

# **Subtree Transfer Operations and their Induced Metrics on Evolutionary Trees**

**BENJAMIN L. ALLEN and MIKE STEEL**

*Biomathematics Research Centre  
University of Canterbury  
Private Bag 4800  
Christchurch, New Zealand*

No. 170

January, 1999

# Subtree Transfer Operations and their Induced Metrics on Evolutionary Trees

Benjamin L. Allen      Mike Steel\*

January 28, 1999

## Abstract

Leaf-labelled trees are widely used to describe evolutionary relationships, particularly in biology. In this setting, extant species label the leaves of the tree, while the internal vertices correspond to ancestral species. Various techniques exist for reconstructing these evolutionary trees from data, and an important problem is to determine how “far apart” two such reconstructed trees are from each other, or indeed from the *true* historical tree. To investigate this question requires tree metrics, and these can be induced by operations that rearrange trees locally. Here we investigate three such operations: *nearest neighbour interchange* (NNI), *subtree prune and regraft* (SPR), and *tree bisection and reconnection* (TBR). The SPR operation is of particular interest as it can be used to model biological processes such as horizontal gene transfer and recombination. We count the number of unrooted binary trees one SPR from any given unrooted binary tree, as well as providing new upper and lower bounds for the diameter of the adjacency graph of trees under SPR and TBR. We also show that the problem of computing the minimum number of TBR operations required to transform one tree to another can be reduced to a problem whose size is a function just of the distance between the trees (and not of the size of the two trees), and thereby establish that the problem is *fixed-parameter tractable*.

**Keywords:** Trees, metrics, subtree transfer, fixed parameter tractability.

## 1 Introduction

Leaf-labelled trees are widely used to represent evolutionary relationships, particularly in biology, but also in other areas of classification (including linguistics and philology). Typically a set  $S$  of extant (present day) species label the leaves (degree one vertices) of the tree and the remaining vertices represent ancestral species. A root vertex may be present, which corresponds to the most recent ancestor of the species under study. It is usually assumed that each “speciation” event leads simply to the appearance of one new lineage, and thus, in this directed tree, each vertex has exactly two outgoing edges.

Given data (such as aligned DNA sequences), numerous methods exist for reconstructing a tree (see [12]) that hopefully approximates the true historical tree of descent of the species under study. However, different data sets and different methods often lead to different trees being reconstructed for the same set of species. Thus it becomes imperative to determine how “close” two reconstructed trees are. This requires the introduction of metrics on trees. Several such metrics have been considered (see, for example, [9]). A particularly natural choice is to say that two trees are “close together” if one can be obtained from the other by a small number of “local” operations. Typically, three types of local rearrangements have been studied and we will consider these in detail in the next chapter. However, little is known about how pairs of trees are distributed according to these metrics, or even how to efficiently calculate them. In this paper we investigate both questions. In particular we:

- establish new results on the diameter and density of the adjacency graph of unrooted trees under the *subtree prune and regraft* and *tree bisection and reconnection* operations, thereby correcting an oversight in [10];

---

\*Biomathematics Research Centre, University of Canterbury, New Zealand. m.steel@math.canterbury.ac.nz

- establish a relationship between the number of *tree bisection and reconnection* operations required to transform one tree into another and the size of the maximum agreement forest for the two trees, thereby correcting an error in [6];
- investigate the computational complexity of the NNI, SPR and TBR Distance Problems, and point out that the TBR-Distance Problem is *NP*-hard;
- show that, for the *tree bisection and reconnection* operation, the question of whether a given unrooted binary tree can be transformed to another given unrooted tree by at most  $k$  operations, namely the Parameterized TBR-Distance Problem, is *fixed parameter tractable* (FPT), and conjecture that the Parameterized SPR-Distance Problem is *FPT* as well.

Two further motivations for analysing these tree edit operations are that (i) they form the basis of tree reconstruction heuristics that attempt to locate the “best” tree according to various criteria (see [9]), and (ii) one of the tree edit operations, the SPR, is useful for modelling horizontal gene transfer and recombination events (see [3], [4], [5], [6] and [8]).

However before we investigate any tree edit operations, we need to introduce technical definitions.

### Definitions

1. An *unrooted binary phylogenetic tree* (or more briefly a *binary tree*) is a tree whose leaves (degree 1 vertices) are labelled bijectively by a (species) set  $S$ , and such that each non-leaf vertex is unlabelled and has degree three. We let  $UB(n)$  denote the set of such trees for  $S = \{1, \dots, n\}$ .
2. An edge of a tree incident with a leaf is a *pendant edge*, otherwise we say it is an *internal edge*. Let  $\mathcal{L}(T)$  denote the leaf set of a tree  $T$ ; the other vertices are said to be *internal*. A *cherry* of a tree  $T$  is subtree containing exactly two leaves and their associated pendant edges, along with the vertex to which both pendant edges are incident.
3. A *forced contraction* is an operation on a tree in which we delete a vertex  $v$  of degree two and replace the two edges incident to  $v$  by a single edge. Given a set  $U \subseteq \mathcal{L}(T)$  for some binary tree  $T$ , let  $T(U)$  denote the minimal subtree of  $T$  connecting leaves from  $U$ , and let  $T|_U$  denote the binary tree obtained from  $T(U)$  by applying forced contractions to remove all vertices of degree 2.

The following results are well known.

**Lemma 1.1** 1. Any tree in  $UB(n)$  has  $2n - 2$  vertices and  $n - 3$  internal edges.

2.  $|UB(n)| = (2n - 5)!! := 1 \times 3 \times 5 \dots \times (2n - 5)$ .

## 2 Subtree Transfer Operations

### 2.1 Definitions

We now recall three commonly used subtree transfer operations, which form a nested sequence. We will describe them from the most restrictive to the most general.

#### 2.1.1 Nearest Neighbour Interchange

**Definition 2.1** Any internal edge of a unrooted binary tree has four subtrees attached to it. A *nearest neighbour interchange* (NNI) occurs when one subtree on one side of an internal edge is swapped with a subtree on the other side of the edge, as illustrated in Fig. 1.

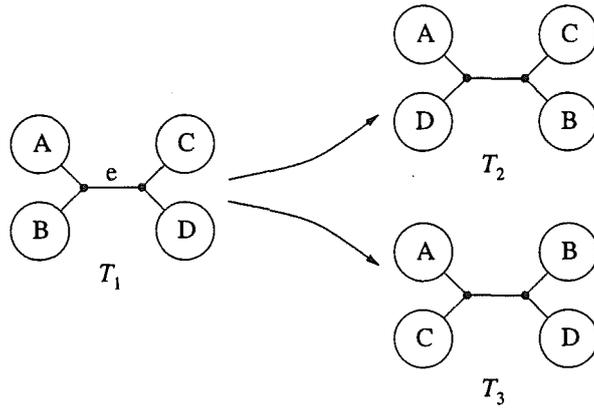


Figure 1: Trees  $T_2$  and  $T_3$  result from the two possible NNI's about edge  $e$  in  $T_1$

### 2.1.2 Subtree Prune and Regraft

The main focus for this paper is the *subtree prune and regraft* operation. This is the subtree transfer operation that is used to model the effect of a horizontal gene transfer or recombination in genomic data sets.

**Definition 2.2** A *subtree prune and regraft* (SPR) on a binary tree  $T$  is defined as cutting any edge and thereby pruning a subtree,  $t$ , and then regrafting the subtree by the same cut edge to a new vertex obtained by subdividing a pre-existing edge in  $T - t$ . We also apply a forced contraction to maintain the binary property of the resulting tree. See Fig. 2 for schematic representation of an SPR.

### 2.1.3 Tree Bisection and Reconnection

**Definition 2.3** A *tree bisection and reconnection* (TBR) on a binary tree  $T$  is defined as removing any edge, giving two new subtrees,  $t_1$  and  $t_2$ , which are then reconnected by creating a new edge between the midpoints of any edge in  $t_1$  and any edge in  $t_2$ . Again forced contractions are applied to ensure the resulting tree is binary. In the case that one of the subtrees is a single leaf, then the edge connecting  $t_1$  and  $t_2$  is incident to the leaf. See Fig. 2 for a schematic representation of a TBR.

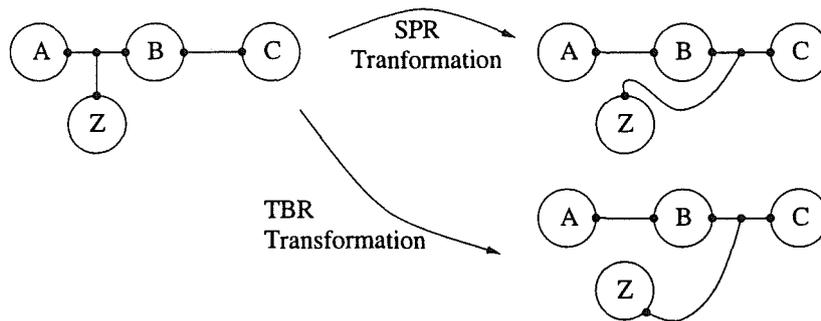


Figure 2: A schematic representation of the SPR and TBR operations. Note that the SPR operation can also be consider as a TBR, but not conversely.

## 2.2 Tree Neighbours

In this subsection we count the number of trees one subtree transfer operation from any given tree. Contrary to the findings of Page ([10]), the number of trees that are induced by one SPR from any given  $T \in UB(n)$  can be described by a simple formula involving just  $n$ .

**Definition 2.4** Two trees  $T, T'$  are said to be *neighbours* under a specific subtree transfer operation if  $T'$  can be obtained from  $T$  in one subtree transfer operation. The *neighbourhood* of a tree  $T$  is all trees that are neighbours with  $T$ .

**Theorem 2.1** *The size of the neighbourhood for  $T \in UB(n)$  is:*

1.  $2n - 6$  for the NNI operation,
2.  $2(n - 3)(2n - 7)$  for the SPR operation,
3. at most  $(2n - 3)(n - 3)^2$ , and dependent on the topology of  $T$  for the TBR operation.

**Proof** (1) was established by Robinson([11]). For (2), when a subtree is pruned and regrafted we cut an edge and then re-attach it to a different edge. The number of edges we can choose to cut is  $2n - 3$  and the number we can re-attach to is  $2n - 4$ . Hence the total number of possible subtree prune and regrafts is  $(2n - 3)(2n - 4)$ . However not all of these subtree prune and regrafts produce distinct trees, or even different trees to  $T$ . We can eliminate over-counts by separating SPR operations into three disjoint cases:

- (i) The edge to which the subtree will be regrafted is adjacent to the cut edge. This results in no change to the tree's topology.
- (ii) The edge to which the subtree will be regrafted is separated by exactly one edge from the edge to be cut. These are precisely the NNI transformations, and so from part (1), precisely  $2n - 6$  trees are generated.
- (iii) Lastly, consider the case where the edge to which the subtree is regrafted is separated by more than one edge away from the cut edge. It can be checked that any such prune and regraft will create a tree that can not be obtained by any other single SPR. Now we must count the number of such SPR operations. If we code an SPR operation by an ordered pair of edges (corresponding to the edge that is cut, and the edge that is attached to) then the the number of SPR operations in this last case is the number of ordered pairs of distinct edges (viz.  $(2n - 3)(2n - 4)$ ) minus the number of ordered pairs that correspond to SPR's covered by cases (i) and (ii). For case (i) the number of such pairs is  $6(n - 2)$  (since we have 6 such pairs associated with each of the  $n - 2$  internal vertices of  $T$ ), while for case (ii) the number of such pairs is  $8(n - 3)$  (since each of the  $n - 3$  internal edges of  $T$  gives rise to 8 such pairs). Thus the number of SPR operations corresponding to case (iii) is  $(2n - 3)(2n - 4) - 6(n - 2) - 8(n - 3) = 4(n - 3)(n - 4)$ , as required.

Combining cases (i), (ii) and (iii) the total number of trees at a distance of one SPR is  $0 + 2(n - 3) + 4(n - 3)(n - 4) = 2(n - 3)(2n - 7)$ .

Finally, for part (3), there is an injection from the set of TBR's on  $T$  to the set of ordered pairs  $(e, \{a, b\})$  where  $e$  is an edge of  $T$ , and where, if  $\{A, B\}$  is the bi-partition of  $\mathcal{L}(T)$  induced about  $e$ ,  $a$  is an edge from subtree  $T|_A$  (or  $a = T|_A$  if  $T|_A$  is a single vertex), and  $b$  is an edge from subtree  $T|_B$  (or  $b = T|_B$  if  $T|_B$  is a single vertex). Furthermore there are  $2n - 3$  choices for  $e$ ,  $|2|A| - 3|$  choices for  $a$  and  $|2|B| - 3|$  choices for  $b$ . Thus, there are at most  $(2n - 3)(|2|A| - 3)(|2|B| - 3)|$  trees, and furthermore  $|A| + |B| = n$ . For  $x + y = n$ ,  $(2x - 3)(2y - 3)$  attains its constrained maximum at  $x = y = n/2$ . Hence the number of trees one TBR from  $T$  is at most  $(2n - 3)(n - 3)^2$ .  $\square$

### 2.3 Tree Metrics

**Definition 2.5** Let the *distance* between two binary trees  $T_1$  and  $T_2$  with respect to a specific subtree transfer operation  $\Theta \in \{NNI, SPR, TBR\}$  be the minimum number of  $\Theta$  operations required to transform  $T_1$  to  $T_2$ . We write this as  $d_\Theta(T_1, T_2)$ .

**Lemma 2.1** 1.  $NNI \subseteq SPR \subseteq TBR$ .

2. For any  $T_1, T_2 \in UB(n)$ :

- (a)  $d_{TBR}(T_1, T_2) \leq d_{SPR}(T_1, T_2) \leq d_{NNI}(T_1, T_2)$ ; and,
- (b)  $d_{SPR}(T_1, T_2) \leq 2 \times d_{TBR}(T_1, T_2)$ .

**Proof** Part (1) was observed by Maddison [9]. Part (2a) follows immediately from Part (1). For Part (2b), consider the TBR of following general form in Fig 2. We can also obtain the same tree after two SPR's. We firstly prune the Z component subtree and regraft it to the correct edge, as in Figure 2. We then reconnect the Z component subtree so that it is joined at the correct vertex. This is achieved by treating the rest of the tree as a subtree to be pruned and regraft to the correct edge in the Z component. Thus we obtain exactly the same binary tree as that obtained from the TBR operation.  $\square$

**Definition 2.6** For  $\Theta \in \{NNI, SPR, TBR\}$ , the  $\Theta$ -adjacency graph  $G_\Theta(n) = (V, E)$  is the graph with  $V = UB(n)$  and  $\{t_u, t_v\} \in E \iff d_\Theta(t_u, t_v) = 1$ . The *diameter* of  $G_\Theta(n)$ , denoted  $\Delta(G_\Theta(n))$ , is the maximum value of  $d_\Theta(T, T')$  over all pairs  $T, T' \in UB(n)$ .

Robinson ([11]) showed that  $G_{NNI}(n)$  is connected — that is  $d_{NNI}(T, T')$  is defined for all  $T, T' \in UB(n)$ , and hence by Lemma 2.1 it follows that  $G_{SPR}(n)$  and  $G_{TBR}(n)$  are also connected. Thus  $\Delta(G_\Theta(n))$  is well defined. For the NNI operation, Li *et al.* [7] established the following nontrivial tight bound on the diameter of  $G_{NNI}(n)$ .

**Theorem 2.2** (from [7])

$$((n-2)/4) \log_2[2(n-2)\sqrt{2/3e}] \leq \Delta(G_{NNI}(n)) \leq n \log_2 n + \mathcal{O}(n).$$

We establish analogues for the SPR and TBR operations as follows.

**Theorem 2.3** For the SPR and TBR adjacency graphs:

- 1.  $n/2 - o(n) \leq \Delta(G_{SPR}(n)) \leq n - 3$ ; and,
- 2.  $n/4 - o(n) \leq \Delta(G_{TBR}(n)) \leq n - 3$ .

**Proof** For the lower bound of Part (1), recall from Theorem 2.1 that in  $UB(n)$  the number of trees one SPR from a given tree is  $2(n-3)(2n-7)$ , and that the number of unrooted binary trees is  $(2n-5)!!$ . Thus if  $d = \Delta(G_{SPR}(n))$ , then (since every vertex in the graph  $G_{SPR}(n)$  lies in a 3-cycle);

$$[2(n-3)(2n-7)]^d \geq (2n-5)!! = \frac{(2n-4)!}{2^{(n-2)}(n-2)!}. \quad (1)$$

By Equation (1) and Stirling's factorial approximation,

$$\begin{aligned} [2(n-3)(2n-7)]^d &\geq \frac{\sqrt{2\pi(2n-4)} (2n-4)^{(2n-4)} e^{-(2n-4)}}{2^{(n-2)} \sqrt{2\pi(n-2)} (n-2)^{(n-2)} e^{-(n-2)} e^{1/(12(n-2))}} \\ &= \sqrt{2} 2^{(n-2)} (n-2)^{(n-2)} e^{-(n-2)} e^{-1/(12(n-2))}. \end{aligned} \quad (2)$$

Taking natural logarithms of both sides of (2) gives:

$$d[\log(4) + \log(n-3)(n-7/2)] \geq (n-2)[\log 2 + \log(n-2) - 1] + \frac{1}{2} \log 2 - \frac{1}{12(n-2)}. \quad (3)$$

Now, for  $n \geq 4$ , we have  $\frac{1}{12(n-2)} \leq \frac{1}{2} \log 2$ , so

$$d[\log(4) + \log(n-3)(n-7/2)] \geq (n-2)[\log 2 + \log(n-2) - 1], \quad (4)$$

and if we let  $n \rightarrow \infty$  we get:

$$\frac{d}{n-2} = \frac{\log 2 + \log(n-2) - 1}{\log(4) + \log(n-3)(n-7/2)} \rightarrow \frac{1}{2}. \quad (5)$$

which establishes the lower bound for Part (1).

For the upper bound of Part (1), we use induction on the number of leaves. There are three binary trees on four leaves, all of which are at distance one SPR from each other. So the hypothesis holds for  $n = 4$ . Assume now that the hypothesis is true for any pair of trees in  $UB(k)$  and suppose  $T_1, T_2 \in UB(k+1)$ . Considering the cherries of  $T_1$  and  $T_2$ , there are two cases:

- (i) There is a cherry that occurs in both  $T_1$  and  $T_2$ . Replace this cherry in both trees by a single leaf to get  $T'_1$  and  $T'_2$ , both on  $k$  leaves. Hence  $T'_1$  can be transformed to  $T'_2$  in at most  $k-3$  operations and therefore, so too for  $T_1$  and  $T_2$ . Hence the hypothesis is valid for  $n = k+1$  in this case.
- (ii) If there is no cherry that occurs in both trees, then distinguish a cherry in  $T_2$ . Let  $T'_1$  be the tree obtained from  $T_1$  after one of the leaves of the distinguished cherry in  $T_2$  has been pruned from  $T_1$  and regrafted so that the distinguished cherry occurs in  $T'_1$  as well. Now apply case (i) to get that  $T'_1$  can be converted to  $T_2$  in at most  $k-3$  SPR. Hence  $T_1$  can be converted to  $T_2$  in at most  $k-2$  ways, hence the hypothesis is valid in this case for  $n = k+1$  as well.

Since cases (i) and (ii) cover all problem instances, the hypothesis is valid for all  $n$  by induction. It immediately follows that  $\Delta(G_{SPR}(n)) \leq n-3$ .

The lower bound in Part (2) follows from Part (1) and Lemma 2.1 (2b), and the upper bound of Part (2) follows from Part (1) and Lemma 2.1 (2a).  $\square$

## 2.4 Induced Subtree Distances

**Lemma 2.2** *Suppose we have  $T, T' \in UB(n)$ . Let  $U \subseteq \mathcal{L}(T)$ . Then  $d_\Theta(T|_U, T'|_U) \leq d_\Theta(T, T')$  for all  $\Theta \in \{NNI, SPR, TBR\}$ .*

**Proof** First note that a  $\Theta$ -operation on  $T$  induces a  $\Theta$ -operation on  $T|_U$  (provided we also allow the identity operation which leaves  $T$  unchanged to count as a  $\Theta$  operation).

Next we establish the result in the case  $d_\Theta(T, T') = 1$ . We will suppose that  $\Theta = SPR$ ; the NNI and TBR cases are similar. Represent the two trees that are one SPR apart as in Fig. 2, and consider the following cases.

- (i) If either  $S \cap \mathcal{L}(B) = \emptyset$  or  $S \cap \mathcal{L}(Z) = \emptyset$  then  $d_{SPR}(T|_U, T'|_U) = 0$ .
- (ii) If  $S \cap \mathcal{L}(C) = \emptyset$ , or  $S \cap \mathcal{L}(A) = \emptyset$  and  $B$  is a pendant subtree, then there is there is no change in the tree, since there is one central vertex from which  $Z$  is pruned and then reconnected to. Hence  $d_{SPR}(T|_U, T'|_U) = 0$ .

(iii) Lastly, if none of the above cases are true then there must be at least one internal vertex that distinguishes the placement of the  $Z_{|U}$  subtree. Hence  $d_{SPR}(T_{|U}, T'_{|U}) = 1$ , as  $Z_{|U}$  can be moved in one SPR.

Now, if  $d_{\Theta}(T, T') = k > 1$ , there are trees  $T^0, T^1, \dots, T^k$  such that  $T^0 = T, T^k = T'$  and  $d_{\Theta}(T^l, T^{l+1}) = 1$  for all  $l \in \{0, 1, \dots, k-1\}$ . Let  $t^l = T^l_{|U}$  for all  $l \in \{0, 1, \dots, k-1\}$ . Then from the particular case above,  $d_{\Theta}(t^l, t^{l+1}) \leq 1$  for all  $l \in \{0, 1, \dots, k-1\}$ . Thus, the trees  $t^1, \dots, t^k$  define a series of at most  $k$   $\Theta$ -operations that transform  $T_{|U}$  to  $T'_{|U}$ , as required.  $\square$

## 2.5 Maximum Agreement Forests

The concept of a (maximum) agreement forest for two binary trees was introduced by Hein et al [6]. As we will see it is particularly useful for analysing the TBR operation.

**Definition 2.7** Suppose we have two binary trees  $T_1$  and  $T_2$  with  $\mathcal{L}(T_1) = \mathcal{L}(T_2) = \mathcal{L}$ . Then (recalling definition 3 from Section 1),

- An *agreement forest* (AF) for  $T_1, T_2$  is a collection  $\mathcal{F} = \{t_1, \dots, t_k\}$  of binary trees such that, if we let  $\mathcal{L}_j := \mathcal{L}(t_j)$  for  $j \in \{1, \dots, k\}$ , then the following are satisfied:
  1.  $\mathcal{L}_1, \dots, \mathcal{L}_k$  partitions  $\mathcal{L}$
  2.  $t_j = T_1|_{\mathcal{L}_j} = T_2|_{\mathcal{L}_j}$  for all  $j \in \{1, \dots, k\}$ ; and
  3. for both  $i = 1$  and  $i = 2$  the trees  $\{T_i(\mathcal{L}_j) : j = 1, \dots, k\}$  are vertex-disjoint subtrees of  $T_i$ .
- A *maximum agreement forest* (MAF) for  $T_1, T_2$  is an agreement forest  $\mathcal{F}$  for  $T_1, T_2$  for which  $|\mathcal{F}|$  is minimal. Let  $m(T_1, T_2) := \min\{|\mathcal{F}| - 1 : \mathcal{F} \text{ is an AF for } T_1, T_2\}$ .

### Remarks:

1. Informally,  $m(T_1, T_2)$  is the smallest number of edges that need to be cut from each of  $T_1$  and  $T_2$  so that the resulting forests agree, once unlabelled vertices of degree less than three are removed (by deletion of unlabelled vertices of degree 1, and forced contractions).
2. For  $T_1, T_2 \in UB(n)$ , the same number of edges must be cut in both  $T_1$  and  $T_2$  to construct their MAF.
3. A MAF for  $T_1, T_2 \in UB(n)$  need not be unique. Suppose  $T_1, T_2$  are two different unrooted binary trees on four leaves. By removing the same leaf from both trees we obtain a MAF, however there are four possible leaves that we can remove, and so four possible MAFs.

## 2.6 MAF Size and SPR- and TBR-Distance

Lemma 7 of [6] states that the size of a MAF for any two given rooted binary trees  $T_1, T_2$  is one more than their SPR-distance. However this is not true for unrooted trees, and indeed, neither is it true for SPR transformations (suitably defined, see [1]) on rooted trees as the counterexamples in Figure 3 show.

Despite these counterexamples, Lemma 7 of [6] becomes true if we consider the TBR operation instead of the SPR operation, as the next theorem shows.

**Theorem 2.4** Suppose we have two binary trees  $T, T'$  with  $\mathcal{L}(T) = \mathcal{L}(T') = \mathcal{L}$ . Then,

$$d_{TBR}(T, T') = m(T, T').$$

In particular,  $m$  is a metric.

**Proof** We first show that  $m(T, T') \leq d_{TBR}(T, T')$  by using induction on  $k = d_{TBR}(T, T')$ . If  $k = 1$ , then only one edge needs to be cut in each of  $T$  and  $T'$  in order to construct a MAF, hence the hypothesis holds.

Now, suppose that the hypothesis holds for pairs of trees with a TBR-distance of  $k \geq 1$  and suppose  $d_{TBR}(T, T') = k + 1$ . Then there is a tree  $T''$  such that  $d_{TBR}(T, T'') = k$  and  $d_{TBR}(T'', T') = 1$ . Thus, by the inductive hypothesis, there exists a partition  $\pi = \{A_1, \dots, A_k\}$  of  $\mathcal{L}$  such that  $\{T''_{|A_i} : i = 1, \dots, k\}$  is a MAF for  $(T, T'')$ , and a bipartition  $\pi' = \{A, B\}$  of  $\mathcal{L}$  such that  $\{T''_{|A}, T''_{|B}\}$  is a MAF for  $(T'', T')$ . Now, by considering the subtrees  $\{T''(A_i) : i = 1, \dots, k\}$  of  $T''$ , we see that  $\pi'$  either splits no set in  $\pi$  (case (i)), or  $\pi'$  splits precisely one set in  $\pi$  - say  $A_j$  (case (ii)). Thus, if we set  $\pi''$  equal to  $\pi$  in case (i), or equal to  $\{\pi - \{A_j\}\} \cup \{A_j \cap A, A_j \cap B\}$  in case (ii), we have that  $\{T''_{|U} : U \in \pi''\}$  forms an agreement forest for  $(T, T'')$  and for  $(T'', T')$  and thereby for  $(T, T')$ . Thus,  $m(T, T') \leq k + 1$ , which completes the induction step.

To show that  $m(T, T') \geq d_{TBR}(T, T')$ , we again use induction, this time on  $m = m(T, T')$ . For  $m = 1$ , the MAF is obtained by deleting a single edge from each of  $T$  and  $T'$ , hence  $d_{TBR}(T, T') = 1$ . Now suppose the inductive hypothesis holds for  $m \leq k - 1$  and that  $m(T, T') = k$ . Let  $\{t_1, \dots, t_{k+1}\}$  be a MAF for  $T, T'$ . For at least one  $i \in \{1, \dots, k + 1\}$ , the subtree  $T(\mathcal{L}_i)$  of  $T$  can be pruned from the rest of  $T$  by deleting one edge only. In  $T'$  there exists at least one  $j \in \{1, \dots, k + 1\}$  such that  $T'(\mathcal{L}_i)$  is joined to  $T'(\mathcal{L}_j)$  by a path that does not include any vertices in  $\cup_{m \neq i, j} T'(\mathcal{L}_m)$ . Note that this last sentence could not also be true with  $T'$  replaced by  $T$ , else we could construct a smaller MAF for  $T, T'$  by amalgamating  $\mathcal{L}_i$  and  $\mathcal{L}_j$ . Now, we can cut the single edge of  $T$  incident with  $T(\mathcal{L}_i)$  and then re-attach  $T(\mathcal{L}_i)$  to  $T(\mathcal{L}_j)$  in such a way that  $T_{|\mathcal{L}_i \cup \mathcal{L}_j} = T'_{|\mathcal{L}_i \cup \mathcal{L}_j}$ . We call this new tree  $T''$  and note that it must differ from  $T$  by exactly one TBR.  $T''$  and  $T'$  now have an AF of size  $k$ , and so  $m(T'', T') \leq k - 1$ . Thus, by the inductive hypothesis,  $d_{TBR}(T'', T') \leq k - 1$ . Thus  $d_{TBR}(T, T') \leq d_{TBR}(T, T'') + d_{TBR}(T'', T') \leq k$  as required to establish the induction step.  $\square$

By Lemma 2.1 (2a) and the first inequality of Theorem 2.4 we have:

$$d_{SPR}(T_1, T_2) \geq m(T_1, T_2). \quad (6)$$

The counterexample at the start of this subsection show that the inequality can be strict.

### 3 Complexity of Computing Distances Between Evolutionary Trees

A fundamental problem is determining the distance between two given trees from  $UB(n)$  with respect to some tree metric. Seemingly the only paper to address the complexity of the computing the SPR-distance between two trees is [6]. However the authors base their treatment on Lemma 7 of [6], which, as pointed out in Section 2.6 is incorrect. Consequently, the complexity of the SPR distance problem remains unresolved. However, Theorem 8 of [6] should not be disregarded since its proof can be used to establish that the TBR-distance problem is *NP*-hard, by invoking Theorem 2.4

#### 3.1 Fixed Parameter Tractability for the $\Theta$ -Distance Problem

##### 3.1.1 Tree Reduction Rules

Despite the fact that the TBR-distance problem is *NP*-hard and the suspicion that so too is the SPR-distance problem, we show here that the Parameterized TBR-distance problem is *fixed parameter tractible* (FPT). That is, we show that the problem of determining the TBR distance between two trees, each with  $n$  leaves and whose TBR distance is at most  $k$  can be solved by an algorithm which runs in polynomial time (in  $n$ ) and for which the degree of this polynomial is independent of  $k$ .

The first step of a typical FPT problem is to *kernelize* the problem, that is, the size of the problem is reduced in such a way that the answer to the reduced problem is the same as the answer to the original problem and so that the size of the reduced problem is some function involving just the parameter  $k$ , i.e. it does not involve  $n$  (see [2]). In our case we wish to kernelize the problem by reducing the size of the two given trees, while still maintaining the SPR or TBR distance between them. We do this by repeatedly applying the following:

- **Rule 1** Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label.
- and
- **Rule 2** Replace any chain of pendant subtrees that occur identically in both trees by three new leaves with new labels correctly oriented to preserve the direction of the chain.

For both rules, the position of attachment of each pendant subtree must be the same in the two trees. Figures 4 and 5 illustrate Rule 1 and Rule 2 respectively.

The following Lemma is easily demonstrated. We will not attempt to do so here, nor quantify the time required. Useful further work might involve finding a fast implementation.

**Lemma 3.1** *For  $T_1, T_2 \in UB(n)$  Rule 1 and Rule 2 can be repeatedly applied to reduce  $T_1$  and  $T_2$ , until they can be reduced no further, in polynomial time in  $n$ .*

### 3.1.2 Preservation of $\Theta$ -Distance

**Definition 3.1** An *abc tree* is a binary tree  $T$  whose leaf set includes three leaves  $a, b, c$  with the following property; if  $v_a, v_b, v_c$  are the three vertices of  $T$  adjacent to  $a, b, c$  (resp.) then  $\{v_a, v_b\}$  and  $\{v_b, v_c\}$  are edges of  $T$ . Trees  $T'_1$  and  $T'_2$  in Figure 5 furnish two examples of *abc trees*.

**Lemma 3.2** *If  $T, T' \in UB(n)$  are two abc trees with  $\mathcal{L}(T) = \mathcal{L}(T')$ , then there exists a MAF  $\mathcal{F}$  for  $T, T'$  in which  $a, b, c$  are contained in the leaf set of one of the trees in  $\mathcal{F}$ .*

**Proof** Suppose  $\mathcal{F}$  is a MAF for  $T, T'$ . Let  $L_a$  (resp.  $L_c$ ) be the set of leaves connected to  $a$  (resp.  $c$ ) once edge  $\{v_a, v_b\}$  (resp.  $\{v_b, v_c\}$ ) is deleted from  $T$ . Let  $L'_a = L_a - \{a\}$ ;  $L'_c = L_c - \{c\}$ . We now distinguish two cases:

1. There exists a tree  $t \in \mathcal{F}$  with leaves from both  $L'_a$  and  $L'_c$ .
2. No tree in  $\mathcal{F}$  contains leaves from both  $L'_a$  and  $L'_c$ .

**Case (1)** Let  $t_a = t_{|L'_a}$  and  $t_c = t_{|L'_c}$ , and let  $I := |\mathcal{L}(t) \cap \{a, b, c\}|$ . If  $I = 0$  then each of  $a, b$  and  $c$  must be isolated point in  $\mathcal{F}$  (by property (3) in the definition of an AF). Let  $\mathcal{F}' := (\mathcal{F} - \{a, b, c, t\}) \cup \{t_a, t_c, t_{abc}\}$  (where  $t_{abc}$  is the tree with the three leaves  $a, b, c$ ). Then  $\mathcal{F}'$  is an agreement forest for  $T, T'$  with fewer trees than  $\mathcal{F}$ , contradicting the minimality of  $\mathcal{F}$  - thus this case does not arise.

If  $I = 1$ , let  $x$  denote the leaf in  $\mathcal{L}(t) \cap \{a, b, c\}$  and  $y, z$  denote the other two leaves. Then,  $y, z$  must be isolated vertices in  $\mathcal{F}$  and so  $\mathcal{F}' := (\mathcal{F} - \{y, z, t\}) \cup \{t_a, t_c, t_{abc}\}$  is also an AF for  $T, T'$  with the same number of trees as  $\mathcal{F}$ . Thus we can replace  $\mathcal{F}$  by  $\mathcal{F}'$  to obtain a MAF in which  $a, b, c$  occur in a single component.

If  $I = 2$ , then one of the leaves,  $x \in \{a, b, c\}$  is an isolated vertex in  $\mathcal{F}$ . Let  $t' := T_{|\mathcal{L}(t) \cup \{x\}}$ . Then  $\mathcal{F}' = (\mathcal{F} - \{x, t\}) \cup \{t'\}$  is also an AF forest for  $T, T'$ , but with fewer trees than  $\mathcal{F}$ , a contradiction, so this case does not arise.

If  $I = 3$ ,  $\mathcal{F}$  already satisfies the condition we want and we are done.

**Case (2)** If  $\mathcal{F}$  contains all three leaves  $a, b, c$  then we are done. Otherwise, we distinguish two subcases:

- (i) at least one leaf  $x \in \{a, b, c\}$  occurs as an isolated vertex in  $\mathcal{F}$ , or

- (ii) leaves  $a, b$  are in one component  $t_1 \in \mathcal{F}$  and leaf  $c$  is in another  $t_2 \in \mathcal{F}$  (or leaves  $b, c$  are in one component, and leaf  $a$  is in another).

In subcase (i), delete  $a, b, c$  from any trees in  $\mathcal{F}$  and replace isolated leaf  $x$  by the tree  $t_{abc}$  to obtain an AF for  $T, T'$  of the same size as  $\mathcal{F}$ . Since this contains  $a, b, c$  in one tree we are done.

In subcase (ii), let  $t := T|_{\mathcal{L}(t_1) \cup \mathcal{L}(t_2)}$ . Then  $\mathcal{F}' := (\mathcal{F} - \{t\}) \cup \{t'\}$  is an AF for  $T, T'$  yet smaller than  $\mathcal{F}$ ; a contradiction.  $\square$

**Theorem 3.1** *Let  $T_1, T_2 \in UB(n)$  and let  $T'_1$  and  $T'_2$  be obtained from  $T_1$  and  $T_2$  respectively by applying Rule 1 or Rule 2. Then  $d_{TBR}(T_1, T_2) = d_{TBR}(T'_1, T'_2)$ .*

**Proof** We establish the result for Rule 2; the corresponding result for Rule 1 is similar but simpler. Label the subtrees in the chain shared by  $T_1$  and  $T_2$  as  $t_1, \dots, t_r$  where  $r \geq 3$  (with this order). Suppose these are replaced by new leaves  $a, b, c$  under Rule 2. Thus  $T'_1$  and  $T'_2$  are both  $abc$  trees, and so there exists a MAF  $\mathcal{F}$  for  $T'_1, T'_2$  satisfying Lemma 3.2. Now, in these trees, let us re-insert the trees  $t_1, \dots, t_r$  in this order in each of  $T'_1, T'_2$  to new vertices that subdivide the edge  $\{v_a, v_b\}$  (where  $v_a, v_b$  are the vertices adjacent to  $a$  and  $b$ ). Call the resulting trees  $T''_1, T''_2$ . Now, any MAF for  $T''_1, T''_2$  which has leaves  $a, b, c$  in the same component  $t$  can be modified to produce an agreement forest for  $T''_1, T''_2$  of the same size, by simply attaching the trees  $t_1, \dots, t_r$  along the edge  $\{v_a, v_b\}$  of  $t$  (or, in case  $v_a = v_b$  in  $t$ , along the edge from  $a$  to  $v_a$ ). Thus, by Theorem 2.4,  $d_{TBR}(T''_1, T''_2) \leq d_{TBR}(T'_1, T'_2)$ . However, since  $T_1, T_2$  are both induced subtrees of  $T''_1, T''_2$ , Lemma 2.2 gives  $d_{TBR}(T_1, T_2) \leq d_{TBR}(T''_1, T''_2)$  and thus  $d_{TBR}(T_1, T_2) \leq d_{TBR}(T'_1, T'_2)$ .

For the inverse inequality, with  $t_1, \dots, t_r$  as before, suppose we select a leaf  $a \in \mathcal{L}(t_1), b \in \mathcal{L}(t_2), c \in \mathcal{L}(t_3)$  and replace the chain  $t_1, \dots, t_r$  in  $T_1, T_2$  by leaves  $a, b, c$  (correctly oriented) to obtain trees  $T'_1, T'_2$ . Let  $U$  denote the set of leaves of  $T_1$  that do not lie in the chain, together with  $a, b, c$ . Then, by Lemma 2.2,  $d_{TBR}(T_{1|U}, T_{2|U}) \leq d_{TBR}(T_1, T_2)$ , and since  $T_{i|U} = T'_i$  for  $i = 1, 2$  we obtain  $d_{TBR}(T'_1, T'_2) \leq d_{TBR}(T_1, T_2)$ , as required.

Combining both inequalities we get  $d_{TBR}(T'_1, T'_2) = d_{TBR}(T_1, T_2)$ , as required.  $\square$

Note that Theorem 3.1 applies only to the TBR operation. Rule 1 is also distance preserving for the SPR operation, however for Rule 2 we offer the following:

**Conjecture 3.2** *Let  $T_1, T_2 \in UB(n)$  and let  $T'_1$  and  $T'_2$  be obtained from  $T_1$  and  $T_2$  by applying Rule 2. Then  $d_{SPR}(T_1, T_2) = d_{SPR}(T'_1, T'_2)$ .*

For the NNI-distance problem, Rule 2 is not distance preserving (see [1]).

### 3.1.3 Maximally Reduced Trees have Bounded Size

Suppose that we are given  $T_1, T_2 \in UB(n)$  such that  $d_\Theta(T_1, T_2) = k$  for  $\Theta \in \{SPR, TBR\}$ , and that  $T_1$  and  $T_2$  can be reduced no further by Rule 1 or Rule 2. In this section, we show that the size of the leaf set of the two trees is bounded by some function  $f$  which depends (linearly!) only on  $k$ , ie  $|\mathcal{L}(T_i)| \leq f(k)$ , where  $i \in \{1, 2\}$ . Given  $T_1, T_2 \in UB(n)$ , and a MAF,  $t_1, \dots, t_k$ , let  $deg^i(t_j)$  for  $i = 1, 2$  denote the number of edges of  $T_i$  that are incident with the subtree  $t_j$ . Now, for both  $i = 1$  or  $i = 2$  we have:

**Lemma 3.3**

$$\sum_j deg^i(t_j) \leq 2k - 2$$

**Proof** In  $T_i$ , collapse each of  $t_j, j = 1, \dots, k$  to a single vertex of degree  $deg^i(t_j)$  thereby obtaining a tree,  $(V, E)$  consisting of these new vertices, and  $n_3 \geq 0$  vertices of degree 3. Thus,  $|V| = n_3 + k$ , and

$\sum_j \deg^i(t_j) + 3n_3 = 2|E|$ . Now, since  $(V, E)$  is a tree,  $|V| = |E| + 1$ , and so  $\sum_j \deg^i(t_j) = 2k - 2 - n_3 \leq 2k - 2$ , as required.  $\square$

**Lemma 3.4** *Let  $t_1, \dots, t_k$  be a MAF for  $T_1, T_2 \in UB(n)$ . Then, by repeatedly applying Rules 1 and 2, the number of leaves in  $t_j$  (for  $j = 1, \dots, k$ ) can be reduced to  $c(\deg^1(t_j) + \deg^2(t_j))$  for a fixed constant  $c \leq 7$ .*

**Proof** Let  $I_j$  be the set of edges of  $t_j$  that are incident with edges of either  $T_1$  or  $T_2$ . Let  $t'_j$  denote the minimal subtree of  $t_j$  containing among its edges the set  $I_j$ . Let  $t''_j$  be the tree obtained from  $t'_j$  by replacing each maximal path that contains no edge from  $I_j$  by a single edge - let  $F_j$  denote this set of new edges. Let  $P_j$  denote the set of pendant edges of  $t''_j$ . Let  $i_j := |I_j|$ ;  $f_j := |F_j|$ ;  $p_j := |P_j|$ . We pause to make several observations.

1.  $P_j \subseteq I_j$ ,
2.  $I_j \cup F_j$  forms a disjoint partition of the edges of  $t''_j$ ,
3. Any vertex of  $t''_j$  of degree 2 is incident with at least one edge from  $I_j$ ,
4. By applying rules 1 and 2 to  $T_1$  and  $T_2$  the subtree  $t_j$  can be reduced to a subtree of size at most  $s$ , where

$$s := p_j + 3f_j. \quad (7)$$

Now we claim that:

$$s \leq 7i_j - 9. \quad (8)$$

To establish this inequality, let  $v_j^{(k)}$  denote the number of vertices of  $t''_j$  of degree  $k$ . Then,  $v_j^{(1)} = p_j$ , and  $v_j^{(k)} = 0$ , for  $k > 3$ . Counting the edges of  $t''_j$  twice by summing degrees we have:

$$v_j^{(1)} + v_j^{(2)} + v_j^{(3)} = 2(i_j + f_j) = 2(v_j^{(1)} + v_j^{(2)} + v_j^{(3)} - 1) \quad (9)$$

where the second equality is because  $t''_j$  is a tree and so has one less edge than its number of vertices. Rearranging Equation (9), and noting that  $v_j^{(1)} = p_j$ , we have:

$$p_j - v_j^{(3)} = 2.$$

Now,  $f_j$  is the total number of edges of  $t''_j$  minus  $i_j$ , and so

$$f_j = (v_j^{(1)} + v_j^{(2)} + v_j^{(3)} - 1) - i_j.$$

Substituting these last two equations into Equation (7) gives,

$$s = p_j + 3(2p_j + v_j^{(2)} - 3 - i_j) = 7p_j + 3(v_j^{(2)} - i_j) - 9.$$

Now, since each edge in  $P_j$  gives rise to at most one vertex of degree 2 and each edge in  $I_j - P_j$  gives rise to at most two vertices of degree 2, and (by observations (1) and (3) above) all vertices of degree 2 are covered in this way we obtain:  $v_j^{(2)} \leq p_j + 2(i_j - p_j)$ . Substituting this inequality into the previous equality gives:

$$s \leq 4p_j + 3i_j - 9 \leq 7i_j - 9,$$

as claimed. This establishes inequality (8). Finally, we have,

$$i_j \leq \deg^1(t_j) + \deg^2(t_j)$$

which combined with inequality (8) completes the proof of the Lemma.  $\square$

**Theorem 3.3** *The Parameterized TBR-Distance Problem is fixed-parameter tractable.*

**Proof** By Lemma 3.1, Rule 1 and Rule 2 can be repeatedly applied to reduce any two trees from  $UB(n)$  in polynomial time. By Lemmas 3.3 and 3.4 the number of leaves in each reduced tree can be bounded above by:

$$\sum_j c(\deg^1(t_j) + \deg^2(t_j)) \leq 4c(k-1)$$

which is independent of  $n$ . Computing the TBR-distance between these reduced trees depends only on their size, and not the size of the leaf set of the original two trees, and this distance is the TBR distance between the original two trees (by Theorem 3.1), completing the proof.  $\square$

Theorem 3.3 shows that, provided the TBR-distance between two trees is sufficiently small we will be able to determine the exact distance in realistic time. Note that if Conjecture 3.2 is true, then the argument in Theorem 3.3 would also establish that the Parameterized SPR-Distance Problem is *FPT*.

## 4 Acknowledgements

The authors wish to thank the generous support of the New Zealand Marsden Fund for this report (UOC516). Thanks also to Mike Fellows for helpful comments concerning *FPT* analysis, and for Jotun Hein and Charles Semple for some comments on an earlier version of this manuscript.

## References

- [1] B. Allen, Subtree transfer operations and their induced metrics on evolutionary trees. MSc thesis, University of Canterbury, Christchurch, New Zealand, 1998.
- [2] R. Downey and M.R. Fellows, Parameterized complexity, Springer-Verlag, New York, 1999.
- [3] P. Gilson and G. McFadden, Something borrowed, something green: lateral transfer of chloroplasts by secondary endosymbiosis, *Trends Ecol. Evol.*, 10, 12–17 (1995).
- [4] J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, *Math. Biosci.*, 98, 185–200 (1990).
- [5] J. Hein, A heuristic method to reconstruct the history of sequences subject to recombination, *J. Mol. Evol.*, 36, 369–405 (1993).
- [6] J. Hein, T. Jiang, L. Wang, and K. Zhang, On the complexity of comparing evolutionary trees. *Discr. Appl. Math.*, 71, 153–169 (1996).
- [7] M. Li, J. Tromp, and L. Zhang, On the nearest neighbour interchange distance between evolutionary trees, *J. Theor. Biol.*, 182, 463–467 (1996).
- [8] W. H. Li and D. Graur, *Fundamentals of Molecular Evolution*, Sinauer Associates, Inc., 1991.
- [9] D. R. Maddison, The discovery and importance of multiple islands of most-parsimonious trees, *Syst. Zool.*, 43(3), 315–328 (1991).
- [10] R. D. M. Page, On islands of trees and the efficacy of different methods of branch swapping in finding most-parsimonious trees, *Syst. Biol.*, 42(2), 200–210 (1993).
- [11] D. F. Robinson, Comparison of labeled trees with valency three, *J. Combin. Theory*, 11, 105–119 (1971).
- [12] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis, *Phylogenetic Inference in Molecular Systematics*, Sinauer Associates, second edition, 1996.

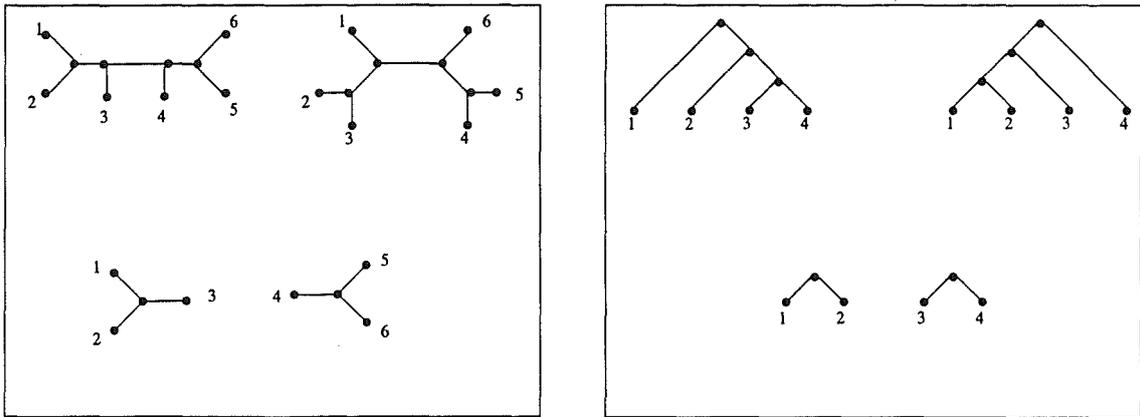


Figure 3: Counterexamples to Lemma 7 of [6]. In the first (resp. second) box there are two unrooted (resp. rooted) binary trees that are more than one SPR apart, yet their MAF requires just one edge deletion.

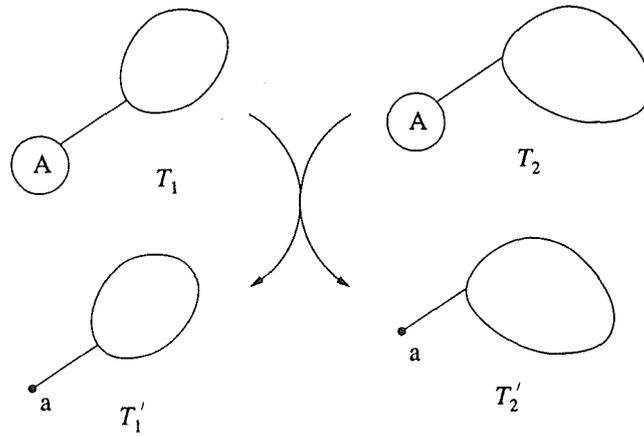


Figure 4: Reduction of two trees using Rule 1.

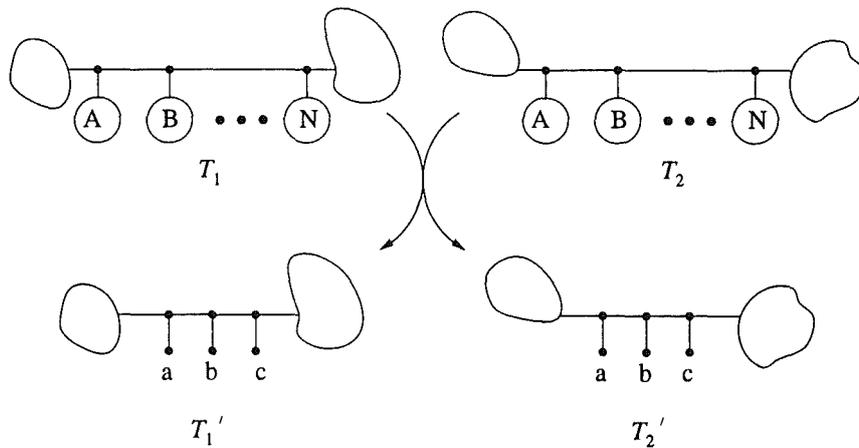


Figure 5: Reduction of two trees using Rule 2.