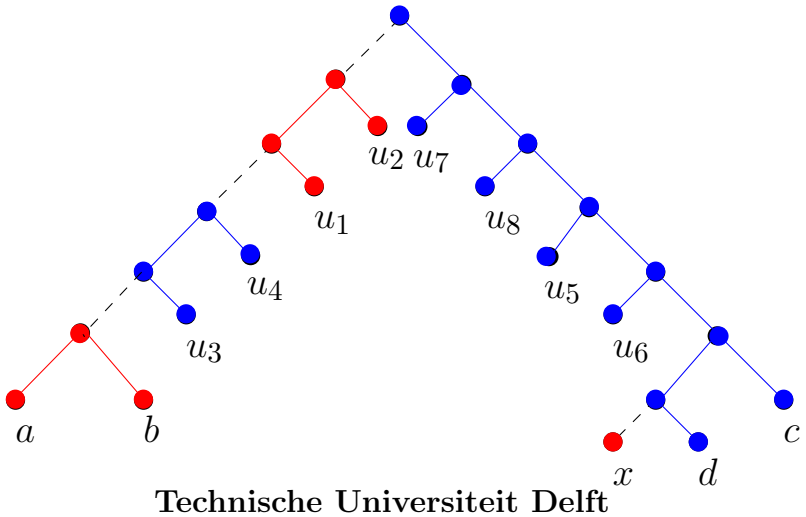
Different reduction rules for the Maximum Parsimony distance on phylogenetic trees

Author

Elise Deen (4896467)

Supervisors

Mark Jones
Leo van Iersel



Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

June 2021

Abstract

In this report, the bounded Maximum Parsimony distance will be considered when applying three different reduction rules. The distance is a measure on how dissimilar two trees are and is calculated based on the number of mutations that occur when looking at heritable traits. The first rule considered, is the chain reduction. For this rule, it is proven that the bounded MP distance is preserved after applying this rule. This is done by adapting the proof from Steven Kelk et al. [10]. For the second rule considered, the generalized subtree reduction, it is also proven that the bounded MP distance is preserved after applying this reduction. Again, this is done by adapting the proof in the paper by Steven Kelk et al. [10]. Then, at last, we looked at a new reduction rule for the TBR distance, introduced by Steven Kelk and Simone Linz [12], the (2,1,2)-reduction. In this report, it is shown with help of a counterexample that this rule does not necessarily reduce the distance with one like it is the case for the TBR distance. However, it can be concluded that the distance is either preserved or reduced with one.

Contents

1 Introduction

Phylogenetics is a study about the evolutionary relationships between species. Thus this study looks at organisms that are probably related and searches for extinct species from which the organisms originate. The best way to look at the relationships is with help of a phylogenetic tree. In such a tree, you can see different species and their ancestors. To make this tree, you can look at (dis)similarities between heritable traits like DNA (for more information about phylogenetics, see [16]).
To compare different trees and to see how dissimilar they are, a distance is needed. There are different distances like the Tree Bisection and Reconnection (TBR) distance [1], [2] or the subtree Prune and Regraft (SPR) distance [3]. These distances are looking at how many actions of a particular kind it requires to transform one tree into the other. The specific action is dependent on the distance. However, there is a relatively new distance that uses a different method, the Maximum Parsimony distance [9] or short MP distance. To calculate this distance, you first assign some distribution of states to the species, the leaves of the tree. This distribution can for example be based on the DNA of the organism. The Parsimony score is then the minimum number of mutations that occurs when also assigning a state to the ancestors, the rest of the nodes. The Maximum Parsimony distance is the difference in mutations maximized over all possible distributions [9]. There are two types of MP distances: unbounded and bounded. A bounded MP distance is a MP distance where there is a bound on the number of states used. An unbounded MP distance is the opposite, without bound on the number of states used. In this report, mostly the bounded distance is considered. The concept of the Maximum Parsimony distance feels more intuitive than the one from the other distances, because the MP distance looks at mutations or differences between species. Over time, organisms can transform into other species by mutations in the DNA, so it make sense to look at this when finding the best possible tree. Moreover, the distance is a metric which can be quite handy. However, this distance is NP-hard to calculate [9].
A solution to this can be to kernelize the problem, so to make the trees smaller without changing the distance. There are different rules which can reduce the trees without changing the TBR distance [11]. The MP distance is related to the TBR distance [6], so the rules might also hold for the Maximum Parsimony distance. There are already some results on this [10]), but in this report we will look further into these reduction rules. The two most used rules are chain reduction and generalized subtree reduction. In this report, we will show that the bounded MP distance is preserved after applying these rules. We will do this by adapting the proof of unbounded MP distance from Steven Kelk et al. [10]. We will also look into one other rule, the (2,1,2)-reduction. This rule is a new rule introduced by Steven Kelk and Simone Linz [12]. Unfortunately, the MP distance is not necessarily preserved for this rule. However, there is something we can say about this: the distance is either preserved or reduced by one after applying the rule.

This report will start with all the needed definitions on phylogenetic trees and on the MP distance. Some more explanation on this will be given. Then, in section 3, 4 and 5, the three reduction rules mentioned earlier will be explained. The report ends in section 6 with a conclusion and some ideas for further research.

2 Definitions and preliminaries

The same definitions as the definitions in the article by Steven Kelk and Mareike Fischer [9] and the article by Vincent Moulton and Taoyang Wu [14] are used, but will be stated here again.

2.1 Phylogenetic trees

First some basic things about graphs. A graph consists of a set of nodes V and a set of edges E that connect the nodes. The *degree* of a node v is the number of edges that have v as an endpoint. In a graph you can denote a *path*. A path is a sequence of nodes where consecutive nodes are connected by an edge. A graph can be connected, which means that any node has a path to any other node. A *tree* is then defined as a connected graph with no cycle. A cycle is a path that has the same endpoint as the starting point of the path. An edge in a graph can be *split* which means that the edge is divided by a node into two parts (see figure 1). Going back from the split edge to the original edge means that you *suppress* the node (see figure 1).

Figure 1: Two trees, where T_2 is the tree with edge e split and T_1 is the tree with node u suppressed. T_2 is a cherry.

An *unrooted phylogenetic tree* is an undirected tree $T = (V, E)$ on some set of taxa X, taxon is a group of organisms that are related. So the leaves of tree T are the taxa in X. A *rooted phylogenetic tree* is a phylogenetic tree with one internal node specified as the root. In such a tree, the *parent* of a node v is the node adjacent to v on the path from v to the root. A child of a node w is then a node from which w is the parent. A phylogenetic tree is *binary* if the root has degree 2 and the other nodes have degree 3 or 1. Thus a binary unrooted tree has only nodes with degree 1 or 3. A node with degree 1 is called a *leaf*, so the set X is also called the leaf set or set of leaves. A *cherry* consists of two leaves with their common parent. In figure 1 (and table 1), an example of a cherry is shown.
If T is such a (binary) tree on a set of taxa X and $Y \subseteq X$, then a *subtree* $T|_Y$ is a minimal connected subgraph of T such that $T|_Y$ contains all the elements in Y where all nodes with degree two, except the root, are suppressed. So $T|_Y$ is a tree on taxa set Y. In figure 2 an example of a subtree is shown.
A *split* $A|B$ on taxa set X is a bipartition such that (i) $A \cap B = \emptyset$, (ii) $A \cup B = X$ and (iii) A, $B \neq \emptyset$. An edge e *induces* such a split if, after removing e, A is the set of taxa appearing in one connected component and B is the set of taxa appearing in the other component. In figure 2, there is a split $A|B$ in tree T with $A = Y$ and $B = X \setminus Y$. The red edge e induces this split.

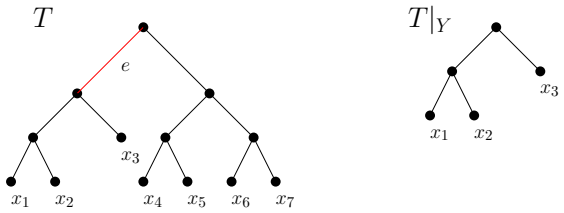
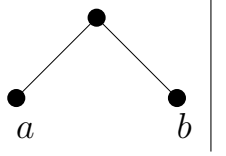
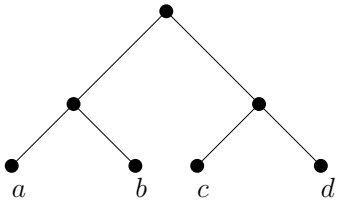
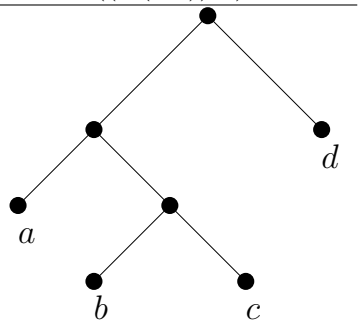


Figure 2: Tree T on leaf set $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ and the subtree $T|_Y$ with $Y = \{x_1, x_2, x_3\} \subseteq X$. In tree T, there is a $Y|X \setminus Y$ split which is induced by the red edge e.

One way to represent a phylogenetic tree is the *Newick format* [4]. The Newick format is used to read trees in a computer program. The format uses parentheses to group species together based on relations between them. In table 1, are some simple trees with their Newick format.

Table 1: Some simple trees with their Newick format.

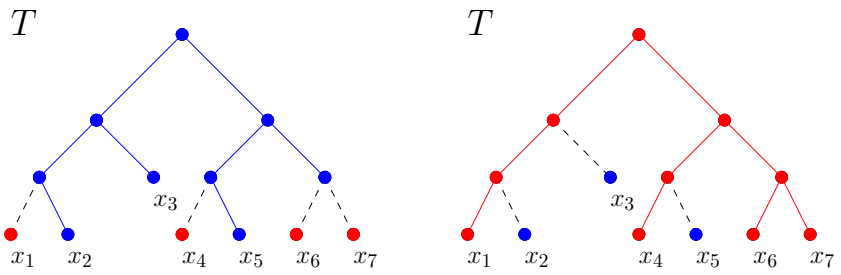
Newick format	(a,b);	((a,b),(c,d));	((a,(b,c)),d);
Tree			

2.2 The Parsimony score

A *character* on X is a surjective function $f : X \to C$ with C the set of states. An example of such a character is shown in figure 3. This character is a function $f : X \to \{\text{red}, \text{blue}\}$ where the leafs x_1, x_4, x_6 and x_7 are coloured red and the leafs x_2, x_3 and x_5 are coloured blue.

An *extension* on such a character f is a function $g : V \to C$ such that $g(x) = f(x)$ for all $x \in X$. In figure 3, two extensions of character f are shown, g_1 and g_2. g_1 is the extension where the leafs are coloured as indicated by the character and the rest of the nodes are coloured blue. g_2 is the extension where all nodes except the leafs are coloured red.

In an extension, a *mutation* can occur. There is a mutation between node v and node w if there appears a substitution in the edge vw i.e. $g(v) \neq g(w)$. The number of mutations



with extension g_1 with extension g_2

Figure 3: Tree T on taxa set $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ with character $f : X \to \{\text{red}, \text{blue}\}$. We have shortened blue to b and red to r. f assigns the colour red to leafs x_1, x_4, x_6 and x_7 and the colour blue to leafs x_2, x_3 and x_5. Two different extensions are shown, g_1 where all nodes except the leaves are coloured blue and g_2 where they are coloured red. The mutations are given by the dotted lines. In the tree with g_2, also the Fitch map is shown.

that appear in an extension g is denoted by $\Delta(g)$. Note that $\Delta(g) \geq 0$.
In both extensions in figure 3, some mutations occur. The dotted edges are the edges where a mutation occurs. So $\Delta(g_1) = 4$ and $\Delta(g_2) = 3$.

The *parsimony score* $l_f(T)$ is this number $\Delta(g)$ minimized over all possible extensions g on the same character f. The extension that minimizes $\Delta(g)$ is called a minimum or optimal extension. The Parsimony score is independent of the root and can also be calculated for unrooted trees. In the figure 3, two extensions were shown. $\Delta(g_1) > \Delta(g_2)$, so g_1 is definitely not an optimal character and we know that the parsimony score $l_f(T) \leq 3$. Knowing the exact score is quite difficult, but there are algorithms to compute it.

A good algorithm to find the parsimony score of a tree and character is the *Fitch algorithm*. We will shortly state Fitch algorithm for binary trees. If the trees are unrooted, one can root it by splitting an arbitrary edge. For some internal node v, we denote its two children by v_l and v_r. Then, first a *Fitch map* $F : V \to 2^C \setminus \{\emptyset\}$ is constructed:

1. For each leaf x, $F(x) = \{f(x)\}$.

2. For each internal node v where $F(v_l)$ and $F(v_r)$ already have been calculated:

$$F(v) = \begin{cases} F(v_l) \cup F(v_r) & \text{if } F(v_l) \cap F(v_r) = \emptyset \\ F(v_l) \cap F(v_r) & \text{otherwise.} \end{cases} \tag{1}$$

An internal node v is called a *union node* if $F(v_l) \cap F(v_r) = \emptyset$ and is called a *intersection node* otherwise. The parsimony score l_f is then equal to the number of union nodes [10]. In figure 4 the Fitch map for character f is shown. It is clear from the picture that there are three union nodes, so $l_f(T) = 3$.
An extension $\overline{f}$ of character f is called a *Fitch extension* if $(i)\overline{f}(v) \in F(v)$ for all $v \in V$, and (ii) $\overline{f}(v) = \overline{f}(v_l)$ or $\overline{f}(v) = \overline{f}(v_l)$ if v is an union node and $\overline{f}(v) = \overline{f}(v_l) = \overline{f}(v_l)$

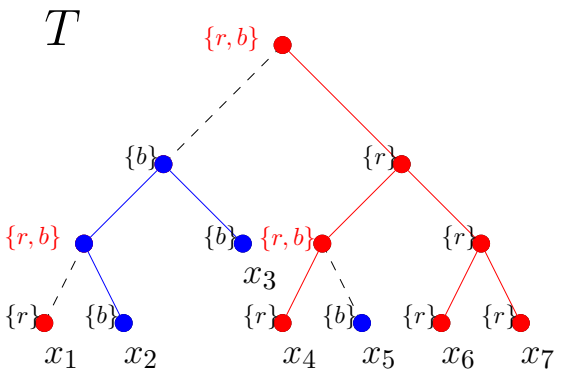


Figure 4: Tree T with its Fitch map and a Fitch extension. The Fitch map of the union nodes is coloured red.

if v is an intersection node. Each Fitch extension is minimal, but not all minimal extensions are Fitch extensions. You can also see this in figure 3. We determined that the parsimony score is equal to 3 and we also found an extension g_2 with $\Delta(g_2) = 3$. So g_2 is minimal. However, this extension is not a Fitch extension since $g_2(v) \notin F(v)$ for some $v \in V$. There is a node which is coloured red, although the Fitch map is equal to $\{b\}$, blue. In figure 4 a Fitch extension is shown. The number of mutations with this extension is three. So this extension is minimum.
To find a Fitch extension, start with a node ρ and choose a state $s \in F(\rho)$ where F is a Fitch map. Let $\overline{f}(\rho) = s$. Then for all nodes w, a child from v from which $\overline{f}(v)$ is known, the Fitch extension is as follows.

$$\overline{f}(w) = \begin{cases} \overline{f}(v) & \text{if } \overline{f}(v) \in F(w) \\ \text{Any state in } F(w) & \text{otherwise.} \end{cases} \tag{2}$$

Let T be a rooted binary tree on X and f a character on X. Let ρ be the root of T and F the Fitch map induced by f. Note that for each state $s \in F(\rho)$, there exists a Fitch-extension $\overline{f}$ of f such that $\overline{f}(\rho) = s$.

2.3 The maximal Parsimony distance (MP)

The maximum Parsimony distance between T_1 and T_2 is denoted as $d_{MP}(T_1, T_2)$ and defined as $d_{MP}(T_1, T_2) = \max_f |l_f(T_1) - l_f(T_2)|$. This is the unbounded MP distance. We will define $d^r_{MP}(T_1, T_2)$ as the bounded maximal Parsimony distance when the number of states used in the character f is bounded by r ($|C| \leq r$). In figure 3, we looked at the parsimony score of tree T with character f. We can also take a different tree T^* on the same set of taxa X and calculate the parsimony score with the same character f on this tree. This is done in figure 5. As can be seen, the parsimony score of T^* with respect to f is $l_f(T^*) = 2$. So the maximum parsimony distance is at least the difference between the two scores, $|l_f(T) - l_f(T^*)| = |3 - 2| = 1$. Knowing the exact Maximum Parsimony

distance is really difficult. In general, it is NP-hard to calculate [9].

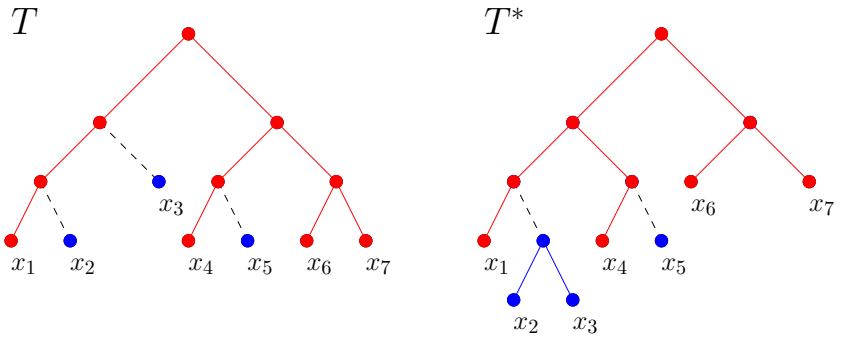


Figure 5: Tree T with character f and a minimum extension g_2 and tree T^* on the same set of taxa X with character f and a minimum extension. The mutations are given by the dotted edges. So the parsimony score of T is 3 and the score of T^* is 2.

There are different properties of the MP distance. The MP distance is for example a metric (see article [9]). Another useful property which we will use later in the proofs of the reduction rules is the following lemma:

Lemma 2.1. *Suppose that T_1 and T_2 are binary phylogenetic trees on the set X and Y is a subset of X. Then $d^r_{MP}(T_1|_Y, T_2|_Y) \leq d^r_{MP}(T_1, T_2)$ for r an arbitrary integer.*

Proof. It is sufficient to prove the lemma for the case $Y = X \setminus \{x\}$ with $x \in X$. Let f be the optimal character on Y and assume without loss of generality that $l_f(T_1|_Y) \leq l_f(T_2|_Y)$. So $d^r_{MP}(T_1|_Y, T_2|_Y) = l_f(T_2|_Y) - l_f(T_1|_Y)$. Now look at the neighbour of x in T_1, call this node p. Root the tree by splitting an edge adjacent to p. Now call the other neighbours of p, p_l and p_r such that p_l is the root of the tree (see also figure 6). Consider now a Fitch map F induced by f. Let f^* be a character on X obtained from f by assigning x to a state s in $F(p_r)$. The Fitch map of p is equal to $F(p) = F(x) \cap F(p_r) = s$ since $F(x) \cap F(p_r) \neq \emptyset$ and x and p_r children of p. So $l_{f^*}(T_1) = l_f(T_1|_Y)$, because no extra mutation occur when adding x. Moreover, f^* introduces two, one or no mutations in T_2. So $l_f(T_2|_Y) \leq l_{f^*}(T_2)$. Note that $l_{f^*}(T_1) \leq l_{f^*}(T_2)$. So by these inequalities and the definition of the maximum Parsimony distance, we have

$$\begin{aligned} d^r_{MP}(T_1|_Y, T_2|_Y) &= l_f(T_2|_Y) - l_f(T_1|_Y) \\ &\leq l_{f^*}(T_2) - l_{f^*}(T_1) \\ &\leq d^r_{MP}(T_1, T_2). \end{aligned}$$

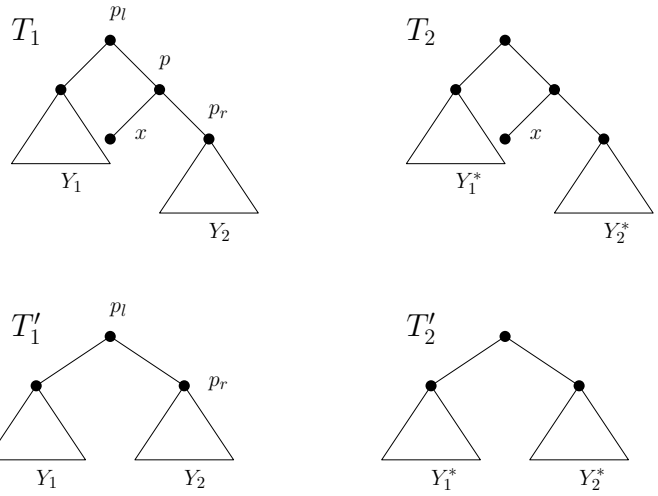


Figure 6: Two trees on leaf set X and the same trees on leaf set $Y = X \setminus x$. So, $Y = Y_1 \vee Y_2 = Y_1^* \vee Y_2^*$.

□

2.4 Less constrained roots argument

Let T_1 and T_2 be two phylogenetic trees each consisting of 2 subtrees as shown in figure 7. Let T_A, T_B, T_C, T_D be these subtrees. For $P \in \{A, B, C, D\}$, let e_P refer to the edge incoming to the root of T_P; let X_P refer to the taxa in subtree T_P; let f_P refer to the character obtained by restricting f to X_P; and let F_P refer to the set of states assigned to the root of T_P by the Fitch map induced by f_P. Note that $X_A \cup X_B = X_C \cup X_D$. Moreover, for both trees, we define the chain region of T to be the set of edges incident to at least one red node. The red nodes are the nodes not in one of the subtrees. Let m_i, $(i = 1, 2)$ be the number of union nodes among red nodes in T_i, which is the same as the number of mutations occurring in the chain region of T_i for a Fitch-extension of f. Then,

$$m_1 = l_f(T_1) - l_{f_A}(T_A) - l_{f_B}(T_B)$$
$$m_2 = l_f(T_2) - l_{f_C}(T_C) - l_{f_D}(T_D)$$

In addition, let $p = m_2 - m_1$ and then we have

$$d^r_{MP}(T_1, T_2) = l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + p. \tag{3}$$

We have now the following lemmas, the less constrained roots arguments:

Lemma 2.2. *If $l_f(T_1) < l_f(T_2)$ and $F_A \subseteq F_C \wedge F_B \subseteq F_D$ for T_1 and T_2, then $p \leq 0$.*

Proof. Consider a Fitch-extension $\overline{f_1}$ of f to T_1. Then by definition $\overline{f_1}$ assigns a state a from F_A to the root of T_A, and a state b from F_B to the root of T_B. Since $a \in F_C$ $(F_A \subseteq F_C)$, we fix a Fitch-extension $\overline{f_C}$ of f_C to T_C that maps the root of T_C to a.

Similarly, we fix a Fitch-extension $\overline{f_D}$ of f_D to T_D that maps the root of T_D to b. Now consider the extension $\overline{f_2}$ of f to T_2 obtained by combining $\overline{f_C}$, $\overline{f_D}$, and exactly mimicking $\overline{f_1}$ for the red nodes of T_2. Then the number of mutations induced by $\overline{f_2}$ in the chain region of T_2 is exactly the same as that by $\overline{f_1}$ in the chain region of T_1. In other words, we have $\Delta(f_2) = l_{f_C}(T_C) + l_{f_D}(T_D) + m_1$, from which we conclude that, if $(F_A \subseteq F_C) \wedge (F_B \subseteq F_D)$, then

$$d^r_{MP}(T_1, T_2) = l_f(T_2) - l_f(T_1) \leq \Delta(f_2) - l_f(T_1) = l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B).$$

In particular, this shows $p \leq 0$.
If one of the subtrees is empty, add a tree with all states in the Fitch map of the root (see figure 8). Adding this tree does not add any mutations in the chain region of the tree since all states are contained in the Fitch map. So also in these cases, $p \leq 0$. □

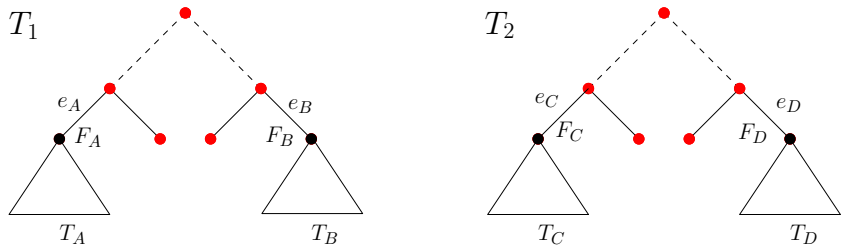


Figure 7: Trees T_1 and T_2 on which lemma 2.2 is applicable when $l_f(T_1) < l_f(T_2)$ and $F_A \subseteq F_C \wedge F_B \subseteq F_D$.

Lemma 2.3. *If $l_f(T_1) < l_f(T_2)$ and $F_A \subseteq F_C$ for T_1 and T_2, then $p \leq 1$.*

Proof. Consider a Fitch-extension $\overline{f_1}$ of f to T_1. Then by definition $\overline{f_1}$ assigns a state a from F_A to the root of T_A. Since $a \in F_C$ ($F_A \subseteq F_C$), we fix a Fitch-extension $\overline{f_C}$ of f_C to T_C that maps the root of T_C to a. Now consider the extension $\overline{f_2}$ of f to T_2 obtained by combining $\overline{f_C}$, a Fitch-extension of f_D to T_D, and exactly mimicking $\overline{f_1}$ for the red nodes of T_2. Then the number of mutations induced by $\overline{f_2}$ in the chain region of T_2 is exactly the same as that by $\overline{f_1}$ in the chain region of T_1 plus one (at e_D). In other words, we have $\Delta(f_2) = l_{f_C}(T_C) + l_{f_D}(T_D) + m_1 + 1$, from which we conclude that, if $(F_A \subseteq F_C)$, then

$$d^r_{MP}(T_1, T_2) = l_f(T_2) - l_f(T_1) \leq \Delta(f_2) - l_f(T_1) = l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + 1.$$

In particular, this shows $p \leq 1$.
If one of the subtrees is empty, add a tree with all states in the Fitch map of the root (see figure 8). Adding this tree does not add any mutations in the chain region of the tree since all states are contained in the Fitch map. So also in these cases, $p \leq 1$. □

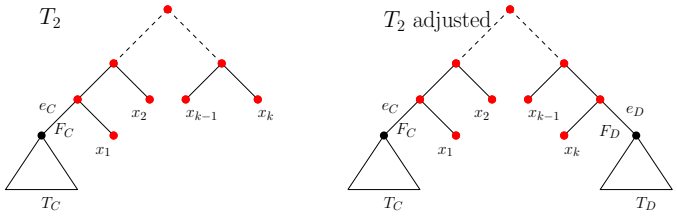


Figure 8: Tree T_2 from figure 11 and the same tree but adjusted with an extra subtree T_D such that F_D contains all states. This does not change the number of mutations in the chain region.

2.5 Tree Bisection and Reconnection distance (TBR)

A common used distance is the Tree Bisection and Reconnection distance, *TBR distance* for short. The TBR distance ($d_{TBR}(T_1, T_2)$) is the number of TBR moves that are needed to transform one tree into the other. A TBR move consists of two parts. First, cut an edge such that two components occur and suppress all nodes of degree 2. Second, split an edge in both components and connect the two components again with a new edge between the new created nodes. In figure 9, an example of a TBR move is shown. First, the red line is cut such that two components occur and nodes u and v are suppressed. This transforms tree T into T'. Second, the blue lines are split with nodes u' and v' and the two components are connected with a new edge between u' and v'. This transforms T' into T''. So tree T is transformed into T'' with one TBR move.

In the article by B Allen and M. Steel [1], it is proven that this distance is a metric,

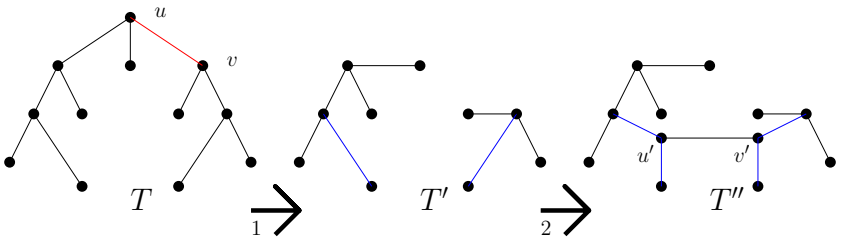


Figure 9: One TBR move that transforms T into T''. The red edge is cut and nodes u and v are suppressed in the first step. In the second step, the blue nodes are split and the components are connected with a new edge between u' and v'.

just like the MP distance. The TBR distance has already been investigated a lot. For example, it has been proven that you can reduce two trees with help of chain reduction or subtree reduction without changing its TBR distance [1]. Moreover, for five new reduction rules, the TBR distance is reduced with one after applying these rules [12]. It is also shown that the TBR distance is in a lot of cases very close to the MP distance [10]. In an article by Mark Jones et al. [6], it is shown that $d_{MP}(T_1, T_2) \leq d_{TBR}(T_1, T_2) \leq 2\alpha d_{MP}(T_1, T_2)$ with α a constant factor. So we suspect that the reduction rules that apply for the TBR distance, also apply for the MP distance.

3 Chain reduction

In this section we will look at chain reduction and we will prove that the bounded MP distance is preserved after applying this rule.
First we will give the definition of chain reduction. For a leaf x_i, let p_i be the (unique) neighbour of x_i. Then an ordered sequence $(x_1, x_2, ..., x_k)$ is called a *chain of length* k if $(p_1, p_2, ..., p_k)$ is a path in T. It is possible to have $p_1 = p_2$ and/or $p_{k-1} = p_k$. If this is the case, the chain is called *pendant* in the tree. A chain is *common* to T_1 and T_2 if it is a chain in both trees. Suppose two trees T_1 and T_2 have a common chain $X(k) = (x_1, ..., x_k)$ with $k \geq 5$, then these trees can be reduced with use of *chain reduction*. Let T_1' and T_2' be trees on leaf set $X' = X \setminus X(k) \cup \{x_1, x_2, x_{k-1}, x_k\}$ where $T_1' = T_1 \mid_{X'}$ and $T_2' = T_2 \mid_{X'}$. T_1' and T_2' are said to be obtained by reducing the chain $X(k)$ to length 4. It can be proved that the unbounded Maximum Parsimony distance is preserved when applying chain reduction [10]. In that proof the number of states is increased, so this proof does not hold for the bounded MP distance. The question now is if the bounded distance is also preserved.

Theorem 3.1. *Let T_1 and T_2 be two unrooted binary trees on the same set of taxa X. Let K be a common chain of length $k \geq 5$. Let T_1' and T_2' be the two trees obtained by reducing K to length 4. Then $d^r_{MP}(T_1, T_2) = d^r_{MP}(T_1', T_2')$ for $r \in \mathbb{N}$.*

The proof is based on the proof by Steven Kelk et al. [10].

Proof. In Lemma 2.1, it is proven that for all $Y \subseteq X$, $d^r_{MP}(T_1|_Y, T_2|_Y) \leq d^r_{MP}(T_1, T_2)$. This corollary in combination with the definition yields that $d^r_{MP}(T_1', T_2') \leq d^r_{MP}(T_1, T_2)$. So we only have to prove $d^r_{MP}(T_1', T_2') \geq d^r_{MP}(T_1, T_2)$.
Without loss of generality, we may assume that $d^r_{MP}(T_1, T_2) > 0$ (i.e., $T_1 \neq T_2$) since otherwise the claim clearly holds. Note that this implies $X \neq K$ and whenever K is pendant in a tree, at least one end of the chain is attached to the main part of the tree. There are now three main cases to consider, the common chain is pendant in neither tree, one tree or both trees.

I: the common chain is pendant in neither tree
Let f be a character that maximizes $|l_f(T_1) - l_f(T_2)|$. Without loss of generality, we can assume that $l_f(T_1) < l_f(T_2)$. So $d^r_{MP}(T_1, T_2) = l_f(T_2) l_f(T_1)$. The trees are not necessarily rooted, but for the proof we root the trees somewhere between p_2 and p_{k-1} (see figure 10).
Let T_A, T_B, T_C, T_D be the 4 subtrees of T_1 and T_2 as shown in figure 10. For $P \in \{A, B, C, D\}$, let e_P refer to the edge incoming to the root of T_P; let X_P refer to the taxa in subtree T_P; let f_P refer to the character obtained by restricting f to X_P; and let F_P refer to the set of states assigned to the root of T_P by the Fitch map induced by f_P. Note that $X_A \cup X_B = X_C \cup X_D$. Moreover, for each tree $T \in \{T_1, T_2, T_1', T_2'\}$ we define the chain region of T to be the set of edges incident to at least one red node. The red nodes are the nodes in the chain $x_1, ..., x_k$ and their parents $p_1, ..., p_k$, as shown in figure 10. Let m_i, $(i = 1, 2)$ be the number of union nodes among red nodes in T_i,

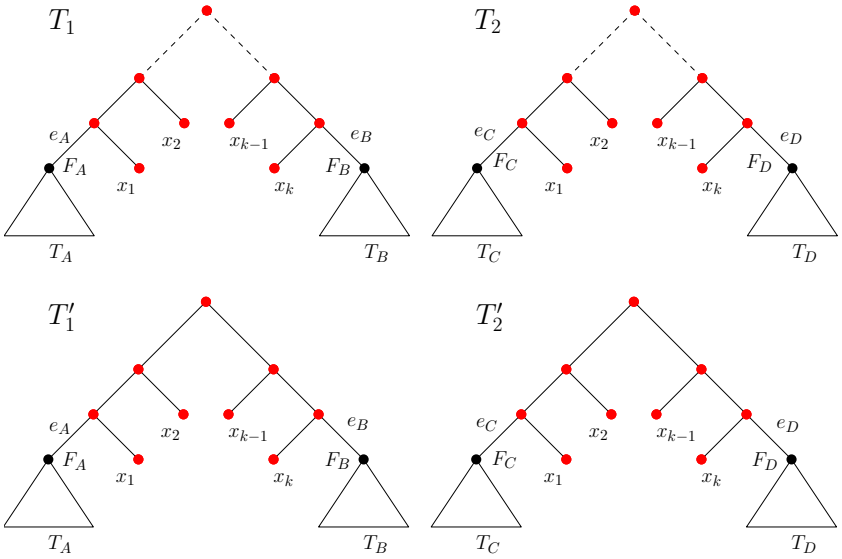


Figure 10: The chain reduction as applied in the case when the common chain K is pendant in neither tree. Note that in T_1 and T_2 a dotted line is used to denote the taxa $\{x_3, ..., x_{k-2}\}$ which are removed by the chain reduction. All the trees in the figure are unrooted, but for the purpose of proving correctness of the chain reduction we have shown them as rooted. T_1' and T_2' must be rooted exactly halfway along the chain, as shown. For T_1 and T_2 it is not so important where the tree is rooted as long as the root is in the same part of the chain in both trees.

which is the same as the number of mutations occurring in the chain region of T_i for a Fitch-extension of f. Then,

$$m_1 = l_f(T_1) - l_{f_A}(T_A) - l_{f_B}(T_B)$$

$$m_2 = l_f(T_2) - l_{f_C}(T_C) - l_{f_D}(T_D).$$

In addition, let $p = m_2 - m_1$ and then we have

$$d^r_{MP}(T_1, T_2) = l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + p. \tag{4}$$

First we shall show that $p \leq 2$. Fix a Fitch-extension f_1 of f to T_1, and consider an extension f_2 of f to T_2 obtained by a minimum extension of f_C to T_C, a minimum extension of f_D to T_D, and exactly mimicking f_1 on the red nodes of T_2. Then compared with f_1, the extension f_2 creates at most two new mutations on the chain region, namely on the edges e_C and e_D. In other words, we have $\Delta(f_2) \leq l_{f_C}(T_C) + l_{f_D}(T_D) + (m_1 + 2)$.

Together with $l_f(T_2) \leq \Delta(f_2)$ and $l_f(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + m_1$, this implies

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &= l_f(T_2) - m_1 - l_{f_C}(T_C) - l_{f_D}(T_D) \\ &\leq \Delta(f_2) - m_1 - l_{f_C}(T_C) - l_{f_D}(T_D) \\ &\leq 2. \end{aligned} \tag{5}$$

Now we will show that $p \geq 0$ holds with r states. Let the states be $(s_1, s_2, ..., s_r)$. Given a character f on X, we will write $f^* = f[a_1, a_2, ..., a_{k-1}, a_k]$ as shorthand for the character on X obtained from f by leaving the states assigned to taxa in $X_A \cup X_B = X_C$ intact and assigning states a_i to the leaves x_i for $i \in \{1, ..., k\}$. Now let $f^* = f[s_1, s_1, ..., s_1, s_1]$ be a (not necessarily optimal) character. Consider a Fitch extension of f^* to T_1 and T_2 and let $m_i^*, i = (1, 2)$ be the number of mutations occurring in the chain region of tree T_i considering the extension of f^*. Then three cases can occur, $m_1^* = 0$, $m_1^* = 1$ or $m_1^* = 2$, since the only mutations that can occur are at e_A and e_B. Note that by the same argument $0 \leq m_2^* \leq 2$.
Case 1: $m_1^* = 0$. There is no mutation in the chain region for tree T_1. Thus, we have

$$l_{f^*}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) \text{ and } l_{f^*}(T_2) \geq l_{f_C}(T_C) + l_{f_D}(T_D). \tag{6}$$

The optimality of f implies that $l_f(T_2) - l_f(T_1) \geq l_{f*}(T_2) - l_{f*}(T_1)$. Then by equation 6 and this inequality, we have

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq l_{f*}(T_2) - l_{f*}(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &= 0. \end{aligned} \tag{7}$$

Case 2: $m_1^* = 1$. There is one mutation in the chain region for tree T_1. Thus, we have

$$l_{f*}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + 1. \tag{8}$$

If there are one or two mutations in the chain region of T_2, it is easy to show $p \geq 0$. We then have $l_{f*}(T_2) \geq l_{f_C}(T_C) + l_{f_D}(T_D) + 1$ and this leads to

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq l_{f*}(T_2) - l_{f*}(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq l_{f_C}(T_C) + l_{f_D}(T_D) + 1 - l_{f_A}(T_A) - l_{f_B}(T_B) - 1 - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &= 0. \end{aligned} \tag{9}$$

In the case that there is no mutation in the chain region of T_2, $s_1 \in F_C$ and $s_1 \in F_D$. Since $m_1^* = 1$, we know that $s_1 \notin F_A$ or $s_1 \notin F_B$ but not both. Assume without loss of

generality that $s_1 \notin F_A$. Then some other state is in F_A, assume $s_2 \in F_A$. Consider now another character $f^{**} = f[s_2, s_1, ..., s_1, s_1]$. The number of mutations in the chain region of T_1 is still one, since the only mutation occurs between x_2 and its parent p_2 or between p_1 and p_2. But now there is also one mutation in the chain region of T_2, one between p_1 and p_2 or one between x_1 and its parent p_1. So $l_{f^{**}}(T_2) = l_{f_C}(T_C) + l_{f_D}(T_D) + 1$ and $l_{f^{**}}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + 1$ and we have

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq l_{f^{**}}(T_2) - l_{f^{**}}(T_1) - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &= l_{f_C}(T_C) + l_{f_D}(T_D) + 1 - l_{f_A}(T_A) - l_{f_B}(T_B) - 1 - l_{f_C}(T_C) - l_{f_D}(T_D) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &= 0. \end{aligned} \tag{10}$$

Case 3: $m_1^* = 2$. There are two mutations in the chain region of T_1. Thus $s_1 \notin F_A$ and $s_1 \notin F_B$ and the mutations are at e_A and e_B. assumes $s_2 \in F_A$. Let $f^{**} = [s_2, s_2, ..., s_2, s_2]$. Now there is at most one mutation in the chain region of T_1 (at e_B) and we have the same situation as in case 1 or as case 2. So also in this case, $p \geq 0$.

By equation 4, the claim $d^r_{MP}(T_1', T_2') \geq d^r_{MP}(T_1, T_2)$ follows from

$$d^r_{MP}(T_1', T_2') \geq l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + p \tag{11}$$

So to prove main case I, it is sufficient to show Equation 11. Then there are three cases to consider, namely $p = 0$, $p = 1$ or $p = 2$, since $0 \leq p \leq 2$. X' is the taxa of the trees T_1' and T_2', so $X' = X \setminus \{x_3, x_4, ..., x_{k-3}, x_{k-2}\}$. For short notation, we will write $f[a, b, c, d]$ to denote the character on X' obtained from f by leaving the states assigned to taxa in $X_A \cup X_B = X_C \cup X_D$ intact and assigning states a, b, c, d to x_1, x_2, x_{k-1}, x_k respectively.

Case 1: $p = 0$. Let $f' = f[s_1, s_1, s_1, s_1]$. Then like before there are three cases, zero, one or two mutations in the chain region of T_1'. Define m_i', $i = (1, 2)$ to be the mutations in the chain region of T_i'.

- $m_1' = 0$. Since $m_2' \geq 0$ and $m_1' = 0$, $l_{f'}(T_1') = l_{f_A}(T_A) + l_{f_B}(T_B)$ and $l_{f'}(T_2') \geq l_{f_C}(T_C) + l_{f_D}(T_D)$. So we have

  $$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &\geq l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) \end{aligned}$$

  So equation 11 holds for this case.

- $m_1' = 1$. Since $m_1' = 1$, $l_{f'}(T_1') = l_{f_A}(T_A) + l_{f_B}(T_B) + 1$. If there is one or two mutations in the chain region of T_2', $l_{f'}(T_2') \geq l_{f_C}(T_C) + l_{f_D}(T_D) + 1$ leads to equation 11. So we only have to prove equation 11 if $m_2' = 0$. In this case, $s_1 \in F_C$ and $s_1 \in F_D$ and we know that $s_1 \notin F_A$ or $s_1 \notin F_B$ but not both. Assume without

loss of generality that $s_1 \notin F_A$. Assume $s_2 \in F_A$. Consider now another character $f'' = f[s_2, s_1, s_1, s_1]$. The number of mutations in the chain region of T_1' is still one (between p_1 and p_2, or between p_1 and x_2), but now there is also one mutation in the chain region of T_2 (between p_1 and p_2, or between p_1 and x_1). So $l_{f''}(T_2') = l_{f_C}(T_C) + l_{f_D}(T_D) + 1$ and equation 11 holds.

- $m_1' = 2$. There are two mutations in the chain region, so $s_1 \notin F_A$ and $s_1 \notin F_B$. Assume $s_2 \in F_A$. Consider $f[s_2, s_2, s_2, s_2]$. Now there is at most one mutation in T_1' (at e_B) and we have the same situation as in case 1a or 1b. So, equation 11 holds.

Case 2: $p = 1$. We will consider three subcases.

- $F_A \setminus F_C \neq \emptyset$. Let $s_1 \in F_A \setminus F_C$ and consider the character $f' = f[s_1, s_1, s_1, s_1]$. Let m_i', $i = (1, 2)$ be the number of mutation that occurs in the chain region of T_i' for character f'. Then $0 \leq m_1' \leq 1$ since there is only one mutation that can occur, namely at e_B ($s_1 \in F_A$). $1 \leq m_2' \leq 2$ since there is at least one mutation ($s_1 \notin F_C$) and at most two (e_C and/or e_D). So there are only four possibilities. For the possibilities $m_1' = 0$, $m_2' = 1$ and $m_1' = 1$, $m_2' = 2$, the difference $m_2' - m_1' = 1$ and we have

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &= l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + 1 \end{aligned}$$

  For $m_1' = 0$ and $m_2' = 2$, the difference $m_1' - m_2' = 2$ and we have

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &= l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + 2 \\ &\geq l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + 1 \end{aligned}$$

  For $m_1' = 1$ and $m_2' = 1$, the difference $m_1' - m_2' = 0$. So we should look for another character f''. We know that $s_1 \in F_A$, $s_1 \notin F_B$, $s_1 \notin F_C$, and $s_1 \in F_D$. Assume $s_2 \in F_B$ and let $f'' = f[s_1, s_1, s_1, s_2]$. Then $m_1'' = 1$ since there is only one mutation, between p_{k-1} and p_k. $m_2'' = 2$, because there is a mutation at e_C and one between x_k and p_k or between p_{k-1} and p_k. Now the difference is one and we have

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f''}(T_2') - l_{f''}(T_1') \\ &= l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + 1 \end{aligned}$$

- $F_B \setminus F_D \neq \emptyset$. This is symmetrical to the previous case.

- $F_A \subseteq F_C \wedge F_B \subseteq F_D$. This case cannot occur. By the less constraint roots argument, lemma 2.2 in section 2.4, $p \leq 0$ since $F_A \subseteq F_C \wedge F_B \subseteq F_D$. This contradicts the assumption that $p = 1$.

Case 3: $p = 2$. Then we have the following three subcases to consider.

- $(F_A \setminus F_C \neq \emptyset) \wedge (F_B \setminus F_D \neq \emptyset)$. Let $a \in F_A \setminus F_C$ and $b \in F_B \setminus F_D$. Take the character $f' = f[a, a, b, b]$ and let m'_i, $i = (1, 2)$ defined as before. If $a \neq b$, $m'_1 = 1$ (between p_2 and p_{k-1}) and $m'_2 = 3$ (e_C, e_D and between p_2 and p_{k-1}). If $a = b$, $m'_1 = 0$ and $m'_2 = 2$ (e_C and e_D). In both situations, the difference between m'_1 and m'_2 is equal to two. Thus we have

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &= l_{f_C}(T_C) + l_{f_D}(T_D) - l_{f_A}(T_A) - l_{f_B}(T_B) + 2 \end{aligned}$$

- $F_A \subseteq F_C$. This case cannot occur. By one of the less constraint roots arguments 2.3, $p \leq 1$. This contradict $p = 2$.

- $F_B \subseteq F_D$. This case cannot occur since it is symmetrical to the previous subcase.

So in all cases, equation 11 holds. So main case 1 holds.

II: the common chain is pendant in exactly one tree
Without loss of generality we assume that K is pendant in T_2 and that the situation is as described in figure 11. Let f be an optimal character. T_P, e_P, F_P, X_P and f_P for $P \in (A, B, C)$ are defined as before. The only difference now is that $X_C = X_A \cup X_B$. Then we have two cases, $l_f(T_1) < l_f(T_2)$ or $l_f(T_1) > l_f(T_2)$.

Case 1: $l_f(T_1) < l_f(T_2)$. So $d^r_{MP}(T_1, T_2) = l_f(T_2) - l_f(T_1)$. As in Equation 6 we have,

$$d^r_{MP}(T_1, T_2) = l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) + p. \tag{12}$$

In this case, $p \leq 1$ because of the usual mimicking construction (i.e. copying the states allocated to the red nodes in T_1, to T_2) used in the proof of Equation 5. That is, at most one extra mutation incurs in T_2 (i.e. on the edge e_C).
We can prove $p \geq 0$ by relabeling f to a new character $f^* = [a, a, ..., a, b]$. If $a = b$, there is no mutation in the chain region of T_1 and at most one mutation in the chain region of T_2 (only possibility is at e_C). So we have

$$l_{f^*}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) \text{ and } l_{f^*}(T_2) \geq l_{f_C}(T_C).$$

Since the optimality of f implies $l_f(T_1) - l_f(T_2) \geq l_{f^*}(T_2) - l_{f^*}(T_1)$, we have

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_C}(T_C) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq l_{f^*}(T_2) - l_{f^*}(T_1) - l_{f_C}(T_C) + l_{f_A}(T_A) + l_{f_B}(T_B) \\ &\geq 0. \end{aligned} \tag{13}$$

If $a \neq b$, then there is exactly one mutation in the chain region of T_1 (between p_{k-1} and p_k) and one or two in the chain region of T_2 (between p_{k-1} and p_k and/or at e_C). Then

$$l_{f^*}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + 1 \text{ and } l_{f^*}(T_2) \geq l_{f_C}(T_C) + 1$$

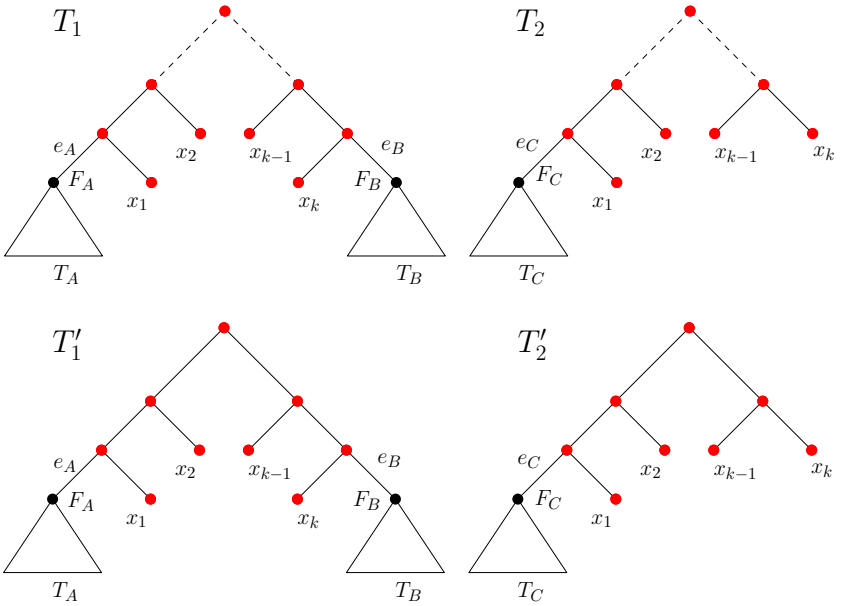


Figure 11: The chain reduction as applied in the case when the common chain K is pendant in tree T_2. Like in fig. 10, the trees are unrooted but shown as rooted and the dotted line denote the taxa $\{x_3, ..., x_{k-2}\}$.

and by the same argument as used for Equation 13, $p \geq 0$ holds. Hence, in Equation 12, we have $p \in \{0, 1\}$, and it remains to prove that

$$d^r_{MP}(T_1', T_2') \geq l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) + p$$

holds, which will be done by considering two subcases.

Case 1.1: $p = 0$. Two cases are considered:

- $F_A \not\subseteq F_C$. Let $s_1 \in F_A \setminus F_C$. Note that $s_1 \notin F_B$ because otherwise the character $f^* = f[s_1, s_1, ..., s_1, s_1]$ would lead to no mutations in T_1 but one mutation in T_2. This contradicts $p = 0$, because $p = m_2 - m_1 \geq m_2^* - m_1^* = 1$ where m_i^*, $i \in (0, 1)$ is the number of mutation in the chain region of T_i considering the character f^*. In other words, such an f^* would give a larger parsimony distance between T_1 and T_2, which contradicts our choice of an optimal f. So this implies that the character $f' = f[s_1, s_1, s_1, s_1]$ leads to one mutation in T_1' (at e_B) and one in T_2' (at e_C). Then,

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &= l_{f_C}(T_C) + 1 - l_{f_A}(T_A) - l_{f_B}(T_B) - 1 \\ &= l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) \end{aligned}$$

  and we are done.

- $F_A \subseteq F_C$. If $F_A \cap F_B \neq \emptyset$, then let $s_1 \in F_A \cap F_B$. Clearly $s_1 \in F_C$. Taking character $f' = f[s_1, s_1, s_1, s_1]$ yields no mutations in both the chain regions of the trees T_1' and T_2'. So the difference between the mutations is zero. Otherwise, $F_A \cap F_B = \emptyset$. In this situation, let $s_1 \in F_A \subseteq F_C$ and let $s_2 \in F_B$. Consider character $f' = f[s_1, s_1, s_2, s_2]$. In this situation, there is one mutation in the chain region of each tree. Again the difference is zero and for both cases it holds that

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2) - l_{f'}(T_1) \\ &= l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) + 0. \end{aligned}$$

  So also in this case, the claim holds.

Case 1.2: $p = 1$.

- $F_A \not\subseteq F_C$. Let $s_1 \in F_A \setminus F_C$. If $s_1 \in F_B$, then consider the character $f' = f[s_1, s_1, s_1, s_1]$. This leads to no mutation in the chain region of T_1 and one mutation in the chain region of T_2. Now suppose $s_1 \notin F_B$. Suppose $s_2 \in F_B$ and take $f' = f[s_1, s_1, s_2, s_2]$, this leads to one mutation in the chain region of T_1 and two in the chain region of T_2. In both situations, the difference between the mutation is one. So we have

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2) - l_{f'}(T_1) \\ &= l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) + 1. \end{aligned}$$

- $F_A \subseteq F_C$. This case cannot happen by the less constrained roots argument (lemma 2.2 in section 2.4). Note that in this case there is no F_D, so we define F_D as the set containing all the states (see figure 8 and the lemma). By lemma 2.2 $p \leq 0$, a contradiction. So we are done.

Case 2: $l_f(T_1) > l_f(T_2)$. so $d^r_{MP}(T_1, T_2) = l_f(T_1) - l_f(T_2)$. In such a case we have

$$d^r_{MP}(T_1, T_2) = l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + p \tag{14}$$

where $p = m_1 - m_2$. We have $p \leq 2$, by the usual mimicking argument, but this time the red nodes in T_1 copy their states from T_2 and not the other way round. (Nodes x_k and its parent p_k in T_1 should both be assigned the state that is assigned to x_k in T_2). We can also show that $p \geq 0$. Consider the character $f^* = f[s_1, s_1, ..., s_1, s_1]$. There are three cases, $m_1^* = 0$, $m_1^* = 1$ or $m_1^* = 2$. Note that $m_2^* \in \{0, 1\}$.

Case 2a: $m_1^* = 0$. Since there are no mutation in the chain region of T_1, we know that $s_1 \in F_A$ and $s_1 \in F_B$. We will now consider three subcases.

- $F_C \not\subseteq F_A$: So we know that there is a s_2 such that $s_2 \in F_C$ and $s_2 \notin F_A$. Consider the character $f^{**} = f[s_2, ..., s_2]$. Then $m_2^{**} = 0$ and $m_1^{**} \geq 1$, since $s_2 \notin F_A$. So,

$$l_{f^{**}}(T_1) \geq l_{f_A}(T_A) + l_{f_B}(T_B) + 1 \text{ and } l_{f^{**}}(T_2) = l_{f_C}(T_C).$$

Since the optimality of f implies $l_f(T_1) - l_f(T_2) \geq l_{f^{**}}(T_1) - l_{f^{**}}(T_2)$, we have

$$\begin{aligned} p &= l_f(T_1) - l_f(T_2) + l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) \\ &\geq l_{f^{**}}(T_1) - l_{f^{**}}(T_2) + l_{f_C}(T_C) - l_{f_A}(T_A) - l_{f_B}(T_B) \\ &\geq 1 \geq 0. \end{aligned} \tag{15}$$

- $F_C \not\subseteq F_B$: This case is similar to the previous one. Consider the character $f^{**} = f[s_2, ..., s_2]$ with $s_2 \in F_C \setminus F_B$. Then $m_2^{**} = 0$ and $m_1^{**} \geq 1$. So $p \geq 0$.
- $F_C \subseteq F_A \wedge F_C \subseteq F_B$: Let $s_1 \in F_C$, then $s_1 \in F_A$ and $s_1 \in F_B$. Consider the character $f^{**} = f[s_1, ..., s_1]$. In both chain regions are no mutations. So

  $$l_{f^{**}}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) \text{ and } l_{f^{**}}(T_2) = l_{f_C}(T_C).$$

  leads to $p \geq 0$ by a similar argument as Equation 15.

Case 2b: $m_1^* = 1$. So either $s_1 \notin F_A$ or $s_1 \notin F_B$ but not both. Assume first that $s_1 \notin F_A$. Let $s_2 \in F_A$. If $s_2 \in F_B$, then consider $f[s_2, ..., s_2]$. With this character, there is no mutation in the chain region of T_1. So we have the same situation as case 2a. So assume $s_2 \notin F_B$. Consider $f^{**} = f[s_2, ..., s_2]$. This character leads to at most one mutation in the chain region of T_2 and one mutation in the chain region of T_1 (at e_B). Then

$$l_{f^{**}}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + 1 \text{ and } l_{f^{**}}(T_2) \leq l_{f_C}(T_C) + 1.$$

By a similarly argument as used by Equation 15, we have $p \geq 0$.
Now assume that $s_1 \notin F_B$. Let $s_2 \in F_B$. If $s_2 \in F_A$, then consider $f[s_2, ..., s_2]$. With this character, there is no mutation in the chain region of T_1. So we have the same situation as case 2a. So assume $s_2 \notin F_A$ and consider the character $f^{**} = f[s_2, ..., s_2]$. This leads to at most one mutation in the chain region of T_2 and one mutation in the chain region of T_1 (at e_A). Then

$$l_{f^{**}}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + 1 \text{ and } l_{f^{**}}(T_2) \leq l_{f_C}(T_C) + 1.$$

By a similarly argument as used by Equation 15, we have $p \geq 0$.

Case 2c: $m_1^* = 2$. We know that $m_2^* \leq 1$ (only possible mutation is at e_c). So

$$l_{f^*}(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + 2 \text{ and } l_{f^*}(T_2) \leq l_{f_C}(T_C) + 1$$

and by a similarly argument as used by Equation 15, we have $p \geq 0$.

So we know that $p \in (0, 1, 2)$ and it remains to prove that

$$d_{MP}(T_1', T_2') \geq l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + p$$

holds for these p, which will be done by considering three subcases.

Case 2.1: $p = 0$. Two subcases will be considered.

- $F_C \not\subseteq F_A$. Let $s_1 \in F_C \setminus F_A$. If $s_1 \in F_B$, then the character $f[s_2, s_1, s_1, s_1]$ with $s_2 \in F_A$ leads to one mutation in the chain regions of both tree (between p_1 and p_2 for T_1 and between x_1 and p_1 for T_2) and the difference in the number of mutations in the chain region is zero. If $s_1 \notin F_B$, then the character $f[s_2, s_2, s_1, s_1]$ with $s_2 \in F_A$ leads to two mutations in the chain regions of both tree (between p_2 and p_{k-1} and at e_B for T_1 and between p_2 and p_{k-1} and at e_C for T_2). Again, the difference in the number of mutations in the chain region is zero. So in both situations,

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_1) - l_{f'}(T'_2) \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) \end{aligned}$$

  and we are done.

- $F_C \subseteq F_A$. Let $s_1 \in F_C$ (clearly $s_1 \in F_A$). If $s_1 \in F_B$, then the character $f[s_1, s_1, s_1, s_1]$ leads to no mutation in the chain regions of both trees and the difference in the number of mutations in the chain region is zero. If $s_1 \notin F_B$, consider the character $f[s_1, s_1, s_2, s_2]$ with $s_2 \in F_B$. This leads to one mutation in the chain regions of both trees (between p_2 and p_{k-1}) and the difference in the number of mutations in the chain region is again zero. So in both situations,

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_1) - l_{f'}(T'_2) \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) \end{aligned}$$

  and we are done.

Case 2.2: $p = 1$. Consider the following subcases:

- $F_C \not\subseteq F_A$. Let $s_1 \in F_C \setminus F_A$. Note that $s_1 \in F_B$ because otherwise the character $f[s_1, ..., s_1]$ would lead to no mutations in the chain region of T_2 and two mutations in the chain region of T_1 (e_A and e_B), contradicting $p = 1$. So then the character $f' = f[s_1, s_1, s_1, s_1]$ leads to one mutation in the chain region of T'_1 and no mutations in the chain region of T'_2. So we have

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_1) - l_{f'}(T'_2) \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) + 1 - l_{f_C}(T_C) \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + p \end{aligned}$$

  and we are done.

- So suppose $F_C \subseteq F_A$. Let $s_1 \in F_C$ (clearly, $s_1 \in F_A$). Suppose first $F_C \subseteq F_B$, then $s_1 \in F_B$. So consider the character $f' = f[s_1, s_1, s_2, s_2]$ where $s_2 \notin F_B$. Then there are two mutations in T_1 and one mutation in T_2. Note that the case that there is no state $s_2 \notin F_B$ cannot occur. Namely by the less constrained roots argument 2.2 and since $F_C \subseteq F_A$ ($l_f(T_1) \geq l_f(T_2)$ and F_D empty), $p \leq 0$. So F_B cannot contain all states when $F_C \subseteq F_A$.

So suppose next $F_C \not\subseteq F_B$, which implies $s_1 \notin F_B$. Consider $f' = f[s_1, s_1, s_1, s_1]$. Then there is one mutation in the chain region of T_1 (at e_B) and no mutation in the chain region of T_2. In both cases, the difference in mutations in chain regions is one. Thus,

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_1) - l_{f'}(T'_2) \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + 1 \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + p \end{aligned}$$

and we are done.

Case 2.3: $p = 2$.

- $F_C \not\subseteq F_A$. Let $s_1 \in F_C \setminus F_A$. If $s_1 \in F_B$, then we know that there is some other state $s_2 \notin F_B$, because otherwise all states are in F_B and this cannot happen. Namely $F_D \subseteq F_B$ with F_D containing all states like in figure 8. Then by the less constraint roots argument (lemma 2.3 in section 2.4), $p \leq 1$, contradicting $p = 2$.
  So $F_C \not\subseteq F_A$ and $s_1 \in F_C \wedge F_B \setminus F_A$ and $s_2 \notin F_B$. Consider the character $f' = f[s_1, s_1, s_2, s_2]$. With this character f' there are three mutations in the chain region of T_1 (e_A, e_B and between p_2 and p_{k-1}) and only one mutation in the chain region of T_2 (between p_2 and p_{k-1}).
  If $s_1 \notin F_B$, consider $f' = f[s_1, s_1, s_1, s_1]$. This leads to two mutation in the chain region of T_1 and no mutations in the chain region of T_2. In both situations the difference in mutations is two and so we have

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + 2 \\ &= l_{f_A}(T_A) + l_{f_B}(T_B) - l_{f_C}(T_C) + p. \end{aligned}$$

- Suppose now that $F_C \subseteq F_A$. This case cannot occur. By the less constraint roots argument 2.3, $p \leq 1$, contradicting $p = 2$.

III: the common chain is pendant in both trees

There are two main situations here: the chains are oriented in the same direction (Figure 12), and the chains are oriented in the opposite direction (Figure 13). Whichever situation occurs, we can assume without loss of generality that $l_f(T_1) < l_f(T_2)$, so $d_{MP}(T_1, T_2) = l_f(T_2) - l_f(T_1)$ with f the optimal character. Everything is defined as before and as in Equation 4 we have,

$$d^r_{MP}(T_1, T_2) = l_{f_C}(T_C) - l_{f_A}(T_A) + p. \tag{16}$$

Note that we have $p \leq 1$ by the mimicking construction used in equation 5 (i.e. copying the states allocated to the red nodes in T_1, to T_2. We will now show that $p \geq 0$. Relabel f to new character $f^* = f[s_1, ..., s_1]$ with $s_1 \in F_A$. Then there is no mutation in the

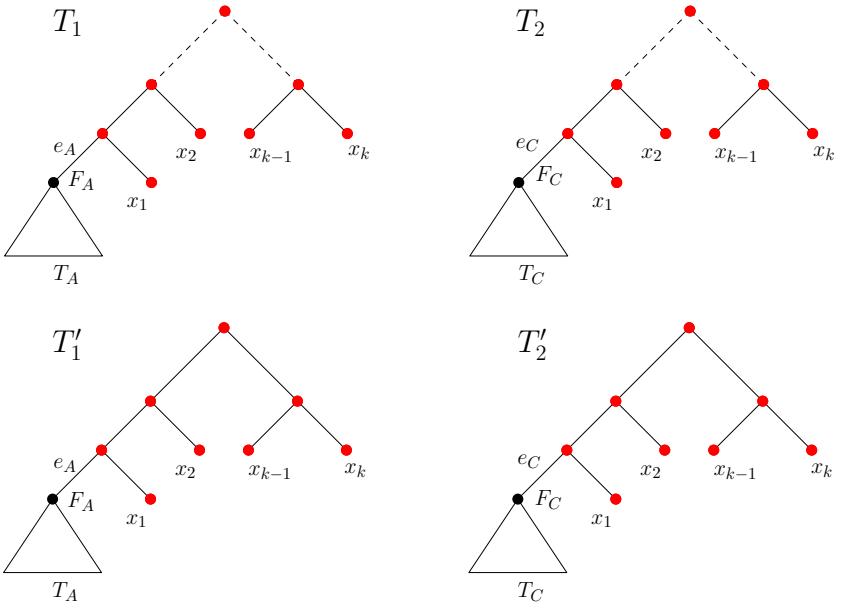


Figure 12: The chain reduction as applied in the case when the common chain K is pendant in both trees oriented in the same direction. Like in fig. 10 and fig. 11, the trees are unrooted but shown as rooted and the dotted line denote the taxa $\{x_3, ..., x_{k-2}\}$.

chain region of T_1 and at most one mutation in the chain region of T_2 (at e_C). So we have

$$l_{f^*}(T_1) = l_{f_A}(T_A) \text{ and } l_{f^*}(T_2) \geq l_{f_C}(T_C).$$

Since the optimality of f implies $l_f(T_2) - l_f(T_1) \geq l_{f*}(T_2) - l_{f*}(T_1)$, we have

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_C}(T_C) + l_{f_A}(T_A) \\ &\geq l_{f^*}(T_2) - l_{f^*}(T_1) - l_{f_C}(T_C) + l_{f_A}(T_A) \\ &\geq 0. \end{aligned} \tag{17}$$

Hence, $p \in (0, 1)$ and it remains to show that

$$d^r_{MP}(T_1', T_2') \geq l_{f_C}(T_C) - l_{f_A}(T_A) + p$$

holds, which can be done by considering two cases.

Case 1: $p = 0$. In this case we can take $f' = f[s_1, s_1, s_1, s_1]$ where $s_1 \in F_A$. Note that $s_1 \in F_C$ because otherwise $f[s_1, ..., s_1]$ would lead to no mutations in the chain region of T_1 but one mutation in the chain region of T_2, contradicting $p = 0$. So we are done since $s_1 \in F_C$ and $s_1 \in F_C$ implies that there are no mutations in both chain regions. So

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &= l_{f_C}(T_C) - l_{f_A}(T_A). \end{aligned}$$

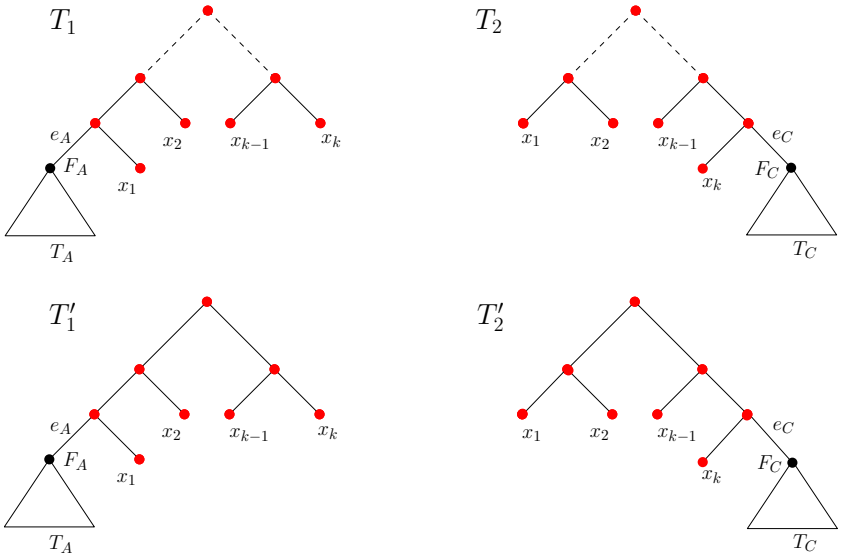


Figure 13: The chain reduction as applied in the case when the common chain K is pendant in both trees oriented in a different direction. Like in fig. 10 and fig. 11, the trees are unrooted but shown as rooted and the dotted line denote the taxa $\{x_3, ..., x_{k-2}\}$.

Case 2: $p = 1$. Consider two subcases, the case where the chains are oriented in the same direction and where they are oriented in different directions.

Case 2a: the chains are oriented in the same direction. In this case, we will look at two different subcases.

- $F_A \not\subseteq F_C$. Let $s_1 \in F_A \setminus F_C$. Consider the character $f' = f[s_1, s_1, s_1, s_1]$. Then there is no mutation in the chain region of T_1 and one mutation in the chain region of T_2 (at e_C). Then

$$\begin{aligned} d^r_{MP}(T_1', T_2') &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &= l_{f_C}(T_C) + 1 - l_{f_A}(T_A) \\ &= l_{f_C}(T_C) - l_{f_A}(T_A) + p \end{aligned}$$

  and we are done.

- $F_A \subseteq F_C$. This cannot hold by the less constrained roots argument, lemma 2.2 in section 2.4. The lemma implies that $p \leq 0$, contradicting $p = 1$.

Case 2a: the chains are oriented in different directions. Let $s_1 \in F_A$ and $s_2 \notin F_C$. Note that there is always a state $s_2 \notin F_C$ since the case that all states are in F_C cannot happen. The less constrained roots argument (lemma 2.2 in section 2.4) implies $p \leq 0$, a contradiction.
So the chains are oriented in different directions and $s_1 \in F_A$ and $s_2 \notin F_C$. Now consider

the character $f' = f[s_1, s_1, s_2, s_2]$. If $s_1 = s_2$, then there is no mutation in the chain region of T_1 and one mutation in the chain region of T_2 (at e_c). If $s_1 \neq s_2$, then there is one mutation in the chain region of T_1 (between p_2 and p_{k-1}) and two mutations in the chain region of T_2 (at e_c and between p_2 and p_{k-1}). In both situations is the difference one and we have:

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &= l_{f_C}(T_C) - l_{f_A}(T_A) + 1 \end{aligned}$$

□

In this section, we proved that we can reduce the length of a common chain to four. This is the best we can do. In the article by steven Kelk et al. [10], it is shown that we can not reduce the chain to length three. They showed a counterexample (figure 5, [10]) where the MP distance of the original trees are 2 and for the trees with common chain reduced to length three, the distance is 1.

4 Generalized subtree reduction

In this section, we will consider generalized subtree reduction. We will prove that this reduction rule can be applied without changing the bounded MP distance.
In two trees T_1 and T_2 you can have a common pendant subtree ignoring root location (i.r.l.) or/and a common pendant subtree. T_1 and T_2 have *common pendant subtree i.r.l.* on $X' \subset X$ if (i) for $i \in (1,2)$, T_i contains an edge e_i that induces a split $(X \setminus X')|X'$ in T_i and (ii) $T_1|_{X'} = T_2|_{X'}$.
Now assume for $i \in (1,2)$ v_i is the endpoint of e_i that is closest to X'. Then v_i can be used as root for the tree $T_i|_{X'}$. Denote this rooted tree by $(T_i|_{X'})^\rho$. T_1 and T_2 have a *common pendant subtree* on X' if $(T_1|_{X'})^\rho = (T_2|_{X'})^\rho$. Clearly, if T_1 and T_2 have a common pendant subtree, they also have a common pendant subtree i.r.l., but the other war around is not always true.

In figure 14, two trees who have a common pendant subtree as well as a common

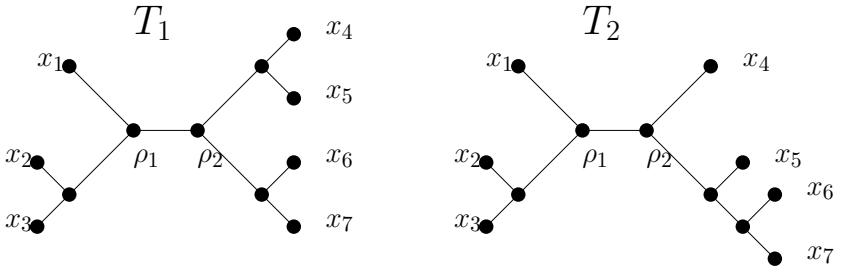


Figure 14: Two trees, T_1 and T_2, with a common pendant subtree on $\{x_1, x_2, x_3\}$, and a common pendant subtree i.r.l. on $\{x_4, x_5, x_6, x_7\}$. Here ρ_1 and ρ_2 are the roots of the subtrees.

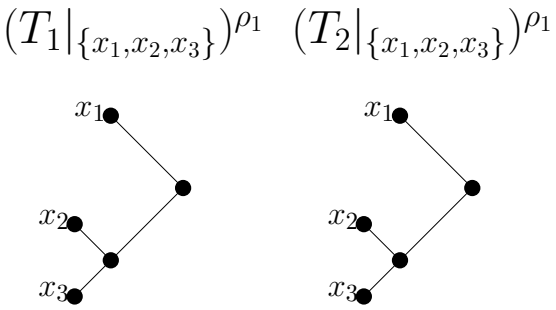


Figure 15: The common pendant subtrees on $\{x_1, x_2, x_3\}$ from T_1 and T_2.

subtree i.r.l. are shown. The trees have a common subtree on $\{x_1, x_2, x_3\}$ as can be seen in figure 15. The trees are exactly the same when rooted at ρ_1. In figure 16, the common subtree i.r.l. can be seen. As we can see, the trees are similar when they are unrooted, but they are not the same when rooted at ρ_2.
Now we can define *generalized subtree reduction* on two trees T_1 and T_2 on taxa set X.

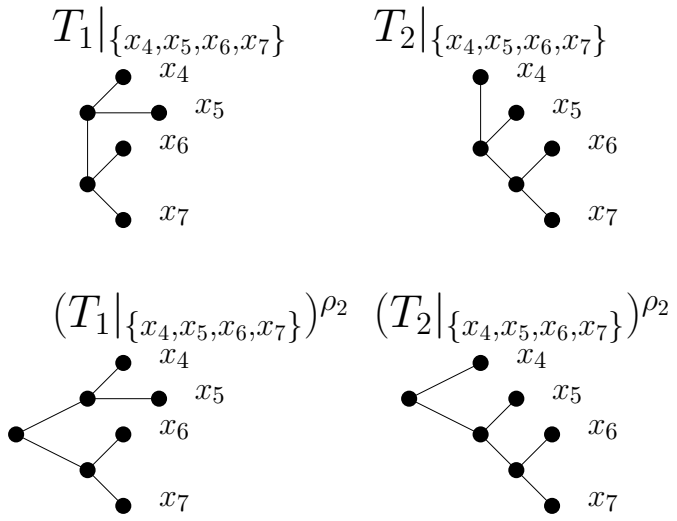


Figure 16: The common pendant subtrees i.r.l. on $\{x_4, x_5, x_6, x_7\}$ from T_1 and T_2. First the unrooted version $T_i|_{\{x_4,x_5,x_6,x_7\}}$ and second the rooted versions $(T_i|_{\{x_4,x_5,x_6,x_7\}})^{\rho_2}$.

Suppose T_1 and T_2 have a common pendant subtree i.r.l. on X' where $X' \subset X$ and $|X'| \geq 2$. We have two cases at which we can construct reduced trees T_1' and T_2'.

- *Traditional case:* T_1 and T_2 have a common pendant subtree on X'. Let $T_1' = T_1|_{(X \setminus X') \cup \{x\}}$ and $T_2' = T_2|_{(X \setminus X') \cup \{x\}}$ where $x \in X'$. (see figure 17)
- *Extended case:* T_1 and T_2 have no common pendant subtree on X' and $|X'| \geq 4$. Let x, y, z be distinct taxa in X' such that in $(T_1|_{X'})^{\rho}$, x and y are on one side of the root ρ, and z on the other, while in $(T_2|_{X'})^{\rho}$, x and z are on one side of the root ρ, and y on the other. These taxa always exist because $(T_1|_{X'})^{\rho} \neq (T_2|_{X'})^{\rho}$. Then let $T_1' = T_1|_{(X \setminus X') \cup \{x,y,z\}}$ and $T_2' = T_2|_{(X \setminus X') \cup \{x,y,z\}}$. (see figure 18)

Theorem 4.1. *Let T_1 and T_2 be two unrooted binary trees on taxa set X and suppose T_1' and T_2' are reduced trees after applying generalized subtree reduction to T_1 and T_2. Then $d^r_{MP}(T_1, T_2) = d^r_{MP}(T_1', T_2')$ for r an arbitrary integer.*

Proof. Note that $d^r_{MP}(T_1', T_2') \leq d^r_{MP}(T_1, T_2)$ follows from lemma 2.1 in section 3 and the definition of generalized subtree reduction. It remains to show that $d^r_{MP}(T_1', T_2') \geq d^r_{MP}(T_1, T_2)$.
We may assume $d^r_{MP}(T_1, T_2) > 0$ as otherwise the theorem clearly holds. Let T_A, T_B, T_C and T_D refer to the four subtrees of T_1 and T_2 as shown in figure 18 and figure 17. For $P \in \{A, B, C, D\}$, let X_P refer to the taxa in the subtree T_P. Let $X_B = X_D = X'$, $X_A = X_C = X \setminus X'$. Note that $T_1|_{X'} = T_2|_{X'}$. That is, T_B and T_D are identical subtrees ignoring the point at which each subtree is connected to the rest of its tree. As indicated in the figures, we root T_1 and T_2 by subdividing the edge that connects each pendant subtree to the rest of the tree. Let f_P denote the character obtained by restricting f to

X_P, and let F_P refer to the set of states assigned to the root of T_P by the Fitch map induced by f_P.
For $i \in (1,2)$, let $m_i = 0$ if the root of T_i is an intersection node, and $m_i = 1$ otherwise (i.e. the root is a union node). Then we have

$$l_f(T_1) = l_{f_A}(T_A) + l_{f_B}(T_B) + m_1 \text{ and } l_f(T_2) = l_{f_C}(T_C) + l_{f_D}(T_D) + m_2.$$

Note that we also have $l_{f_B}(T_B) = l_{f_D}(T_D)$ because T_B and T_D are (from an unrooted perspective) identical. Let f be an optimal character for T_1 and T_2 and assume that $l_f(T_1) < l_f(T_2)$, as the other case is symmetrical. So $d^r_{MP}(T_1, T_2) = l_f(T_2) - l_f(T_1)$. Let $p = m_2 - m_1$. Then we have

$$\begin{aligned} d^r_{MP}(T_1, T_2) &= l_f(T_2) - l_f(T_1) \\ &= (l_{f_C}(T_C) + l_{f_D}(T_D) + m_2) - (l_{f_A}(T_A) + l_{f_B}(T_B) + m_1) \\ &= l_{f_C}(T_C) - l_{f_A}(T_A) + p \end{aligned} \tag{18}$$

Now we know by definition of p, that $p \leq 1$ since $m_2 \leq 1$ and $m_1 \geq 0$. Now we will show that $p \geq 0$. Let s be a state such that $s \in F_A$. Consider now the character f^* obtained from modifying f by reassigning all the taxa in X' to the state s (note that $l_{f_B}(T_B) = l_{f_D}(T_D) = 0$). Then we have $l_{f^*}(T_1) = l_{f_A}(T_A)$. Note that $l_{f^*}(T_2) \geq l_{f_C}(T_C)$. So we can conclude that $d^r_{MP}(T_1, T_2) \geq l_{f^*}(T_2) - l_{f^*}(T_1) \geq l_{f_C}(T_C) - l_{f_A}(T_A)$, and hence $p \geq 0$.
In order to show $d^r_{MP}(T'_1, T'_2) \geq d^r_{MP}(T_1, T_2)$, by Equation 18 it suffices to show that

$$d^r_{MP}(T'_1, T'_2) \geq l_{f_C}(T_C) - l_{f_A}(T_A) + p \tag{19}$$

for $p \in \{0, 1\}$. We will now consider the two cases, the traditional case and the extended case.

I: The traditional case
To shorten notation we will write $f[a]$ to denote the character on $(X \setminus X') \cup \{x\}$ obtained from f by leaving the states assigned to taxa in $X_A = X_C = (X \setminus X')$ intact and assigning state a to x. We have two cases:
Case 1: $p = 0$. Let $s \in F_A$ and consider the character $f' = f[s]$. Then $l_{f'}(T'_1) = l_{f_A}(T_A)$ and $l_{f'}(T'_2) \geq l_{f_C}(T_C)$. This implies

$$\begin{aligned} d^r_{MP}(T_1, T_2) &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &\geq l_{f_C}(T_C) - l_{f_A}(T_A). \end{aligned} \tag{20}$$

Equation 19 follows from this and we are done.
Case 2: $p = 1$. Consider two subcases.

- $F_A \not\subseteq F_C$. Let $s \in F_A \setminus F_C$ and consider the character $f' = f[s]$. Then $l_{f'}(T'_1) = l_{f_A}(T_A)$ and $l_{f'}(T'_2) = l_{f_C}(T_C)+1$. Now equation 19 follows from a similar argument as equation 20 and we are done.

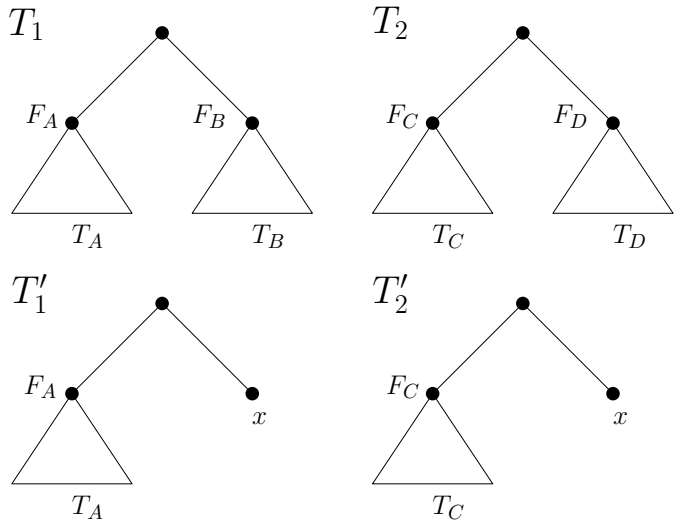


Figure 17: The generalized subtree reduction in the traditional case. That is, $(T_1|_{X'})^\rho = (T_2|_{X'})^\rho$. The trees are rooted in the edge between the common pendant subtree and the rest of the tree.

- $F_A \subseteq F_C$. This case cannot occur. By the less constrained roots argument, lemma 2.2, if $F_A \subseteq F_C \wedge F_B \subseteq F_D$, then $p \leq 0$. In this case, there is no F_B and F_D. However, we can define them as the set containing all states (so $F_B \subseteq F_D$) in the same way as figure 8 and described in the lemma. So $p \leq 0$, a contradiction to $p = 1$. And we are done.

II: The extended case
To shorten notation we will write $f[a, b, c]$ to denote the character on $(X \setminus X') \cup \{x, y, z\}$ obtained from f by leaving the states assigned to taxa in $X_A = X_C = (X \setminus X')$ intact and assigning states a, b, c to x, y, z respectively. We have two cases:
Case 1: $p = 0$. Let $s \in F_A$ and consider the character $f' = f[s, s, s]$. Then $l_{f'}(T'_1) = l_{f_A}(T_A)$ and $l_{f'}(T'_2) \geq l_{f_C}(T_C)$. This implies

$$\begin{aligned} d^r_{MP} &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &\geq l_{f_C}(T_C) - l_{f_A}(T_A). \end{aligned} \tag{21}$$

Equation 19 follows from this and we are done.
Case 2: $p = 1$. Consider two subcases.

- $F_A \not\subseteq F_C$. Let $s \in F_A \setminus F_C$ and consider the character $f' = f[s, s, s]$. Then $l_{f'}(T'_1) = l_{f_A}(T_A)$ and $l_{f'}(T'_2) = l_{f_C}(T_C) + 1$. Now equation 19 follows from a similar argument as equation 21 and we are done.

- $F_A \subseteq F_C$. Let $s_1 \in F_A$, then clearly $s_1 \in F_C$. Let $s_2 \notin F_C$. There is always such a state s_2 since otherwise F_C will contain all states, which cannot occur since then $m_2 = 0$. So $p = m_2 - m_1 = 0 - m_1 \leq 0$, contradicting $p = 1$.

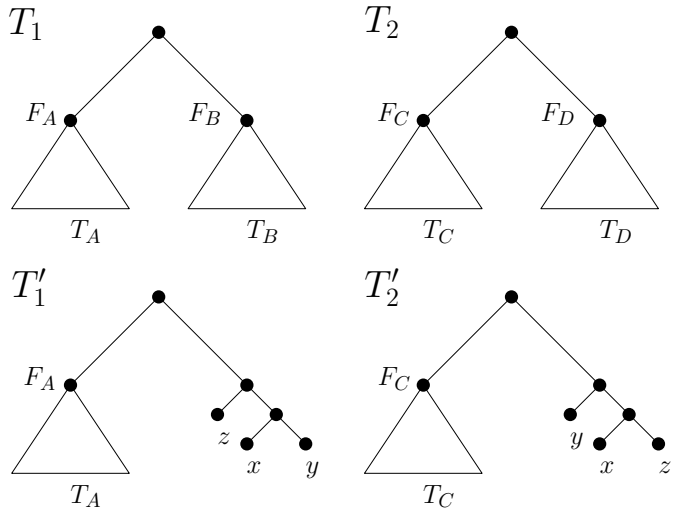


Figure 18: The generalized subtree reduction in the extended case. That is, $|X'| \geq 4$, $T_1|_{X'} = T_2|_{X'}$ but $(T_1|_{X'})^\rho \neq (T_2|_{X'})^\rho$. The trees are rooted in the edge between the common subtree and the rest of the tree.

Consider now the character $f' = f[s_2, s_2, s_1]$. This character f' introduces one mutation in the chain region of T_1' and two mutations in the chain region of T_2'. So we have

$$\begin{aligned} d^r_{MP} &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &\geq l_{f_C}(T_C) + 2 - l_{f_A}(T_A) - 1 \\ &= l_{f_c}(T_C) - l_{f_A}(T_A) + p. \end{aligned} \tag{22}$$

and we are done.

□

5 (2, 1, 2)-reduction

In a recent article by Steven Kelk and Simone Linz [12], five new reduction rules are introduced. In this section, I will look into the (2, 1, 2)-reduction.
Let T_1 and T_2 be two binary untooted trees with two common chains of length 2, $C_1 = (a, b)$ and $C_2 = (c, d)$. Then a *(2, 1, 2)-reduction* can be applied if T_1 has cherries (b, x) and (c, d) and T_2 has cherries (a, b) and (d, x) for some element $x \in X \setminus (C_1 \cup C_2)$. Such a (2, 1, 2)-reduction is the operation of deleting x from T_1 and T_2, i.e. we set $T_1' = T_1|_{X\setminus\{x\}}$ and $T_2' = T_2|_{X\setminus\{x\}}$. (see figure 20)
In the article by Steven Kelk and Simone Linz ([12]), it is proved that the TBR distance is reduced by one after applying any of the new rules including the $(2, 1, 2)$-reduction rule. Since the Maximum Parsimony distance has some relation to the TBR distance [6], it is likely that also the MP distance is reduced by one after applying this rule. However, it turns out that there are some cases for which it does not necessarily hold. Namely for the case where $F_A = \{s_1\}$, $F_B = \{s_1\}$, $F_C = \{s_2\}$, $F_D = \{s_1, s_2\}$ and the number of mutations among red nodes is 1 in the reduced trees (see figure 20 and the text in the proof of theorem 5.1 for the definitions) and the case where $F_A = \{s_1\}$, $F_B = \{s_1\}$, $F_C = \{s_2\}$, $F_D = \{s_2\}$ and the number of mutations among red nodes is 2 in the reduced trees. So instead of proving that the MP distance is reduced by one after applying the (2, 1, 2)-reduction rule, we found an example where the distance is preserved after applying (2,1,2)-reduction, a counterexample, as shown in figure 19.

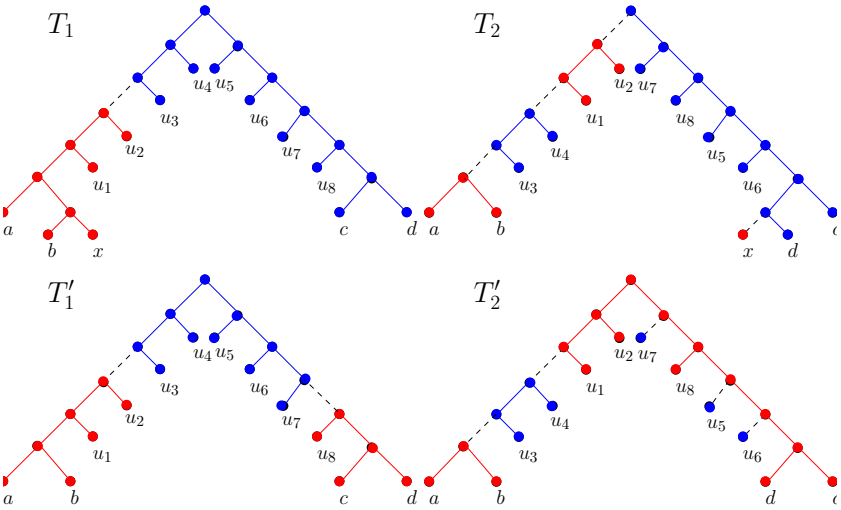


Figure 19: Two trees and the reduced versions of them after applying the (2, 1, 2)-reduction rule with a character f, and optimal extensions. The dotted lines are the lines where a mutation occurs, the red lines have both endpoint coloured red and the blue lines have both endpoints blue. This is a counterexample of theorem 5.1.

In the figure, there is an example of a character that gives a distance of 3 for both the reduced trees as the original trees. To verify that the distances are indeed equal, we used the java code from Steven Kelk [7] that calculates the Maximum Parsimony distance between certain trees (for more programs which calculates MP distances, see the site of Steven Kelk [8]). The program formulates a ILP, integer linear program (see [15] for more information about this). An ILP can easily be solved with help of a solver. We used GLPK package for this [13]. The program from Steven Kelk uses the Newick representation of a tree (see section 2). The Newick format of the trees from figure 19 are shown in table 5.

Trees	Newick format
T_1	((u4,(u3,(u2,(u1,(a,(b,x))))))),(u5,(u6,(u7,(u8,(c,d))))));
T_2	((u2,(u1,(u4,(u3,(a,b))))),(u7,(u8,(u5,(u6,(c,(d,x)))))));
T_1'	((u4,(u3,(u2,(u1,(a,b))))),(u5,(u6,(u7,(u8,(c,d))))));
T_2'	((u2,(u1,(u4,(u3,(a,b))))),(u7,(u8,(u5,(u6,(c,d))))));

Note that the code only calculates $max_f(l_f(T_2) - l_f(T_1))$ where T_1 is the first tree given in the text document and T_2 the second. So we have to calculate the distance twice and take the maximum of these. Also, we can specify the number of states used, in this case we will look at the d^2_{MP}. It turns out that indeed the distances are equal, $d^2_{MP}(T_1, T_2) = d^2_{MP}(T_1', T_2') = 3$ (see appendix for the outcome of the program). So the distance is not always reduced by one after applying the reduction rule, but maybe it is always preserved.
Unfortunately, it does not. We run the software also for the MP distance with three states, d^3_{MP}. It gives a distance of 4 for the original trees and 3 for the reduced trees. So $d^3_{MP}(T_1', T_2') = 3 = 4 - 1 = d^3_{MP}(T_1, T_2) - 1$. Now we can also conclude that the distance is not always preserved after (2,1,2)-reduction.
Despite that the distance is not always reduced by one after applying (2,1,2)-reduction, we can say something about the (bounded) MP distance after applying this rule.

Theorem 5.1. *Let T_1 and T_2 be binary unrooted trees and let T_1' and T_2' the trees obtained from T_1 and T_2 respectively by applying (2, 1, 2)-reduction. Then $d^r_{MP}(T_1, T_2) - 1 \leq d^r_{MP}(T_1', T_2') \leq d^r_{MP}(T_1, T_2)$ for any integer r.*

Proof. From lemma 2.1, we know that $d^r_{MP}(T_1', T_2') \leq d^r_{MP}(T_1, T_2)$ since the leaf set from T_1', T_2' a subset is from the leaf set of T_1 and T_2. So now we have to prove that $d^r_{MP}(T_1', T_2') \geq d^r_{MP}(T_1, T_2) - 1$.
Let f be the optimal character. Assume without loss of generalization that $l_f(T_2) \geq l_f(T_1)$. So we have $d^r_{MP}(T_1, T_2) = l_f(T_2) - l_f(T_1)$. Let T_A, T_B, T_C and T_D be the subtrees in T_1 and T_2 as shown in figure 20. For $P \in \{A, B, C, D\}$, let X_P refer to the taxa in subtree T_P. Note that $X_A \cup X_B = X_C \cup X_D$. Let u_P refer to the point where the cherries are connected to the rest of the tree (see figure 20. Let $f_{A,B}$ refer to the optimal character obtained by restricting f to X_A and X_B, and let $F_{A,B}$ refer to the set containing pairs of states (s_1, s_2). A pair (s_1, s_2) is in $F_{A,B}$ if there is a minimum

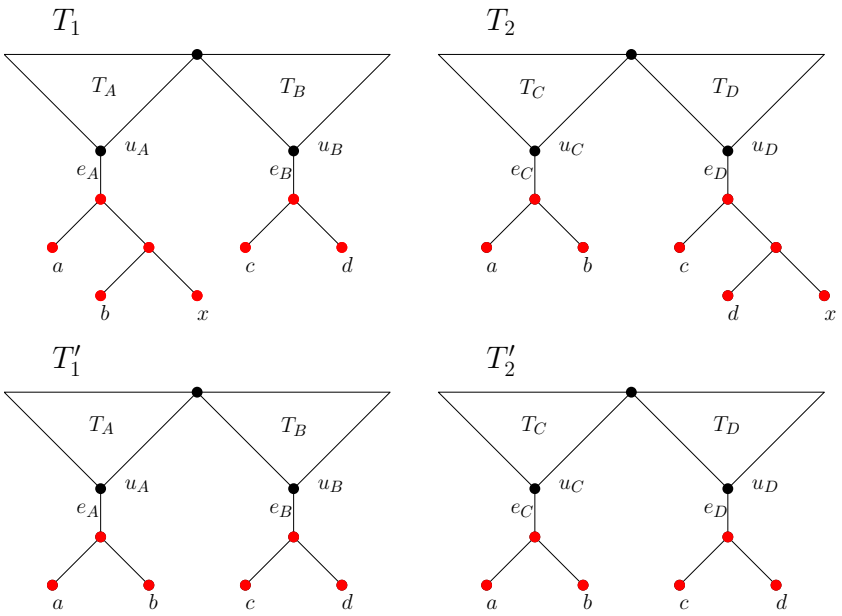


Figure 20: The $(2, 1, 2)$-reduction.

extension of $f_{A,B}$ that assigned s_1 to u_A and s_2 to u_B. Define F_A as $\{s_1 : (s_1, s_2) \in F_{A,B}\}$ and F_B as $\{s_2 : (s_1, s_2) \in F_{A,B}\}$. Define $f_{A,B}$, $F_{C,D}$, F_C and F_D similarly. Moreover, for each tree $T \in \{T_1, T_2, T_1', T_2'\}$ we define the chain region of T to be the set of edges incident to at least one red node. The red nodes are the nodes a, b, c, d, x and their parents, as shown in figure 20. Let m_i, $(i = 1, 2)$ be the number of mutations occurring in the chain region of T_i for a minimum extension of f. Then,

$$m_1 = l_f(T_1) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B)$$

$$m_2 = l_f(T_2) - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D)$$

In addition, let $p = m_2 - m_1$ and then we have

$$d^r_{MP}(T_1, T_2) = l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + p. \tag{23}$$

First we are going to prove $p \leq 3$. Consider an optimal extension f_1 of f to T_1. Let f_2 be an extension obtained by combining a minimum extension of $f_{C,D}$ to T_C and T_D and exactly mimicking f_1 on the red nodes of T_2. So the parents of c and d in T_2 both get the same state assigned to it as the parent of c in T_1. Now there are two cases, the parents of a and b in T_1 have the same state assigned by f_1 or they don't. In the first case, let the parent of a in T_2 have the same state assigned to it as the parents of a and b in T_1. In this case, compared with f_1, the extension f_2 creates at most three mutations extra on the chain region of T_2, namely at e_C, at e_D and between x and the parent of d. So we have $\Delta(f_2) \leq l_{f_{C,D}}(T_C) + l_{f_{c,D}}(T_D) + (m_1 + 3)$. Together with $l_f(T_2) \leq \Delta(f_2)$

and $l_f(T_1) = l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) + m_1$, this implies

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) + l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) \\ &= l_f(T_2) - m_1 - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) \\ &\leq \Delta(f_2) - m_1 - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) \\ &\leq 3. \end{aligned} \tag{24}$$

In the other case, where the parents of a and b in T_1 have another state assigned by f_1, let the parent of a and b in T_2 have the same state as the parent of a. Now, compared with f_1, the extension f_2 creates at most four mutations extra on the chain region of T_2, namely at e_C, between b and its parent, at e_D and between x and the parent of d. Note that the chain region of T_2 has also one mutation less compared to T_1. In T_1, there is a mutation on the edge between the parent of a and the parent of b, which cannot occur in T_2 since this edge does not exist in T_2. So again we have $\Delta(f_2) \leq l_{f_{C,D}}(T_D) + l_{f_{C,D}}(T_D) + (m_1 + 3)$ and this implies $p \leq 3$ by the same arguments as in equation 24.
Now we will show $p \geq 1$. Let f^* be the character obtained from f, by reassigning taxa a, b, x to a state s_1 and reassigning c, d to a state s_2 with $(s_1, s_2) \in F_B$. Now there are two cases, $s_1 = s_2$ and $s_1 \neq s_2$.
In the situation that $s_1 \neq s_2$, there is no mutation in the chain region of T_1 and at least one mutation in the chain region of T_2 (between x and its parent). So we have

$$l_{f^*}(T_1) = l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) \text{ and } l_{f^*}(T_2) \geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + 1. \tag{25}$$

The optimality of f implies that $l_f(T_2) - l_f(T_1) \geq l_{f^*}(T_2) - l_{f^*}(T_1)$. Then by equation 25 and this inequality, we have

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) + l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) \\ &\geq l_{f^*}(T_2) - l_{f^*}(T_1) - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) + l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) \\ &\geq 1. \end{aligned} \tag{26}$$

In the case that $s_1 = s_2$, let $s_3 \neq s_1$ an existing state and consider the character f^{**} obtained from f, by reassigning taxa a, c, d to s_1 and reassigning b, x to s_3. Then there is one mutation in the chain region of T_1 (between the parent of b and the parent of a) and at least two mutations in the chain region of T_2 (between b and its parent (or a) and between x and its parent). So we have

$$l_{f^{**}}(T_1) = l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) + 1 \text{ and } l_{f^{**}}(T_2) \geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + 2. \tag{27}$$

The optimality of f implies that $l_f(T_2) - l_f(T_1) \geq l_{f^{**}}(T_2) - l_{f^{**}}(T_1)$. Then by equation 27 and this inequality, we have

$$\begin{aligned} p &= l_f(T_2) - l_f(T_1) - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) + l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) \\ &\geq l_{f^{**}}(T_2) - l_{f^{**}}(T_1) - l_{f_{C,D}}(T_C) - l_{f_{C,D}}(T_D) + l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) \\ &\geq 1. \end{aligned} \tag{28}$$

By equation 23, the claim follows from

$$d^r_{MP}(T'_1, T'_2) \geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + p - 1 \tag{29}$$

So to prove the theorem, it is sufficient to show Equation 29. Then there are three cases to consider, namely $p = 1$, $p = 2$ or $p = 3$, since $1 \leq p \leq 3$. X' is the taxa of the trees T'_1 and T'_2. For short notation, we will write $f[s_a, s_b, s_c, s_d]$ to denote the character on X' obtained from f by leaving the states assigned to taxa in $X_A \cup X_B = X_C \cup X_D$ intact and assigning states s_a, s_b, s_c, s_d to a, b, c, d respectively.

Case 1: $p = 1$. Let $(s_1, s_2) \in F_{A,B}$ and consider the character $f' = f[s_1, s_1, s_2, s_2]$. Then there is no mutation in the chain region of T'_1, so $l_{f'}(T'_1) = l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B)$. Note that $l_{f'}(T'_2) \geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D)$. Thus,

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &= l_f(T'_2) - l_f(T'_1) \\ &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &\geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B)) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + p - 1 \end{aligned} \tag{30}$$

and we are done.

Case 2: $p = 2$. We will consider two subcases.

- $F_A \not\subseteq F_C$. Let $(s_1, s_2) \in F_{A,B}$ with $s_1 \notin F_C$. Consider the character $f' = f[s_1, s_1, s_2, s_2]$. With this character there is no mutation in the chain region of T'_1 and at least one mutation in the chain region of T'_2 (at e_C). Then we have

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &= l_f(T'_2) - l_f(T'_1) \\ &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &\geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + 1 - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B)) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + p - 1 \end{aligned} \tag{31}$$

  and we are done.

- $F_A \subseteq F_C$. Let $(s_1, s_2) \in F_{A,B}$ (clearly $s_1 \in F_C$). Now two situations can occur. First assume that $(s_1, s_2) \notin F_{C,D}$. Consider in this situation the character $f' = f[s_1, s_1, s_2, s_2]$. With this character there is no mutation in the chain region of T'_1 and at least one mutation in the chain region of T'_2 (at e_D or at e_C). Then we have

$$\begin{aligned} d^r_{MP}(T'_1, T'_2) &= l_f(T'_2) - l_f(T'_1) \\ &\geq l_{f'}(T'_2) - l_{f'}(T'_1) \\ &\geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + 1 - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B)) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + p - 1. \end{aligned} \tag{32}$$

So we are done.
Now assume that $(s_1, s_2) \in F_{C,D}$. This situation cannot occur. Consider a minimum extension f_1 of f to T_1 that maps s_1 to u_A (and s_2 to u_B). Since $(s_1, s_2) \in F_{C,D}$, there exists a minimum extension $\overline{f}_{C,D}$ of $f_{C,D}$ to T_C and T_D that assigns s_1 to u_C and s_2 to u_D. Consider now an extension f_2 obtained from combining $\overline{f}_{C,D}$ and exactly mimicking the red nodes of f_1 to T_2. Then the mutations in the chain region of T_2 is equal to the number of mutations in the chain region of T_1 plus one (between x and its parent). In other words, $\Delta(f_2) = l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + m_1 + 1$. So we have

$$\begin{aligned} d^r_{MP}(T_1, T_2) &= l_f(T_2) - l_f(T_1) \\ &\leq \delta(f_2) - l_f(T_1) \\ &\leq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + m_1 + 1 - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) + m_1) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + 1. \end{aligned}$$

This shows in particular $p \leq 1$, a contradiction.

Case 3: $p = 3$. We will consider two subcases.

- $F_A \subseteq F_C$. This case cannot occur. Consider a minimum extension f_1 of f to T_1. Then this f_1 assigns a state a from F_A to u_A. Since $a \in F_C$, there exists a minimum extension $\overline{f_{C,D}}$ of $f_{C,D}$ to T_C and T_D such that this extension $\overline{f_{C,D}}$ assigns the state a to u_C. Consider now an extension f_2 obtained from $\overline{f_{C,D}}$ and exactly mimicking f_1 to the red nodes of T_2. Then the mutations in the chain region of T_2 is less or equal to the number of mutations in the chain region of T_1 plus two (between x and its parent and at e_D). In other words, $\Delta(f_2) \leq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + m_1 + 2$. So we have

  $$\begin{aligned} d^r_{MP}(T_1, T_2) &= l_f(T_2) - l_f(T_1) \\ &\leq \delta(f_2) - l_f(T_1) \\ &\leq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + m_1 + 2 - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) + m_1) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + 2. \end{aligned}$$

  This shows in particular $p \leq 2$, a contradiction.

- $F_A \not\subseteq F_C$. Let $(s_1, s_2) \in F_{A,B}$ with $s_1 \notin F_C$. Now we will prove that $s_2 \notin F_D$. Suppose that $s_2 \in F_D$. Then consider a minimum extension f_1 of f to T_1 that assigns the state s_1 to u_A and s_2 to u_B. Since $s_2 \in F_D$, there exists a minimum extension $\overline{f_{C,D}}$ of $f_{C,D}$ to T_C and T_D such that this extension $\overline{f_{C,D}}$ assigns the state s_2 to u_D. Consider now an extension f_2 obtained from $\overline{f_{C,D}}$ and exactly mimicking f_1 to the red nodes of T_2. Then the mutations in the chain region of T_2 is less or equal to the number of mutations in the chain region of T_1 plus two (between x and its parent and at e_C). In other words, $\Delta(f_2) \leq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + m_1 + 2$.

So we have

$$\begin{aligned} d^r_{MP}(T_1, T_2) &= l_f(T_2) - l_f(T_1) \\ &\leq \Delta(f_2) - l_f(T_1) \\ &\leq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + m_1 + 2 - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B) + m_1) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + 2. \end{aligned}$$

This shows in particular $p \leq 2$, a contradiction. So $s_2 \notin F_D$.
So we know that for all pairs $(s_3, s_4) \in F_{C,D}$, $s_1 \neq s_3$ and $s_2 \neq s_4$. Consider the character $f' = f[s_1, s_1, s_2, s_2]$. This character leads to no mutations in the chain region of T_1' and two mutations in the chain region of T_2' (at e_C and e_D). Thus we have

$$\begin{aligned} d^r_{MP}(T_1', T_2') &= l_f(T_2') - l_f(T_1') \\ &\geq l_{f'}(T_2') - l_{f'}(T_1') \\ &\geq l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) + 2 - (l_{f_{A,B}}(T_A) + l_{f_{A,B}}(T_B)) \\ &= l_{f_{C,D}}(T_C) + l_{f_{C,D}}(T_D) - l_{f_{A,B}}(T_A) - l_{f_{A,B}}(T_B) + p - 1 \end{aligned} \tag{33}$$

and we are done.

So we proved $d^r_{MP}(T_1', T_2') \geq d^r_{MP}(T_1, T_2) - 1$ and the theorem holds. □

6 Conclusion

The Maximum Parsimony (MP) distance is a distance between two trees [9]. There are two types of MP distances, unbounded and bounded. In this report, only the bounded distance is considered. The MP distance is related to the TBR distance [6]. For the TBR distance there are some reduction rules that reduces the trees without changing its TBR distance [1], [2]. This is quite handy, because it makes the calculations easier. Since there is a relation between the two distances, we suspected that these reduction rules can also be applied without changing the Maximum Parsimony distance. In this report, three different reduction rules were considered, chain reduction, generalized subtree reduction and $(2, 1, 2)$-reduction.

By chain reduction, a common chain except two leaves at the beginning and two at the end of the chain can be deleted from both trees. So the common chains of the trees are then reduced to four (see exact definition in section 3). In paper [10], it was proven that the unbounded MP distance is preserved after applying chain reduction. We have used this proof and adapt it, to prove that also the bounded Maximum Parsimony distance is preserved after applying chain reduction. The chain in this case is reduced to length four which is the best we can do. In the paper by Steven Kelk et al [10], a counterexample for reduction of the chain to length three was given.

The second reduction we considered, is the generalized subtree reduction. The exact rule is dependent on the type of subtree. The subtree can be a pendant subtree ignoring root location (i.r.l) or just a pendant subtree. The subtree is not i.r.l. when the subtrees are also common when rooted at the end of the edges that induces the split. In this case, you can reduce the tree by removing the whole subtree except one leaf. For the pendant subtree i.r.l., you can remove the subtree except three leaves (see exact definition in section 4). In paper [10], it was proven that the unbounded MP distance is preserved after applying subtree reduction. Like chain reduction, we have adapt this proof, to prove that also the bounded Maximum Parsimony distance is preserved after applying generalized subtree reduction.

At last we considered a rule from the paper [12], the $(2, 1, 2)$-reduction. (see exact definition in section 5) In that paper, only a proof that the TBR distance is reduced by one after applying this rule, is given. So to prove that this is also the case for the MP distance, we used the same methods as used in the proofs for the chain reduction and subtree reduction. However, it turned out that the MP distance is not necessarily reduced with one after applying $(2, 1, 2)$-reduction. We found a counterexample, see figure 19 in section 5. We do know that the bounded MP distance is either preserved or reduced with one after applying $(2, 1, 2)$-reduction.

There is a lot more you can investigate on this subject. It is for example interesting to see if the MP distance is preserved or reduced with one after applying $(2, 1, 2)$-reduction for all bounds. The counterexample given, works only for a bound of two. So can we say more about the MP distance on three states after applying $(2, 1, 2)$-reduction? And what about the unbounded MP distance?

In paper [12], in total five reduction rules are introduced. So another thing to look into

is what is happening with the MP distance after applying the other four rules.
Furthermore, it is interesting to see if the kernel changes, now we know that these rules can/cannot be applied. The kernel is a problem that is the same as your original problem but easier to solve/calculate. The size of the kernel, time to solve the problem, is related to a parameter, for example the MP distance (see for more information on kernels [5]). With the proof in section 3 in this report, it is already proven that there exist a kernel for d^2_{MP} which was not known before. This is done by Elise Deen, Leo van Iersel, Remie Janssen, Mark Jones, Yuki Murakami and Norbert Zeh and will be published in a forthcoming paper.

References

[1] B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of combinatorics*, 2001.

[2] R. Atkins and C. McDiarmid. Extremal distances for subtree transfer operations in binary trees. *Annals of Combinatorics*, 23(1), 2019. Cited By :3.

[3] M. L. Bonet and K. St. John. On the complexity of uspr distance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):572–576, 2010. Cited By :13.

[4] J Felsenstein. The newick tree format. `http://evolution.genetics.washington.edu/phylip/newicktree.html`. (1986).

[5] F.V. Fomin, D. Lokshtanov, S. Saurabh, and M. Zehavi. *Kernelization, Theory of Parameterized Preprocessing*. Cambridge University Press, 2018.

[6] M. Jones, S. Kelk, and L. Stougie. Maximum parsimony distance on phylogenetic trees: Alinear kernel and constant factor approximation algorithm. *Journal of computer and system siences*, 2020.

[7] S. Kelk. Program that calculates the bounded or unbounded maximum parsimony distance. `http://skelk.sdf-eu.org/mpdistbinary/`. Accessed: 16-06-2021.

[8] S. Kelk. Steven kelk's homepage. `http://skelk.sdf-eu.org/`. Accessed: 17-06-2021.

[9] S. Kelk and M. Fischer. On the maximum parsimony distance between phylogenetic trees. *Annals of Combinatorics*, 2015.

[10] S. Kelk, M. Fischer, V. Moulton, and T. Wu. Reduction rules for the maximum parsimony distance on phylogenetic trees. *Theoretical Computer Science*, 2016.

[11] 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. Cited By :2.

[12] S. Kelk and S. Linz. New reduction rules for the tree bisection and reconnection distance. *Annals of combinatorics*, 2020.

[13] A. Makhorin. Gnu linear programming kit. `https://www.gnu.org/software/glpk/`. Accessed: 17-06-2021.

[14] V. Moulton and T. Wu. A parsimony-based metric for phylogenetic trees. *Advances in Applied Mathematics*, 2015.

[15] A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[16] C. Semple, M. Steel, et al. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.

Appendices

Result Program

$max_f(l_f(T_2) - l_f(T_1)$ for two states:
Long-step dual simplex will be used
+ 285: mip = not found yet <= +inf (1; 0)
+ 2915: >>>>> 1.000000000e+00 <= 8.000000000e+00 700.0% (158; 23)
+ 8564: >>>>> 2.000000000e+00 <= 5.000000000e+00 150.0% (443; 147)
+ 13428: >>>>> 3.000000000e+00 ¡= 4.000000000e+00 33.3% (536; 416)
+ 18794: mip = 3.000000000e+00 <= tree is empty 0.0% (0; 1941)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 1.9 secs
Memory used: 2.2 Mb (2269812 bytes)

$max_f(l_f(T_1) - l_f(T_2)$ for two states:
Long-step dual simplex will be used
+ 277: mip = not found yet <= +inf (1; 0)
+ 3511: >>>>> -1.110223025e-16 <= 7.000000000e+00 (186; 39)
+ 5066: >>>>> 2.000000000e+00 <= 6.000000000e+00 200.0% (260; 71)
+ 6532: >>>>> 3.000000000e+00 <= 6.000000000e+00 100.0% (283; 174)
+ 15142: mip = 3.000000000e+00 <= tree is empty 0.0% (0; 1355)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 1.5 secs
Memory used: 1.6 Mb (1662090 bytes)

$max_f(l_f(T_1') - l_f(T_2')$ for two states:
Long-step dual simplex will be used
+ 283: mip = not found yet <= +inf (1; 0)
+ 2566: >>>>> 2.000000000e+00 <= 7.000000000e+00 250.0% (131; 26)
+ 11659: >>>>> 3.000000000e+00 <= 3.000000000e+00 0.0% (285; 395)
+ 11659: mip = 3.000000000e+00 <= tree is empty 0.0% (0; 1255)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 1.2 secs
Memory used: 1.7 Mb (1807938 bytes)

$max_f(l_f(T_2') - l_f(T_1')$ for two states:
Long-step dual simplex will be used
+ 283: mip = not found yet <= +inf (1; 0)
+ 2566: >>>>> 2.000000000e+00 <= 7.000000000e+00 250.0% (131; 26)
+ 11659: >>>>> 3.000000000e+00 <= 3.000000000e+00 0.0% (285; 395)
+ 11659: mip = 3.000000000e+00 <= tree is empty 0.0% (0; 1255)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 1.2 secs

Memory used: 1.7 Mb (1807938 bytes)

$max_f(l_f(T_2) - l_f(T_1)$ for three states:
Long-step dual simplex will be used + 436: mip = not found yet <= +inf (1; 0)
+ 6855: >>>>> 1.000000000e+00 <= 1.000000000e+01 900.0% (256; 17)
+ 16234: >>>>> 2.000000000e+00 <= 9.000000000e+00 350.0% (610; 57)
+ 25121: >>>>> 3.000000000e+00 <= 8.000000000e+00 166.7% (896; 168)
+ 57508: mip = 3.000000000e+00 <= 7.000000000e+00 133.3% (1905; 526)
+ 59024: >>>>> 4.000000000e+00 <= 7.000000000e+00 75.0% (1963; 537)
+ 93685: mip = 4.000000000e+00 <= 7.000000000e+00 75.0% (2360; 1650)
+127789: mip = 4.000000000e+00 <= 7.000000000e+00 75.0% (3144; 1951)
+167308: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (3702; 2468)
+204530: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (4088; 3011)
+241730: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (4486; 3518)
+278029: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (4858; 4023)
+323826: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (4645; 5805)
+380363: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (3406; 9717)
+434861: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (2317; 13364)
+486363: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (1338; 17036)
Time used: 60.0 secs. Memory used: 16.1 Mb.
+538886: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (404; 21039)
+559671: mip = 4.000000000e+00 <= tree is empty 0.0% (0; 23297)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 65.9 secs
Memory used: 16.3 Mb (17116189 bytes)

$max_f(l_f(T_1) - l_f(T_2)$ for three states:
Long-step dual simplex will be used
+ 430: mip = not found yet <= +inf (1; 0)
+ 2605: >>>>> 3.000000000e+00 <= 1.100000000e+01 266.7% (85; 4)
+ 36056: mip = 3.000000000e+00 <= 8.000000000e+00 166.7% (1148; 186)
+ 69177: mip = 3.000000000e+00 <= 7.000000000e+00 133.3% (2178; 386)
+ 81687: >>>>> 4.000000000e+00 <= 7.000000000e+00 75.0% (2554; 467)
+113703: mip = 4.000000000e+00 <= 7.000000000e+00 75.0% (2524; 2074)
+148156: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (2983; 2502)
+181066: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (3355; 2921)
+213672: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (3682; 3348)
+245319: mip = 4.000000000e+00 <= 6.000000000e+00 50.0% (4008; 3772)
+285871: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (3961; 5078)
+334283: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (2967; 8141)
+382409: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (2041; 11312)
+428021: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (1147; 14593)
Time used: 60.0 secs. Memory used: 13.6 Mb.

+479303: mip = 4.000000000e+00 <= 5.000000000e+00 25.0% (261; 18437)
+493799: mip = 4.000000000e+00 <= tree is empty 0.0% (0; 20091)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 63.8 secs
Memory used: 13.8 Mb (14522315 bytes)

$max_f(l_f(T_2') - l_f(T_1')$ for three states:
Long-step dual simplex will be used
+ 418: mip = not found yet <= +inf (1; 0)
+ 2645: >>>>> 1.000000000e+00 <= 1.000000000e+01 900.0% (95; 4)
+ 3757: >>>>> 2.000000000e+00 <= 9.000000000e+00 350.0% (130; 23)
+ 41295: mip = 2.000000000e+00 <= 7.000000000e+00 250.0% (1420; 244)
+ 79505: mip = 2.000000000e+00 <= 6.000000000e+00 200.0% (2756; 539)
+115795: mip = 2.000000000e+00 <= 6.000000000e+00 200.0% (3975; 818)
+122917: >>>>> 3.000000000e+00 <= 5.000000000e+00 66.7% (4228; 874)
+166737: mip = 3.000000000e+00 <= 5.000000000e+00 66.7% (3354; 4125)
+208230: mip = 3.000000000e+00 <= 5.000000000e+00 66.7% (3813; 4761)
+248111: mip = 3.000000000e+00 <= 5.000000000e+00 66.7% (4316; 5401)
+310023: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (2968; 9860)
+367334: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (1738; 14241)
+423706: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (595; 19057)
+455060: mip = 3.000000000e+00 <= tree is empty 0.0% (0; 22261)
INTEGER OPTIMAL SOLUTION FOUND
Time used: 49.3 secs
Memory used: 14.1 Mb (14777302 bytes)

$max_f(l_f(T_1') - l_f(T_2')$ for three states:
Long-step dual simplex will be used
+ 418: mip = not found yet <= +inf (1; 0)
+ 2645: >>>>> 1.000000000e+00 <= 1.000000000e+01 900.0% (95; 4)
+ 3757: >>>>> 2.000000000e+00 <= 9.000000000e+00 350.0% (130; 23) +
41579: mip = 2.000000000e+00 <= 7.000000000e+00 250.0% (1433; 246)
+ 78510: mip = 2.000000000e+00 <= 6.000000000e+00 200.0% (2729; 529)
+109797: mip = 2.000000000e+00 <= 6.000000000e+00 200.0% (3767; 774)
+122917: >>>>> 3.000000000e+00 <= 5.000000000e+00 66.7% (4228; 874)
+163251: mip = 3.000000000e+00 <= 5.000000000e+00 66.7% (3320; 4078)
+197517: mip = 3.000000000e+00 <= 5.000000000e+00 66.7% (3706; 4599)
+233673: mip = 3.000000000e+00 <= 5.000000000e+00 66.7% (4157; 5156)
+285055: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (3587; 7932)
+342297: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (2262; 12298)
+397605: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (1130; 16690)
+452588: mip = 3.000000000e+00 <= 4.000000000e+00 33.3% (49; 21853)
+455060: mip = 3.000000000e+00 <= tree is empty 0.0% (0; 22261)

INTEGER OPTIMAL SOLUTION FOUND
Time used: 52.8 secs
Memory used: 14.1 Mb (14777302 bytes)