *sides*. Intuitively, given a phylogenetic network N with $r(N) = k$, we can obtain a k -generator by, repeatedly, deleting all (labeled and unlabeled) leaves and suppressing any resulting degree-2 vertices. We say that the generator obtained in this way *underlies* N . Now, let G be a k -generator, let $\{u, v\}$ be a side of G , and let Y be a set of leaves. The operation of subdividing $\{u, v\}$ with $|Y|$ new vertices and, for each such new vertex w , adding a new edge $\{w, \ell\}$, where $\ell \in Y$ and Y bijectively labels the new leaves, is referred to as *attaching* Y to $\{u, v\}$ or as *decorating* $\{u, v\}$ with Y . Lastly, if at least one new leaf is attached to each loop and to each pair of parallel edges in G , then the resulting graph is a phylogenetic network N with $r(N) = k$. Note that N has no pendant subtree with more than a single leaf.

Hence, we have the following observation.

Observation 1 *Let N be a phylogenetic network that has no pendant subtree with at least two leaves, and let G be a generator. Then G underlies N if and only if N can be obtained from G by attaching a (possibly empty) set of leaves to each side of G .*

Unrooted minimum hybridization. In [22], it was shown that computing the TBR distance for a pair of phylogenetic trees T and T' on X is equivalent to computing the minimum number of extra edges required to simultaneously explain T and T' . More precisely, we set

$$h^u(T, T') = \min_N \{r(N)\},$$

where the minimum is taken over all phylogenetic networks N on X that display T and T' . The value $h^u(T, T')$ is known as the (*unrooted*) *hybridization number* of T and T' [22].

The aforementioned equivalence is given in the next theorem that was established in [22, Theorem 3].

Theorem 2. *Let T and T' be two phylogenetic trees on X . Then*

$$d_{\text{TBR}}(T, T') = h^u(T, T').$$

This means that $d_{\text{TBR}}(T, T') = k$ if and only if there exists a phylogenetic network N with $r(N) = k$ that displays both T and T' . Such an N can be obtained from its underlying generator, which has exactly $3(k-1)$ sides [14, Lemma 1], by attaching taxa to sides. The articles [14,15] use this fact extensively to derive a bound on the size of the kernelized instance. We will use the same generator-based framework for our results.

Parameterized algorithms. A *parameterized problem* is a problem for which the inputs are of the form (x, k) , where k is a non-negative integer, called the *parameter*. A parameterized problem is *fixed-parameter tractable* (FPT) if there exists an algorithm that solves³ any instance (x, k) in $f(k) \cdot |x|^{O(1)}$ time, where $f(\cdot)$ is a computable function depending only on k . A parameterized problem has a *kernel* of size $g(k)$, where $g(\cdot)$ is a computable function depending only on k , if there exists a polynomial time algorithm transforming any instance (x, k) into an equivalent problem (x', k') , with $|x'|, k' \leq g(k)$. Informally, this polynomial-time algorithm usually consists of reduction rules that are applied to an instance (x, k) to transform it into an equivalent but smaller instance (x', k') . If $g(k)$ is a polynomial in k then we call this a *polynomial kernel*; if $g(k) = O(k)$ then it is a *linear kernel*. It is well-known that a parameterized problem is fixed-parameter tractable if and only if it has a (not necessarily polynomial) kernel. For more background information on fixed parameter tractability and kernelization, we refer the reader to standard texts such as [5,6,10].

Let T and T' be two phylogenetic trees on X . To compute $d_{\text{TBR}}(T, T')$, we take d_{TBR} as the parameter k and take $|X|$, the number of leaves, as the size of the instance $|x|$. The reduction rules described in the following section produce a linear kernel and run in $\text{poly}(|X|)$ time.

2.2 Seven reductions to kernelize the TBR distance

We start this section by describing the existing seven reductions that have previously been used to establish kernelization results for computing the TBR distance. These existing reductions will be extended to ten reductions in Section 5.

³ Note that the formalism described here actually concerns *decision* (i.e. yes/no) problems, which in the context of the current article is most naturally “Is $d_{\text{TBR}} \leq k$?”. An FPT algorithm for answering this question can easily be transformed into an algorithm for computing d_{TBR} with similar asymptotic time complexity by increasing k incrementally from 0 until a yes-answer is obtained.

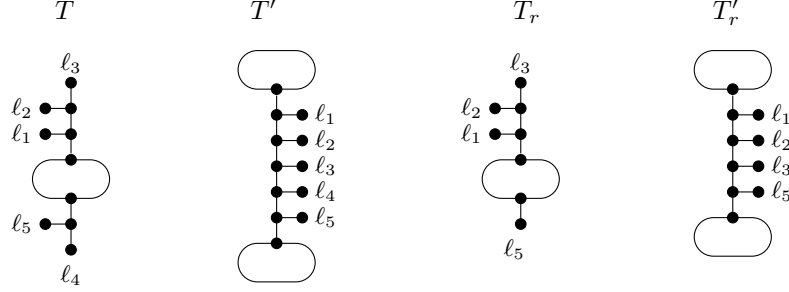


Fig. 3: An example of Reduction 7. Ovals indicate subtrees.

Let T and T' be two phylogenetic trees on X . The seven reductions are as follows.

Reduction 1. [1] If T and T' have a maximal common pendant subtree S with at least two leaves, then reduce T and T' to T_r and T'_r , respectively, by replacing S with a single leaf with a new label.

Reduction 2. [1] If T and T' have a maximal common n -chain $C = (\ell_1, \ell_2, \dots, \ell_n)$ with $n \geq 4$, then reduce T and T' to $T_r = T|X \setminus \{\ell_4, \ell_5, \dots, \ell_n\}$ and $T'_r = T'|X \setminus \{\ell_4, \ell_5, \dots, \ell_n\}$, respectively.

Reduction 3. [15] If T and T' have a common 3-chain $C = (\ell_1, \ell_2, \ell_3)$ such that $\{\ell_1, \ell_2\}$ is a cherry in T and $\{\ell_2, \ell_3\}$ is a cherry in T' , then reduce T and T' to $T_r = T|X \setminus C$ and $T'_r = T'|X \setminus C$, respectively.

Reduction 4. [15] If T and T' have a common 3-chain $C = (\ell_1, \ell_2, \ell_3)$ such that $\{\ell_2, \ell_3\}$ is a cherry in T and $\{\ell_3, x\}$ is a cherry in T' with $x \in X \setminus C$, then reduce T and T' to $T_r = T|X \setminus \{x\}$ and $T'_r = T'|X \setminus \{x\}$, respectively.

Reduction 5. [15] If T and T' have two common 2-chains $C_1 = (\ell_1, \ell_2)$ and $C_2 = (\ell_3, \ell_4)$ such that T has cherries $\{\ell_2, x\}$ and $\{\ell_3, \ell_4\}$, and T' has cherries $\{\ell_1, \ell_2\}$ and $\{\ell_4, x\}$ with $x \in X \setminus (C_1 \cup C_2)$, then reduce T and T' to $T_r = T|X \setminus \{x\}$ and $T'_r = T'|X \setminus \{x\}$, respectively.

Reduction 6. [15] If T and T' have two common 3-chains $C_1 = (\ell_1, \ell_2, \ell_3)$ and $C_2 = (\ell_4, \ell_5, \ell_6)$ such that T has cherries $\{\ell_2, \ell_3\}$ and $\{\ell_4, \ell_5\}$, and $(\ell_1, \ell_2, \dots, \ell_6)$ is a 6-chain of T' , then reduce T and T' to $T_r = T|X \setminus \{\ell_4, \ell_5\}$ and $T'_r = T'|X \setminus \{\ell_4, \ell_5\}$, respectively.

Reduction 7. [15] If T and T' have common chains $C_1 = (\ell_1, \ell_2, \ell_3)$ and $C_2 = (\ell_4, \ell_5)$ such that T has cherries $\{\ell_2, \ell_3\}$ and $\{\ell_4, \ell_5\}$, and $(\ell_1, \ell_2, \dots, \ell_5)$ is a 5-chain of T' , then reduce T and T' to $T_r = T|X \setminus \{\ell_4\}$ and $T'_r = T'|X \setminus \{\ell_4\}$, respectively.

An example of Reduction 7 is illustrated in Figure 3.

Reduction 1 is known as *subtree reduction* while Reduction 2 is known as *chain reduction* in the literature. Now, suppose that two phylogenetic trees T_r and T'_r have a common 3-chain $C = (\ell_1, \ell_2, \ell_3)$. We refer to the reverse of Reduction 2 which is the process of obtaining T and T' from T_r and T'_r , respectively, as *extending C to an n -chain* for $n > 3$. We will always explicitly say in which order and to which end of C we add the new leaves $\ell_4, \ell_5, \dots, \ell_n$.

The following lemma and theorem summarize results established in [1,14,15].

Lemma 1. *Let T and T' be two phylogenetic trees on X . If T_r and T'_r are two phylogenetic trees obtained from T and T' , respectively, by a single application of Reduction 1, 2, 6, or 7, then $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(T_r, T'_r)$. Moreover, if T_r and T'_r are two trees obtained from T*

and T' , respectively, by a single application of Reduction 3, 4, or 5, then $d_{\text{TBR}}(T, T') - 1 = d_{\text{TBR}}(T_r, T'_r)$.

Theorem 3. *Let S and S' be two phylogenetic trees on X that cannot be reduced by Reduction 1 or 2, and let T and T' be two phylogenetic trees on Y that cannot be reduced by any of Reductions 1–7. If $d_{\text{TBR}}(S, S') \geq 2$, then $|X| \leq 15d_{\text{TBR}}(S, S') - 9$. Furthermore, if $d_{\text{TBR}}(T, T') \geq 2$, then $|Y| \leq 11d_{\text{TBR}}(T, T') - 9$.*

Note that each of Reductions 3, 4, and 5 triggers a *parameter reduction*, whereby the TBR distance is reduced by one. In these cases, an element of X is located which definitely comprises a singleton component in some maximum agreement forest, and whose deletion thus lowers the TBR distance by 1. Reductions 1, 2, 6 and 7, on the other hand, preserve the TBR distance. Reduction 6 and 7 work by truncating short chains, i.e. chains which escape Reduction 2, to be even shorter.

The following minor observation is worth noting.

Observation 2 *Assume that Reductions 1–7 have been applied to exhaustion. Suppose T and T' have a common chain $C = (b, c, d)$ that is pendant in T' . Then C is not pendant in T .*

Proof. If C was pendant in T then at least one of the subtree reduction or Reduction 3 would be applicable on C , contradicting the assumption that the reduction rules had been applied to exhaustion. \square

We end this section by outlining some of the machinery used in [14,15] to kernelize the TBR distance. This article builds on that machinery and further refines it. Let T and T' be two phylogenetic trees on X that cannot be reduced under Reduction 1 or 2, and let N be a phylogenetic network on X that displays T and T' . Let R and R' be spanning trees of N obtained by greedily extending an embedding of T (respectively, T') to become a spanning tree, if it is not that already. Since N displays T and T' , R and R' exist. Furthermore, let G be the generator that underlies N . Since T and T' are subtree and chain reduced, N does not have a pendant subtree of size at least two. Hence, by Observation 1, we can obtain N from G by attaching leaves to G . Let $S = \{u, w\}$ be a side of G . Let $Y = \{\ell_1, \ell_2, \dots, \ell_m\}$ be the set of leaves that are attached to S in obtaining N from G . Recall that $m \geq 0$. Then there exists a path

$$u = v_0, v_1, v_2, \dots, v_m, v_{m+1} = w$$

of vertices in N such that, for each $i \in \{1, 2, \dots, m\}$, v_i is the unique neighbor of ℓ_i . We refer to this path as the *path associated with S* and denote it by P_S . Importantly, for a path P_S in N that is associated with a side S of G , there is at most one edge in P_S that is not contained in R , and there is at most one (not necessarily distinct) edge in P_S that is not contained in R' . We make this precise in the following definition and say that S is a *b-breakpoint side* relative to R and R' , where

1. $b = 0$ if R and R' both contain all edges of P_S ,
2. $b = 1$ if one element in $\{R, R'\}$ contains all edges of P_S while the other element contains all but one edge of P_S , and
3. $b = 2$ if each of R and R' contains all but one edge of P_S .

Since R and R' span N , note that S cannot have more than two breakpoints relative to R and R' . Let $S = \{u, w\}$ be a side of G to which four taxa get attached in obtaining N

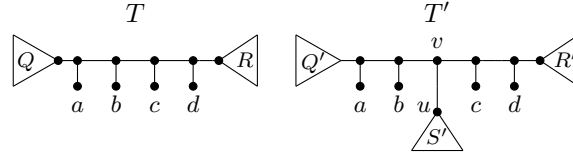


Fig. 4: Here (a, b, c, d) is an interrupted 4-chain of T and T' . Triangles indicates subtrees of T and T' . The sets S, Q, R, Q', R' are referred to in the proof of Theorem 5, which proves that at least one maximum agreement forest of T, T' does *not* use the edge $\{u, v\}$ in T' . Note that S' must contain at least one leaf, but the leaf set of any of the subtrees Q, R, Q' , and R' may be empty, in which case $\{a, b\}$ or $\{c, d\}$ can become cherries.

from G , and let $P_S = u, p_a, p_b, p_c, p_d, w$ be the path associated with S . For shorthand we will throughout this article use notation such as $2|2$ or $S = ab|cd$ to refer to a side S if P_S has a single breakpoint such that one of R and R' does not contain the edge $\{p_b, p_c\}$, and $2|1|1$ or $S = ab|c|d$ to refer to a side S if P_S has two breakpoints such that one of R and R' does not contain the edge $\{p_b, p_c\}$ and the other does not contain the edge $\{p_c, p_d\}$. If R and R' both have the same breakpoint (i.e. there exists an edge of P_S that neither R nor R' contains), then we write, for example, $2||2$ or $1||3$. Lastly, note that there also may exist a side such that R or R' does not contain the edge $\{u, p_a\}$ or $\{p_d, w\}$ in which case we write, for example, $0|2|2$, $0|4|0$, or $0|4$. Similar notation extends to sides in G to which three taxa get attached in obtaining N from G .

3 Two technical results about short chains

This section present two technical but powerful theorems that play a crucial part in the upcoming sections. The first, Theorem 4, was established in [15, Theorem 5], while the second, Theorem 5, is new to this paper.

Let $F = \{B_0, B_1, B_2, \dots, B_k\}$ be an agreement forest for two phylogenetic trees T and T' on X , and let Y be a subset of X . We say that Y is *preserved* in F if there exists an element B_i in F with $i \in \{0, 1, 2, \dots, k\}$ such that $Y \subseteq B_i$. Throughout the article we will make heavy use of the following theorem, referred to as the *chain preservation theorem (CPT)*.

Theorem 4. *Let T and T' be two phylogenetic trees on X . Let K be an (arbitrary) set of mutually taxa-disjoint chains that are common to T and T' . Then there exists a maximum agreement forest F of T and T' such that*

1. *every n -chain in K with $n \geq 3$ is preserved in F , and*
2. *every 2-chain in K that is pendant in at least one of T and T' is preserved in F .*

Following on from the last theorem, we say that common n -chains with $n \geq 3$, and common 2-chains that are pendant in at least one of T and T' are *CPT-eligible* chains. In our proofs, CPT-eligible chains will function as ‘obstructions’ that allow us to reason about the structure of maximum agreement forests.

We now turn to the second technical result whose proof is given in the appendix. Let $C = (a, b, c, d)$ be a 4-chain. We say that C is an *interrupted 4-chain* of two phylogenetic trees

T and T' on X if C is a chain of T and, in T' , there exists a walk p_a, p_b, v, p_c, p_d such that p_a , v , and p_c are three pairwise distinct vertices. Note that v is not necessarily the parent of a leaf in T' . Furthermore, by deleting the edge $e = \{u, v\}$ in T' with $u \notin \{p_a, p_c\}$ and suppressing v , the resulting tree and T have C as a common 4-chain. We call e the *interrupter* of C . An example of an interrupted 4-chain is shown in Figure 4.

Theorem 5. *Let T and T' be two phylogenetic X -trees, and let $C = (a, b, c, d)$ be an interrupted 4-chain of T and T' . Then there exists a maximum agreement forest F for T and T' such that, for each $B \in F$ with $T'[B]$ does not use the interrupter of C in T' .*

4 Main result and a bird's-eye view of the main arguments

In this section, we state the main result of this paper and give an overview of our approach to establish it. The following lemma summarizes the situation after Reductions 1–7 have been applied to exhaustion and is the foundation of Theorem 3.

Lemma 2. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X such that $r(N) = d_{\text{TBR}}(T, T')$. Then, in obtaining N from G , the following statements hold.*

- (a) *At most four taxa can be attached to each side of G .*
- (b) *At most three taxa can be attached to each 0-breakpoint side of G .*
- (c) *At most four taxa can be attached to each 1-breakpoint side of G and only sides of the form 1|3 and 2|2 can achieve this upper bound.*
- (d) *At most four taxa can be attached to each 2-breakpoint side of G and only sides of the form 2|1|1 can achieve this upper bound.*

Proof. In [15, Lemma 7], the authors showed that each 0-breakpoint side of G has at most three taxa and that each other side of G has at most four taxa. Consider a 1-breakpoint side S of G . Suppose that four leaves get attached to S in obtaining N from G . If S is a 0|4 side, then T and T' have a common 4-chain, contradicting that Reduction 2 has been applied to exhaustion. Hence S is either a 1|3 or 2|2 side. Next, consider a 2-breakpoint side S of G . Again, suppose that four leaves get attached to S in obtaining N from G . If S is a 0||4, 1||3, or 2||2 side on which the breakpoints of T and T' coincide, then T and T' have a common subtree with at least two leaves, contradicting that Reduction 1 has been applied to exhaustion. Otherwise, if the breakpoints of T and T' do not coincide and S is a 0|2|2, 0|1|3, 0|3|1, or 0|4|0 side, then it is straightforward to check that T and T' would have been reduced by Reduction 1, 4, 3 or 2 respectively, again a contradiction. It now follows that S is a 2|1|1 side. \square

Let G be a k -generator as described in Lemma 2. By [14, Lemma 1], G has $3(k - 1)$ sides, and there are $2k$ breakpoints to divide across these sides. Intuitively, after attaching the elements in X to sides of G in order to obtain N , we delete k edges in N to obtain a subdivision of T and k edges to obtain a subdivision of T' . Instead of deleting edges in N , we think of this as placing breakpoints on the sides of G . Now, with a view towards obtaining an upper bound on the number of leaves that can be attached to any k -generator G , the best we can do after applying Reductions 1–7 to exhaustion is to have $2k$ 1-breakpoint sides with four taxa each, and $(k - 3)$ 0-breakpoint sides with three taxa each; this is the origin of the $11k - 9$ kernel in [15]. The main bottleneck in achieving a kernel that is smaller than $11k - 9$

are sides with four taxa and, in particular, those that only have one breakpoint. This explains the heavy emphasis on $1|3$, $2|2$ and $2|1|1$ sides in the rest of the article.

The high-level idea to achieve a kernel for d_{TBR} that is smaller than $11k - 9$ is as follows. In Section 5, we present new Reductions 8, 9, and 10.

1. A $1|3$ side triggers Reduction 8 that, as long as a ‘secondary’ common 3-chain is available, reduces the number of taxa by one. The ‘3’ part in a $1|3$ side can itself function as a secondary chain for another $1|3$ side, so this reduction rule eliminates all but at most one $1|3$ side.
2. A $2|2$ side that (informally) has relatively many taxa on the adjacent sides can be reduced by Reduction 9, which essentially first transforms such a side into a $1|3$ side before executing Reduction 8 if another $1|3$ side (and therefore a ‘secondary’ common 3-chain) is available.
3. A $2|1|1$ side that (informally) has relatively many taxa on the adjacent sides triggers the parameter-reducing Reduction 10.

After applying Reductions 1–10 to exhaustion, all sides with four taxa (apart from possibly one single exception) do *not* have many taxa on the adjacent sides. This has the consequence that, once these sparse adjacent sides are taken into account, 4-taxa sides contribute (on average) significantly fewer than four taxa per side. With some careful counting this leads to an improved kernel of size $9k - 8$.

We are now in a position to state the main result of this paper. The proof is deferred until Section 7.

Theorem 6. *Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–10. If $d_{\text{TBR}}(T, T') \geq 2$, then $|X| \leq 9d_{\text{TBR}}(T, T') - 8$.*

We finish this section by noting that the portfolio of reduction rules should always be executed in the order 1, 2, \dots , 10. Every time a reduction rule executes, the sequence should be restarted from Reduction 1. The fact that in all cases the number of taxa (and sometimes the TBR distance) is reduced by at least one, and the fact that the reduction rules themselves can be executed in polynomial-time, ensures polynomial-time execution overall.

5 Three new reduction rules

5.1 Reduction 8: A reduction rule to reduce a $1|3$ side if there is a spare common 3-chain available.

The reduction rule that we describe first is designed to target the structures of two phylogenetic trees that are induced by a $1|3$ side $S = a|bcd$. We start by describing the first of two parts of Reduction 8 and already note here that the second part is an application of Reduction 1.

Reduction 8A. Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–7. Suppose that T and T' have two leaf-disjoint common 3-chains $C = (b, c, d)$ and $D = (e, f, g)$ such that C is pendant in T' with cherry $\{b, c\}$ and C is not pendant in T , and there exists a taxon a such that (a, b, c, d) is a chain of T and not a chain

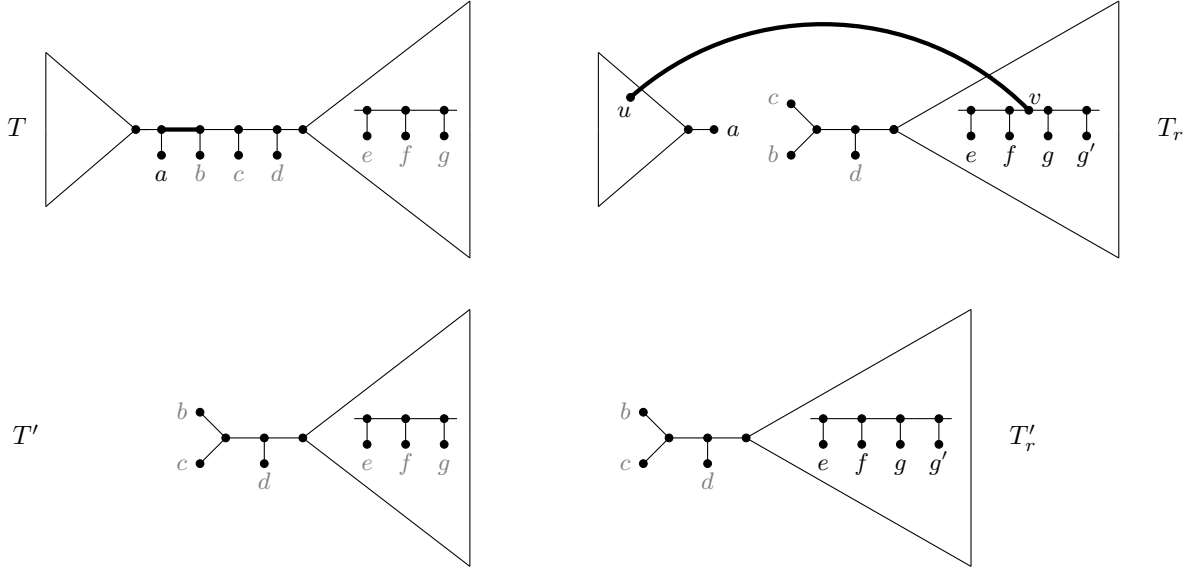


Fig. 5: Reduction 8A can be used to reduce a 1|3 side $a|bcd$, as long as a secondary common 3-chain (here $\{e, f, g\}$) is available. Swapping the bold edge in T with the bold edge in T_r preserves d_{TBR} and creates a common pendant subtree $\{b, c, d\}$ which can be reduced under Reduction 1. Observe that deleting the edge $\{p_a, p_b\}$ in T disconnects T into two smaller trees and, in the example above, (b, c, d) and (e, f, g) are leaves of the same smaller tree. In general, this does not need to be the case: (b, c, d) and (e, f, g) can be in different subtrees.

of T' . Then obtain T_r from T by extending D to the 4-chain (e, f, g, g') such that $g' \notin X$, deleting the edge $\{p_a, p_b\}$, suppressing p_a and p_b , subdividing the edge $\{p_f, p_g\}$ with a new vertex v , and adding the edge $\{u, v\}$, where u is a new vertex that subdivides an arbitrary edge of the component that does not contain e such that (b, c, d) is a pendant 3-chain in T_r . Finally, obtain T'_r from T' by extending D to the 4-chain (e, f, g, g') . Then T_r and T'_r are two phylogenetic trees on $X \cup \{g'\}$. In what follows, we will call D the *secondary common 3-chain* when executing Reduction 8A. An application of this reduction is shown in Figure 5.

Lemma 3. *Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–7. If T_r and T'_r are two phylogenetic trees obtained from T and T' , respectively, by a single application of Reduction 8A, then $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(T_r, T'_r)$, and T_r and T'_r have a common pendant subtree of size three.*

Proof. We establish the theorem using the same notation as in the definition of Reduction 8A. Since T and T' cannot be reduced under any of Reductions 1–7, neither C nor D is the leaf set of a common subtree of T and T' . Furthermore, by Observation 2, D is pendant in at most one of T and T' . Hence, without loss of generality, we assume that $\{f, g\}$ is not a cherry in either T or T' . Now, let S and S' be the two phylogenetic trees obtained from T and T' , respectively, by extending D to the 4-chain (e, f, g, g') . It follows from applying Lemma 1 to Reduction 2 that $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(S, S')$. It remains to show that $d_{\text{TBR}}(S, S') = d_{\text{TBR}}(T_r, T'_r)$. First, let $F = \{B_0, B_1, B_2, \dots, B_k\}$ be a maximum agreement forest for S and S' . By CPT, we may assume that $C \subseteq B_i$ for some $i \in \{0, 1, 2, \dots, k\}$. Let S_1 and S_2 be the two phylogenetic trees obtained from S by deleting the edge $e_1 = \{p_a, p_b\}$ such that S_1 does not contain b . Since C is

pendant in S' , it follows that B_i does not contain a taxon of S_1 . This in turn implies that e_1 is not used by any embedding $S[B_j]$ with $B_j \in F \setminus \{B_i\}$. Hence F is an agreement forest for T_r and T'_r . Second, let $F_r = \{B_0, B_1, B_2, \dots, B_k\}$ be a maximum agreement forest for T_r and T'_r . By Theorem 5, we may assume that there exists no component $B_i \in F_r$ with $i \in \{0, 1, 2, \dots, k\}$ such that $T_r[B_i]$ uses the interrupter of (e, f, g, g') . Hence F_r is an agreement forest for S and S' . Combining both cases, establishes that $d_{\text{TBR}}(S, S') = d_{\text{TBR}}(T_r, T'_r)$. Moreover, by construction, $\{b, c, d\}$ is the leaf set of a common pendant subtree of T_r and T'_r . \square

We are now in a position to describe Reduction 8.

Reduction 8. Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–7. If T and T' can be reduced under Reduction 8A, then reduce T and T' to T_r and T'_r , respectively, by an application of Reduction 8A followed by an application of Reduction 1.

If T_r and T'_r are obtained from T and T' as described in Reduction 8, we say that *Reduction 8 is applied to $C = (b, c, d)$ and $D = (e, f, g)$* , where C and D are as defined in Reduction 8A.

The next theorem shows that an application of Reduction 8 preserves the TBR distance and reduces the number of taxa by one. Furthermore, this reduction can be executed in polynomial time by trying all possible candidates for the taxa $\{a, b, c, d, e, f, g\}$ as defined in Reduction 8A.

Theorem 7. *Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–7. Suppose that T and T' can be reduced under Reduction 8A. Let T_r and T'_r be two phylogenetic trees on X' that are obtained from T and T' , respectively, by a single application of Reduction 8. Then $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(T_r, T'_r)$ and $|X'| = |X| - 1$.*

Proof. Let S and S' be the two phylogenetic trees on $|X| + 1$ leaves obtained from T and T' , respectively, by a single application of Reduction 8A. By Lemma 3, we have $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(S, S')$. Moreover, using the same notation as in the definition of Reduction 8A, it follows that S and S' have a common pendant subtree with leaf set $\{b, c, d\}$. Setting T_r and T'_r to be the two phylogenetic trees obtained from S and S' , respectively, by applying Reduction 1 to $\{b, c, d\}$ and noting that $|X'| = |X| + 1 - 2 = |X| - 1$ establishes the theorem. \square

Reduction 8 also leads to the following observation, which we will need later.

Observation 3 *Let N be a phylogenetic network on X that displays two phylogenetic trees T and T' on X that cannot be reduced under any of Reductions 1–8. Let G be the generator that underlies N . Then G has at most one 1|3 side.*

Proof. Suppose that there are two distinct such sides, $a|bcd$ and $.|efg$ where “.” denotes a single taxon. Clearly, $\{b, c, d\} \cap \{e, f, g\} = \emptyset$ because the taxa are from distinct sides of G . Then $C = \{b, c, d\}$ and $D = \{e, f, g\}$ are two common 3-chains of T and T' that satisfy the three properties described in the definition of Reduction 8A. Hence, T and T' can be further reduced under Reduction 8A and, therefore, under Reduction 8, a contradiction. \square

5.2 Reduction 9: A reduction rule that triggers Reduction 8 by transforming certain 2|2 sides into 1|3 sides.

The next reduction rule targets the structures of two phylogenetic trees that are induced by a 2|2 side $S = ab|cd$. We start by introducing an operation that does not reduce the number of leaves in the trees. Instead, this operation transforms certain 2|2 sides of a generator into 1|3 sides and is a precursor (hence the name P) to Reduction 9 that is described towards the end of this subsection.

Operation P. Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–8. Suppose that T has a non-pendant chain (a, b, c, d) , T' has cherries $\{a, b\}$ and $\{c, d\}$, and there exists a maximum agreement forest F for T and T' such that $\{a, b\}$ and $\{c, d\}$ are each preserved in F , but $\{a, b, c, d\}$ is not preserved in F . Then let $S = T$, and let S' be the tree obtained from T' by deleting b , suppressing p_b , subdividing the edge incident with c with a new vertex v , and adding the edge $\{v, b\}$.

If $\{a, b, c, d\}$ satisfies all properties in the description of Operation P, we say that $\{a, b, c, d\}$ is *eligible for Operation P*. Moreover, if S and S' are obtained from T and T' as described above, we say that *Operation P is applied to $\{a, b, c, d\}$* .

Theorem 8. *Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–8. Furthermore, let S and S' be two phylogenetic trees obtained from T and T' , respectively, by applying Operation P to $\{a, b, c, d\} \subseteq X$. Then $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(S, S')$. Moreover, (b, c, d) is a common chain of S and S' , and S' has cherry $\{b, c\}$.*

Proof. We establish the theorem using the same notation as in the definition of Operation P. First, let $F = \{B_0, B_1, B_2, \dots, B_k\}$ be a maximum agreement forest for T and T' such that $\{a, b\}$ and $\{c, d\}$ are each preserved in F , but $\{a, b, c, d\}$ is not preserved in F . Since the elements in $\{T[B_i] : i \in \{0, 1, 2, \dots, k\}\}$ are pairwise vertex disjoint, no element $T[B_i]$ uses the edge $\{p_b, p_c\}$. Let B_j be the element in F such that $\{a, b\} \subseteq B_j$ and, similarly, let $B_{j'}$ be the element in F such that $\{c, d\} \subseteq B_{j'}$. Then

$$(F \setminus \{B_j, B_{j'}\}) \cup \{B_j \setminus \{b\}, B_{j'} \cup \{b\}\}$$

is an agreement forest for S and S' that has the same size as F . Hence $d_{\text{TBR}}(S, S') \leq d_{\text{TBR}}(T, T')$. Second, let $F = \{B_0, B_1, B_2, \dots, B_k\}$ be a maximum agreement forest for S and S' . By CPT, we may assume that $\{b, c, d\} \subseteq B_j$ for some $j \in \{0, 1, 2, \dots, k\}$. Since $S[B_j] = S'[B_j]$ and the elements in $\{S[B_i] : i \in \{0, 1, 2, \dots, k\}\}$ are pairwise vertex disjoint, it follows that the edge $\{p_a, p_b\}$ is not used by $S[B_i]$ for any $i \in \{0, 1, \dots, k\}$. Now let $B_{j'}$ be the element in $F \setminus \{B_j\}$ such that $a \in B_{j'}$. Then

$$(F \setminus \{B_j, B_{j'}\}) \cup \{B_j \setminus \{b\}, B_{j'} \cup \{b\}\}$$

is an agreement forest for T and T' that has the same size as F . Thus $d_{\text{TBR}}(S, S') \geq d_{\text{TBR}}(T, T')$. Combining both cases establishes that $d_{\text{TBR}}(S, S') = d_{\text{TBR}}(T, T')$. Moreover, by construction of S and S' it follows immediately that (b, c, d) is a common chain of S and S' , and S' has cherry $\{b, c\}$. \square

Let T and T' be two phylogenetic trees on X . Following on from the description of Operation P, we present an explicit, polynomial-time algorithm—called Algorithm 1—in the appendix for testing whether or not, given a subset $\{a, b, c, d\}$ of X , there exists a maximum agreement forest F for T and T' such that $\{a, b\}$ and $\{c, d\}$ are each preserved in F , but $\{a, b, c, d\}$ is not preserved in F . Although the algorithm does not necessarily catch *all* situations when F exists, it is enough for our purposes. The high-level idea is that, as soon as a side $ab|cd$ has ‘many taxa on its surrounding sides’, then T , and T' will contain easily detectable structures that constitute a certificate for the existence of F and Algorithm 1 will find them.

The following corollary to Theorem 8 is useful later.

Corollary 1. *Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–8, and let S and S' be the two phylogenetic trees on X that are obtained from T and T' , respectively, by applying Operation P to $\{a, b, c, d\} \subseteq X$. Furthermore, let N be a phylogenetic network on X that displays T and T' . If the generator G that underlies N has a 2|2 side $S = ab|cd$, then N also displays S and S' .*

Proof. Using the same notation as in the definition of Operation P, observe that the breakpoint on S is relative to T' . To see that N also displays S and S' , we view S as the 1|3 side $a|bcd$, where the breakpoint is now relative to S' . \square

Theorem 8 and Corollary 1 are the theoretical foundation for Operation P. Once Operation P is applied, Reduction 8 may be triggered if another 1|3 side is available. This can happen in two slightly different ways which we describe next as Reduction 9.1 and 9.2. Reduction 9.1 is tried first and, if it fails, Reduction 9.2 is tried. Essentially Reduction 9.1 converts one 2|2 side into a 1|3 side and Reduction 9.2 converts two 2|2 sides into 1|3 sides. In both cases, the new 1|3 sides trigger Reduction 8. Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–8. Then Reductions 9.1 and 9.2 are defined as follows.

Reduction 9.1. Suppose that there exists $\{a, b, c, d\} \subseteq X$ such that $C = (b, c, d)$ is a common chain of T and T' , C is pendant in T' with cherry $\{b, c\}$ and not pendant in T , and (a, b, c, d) is a chain of T and not a chain of T' . Suppose furthermore that there exists $\{a', b', c', d'\} \subseteq X$ which is eligible for Operation P, where $\{a, b, c, d\} \cap \{a', b', c', d'\} = \emptyset$. Then an application of Reduction 9.1 to T and T' consists of an application of Operation P to $\{a', b', c', d'\}$, thereby creating two phylogenetic trees with a common 3-chain (b', c', d') , and a subsequent application of Reduction 8 to C and the newly created secondary common 3-chain $D = (b', c', d')$.

Reduction 9.2. Suppose that there exist two disjoint subsets $\{a', b', c', d'\}$ and $\{a'', b'', c'', d''\}$ of X , such that both are eligible for Operation P. Then, an application of Reduction 9.2 to T and T' consists of an application of Operation P to $\{a', b', c', d'\}$ followed by an application of the same operation to $\{a'', b'', c'', d''\}$ if it is still eligible⁴, and finally an application of Reduction 8 to $C = \{b', c', d'\}$ and the secondary common chain $D = (b'', c'', d'')$.

It is important to note that Reduction 9.2 is an ‘all-or-nothing’ reduction, i.e. it either executes fully or not at all. Specifically, it does not execute the first application of Operation P but not

⁴ In our analysis later in the article, we apply Reduction 9.2 in a situation where $\{a'', b'', c'', d''\}$ is definitely still eligible for transformation after $\{a', b', c', d'\}$ has been transformed.

the second. As Algorithm 1 (see appendix) runs in polynomial time, it follows that Reductions 9.1 and 9.2 can be executed in polynomial time by trying all possible candidates for the taxa $\{a, b, c, d, a', b', c', d'\}$ and $\{a', b', c', d', a'', b'', c'', d''\}$, respectively. Lastly, since we do not always need to distinguish between Reduction 9.1 and Reduction 9.2, we refer to an application of one of the two reductions as *Reduction 9*.

5.3 Reduction 10: A reduction rule to reduce certain 2|1|1 sides.

The last new reduction rule targets the structures of two phylogenetic trees that are induced by a 2|1|1 side $S = ab|c|d$. Reduction 10 is much more straightforward than Reductions 8 and 9.

Reduction 10. Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–9. If T has two cherries $\{a, b\}$ and $\{c, d\}$, T' has the 3-chain (a, b, c) such that $\{b, c\}$ is a cherry, and there exists a maximum agreement forest F for T and T' such that $\{c\} \in F$, then reduce T and T' to $T_r = T|X \setminus \{c\}$ and $T'_r = T'|X \setminus \{c\}$, respectively.

If $\{a, b, c, d\}$ satisfies all properties in the description of Reduction 10, we say that $\{a, b, c, d\}$ is *eligible for Reduction 10*. Moreover, if T_r and T'_r are obtained from T and T' as described above, we say that *Reduction 10 is applied to $\{a, b, c, d\}$* .

The next theorem shows that Reduction 10 is parameter reducing. Its proof is straightforward and omitted.

Theorem 9. Let T and T' be two phylogenetic trees on X that cannot be reduced under any of Reductions 1–9. Furthermore, let T_r and T'_r be two phylogenetic trees obtained from T and T' , respectively, by a single application of Reduction 10. Then $d_{\text{TBR}}(T_r, T'_r) = d_{\text{TBR}}(T, T') - 1$.

It remains to establish that Reduction 10 can be executed in polynomial time. In the appendix, we present an explicit, polynomial-time algorithm—called Algorithm 2—for testing whether there exists a maximum agreement forest F for T and T' such that $\{c\} \in F$. As Algorithm 2 runs in polynomial time, it follows that Reduction 10 can be executed in polynomial time by trying all possible candidates for the taxa $\{a, b, c, d\}$ as defined in Reduction 10. Moreover, an application of Reduction 10 decreases the number of taxa and the TBR distance both by exactly 1.

6 A win-win scenario

In this section we explore the interplay of sides of a generator that are adjacent to each other. We will see that a generator side whose adjacent sides are densely decorated with taxa trigger reduction rules and that a generator side that does not trigger a reduction rule has adjacent sides that are, on average, only sparsely decorated with taxa. To this end, we establish several results that pinpoint when a subset of taxa is eligible for Operation P or Reduction 10. We begin with a key insight.

Observation 4 Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X such that $r(N) = d_{\text{TBR}}(T, T')$, and let S be a side of G . If at least three taxa are attached to S in obtaining N from G , then T and T' have a CPT-eligible chain unless S is a 1|1|1 side.

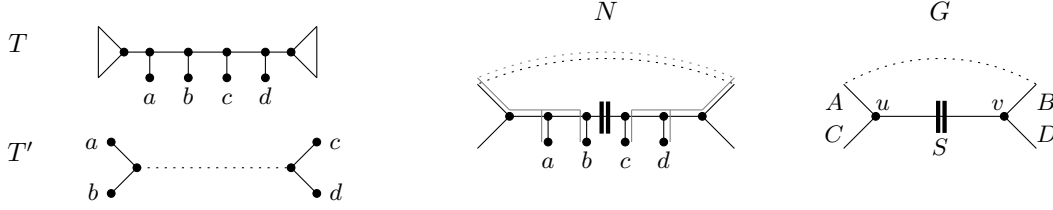


Fig. 6: The situation described in Theorem 10, which concerns $2|2$ sides $S = ab|cd$, where S is not part of a multi-edge. The grey edges in N indicate an image of T' .

Proof. If S is a 0-breakpoint side, then it immediately follows that T and T' have a common 3-chain that is CPT eligible. Suppose that S is a 1-breakpoint side. Then S is a $n_1|n_2$ side, where $n_1 \geq 0$ and $n_2 \geq 0$ denote the number of taxa attached to S on either side of the breakpoint. Since $n_1 + n_2 \geq 3$, either $n_1 \geq 2$ or $n_2 \geq 2$. Hence T and T' have a common 2-chain that is pendant in one of T and T' and therefore CPT eligible. Lastly, suppose that S is a 2-breakpoint side. Similar to the 1-breakpoint case, S is a $n_1|n_2|n_3$ side, where $n_1 \geq 0$, $n_2 \geq 0$, and $n_3 \geq 0$ denote the number of taxa attached to S before the first breakpoint, after the first and before the second breakpoint, and after the second breakpoint, respectively. Since $n_1 + n_2 + n_3 \geq 3$ and S is not a $1|1|1$ side it again follows that T and T' have a common 2-chain that is pendant in one of T and T' and therefore CPT eligible. \square

In the remainder of this section, we carefully analyze $2|2$ and $2|1|1$ sides, and establish sufficient conditions under which a subset of taxa that decorates such a side is eligible for Operation P or Reduction 10. Let S be a side of a generator G . Viewing G as a graph, S can either be a *simple edge*, i.e. an edge that is not part of a multi-edge, an edge that is part of a *multi-edge*, or a *loop*. A multi-edge of G contains at most two edges, due to the fact that each vertex of G has degree three. The only exception is if G has exactly two vertices, and one multi-edge consisting of three edges. This implies that $d_{\text{TBR}} \leq 2$. By assuming throughout the rest of the paper that $d_{\text{TBR}} \geq 3$ we can exclude this case⁵. The following analyses depend on whether S is a simple edge, an edge of a multi-edge, or a loop.

Observe that a $2|2$ side cannot be a loop because the phylogenetic tree that does not have a breakpoint on that side would contain a cycle.

6.1 $2|2$ sides

Theorem 10. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. Furthermore, let $S = ab|cd$ be a side of G that is a simple edge $\{u, v\}$. Let A , B , C , and D be the four sides incident with S such that A and C are both incident with u , and B and D are both incident with v . If each of A and C is decorated with at least two taxa, or each of B and D is decorated with at least two taxa in obtaining N from G , then $\{a, b, c, d\}$ is eligible for Operation P.*

Proof. Assume without loss of generality that T' has cherries $\{a, b\}$ and $\{c, d\}$ and that T has a non-pendant chain (a, b, c, d) as illustrated in Figure 6. Then (a, b) and (c, d) are two

⁵ This does not harm the final upper bound $9k - 8$ on the size of the kernel, because we can easily test in polynomial time whether $d_{\text{TBR}} \leq 2$ and if so exactly compute d_{TBR} in the same time bound. Subsequently we can output a trivial YES/NO instance to complete the kernelization.

common 2-chains of T and T' . Both of these chains are pendant in T' and therefore CPT eligible. For each $Y \in \{A, B, C, D\}$, let P_Y be the path associated with Y in N . Now, consider an image I of T' in N . Let P be the path from a to c in I . Since S is a 1-breakpoint side, one of P_A and P_C is a subpath of P and, similarly, one of P_B and P_D is a subpath of P . We assume without loss of generality that P_A and P_B are both subpaths of P . It follows that T' has no breakpoint on A or B and, hence each of A and B has at most one breakpoint relative to T . Moreover, by the assumption in the statement of the theorem, at least one of A and B is decorated with at least two taxa in the process of obtaining N from G .

Suppose that A is decorated with at least three taxa. Since A has at most one breakpoint, it follows from Observation 4 that T and T' have a CPT-eligible chain Z whose elements are attached to A in obtaining N from G . Let $K = \{\{a, b\}, \{c, d\}, Z\}$. By applying the CPT to K , there exists a maximum agreement forest F for T and T' such that each element in K is preserved in F . Assume that there exists an element B in F such that $\{a, b, c, d\} \subseteq B$. Let B' be the element in F such that $Z \subseteq B'$. If $B \neq B'$, then $T'[B]$ and $T'[B']$ are not vertex disjoint, a contradiction. Hence $B = B'$. But then $T|(\{a, b, c, d\} \cup Z) \neq T'|(\{a, b, c, d\} \cup Z)$, another contradiction. It follows that B does not exist and $\{a, b, c, d\}$ is eligible for Operation P. An identical analysis holds for when B is decorated with at least three taxa.

We can now assume that neither A nor B is decorated with at least three taxa. Then, by the statement of the theorem, A or B is decorated with exactly two taxa. We establish the theorem for when A is decorated with exactly two taxa. An analogous and symmetric argument holds for when B is decorated with exactly two taxa. Let e and f be the two taxa that are attached to A in obtaining N from G such that the path from p_e to p_a in T' does not pass through p_f . Intuitively, e is closer than f to a in N . If $A = |ef$ or $A = ef|$, then T and T' have a common 2-chain (e, f) that is pendant in T . By applying an argument that is similar to that of the last paragraph and setting $Z = \{e, f\}$, we deduce that $\{a, b, c, d\}$ is eligible for Operation P. Hence $A = ef$ or $A = e|f$. Let F be a maximum agreement forest for T and T' . Assume that there exists an element B in F such that $\{a, b, c, d\} \subseteq B$. Since $T|B = T'|B$, we have $B = \{a, b, c, d\}$. Furthermore, each of e and f is a singleton in F . We now consider two cases, depending on whether A has one or zero breakpoints, and show that there exists another maximum agreement forest for T and T' that has the desired properties such that $\{a, b, c, d\}$ is eligible for Operation P.

First, suppose that $A = ef$. Let

$$F' = (F \setminus \{B, \{e\}, \{f\}\}) \cup \{\{a, b, e, f\}, \{c, d\}\}$$

be a forest. Noting that $|F'| < |F|$, it follows by the maximality of F that F' is not an agreement forest for T and T' . Hence, by construction of F' , there exists an element B' in $F' \setminus \{\{a, b, e, f\}\}$ such that $T[B']$ uses the edge $\{p_e, p_f\}$ in T . Let B'_1, B'_2 be a bipartition of B' such that neither $T[B'_1]$ nor $T[B'_2]$ uses the edge $\{p_e, p_f\}$ in T . As B' is also an element of F , it now follows that

$$F'' = (F \setminus \{B, B', \{e\}, \{f\}\}) \cup \{\{a, b, e, f\}, \{c, d\}, B'_1, B'_2\}$$

is an agreement forest for T and T' with $|F| = |F''|$ and in which $\{a, b\}$ and $\{c, d\}$ are both preserved, and $\{a, b, c, d\}$ is not preserved. Hence, $\{a, b, c, d\}$ is eligible for Operation P.

Second, suppose that $A = e|f$. Observe that (e, a, b, c, d) is a chain of T . Let

$$F' = (F \setminus \{B, \{e\}\}) \cup \{\{a, b, e\}, \{c, d\}\}$$

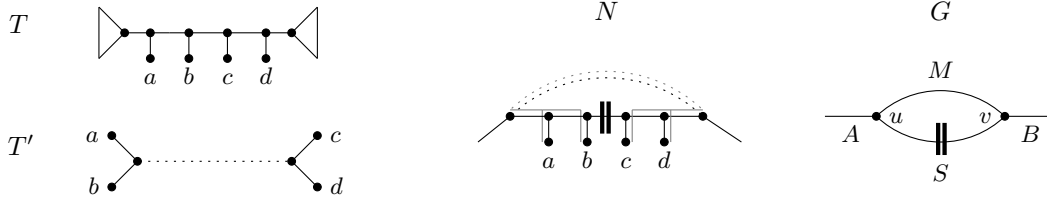


Fig. 7: The situation described in Theorem 11, which concerns $2|2$ sides $S = ab|cd$, where S is part of a multi-edge. The grey edges in N indicate an image of T' that passes through the path in N that is associated with M .

be a forest. Since there exists no element in $F \setminus \{B\}$ whose embedding in T uses p_e , F' is an agreement forest for T and T' . As $|F'| = |F|$ it now follows again that $\{a, b, c, d\}$ is eligible for Operation P. \square

Theorem 11. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. Furthermore, let $S = ab|cd$ be a side of G that is part of a multi-edge $\{u, v\}$. Let A, B , and M be the three sides incident with S such that M is incident with u and v , A is only incident with u , and B is only incident with v . Then $\{a, b, c, d\}$ is eligible for Operation P if each of the following conditions hold in obtaining N from G :*

1. *one of A and B is decorated with at least two taxa; and*
2. *M is decorated with at least one taxon.*

Proof. Assume without loss of generality that T' has cherries $\{a, b\}$ and $\{c, d\}$ and that T has a non-pendant chain (a, b, c, d) as illustrated in Figure 7. Then (a, b) and (c, d) are two common 2-chains of T and T' . Both of these chains are pendant in T' and therefore CPT eligible. For each $Y \in \{A, B, M\}$, let P_Y be the path associated with Y in N . Now, consider an image I of T' in N . Let P be the path from a to c in I . Then either P_A and P_B are subpaths of P , or P_M is a subpath of P . If P_A and P_B are subpaths of P then, since the first condition in the statement of the theorem is satisfied, we can apply the same argument as in the proof of Theorem 10 to establish that $\{a, b, c, d\}$ is eligible for Operation P. We may therefore assume that P_M is a subpath of P . Since T does not contain a cycle and the breakpoint on S is relative to T' , it follows that M has a single breakpoint that is relative to T .

Let Z be the set of taxa that is attached to M in obtaining N from G . First, assume that Z is CPT eligible. Then $|Z| \geq 2$, and there exists a maximum agreement forest F for T and T' that preserves each element in $\{\{a, b\}, \{c, d\}, Z\}$. If there exists an element B in F such that $\{a, b, c, d\} \subseteq B$, then B also contains Z since, otherwise, $T'[B]$ and $T'[B']$ are not vertex disjoint, where B' is the element in $F \setminus \{B\}$ such that $Z \subseteq B'$. Hence $B = B'$, thereby implying that $T[B] \neq T'[B]$, a contradiction. It now follows that $\{a, b, c, d\}$ is not a subset of any element in F and, thus, $\{a, b, c, d\}$ is eligible for Operation P. Second, assume that Z is not CPT eligible. Since the second condition in the statement of the theorem is satisfied, it follows from the fact that M is a 1-breakpoint side and from the contrapositive of Observation 4 that $1 \leq |Z| \leq 2$. Let F be a maximum agreement forest that preserves $\{a, b\}$

and $\{c, d\}$. Again assume that there exists an element B in F such that $\{a, b, c, d\} \subseteq B$. We next consider two cases.

First suppose that $Z = \{e, f\}$. Since Z is not CPT eligible, it follows that $M = e|f$. Without loss of generality, we assume that the path from p_a to p_e in T' does not pass through p_f . Since $T|B = T'|B$, it follows that $\{e\}$ and $\{f\}$ are elements in F . Observe that (e, a, b, c, d, f) is a chain of T . Now, let

$$F' = (F \setminus \{\{B, \{e\}, \{f\}\}\}) \cup \{\{a, b, e\}, \{c, d, f\}\}.$$

As F is an agreement forest for T and T' and each edge of P is used by $T'[B]$, F' is such a forest as well, contradicting the minimality of F . Hence B does not exist in F and $\{a, b, c, d\}$ is therefore eligible for Operation P.

Second suppose that $Z = \{e\}$. Then $M = |e$ or $M = e|$. Since $T|B = T'|B$, it follows that $\{e\}$ is an element in F . Observe that either (e, a, b, c, d) or (a, b, c, d, e) is a chain of T . Now, if (e, a, b, c, d) is a chain in T , let

$$F' = (F \setminus \{B, \{e\}\}) \cup \{\{a, b, e\}, \{c, d\}\}$$

and, if (a, b, c, d, e) is a chain in T , let

$$F' = (F \setminus \{B, \{e\}\}) \cup \{\{a, b\}, \{c, d, e\}\}.$$

As F is a maximum agreement forest for T and T' , it follows that, regardless of which case applies, F' is also such a forest. Thus, $\{a, b, c, d\}$ is eligible for Operation P. \square

6.2 2|1|1 sides

Theorem 12. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. Furthermore, let $S = ab|c|d$ be a side of G that is a simple edge $\{u, v\}$. Let A, B, C , and D be the four sides incident with S such that A and C are both incident with u , and B and D are both incident with v . If each of A and C is decorated with at least two taxa, or each of B and D is decorated with at least two taxa in obtaining N from G , then $\{a, b, c, d\}$ is eligible for Reduction 10.*

Proof. Assume without loss of generality that T has cherries $\{a, b\}$ and $\{c, d\}$ and that T' has a pendant 3-chain (a, b, c) with cherry $\{b, c\}$ as illustrated in Figure 8. Then the 2-chain (a, b) is CPT eligible. Let F be a maximum agreement forest for T and T' such that $\{a, b\}$ is preserved in F . Let B be the element in F with $\{a, b\} \subseteq B$. Then either $\{c\} \in F$ or $\{a, b, c\} \subseteq B$. Assume that the latter holds. Then, as $T|B = T'|B'$, we have $B = \{a, b, c\}$. We freely use this observation throughout the rest of the proof.

Now, for each $Y \in \{A, B, C, D\}$, let P_Y be the path associated with Y in N . Furthermore, let I be an image of T in N , and let P be the path from a to c in I . Since S is a 2-breakpoint side, either P_A or P_C is a subpath of P and, similarly, one of P_B or P_D is a subpath of P . We assume without loss of generality that P_A and P_B are both subpaths of P . It follows that T has no breakpoint on A or B and, hence each of A and B has at most one breakpoint relative to T .

Suppose that A is a 1-breakpoint side $A = |ef$ or $A = ef|$ that is decorated with exactly two taxa e and f , or that A is decorated with at least three taxa. Since A has at most one

breakpoint, it follows from Observation 4 that T and T' have a CPT-eligible chain Z whose elements are attached to A in obtaining N from G . Let $K = \{\{a, b\}, Z\}$. By applying the CPT to K , there exists a maximum agreement forest F' for T and T' such that each element in K is preserved in F' . Let B and B' be the elements of F' such that $\{a, b\} \subseteq B$ and $Z \subseteq B'$. Assume that $\{c\} \notin F'$. Then, by the observation in the first paragraph of the proof, we have $B = \{a, b, c\}$. Since $T[B]$ and $T[B']$ are vertex disjoint, it follows that $B = B'$. In turn, this implies that $T[(\{a, b, c\} \cup Z) \neq T'[(\{a, b, c\} \cup Z)]$, thereby contradicting that F' is an agreement forest for T and T' . Hence $\{c\} \in F'$ and, so, $\{a, b, c, d\}$ is eligible for Reduction 10. An identical analysis holds for when B is decorated with at least three taxa. Hence, one of A and B is decorated with exactly two taxa.

Now, reconsider F . If $\{c\} \in F$, then $\{a, b, c, d\}$ is clearly eligible for Reduction 10. We may therefore assume that $B = \{a, b, c\}$ and, consequently, $\{d\} \in F$. We next distinguish two cases and show that there always exists another maximum agreement forest for T and T' that has the desired property such that $\{a, b, c, d\}$ is eligible for Reduction 10.

- (1) Suppose that A is decorated with exactly two taxa e and f such that the path from p_e to p_a in T does not pass through p_f . By the definition of an agreement forest, it follows that $\{e\}$ and $\{f\}$ are elements of F . Recall that A has at most one breakpoint and that this breakpoint is, if it exists, relative to T' . Hence A is either a 0-breakpoint side $A = ef$ or a 1-breakpoint side $A = e|f$. First, if $A = ef$, let

$$F' = (F \setminus \{B, \{e\}, \{f\}\}) \cup \{\{a, b, e, f\}, \{c\}\}$$

be a forest. Noting that $|F'| < |F|$, it follows by the maximality of F that F' is not an agreement forest for T and T' . Hence, by construction of F' , there exists an element B' in $F' \setminus \{\{a, b, e, f\}\}$ such that $T'[B']$ uses the edge $\{p_e, p_f\}$ in T' . Let B'_1, B'_2 be a bipartition of B' such that neither $T'[B'_1]$ nor $T'[B'_2]$ uses the edge $\{p_e, p_f\}$ in T' . As B' is also an element of $F \setminus \{B, \{e\}, \{f\}\}$, it now follows that

$$F'' = (F \setminus \{B, B', \{e\}, \{f\}\}) \cup \{\{a, b, e, f\}, \{c\}, B'_1, B'_2\}$$

is another maximum agreement forest for T and T' in which $\{c\}$ is a singleton. Hence, $\{a, b, c, d\}$ is eligible for Reduction 10. Second, if $A = e|f$, let

$$F' = (F \setminus \{B, \{e\}\}) \cup \{\{a, b, e\}, \{c\}\}$$

be a forest. Since there exists no element in $F \setminus \{B\}$ whose embedding in T' uses p_e , F' is another maximum agreement forest for T and T' . It follows again that $\{a, b, c, d\}$ is eligible for Reduction 10.

- (2) Suppose that B is decorated with exactly two taxa e and f such that the path from p_e to p_d in T does not pass through p_f . As in Case (1), $\{e\}$ and $\{f\}$ are elements of F . Moreover, B is either a 0-breakpoint side $B = ef$ or a 1-breakpoint side $B = e|f$, where the breakpoint is relative to T' . First, if $B = ef$, let

$$F' = (F \setminus \{B, \{d\}, \{e\}, \{f\}\}) \cup \{\{a, b\}, \{c\}, \{d, e, f\}\}$$

be a forest. Noting that $|F'| < |F|$, it follows by the maximality of F that F' is not an agreement forest for T and T' . Hence, by construction of F' , there exists an element B' in $F' \setminus \{\{d, e, f\}\}$ such that $T'[B']$ uses the edge $\{p_e, p_f\}$ in T' . Let B'_1, B'_2 be a bipartition

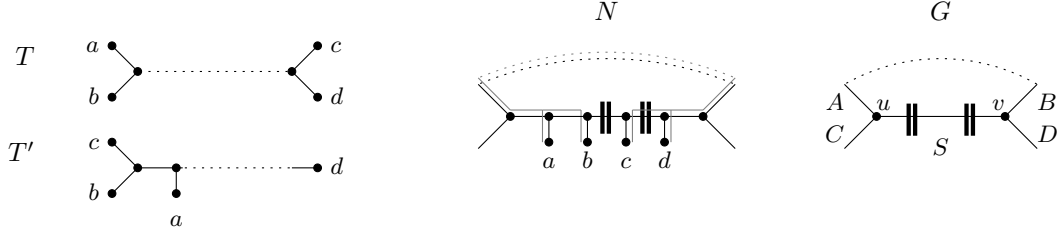


Fig. 8: The situation described in Theorem 12, which concerns $2|1|1$ sides $S = ab|c|d$, where S is not part of a multi-edge. The grey edges in N indicate an image of T .

of B' such that neither $T'[B'_1]$ nor $T'[B'_2]$ uses the edge $\{p_e, p_f\}$ in T' . As B' is also an element of $F \setminus \{B, \{d\}, \{e\}, \{f\}\}$, it now follows that

$$F'' = (F \setminus \{B, B', \{d\}, \{e\}, \{f\}\}) \cup \{\{a, b\}, \{c\}, \{d, e, f\}, B'_1, B'_2\}$$

is another maximum agreement forest for T and T' in which $\{c\}$ is a singleton. Hence, $\{a, b, c, d\}$ is eligible for Reduction 10. Second, if $B = e|f$, let

$$F' = (F \setminus \{B, \{d\}, \{e\}\}) \cup \{\{a, b\}, \{c\}, \{d, e\}\}.$$

Since there exists no element in $F \setminus \{B\}$ whose embedding in T' uses p_e , F' is another maximum agreement forest for T and T' . Thus $\{a, b, c, d\}$ is eligible for Reduction 10. \square

Theorem 13. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. Furthermore, let $S = ab|c|d$ be a side of G that is part of a multi-edge $\{u, v\}$. Let A , B , and M be the three sides incident with S such that M is incident with u and v , A is only incident with u , and B is only incident with v . Then $\{a, b, c, d\}$ is eligible for Reduction 10 if each of the following conditions hold in obtaining N from G :*

1. *one of A and B is decorated with at least two taxa; and*
2. *M is decorated with at least one taxon.*

Proof. Assume without loss of generality that T has cherries $\{a, b\}$ and $\{c, d\}$ and that T' has a pendant 3-chain (a, b, c) with cherry $\{b, c\}$ as illustrated in Figure 9. Then the 2-chain (a, b) is CPT eligible. Let F be a maximum agreement forest for T and T' such that $\{a, b\} \subseteq B$. As in the proof of Theorem 12, we can assume that, if $\{c\} \notin F$, then $B = \{a, b, c\}$.

Now, for each $Y \in \{A, B, M\}$, let P_Y be the path associated with Y in N . Furthermore, let I be an image of T in N , and let P be the path from a to c in I . Since S is a 2-breakpoint side, either P_A and P_B is a subpath of P , or P_M is a subpath of P . If P_A and P_B are subpaths of P then, since the first condition in the statement of the theorem is satisfied, we can apply the same argument as in the proof of Theorem 12 to establish that $\{a, b, c, d\}$ is eligible for Reduction 10. We may therefore assume that P_M is a subpath of P . As M does not have a breakpoint relative to T , M is a 0-breakpoint side or a 1-breakpoint side in which case the breakpoint is relative to T' .

Suppose that M is a 1-breakpoint side $M = |ef|$ or $M = ef|$ that is decorated with exactly two taxa e and f , or that M is decorated with at least three taxa. Since M has at most one

breakpoint, it follows from Observation 4 that T and T' have a CPT-eligible chain Z whose elements are attached to M in obtaining N from G . Applying the same argument as in the third paragraph of the proof of Theorem 12 establishes that $\{a, b, c, d\}$ is eligible for Reduction 10.

Since the second condition in the statement of the theorem holds, we complete the proof by considering two cases depending on whether M is decorated with one or two taxa. For both cases, reconsider F and assume that $B = \{a, b, c\}$. We will see that there exists another maximum agreement forest for T and T' that has the desired property such that $\{a, b, c, d\}$ is eligible for Reduction 10.

First suppose that M is decorated with only a single taxon e . Clearly, $\{d\}$ and $\{e\}$ are elements of F . If M is a 0-breakpoint side, let

$$F' = (F \setminus \{B, \{d\}, \{e\}\}) \cup \{\{a, b, d, e\}, \{c\}\}$$

be a forest. Since $|F'| < |F|$, it follows from the maximality of F that F' is not an agreement forest for T and T' . Hence there exists an element B' in F' (as well as in F) such that $T'[B']$ uses the two edges f and f' that are both incident with p_e and not incident with e . Let B'_1, B'_2 be a bipartition of B' such that neither $T'[B'_1]$ nor $T'[B'_2]$ uses f or f' . Then

$$F'' = (F \setminus \{B, B', \{d\}, \{e\}\}) \cup \{\{a, b, d, e\}, \{c\}, B'_1, B'_2\}$$

is an agreement forest for T and T' with $|F''| = |F|$ and, hence, $\{a, b, c, d\}$ is eligible for Reduction 10. On the other hand, if M is a 1-breakpoint side, then either (c, b, a, e) or (d, e) is a pendant chain of T' . In the former case, let

$$F' = (F \setminus \{B, \{e\}\}) \cup \{\{a, b, e\}, \{c\}\}$$

be a forest and, in the latter case let

$$F' = (F \setminus \{B, \{e\}, \{d\}\}) \cup \{\{a, b\}, \{c\}, \{d, e\}\}$$

be a forest. Regardless which applies, F' is an agreement forest and, again, $\{a, b, c, d\}$ is eligible for Reduction 10.

Second suppose that M is decorated with exactly two taxa e and f such that the path from p_a to p_e in T does not pass through p_f . Clearly, $\{e\}$ and $\{f\}$ are elements of F . If M is a 0-breakpoint side, let

$$F' = (F \setminus \{B, \{e\}, \{f\}\}) \cup \{\{a, b, e, f\}, \{c\}\}$$

be a forest. As usual, the size of F' contradicts the maximality of F . Hence, there exists an element B' in F' (as well as in F) such that $T'[B']$ uses an edge $\{p_e, p_f\}$. Let B'_1, B'_2 be a bipartition of B' such that neither $T'[B'_1]$ nor $T'[B'_2]$ uses $\{p_e, p_f\}$. Then

$$F'' = (F \setminus \{B, B', \{e\}, \{f\}\}) \cup \{\{a, b, e, f\}, \{c\}, B'_1, B'_2\}$$

is an agreement forest for T and T' with $|F''| = |F|$ and, hence, $\{a, b, c, d\}$ is eligible for Reduction 10.

Finally, if M is a 1-breakpoint side with $M = e|f$, let

$$F' = (F \setminus \{B, \{e\}\}) \cup \{\{a, b, e\}, \{c\}\}$$

be a forest. As F is an agreement forest for T and T' and (c, b, a, e) is a chain of T' , F' is such a forest as well. Thus, $\{a, b, c, d\}$ is eligible for Reduction 10. \square

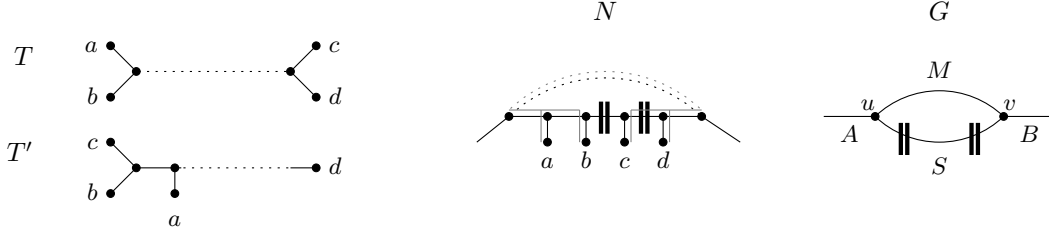


Fig. 9: The situation described in Theorem 13, which concerns $2|1|1$ sides $S = ab|c|d$, where S is part of a multi-edge. The grey edges in N indicate an image of T that uses the path in N that is associated with M .

Finally, we turn to loops. Consider two phylogenetic trees T and T' and a phylogenetic network N that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. Let S be a loop side of the generator G that underlies N . Then S is decorated with at least one taxon since, otherwise, there exists a phylogenetic network with strictly fewer than $r(N)$ reticulations that displays T and T' . Moreover if A denotes the side of G that is incident to S then, because T and T' are connected, A has no breakpoint and S has two breakpoints. Hence every loop is adjacent to a 0-breakpoint side. These observations as well as the next lemma, which shows that loop sides exhibit clean behavior in terms of being eligible for Reduction 10, will be convenient for the bounding argument of the next section.

Theorem 14. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–7. Let G be the generator underlying a phylogenetic network N on X that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. If $S = ab|c|d$ is a side of G that is a loop, then $\{a, b, c, d\}$ is eligible for Reduction 10.*

Proof. Assume without loss of generality that T has cherries $\{a, b\}$ and $\{c, d\}$ and that T' has a pendant 3-chain (a, b, c) with cherry $\{b, c\}$. Then (a, b) is a CPT-eligible 2-chain. Let F be a maximum agreement forest for T and T' such that $\{a, b\} \subseteq B$ for some element B in F . Assume that $\{c\} \notin F$. Then, as before, $c \in B$. Since $T|B = T'|B$, we have $B = \{a, b, c\}$ and, therefore, $\{d\} \in F$. It follows that

$$F' = (F \setminus \{B, \{d\}\}) \cup \{\{a, b, d\}, \{c\}\}$$

is an agreement forest for T . Moreover, as $|F'| = |F|$ and $\{c\} \in F'$, $\{a, b, c, d\}$ is eligible for Reduction 10. \square

7 Putting it all together and bounding the size of the kernel

In this section, we establish an improved kernel result for computing the TBR distance that is based on Reductions 1–10. We start by bounding the number of certain types of sides in a generator.

Lemma 4. *Let T and T' be two phylogenetic trees on X that cannot be reduced under Reductions 1–10, and let G be a generator that underlies a phylogenetic network N that displays T and T' such that $r(N) = d_{\text{TBR}}(T, T')$. Furthermore, let s_2 be the number of $2|2$ sides of G , each being decorated with four taxa that are eligible for Operation P, and let s_1 be the number of $1|3$ sides of G . Then $s_1 + s_2 \leq 1$.*

Proof. Suppose that $s_1 + s_2 \geq 2$. By Observation 3, we have $s_1 \leq 1$. If $s_1 = 1$ and $s_2 \geq 1$, then T and T' can be reduced by an application of Reduction 9.1. Hence, we may assume that $s_1 = 0$ and $s_2 \geq 2$. Let $S_1 = a'b'|c'd'$ and $S_2 = a''b''|c''d''$ be two 2|2 sides of G such that each of $\{a', b', c', d'\}$ and $\{a'', b'', c'', d''\}$ are both eligible for Operation P. We establish the lemma by showing that we can apply Reduction 9.2, thereby contradicting that T and T' cannot be reduced under Reductions 1–10. Since $\{a'', b'', c'', d''\}$ is eligible for Operation P before this operation is applied to $\{a', b', c', d'\}$, note first that S_2 satisfies the conditions in the statement of Theorem 10 or 11, depending on whether S_2 is part of a multi-edge or not. Now let S and S' be the two phylogenetic trees obtained from T and T' , respectively, by applying Operation P to $\{a', b', c', d'\}$. Crucially, by Corollary 1, N displays S and S' - as noted in the corollary the only change is that on side S_1 a breakpoint moves slightly. Furthermore, by Theorem 8, $d_{\text{TBR}}(T, T') = d_{\text{TBR}}(S, S')$ which implies that there exists no phylogenetic network that displays S and S' and has strictly fewer than $r(N)$ reticulations. It now follows that $\{a'', b'', c'', d''\}$ is still eligible for Operation P after this operation has been applied to $\{a', b', c', d'\}$ because S_2 still satisfies the conditions in the statement of Theorem 10 or 11, depending on whether S_2 is part of a multi-edge or not. \square

We next use a pessimistic, but safe, counting argument to finally bound the size of the kernel for computing the TBR distance.

Theorem 6. *Let T and T' be two phylogenetic trees on X with $d_{\text{TBR}}(T, T') \geq 2$ that cannot be reduced under Reductions 1–10. Then $|X| \leq 9d_{\text{TBR}}(T, T') - 8$.*

Proof. Let N be a phylogenetic network that displays T and T' such that

$$k = r(N) = d_{\text{TBR}}(T, T'),$$

and let G be the generator that underlies N . By [14, Lemma 1], G has $3k - 3$ sides. By Lemma 2, it immediately follows that each side of G is decorated with at most four taxa when obtaining N from G . Additionally, by the same lemma, each side that is decorated with four taxa is a 1|3, 2|2, or 2|1|1 side. Moreover, by Lemma 4, the number of 2|2 sides that are eligible for Operation P plus the number of 1|3 sides is at most one. We next use the results established in Section 6 to derive the following *adjacency rules* for sides of G that are decorated with four taxa.

- A1. From the contrapositives of Theorems 10 and 12, it follows that each side of G (with possibly one exception by Lemma 4) that is decorated with four taxa, and is not a loop or part of a multi-edge, is incident to at least two distinct sides that are each decorated with at most one taxon.
- A2. From the contrapositives of Theorems 11 and 13, it follows that each side of G (with possibly one exception by Lemma 4) that is decorated with four taxa and is part of a multi-edge is either incident to at least two distinct sides that are not part of the same multi-edge and each decorated with at most one taxon, or the second side in the multi-edge is decorated with zero taxa.
- A3. From the contrapositive of Theorem 14, it follows that each side of G that is a loop is decorated at most three taxa. Furthermore, each such loop side is incident to a 0-breakpoint side that, by Lemma 2(b), is decorated with at most three taxa.

We first deal with the exceptional situation that there is a single side S of G that is decorated with four taxa and does not obey A1 or A2. For the purpose of the upcoming

counting argument, we view S as a side that is only decorated with three taxa. This does not affect A1 or A2 because these rules only consider adjacent sides that are decorated with at most one taxon. Furthermore, recalling that S is not a 0-breakpoint side, viewing S as a side that is decorated with three taxa does not affect A3 either. To avoid an underestimate of the final kernel size, we add one to the counting formula below. Next, we consider each side S of G that is decorated with zero or two taxa and view it in one of the following ways for counting purposes.

1. If S is not a loop, not part of a multi-edge, and decorated with zero taxa, we view S as a side that is decorated with one taxon. This does not affect A1–A3.
2. If S is part of a multi-edge and decorated with zero taxa, we view S as a side that is decorated with three taxa and, if subsequently any side S' of G that is incident to S is decorated with four taxa, then we view S' as a side that is decorated with three taxa. This cannot decrease the total number of taxa because S is incident to at most three sides that are each decorated with four taxa. Also, we still obey A1–A3, because any side decorated with four taxa that needed S as a side decorated with zero taxa is now viewed as a side decorated with three taxa.
3. If S is decorated with two taxa, we view S as a side that is decorated with three taxa. Again, this does not affect any of A1–A3.

Now we still have a valid upper bound on the total number of taxa that decorate sides of G , but a simplified counting system because every side is decorated with four, three, or one taxa. Let p , q , and r be the number of sides of G that are decorated with with four, three, and one taxa respectively. Then we have the following optimization problem, where the $+1$ in the objective function is due to the possibly undercounted side decorated with four taxa that is mentioned above and that we view as a side decorated with three taxa.

Maximize $4p + 3q + 1r + 1$
subject to
 $p+q+r = 3k - 3$
 $p \leq 2k$
 $r \geq (2/4)p$ and
 $p, r, q \geq 0$ (and integer)

The $p \leq 2k$ inequality occurs because, by Lemma 2, a side that is decorated with four taxa has at least one breakpoints and there are $2k$ breakpoints in total (i.e., k breakpoints for each tree). Furthermore, each side that is decorated with one taxon can be incident to at most four sides that are each decorated with four taxa. On the other hand, since T and T' cannot be further reduced under any of Reductions 1–10, each side that is decorated with four taxa needs to be incident to at least two sides decorated with one taxon. This implies that $r \geq (2/4)p$. We next substitute $q = (3k - 3) - p - r$ and this gives

Maximize $9k + p - 2r - 8$
subject to
 $p \leq 2k$
 $r \geq (1/2)p$ and
 $p, r, q \geq 0$ (and integer).

The fact that $r \geq (1/2)p$ implies that the term $(p - 2r)$ in the objective function is at most 0. We conclude that $|X| \leq 9k - 8 = 9d_{\text{TBR}}(T, T') - 8$ is an upper bound on the size of our kernel. \square

The bound $9k - 8$ is tight up to an additive term of 1, as the following theorem shows. The additive term is due, in the above analysis, to the at most one generator side with four taxa that does not obey A1 or A2.

Theorem 15. *For each $k \geq 3$ there exist two phylogenetic trees T_k and T'_k with $9k - 9$ taxa and $d_{\text{TBR}}(T_k, T'_k) = k$ that cannot be reduced under Reductions 1–10.*

Proof. Let $k \geq 3$. We proceed by building a specific ladder-like generator G_k , converting this to a phylogenetic network N_k , and extracting the trees T_k and T'_k from this. We will then prove that $d_{\text{TBR}}(T_k, T'_k) = k$ and that they are irreducible under Reductions 1–10.

Generator G_k is built as follows. We take the rectangular $2 \times (k + 1)$ grid on $2(k + 1)$ nodes and suppress the four corner vertices of degree 2. This creates a cubic multigraph G_k with $3(k - 1)$ sides. Note that G_k has exactly two pairs of multi-edges. We create N_k by decorating each side of G_k with 3 taxa. Let X_k be the set of all taxa added; we have $|X_k| = 9k - 9$. By construction, $r(N_k) = k$. See Figure 10 for the situation $k = 5$. Let T_k (respectively, T'_k) be the tree displayed by N_k that is induced by the k solid (respectively, hollow) breakpoints as indicated in the figure. Given that $r(N_k) = k$ we have $d_{\text{TBR}}(T_k, T'_k) \leq k$.

To prove that $d_{\text{TBR}}(T_k, T'_k) \geq k$ we use the same lower-bounding technique as [14,15]. Specifically, a *binary character* f on X is a function that assigns each element in X to an element in $\{0, 1\}$. Let T be an unrooted binary phylogenetic X -tree with vertex set V . An *extension* g of f to V is a function g that assigns each element in V to an element in $\{0, 1\}$ such that $g(x) = f(x)$ for each $x \in X$. The *parsimony score* of f on T , denoted by $l_f(T)$, denotes the minimum number of edges $\{u, v\}$ in T such that $g(u) \neq g(v)$, ranging over all extensions of f . Now, for two unrooted binary phylogenetic X -trees T and T' , the *maximum parsimony distance on binary characters* d_{MP}^2 is defined as $d_{\text{MP}}^2(T, T') = \max_f |l_f(T) - l_f(T')|$ where f ranges over all binary characters on X . It is well-known that $d_{\text{TBR}}(T, T') \geq d_{\text{MP}}^2(T, T')$ [8]. Hence, to prove that $d_{\text{TBR}}(T_k, T'_k) \geq k$ it is sufficient to give a binary character f on X_k such that $|l_f(T_k) - l_f(T'_k)| \geq k$. We define f by assigning 0 to each taxon to the left of the grey line, as indicated in Figure 10, and assigning 1 to all other taxa, i.e. those to the right of the grey line. It is easy to check that $l_f(T_k) = 1$ and that, by Fitch's algorithm or similar [9], $l_f(T'_k) \geq (k + 1)$, so $d_{\text{TBR}}(T_k, T'_k) \geq |l_f(T_k) - l_f(T'_k)| \geq k$ as required. This concludes the proof that $d_{\text{TBR}}(T_k, T'_k) = k$.

Regarding irreducibility, it is helpful to first inventarise some topological features of T_k and T'_k . They have no common pendant subtrees of size 2 or larger, so Reduction 1 is excluded, and they have no common chains of length 4 or longer, so Reduction 2 is excluded. Crucially, each tree has exactly one pendant 3-chain but this is *not* common with the other tree. For T_k this is (p, q, r) , where $\{q, r\}$ is its cherry, and for T'_k this is (s, t, u) , where $\{t, u\}$ is its cherry. Hence, Reductions 3, 4, 6 and 7 are excluded. Recalling the definition of Reduction 5, we see that if the preconditions for this reduction rule hold, then it also follows that (ℓ_1, ℓ_2, x) is a pendant 3-chain in one tree (with $\{\ell_2, x\}$ the cherry) and (ℓ_1, ℓ_2) is a 2-chain common to both trees. As noted already each tree has exactly one pendant 3-chain. Without loss of generality (due to symmetry between T_k and T'_k), observe that the single pendant 3-chain (p, q, r) in T_k , where $\{q, r\}$ is the cherry, has the property that (p, q) is *not* a 2-chain in T'_k , so Reduction 5 cannot apply.

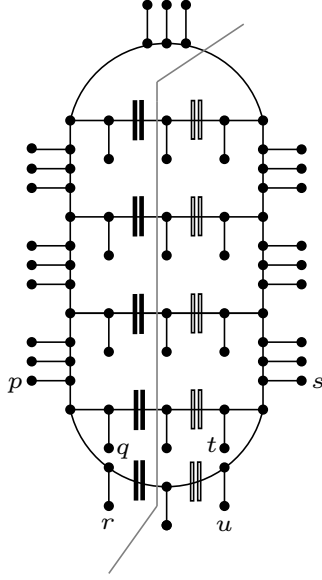


Fig. 10: The phylogenetic network N_k for $k = 5$, as constructed in the proof of Theorem 15. The k solid double bars represent the breakpoints for the first tree T_k and the k hollow double bars represent the breakpoints for the second tree T'_k . Taxa p , q , and r form the unique pendant 3-chain of T_k and s , t , and u form the unique pendant 3-chain of T'_k . The grey line is used in the proof that $d_{\text{TBR}}(T_k, T'_k) \geq k$.

We now turn to the new reduction rules. Consider Reduction 8. This is built on Reduction 8A, which requires a common 3-chain that is pendant in one tree: again, this does not exist, so the reduction is excluded. The same fact immediately excludes Reduction 9.1. Reduction 9.2 requires Operation P to execute, and this operation requires one of the trees to have a non-pendant 4-chain (a, b, c, d) and the other tree to have cherries $\{a, b\}$ and $\{c, d\}$. Each of T_k and T'_k has exactly $(k + 1)$ cherries, but no pair of these cherries combine to form a 4-chain in the other tree, so Operation P cannot apply. (Viewed from the contrapositive perspective: any 4-chain must contain at least one taxon that is not in a cherry in the other tree). Finally, consider Reduction 10. The preconditions here require one tree to have a pendant 3-chain (a, b, c) where $\{b, c\}$ is the cherry, and the other tree to have the cherry $\{a, b\}$. However, the single pendant 3-chain in (without loss of generality) T_k , (p, q, r) where $\{q, r\}$ is the cherry has the property that the first taxon p on the chain is definitely not in a cherry in the other tree, so Reduction 10 cannot execute. We are done. \square

8 Conclusion and future work

There are a number of interesting future research directions. The most obvious direction is to design new reduction rules capable of further improving the current $9k - 8$ bound. How far below $9k - 8$ can we go, and what is the trade off between proof complexity and the obtained decrease in kernel size? Specifically, the existing reduction rules and their associated proofs are already rather complex, requiring extensive auxiliary mathematical machinery and quite some case-checking. It is natural to ask whether the design of further rules and the related proofs can be streamlined and simplified in some fashion by deepening our understanding of

the combinatorial behaviour of agreement forests. In the meantime (semi-)automated tools for proof verification could be utilized to help keep case-checking under control. We note also that Reduction 8 hints at a wider family of reduction rules. Essentially, it gives us a general recipe for moving certain edges around in the trees such that d_{TBR} is preserved: this allows us to rearrange the trees in such a way that *other* reduction rules are triggered. We expect that such ‘indirect’ reduction rules will be very useful in the future.

Next, an empirical study in the spirit of [23] could investigate how much extra reductive power the new $9k - 8$ rules have in practice; the rules for the $11k - 9$ kernel do have more practical effect than the $15k - 9$ rules, does this trend continue? Another angle to explore is to translate the new reduction rules onto other agreement-forest based phylogenetic distances to obtain smaller kernels there; this has already been effective in designing new reduction rules for Rooted Subtree Prune and Regraft distance [16].

9 Acknowledgements

Steven Kelk and Simone Linz were supported by the New Zealand Marsden Fund. Ruben Meuwese was supported by the Dutch Research Council (NWO) KLEIN 1 grant *Deep kernelization for phylogenetic discordance*, project number OCENW.KLEIN.305.

References

1. B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–15, 2001.
2. R. Atkins and C. McDiarmid. Extremal distances for subtree transfer operations in binary trees. *Annals of Combinatorics*, 23(1):1–26, 2019.
3. L. Bulteau and M. Weller. Parameterized algorithms in bioinformatics: an overview. *Algorithms*, 12(12):256, 2019.
4. J. Chen, J.-H. Fan, and S.-H. Sze. Parameterized and approximation algorithms for maximum agreement forest in multifurcating trees. *Theoretical Computer Science*, 562:496–512, 2015.
5. M. Cygan, F. Fomin, L. Kowalik, D. Lokshtanov, D. Marx, M. Pilipczuk, M. Pilipczuk, and S. Saurabh. *Parameterized Algorithms*. Springer Publishing Company, Incorporated, 1st edition, 2015.
6. R. Downey and M. Fellows. *Fundamentals of parameterized complexity*, volume 4. Springer, 2013.
7. J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Incorporated, 2004.
8. M. Fischer and S. Kelk. On the Maximum Parsimony distance between phylogenetic trees. *Annals of Combinatorics*, 20(1):87–113, 2016.
9. W. M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.
10. F. Fomin, D. Lokshtanov, S. Saurabh, and M. Zehavi. *Kernelization: Theory of Parameterized Preprocessing*. Cambridge University Press, 2019.
11. J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71(1-3):153–169, 1996.
12. D. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2011.
13. K. St John. The shape of phylogenetic treespace. *Systematic Biology*, 66(1):e83, 2017.
14. S. Kelk and S. Linz. A tight kernel for computing the tree bisection and reconnection distance between two phylogenetic trees. *SIAM Journal on Discrete Mathematics*, 33(3):1556–1574, 2019.
15. S. Kelk and S. Linz. New reduction rules for the tree bisection and reconnection distance. *Annals of Combinatorics*, 24(3):475–502, 2020.
16. S. Kelk, S. Linz, and R. Meuwese. Cyclic generators and an improved linear kernel for the rooted subtree prune and regraft distance. *arXiv preprint arXiv:2202.09904*, 2022.
17. S. Kelk, L. van Iersel, C. Scornavacca, and M. Weller. Phylogenetic incongruence through the lens of monadic second order logic. *Journal of Graph Algorithms and Applications*, 20(2):189–215, 2016.

18. D. Money and S. Whelan. Characterizing the phylogenetic tree-search problem. *Systematic Biology*, 61(2):228, 2012.
19. N. Olver, F. Schalekamp, S. van der Ster, L. Stougie, and A. van Zuylen. A duality based 2-approximation algorithm for maximum agreement forest. *Mathematical Programming*, pages 1–43, 2022.
20. C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.
21. A. Som. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, 16(3):536–548, 2015.
22. L. van Iersel, S. Kelk, G. Stamoulis, L. Stougie, and O. Boes. On unrooted and root-uncertain variants of several well-known phylogenetic network problems. *Algorithmica*, 80(11):2993–3022, 2018.
23. R. van Wersch, S. Kelk, S. Linz, and G. Stamoulis. Reflections on kernelizing and computing unrooted agreement forests. *Annals of Operations Research*, 309(1):425–451, 2022.
24. C. Whidden, R. G. Beiko, and N. Zeh. Fixed-parameter algorithms for maximum agreement forests. *SIAM Journal on Computing*, 42(4):1431–1466, 2013.
25. C. Whidden, N. Zeh, and R. G. Beiko. Supertrees based on the subtree prune-and-regraft distance. *Systematic Biology*, 63(4):566–581, 2014.

Appendix

A Proof of Theorem 5

Proof of Theorem 5. Let $e = \{u, v\}$ be the interrupter in T' of C , and let v be the unique common neighbor of p_b and p_c in T' . Towards a contradiction assume that the result does not hold. Let F^* be a maximum agreement forest for T and T' . Then there exists an element $B \in F^*$ such that $T'[B]$ uses e . Let Q, R be the bipartition of $X \setminus \{a, b, c, d\}$ such that, in T , the path from each element in Q to a is shorter than its path to d , and the path from each element in R to d is shorter than its path to a . Similarly, let Q', R', S' be the tripartition of $X \setminus \{a, b, c, d\}$ such that, in T' , the path from each element in Q' to a is shorter than its path to d , the path from each element in R' to d is shorter than its path to a , and the path from each element in S' to a has the same length than its path to d . This setup is illustrated in Figure 4. We next define five sets that will be useful throughout the proof. Specifically, let $B_Q = B \cap Q$, $B_R = B \cap R$, $B_{Q'} = B \cap Q'$, $B_{R'} = B \cap R'$, and $B_{S'} = B \cap S'$. As $T'[B]$ uses e , note that $B_{S'}$ is non empty. Moreover, since $T[B] = T'[B]$, it follows that $|B \cap C| \leq 3$. We freely use the previous two properties of $B_{S'}$ and $B \cap C$, respectively, throughout the remainder of the proof. To establish the result, we next consider three cases that each have several subcases. In all (sub)cases we will show that there exists a maximum agreement forest F that does not use e .

Case 1. $B_{Q'} = \emptyset$, and $B_{R'} = \emptyset$

Since $T'[B]$ uses e , we have $1 \leq |B \cap C| \leq 3$. Furthermore, there is no element in $F^* \setminus \{B\}$ that contains an element of Q' and an element of R' . However, if $B_{S'} \subseteq Q$ or $B_{S'} \subseteq R$, then there can be an element in $F^* \setminus B$ that has a non-empty intersection with C and a non-empty intersection with one of Q and R . There are three subcases to consider for Case 1.

First suppose that $B \cap \{a, b\} \neq \emptyset$ and $B \cap \{c, d\} \neq \emptyset$. If $|B \cap C| = 2$, then $B_{S'} \subseteq Q$ or $B_{S'} \subseteq R$. On the other hand, if $|B \cap C| = 3$ then, because $T[B] = T'[B]$, we have $B_{S'} \subseteq R$ when $\{a, b\} \subset B$, and $B_{S'} \subseteq Q$ when $\{c, d\} \subset B$. Considering $T[B]$, it follows that there exists an element $\ell \in C \setminus B$ such that $\{\ell\} \in F^*$. Hence

$$F = (F^* \setminus \{B, \{\ell\}\}) \cup \{B_{S'}, (B \setminus B_{S'}) \cup \{\ell\}\}$$

is an agreement forest for T and T' with $|F| = |F^*|$.

Second suppose that $B \cap \{a, b\} = \emptyset$ and $|B \cap \{c, d\}| = 2$. Then again $B_{S'} \subseteq Q$ or $B_{S'} \subseteq R$. Let B' be the element in F^* that contains b . Note that $|B'| \geq 1$ if $B_{S'} \subseteq R$ and $|B'| = 1$ if $B_{S'} \subseteq Q$. Moreover if B' contains an element in $X \setminus C$, then $B' \setminus C \subseteq Q$ and $B' \setminus C \subseteq Q'$. Hence

$$F = (F^* \setminus \{B, B'\}) \cup \{B_{S'}, (B \setminus B_{S'}) \cup B'\}$$

is an agreement forest for T and T' with $|F| = |F^*|$. An analogous symmetric analysis applies when $|B \cap \{a, b\}| = 2$ and $B \cap \{c, d\} = \emptyset$.

Third suppose that $B \cap \{a, b\} = \emptyset$ and $|B \cap \{c, d\}| = 1$. If $\{c, d\} \cap B = \{c\}$, let B' be the element in F^* that contains b , and if $\{c, d\} \cap B = \{d\}$, let $B' = \{c\}$. In the latter case, note that $B' \in F^*$ because of $T'[B]$. Now, under the assumption that $B_Q = \emptyset$ or $B_R = \emptyset$, it follows that

$$F = (F^* \setminus \{B, B'\}) \cup \{B_{S'}, (B \setminus B_{S'}) \cup B'\}$$

is an agreement forest for T and T' with $|F| = |F^*|$. We may therefore assume that $B_Q \neq \emptyset$ and $B_R \neq \emptyset$, that is $B_{S'} = B_Q \cup B_R$. Then, because of $T[B]$, there are three singletons $\{\ell\}, \{\ell'\},$ and $\{\ell''\}$ in F^* such that $\{\ell, \ell', \ell''\} = C \setminus B$. In other words, there is no element in $F^* \setminus \{B\}$ that contains an element in C and an element not in C . Hence

$$F = (F^* \setminus \{B, \{\ell\}, \{\ell'\}, \{\ell''\}\}) \cup \{B_Q, B_R, (B \setminus B_{S'}) \cup \{\ell, \ell', \ell''\}\}$$

is an agreement forest for T and T' with $|F| < |F^*|$. An analogous symmetric analysis applies when $|B \cap \{a, b\}| = 1$ and $B \cap \{c, d\} = \emptyset$.

Case 2. $B_I = \emptyset$ and $B_J \neq \emptyset$ with $\{I, J\} = \{Q', R'\}$

Without loss of generality, we may assume that $B_{Q'} = \emptyset$ and $B_{R'} \neq \emptyset$. There are four subcases to consider for Case 2.

First suppose that $B \cap \{a, b\} \neq \emptyset$ and $B \cap \{c, d\} \neq \emptyset$. Since $T|B = T'|B$, it follows that $B_{S'} = B_Q$ and $B_{R'} = B_R$. Moreover, if $\{a, b\} \subset B$, then $T|B \neq T'|B$, which implies that there exists an element $\ell \in \{a, b\}$ such that $\{\ell\} \in F^*$. Hence

$$F = (F^* \setminus \{B, \{\ell\}\}) \cup \{B_{S'}, (B \setminus B_{S'}) \cup \{\ell\}\}$$

is an agreement forest for T and T' with $|F| = |F^*|$.

Second suppose that $B \cap \{a, b\} \neq \emptyset$ and $B \cap \{c, d\} = \emptyset$. Then clearly $\{c\}, \{d\} \in F^*$. Since $B \cap C \neq \emptyset$, there exists no element in $F^* \setminus \{B\}$ that contains an element of Q and an element of R . Moreover, if $B \cap C \in \{\{a\}, \{a, b\}\}$, then no element in $F^* \setminus \{B\}$ contains an element in C and an element in $X \setminus C$. Hence,

$$F = (F^* \setminus \{B, \{c\}, \{d\}\}) \cup \{B_{S'}, B_{R'}, (B \cap C) \cup \{c, d\}\}$$

is an agreement forest for T and T' with $|F| = |F^*|$. For the remainder of this subcase, assume that $B \cap \{a, b\} = \{b\}$. If $B_Q \neq \emptyset$ then $\{a\} \in F^*$ and F is again an agreement forest for T and T' . Lastly, if $B_Q = \emptyset$, let B' be the element in $F^* \setminus \{B\}$ such that $a \in B'$. As $T'[B]$ and $T'[B']$ are vertex disjoint, we have $B' \setminus \{a\} \subseteq Q'$. It is now straightforward to check that F is an agreement forest for T and T' .

Third suppose that $B \cap \{a, b\} = \emptyset$ and $B \cap \{c, d\} \neq \emptyset$. If $B \cap \{c, d\} = \{c, d\}$, then $B_{S'} = B_Q$ and $B_{R'} = B_R$. It follows that $\{a\}, \{b\} \in F^*$. On the other hand, if $B \cap \{c, d\} = \{c\}$ (resp. $B \cap \{c, d\} = \{d\}$), then $\{d\} \in F^*$ (resp. $\{c\} \in F^*$). Hence,

$$F = (F^* \setminus \{B, \{\ell\}\}) \cup \{B_{S'}, (B \setminus B_{S'}) \cup \{\ell\}\}$$

is an agreement forest for T and T' with $|F| = |F^*|$ and where $\ell \in \{b, c, d\}$ depending on which of the three elements is a singleton in F^* .

Fourth suppose that $B \cap C = \emptyset$. Clearly, $\{c\}, \{d\} \in F^*$. If there exists no $B' \in F^*$ such that $B' \cap (Q \cup \{a, b\}) \neq \emptyset$ and $B' \cap R \neq \emptyset$, then

$$F = (F^* \setminus \{B, \{c\}, \{d\}\}) \cup \{B_{S'}, B \setminus B_{S'}, \{c, d\}\}$$

is an agreement forest for T and T' with $|F| = |F^*|$. Hence, we may assume that B' exists. Furthermore, assume first that $B' \neq B$. Since $T|B' = T'|B'$, one of the following properties holds depending on which of a and b is contained in B' .

1. If $B' \cap \{a, b\} = \emptyset$, then $\{a\}, \{b\} \in F^*$ because $T[B']$ uses $\{p_a, p_b\}$.

2. If $B' \cap \{a, b\} = \{a\}$, then $\{b\} \in F^*$ because $T[B']$ uses $\{p_a, p_b\}$.
3. If $B' \cap \{a, b\} = \{b\}$, then $\{a\} \in F^*$ because $T'[B]$ and $T'[B']$ are vertex disjoint.
4. If $B' \cap \{a, b\} = \{a, b\}$, then $B' \cap Q = \emptyset$ because $T'[B]$ and $T'[B']$ are vertex disjoint.

It now follows that

$$F = (F^* \setminus \{B, B', \{a\}, \{b\}, \{c\}, \{d\}\}) \cup \{B_{S'}, B \setminus B_{S'}, C, B' \cap Q, B' \cap R\}$$

is an agreement forest for T and T' with $|F| < |F^*|$ if Property (1) applies,

$$F = (F^* \setminus \{B, B', \{\ell\}, \{c\}, \{d\}\}) \cup \{B_{S'}, B \setminus B_{S'}, C, B' \cap Q, B' \cap R\}$$

is an agreement forest for T and T' with $|F| \leq |F^*|$ and $\ell = b$ (resp. $\ell = a$) if Property (2) (resp. Property (3)) applies, and

$$F = (F^* \setminus \{B, B', \{c\}, \{d\}\}) \cup \{B_{S'}, B \setminus B_{S'}, C, B' \cap R\}$$

is an agreement forest for T and T' with $|F| = |F^*|$ if Property (4) applies. Now assume that $B' = B$. Clearly $\{a\}, \{b\} \in F^*$. Consider the bipartition $B_{S'}, B_{R'}$ of B . Let $B_{S'}^Q = B_{S'} \cap Q$, $B_{S'}^R = B_{S'} \cap R$, $B_{R'}^Q = B_{R'} \cap Q$, and $B_{R'}^R = B_{R'} \cap R$. It now follows that

$$F = (F^* \setminus \{B, \{a\}, \{b\}, \{c\}, \{d\}\}) \cup \{B_{S'}^Q, B_{S'}^R, B_{R'}^Q, B_{R'}^R, C\}$$

is an agreement forest for T and T' . In particular, since $T|B = T'|B$, at least one element in $\{B_{S'}^Q, B_{S'}^R, B_{R'}^Q, B_{R'}^R\}$ is the empty set and, so, $|F| < |F^*|$.

Case 3. $B_{Q'} \neq \emptyset$ and $B_{R'} \neq \emptyset$

Since $T|B = T'|B$, we have that $B \cap \{a, b\} = \emptyset$ or $B \cap \{c, d\} = \emptyset$. Hence $|B \cap C| \leq 2$. There are three subcases to consider for Case 3.

First suppose that $|B \cap C| = 2$. Then there exist two distinct element $\ell, \ell' \in C$ such that $\{\ell\}, \{\ell'\} \in F^*$. If $B \cap C = \{a, b\}$ (resp. $B \cap C = \{c, d\}$), then $B_{S'} \cup B_{R'} \subseteq R$ and $B_{Q'} \subseteq Q$ (resp. $B_{S'} \cup B_{Q'} \subseteq Q$ and $B_{R'} \subseteq R$). Hence,

$$F = (F^* \setminus \{B, \{\ell\}, \{\ell'\}\}) \cup \{B_{S'}, (B \setminus B_{S'}) \cup \{\ell, \ell'\}\}$$

is an agreement forest for T and T' with $|F| < |F^*|$.

Second suppose that $|B \cap C| = 1$. Then there exist three distinct element $\ell, \ell', \ell'' \in C$ such that $\{\ell\}, \{\ell'\}, \{\ell''\} \in F^*$. In turn, because there is no element in $F^* \setminus B$ that contains an element in Q and an element in R , this implies that, except for possibly B , no other element in F^* uses any of the three edges $\{p_a, p_b\}$, $\{p_b, p_c\}$, and $\{p_c, p_d\}$. Lastly, since $B \cap C \neq \emptyset$, $B_{S'}$ is either contained in Q or R . It now follows that

$$F = (F^* \setminus \{B, \{\ell\}, \{\ell'\}, \{\ell''\}\}) \cup \{B_{Q'}, B_{R'}, B_{S'}, C\}$$

is an agreement forest for T and T' with $|F| = |F^*|$.

Third suppose that $|B \cap C| = 0$. Then each element in C is a singleton in F^* . If there exists no element $B' \in F^*$ such that $T[B']$ uses an edge $\{p_\ell, p_{\ell'}\}$ for two distinct $\ell, \ell' \in C$, then

$$F = (F^* \setminus \{B, \{a\}, \{b\}, \{c\}, \{d\}\}) \cup \{B_{Q'}, B_{R'}, B_{S'}, C\}$$

is an agreement forest for T and T' with $|F| < |F^*|$. Otherwise, if B' exists, then B' is the unique such element since $B' \cap Q \neq \emptyset$ and $B' \cap R \neq \emptyset$. Hence, assuming that $B' \neq B$,

$$F = (F^* \setminus \{B, B', \{a\}, \{b\}, \{c\}, \{d\}\}) \cup \{B_{Q'}, B_{R'}, B_{S'}, B' \cap Q, B' \cap R, C\}$$

is an agreement forest for T and T' with $|F| = |F^*|$. Lastly, if $B' = B$, consider the three sets $B_{Q'}$, $B_{R'}$, and $B_{S'}$. Since $T|B = T'|B$, at most one of these three sets, say $B_{S'}$, has a non-empty intersection with Q and a non-empty intersection with R . Then

$$F = (F^* \setminus \{B, \{a\}, \{b\}, \{c\}, \{d\}\}) \cup \{B_{Q'}, B_{R'}, B_{S'} \cap Q, B_{S'} \cap R, C\}$$

is an agreement forest for T and T' with $|F| = |F^*|$. An analogous argument holds if $B_{Q'}$ or $B_{R'}$ has a non-empty intersection with Q and a non-empty intersection with R . \square

B Explicit descriptions of Algorithm 1 and 2 for testing eligibility for Operation P or Reduction 10.

Reductions 9 and 10 rely on Algorithms 1 and 2 to test eligibility. These algorithms do not have access to the underlying generator and have to search for the corresponding structures in two phylogenetic trees. The algorithms closely mirror the analyses in the proofs of Theorems 10–14.

B.1 Algorithm 1 tests whether $\{a, b, c, d\}$ is eligible for Operation P

Assume that T' has cherries $\{a, b\}$ and $\{c, d\}$ and T has a non-pendant chain (a, b, c, d) where a and d are the outermost leaves on the chain. This can easily be confirmed in polynomial time.

If at least one of the following polynomial-time checkable conditions is true, return YES i.e. Operation P can be applied. If none of them are true, return NO/DON'T KNOW.⁶

1. In T' , the path from p_a to p_c passes through at least one CPT-eligible chain Z where $Z \cap \{a, b, c, d\} = \emptyset$.
2. Any of situations (a)–(g) from Figure 11 occur.

Step 1 captures the parts of Theorems 10 and 11 when the path P , passing through side A , B or M depending on the situation, contains at least one CPT-eligible chain. (It does not matter if the chain found does not actually lie on A , B or M : it is still correct in this case to conclude that Operation P is eligible). Situation (a) of Step 2 covers the situation in Theorem 10 when the side A has 0 breakpoints and two taxa e and f . Situation (b) is when side A has the form $e|f$. Situations (c) and (d) are symmetrical to (a) and (b): when the path P uses side B rather than A . Situations (e)–(g) concern the cases in Theorem 11 where M is a side $e|f$, $e|$ or $|e$ respectively (and the breakpoint is with respect to T).

⁶ We write NO/DON'T KNOW because it might still be possible that $\{a, b, c, d\}$ is eligible for Operation P but for reasons that fall outside the conditions described in Theorems 10 and 11 (which are those checked by the algorithm). However, we do not care about such cases. Functionally speaking a NO/DON'T KNOW answer is therefore interpreted simply as NO. The same comment holds for Algorithm 2.

B.2 Algorithm 2 tests whether $\{a, b, c, d\}$ is eligible for Reduction 10

Assume without loss of generality that T' has a pendant 3-chain (a, b, c) where $\{b, c\}$ is the cherry, and T has two cherries $\{a, b\}$ and $\{c, d\}$. This can easily be confirmed in polynomial time.

If at least one of the following polynomial-time checkable conditions is true, return YES i.e. Reduction 10 can be applied. If none of them are true, return NO/DON'T KNOW.

1. In T , the path from p_a to p_c passes through at least one CPT-eligible chain Z where $Z \cap \{a, b, c, d\} = \emptyset$.
2. Any of situations (a)–(j) from Figure 12 occur.

Step 1 captures the parts of Theorems 12 and 13 when the path P , passing through side A , B or M depending on the situation, contains at least one CPT-eligible chain. (It does not matter if the chain found does not actually lie on A , B or M : it is still correct in this case to conclude that Reduction 10 is eligible). Situation (a) of Step 2 covers the situation in Theorem 12 when side A is a 0-breakpoint side with two taxa e and f , and (b) when A is a side $e|f$. Situation (c) covers the situation when side B is a 0-breakpoint side with taxa e and f , and situation (d) when side B is a $e|f$ side; note that situations (c) and (d) are not entirely symmetrical to situations (a) and (b) due to the inherent asymmetry of 2|1|1 sides. Situations (e)–(i) concern Theorem 13. In particular, (e) is when M is a 0-breakpoint side with two taxa e and f and (f) is when M is a 1-breakpoint side $e|f$ (where the breakpoint is with respect to T'). Situation (g) is when M is a 0-breakpoint side with only one taxon e , (h) is when M is a side $e|$, and (i) is when M is a side $|e$. Again, the breakpoints here are with respect to T' . Situation (j) reflects Theorem 14.

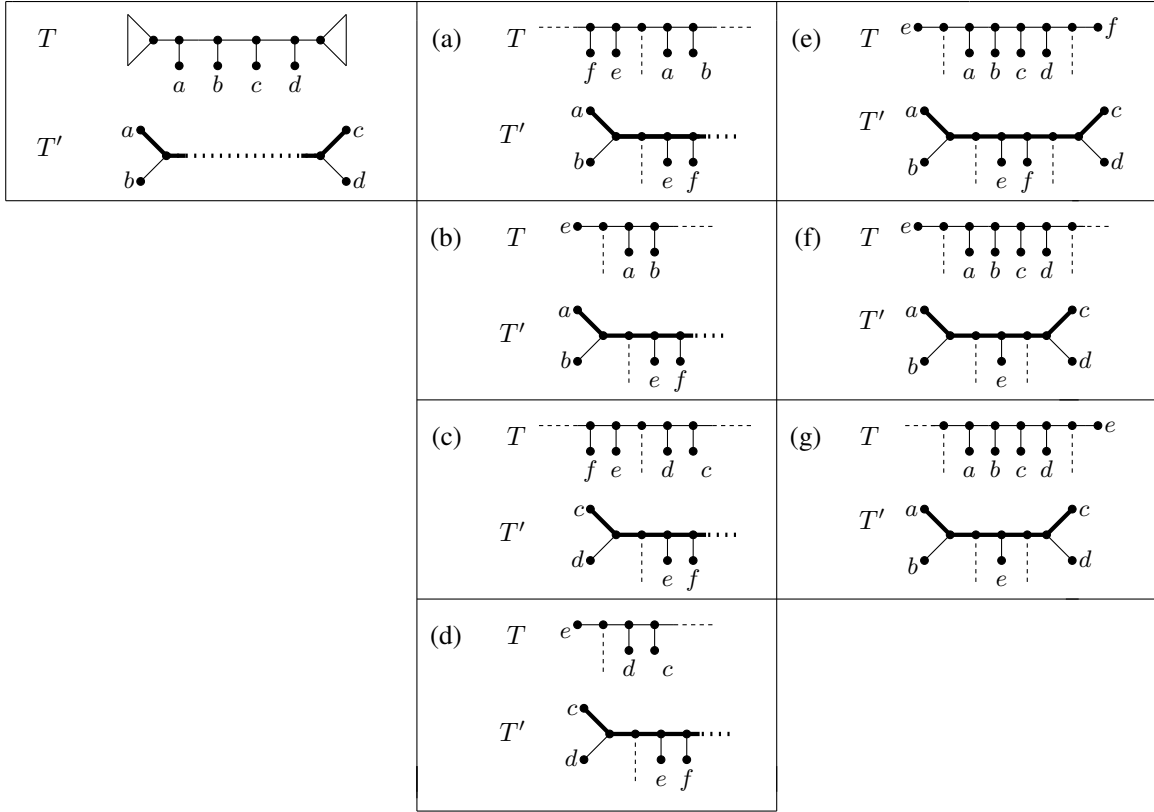


Fig. 11: Tree topologies checked by Algorithm 1. Path P , as used in Theorems 10 and 11, is indicated in bold. Solid lines are edges. Dotted and dashed lines are subtrees that can be optionally present in the tree. Figures (a)–(d) correspond to the situation when the 2|2 side in the underlying generator is a simple edge, and (e)–(g) to when it is a multi-edge.

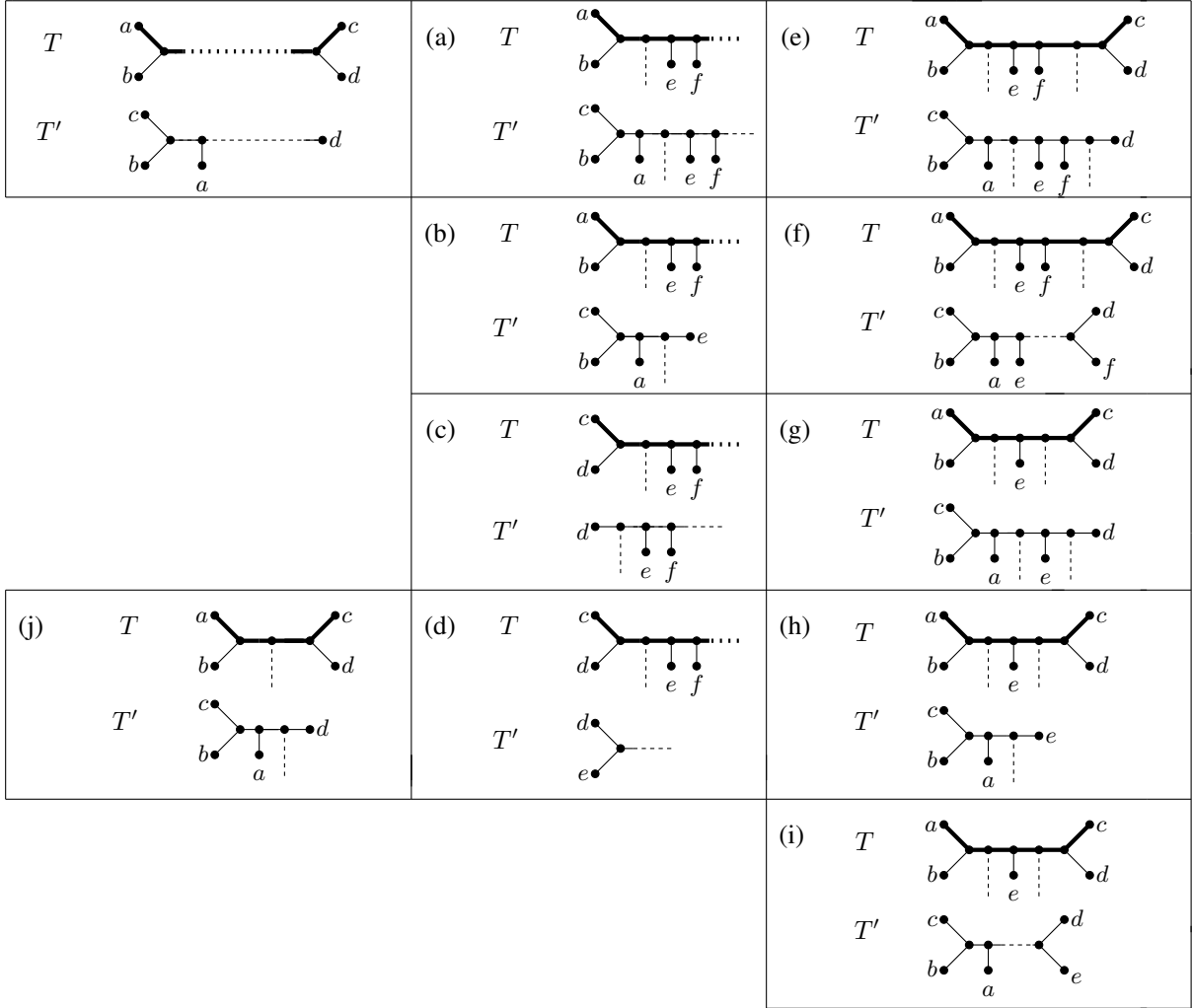


Fig.12: Tree topologies checked by Algorithm 2. Path P , as used in Theorems 12 and 13, is indicated in bold. Solid lines are edges. Dotted and dashed lines are subtrees that can be optionally present in the tree. Figures (a)–(d) correspond to the situation when the $2|1|1$ side in the underlying generator is a simple edge, and (e)–(i) to when it is a multi-edge. Figure (j) corresponds to Theorem 14, which deals with the situation when the side in the underlying generator is a loop.