On the Complexity of uSPR Distance

Maria Luisa Bonet* Katherine St. John†‡

January 11, 2009

Abstract

We show that subtree prune and regraft (uSPR) distance on unrooted trees is fixed parameter tractable with respect to the distance. We also make progress on a conjecture of Steel [9] on the preservation of uSPR distance under chain reduction, improving on lower bounds of Hickey *et al.* [7].

Keywords: phylogeny, analysis of algorithms, fixed parameter tractability

1 Introduction

Phylogenies, or evolutionary histories, are a crucial tool in biology. Often thousands of phylogenies are generated for a set of taxa [10]; comparing these phylogenetic trees is an essential step in determining the topology of the optimal tree. This paper focuses on the computational complexity of the popular phylogenetic distance based on the subtree prune and regraft (SPR) tree operation (defined in Section 2). Roughly, an SPR move between two trees breaks a subtree from the first tree and regrafts it to an edge of the second tree (see Figure 1), and the SPR distance between two trees is the minimal number of SPR moves that transforms one tree to the other. Another popular distance is based on the tree bisection and reconnection (TBR) tree operation (defined formally in Section 2). Informally, a TBR move removes an edge in a tree creating two trees, and then reconnects the two trees using an edge from each (see Figure 1 for an example and comparison between the two measures). Calculating the TBR distance is NP-hard [6]. Allen and Steel [1] showed that the distance is also fixed parameter tractable (FPT) (Hallett and McCartin [5] improved the FPT algorithm for TBR, which also gives a 4-approximation for calculating this distance). Allen and Steel used a correspondence between the TBR distance and the size of the maximum agreement forest (roughly, disjoint subtrees that can be arranged to form both of the trees) [6]. Then, they showed that natural rules for reducing trees preserve the distance (also defined in Section 2). SPR distance is preserved under one of these rules; Steel [9] conjectures that the second rule is also preserved.

SPR distance differs between rooted and unrooted trees [4]. Bordewich and Semple [4] examined these issues for SPR distance on rooted trees (rSPR). Namely, they showed that calculating the

*Lenguajes y Sistemas Informaticos (LSI), Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain, bonet@lsi.upc.edu.

†Department of Math & Computer Science, Lehman College– City University of New York (CUNY), Bronx, NY 12581, United States, {stjohn}@lehman.cuny.edu.

‡Department of Computer Science, CUNY Graduate Center, New York, NY 10016.

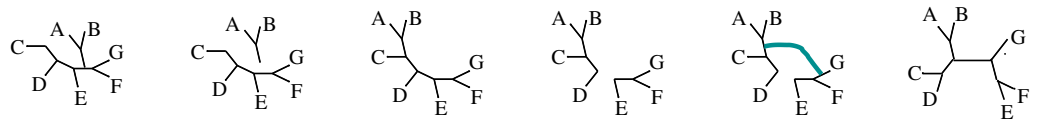


Figure 1: Left: A SPR move: a subtree is "pruned" and regrafted. Right: a TBR move: an edge is removed, creating two subtrees, and a new edge is added "reconnecting" the trees. All resulting vertices of degree 2 are contracted, so, the end result of both moves is a phylogenetic tree.

rSPR distance is NP-hard and FPT. These results rely on a correspondence between the rSPR distance between two trees and an agreement forest of those two trees. Using the agreement forests, they showed that natural reduction rules of Allen and Steel [1] preserve rSPR distance, yielding both the FPT result and solving Steel's conjecture in the affirmative for rooted trees. Approximation algorithms, with provable bounds, for rSPR have been developed by Bonet *et al.* [2] and Bordewich *et al.* [3].

While the first reduction rule (subtree reduction) preserves most distances between trees, it is unknown whether the second reduction rule (chain reduction) does for uSPR distance. The latter is Steels conjecture for SPR distance [9]. The work of Bordewich and Semple [4] proves the conjecture true for rSPR distance. For uSPR distance, Hickey *et al.* [7] proved that the uSPR distance is NP-hard and gave insight into Steels conjecture. They show that applications of the chain reduction rule can reduce the uSPR distance by at most two. The latter follows by an elegant argument that reduces applications of the chain reduction rule to that of a known distance-preserving reduction: subtree reduction rule (see Figure 2). Previously, the FPT of uSPR was not known, though the work of Hallett and McCartin [[5] provides an 8-approximation for calculating the uSPR distance [8].

In the present paper, we prove two new results about the uSPR distance: uSPR distance is fixed parameter tractable (with parameter, the distance between the trees), and applications of the chain reduction rule can reduce the uSPR distance by at most one, making progress towards solving the conjecture of Steel [9]. Unlike previous proofs, our FPT result does not rely on a correspondence between agreement forests and distances and the preservation of distances under the chain reduction rule. So, while we give a proof for FPT for uSPR (that also gives alternative proofs for TBR and rSPR), the conjecture itself remains an open problem.

2 Background and Definitions

The following definitions follow those of Allen and Steel [1] and Bordewich and Semple [4].

Definition 1 *An* **phylogenetic tree**, T, *is a binary tree whose leaves or degree one vertices are labeled by a set of species, and the non-leaf vertices are unlabelled and have degree three. For rooted trees, we have that the root vertex has degree two. An edge incident to a leaf is called a* **pendant edge**, *and otherwise it is an* **internal edge**. *We define the* **size of** T**,** $|T|$**,** *to be the number of leaves of* T.

Definition 2 *Given two phylogenetic trees T_1 and T_2 on the same leaf set, a* **(common) chain** *is three or more adjacent subtrees that occur identically in both trees (see T_1 and T_2 in Figure 3).*

Definition 3 *A* **subtree prune and regraft (SPR)** *operation on a phylogenetic tree is defined as cutting any edge and pruning a subtree t, and then regrafting the subtree by the same cut edge to a new vertex obtained by subdividing an edge in $T - t$. Any resulting vertex of degree two is contracted, so that the result is a phylogenetic tree (see Figure 1).*

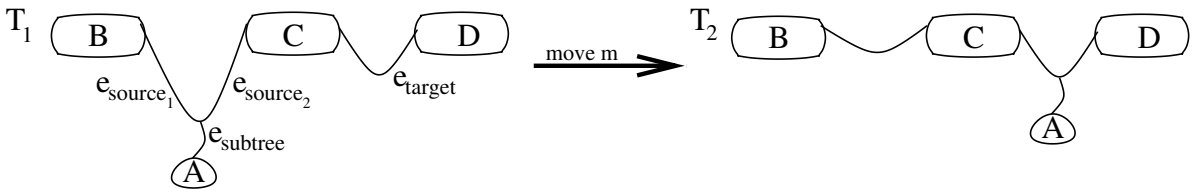
The **SPR distance** *between two phylogenetic trees T_1 and T_2 on the same leaf set, denoted by $d_{SPR}(T_1, T_2)$, is the minimum number of SPR operations required to convert T_1 into T_2. To distinguish between the distances on rooted and unrooted trees, we will refer to this measure as* **rSPR** *and* **uSPR***, respectively.*

Definition 4 *A* **tree bisection and reconnection (TBR)** *operation on an unrooted phylogenetic 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 betweeen the midpoints of any edge in t_1 and any edge in t_2. Any resulting vertex of degree two is contracted, so that the result is a phylogenetic tree (see Figure 1).*

The **TBR distance** *between two trees T_1 and T_2, $d_{TBR}(T_1, T_2)$, is the minimal number of TBR moves needed to transform T_1 into T_2.*

In our proofs, we will distinguish edges that are changed or altered by a sequence of moves. More formally,

Definition 5 *The T_1 be phylogenetic trees, and m_1 an uSPR move, that when applied to T_1 yields the tree T_2. Label the edges involved in the uSPR tree as follows: $e_{subtree}$ is the edge cut by move m and regrafted to the edge, e_{target}, and the initial neighboring edges in T_1 to $e_{subtree}$ are e_{source_1} and e_{source_2}. We say that the edge $e_{subtree}$ is* **broken** *by the move m_1, and there was an* **insertion** *on e_{target}. We will call all four edges, $e_{subtree}$, e_{target}, e_{source_1}, e_{source_2},* **altered** *by the move m.*



We extend the definition of altered edges (and similarly broken edges and insertions) to a sequence of moves, inductively from above. For a sequence of moves, $m_1, m_2, \ldots, m_{k+1}$ applied to T_1, let T_k be the result of applying $m_1, m_2, \ldots, m_k$ to T_1. Define the **altered** *edges of T_1, under moves $m_1, m_2, \ldots, m_{k+1}$ to be the union of the altered edges of T_1 under moves $m_1, m_2, \ldots, m_k$ with the altered edges of T_k under move m_{k+1}.*

Originally linked to tree measures by Hein *et al.* [6], agreement forests are an useful tool for calculating and showing hardness for tree measures.

Definition 6 *Let T_1 and T_2 be two phylogenetic trees on the same leaf set, L. An* **agreement forest (AF)** *for T_1, T_2 is a collection $\mathcal{F} = \{t_1, \ldots, t_k\}$ of phylogenetic trees such that if we let L_j be the leaves of tree t_j for $j \in \{1, \ldots, k\}$, then the following are satisfied:*

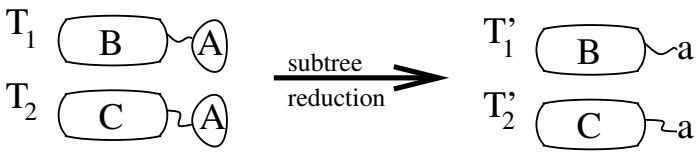


Figure 2: Subtree Reduction Rule: Common subtrees in the phylogenetic trees T_1 and T_2 are replaced by a single leaf, to yield new trees T_1' and T_2'.

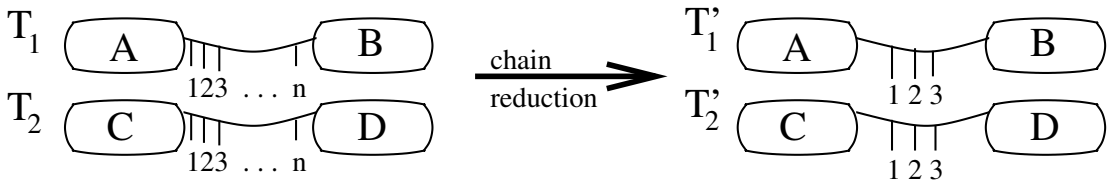


Figure 3: Chain Reduction Rule: Common chains in trees T_1 and T_2 are replaced by chains of length 3, yielding the new trees T_1' and T_2'.

1. *$L_1, \ldots, L_k$ partitions L (that is, the subtrees partition the leaf set),*
2. *$t_j = T_1|_{L_j} = T_2|_{L_j}$ for all $j \in \{1, \ldots, k\}$ (that is, the subtrees occur as induced subtrees of T_1 and T_2), and*
3. *for both $i = 1$ and $i = 2$, the trees $\{T_i|_{L_j} \mid j = 1, \ldots, k\}$ are vertex-disjoint subtrees of T_i (that is, the subtrees are vertex disjoint in both T_1 and T_2).*

A **maximum agreement forest (MAF)** *for T_1 and T_2 is an agreement forest for T_1 and T_2 with minimal number of subtrees in the forest over all possible agreement forests for T_1 and T_2.*

Allen and Steel [1] showed that the TBR distance between two trees is one less than the size of the TBR maximum agreement forest of the two trees.

We will now define the reduction rules used on pairs of trees:

Definition 7 *We consider three reduction rules for pairs of trees T_1 and T_2:*

- **Subtree Reduction Rule***: Replace a subtree that occurs identically in both trees by a single leaf with a new label (see Figure 2).*
- **Chain Reduction Rule***: Replace a chain of subtrees that occur identically in both trees by three new leaves with new labels correctly oriented to preserve the direction of the chain. (see Figure 3).*
- ***c*-Chain Reduction Rule:** *Replace a chain of pendant leaves that occur identically in both trees by c new leaves with new labels correctly oriented to preserve the direction of the chain.*

The first two reduction rules have been important tools in showing fixed parameter tractability results for tree distances. The third rule, a variant on the second, is introduced here to show the FPT result for calculating the uSPR distance. The subtree reduction rule is distance preserving for

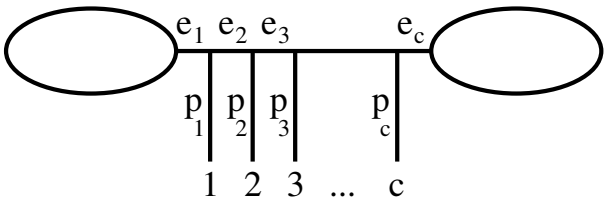


Figure 4: Notation for common chains: We call the pendant edges of the chain: $p_1, \ldots, p_c$, the internal edges of the chain: $e_1, \ldots, e_{c-1}$. e_0 and e_c refer to the edges incident, but not part of, the common chain.

the uSPR operation, as well as TBR and rSPR operations [1]. The chain rule is distance preserving for the TBR and rSPR operations [1, 4], and it is unknown whether it is also distance preserving for the uSPR operation (this is Steel's conjecture [9] on which this paper presents some progress). These reductions are essential to proofs of fixed parameter tractability for TBR [1] and rSPR [4], providing a way to reduce the initial trees to smaller trees (with equivalent distance) whose size is bounded by the parameter, the distance between the trees. The usefulness of this technique and similarity of the measures suggests Steel's conjecture [9] that the chain rule preserves SPR distance is true for unrooted trees. Hickey *et al.* [7] show that the applications of the chain rule reduce the distances by at most two. We improve that bound by one, making progress on the conjecture.

The following additional notation will be used in the paper: let T_1, T_2 be unrooted phylogenetic trees, labeled by the same leaf set, with a common chain of elements, $1, 2, \ldots, c$. We will refer to the pendant edges of the chain as $p_1, \ldots, p_c$, and the internal edges of the chain as $e_1, \ldots, e_{c-1}$. e_0 and e_c refer to the edges incident, but not part of, the common chain (see Figure 4). Finally, we will use the notation T_1^j and T_2^j to make explicit the fact that the trees have a common chain of length j: $1, 2, \ldots, j$.

3 Fixed Parameter Tractability of uSPR

We show the fixed parameter tractability of uSPR distance, with parameter, the distance, k. We note that this argument also applies to TBR and rSPR distances, giving an alternative proof of these results.

Theorem 1 *Let T_1 and T_2 be two unrooted phylogenetic trees on the same taxa set. Let n be the number of taxa in the trees, and let k be the uSPR distance between T_1 and T_2. The uSPR distance between T_1 and T_2 can be calculated in $O(f(k)n)$ time, where $f(k)$ is a function that does not depend on n.*

The proof of the theorem relies on two straightforward lemmas, that also generalize to other distances, such as rSPR and TBR distance. The first lemma states that if the common chain between two trees is sufficiently larger than the minimal number of moves to transform one tree to the other, then all trees with larger chains have the same distance.

Lemma 1 *Let T_1^j and T_2^j be two trees such that $d_{uSPR}(T_1^j, T_2^j) = k$. Suppose also $j > 9k$. Then for all $m \geq j$, $d_{uSPR}(T_1^m, T_2^m) = k$.*

Proof: Let $m_1, \ldots, m_k : T_1^j \to T_2^j$ be a minimal sequence of moves. We first show that there exists $i < j$ such that the subsequence, p_i, e_i, and p_{i+1}, are not altered by any of the moves $m_1, \ldots, m_k$. As noted in the definition of altered edges, at most $4k$ edges of T_1 are altered by a sequence of k moves. By the Pigeonhole Principle, if there are more than two times the number of chain elements as there are altered edges, the desired subsequence of unaltered edges existed. As noted in the definition of altered edges, at most $4k$ edges of T_1 are altered by a sequence of k moves. So, the length of the common chain needs to be greater than $2(4k+1)$. By hypothesis, $j > 9k > 2(4k+1)$, and thus, there exists an i such that p_i, e_i, and p_{i+1} are not altered by any of the moves $m_1, \ldots, m_k$.

At this point, we can insert $m - j$ new elements onto the edge e_i to create trees T_1' and T_2'. By construction, T_1' and T_2' have a common chain of length m and are isomorphic to T_1^m and T_2^m. Further, $m_1, \ldots, m_k : T_1' \to T_2'$. So, by a simple relabeling, there exists $m_1', \ldots, m_k' : T_1^m \to T_2^m$. This gives $d_{uSPR}(T_1^m, T_2^m) \leq k$. Since $d_{uSPR}(T_1^m, T_2^m) \geq d_{uSPR}(T_1^j, T_2^j)$ for all $m > j$, we have $d_{uSPR}(T_1^m, T_2^m) = k$. □

To show the theorem, we prove a second lemma, similar to Lemma 3.4 of [1] that shows that completely reduced pairs of trees are bounded by a function depending only on k, the uSPR distance (and not on n, the number of taxa in the trees). The original lemma of [1] bounds the size of the reduced trees under the subtree and chain rules. We extend the proof to give bounds on the size of the reduced trees when using the c-chain rule.

Lemma 2 *Let T_1 and T_2 be phylogenetic trees on the same leaf set. Let T_1' and T_2' be the reduction of T_1 and T_2 under the subtree and $9k$-chain reduction rules. Then if $d_{uSPR}(T_1, T_2) \leq k$, then $|T_1'| \leq 76k^2$.*

Proof: First, we will show that if $d_{TBR}(T_1, T_2) \leq k$, then $|T_1'| \leq 76k^2$. We note that since $d_{TBR}(T_1, T_2) \leq 2d_{uSPR}(T_1, T_2)$ [1] that this will give the desired result. We will do this by modifying the proof of Lemma 3.4 of [1] to hold for trees with chains of length at most $9k$.

Suppose that $d_{TBR}(T_1, T_2) \leq k$. Then $d_{TBR}(T_1', T_2') = k' \leq k$ and there exists a MAF for T_1' and T_2' such that $\mathcal{F} = \{t_0, \ldots, t_{k'}\}$ of size $k'+1$. By analyzing the maximum number of possible edges and leaves in $\mathcal{F}$ (that is, modulo the reduction rules), we get the desired bounds. We first count edges related to each $t_j \in \mathcal{F}$.

Let t_j be a tree of the MAF $\mathcal{F}$ for T_1' and T_2'. Let I_j be the set of edges of t_j that are incident with edges of either T_1 or T_2 that are not in t_j. Let t_j' be the minimal subtree of t_j containing among its edges the set I_j. Note that t_j' could equal t_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 be the set of these new edges, and P_j the set of pendant edges of t_j''. At this point, it is clear that:

Fact 1: $P_j \subseteq I_j$,

Fact 2: $I_j \cup F_j$ is a disjoint partition of the edges of t_j'', and

Fact 3: Any vertex of t_j'' of degree two is incident with at least one edge from I_j.

Also let $i_j = |I_j|$, $f_j = |F_j|$ and $p_j = |P_j|$. Then since t_j has chains of length at most $9k$ and the subtree reduction rule has been applied, the size of t_j is at most

$$|t_j| \leq p_j + 9k \cdot f_j \tag{1}$$

This follows from 1) the subtrees of t_j corresponding to an edge P_j can be replaced by a single leaf by the subtree reduction rule; and 2) the collection of subtrees corresponding to an edge in F_j can be replaced by at most $9k$ leaves by the $9k$-chain rule.

Now, let v^d denote the number of vertices of t''_j of degree d. Then,

$$v^1 + 2v^2 + 3v^3 = 2(i_j + f_j) = 2(v^1 + v^2 + v^3 - 1).$$

The first equality is by counting the edges of t''_j twice and summing the degrees of the nodes. The second is because in a tree the number of edges is one less the number of vertices.

By rearranging the previous equation and noting that $v^1 = p_j$, we have: $v_3 = p_j - 2$. We also have that f_j is the total number of edges of t''_j minus i_j (by Fact 2), so,

$$f_j = (v^1 + v^2 + v^3 - 1) - i_j.$$

By Fact 1 and Fact 3 and the fact that each edge in P_j gives rise to at most one vertex of degree 2, we have that

$$v^2 \leq p_j + 2(i_j - p_j).$$

Substituting the last three equations into Equation 1 (and noting that $v^1 = p_j \leq i_j$),we obtain that the size of t_j is at most:

$$\begin{aligned} |t_j| &\leq p_j + 9k \cdot [(v^1 + v^2 + v^3 - 1) - i_j] \\ &\leq p_j + 9k \cdot [(p_j + (p_j + 2(i_j - p_j) + (p_j - 2) - 1) - i_j] \\ &= p_j + 9k \cdot [i_j + p_j - 3] \\ &= (9k + 1)p_j + 9ki_j - 9k \cdot 3 \\ &\leq 19ki_j \end{aligned}$$

Summing over the $k + 1$ trees of the MAF of T_1 and T_2, we have:

$$|T'_1| = \sum_{j=0}^{k} 19ki_j = 19k \sum_{j=0}^{k} i_j \leq 19k \cdot 4k = 76k^2.$$

The inequality follows from Lemma 3.3 of [1] which states that the possible incident edges to the components of the MAF in T_1 (and similarly for T_2) are bounded by $2k-2$, and thus $\sum_{j=0}^{k} i_j < 4k$. □

The fixed parameter tractability follows directly from the previous two lemmas:

Proof of Theorem: Let T_1 and T_2 be phylogenetic trees on the same leaf set, and let k be an integer. The algorithm to decide if $d_{uSPR}(T_1, T_2) \leq k$ is:

1. Reduce T_1 and T_2 using the subtree reduction rule and the $9k$-chain reduction rule. Let T'_1 and T'_2 be the result of repeatedly applying the rules to T_1 and T_2 until no further reduction is possible. This part of the algorithm is linear in the size of T_1 and T_2 [2]. By Lemma 1, if $d_{uSPR}(T'_1, T'_2) = k$, then $d_{uSPR}(T_1, T_2) = k$.
2. if $|T'_1| > 76k^2$, then the distance of the original trees is greater than k (by Lemma 2), and return the answer no.
3. Otherwise, $|T'_1| \leq 76k^2$. Look at all sequences of k moves that start with T'_1. Since $|T'_1|$ is bounded by $76k^2$, the number of sequences is bounded (exponentially) by k, and does not depend on n, the number of leaves. If some sequence of moves transforms T'_1 into T'_2, return the answer yes, else return no.

□

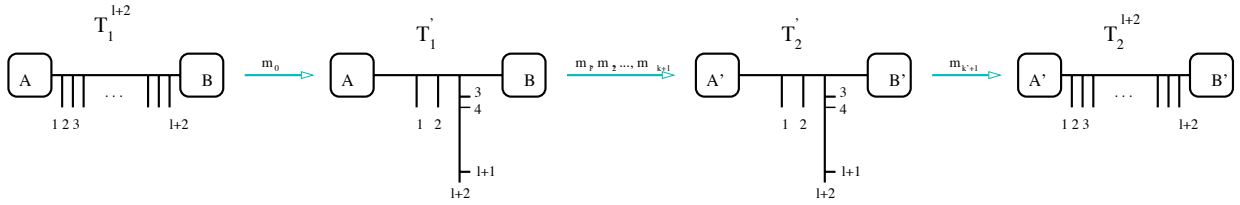


Figure 5: The intermediate trees in Case 2 of Theorem 2.

4 Bounding uSPR

The proof of the lower bound on uSPR distance is inspired by the proof of Hickey *et al.* [7]. They alter the initial trees to obtain trees, distance one from the originals, with the common chain as subtrees (similar to T_1' in Figure 5) and apply the subtree reduction rule. This clever transformation gives a lower bound of two less than the distance. By careful analysis of the sequence of moves that transforms the first tree into the second, we give a sharper lower bound for uSPR distance under the chain reduction rule:

Theorem 2 *An application of the chain reduction rule to common chains cannot reduce the uSPR-distance of two phylogenetic trees by more than 1.*

Proof: Let T_1 and T_2 be the given trees with a common chain of $1, \ldots, l$, and let T_1^3 and T_2^3 be the result of applying the chain rule. Let $d_{uSPR}(T_1, T_2) = k$.

To show that $d_{uSPR}(T_1^3, T_2^3) \geq k - 1$, we go by cases on the minimal sequences of moves, $m_1, \ldots, m_{k'}$, that transform T_1^3 into T_2^3:

Case 1: There exists a minimal sequence of moves that transforms T_1^3 into T_2^3 that either does not break the chain edges, p_1, e_1, p_2 or does not break the chain edges p_2, e_2, p_3. If this is the case, then insert elements $4, \ldots, l$ on e_1 (e_2 respectively) to yield trees isomorphic to T_1 and T_2. Other elements may be inserted on e_1 (e_2, respectively), but since p_1, e_1, p_2 (p_2, e_2, p_3, respectively) are not broken, these additional elements are moved by subsequent moves. Thus, the distance $k = d_{uSPR}(T_1, T_2) = d_{uSPR}(T_1^3, T_2^3) = k'$.

Case 2: The minimal sequence of moves that transforms T_1^3 into T_2^3 breaks all three pendant edges: $\{p_1, p_2, p_3\}$. We can show a stronger result: $d_{uSPR}(T_1, T_2) = d_{uSPR}(T_1^3, T_2^3)$. Towards a contradiction, assume that $d_{uSPR}(T_1, T_2) > d_{uSPR}(T_1^3, T_2^3)$. Then there exists a sequence of moves, $m_1, \ldots, m_{k'} : T_1^3 \to T_2^3$ where $k' < k$. In T_1^3 and T_2^3, replace the chain element 3 by a chain subtree with leaves $3, \ldots, l+2$. Call the resulting trees T_1' and T_2' (see T_1' and T_2' in Figure 5). Since T_1^3 and T_2^3 can be obtained from T_1' and T_2' by applying the subtree reduction rule to the subtree chain with leaves $3, \ldots, l+2$, $d_{uSPR}(T_1', T_2') = d_{uSPR}(T_1^3, T_2^3)$.

Now, consider the trees T_1^{l+2} and T_2^{l+2} as defined in Section 2. T_1^{l+2} can be converted to T_1' by a single move (break e_{l+2} and connect to e_2, illustrated in Figure 5 between T_1^{l+2} and T_1'). Call this move m_0. Similarly, T_2^{l+2} can be converted to T_2' by a single move, $m_{k'+1}$ (see trees T_2' and T_2^{l+2} in Figure 5). Now, the sequence of moves $m_0, m_1, \ldots, m_{k'}, m_{k'+1}$ transforms T_1^{l+2} into T_2^{l+2}.

Let T_1'' be the tree T_1^{l+2} with the chain elements 1 and 2 removed. T_2'' is defined similarly from T_2^{l+2}. By hypothesis, p_1 and p_2 are broken by moves in $m_0, m_1, \ldots, m_{k'}, m_{k'+1}$. Let $m_1', \ldots, m_{k'}'$ be the result of removing these two moves, with the natural change: any edge that gets attached

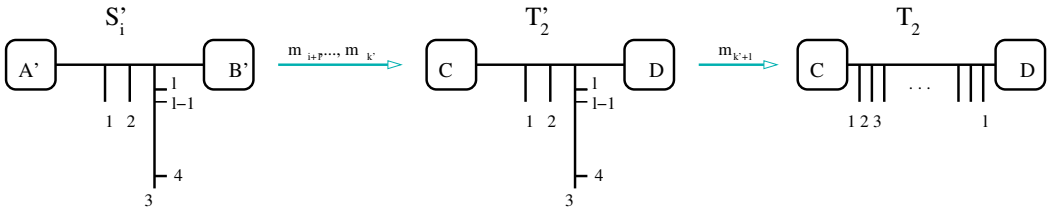


Figure 6: The intermediate trees in Case 5 of Theorem 2.

into edges e_0, e_1, e_2, p_1 or p_2 in the moves $m'_1, \ldots, m'_{k'}$, will now get attached to an edge that will be the contraction of e_1, e_2 and e_3. So, $m'_1, \ldots, m'_{k'}$ transforms T''_1 into T''_2. Finally, using a simple relabeling of chain elements, T''_1 is isomorphic to T_1 and T''_2 is isomorphic to T_2. This implies $d_{uSPR}(T_1, T_2) = d_{uSPR}(T''_1, T''_2) \leq k'$, which is a contradiction since we assumed that $d_{uSPR}(T_1, T_2) = k > k'$.

Case 3: There exists a minimal sequence of moves $m_1, \ldots, m_{k'}$ transforming T_1^3 into T_2^3 where exactly two of the pendant edges, $\{p_1, p_2, p_3\}$, are broken. This follows by a similar argument to above, where the chain element that is replaced by the subtree is the one not broken by the minimal sequence of moves.

Case 4: There exists a minimal sequence of moves $m_1, \ldots, m_{k'}$ transforming T_1^3 into T_2^3 where exactly one of the pendant edges, $\{p_1, p_2, p_3\}$, is broken. The argument is similar to Case 2 but does not yield as strong a result. We show $d_{uSPR}(T_1^3, T_2^3) \geq k-1$, instead of a strict inequality. We create trees as in Case 2, inserting a subtree with a chain of $l+1$ new elements on one of the pendant edges that is not broken. This gives a sequence of moves $m_0, m_1, \ldots, m_{k'}, m_{k'+1} : T_1^{l+1} \to T_2^{l+1}$. Unlike the previous two cases, there is only one pendant edge that is broken by $m_1, \ldots, m_{k'}$. Let $m'_1, \ldots, m'_{k'+1}$ be the result of removing this move, and T''_1 and T''_2 the result of removing the corresponding chain element from T'_1 and T'_2. Then, we have $m'_1, \ldots, m'_{k'+1} : T''_1 \to T''_2$. Under a simple relabeling of chain elements, T''_1 is isomorphic to T_1 and T''_2 is isomorphic to T_2. This implies $k = d_{uSPR}(T_1, T_2) = d_{uSPR}(T''_1, T''_2) \leq k'+1$, which gives $d_{uSPR}(T_1^3, T_2^3) = k' \geq k-1$.

Case 5: None of the above. Then every sequence of minimal moves transforming T_1^3 into T_2^3 breaks at leaset one of $\{e_1, e_2\}$, and it does not contain any moves breaking any pendant chain edges p_1 or p_2 or p_3.

Let $m_1, \ldots, m_{k'} : T_1^3 \to T_2^3$ be a minimal sequence of moves. Assume, without loss of generality that the edge, e_2 is broken in the direction of p_3, by move m_i, that is the component containing p_3 is the target of the move m_i. Also, assume that this is the first move where either e_1 or e_2 is broken. Let S_{i-1} be the result of applying moves $m_1, m_2, \ldots, m_{i-1}$ to T_1^3. Then, $d_{uSPR}(S_i, T_2^3) = k' - i - 1$ via the moves $m_i, m_{i+1}, \ldots, m_{k'}$.

In S_{i-1} and T_2^3, replace the chain element 3 by a subtree chain with leaves $l, \ldots, 3$. Call the resulting trees S'_{i-1} and T'_2 (see Figure 6). Since S_i and T_2^3 can be obtained from S'_i and T'_2 by applying subtree reduction rule to the chain, $l, \ldots, 3$, $d_{uSPR}(S'_{i-1}, T'_2) = d_{uSPR}(S_{i-1}, T_2^3) = k' - i - 1$. Now, let $m_{k'+1}$ break e_2 and connect it to the pendant edge, p_3, above the last chain element 3. This gives: $m'_i, \ldots, m'_{k'+1} : S'_i \to T_2$ and $d_{uSPR}(S'_i, T_2) = k' - i + 1$.

We note that changing slightly the moves $m_1, m_2, \ldots, m_{i-1}$, we can transform tree T_1 into S'_{i-1}. Recall that e_1, e_2, p_1, p_2, p_3, are not broken in steps $m_1, m_2, \ldots, m_{i-1}$. We transform the moves by: if some edge gets attached into p_3, e_2 (and in the move m_i stays with leaf 3), or e_3, in the

new move, attach it to e_l. The other moves are identical. Let us call this new set of moves: $m'_1, m'_2, \ldots, m'_{i-1}$. By construction, these new moves that transform T_1 into S'_{i-1}. Combining the two sets of moves, we obtain $m'_1, \ldots, m'_{i-1}, m'_i, \ldots m'_{k'+1} : T_1 \rightarrow T_2$. Thus, $k = d_{uSPR}(T_1, T_2) \leq k' + 1 = d_{uSPR}(T_1^3, T^3, 2) + 1$.

Therefore, $d_{uSPR}(T_1, T_2) \geq k - 1$. □

5 Acknowledgements

We would like to thank the anonymous refereees for their thoughtful comments and Charles Semple for insightful conversations. The first author was partially supported by the Spanish grant TIN2004-04343. Also, this project was partially supported by USA NSF grants ITR 0121651 and SEI 0513660. The second author would like to thank the Centre de Recerca Matemàtica at the Barcelona for hosting her visit for Spring 2005 and Fall 2006, and the Isaac Newton Institute, Cambridge University for hosting her in Fall 2007.

References

[1] Benjamin L. Allen and Mike Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5(1):1–15, 2001.

[2] Maria Luisa Bonet, Katherine St. John, Ruchi Mahindru, and Nina Amenta. Approximating subtree distances between phylogenies. *Journal of Computational Biology*, 13(8):1419–1434, 2006.

[3] Magnus Bordewich, Catherine McCartin, and Charles Semple. A 3-approximation algorithm for the subtree distance between phylogenies. *Journal of Discrete Algorithms*, In press: available online (5 January 2008), 2008.

[4] Magnus Bordewich and Charles Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combintorics*, 8:409–423, 2004.

[5] Michael Hallett and Catherine McCartin. A faster FPT algorithm for the maximum agreement forest problem. *Theory Comput. Syst.*, 41(3):539–550, 2007.

[6] Jotun Hein, Tao Jiang, Lusheng Wang, and Kaizhong Zhang. On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71(1-3):153–169, 1996.

[7] Glenn Hickey, Frank Dehne, Andrew Rau-Chaplin, and Christian Blouin. SPR distance computation for unrooted trees. *Evolutionary Bioinformatics*, 4:17–27, 2008.

[8] Catherine McCartin, 2008. personal communication.

[9] Mike Steel. SPR conjecture, 2002. www.math.canterbury.ac.nz/~m.steel/research/reward2.pdf.

[10] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In *Molecular Systematics, Second Edition*, pages 407–514. Sinauer, 1996.