On the Subnet Prune and Regraft distance

Jonathan Klawitter and Simone Linz

Department of Computer Science, University of Auckland, New Zealand
jo.klawitter@gmail.com, s.linz@auckland.ac.nz

Abstract

Phylogenetic networks are rooted directed acyclic graphs that represent evolutionary relationships between species whose past includes reticulation events such as hybridisation and horizontal gene transfer. To search the space of phylogenetic networks, the popular tree rearrangement operation rooted subtree prune and regraft (rSPR) was recently generalised to phylogenetic networks. This new operation - called subnet prune and regraft (SNPR) - induces a metric on the space of all phylogenetic networks as well as on several widely-used network classes. In this paper, we investigate several problems that arise in the context of computing the SNPR-distance. For a phylogenetic tree T and a phylogenetic network N, we show how this distance can be computed by considering the set of trees that are embedded in N and then use this result to characterise the SNPR-distance between T and N in terms of agreement forests. Furthermore, we analyse properties of shortest SNPR-sequences between two phylogenetic networks N and N', and answer the question whether or not any of the classes of tree-child, reticulation-visible, or tree-based networks isometrically embeds into the class of all phylogenetic networks under SNPR.

1 Introduction

Many algorithms that have been developed to reconstruct phylogenetic trees from molecular sequence data require a (heuristic) search of the space of all phylogenetic trees [Fel04]. To this end, local rearrangement operations, such as nearest neighbor interchange, subtree prune and regraft, and tree bisection and reconnection, have been introduced that induce metrics on the space of phylogenetic trees [SOW96]. More recently, rooted phylogenetic networks, which are leaf-labelled rooted directed acyclic graphs, have become increasingly popular in the analysis of ancestral relationships between species whose past includes speciation as well as reticulation events such as hybridisation and horizontal gene transfer [Gus14, HRS10]. In particular, each vertex in a rooted phylogenetic network whose in-degree is at least two represents a reticulation event and is referred to as a *reticulation*. In comparison to tree space, the space of phylogenetic networks is significantly larger and searching this space remains poorly understood although the

above-mentioned rearrangement operations on phylogenetic trees have been generalised to rooted (and unrooted) phylogenetic networks [BLS17, FHMW18, GvIJ$^+$17, HLMW16, HMW16, JJE$^+$17, Kla17].

The goal of this paper is to advance our understanding of the subnet prune and regraft (SNPR) operation [BLS17] with a particular focus on the induced distance. For two phylogenetic networks, this distance equates to the minimum number of SNPR operations that is required to transform one network into the other one. SNPR generalises the rooted subtree prune and regraft (rSPR) operation [AS01, BS05, SOW96] from rooted phylogenetic trees to rooted phylogenetic networks. A second generalisation of the rSPR operation from trees to networks was recently introduced by Gambette et al. [GvIJ$^+$17]. Both generalisations are similar in the sense that they allow horizontal as well as vertical rearrangement moves. From a practical perspective, the space of phylogenetic networks can be searched horizontally in tiers, where a tier contains all phylogenetic networks with a fixed number of reticulations, as well as vertically among different tiers since a single operation can increase or decrease the number of reticulations by at most one. On the other hand, there are also subtle differences between the two operations. While SNPR is defined on rooted phylogenetic networks that allow for parallel edges [BLS17], the generalisation of rSPR to networks as introduced by Gambette et al. [GvIJ$^+$17] is defined on networks that do not allow for parallel edges. Moreover, the latter operation allows for the switching of a parent vertex (referred to as a *tail moves*) and for the switching of a child of a reticulation (referred to as a *head moves*) while SNPR only allows for tail moves. Under SNPR, tail moves are sufficient to establish that the operation induces a metric on the space of all rooted phylogenetic networks. Moreover, SNPR also induces a metric on the space of several popular classes of phylogenetic networks, such as tree-child, reticulation-visible, and tree-based networks [CRV09, FS15], regardless of whether or not one restricts to subclasses of these networks that have a fixed number of reticulations.

Since computing the rSPR-distance between two phylogenetic trees is NP-hard [BS05], it is not surprising that calculating the SNPR-distance as well as the distance induced by the operation introduced by Gambette et al. [GvIJ$^+$17] and further investigated by Janssen et al. [JJE$^+$17] is also NP-hard. In this paper, we investigate problems that arise in the context of computing the SNPR-distance. Bordewich et al. [BLS17] established several bounds on the SNPR-distance and showed that, for a rooted phylogenetic tree T and a rooted phylogenetic network N, the SNPR-distance $\mathrm{d_{SNPR}}(T, N)$ between T and N is equal to the number of reticulations in N if T is embedded in N. In the first part of this paper, we extend their result by showing how $\mathrm{d_{SNPR}}(T, N)$ can be computed regardless of whether or not T is embedded in N. Roughly speaking, the problem of computing the SNPR-distance is equivalent to computing the minimum rSPR-distance between all tree pairs consisting of T and a tree that is embedded in N. Hence, one way of computing $\mathrm{d_{SNPR}}(T, N)$ is by repeatedly solving the rSPR-distance problem between two trees. We use this result to show that computing $\mathrm{d_{SNPR}}(T, N)$ is fixed-parameter tractable. We then show that $\mathrm{d_{SNPR}}(T, N)$ can also be characterised in terms of agreement forests. The notion of agreement forests is the underpinning concepts for almost all theoretical results as well as practical algorithms that are related to computing the rSPR-distance

between two rooted phylogenetic trees [BS05, CFS15, WBZ16, Wu09]. We extend this notion to computing $\mathrm{d_{SNPR}}(T, N)$, which allows us to work directly on T and N instead of different tree pairs. In the second part of this paper, we turn to problems that are related to finding shortest SNPR-sequences for two rooted phylogenetic networks N and N' with r and r' reticulations, respectively. While the length of a shortest SNPR-sequence for N and N' is equal to $\mathrm{d_{SNPR}}(N, N')$ it additionally contains all intermediate networks that are obtained by a sequence of SNPR operations that transform N into N'. Here, we are interested in revealing properties of the intermediate networks in a shortest SNPR-sequence for N and N'. For example, if there is always a sequence with the property that each network in the sequence has at least $\min(r, r')$ and at most $\max(r, r')$ reticulations, then this might have positive implications in devising practical search algorithms because the search space could be pruned appropriately. Surprisingly, we find that, even if $r = r'$, it is possible that every shortest SNPR-sequence for N and N' contains a network with strictly more than r' reticulations. Moreover, for each r with $r \geq 1$, there exist two rooted phylogenetic networks that both have r reticulations and for which every shortest SNPR-sequence contains a rooted phylogenetic tree.

The paper is organised as follows. The next section contains notation and terminology that is used throughout the rest of this paper. Section 3 establishes a new result that equates the SNPR-distance between a phylogenetic tree T and a phylogenetic network N to the rSPR-distance between pairs of trees. This result is used in Section 4 to characterise the SNPR-distance between T and N in terms of agreement forests. We then investigate properties of shortest SNPR-sequences between two phylogenetic networks in Section 5. We end this paper with some concluding remarks in Section 6

2 Preliminaries

This section provides notation and terminology that is used in the remainder of the paper. In particular, we will introduce notation in the context of phylogenetic networks as well as the SNPR operation. Throughout this paper, $X = \{1, 2, \ldots, n\}$ denotes a finite set.

Phylogenetic networks. A *rooted binary phylogenetic network* N on X is a rooted directed acyclic graph with the following vertices:

- the unique *root* ρ with in-degree zero and out-degree one,
- *leaves* with in-degree one and out-degree zero bijectively labelled with X,
- *inner tree vertices* with in-degree one and out-degree two, and
- *reticulations* with in-degree two and out-degree one.

The *tree vertices* of N are the union of the inner tree vertices, the leaves and the root. An edge $e = (u, v)$ is called *reticulation edge*, if v is a reticulation, and *tree edge*, if v is a tree vertex. The set X is referred to as the *label set* of N and is sometimes denoted by

$L(N)$. Following Bordewich et al. [BLS17], we allow edges in N to be in *parallel*, that is, two distinct edges join the same pair of vertices. Also note that our definition of the root is known as *pendant root* [BLS17] and it differs from another common definition where the root has out-degree two. Our variation serves both elegance and technical reasons.

Let N be a rooted binary phylogenetic network on X. For two vertices u and v in N, we say that u is a *parent* of v and v is a *child* of u if there is an edge (u, v) in N. Similarly, we say that u is *ancestor* of v and v is *descendant* of u if there is a directed path from u to v in N. Lastly, u and v are *siblings* if they have a common parent.

A *rooted binary phylogenetic tree* on X is a rooted binary phylogenetic network that has no reticulations.

To ease reading, we refer to a rooted binary phylogenetic network (resp. rooted binary phylogenetic tree) on X simply as a phylogenetic network or network (resp. phylogenetic tree or tree). Furthermore, we denote the set of all phylogenetic networks on X by $\mathcal{N}_n$ and the set of all phylogenetic trees on X by $\mathcal{T}_n$.

Let $T \in \mathcal{T}_n$ and $N \in \mathcal{N}_n$. Then we say N *displays* T if, up to suppressing degree-two vertices, T can be obtained from N by deleting edges and vertices, in which case the underlying directed graph is an *embedding* of T into N. The set of all phylogenetic trees that are displayed by N is denoted by $D(N)$.

Classes of phylogenetic networks. Let $N \in \mathcal{N}_n$. The network N is a *tree-child* network if each of its non-leaf vertices has a tree vertex as child. A vertex v of N is called *visible* if there is a leaf l in N such that every directed path from the root of N to l traverses v. We say that N is a *reticulation-visible* network if every reticulation of N is visible. Lastly, N is *tree based* if there exists a direction-preserving embedding of a phylogenetic tree T in N that covers every vertex of N. For a fixed n, the class of tree-child networks is denoted by $\mathcal{TC}_n$, of reticulation-visible networks by $\mathcal{RV}_n$, and of tree-based networks by $\mathcal{TB}_n$. Each tree-child network is also a reticulation-visible network [HRS10] and each reticulation-visible network is also a tree-based network [GGL$^+$15, FS15].

SNPR. Let $N \in \mathcal{N}_n$ with root ρ and let $e = (u, v)$ be an edge of N. Bordewich et al. [BLS17] introduced the *SubNet Prune and Regraft (*SNPR*)* operation that transforms N into a phylogenetic network N' in one of the following three ways:

(SNPR^0) If u is a tree vertex (and $u \neq \rho$), then delete e, suppress u, subdivide an edge that is not a descendant of v with a new vertex u', and add the edge (u', v).

(SNPR$^+$) Subdivide (u, v) with a new vertex v', subdivide an edge in the resulting network that is not a descendant of v' with a new vertex u', and add the edge (u', v').

(SNPR$^-$) If u is a tree vertex and v is a reticulation, then delete e, and suppress u and v.

In what follows, we sometimes need to specify which of the three operations we consider, in which case we use 0, +, or − as a superscript to indicate the type of operation.

The three types of operations are illustrated in Figure 1. We note that an SNPR^0 does not change the number of reticulations, while an SNPR^- decreases it by one and an SNPR^+ increases it by one. Lastly, it is worth noting that the well known rSPR operation [BS05] on phylogenetic trees is a restriction of SNPR in which N and N' are phylogenetic trees and N is transformed into N' by SNPR^0 operations.

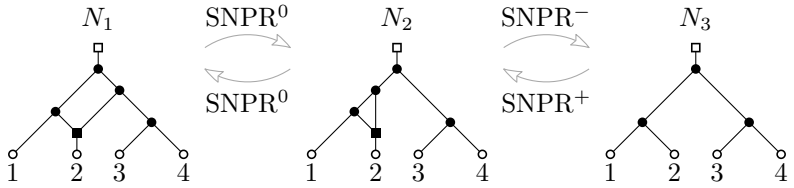

Figure 1: The phylogenetic network N_2 can be obtained from N_1 by an SNPR^0 and the phylogenetic network N_3 can be obtained from N_2 by an SNPR^-. Both operations have a corresponding SNPR^0 and SNPR^+, respectively, that reverses the transformation.

SNPR-distance. Let $N, N' \in \mathcal{N}_n$. An SNPR-*sequence* from N to N' is a sequence

$$\sigma = (N = N_0, N_1, N_2, \ldots, N_k = N')$$

of phylogenetic networks such that, for each $i \in \{1, 2, \ldots, k\}$, we can obtain N_i from N_{i-1} by a single SNPR. The *length* of σ is k.

Now, let $\mathcal{C}$ be a class of phylogenetic networks. Then $\mathcal{C}$ is said to be *connected* (under SNPR) if, for all pairs N and N' of networks in $\mathcal{C}$, there exists an SNPR-sequence from N to N'. Moreover, if $\mathcal{C}$ is connected, then the SNPR-*distance* between two elements in $\mathcal{C}$, say N and N', is the length of a shortest SNPR-sequence from N to N' with the property that each network of the sequence is in $\mathcal{C}$. We denote this distance by $d_{\mathrm{SNPR}_\mathcal{C}}(N, N')$ or simple $\mathrm{d}_{\mathrm{SNPR}}(N, N')$ if the class under consideration is clear from the context. Finally, let $\mathcal{C}$ and $\mathcal{C}'$ be two classes of phylogenetic networks such that all elements in $\mathcal{C}$ are also contained in $\mathcal{C}'$. We say that $\mathcal{C}$ *isometrically embeds* into $\mathcal{C}'$ if $d_{\mathrm{SNPR}_\mathcal{C}}(N, N') = d_{\mathrm{SNPR}_{\mathcal{C}'}}(N, N')$ for all pairs N and N' of networks in $\mathcal{C}$.

The following proposition, which was established by Bordewich et al. [BLS17], summarises some properties of the SNPR operation.

Proposition 2.1 ([BLS17, Corollary 3.3])**.**
The SNPR *operation induces a metric on* $\mathcal{N}_n, \mathcal{TC}_n, \mathcal{RV}_n, \mathcal{TB}_n$.

Note that Proposition 2.1 implies that the four different classes of phylogenetic networks mentioned in the statement are connected under SNPR and that the SNPR operation is reversible, that is, if a phylogenetic network N' can be obtained from a phylogenetic network N by a single SNPR, then N can also be obtained from N' by a single SNPR.

3 Characterising the SNPR-distance between a network and a tree

In this section, we characterise the SNPR-distance $\mathrm{d_{SNPR}}(T, N)$ between a phylogenetic network N and a phylogenetic tree T in terms of $D(N)$, the set of phylogenetic trees that are displayed by N. Bordewich et al. [BLS17] have shown how to compute this distance if T is displayed by N. To give a full characterisation of $\mathrm{d_{SNPR}}(T, N)$ regardless of whether or not T is displayed by N, we make use of the following three lemmata.

Lemma 3.1 ([BLS17, Lemma 7.4])**.**
Let $N \in \mathcal{N}_n$ with r reticulations. Let $T \in D(N)$. Then

$$\mathrm{d_{SNPR}}(T, N) = r.$$

Lemma 3.2 ([BLS17, Proposition 7.1])**.**
Let $T, T' \in \mathcal{T}_n$. Then

$$\mathrm{d_{rSPR}}(T, T') = \mathrm{d_{SNPR}}(T, T').$$

Moreover, the class of all phylogenetic trees $\mathcal{T}_n$ isometrically embeds into the class of all phylogenetic networks $\mathcal{N}_n$ under the SNPR-distance.

Lemma 3.3 ([BLS17, Proposition 7.7])**.**
Let $N, N' \in \mathcal{N}_n$ such that $\mathrm{d_{SNPR}}(N, N') = k$. Let $T \in D(N)$.
Then there exists a phylogenetic tree $T' \in D(N)$ such that $\mathrm{d_{SNPR}}(T, T') \leq k$.

By setting one of the two networks in the previous lemma to be a phylogenetic tree and noting that the roles of N and N' are interchangeable, the next two corollaries are immediate consequences of Lemmata 3.1 and 3.3.

Corollary 3.4.
Let $T \in \mathcal{T}_n$ and $N \in \mathcal{N}_n$ with $\mathrm{d_{SNPR}}(T, N) = k$.
Then $\mathrm{d_{SNPR}}(T, T') \leq k$ for each $T' \in D(N)$.

Corollary 3.5.
Let $N \in \mathcal{N}_n$ with r reticulations. Let $T, T' \in D(N)$. Then $\mathrm{d_{SNPR}}(T, T') \leq r$.

The main result of this section is the following theorem that characterises the SNPR-distance between a phylogenetic tree and a phylogenetic network.

Theorem 3.6.
Let $T \in \mathcal{T}_n$. Let $N \in \mathcal{N}_n$ with r reticulations. Then

$$\mathrm{d_{SNPR}}(T, N) = \min_{T' \in D(N)} \mathrm{d_{SNPR}}(T, T') + r.$$

Proof. Let $T^* \in D(N)$ such that $\mathrm{d_{SNPR}}(T, T^*) \leq \mathrm{d_{SNPR}}(T, T')$ for each $T' \in D(N)$. Then, by Lemmata 3.1 and 3.2, it follows that

$$\mathrm{d_{SNPR}}(T, N) \leq d_{\mathrm{SNPR}}(T, T^*) + d_{\mathrm{SNPR}}(T^*, N) = \min_{T' \in D(N)} \mathrm{d_{SNPR}}(T, T') + r. \qquad (1)$$

We next show that

$$\mathrm{d_{SNPR}}(T, N) \geq \min_{T' \in D(N)} \mathrm{d_{SNPR}}(T, T') + r.$$

Suppose that $\mathrm{d_{SNPR}}(T, N) = k$. Let $\sigma = (T = N_0, N_1, N_2, \ldots, N_k = N)$ be an SNPR-sequence from T to N. For each $i \in \{1, 2, \ldots, k\}$, consider the two networks N_{i-1} and N_i in σ. If N_i has been obtained from N_{i-1} by applying an SNPR$^+$ operation, then $D(N_{i-1}) \subseteq D(N_i)$. Furthermore, it follows from Lemma 3.3 that, for each tree $T_{i-1} \in D(N_{i-1})$, there exists a tree T_i in $D(N_i)$ such that $\mathrm{d_{SNPR}}(T_{i-1}, T_i) \leq 1$. It is now straightforward to check that we can construct a sequence $S = (T_0, T_1, T_2, \ldots, T_k)$ of phylogenetic trees on X from σ that satisfies the following properties.

(i) For each $i \in \{0, 1, \ldots, k\}$, we have $T_i \in D(N_i)$.

(ii) For each $i \in \{1, 2, \ldots, k\}$, if N_i has been obtained from N_{i-1} by applying an SNPR$^+$ operation, then $T_i = T_{i-1}$.

(iii) For each $i \in \{1, 2, \ldots, k\}$, we have $d_{\mathrm{SNPR}}(T_{i-1}, T_i) \leq 1$.

By construction and since σ contains at least r SNPR$^+$ operations, there exists a subsequence of S of length $k - r$ that is an SNPR-sequence from T_0 to T_k. Hence, we have $\mathrm{d_{SNPR}}(T, T_k) \leq k - r$. Moreover, noting that $T_k \in D(N)$ it follows from Lemma 3.1 that $\mathrm{d_{SNPR}}(T_k, N) = r$ and, thus,

$$\begin{aligned} \min_{T' \in D(N)} \mathrm{d_{SNPR}}(T, T') + r &\leq \mathrm{d_{SNPR}}(T, T_k) + \mathrm{d_{SNPR}}(T_k, N) \\ &= k - r + r = k = \mathrm{d_{SNPR}}(T, N). \end{aligned} \tag{2}$$

Combining Inequalities 1 and 2 establishes the theorem. □

Given Lemma 3.2 and Theorem 3.6 and that $\mathrm{d_{SNPR}}(T, T') = \mathrm{d_{rSPR}}(T, T')$, it is worth noting that the problem of computing the SNPR-distance between a phylogenetic network and a phylogenetic tree can be reduced to computing the rSPR-distance between pairs of trees. Calculating the rSPR-distance between two phylogenetic trees is a well understood problem and several exact algorithms exist (e.g. [BS05, WBZ16]). Furthermore, this problem is known to be fixed-parameter tractable with the rSPR-distance itself as parameter [BS05, Theorem 3.4]. We use this result to establish the following theorem.

Theorem 3.7.
Let $T \in \mathcal{T}_n$ and $N \in \mathcal{N}_n$. Then computing $\mathrm{d_{SNPR}}(T, N)$ is fixed-parameter tractable when parameterised by $\mathrm{d_{SNPR}}(T, N)$.

Proof. Let $d = \mathrm{d_{SNPR}}(T, N)$ and let r be the number of reticulations of N. Note that $|D(N)| \leq 2^r \leq 2^d$, since we know by Theorem 3.6 that $r \leq d$. Next, by Corollary 3.5 we know that $\mathrm{d_{rSPR}}(T, T') = \mathrm{d_{SNPR}}(T, T') \leq d$ for all $T' \in D(N)$. Therefore, computing $\mathrm{d_{rSPR}}(T, T')$ is not only fixed-parameter tractable when parameterised by $\mathrm{d_{rSPR}}(T, T')$

but also when parameterised by d. By Theorem 3.6 computing $\mathrm{d}_{\mathrm{SNPR}}(T, N)$ can be done by computing $\mathrm{d}_{\mathrm{rSPR}}(T, T')$ for each $T' \in D(N)$. Since computing each of these at most 2^d distances is fixed-parameter tractable when parameterised by d, so is computing all of them. This establishes the theorem. □

4 Using agreement forests to characterise the SNPR-distance

We now show how agreement forests can be used to characterise the SNPR-distance between a phylogenetic tree T and a phylogenetic network N. Importantly, this characterisation allows us to compute the SNPR-distance between T and N directly without having to compute the rSPR-distance between T and each tree that is displayed by N as suggested by Theorem 3.6.

Let $T \in \mathcal{T}_n$ and let $N \in \mathcal{N}_n$ with r reticulations. For the purpose of the upcoming definition and much of this section, we regard the root ρ of T and N as an element of $L(T)$ and $L(N)$, respectively. An *agreement forest* F for T and N is a collection $\{T_\rho, T_1, T_2, \ldots, T_k, E_1, E_2, \ldots, E_r\}$, where T_ρ is an isolated vertex labelled ρ, or a phylogenetic tree whose label set includes ρ, each T_i with $i \in \{1, 2, \ldots, k\}$ is a phylogenetic tree, and each E_i with $i \in \{1, 2, \ldots, r\}$ is a single directed edge such that the following properties hold:

(i) The label sets $L(T_\rho), L(T_1), L(T_2), \ldots, L(T_k)$ partition $X \cup \{\rho\}$.

(ii) There exist simultaneous edge-disjoint embeddings of all trees in

$$\{T_\rho, T_1, T_2, \ldots, T_k\}$$

into T that cover all edges of T.

(iii) There exist simultaneous edge-disjoint embeddings of all trees in

$$\{T_\rho, T_1, T_2, \ldots, T_k, E_1, E_2, \ldots, E_r\}$$

into N that cover all edges of N.

We refer to each element in $\{E_1, E_2, \ldots, E_r\}$ as a *disagreement edge*. To illustrate, Figure 2 shows an agreement forest F for a phylogenetic tree and a phylogenetic network.

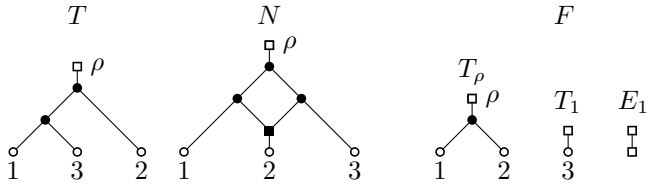

Figure 2: An agreement forest F for the phylogenetic tree T and the phylogenetic network N.

Let $F = \{T_\rho, T_1, T_2, \ldots, T_k, E_1, E_2, \ldots, E_r\}$ be an agreement forest for T and N. Then F is called a *maximum agreement forest* for T and N if the number of elements in the subset $\{T_\rho, T_1, T_2, \ldots, T_k\}$ of F or, equivalently, k is minimised. Moreover, for k being minimum, we set $m(T, N) = k + r$. Referring back to Figure 2, the agreement forest F is a maximum agreement forest for T and N.

For readers familiar with the notion of agreement forests for two phylogenetic trees T and T', we note that the aforementioned definition of a maximum agreement forest coincides with its namesake concept for T and T' as introduced by Bordewich and Semple [BS05]. The importance of the notion of maximum agreement forests for two phylogenetic trees lies in the following theorem.

Theorem 4.1 (Bordewich and Semple [BS05, Theorem 2.1])**.**
Let $T, T' \in \mathcal{T}_n$. Then $\mathrm{d_{rSPR}}(T, T') = m(T, T)$.

Next we show how the more general definition of agreement forests that is introduced in this paper can be employed to characterise the SNPR-distance between a phylogenetic tree T and a phylogenetic network N. We start with a 'warm-up' for when T is displayed by N.

Lemma 4.2.
Let $N \in \mathcal{N}_n$ with r reticulations. Let $T \in D(N)$.
Then $\mathrm{d_{SNPR}}(T, N) = m(T, N) = r$.

Proof. By Lemma 3.1, we have $\mathrm{d_{SNPR}}(T, N) = r$. Let $F = \{T, E_1, \ldots, E_r\}$. As T can be transformed into N by applying r SNPR$^+$, it follows that F can be embedded into N such that all edges in N are covered. It is now easily checked that F is an agreement forest for T and N. Hence,

$$r = \mathrm{d_{SNPR}}(T, N) = |F| - 1 \geq m(T, N).$$

To establish the other direction, let F be a maximum agreement forest for N and T. Recall that, by definition, F contains r disagreement edges and at least one other element. Thus,

$$m(T, N) = |F| - 1 \geq r + 1 - 1 = r = \mathrm{d_{SNPR}}(T, N).$$

This completes the proof of the lemma. □

We are now in a position to establish the main result of this section.

Theorem 4.3.
Let $T \in \mathcal{T}_n, N \in \mathcal{N}_n$. Then $\mathrm{d_{SNPR}}(T, N) = m(T, N)$.

Proof. Let r be the number of reticulations in N. We first show that $m(T, N) \leq \mathrm{d_{SNPR}}(T, N)$. By Theorem 3.6, there exists a phylogenetic tree T' that is displayed by N such that

$$\mathrm{d_{SNPR}}(T, N) = \mathrm{d_{SNPR}}(T, T') + \mathrm{d_{SNPR}}(T', N) = \mathrm{d_{SNPR}}(T, T') + r.$$

Hence, we have $m(T, T') = \mathrm{d_{SNPR}}(T, T') = \mathrm{d_{SNPR}}(T, N) - r$, where the first equality follows from Lemma 3.2 and Theorem 4.1. Moreover, by Lemma 4.2, we have $m(T', N) = \mathrm{d_{SNPR}}(T', N) = r$. Let F' be a maximum agreement forest for T and T', and let F'' be a maximum agreement forest for T' and N. Since F'' contains r disagreement edges, it follows that $T' \in F''$. Now, let

$$F = F' \cup (F'' - \{T'\}).$$

Recall that F' has an embedding into T' that covers all edges. Since T' is displayed by N, it follows that F' has an embedding into N such that all edges that are uncovered by this embedding can be partitioned into at most r directed paths each ending at a reticulation in N. Now, as F' and F'' are agreement forests for T and T', and T' and N, respectively, it is straightforward to check that F is an agreement forest for T and N. Hence,

$$m(T, N) \leq |F| - 1 = |F'| + |F''| - 2 = \mathrm{d_{SNPR}}(T, T') + \mathrm{d_{SNPR}}(T', N) = \mathrm{d_{SNPR}}(T, N).$$

We next show that $\mathrm{d_{SNPR}}(T, N) \leq m(T, N)$. Let

$$F = \{T_\rho, T_1, T_2, \ldots, T_k, E_1, , E_2, \ldots, E_r\}$$

be a maximum agreement forest for T and N. The proof is by induction on $|F|$. If $|F| = 1$, that is $F = \{T\}$, then $N = T$ and, so, $\mathrm{d_{SNPR}}(T, N) = 0$. Now assume that the inequality holds for all pairs of a phylogenetic tree and a phylogenetic network on the same leaf set for which there exists a maximum agreement forest whose number of components is at most $k + r$. If $r = 0$, then the induction step is the same as in the proof of of Theorem 2.1 by Bordewich and Semple [BS05]. We may therefore assume that $r > 0$. Let v be a reticulation in N that has no distinct reticulation as an ancestor. For each component $C_i \in F$, let $\epsilon(C_i)$ be the set of edges in N that is used to embed C_i into N such that

$$\mathcal{E} = \{\epsilon(T_\rho), \epsilon(T_1), \epsilon(T_2), \ldots, \epsilon(T_k), \epsilon(E_1), \epsilon(E_2), \ldots, \epsilon(E_r)\}$$

is a partition of the edge set of N. Since F is an agreement forest for T and N such a partition exists.

Now, let (u, v) and (u', v) be the reticulation edges incident with v. Without loss of generality, we may assume that $(u, v) \in \epsilon(E_i)$ for some $i \in \{1, 2, \ldots, r\}$. Let $F' = F - \{E_i\}$, and let N' be the phylogenetic network obtained from N be deleting (u, v) and suppressing the resulting two degree-2 vertices. We next show that F' is an agreement forest for T and N'. By the choice of v, recall that u is a tree vertex. Let w be the child of u such that $v \neq w$, and let C_j be the component in F such that $(u, w) \in \epsilon(C_j)$. Set $\epsilon'(C_j) = \epsilon(C_j) \cup (\epsilon(E_i) - \{(u, v)\})$. Note that $\epsilon'(C_j) = \epsilon(C_j)$ precisely if $\epsilon(E_i) = \{(u, v)\}$. As $\mathcal{E}$ is an embedding of F into N that partitions the edge set of N,

$$\mathcal{E}' = (\mathcal{E} - \{\epsilon(C_j), \epsilon(E_i)\}) \cup \{\epsilon'(C_j)\}$$

partitions the edge set of N' and induces an embedding of F' in N'. Hence, F' is an agreement forest for N' and T'. Since $|F'| < |F|$, it now follows from the induction hypothesis that there exists an SNPR-sequence from T to N' whose length is at most $|F'| = k + r - 1$. Furthermore, by construction, N can be obtained from N' by a single SNPR$^+$. Taken together, this implies that

$$\mathrm{d}_{\mathrm{SNPR}}(T, N) \leq \mathrm{d}_{\mathrm{SNPR}}(T, N') + 1 \leq k + r - 1 + 1 = m(T, N).$$

Combining both inequalities establishes the theorem. □

5 Properties of shortest SNPR-sequences connecting two networks

In this section, we analyse properties of shortest SNPR-sequences that connect a pair of phylogenetic networks and investigate whether or not the three classes of tree-child, reticulation-visible, and tree-based networks isometrically embed into the class of all phylogenetic networks. We start with some definitions that are used throughout this section. For any non-negative integer r, *tier* r of $\mathcal{N}_n$ is the subset of networks in $\mathcal{N}_n$ that have exactly r reticulations. Note that tier 0 equals $\mathcal{T}_n$. For $N, N' \in \mathcal{N}_n$, let $\sigma = (N = N_0, N_1, \ldots, N_k = N')$ be an SNPR-sequence from N to N'. We say that σ *horizontally traverses* tier r if σ contains two networks N_{i-1} and N_i with $i \in \{1, 2, \ldots, k\}$ such that both have r reticulations; i.e., N_i can be obtained from N_{i-1} by a single SNPR^0.

Let $N, N' \in \mathcal{N}_n$ with r and r' reticulations, respectively. Without loss of generality, we may assume that $r \leq r'$. From a computational viewpoint and in trying to shrink the search space when computing $\mathrm{d}_{\mathrm{SNPR}}(N, N')$, it would be favorably if there always exists a shortest SNPR-sequence connecting N and N' that traverses exactly one tier horizontally. In particular, if $r < r'$ it would have positive implications for computing $\mathrm{d}_{\mathrm{SNPR}}(N, N')$ if all SNPR^0 operations could be pushed to be the beginning or the end of a shortest SNPR-sequence for N and N'. On the other hand, if $r = r'$, then the existence of a shortest SNPR-sequence from N to N' whose networks all belong to tier r would allow us to compute $\mathrm{d}_{\mathrm{SNPR}}(N, N')$ by considering only tier r. In what follows, we present several results showing that the existence of a shortest SNPR-sequence with such properties cannot be guaranteed.

Lemma 5.1.
Let $n \geq 4$. Let $N, N' \in \mathcal{N}_n$ with r and r' reticulations, respectively, such that $r < r'$. Then there does not necessarily exist a shortest SNPR*-sequence from N to N' that traverses at most one tier horizontally.*

Proof. To prove the statement, we show that every shortest SNPR-sequence for the two phylogenetic networks N and N' that are depicted in Figure 3 traverses at least two tiers horizontally.

We start by observing four differences between N and N':

(1) Leaf 1 is a descendant of a reticulation in N, but not in N'.

(2) Leaves 1 and 4 are siblings in N', but not in N.

(3) Leaves 2 and 3 are siblings in N', but not in N.

(4) Leaves 2 and 3 are descendants of two reticulations in N', but not in N.

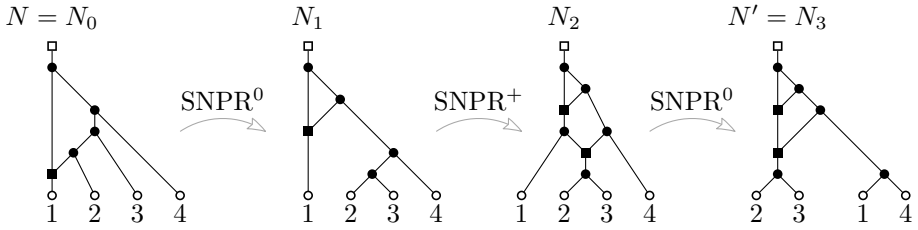

Figure 3: Example that is used in the proof Lemma 5.1 to show that every shortest SNPR-sequence from N to N' traverses two tiers horizontally.

We first note that $\mathrm{d}_{\mathrm{SNPR}}(N, N') > 2$, since there is no phylogenetic network M such that $\mathrm{d}_{\mathrm{SNPR}}(N, M) = 1$ and $\mathrm{d}_{\mathrm{SNPR}}(M, N') = 1$. Consequently, referring back to the networks shown in Figure 3,

$$\sigma = (N = N_0, N_1, N_2, N_3 = N')$$

is a shortest SNPR-sequence from N to N' that horizontally traverses tier 1 and tier 2. Since N' has one reticulation more than N, note that at least one SNPR$^+$ is required to transform N into N'.

To establish the statement, it is therefore sufficient to show that there exists no SNPR-sequence, say

$$\sigma^* = (N, M, M', N'),$$

such that M can be obtained from N by an SNPR$^+$, or N' can be obtained from M' by an SNPR$^+$. We proceed by distinguishing these two cases.

First, assume that σ^* exists and that M has been obtained from N by an SNPR$^+$. Then M and N' have the same four differences as listed above for N and N' with the exception that either leaf 2 or 3 (but not both) is possibly a descendant of two reticulations in M. Suppose that M is indeed obtained from N by (i) subdividing the edge directed into 1 with a new vertex u, subdividing the edge directed into 2 with a new vertex v, and adding the new edge (u, v), or (ii) subdividing the edge directed into 1 with a new vertex u, subdividing the edge directed into 3 with a new vertex v, and adding the new edge (u, v). In both cases, it is easy to check that it requires two SNPR operations to transform M into a network, say M^*, in which leaf 1 is not a descendant of any reticulation and leaves 2 and 3 are descendants of two reticulations. However, $M^* \neq N'$ and, so, it takes at least three SNPR operations to transform M into N'. Now, suppose that M is obtained from N by an SNPR$^+$ other than (i) or (ii). Then, it is easy to see that it takes again at least three SNPR operations to transform M into N'. Hence, we conclude that M has not been obtained from N by an SNPR$^+$.

Second, assume that σ^* exists and that N' has been obtained from M' by an SNPR$^+$ or, equivalently, M' has been obtained from N' by an SNPR$^-$. Clearly, M' is uniquely defined because each of the three SNPR$^-$ operations that can be applied to N' results in the same network M'. Now, again using the aforementioned differences between N and N' that are also differences between N and M' with the exception that leaves 2 and 3 are descendants of only a single reticulation in M', it takes at least three SNPR operations to transform N into M'. Consequently, N' has not been obtained from M' in σ^* by an SNPR$^+$.

Lastly, since neither M nor N' has been obtained from N and M', respectively, by an SNPR$^+$, it follows that σ^* cannot be chosen so that no tier is horizontally traversed. This completes the proof. □

We next shows that, for two phylogenetic networks N and N' that both have r reticulations, every shortest SNPR-sequence from N to N' may contain a phylogenetic tree. Hence, to compute $\mathrm{d}_{\mathrm{SNPR}}(N, N')$ it is necessary to search all phylogenetic networks with at most r reticulations.

Lemma 5.2.
Let $r \geq 2$ and $n \geq 2r + 2$. There exist $\bar{N}_r, \bar{N}'_r \in \mathcal{N}_n$ with r reticulations such that every shortest SNPR*-sequence from $\bar{N}_r$ to $\bar{N}'_r$ contains a phylogenetic tree.*

Proof. To prove the statement, we show that every shortest SNPR-sequence

$$\sigma = (\bar{N}_r = N_0, N_1, \ldots, N_k = \bar{N}'_r)$$

connecting the two phylogenetic networks $\bar{N}_r$ and $\bar{N}'_r$ depicted in Figure 4 has length $2k$, for each $i \in \{1, 2, \ldots, r\}$, N_i is obtained from N_{i-1} by an SNPR$^-$ and for each $i \in \{r + 1, r + 2, \ldots, 2r\}$, N_i is obtained from N_{i-1} by an SNPR$^+$. Since $\bar{N}_r$ and $\bar{N}'_r$ both have r reticulations, this implies that σ contains a phylogenetic tree. Note that σ exists because we can transform $\bar{N}_r$ into $\bar{N}'_r$ by removing each reticulation edge in $\{e_1, e_2, \ldots, e_r\}$ with an SNPR$^-$ and then adding each edge $\{e'_1, e'_2, \ldots, e'_r\}$ with an SNPR$^+$.

We pause to observe three properties of $\bar{N}'_r$ that will be crucial for the remainder of this proof:

(P1) For each $i \in \{1, 2, \ldots, r\}$, the leaf l_i is a sibling of a reticulation.

(P2) Leaves 1 and 2 are siblings and descendants of all reticulations.

(P3) There exists a directed path $(\rho, v_1, v_2, \ldots, v_r)$, where ρ is the root, and each v_i with $i \in \{1, 2, \ldots, r\}$ is a reticulation.

To illustrate, for $r = 2$, the networks $\bar{N}_2$ and $\bar{N}'_2$ are shown in Figure 4.

Now assume that there exists an SNPR-sequence

$$\sigma^* = (\bar{N}_r = M_0, M_1, M_2, \ldots, M_{k'} = \bar{N}'_r)$$

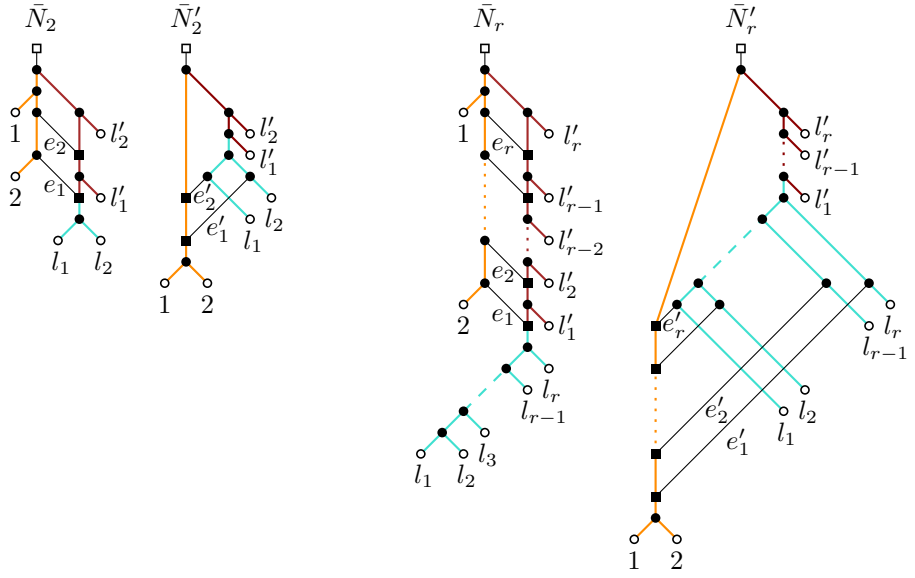

Figure 4: Example that is used in the proof of Lemma 5.2 to show that, for each $r \geq 2$, there exist two phylogenetic networks $\bar{N}_r$ and $\bar{N}'_r$ such that every shortest SNPR-sequence from $\bar{N}_r$ to $\bar{N}'_r$ contains a phylogenetic tree.

from $\bar{N}_r$ to $\bar{N}'_r$ of length $k' \leq 2r$ that is distinct from σ. Let

$$O^* = (o_1, o_2, \ldots, o_{k'})$$

be the sequence obtained from σ^* such that for each $i \in \{1, 2, \ldots, k'\}$ the following holds:

- $o_i = 0$ if M_i is obtained from M_{i-1} by an SNPR^0,
- $o_i = +$ if M_i is obtained from M_{i-1} by an SNPR$^+$, or
- $o_i = -$ if M_i is obtained from M_{i-1} by an SNPR$^-$.

Let m be the number of elements in O^* that are equal to $-$.
Case 1. Assume that $m > r$. Since $\bar{N}_r$ and $\bar{N}'_r$ both have r reticulations, O^* contains exactly m elements that are equal to $+$. Hence, $k' \geq 2m > 2r$; a contradiction.
Case 2. Assume that $m < r$. Again, since $\bar{N}_r$ and $\bar{N}'_r$ both have r reticulations, O^* contains exactly m elements that are equal to $+$. Thus, with $k' \leq 2r$, it follows that O^* contains at most $2(r - m)$ elements that are equal to 0. Let i be an element in $\{1, 2, \ldots, k'\}$ such that $o_i = +$. Then, the number of leaves in $\{l_1, l_2, \ldots, l_r\}$ that are siblings of reticulations in M_{i-1} and M_i differs by at most one. Therefore, we need at least $k_1 \geq r - m$ SNPR^0 operations to obtain a network from $\bar{N}_r$ that satisfies (P1). Similarly, the number of vertices on a directed path that consists only of reticulations in $\mathcal{M}_{i-1}$ and $\mathcal{M}_i$ differs by at most one. Therefore, we need at least $k_2 \geq r - m$ SNPR^0 operations to obtain a network from $\bar{N}_r$ that satisfies (P3). Let $i \in \{1, 2, \ldots, k'\}$ such that $o_i = 0$. If the number of leaves in $\{l_1, l_2, \ldots, l_r\}$ that are siblings of reticulations

in M_i is greater than that number in M_{i-1}, then the number of vertices that lie on a directed path of reticulations in M_i is not greater than that number in M_{i-1}. Similarly, if the number of vertices that lie on a directed path of reticulations in M_i is greater than that number in M_{i-1}, then the number of leaves in $\{l_1, l_2, \ldots, l_r\}$ that are siblings of reticulations in M_i is not greater than that number in M_{i-1}. This is the case since the k_1 SNPR^0 used to satisfy property (P1) affect the leaves l_j and reticulation edges, whereas the k_2 SNPR^0 used to satisfy property (P3) affect the leaves l'_j and (possibly) leaf 1. It follows that $k_1 = k_2 = (r - m)$ and, so, $k' = 2r$. To see that $M_{k'}$ does not satisfy property (P2), observe that neither the $k_1 + k_2$ SNPR^0 operations nor the $2m$ SNPR$^-$ and SNPR$^+$ operations that are used to satisfy (P1) and (P3) result in a network that simultaneously satisfies (P2). Hence, it follows that at least one additional SNPR^0 is needed to transform $\bar{N}_r$ into $\bar{N}'_r$; thereby contradicting that $k' \leq 2r$.

Case 3. Assume that $m = r$. Since $\bar{N}_r$ and $\bar{N}'_r$ both have r reticulations and $k' \leq 2r$, it follows that $k' = 2r$. We complete the proof by showing that, for each $i \in \{1, 2, \ldots, r\}$, we have $o_i = -$ and, for each $i \in \{r+1, r+2, \ldots, 2r\}$, we have $o_i = +$. Assume that, for some $i \leq r$, we have $o_i = +$. Choose i to be as small as possible. Let v be the unique reticulation in M_i that is not a reticulation in M_{i-1}. Then v does not have leaves 1 and 2 as descendants and a leaf in $\{l_1, l_2, \ldots, l_r\}$ as a sibling of a reticulation. Now, as O^* does not contain an element equal to 0, there exists an element $o_j = -$ with $j > i$ such that M_j does not contain the reticulation edge that was added in transforming M_{i-1} into M_i. In turn, this implies that the remaining $r - 1$ SNPR$^+$ cannot transform $\bar{N}_r$ into a network that satisfies (P1) and (P3). Hence, if $m = r$, then $\sigma^* = \sigma$.

Combining all three cases establishes the statement. □

Recall that the statement of Lemma 5.2 requires $\bar{N}_r$ and $\bar{N}'_r$ to have at least two reticulations. Nevertheless, using a slightly different construction than that for $\bar{N}_r$ and $\bar{N}'_r$, Figure 5 shows two phylogenetic networks that both have one reticulation such that every shortest SNPR-sequence connecting these two networks contains a phylogenetic tree. While omitting a formal proof, we note that it can be checked by following the same ideas as in the proof of Lemma 5.2.

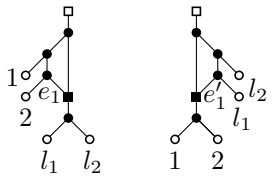

Figure 5: Two phylogenetic networks with one reticulation such that every shortest SNPR-sequence connecting them contains a phylogenetic tree.

Bordewich et al. [BLS17, Proposition 7.5] showed that

$$\mathrm{d}_{\mathrm{SNPR}}(N, N') \leq \min\{\mathrm{d}_{\mathrm{SNPR}}(T, T') \colon T \in D(N) \text{ and } T' \in D(N')\} + r + r',$$

where $N, N' \in \mathcal{N}_n$ with r and r' reticulations, respectively. Lemma 5.2 implies that this upper bound is sharp.

The next lemma shows that, for two phylogenetic networks N and N' that both have r reticulations, every shortest SNPR-sequence from N to N' may contain a network that has more than r reticulations. In particular, to compute $\mathrm{d}_{\mathrm{SNPR}}(N, N')$ it is not sufficient to only search the space of all phylogenetic networks that have at most r reticulations.

Lemma 5.3.
Let $n \geq 2$, $r \geq 3$, and let $N, N' \in \mathcal{N}_n$ with r reticulations.
There does not necessarily exist a shortest SNPR*-sequence from N to N' such that each network in the sequence has at most r reticulations.*

Proof. To establish the lemma, we show that every shortest SNPR-sequence that connects the two phylogenetic networks N and N' as depicted in Figure 6 contains a network with four reticulations. First observe that $\mathrm{d}_{\mathrm{SNPR}}(N, N') \geq 2$ and, so, the SNPR-sequence (N, N_1, N') is of minimum length.

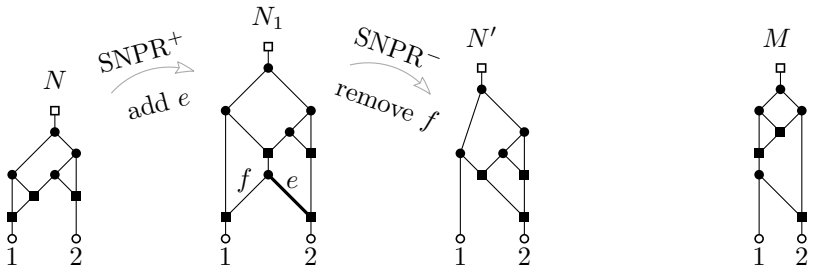

Figure 6: Example that is used in the proof of Lemma 5.3 to show that every shortest SNPR-sequence from N to N' that both have three reticulations contains a network with four reticulations.

We complete the proof by showing that there exists no SNPR-sequence (N, M, N') such that M is obtained from N by an SNPR^- or SNPR^0. Towards a contradiction, assume that M is obtained from N by an SNPR^-. Clearly, leaf 1 is a child of a reticulation in M. Moreover, as M has two reticulations, it follows that N' is obtained from M by an SNPR^+ and that leaf 1 is still a child of a reticulation in N'; a contradiction. Now assume that M is obtained from N by an SNPR^0. If leaf 1 is a child of a reticulation in M, then $\mathrm{d}_{\mathrm{SNPR}}(M, N') > 1$. We may therefore assume that leaf 1 is not a child of a reticulation in M. Hence, M is the network that is shown on the right-hand side of Figure 6 in which all three reticulations lie on a directed path. It now follows that $\mathrm{d}_{\mathrm{SNPR}}(M, N') > 1$ because it requires at least two SNPR operations to transform M into a network in which not all three reticulations are on a directed path and where leaf 1 is not a descendant of any reticulation; again a contradiction. □

The next theorem is an immediate consequence of Lemmata 5.2 and 5.3 and Figure 5.

Theorem 5.4.
Let $\mathcal{C}_r$ be the class of all phylogenetic networks in $\mathcal{N}_n$ that have r reticulations.

If $n \geq 4$ and $r \geq 1$, then $\mathcal{C}_r$ does not isometrically embed into the class of all phylogenetic networks $\mathcal{N}_n$. Moreover, if $n \geq 2$ and $r \geq 3$, then $\mathcal{C}_r$ does not isometrically embed into the class of all phylogenetic networks in $\mathcal{N}_n$ with at most r reticulations.

We now consider different classes of phylogenetic networks and ask if they isometrically embed into the class of all phylogenetic networks. As we will see, we answer this question negatively for tree-child networks $\mathcal{TC}_n$, reticulation-visible networks $\mathcal{RV}_n$, and tree-based networks $\mathcal{TB}_n$.

Proposition 5.5.
Let $\mathcal{C}_n \in \{\mathcal{TC}_n, \mathcal{RV}_n, \mathcal{TB}_n\}$ with $n \geq 4$.
Then $\mathcal{C}_n$ does not embed isometrically into $\mathcal{N}_n$ under SNPR.

Proof. To establish the theorem, we give explicit examples of two networks N and N' that are in $\mathcal{C}_n$ such that $d_{\mathrm{SNPR}_{\mathcal{C}_n}}(N, N') > d_{\mathrm{SNPR}_{\mathcal{N}_n}}(N, N')$.

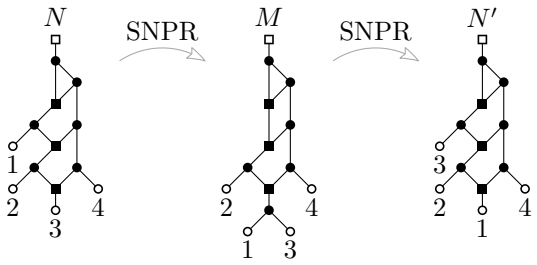

Figure 7: Example that is used in the proof of Proposition 5.5 to show that neither $\mathcal{TC}_n$ nor $\mathcal{RV}_n$ embeds isometrically into $\mathcal{N}_n$.

Let $\mathcal{C}_n = \mathcal{TC}_n$ (resp. $\mathcal{C}_n = \mathcal{TB}_n$). Consider the two tree-child (resp. tree-based) networks N and N' that are shown in Figure 7 (resp. Figure 8). Then $\sigma = (N, M, N')$ is an SNPR-sequence for N and N'. Indeed, as N' can be obtained from N by interchanging the roles of leaves 1 and 3, and the parent of 3 is a reticulation in N, it is straightforward to check that σ is the unique SNPR-sequence in $\mathcal{N}_n$ of length two that connects N and N'. Hence, as M is not tree child (resp. tree based), we have

$$d_{\mathrm{SNPR}_{\mathcal{C}_n}}(N, N') > d_{\mathrm{SNPR}_{\mathcal{N}_n}}(N, N') = 2.$$

Noting that M in Proposition 5.5 is not reticulation visible, the same argument holds for when $\mathcal{C}_n = \mathcal{RV}_n$.

□

While Francis and Steel [FS15] allow tree-based networks to have edges in parallel, it is easily seen that one can add a single edge to each of the three networks depicted in Figure 8 to obtain an example showing that the class of all tree-based networks without parallel edges on n leaves is not isometrically embedded into the class of all phylogenetic networks on n leaves either.

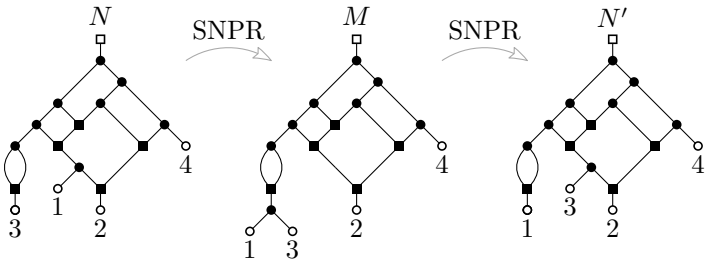

Figure 8: Example that is used in the proof of Proposition 5.5 to show that $\mathcal{TB}_n$ does not embed isometrically into $\mathcal{N}_n$.

Lastly, the networks presented in this section may seem rather small. However, they can be regarded as skeletons of larger networks with the same properties. For instance, in all examples that we used to establish the results of this section, leaves can be replaced with subtrees and subnetworks. Furthermore, some edges can be subdivided to add further reticulation edges or subtrees to obtain larger networks with the same properties.

6 Concluding remarks

In this paper, we have established the first results related to calculating the SNPR-distance, which is an NP-hard problem. In the first part, we have considered the special case of computing this distance between a phylogenetic tree T and a phylogenetic network N. In this particular case, computing the SNPR-distance is fixed-parameter tractable when parameterised by this distance and can be calculated by solving several instances of the rSPR-distance problem. Additionally, we have characterised the SNPR-distance of T and N in terms of agreement forests. This result lends itself to an algorithm that works directly on T and N without having to solve multiple instances of the rSPR-distance problem between two trees. In the second part, we have turned to the SNPR-distance problem between two phylogenetic networks N and N' and presented several results on shortest SNPR-sequences for N and N' with r and r' reticulations, respectively. These results show that the search space for computing the SNPR-distance of N and N' can in general not be pruned to networks whose number of reticulations is at least $\min\{r, r'\}$ or at most $\max\{r, r'\}$. Furthermore, if N and N' are both tree child, reticulation visible, or tree based, the search space can in general not be restricted to these network classes.

As alluded to in the introduction, Gambette et al. [GvIJ$^+$17] have introduced a slightly different operation that generalises rSPR to phylogenetic networks. The main difference between their operation and SNPR is that they allow for an additional operation which is called a head move. In the language of this paper, let N be a phylogenetic network, and let (u, v) be an edge of N such that v is a reticulation. Then, the operation of deleting (u, v), suppressing u, subdividing an edge that is not an ancestor of v with a new vertex u', and adding the edge (v, u') is referred to as a *head move*. Interestingly, if we generalise the SNPR operation by, additionally, allowing for head moves, the properties of shortest SNPR-sequences that we have revealed in Section 5 and that may appear

to be undesirable with regards to practical search algorithm do not change. On the positive side, a characterisation of the SNPR-distance between a phylogenetic tree and a phylogenetic network in terms of agreement forest is possible and a result equivalent to Theorem 4.3 can be established. For further details, we refer the interested reader to the first author's PhD thesis [Kla] which establishes results equivalent to the ones presented in this paper for when one allows for head moves.

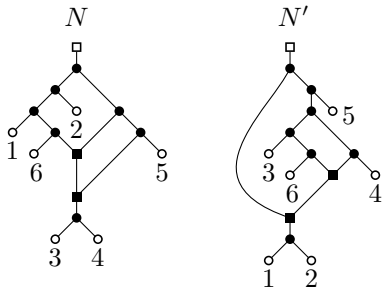

Figure 9: Two phylogenetic networks N and N' for which every shortest SNPR-sequence prunes at least one edge twice.

We close this paper with the question if the notion of agreement forests can be further generalised to computing the SNPR-distance between two phylogenetic networks, regardless of whether head moves are also allowed. As mentioned above, a shortest sequence between N and N' might have to traverse a tier with more or less reticulations than N and N'. It is unclear how an agreement forest could capture edges that first get added and then removed again (or vice versa), as this seems to be beyond embeddings of an agreement forest into N and N', respectively. Furthermore, Figure 9 shows two networks for which every shortest SNPR-sequences prunes at least one edge twice. A similar problem exists for the subtree prune and regraft operation on unrooted phylogenetic trees for which a characterisation in terms of agreement forests appears to be problematic as well [WM18].

Acknowledgments. We thank the New Zealand Marsden Fund for their financial support.

References

[AS01] B. L. Allen and M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees," *Annals of Combinatorics*, vol. 5, no. 1, pp. 1–15, 2001. [Online]. Available: https://dx.doi.org/10.1007/s00026-001-8006-8

[BLS17] M. Bordewich, S. Linz, and C. Semple, "Lost in space? Generalising subtree prune and regraft to spaces of phylogenetic networks," *Journal*

of Theoretical Biology, vol. 423, pp. 1 – 12, 2017. [Online]. Available: https://dx.doi.org/10.1016/j.jtbi.2017.03.032

[BS05] M. Bordewich and C. Semple, "On the computational complexity of the rooted subtree prune and regraft distance," *Annals of Combinatorics*, vol. 8, no. 4, pp. 409–423, Jan 2005. [Online]. Available: https://dx.doi.org/10.1007/s00026-004-0229-z

[CFS15] J. Chen, J.-H. Fan, and S.-H. Sze, "Parameterized and approximation algorithms for maximum agreement forest in multifurcating trees," *Theoretical Computer Science*, vol. 562, pp. 496–512, 2015. [Online]. Available: https://dx.doi.org/10.1016/j.tcs.2014.10.031

[CRV09] G. Cardona, F. Rossello, and G. Valiente, "Comparison of tree-child phylogenetic networks," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 6, no. 4, pp. 552–569, Oct. 2009. [Online]. Available: https://dx.doi.org/10.1109/TCBB.2007.70270

[Fel04] J. Felsenstein, *Inferring phylogenies*. Sinauer Associates, 2004, vol. 2. [Online]. Available: http://www.sinauer.com/media/wysiwyg/tocs/InferringPhylogenies.pdf

[FHMW18] A. Francis, K. T. Huber, V. Moulton, and T. Wu, "Bounds for phylogenetic network space metrics," *Journal of Mathematical Biology*, vol. 76, no. 5, pp. 1229–1248, Apr 2018. [Online]. Available: https://dx.doi.org/10.1007/s00285-017-1171-0

[FS15] A. R. Francis and M. Steel, "Which phylogenetic networks are merely trees with additional arcs?" *Systematic Biology*, vol. 64, no. 5, pp. 768–777, 2015. [Online]. Available: https://dx.doi.org/10.1093/sysbio/syv037

[GGL+15] P. Gambette, A. D. Gunawan, A. Labarre, S. Vialette, and L. Zhang, "Locating a tree in a phylogenetic network in quadratic time," in *RECOMB*, 2015, pp. 96–107. [Online]. Available: https://dx.doi.org/10.1007/978-3-319-16706-0_12

[Gus14] D. Gusfield, *ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT Press, 2014. [Online]. Available: https://mitpress.mit.edu/books/recombinatorics

[GvIJ+17] P. Gambette, L. van Iersel, M. Jones, M. Lafond, F. Pardi, and C. Scornavacca, "Rearrangement moves on rooted phylogenetic networks," *PLOS Computational Biology*, vol. 13, no. 8, pp. 1–21, 08 2017. [Online]. Available: https://dx.doi.org/10.1371/journal.pcbi.1005611

[HLMW16] K. T. Huber, S. Linz, V. Moulton, and T. Wu, "Spaces of phylogenetic networks from generalized nearest-neighbor interchange operations,"

Journal of Mathematical Biology, vol. 72, no. 3, pp. 699–725, 2016. [Online]. Available: https://dx.doi.org/10.1007/s00285-015-0899-7

[HMW16] K. T. Huber, V. Moulton, and T. Wu, "Transforming phylogenetic networks: Moving beyond tree space," *Journal of Theoretical Biology*, vol. 404, pp. 30 – 39, 2016. [Online]. Available: https://dx.doi.org/10.1016/j.jtbi.2016.05.030

[HRS10] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[JJE+17] R. Janssen, M. Jones, P. L. Erdős, L. van Iersel, and C. Scornavacca, "Exploring the tiers of rooted phylogenetic network space using tail moves," *arXiv preprint arXiv:1708.07656*, 2017. [Online]. Available: https://arxiv.org/abs/1708.07656

[Kla] J. Klawitter, "Spaces of phylogenetic networks," PhD thesis, University of Auckland, in preparation.

[Kla17] ——, "The SNPR neighbourhood of tree-child networks," *arXiv preprint arXiv:1707.09579*, 2017. [Online]. Available: https://arxiv.org/abs/1707.09579

[SOW96] D. L. Swofford, G. J. Olsen, and P. J. Waddell, "Phylogenetic inference," in *Molecular Systematics*, D. M. Hillis, C. Moritz, and B. K. Mable, Eds. Sinauer Associates, 1996, ch. 11, pp. 407–514.

[WBZ16] C. Whidden, R. G. Beiko, and N. Zeh, "Fixed-parameter and approximation algorithms for maximum agreement forests of multifurcating trees," *Algorithmica*, vol. 74, no. 3, pp. 1019–1054, Mar 2016. [Online]. Available: https://dx.doi.org/10.1007/s00453-015-9983-z

[WM18] C. Whidden and F. A. Matsen IV, "Calculating the unrooted subtree prune-and-regraft distance," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018. [Online]. Available: https://dx.doi.org/10.1109/TCBB.2018.2802911

[Wu09] Y. Wu, "A practical method for exact computation of subtree prune and regraft distance," *Bioinformatics*, vol. 25, no. 2, pp. 190–196, 2009. [Online]. Available: https://dx.doi.org/10.1093/bioinformatics/btn606