The Computational Complexity of the Unrooted Subtree Prune and Regraft Distance

Glenn Hickey
Frank Dehne
Andrew Rau-Chaplin
Christian Blouin

Technical Report CS-2006-06

July 21, 2006

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

The Computational Complexity of the Unrooted Subtree Prune and Regraft Distance [*]

Glenn Hickey [†] Frank Dehne [‡] Andrew Rau-Chaplin [§] Christian Blouin [¶]

Abstract

We show that computation of the subtree prune and regraft (SPR) distance between unrooted binary phylogenetic trees is NP-Hard and fixed parameter tractable. Similar results exist for the related tree bisection and reconnection (TBR) distance, as well as the SPR distance between rooted trees but the complexity of the unrooted SPR case has heretofore remained unknown.

1 Introduction

Binary phylogenetic trees are used to describe evolutionary relationships between organisms. Typically, species represented by DNA or protein sequence information are associated with the leaves of the tree and the internal nodes correspond to speciation events. In order to model ancestor-descendant relationships on the tree, a direction must be associated with its edges. This is achieved by rooting the tree with a vertex of degree 2, representing a common ancestor to all species in the tree. Often, insufficient information exists to determine the roots of trees and, as such, they are left unrooted. Unrooted trees still provide a notion of evolutionary similarity between organisms even if the direction of descent remains unknown.

The phylogenetic tree representation has recently come under scrutiny with critics claiming that it is too simple to properly represent microbial evolution, particularly in the presence of Lateral Gene Transfer (LGT) events [4]. An LGT is the transfer of genetic material between species by other means than inheritance. Thus, LGT events cannot be represented in a tree as they would form cycles. The prevalence of LGT events in microbial evolution can, however, be studied using phylogenetic trees. Given a pair of trees of the same species, each constructed using different sets of genes, an LGT event will correspond to a displacement of a common subtree, referred to a subtree prune and regraft (SPR) operation. Determining the minimum number of SPR operations that explain the topological differences between a pair of trees thus yields the most parsimonious LGT scenario [2]. Computing this number of SPR operations, known as the SPR distance, is thus key to the study of the prevalence of LGT in bacterial evolution. In this paper, we investigate the computational complexity of this problem for unrooted trees. The outline is as follows. Basic definitions are provided in the remainder of this section. In Section 2 we show that SPR distance computation is NP-Hard for unrooted trees and in Section 3 we show that it is fixed parameter tractable.

[*] Research partially supported by the Natural Sciences and Engineering Research Council of Canada and Genome Atlantic.

[†] Faculty of Computer Science, Dalhousie University, Halifax, Canada, hickey@cs.dal.ca

[‡] School of Computer Science, Carleton University, Ottawa, Canada, frank@dehne.net

[§] Faculty of Computer Science, Dalhousie University, Halifax, Canada, arc@cs.dal.ca

[¶] Faculty of Computer Science, Dalhousie University, Halifax, Canada, cblouin@cs.dal.ca

Definition 1.1. **(1)** An unrooted binary phylogenetic tree T (or more briefly a phylogenetic tree) is a tree whose leaves (degree 1 vertices) are labeled bijectively by a (species) set S, and such that each non-leaf vertex is unlabeled and has degree three. **(2)** An edge of a tree T incident with a leaf is a pendant edge, otherwise we say it is an internal edge. Let $L(T)$ denote the leaf set of a tree T; the other vertices are said to be internal. **(3)** A forced contraction is an operation on a tree T in which we delete a vertex v of degree two and replace the two edges incident to v by a single edge.

Definition 1.2. A *subtree prune and regraft* (SPR) operation on a phylogenetic tree T is defined as cutting any edge and thereby pruning a subtree, t, and then regrafting the subtree t by the same cut edge to a new vertex obtained by subdividing a preexisting-existing edge in T. All vertices of degree 2 are then deleted using forced contractions.

Definition 1.3. A *tree bisection and reconnection* (TBR) operation on a 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 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.

Definition 1.4. An *SPR path* or *TBR path* between two trees T_1 and T_2 is a sequence of SPR operations or TBR operations, respectively, that converts T_1 into T_2.

Definition 1.5. The *SPR distance* (ΔSPR) and *TBR distance* (ΔTBR) between two trees T_1 and T_2 is the minimum number of SPR operations and TBR operations, respectively, required to convert T_1 into T_2.

2 SPR Distance Computation is NP-Hard for Unrooted Trees

Hein et al. [6] showed that determining if $\Delta SPR(T_1, T_2)$ is equal to some constant k is NP-Complete by providing two polynomial time reductions. The first transforms an instance of known NP-Complete problem *Exact Cover by 3-Sets* (X3C) into an instance of *Maximum Agreement Forest* (MAF) size of two rooted phylogenetic trees. The second transforms MAF size into SPR distance. This reduction from MAF to SPR is specified in Lemma 7 from [6] which states that $\Delta SPR(T_1, T_2) = |MAF(T_1, T_2)| - 1$ for any pair of rooted (or unrooted) phylogenetic trees T_1 and T_2.

Allen and Steel [1], however, have since provided a counterexample to Lemma 7 in [6], showing that the equation $\Delta SPR = |MAF| - 1$ does not always hold for rooted and unrooted trees, invalidating the NP-Hardness proof in Hein et al. [6]. Allen and Steel [1] were able to show that the relationship holds for the TBR distance (ΔTBR) of unrooted trees. Bordewich and Semple use a revised definition of the MAF to show that SPR distance computation is NP-hard for *rooted* trees [3]. However, the complexity of computing the SPR distance between unrooted trees has, to the best of our knowledge, remained an open problem.

The remainder of this section is devoted to proving that unrooted SPR is indeed NP-Hard. We first review the reduction from X3C to MAF in [6] and verify that the trees used in this reduction can be unrooted without altering the result. We then show that the SPR distance of the tree instances used in this reduction is equal to their TBR distance, thus showing that even though Lemma 7 from [6] is incorrect in the *general* case, it is valid for the trees in the reduction from X3C.

Definition 2.1. An *agreement forest* for two trees is any common forest that can be obtained from both trees by cutting the same number of edges from each tree, applying forced contractions after each cut. A *maximum agreement forest* (MAF) for two trees is an agreement forest with a minimum number of components. [6]

Definition 2.2. The *exact cover by 3-sets* (X3C) problem is defined as follows [5]: Given a set X with $|X| = n = 3q$ and a collection C of m 3-element subsets of X. Does C contain an exact cover for X, i.e., a sub-collection $C' \subseteq C$ such that every element of X occurs in exactly one member of C'?

NOTE: Remains NP-Complete if no element occurs in more than three subsets. Also note that this problem remains NP-Complete if each element occurs in *exactly* three subsets. This second property is implied by [6] though never explicitly stated. A supplemental proof is provided in Lemma A.1 of Appendix A.

2.1 Reduction from X3C to Rooted MAF

We now review the polynomial-time reduction from X3C to MAF for rooted trees provided by Hein et. al. [6], clarifying their notation to reflect that each element of X belongs to *exactly* three subsets in C, i.e. $|X| = |C| = 3q = m = n$, a fact implied but not clearly stated in their paper.

An instance of X3C is transformed into two rooted phylogenetic trees shown in Figure 1. Each element of X is represented by a triplet of the form $\{a, u, v\}$ and each triplet appears 3 times in each tree, once for each occurrence in a subset in C.

Tree T_1 is illustrated in Figure 1(a). Each subtree $A_i \in T_1$, shown in Figure 1(b) corresponds to a subset $c_i \in C$. Each subtree of A_i induced by the triple $\{a_{i,j}, u_{i,j}, v_{i,j}\}$ where $j \in \{1, 2, 3\}$ corresponds to a single element of X.

Tree T_2, shown in Figure 1(c), has the same leaf set as T_1 but a different topology. Each D_i subtree of T_2, as seen in Figure 1(e), corresponds to a subset in C except only the a-part of each triplet is present. Each B_i subtree of T_2, as seen in Figure 1(d), corresponds to an element in X. Each such element $x = \{a, u, v\}$ in the set X appears in three different subsets of C: c_j, c_k, and c_l. Without loss of generality, assume it consists of the first element of c_j, second element of c_k, and third element of c_l. The corresponding B tree would have leaves $\{u_{j,j'}, u_{k,k'}, u_{l,l'}, v_{j,j'}, v_{k,k'}, v_{l,l'}\}$ where $j' = 1, k' = 2, l' = 3$.

Hein et. al. show that $|MAF(T_1, T_2)| = 20q + 1$ if and only if C contains an exact cover of X.

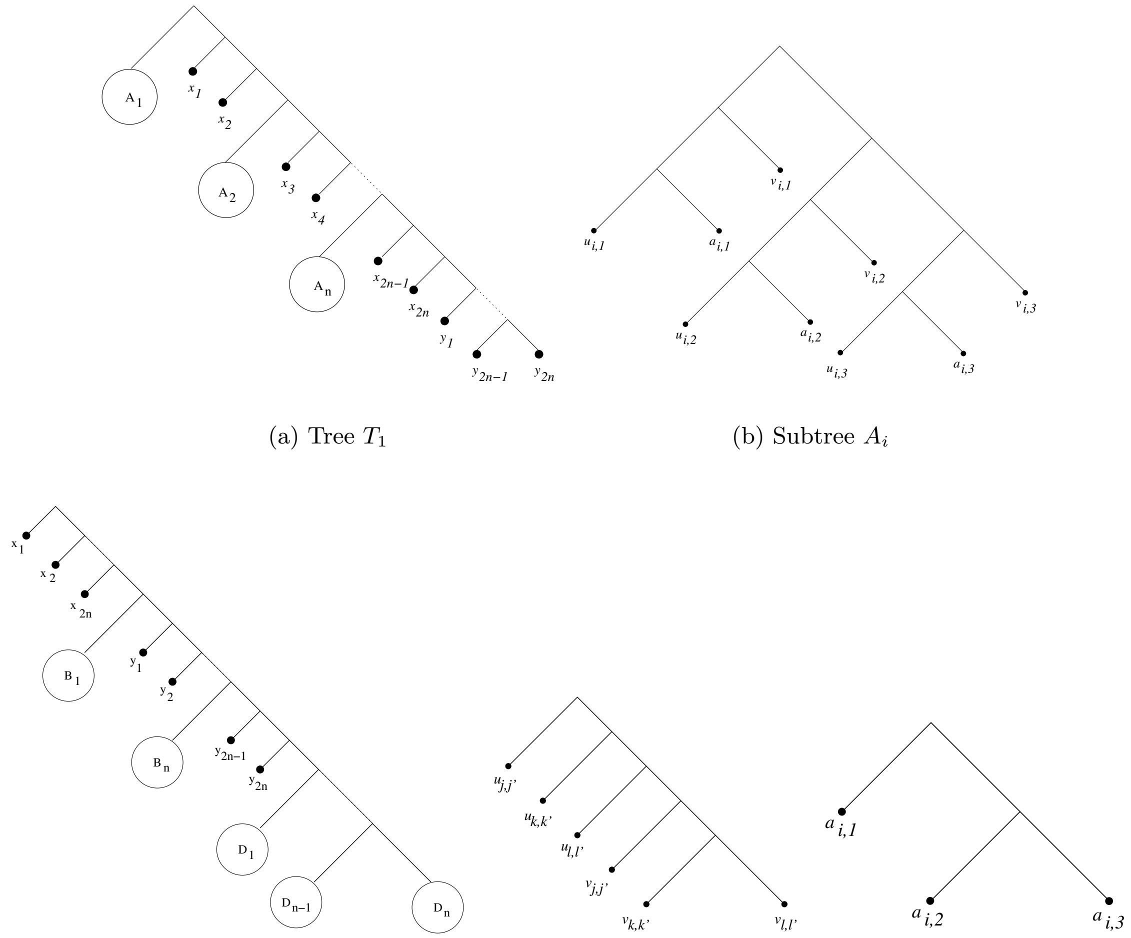
2.2 Reduction from X3C to Unrooted SPR

We begin by verifying that the reduction from X3C to rooted MAF from [6] described above can be trivially applied to the unrooted case.

Lemma 2.1. *Given an instance of X3C where $|X| = |C| = 3q$,*

$$|MAF(T_1, T_2)| = |MAF(U_1, U_2)|$$

where T_1 and T_2 are the trees obtained by the reduction in [6] (Figure 1) and U_1 and U_2 are unrooted versions of T_1 and T_2 displayed in Figure 2. Accordingly, the instance of X3C has a solution if and only if $|MAF(U_1, U_2)| = 20q + 1$ as this equality was shown for $|MAF(T_1, T_2)|$ in [6].



(a) Tree T_1

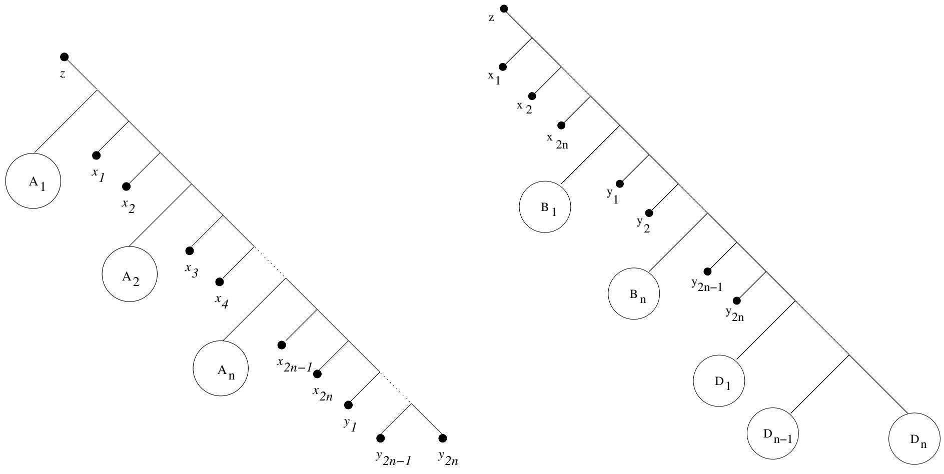
(b) Subtree A_i

(c) Tree T_2

(d) Subtree B_i

(e) Subtree D_i

Figure 1: Reduction of an instance of X3C to $|MAF(T_1, T_2)|$ from [6]. Each element of X corresponds to an $\{a, u, v\}$ triplet. The instance of X3C has a solution if and only if $|MAF(T_1, T_2)| = 20q + 1$ (where $n = 3q$).



(a) Tree U_1 (b) Tree U_2

Figure 2: Unrooted version of T_1 and T_2 from Figure 1. The unrootedness does not affect the number of components in the MAF.

Proof. Trees T_1 and T_2 can be unrooted by adding a leaf z as a pendant to the root, creating trees U_1 and U_2 shown in Figure 2. Recall from [6] that $MAF(T_1, T_2)$ contains a component consisting of the chain $x_1, ...x_{2n}, y_1, ..., y_{2n}$. Observe that $|MAF(U_1, U_2)| \leq |MAF(T_1, T_2)|$ as an agreement forest for U_1 and U_2 can be created from $MAF(T_1, T_2)$ by adding z to the xy chain. Furthermore, $|MAF(U_1, U_2)| \geq |MAF(T_1, T_2)|$ as, for the same reasons outlined in the proof of Lemma 3.2, adding a leaf to both trees cannot decrease their TBR distance. It follows that $|MAF(T_1, T_2)| = |MAF(U_1, U_2)|$. □

We now provide a transformation from X3C to unrooted SPR.

Lemma 2.2. *For any instance of X3C where $|X| = |C| = 3q$,*

$$\Delta SPR(U_1, U_2) = |MAF(U_1, U_2) - 1|$$

Where U_1 and U_2 are unrooted versions of the trees obtained using the reduction in [6] as described above. Note that this is a very restricted version of Lemma 7 from [6].

Proof. It is sufficient to show that the inequality $\Delta SPR(U_1, U_2) \leq |MAF(U_1, U_2) - 1|$ is true as $\Delta SPR(U_1, U_2) \geq |MAF(U_1, U_2) - 1|$ follows from Lemma 2.7(b) and Theorem 2.13 from [1].

$MAF(U_1, U_2)$ is formed by the cutting edges from A_i subtrees (and the corresponding subtrees in U_2) in either of two possible ways [6]:

1. Cut leaves $u_{i,1}, v_{i,1}, u_{i,2}, v_{i,2}, u_{i,3}, v_{i,3}$ and then prune the remaining subtree formed by leaves $\{a_{i,1}, a_{i,2}, a_{i,3}\}$. Such a procedure contributes 7 components to the MAF.

2. Cut the leaves $a_{i,1}, a_{i,2}, a_{i,3}$ then cut each of the remaining two-leaf subtrees: $\{u_{i,1}, v_{i,1}\}$, $\{u_{i,2}, v_{i,2}\}$, and $\{u_{i,3}, v_{i,3}\}$. These operations contribute 6 components to the MAF

We now show that given two trees U_1 and U_2 and their MAF, which was created using the above cut operations, there exists $|MAF| - 1$ SPR operations that can transform U_1 to U_2. In particular, for each set of cut operations, there exists an equivalent set of SPR operations.

1. Prune leaves $u_{i,1}, v_{i,1}, u_{i,2}, v_{i,2}, u_{i,3}, v_{i,3}$ from A_i and regraft them onto the chain, forming B_i subtrees in the required positions. Prune the subtree $\{a_{i,1}, a_{i,2}, a_{i,3}\}$ and regraft into the position of D_i. In this case, 7 SPR operations are performed.

2. Prune the leaves $a_{i,1}, a_{i,2}, a_{i,3}$ and regraft them onto the chain, forming a D_i subtree in the proper position. Prune the remaining two-leaf subtrees: $\{u_{i,1}, v_{i,1}\}$, $\{u_{i,2}, v_{i,2}\}$, and $\{u_{i,3}, v_{i,3}\}$ and regraft them onto the chain, forming B_i subtree components in the required position. 6 SPR operations are used.

There is a one-to-one correspondence between cuts formed when creating the MAF and SPR operations that can transform U_1 to U_2. Thus $\Delta SPR(U_1, U_2) \leq |MAF(U_1, U_2)| - 1$. □

Theorem 2.3. *SPR distance is NP-Hard for unrooted phylogenetic trees.*

Proof. By Lemma 2.1, an instance of X3C with $|X| = |C| = 3q$ can be reduced to a pair of unrooted trees U_1 and U_2 such that X3C has a cover if and only if $|MAF(U_1, U_2)| = 20q + 1$. Lemma 2.2 shows that $\Delta SPR(U_1, U_2) = |MAF(U_1, U_2)| - 1$, completing the reduction from X3C to SPR distance. Note that, in fact, we have actually showed that deciding if the SPR distance between unrooted trees equals a given constant is NP-Complete. Since actually finding the distance would solve this decision problem, SPR distance computation is NP-Hard for unrooted trees. □

3 SPR Distance Computation is Fixed-Parameter Tractable for Unrooted Trees

We will now show that SPR distance computation for unrooted trees is fixed-parameter tractable, where the parameter is the SPR distance. This was conjectured in [1] but could not be proven so far.

We will re-use the kernelization for TBR distance calculation introduced by Allen and Steel [1]. As they show, the repeated application of the following Rule 1 and Rule 2 operations to a set of two trees reduces their sizes to a linear function of their TBR distance while preserving the TBR distance.

Definition 3.1. of RULE 1: Replace any pendant subtree that occurs in both trees by a single leaf with a new label. See Figure 3(a).

Definition 3.2. of 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. See Figure 3(b).

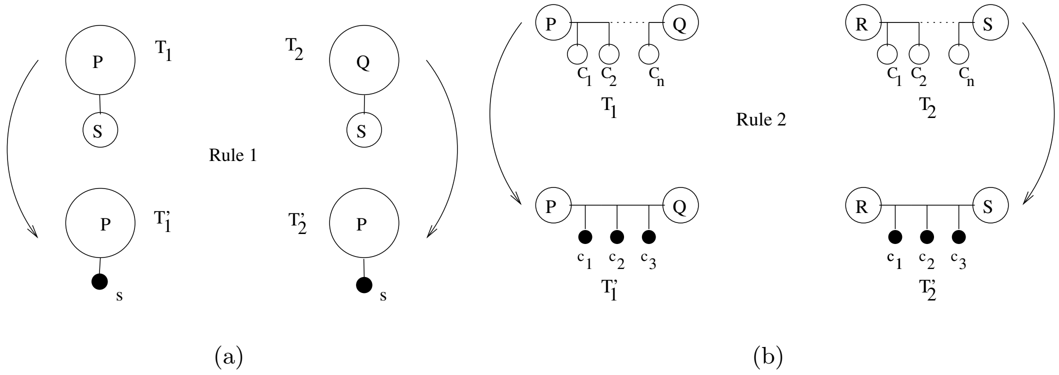


Figure 3: Rules 1 and 2 from Allen and Steel [1].

Allen and Steel [1] showed that Rule 1 also leaves the SPR distance unchanged. They conjectured the same to be true for Rule 2.

Lemma 3.1. *Rule 1 preserves SPR-Distance [1].*

In the remainder of this section, we prove Allen and Steel's conjecture that Rule 2 preserves SPR-Distance.

Lemma 3.2. *If any single common leaf is pruned from both T_1 and T_2, their SPR distance is either unaffected or reduced by 1.*

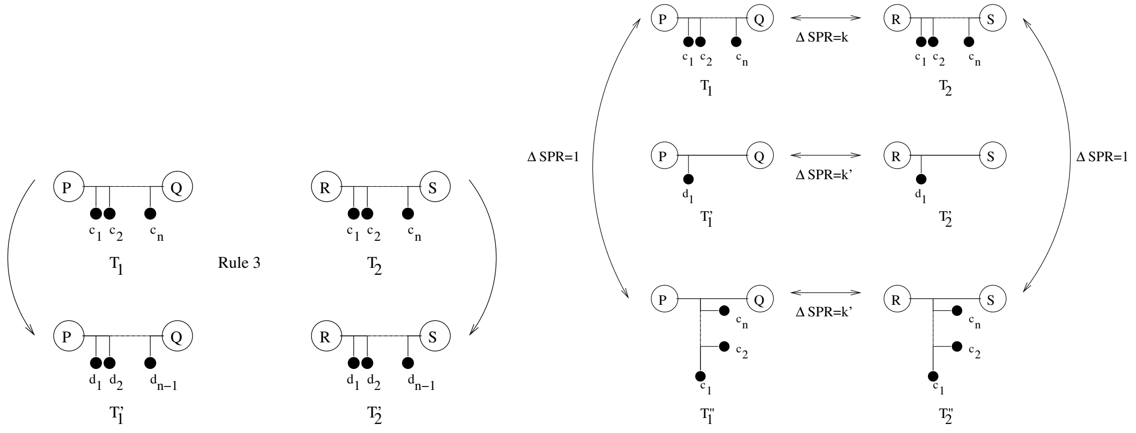
Proof. If any single common leaf is pruned from both T_1 and T_2, the same SPR path still converts T_1 into T_2. Hence, the SPR distance can not increase. Suppose that the SPR distance decreases by 2 (or more). Take the SPR path for the reduced trees and add a single SPR operation to handle the deleted leaf. This results in a SPR path between T_1 and T_2 that is shorter than their SPR distance, a contradiction. □

We now introduce a new Rule 3. Its properties will be used to show that Rule 2 preserves SPR distance.

Definition 3.3. of RULE 3: Replace any (non-empty) chain of pendant leaves of length n that occur identically in both trees by a chain of $n-1$ new leaves with new labels correctly oriented to preserve the direction of the chain as shown in Figure 4(a).

Lemma 3.3. *Successive applications of Rule 3 to common chains of length* > 1 *cannot reduce the SPR-Distance by more than 2.*

Proof. Let T_1 and T_2 be two phylogenetic trees sharing a common chain $C = \{c_1, c_2, ..., c_n\}$ of length $n > 1$ whose SPR distance is k as shown in Figure 4(b). Repeatedly applying Rule 3 to this chain $n-1$ times yields T_1' and T_2' whose common chain contains a single element and whose SPR distance is k'. We now show that $k \leq k'+2$. Let P be the pendant subtree of T_1 rooted at the vertex adjacent to c_1 such that $P \cap C = \emptyset$. Similarly, let Q be the pendant subtree of T_1 rooted at the vertex adjacent to c_n such that $Q \cap C = \emptyset$. Let R and S be the subtrees of T_2 analogous to P and Q respectively. Pruning P and regrafting to the stem of Q and pruning Q and regrafting to the stem of R yields T_1'' and T_2'' respectively as shown in Figure 4(b). Observe that C is now a common subtree of both T_1'' and T_2'' and thus, by Lemma 3.1, $\Delta SPR(T_1'', T_2'') = \Delta SPR(T_1', T_2') = k'$ as C can be reduced via Rule 1 to a chain with a single leaf without affecting the distance. Thus, T_1 can be transformed to T_2 by performing a single SPR to yield T_1'', then k' SPR's to give T_2'' and then a single SPR to transform T_2'' into T_2. Therefore $k \leq k' + 2$. The above steps have shown that reducing a chain of length n to a chain of length 1 using Rule 3 cannot decrease the SPR distance by more than 2. It follows from Lemma 3.2 that using Rule 3 to reduce a chain of length n to a chain of any length less than n cannot decrease the SPR distance by more than 2. □



(a) Rule 3.

(b) Illustration of Lemma 3.3: Application of Rule 3 cannot reduce $\Delta SPR(T_1, T_2)$ by more than 2 since T_1 can always be transformed into T_2 with $\leq k' + 2$ SPR operations regardless of the length of the common chain.

Figure 4: A New Rule 3 and Illustration of Lemma 3.3.

Lemma 3.4. *If Rule 3 preserves SPR distance on common chains of length $n \geq 1$, then it preserves distance for common chains of any length $> n$.*

Proof. Suppose the two trees share the chain $C = \{c_1, c_2, ..., c_{n+1}\}$. Note that these trees also share a common subchain $C' = \{c_1, c_2, ..., c_n\}$. If Rule 3 is SPR distance preserving for chains of length n, then we can apply it to C' yielding a new common chain: $D = \{d_1, d_2, ..., d_{n-1}\}$. The only element left from C is c_{n+1} which follows d_n in both trees. C has thus been transformed into the common chain $C'' = \{d_1, d_2, ..., d_{n-1}, c_{n+1}\}$. There are n elements in C''. C has been reduced from $n + 1$ to n elements. We have shown that if Rule 3 is SPR distance preserving for chains of length n, then it is also distance preserving for chains of length $n + 1$. Thus, if Rule 3 is distance preserving for chains of length n, then by induction it is distance preserving for chains of any length $> n$. □

Lemma 3.5. *Rule 2 preserves SPR-distance. (This was conjectured in [1].)*

Proof. Rule 2 can be decomposed into applications of Rule 1 and Rule 3, the latter only on chain lengths greater than 3. Thus, proving that Rule 3 is distance preserving on such chain lengths is sufficient to prove that Rule 2 is always distance preserving. Let T_1 and T_2 be two trees sharing a common chain $C = \{c_1, c_2, ..., c_n\}$. We will use induction on chain length

n to show that Rule 3 can be applied for any $n > 3$ without altering the SPR distance. Lemma 3.4 has already proved the induction step so all that remains is to show that Rule 2 is distance preserving in the base case.

The base case will be when $n = 4$. Let $\Delta SPR = k$ in this case. Now suppose that Rule 3 *Does Not* preserve the SPR distance when $n = 4$. Proof that this is a contradiction follows. If Rule 3 is not distance preserving for $n = 4$ then it is not distance preserving for $1 < n < 4$. Otherwise it would contradict Lemma 3.4: Suppose that Rule 3 preserves distance for some i, $1 < i < 4$. Then Rule 3 must preserve distance for any $n > i$ which includes $n = 4$. This is a contradiction. Recall that if Rule 3 is not distance preserving then by Lemma 3.2, its application will reduce the SPR distance by 1 whenever it is applied to chains of length $n > 1$. Now back to our chain of length 4. We have shown that each time we reduce the length of this chain with Rule 3, we decrease the SPR distance. That is, for $n = 4$ with $\Delta SPR = k$ by applying Rule 3 we obtain $n = 3$ with $\Delta SPR = k - 1$, and by applying Rule 3 again we obtain $n = 2$ with $\Delta SPR = k - 2$, and by applying Rule 3 a third time we obtain $n = 1$ with $\Delta SPR = k - 3$. However, Lemma 3.3 states that a difference in chain length can change ΔSPR by at most 2! Assuming that Rule 3 *Does Not* preserve SPR distance for $n = 4$ is therefore a contradiction because it implies that the SPR distance can be reduced by 3 using chain reductions. Thus, there *must be* some chain length ≤ 4 such that application of Rule 3 does not affect the SPR distance. As a consequence of Lemma 3.4, Rule 3 does not affect distance when $n = 4$ and the base case is valid. □

Theorem 3.6. *SPR distance computation for unrooted trees is fixed-parameter tractable.*

Proof. Lemmas 3.1 and 3.5 show that Rule 1 and Rule 2 preserve the SPR distance. What remains to be shown is that the tree size after kernelization is bounded by a function of the SPR distance only. This follows by applying two results from [1] for two unrooted phylogenetic trees T_1 and T_2: $\Delta TBR(T_1, T_2) \leq \Delta SPR(T_1, T_2)$ ([1], Lemma 2.4) and $n' \leq 4c(\Delta TBR(T_1, T_2) - 1)$ ([1], Theorem 3.8) where n' is the size of the trees after kernelization and $c \leq 7$ is a constant. Combining these we obtain $n' \leq 28(\Delta SPR(T_1, T_2) - 1)$ which completes the proof. □

4 Conclusion

We have shown that computing the SPR distance between two unrooted phylogenetic trees is NP-Hard and FPT. These results are not altogether surprising given that the same properties have been shown for the related TBR and rooted SPR metrics [1, 3], but solve a long-standing open problem none the less. Presently, we are developing an algorithm for unrooted SPR distance based on the FPT kernelization presented in Section 3 with the intent of testing its performance on real data.

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] Robert G. Beiko and Nicholas Hamilton. Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology*, 15(6), 2006.

[3] Magnus Bordewich and Charles Semple. On the compuational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8(4):409 – 423, 2004.

[4] W. Ford Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124–2128, 1999.

[5] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company, 1979.

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

Appendix

A X3C Clarification

Lemma A.1. *X3C is NP-Complete if each element occurs in* exactly *three subsets.*

Proof. Consider an instance of X3C in which no element occurs in more than three subsets. We provide a polynomial time reduction from such an instance, known to be NP-Complete, into an instance in which each element occurs in exactly three subsets. Let:

$Y_1 \subseteq X$: Elements of X that appear in exactly one subset
$Y_2 \subseteq X$: Elements of X that appear in exactly two subsets
$Y_3 \subseteq X$: Elements of X that appear in exactly three subsets

So $|Y_1| + 2|Y_2| + 3|Y_3| = |X| = 3q$

For each element to appear in exactly three subsets, we must add $2|Y_1| + |Y_2|$ elements to subsets in C.

Let multiset $Z = \{z_0, z_1, \ldots, z_{3p-1}\} = Y_1 + Y_1 + Y_2$ be these elements we have to add. Note that $|Z| = 3p$ where $p = 2(q - |Y_3|) - |Y_2|$.

Let $X' = \{x'_0, x'_1, \ldots, x'_{3p-1}\}$ be a set of new elements such that $|X'| = 3p$ and $X \cap X' = \emptyset$.

We now create a collection C' of new subsets out of Z and X' so that each element in $X \cup X'$ appears in a subset in $C + C'$ exactly three times.

For each $i = 0, 3, \ldots, 3p - 1$, we add four subsets to C':
$c'_{4i} = \{x'_i,\ x'_{i+1},\ x'_{x+2}\}$
$c'_{4i+1} = \{z_i,\ x'_i,\ x'_{i+1}\}$
$c'_{4i+2} = \{z_{i+1},\ x'_{i+1},\ x'_{i+2}\}$
$c'_{4i+3} = \{z_{i+2},\ x'_{i+2},\ x'_i\}$

We now show that $X \cup X'$ and $C + C'$ form an instance of X3C such that every element of $X \cup X'$ appears in 3 subsets in $C + C'$ and X has a cover in C if and only if $X \cup X'$ has a cover in $C + C'$.

(if): If X has a cover in C, then $X \cup X'$ has a cover in $C + C'$: Let $S \subseteq C$ be the cover of X. Then $S + c'_0 + c'_4 + c'_8 + \ldots + c'_{12p-1}$ is a cover $X \cup X'$.

(only if): If $X \cup X'$ has a cover in $C + C'$, then X has a cover in C: Similar to above, the only way to cover X' is with $c'_0 + c'_4 + c'_8 + \ldots + c'_{12p-1}$ and no other elements of C' can be part of an exact cover. This means that X is covered entirely by subsets in C so X is exactly covered by C. □