

# Order-Dependent Dissimilarity Measures on Phylogenetic Trees

Simone Linz<sup>1</sup>, Katherine St. John<sup>2,3</sup>, Charles Semple<sup>4</sup>, and Kristina Wicke<sup>5</sup>

<sup>1</sup>School of Computer Science, University of Auckland, New Zealand

<sup>2</sup>Department of Computer Science, Hunter College, City University of New York, USA

<sup>3</sup>Division of Invertebrate Zoology, American Museum of Natural History, USA

<sup>4</sup>School of Mathematics and Statistics, University of Canterbury, New Zealand

<sup>5</sup>Department of Mathematical Sciences, New Jersey Institute of Technology, USA

July 16, 2025

## Abstract

Ordered leaf attachment, Phylo2Vec, and HOP are three recently introduced vector representations for rooted phylogenetic trees where the representation is determined by an ordering of the underlying leaf set  $X$ . Comparing the vectors of two rooted phylogenetic  $X$ -trees  $T$  and  $T'$  for a fixed ordering on  $X$  leads to polynomial-time computable measure for the dissimilarity of  $T$  and  $T'$ , albeit dependent on the choice of the leaf ordering. For each of ordered leaf attachment, Phylo2Vec, and HOP, we compare this measure with the rooted subtree prune and regraft distance (rSPR), the hybrid number, and the temporal tree-child hybrid number of  $T$  and  $T'$ . Although there is no direct relationship between rSPR and any of the three vector-based measures, we show that, when minimized over all orderings, the hybrid number is equivalent to HOP, and an upper bound on the other two. Moreover, when minimized over all orderings induced by common cherry-picking sequences of  $T$  and  $T'$ , the temporal tree-child hybrid number of  $T$  and  $T'$  is equivalent to each of the three vector-based measures.

*Keywords:* phylogenetic trees, vector representations, order-dependent measures.

*MSC:* 05C05 (Combinatorics: Trees), 92C42 (Systems Biology, Networks).

## 1 Introduction

The task of quantifying the disagreement between phylogenetic trees is essential for evaluating the accuracy of tree inference methods and for comparing phylogenetic trees [? ? ]. One of the most common ways of doing this for two rooted binary phylogenetic trees is by using the tree rearrangement operation of rooted subtree prune and regraft (rSPR) [? ]. The rSPR distance between two arbitrary rooted binary phylogenetic  $X$ -trees  $T$  and  $T'$  is the minimum number of rSPR operations that transforms  $T$  into  $T'$ . However, in general, computing the rSPR distance between  $T$  and  $T'$  is computationally hard [? ]. To circumvent this computational hardness but to also replicate the applicability of rSPR, three new approaches have been recently introduced for this task. All three approaches, namely, ordered leaf attachment (OLA) [? ], Phylo2Vec (P2V) [? ], and HOP [? ], are based on imposing an external ordering  $\sigma$  on  $X$  and, depending on  $T$  and  $\sigma$ , using this ordering to assign a vector to  $T$ . For a particular measure, the distance between  $T$  and  $T'$  under  $\sigma$  is made by a comparison of the vectors assigned to  $T$  and  $T'$ . Regardless of the measure, the assignment of a vector to  $T$  and the comparison of the vectors assigned to  $T$  and  $T'$  can be computed in polynomial time, thereby eliminating the hardness of calculating the rSPR distance.

The purpose of this paper is to investigate the relationship between each of the three order-dependent measures (OLA, P2V, and HOP) and the tree rearrangement operation rSPR that they seek to replicate. If

we are allowed to choose an ordering on  $X$ , then there is little relationship between any of the three measures and rSPR. However, if we choose an ordering that minimizes the measure and broaden the investigation to include the hybridization number [? ], a notion closely related to rSPR, then more direct relationships exist for OLA and HOP. In particular, across all orderings of  $X$ , we show that the minimum OLA measure of two rooted binary phylogenetic  $X$ -trees  $T$  and  $T'$  is bounded above by a linear function in the rSPR distance between  $T$  and  $T'$ , while the minimum HOP measure of  $T$  and  $T'$  equates (exactly) to the hybrid number of  $T$  and  $T'$ . The latter resolves a conjecture in [? ] that computing the minimum HOP measure between two rooted binary phylogenetic trees is NP-hard as computing the hybrid number of two such trees is NP-hard [? ]. Furthermore, by only considering the orderings of  $X$  that are induced by a common cherry-picking sequence of  $T$  and  $T'$  and minimizing across these orderings, we show that the minimum OLA, P2V, and HOP measures all equate to the temporal tree-child hybrid number of  $T$  and  $T'$  [? ].

The above results require a number of concepts that are localized in their use, and so we delay their introduction until the relevant sections of the paper. The paper is organized as follows. In the next section, we formally define each of OLA, P2V, and HOP. Section ?? establishes the above relationships between OLA and rSPR, and between HOP and the hybrid number, while Section ?? establishes the relationship between all three order-dependent measures and the temporal tree-child hybrid number. The relationships in Section ?? make use of agreement forests, while the relationship in Section ?? involves cherry-picking sequences. We end the paper with a discussion and some open problems.

## 2 Order-dependent measures

Throughout the paper,  $X$  denotes a non-empty finite set with  $|X| = n$ . We begin with some concepts on phylogenetic trees, and orderings and vectors, and then turn to defining the three order-dependent measures OLA, P2V, and HOP.

**Phylogenetic trees.** A *rooted binary phylogenetic  $X$ -tree*  $T$  is a rooted tree that satisfies the following three properties: (i) the unique root has in-degree zero and out-degree one, (ii) the leaves are bijectively labeled with the elements in  $X$ , and (iii) each remaining vertex has in-degree one and out-degree two. The set  $X$  is the *label set* of  $T$  and denoted by  $L(T)$ . If  $|X| = 1$ , then  $T$  contains a single edge that is incident with the root and the unique element in  $X$ . An example of a phylogenetic tree with  $X = \{a, b, c, d, e\}$  is shown in Figure ?? . Let  $T$  and  $T'$  be two rooted binary phylogenetic  $X$ -trees with vertex set  $V$  and  $V'$ , respectively. We say that  $T$  and  $T'$  are *isomorphic* if there exists a bijection  $\phi : V \rightarrow V'$  with  $\phi(x) = x$  for all  $x \in X$  and  $(u, v)$  is an edge in  $T$  if and only if  $(\phi(u), \phi(v))$  is an edge in  $T'$  for all  $u, v \in V$ . If  $T$  and  $T'$  are isomorphic, we write  $T \simeq T'$ . Since all phylogenetic trees in this paper are rooted and binary, we refer to a rooted binary phylogenetic  $X$ -tree as simply a *phylogenetic tree* throughout the remainder of the paper.

Let  $T$  be phylogenetic  $X$ -tree with root  $\rho$ , and let  $Y \subseteq X \cup \{\rho\}$ . We denote by  $T(Y)$  the minimal rooted subtree of  $T$  that connects the elements in  $Y$ . Furthermore, the *restriction of  $T$  to  $Y$* , denoted by  $T|Y$ , is the rooted tree that is obtained from  $T(Y)$  by suppressing all vertices of in-degree one and out-degree one. We note that the definitions of the label set of a phylogenetic tree and of two phylogenetic trees being isomorphic naturally extend to restrictions of phylogenetic trees. If  $T|Y$  and  $T|(X \cup \{\rho\}) - Y$  are vertex disjoint, we call  $T|Y$  a *pendant subtree* of  $T$ . A pair of leaves  $\{a, b\}$  of  $T$  is a *cherry* if  $a$  and  $b$  have a common parent. More generally, two vertices  $u$  and  $v$  of  $T$  are *siblings* if they have a common parent. Lastly, let  $T$  be a phylogenetic  $X$ -tree and let  $z$  be an element not in  $X$ . We say that  $z$  has been *adjoined* to  $T$  if we subdivide an edge of  $T$  with a new vertex,  $u$  say, and join  $u$  and  $z$  with a new edge  $(u, z)$ .

**Orderings and vectors.** An *ordering* on  $X$  is a bijection  $\sigma : X \rightarrow \{1, 2, \dots, n\}$ . For each  $x \in X$ , we refer to  $\sigma(x)$  as the *rank* of  $x$  under  $\sigma$ . To illustrate, consider the phylogenetic tree in Figure ?? . Here the ranking of  $a, b, c, d$ , and  $e$  is 1, 2, 3, 4, and 5, respectively. Let  $\sigma$  be an ordering on  $X$ , and let  $Y = \{y_1, y_2, \dots, y_m\}$  be a subset of  $X$ . We say that the elements in  $Y$  have *consecutive ranks* under  $\sigma$  if

$$\{\sigma(y_1), \sigma(y_2), \dots, \sigma(y_m)\} = \{i, i+1, \dots, i+m-1\}$$

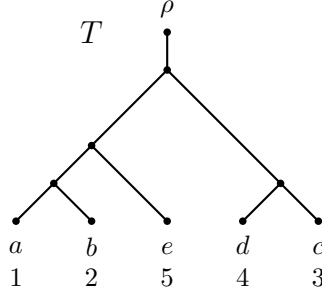


Figure 1: A rooted binary phylogenetic  $X$ -tree with root  $\rho$  and  $X = \{a, b, c, d, e\}$  together with an ordering  $\sigma$  on  $X$  defined by setting  $\sigma(a) = 1, \sigma(b) = 2, \dots, \sigma(e) = 5$ . The rank of each element  $x \in X$  under  $\sigma$  is given below the leaf label.

for some  $i \leq n - m + 1$ . Furthermore, if  $x$  is the element in  $X$  such that  $\sigma(x) = n$ , we use  $\sigma_{-x}$  to denote the ordering on  $X - \{x\}$  such that  $\sigma_{-x}(y) = \sigma(y)$  for each element  $y \in X - \{x\}$ .

Let  $\mathbf{v}$  be a vector. For the purposes of concatenating vectors, we write  $[u, \mathbf{v}, w]$  for  $[u, v_1, v_2, \dots, v_k, w]$ , where  $\mathbf{v} = [v_1, v_2, \dots, v_k]$ . Now let  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  be a vector of length  $n$  and let  $\sigma$  be an ordering on  $X$ . For each  $i \in \{1, 2, \dots, n\}$ , we say that  $v_i$  is *associated* with the element  $x \in X$  if  $\sigma(x) = i$ . Furthermore, for a subset  $Y$  of  $X$ , the vector  $\mathbf{v}$  *restricted to*  $Y$  is the vector with  $|Y|$  coordinates that can be obtained from  $\mathbf{v}$  by deleting each coordinate that is associated with an element in  $X - Y$ .

Lastly, we say that a vector  $\mathbf{u}$  is a *subsequence* of a vector  $\mathbf{v}$  if  $\mathbf{u}$  can be obtained from  $\mathbf{v}$  by the deletion of zero or more coordinates. More generally, let  $S$  be a set of vectors. A vector is a *common subsequence* of  $S$  if it is a subsequence of each vector in  $S$ . A vector is a *longest common subsequence (LCS)* of  $S$  if it is (i) a subsequence of each vector in  $S$  and (ii), among all such common subsequences, it has maximum length.

We next define the three order-dependent measures. In what follows, let  $T$  be a phylogenetic  $X$ -tree with root  $\rho$ , and let  $\sigma$  be an ordering on  $X$ . A *labeling* of  $T$  will be a map  $f : V(T) \rightarrow \mathcal{C}$ , where  $\mathcal{C}$  is a set of symbols. For each of the order-dependent measures, we present an algorithm whose input is a phylogenetic  $X$ -tree  $T$  and an ordering  $\sigma$  on  $X$ , and whose output is a vector associated with  $T$  and  $\sigma$ . Each algorithm proceeds by first defining a labeling of  $T$  or, in the case of P2V, a sequence of labelings of restrictions of  $T$ , and then using this labeling, respectively these labelings, to construct the outputted vector.

## 2.1 Ordered leaf attachment (OLA)

OLA [?] is the simplest of the three ordered-dependent measures to define. For OLA, the *OLA labeling* of  $T$  under  $\sigma$  is a map:

$$f_{\text{OLA}} : (V(T) - \{\rho\}) \rightarrow \{1, 2, \dots, n\} \cup \{-2, -3, \dots, -n\}.$$

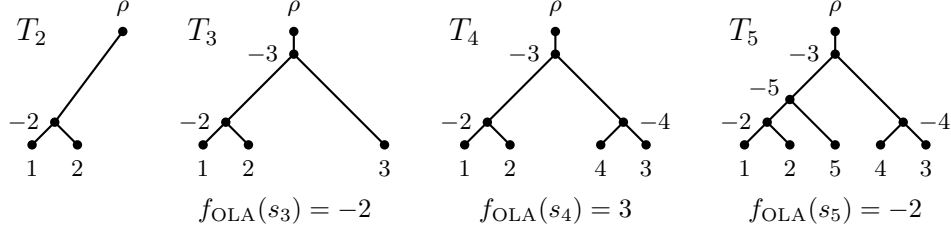


Figure 2: Illustration of the OLA vector construction (Algorithm ??) for the rooted phylogenetic  $X$ -tree on  $X = \{a, b, c, d, e\}$  depicted in Figure ?? under the ordering  $\sigma$  with  $\sigma(a) = 1, \sigma(b) = 2, \dots, \sigma(e) = 5$ . The labels of the interior vertices correspond to the OLA labeling,  $f_{\text{OLA}}$ , and for  $i \in \{3, 4, 5\}$ , we indicate the  $i$ -th vector coordinate of the OLA vector  $\mathbf{v}$  of  $T$  below tree  $T_i$ . It follows that  $\mathbf{v} = [0, 1, f_{\text{OLA}}(s_3), f_{\text{OLA}}(s_4), f_{\text{OLA}}(s_5)] = [0, 1, -2, 3, -2]$ .

---

**Algorithm 1** CONSTRUCT OLA VECTOR

---

- 1: **Input:** A phylogenetic  $X$ -tree with root  $\rho$  and an ordering  $\sigma$  on  $X$ .
  - 2: **Output:** The OLA vector  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  of  $T$  under  $\sigma$ .
  - 3: **for all**  $x \in X$  **do**
  - 4:   Set  $f_{\text{OLA}}(x) = \sigma(x)$ , i.e., each element of  $X$  is labeled by its rank under  $\sigma$ .
  - 5: **end for**
  - 6: Set  $T_2 \simeq T \setminus \{\rho, \sigma^{-1}(1), \sigma^{-1}(2)\}$ ;
  - 7: Set  $u_2$  to be the unique child of  $\rho$ ;
  - 8: Set  $f_{\text{OLA}}(u_2) = -2$ ;
  - 9: **for**  $i = 3, 4, \dots, n$  **do**
  - 10:   Set  $T_i$  to be the phylogenetic tree obtained from  $T_{i-1}$  by adjoining  $\sigma^{-1}(i)$  with the new edge  $(u_i, \sigma^{-1}(i))$  so that  $T_i \simeq T \setminus \{\rho, \sigma^{-1}(1), \sigma^{-1}(2), \dots, \sigma^{-1}(i)\}$ ;
  - 11:   Set  $f_{\text{OLA}}(u_i) = -i$ ;
  - 12:   Set  $s_i$  to be the sibling of  $\sigma^{-1}(i)$  in  $T_i$ ;
  - 13: **end for**
  - 14: **return**  $\mathbf{v} = [0, 1, f_{\text{OLA}}(s_3), f_{\text{OLA}}(s_4), \dots, f_{\text{OLA}}(s_n)]$
- 

The vector returned by Algorithm ?? is the *OLA vector* of  $T$  under  $\sigma$  (see Figure ?? for an illustration).

## 2.2 Phylo2Vec (P2V)

Similar to OLA, the P2V labeling [? ? ] iteratively assigns labels to the interior vertices of a phylogenetic tree based on an ordering of  $X$ . However, while OLA assigns a fixed negative integer to each interior vertex that remains unchanged as new leaves are adjoined, P2V recomputes the labels of these vertices after each new leaf is adjoined.

For all  $i \in \{2, 3, \dots, n\}$ , the  $i$ -th P2V labeling of  $T_i \simeq T \setminus \{\rho, \sigma^{-1}(1), \sigma^{-1}(2), \dots, \sigma^{-1}(i)\}$  under  $\sigma$  is a map

$$f_{\text{P2V}}^i : (V(T_i) - \{\rho\}) \rightarrow \{1, 2, \dots, i, i+1, \dots, 2i-1\}.$$

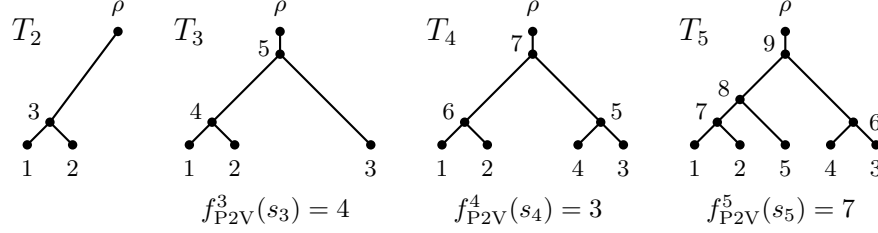


Figure 3: Illustration of the P2V vector construction (Algorithm ??) for the rooted phylogenetic  $X$ -tree on  $X = \{a, b, c, d, e\}$  depicted in Figure ?? under the ordering  $\sigma$  with  $\sigma(a) = 1, \sigma(b) = 2, \dots, \sigma(e) = 5$ . For  $i \in \{2, 3, 4, 5\}$ , the labels of the interior vertices of  $T_i$  correspond to the P2V labeling  $f_{P2V}^i$ . Notice that, in contrast to OLA (Figure ??), the interior vertices are re-labeled in each step. Furthermore, for  $i \in \{3, 4, 5\}$ , we indicate the  $i$ -th vector coordinate of the P2V vector  $\mathbf{v}$  of  $T$  below tree  $T_i$ . It follows that  $\mathbf{v} = [0, 1, f_{P2V}^3(s_3), f_{P2V}^4(s_4), f_{P2V}^5(s_5)] = [0, 1, 4, 3, 7]$ .

---

**Algorithm 2** CONSTRUCT P2V VECTOR

---

- 1: **Input:** A phylogenetic  $X$ -tree  $T$  with root  $\rho$  and an ordering  $\sigma$  on  $X$ .
  - 2: **Output:** The P2V vector  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  of  $T$  under  $\sigma$ .
  - 3: **for all**  $i \in \{2, 3, \dots, n\}$  and  $x \in X$  **do**
  - 4:   Set  $f_{P2V}^i(x) = \sigma(x)$ , i.e., each element of  $X$  is labeled by its rank under  $\sigma$ ;
  - 5: **end for**
  - 6: Set  $T_2 \simeq T \setminus \{\rho, \sigma^{-1}(1), \sigma^{-1}(2)\}$ ;
  - 7: Set  $u_2$  to be the child of  $\rho$ ;
  - 8: Set  $f_{P2V}^2(u_2) = 3$ ;
  - 9: **for**  $i = 3, 4, \dots, n$  **do**
  - 10:   Set  $T_i$  to be the phylogenetic tree obtained from  $T_{i-1}$  by adjoining  $\sigma^{-1}(i)$  with the new edge  $(u_i, \sigma^{-1}(i))$  so that  $T_i \simeq T \setminus \{\rho, \sigma^{-1}(1), \sigma^{-1}(2), \dots, \sigma^{-1}(i)\}$ ;
  - 11:   **for**  $j = i + 1, i + 2, \dots, 2i - 1$  **do**
  - 12:     Assign  $j$  to the vertex,  $w_j$  say, in  $\{u_2, u_3, \dots, u_i\}$  that has no  $i$ -th P2V label but whose two children have an  $i$ -th P2V label and, amongst all such vertices, has the highest  $i$ -th P2V labeled child;
  - 13:     Set  $f_{P2V}^i(w_j) = j$ ;
  - 14:   **end for**
  - 15:   Set  $s_i$  to be the sibling of  $\sigma^{-1}(i)$  in  $T_i$ ;
  - 16: **end for**
  - 17: **return**  $\mathbf{v} = [0, 1, f_{P2V}^3(s_3), f_{P2V}^4(s_4), \dots, f_{P2V}^n(s_n)]$
- 

The vector returned by Algorithm ?? is the *P2V vector* of  $T$  under  $\sigma$  (see Figure ?? for an illustration). Note that Step ?? is well defined as every phylogenetic tree has a cherry and so there is always at least one interior vertex with both of its children labeled. Furthermore, noting that  $T_n \simeq T$ , we denote the  $n$ -th P2V labeling of  $T_n$  by omitting the superscript and writing  $f_{P2V}$ .

## 2.3 HOP

Introduced in [? ], the HOP labeling assigns a positive integer to each interior vertex of  $T$  including the root such that there is a bijection between the interior vertices of  $T$  and the elements in  $\{1, 2, \dots, n\}$ . The associated vector with  $2n$  coordinates that arise from the  $n$  leaves and the  $n$  interior vertices of  $T$  is then obtained by sequentially decomposing  $T$  into  $n$  edge-disjoint paths.

The *HOP labeling* of  $T$  under  $\sigma$  is a map

$$f_{\text{HOP}} : V(T) \rightarrow \{\underline{1}, \underline{2}, \dots, \underline{n}, 1, 2, \dots, n\},$$

where in keeping with [?] the underlined numerals correspond to the leaves of  $T$  and the numerals not underlined correspond to the interior vertices of  $T$ .

---

**Algorithm 3** CONSTRUCT HOP VECTOR

---

```

1: Input: A phylogenetic  $X$ -tree  $T$  with root  $\rho$  and an ordering  $\sigma$  on  $X$ .
2: Output: The HOP vector  $\mathbf{v} = [v_1, v_2, \dots, v_{2n}]$  of  $T$  under  $\sigma$ .
3: for all  $x \in X$  do
4:   Set  $f_{\text{HOP}}(x) = \underline{\sigma(x)}$ ;
5: end for
6: for all  $v \in V(T)$  do
7:   Set  $C(v)$  to be the set of leaves  $x \in X$  with a directed path from  $v$  to  $x$  in  $T$ ;
8:   Compute  $m(v) = \min_{x \in C(v)} f_{\text{HOP}}(x)$ , i.e., find the minimum HOP label among the leaves in  $C(v)$ ;
9: end for
10: for all  $v \in V(T) - (L(T) \cup \{\rho\})$  do
11:   Label  $v$  by  $f_{\text{HOP}}(v) = \max\{m(v_1), m(v_2)\}$ , where  $v_1$  and  $v_2$  are the children of  $v$ ;
12: end for
13: Set  $f_{\text{HOP}}(\rho) = 1$ ;
14: for  $i = 1, 2, \dots, n$  do
15:   Let  $x$  be the leaf with  $\sigma(i) = x$  and  $v$  be the non-leaf vertex with  $f_{\text{HOP}}(v) = i$ ;
16:   Let  $P_i = (v = u_1, u_2, \dots, u_{k-1}, u_k = x)$  be the unique path from  $v$  to  $x$  in  $T$ ;
17:   Set  $\mathbf{v}(P_i) = [f_{\text{HOP}}(u_2), f_{\text{HOP}}(u_3), \dots, f_{\text{HOP}}(u_{k-1})]$ ;
18: end for
19: return  $\mathbf{v} = [1, \mathbf{v}(P_1), \underline{1}, \mathbf{v}(P_2), \underline{2}, \mathbf{v}(P_3), \underline{3}, \dots, \mathbf{v}(P_n), \underline{n}]$ 

```

---

The vector returned by Algorithm ?? is the *HOP vector* of  $T$  under  $\sigma$ . We sometimes abbreviate  $\mathbf{v}(P_i)$  in Line 17 of Algorithm ?? as  $\mathbf{v}_i$ . Note that, by definition,  $\mathbf{v}(P_n) = \mathbf{v}_n$  is empty, and so this leads to  $\mathbf{v} = [1, \mathbf{v}_1, \underline{1}, \mathbf{v}_2, \underline{2}, \mathbf{v}_3, \underline{3}, \dots, \mathbf{v}_{n-1}, \underline{n-1}, \underline{n}]$ .

We remark here that the HOP labeling of  $T$  under  $\sigma$  can alternatively be derived through a tree-growing process similar to the one described for OLA. As this derivation is not needed for the paper, we omit the details.

## 2.4 Order-dependent distances between phylogenetic trees

The primary motivation for the introduction of the OLA, P2V, and HOP vector representations is to quantify the dissimilarity between two phylogenetic  $X$ -trees  $T$  and  $T'$ . This dissimilarity depends on the choice of  $\sigma$ , an ordering of  $X$ , and is computed as follows:

1. The *OLA distance with respect to  $\sigma$*  between  $T$  and  $T'$ , denoted as  $d_{\text{OLA}}^\sigma(T, T')$ , is defined as the Hamming distance between the OLA vectors of  $T$  and  $T'$  under  $\sigma$ .
2. Similarly, the *P2V distance with respect to  $\sigma$*  between  $T$  and  $T'$ , denoted as  $d_{\text{P2V}}^\sigma(T, T')$ , is defined as the Hamming distance between the P2V vectors of  $T$  and  $T'$  under  $\sigma$ .
3. The analogue for HOP is a two-step process. Let  $\mathbf{u} = [1, \mathbf{u}_1, \underline{1}, \mathbf{u}_2, \underline{2}, \dots, \mathbf{u}_{n-1}, \underline{n-1}, \underline{n}]$  and  $\mathbf{v} = [1, \mathbf{v}_1, \underline{1}, \mathbf{v}_2, \underline{2}, \dots, \mathbf{v}_{n-1}, \underline{n-1}, \underline{n}]$  be the HOP vectors of  $T$  and  $T'$  under  $\sigma$ . The *HOP similarity with respect to  $\sigma$*  is defined as

$$\text{Sim}_{\text{HOP}}^\sigma(T, T') = \sum_{1 \leq i \leq n-1} |\text{LCS}(\mathbf{u}_i, \mathbf{v}_i)|.$$

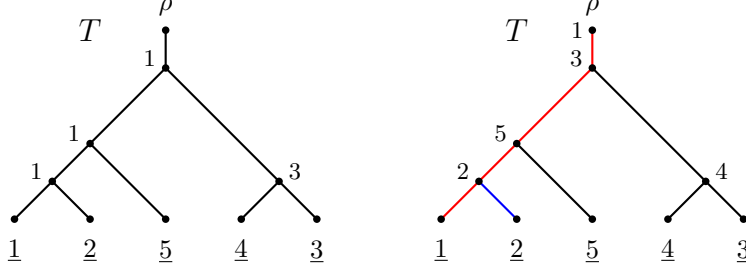


Figure 4: Illustration of the HOP vector construction (Algorithm ??) for the rooted phylogenetic  $X$ -tree on  $X = \{a, b, c, d, e\}$  depicted in Figure ?? under the ordering  $\sigma$  with  $\sigma(a) = 1, \sigma(b) = 2, \dots, \sigma(e) = 5$ . On the left, each vertex  $v \in V(T) - \{\rho\}$  is labeled by  $m(v)$ , the minimum HOP label among the leaves below  $v$ , with leaf labels underlined. On the right, the root is labeled by 1, the leaves are labeled by their ranks (underlined), and each vertex  $v \in V(T) - (L(T) \cup \{\rho\})$  with children  $v_1, v_2$  is labeled by  $\max\{m(v_1), m(v_2)\}$ . For further illustration, the red edges correspond to the path  $P_1$ , and similarly, the blue edge corresponds to the path  $P_2$ . Finally, the HOP vector of  $T$  is given by  $\mathbf{v} = [1, 3, 5, 2, \underline{1}, \underline{2}, 4, \underline{3}, \underline{4}, \underline{5}]$ .

In turn, the *HOP distance with respect to  $\sigma$*  between  $T$  and  $T'$  is given by

$$d_{\text{HOP}}^\sigma(T, T') = n - \text{Sim}_{\text{HOP}}^\sigma(\mathbf{u}, \mathbf{v}) = n - \sum_{1 \leq i \leq n-1} |\text{LCS}(\mathbf{u}_i, \mathbf{v}_i)|.$$

It has been shown in [? ], [? ], and [? ], respectively, that under a fixed  $\sigma$  each of  $d_{\text{OLA}}^\sigma(T, T')$ ,  $d_{\text{P2V}}^\sigma(T, T')$ , and  $d_{\text{HOP}}^\sigma(T, T')$  satisfies the triangle inequality and, therefore, is a distance on the set of phylogenetic  $X$ -trees.

Lastly, for each  $\Theta \in \{\text{HOP}, \text{OLA}, \text{P2V}\}$ , we define

$$d_\Theta^*(T, T') = \min \{d_\Theta^\sigma(T, T') : \sigma \text{ is an ordering on } X\},$$

and refer to  $d_\Theta^*(T, T')$  as the  $\Theta$  *measure between  $T$  and  $T'$* .

The reason for calling it a measure is because  $d_\Theta^*(T, T')$  is not a distance on the set of phylogenetic  $X$ -trees. We give concrete examples showing that  $d_\Theta^*(T, T')$  does not satisfy the triangle inequality in Figure ?? (for OLA) and Figure ?? (for HOP and P2V).

### 3 Bounding order-dependent measures by agreement forests

There are several measures to compute the dissimilarities between two phylogenetic trees on the same label set that are based on agreement forests. These measures include the rooted subtree prune and regraft distance and the hybrid number that we formally define next.

Let  $T$  be a rooted phylogenetic  $X$ -tree with root  $\rho$ . Furthermore, let  $T'$  be a rooted phylogenetic tree that can be obtained from  $T$  by deleting an edge  $(u, v)$  in  $T$  with  $u \neq \rho$ , suppressing  $u$ , and then *adjoining* the subtree with root  $v$  to the phylogenetic tree that contains  $\rho$  by subdividing an edge of the latter with a new vertex,  $u'$  say, and then adding the edge  $(u', v)$ . We say that  $T'$  has been obtained from  $T$  by a *rooted subtree prune and regraft (rSPR) move*. Moreover, we define the *rSPR distance*, denoted  $d_{\text{rSPR}}(T, T')$ , to be the minimum number of rSPR moves that transforms  $T$  to  $T'$ . Note that an rSPR move is reversible, so  $d_{\text{rSPR}}(T, T') = d_{\text{rSPR}}(T', T)$ .

Now, let  $T$  and  $T'$  be two phylogenetic  $X$ -trees whose roots are labeled with  $\rho$ . For the purpose of the upcoming definition, we view  $\rho$  as an element of the label set of  $T$  and  $T'$ . An *agreement forest* for  $T$  and  $T'$

is a collection  $\{T_\rho, T_1, T_2, \dots, T_k\}$  of rooted trees with label sets  $L_\rho, L_1, L_2, \dots, L_k$  such that the following properties are satisfied:

1. The label sets  $L_\rho, L_1, L_2, \dots, L_k$  partition  $X \cup \{\rho\}$  and, in particular,  $\rho \in L_\rho$ .
2. For all  $i \in \{\rho, 1, 2, \dots, k\}$ , the rooted tree  $T_i \simeq T|_{L_i} \simeq T'|_{L_i}$ .
3. The trees in  $\{T(L_i) : i \in \{\rho, 1, 2, \dots, k\}\}$  and  $\{T'(L_i) : i \in \{\rho, 1, 2, \dots, k\}\}$  are vertex-disjoint rooted subtrees of  $T$  and  $T'$ , respectively.

A *maximum agreement forest* for  $T$  and  $T'$  is an agreement forest  $\{T_\rho, T_1, T_2, \dots, T_k\}$  in which  $k$  (the number of components minus one) is minimized. The minimum possible value for  $k$  is denoted by  $m(T, T')$ . The next theorem is due to Bordewich and Semple [?, Theorem 2.1] and establishes a characterization of the rSPR distance between two phylogenetic  $X$ -trees in terms of agreement forests.

**Theorem 1.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then  $d_{\text{rSPR}}(T, T') = m(T, T')$ .*

Following on from Theorem ??, Bordewich and Semple [?] showed that computing the rSPR distance between two phylogenetic trees is NP-hard via a reduction from Exact Cover by 3-Sets to maximum agreement forests.

In order to define the hybrid number of two phylogenetic  $X$ -trees  $T$  and  $T'$ , we first introduce phylogenetic networks. A *rooted binary phylogenetic network*  $N$  on  $X$  is a rooted acyclic directed graph with no parallel edges that satisfies the following properties:

1. the unique root has in-degree zero and out-degree one,
2. vertices with out-degree zero have in-degree one, and the set of vertices with out-degree zero is  $X$ , and
3. all other vertices have either in-degree one and out-degree two or in-degree two and out-degree one.

Furthermore, we use  $h(N)$  to denote the number of vertices with in-degree two in  $N$ . Since all phylogenetic networks in this paper are rooted and binary, we refer to a rooted binary phylogenetic network on  $X$  as simply a *phylogenetic network* from now on.

We next define two particular classes of phylogenetic networks. The first features prominently in the literature. Let  $N$  be a phylogenetic network on  $X$ . We say that  $N$  is *tree-child* if each non-leaf vertex has a child with in-degree one. Moreover, we say that  $N$  is *temporal* if there exists a map  $t : V \rightarrow \mathbb{R}^+$  such that, for each edge  $(u, v)$  in  $N$ , we have  $t(u) = t(v)$  if  $v$  has in-degree two, and  $t(u) < t(v)$  if  $v$  has in-degree one.

Now, let  $N$  be a phylogenetic network on  $X$ , and let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. We say that  $N$  *displays*  $T$  if there exists a subtree of  $N$  that is a subdivision of  $T$ .

With this definition in hand, we set

$$h(T, T') = \min\{h(N) : N \text{ is a phylogenetic network on } X \text{ that displays } T \text{ and } T'\}$$

and refer to  $h(T, T')$  as the *hybrid number* of  $T$  and  $T'$ . Similarly, we set

$$h_{tc}(T, T') = \min\{h(N) : N \text{ is a tree-child network on } X \text{ that displays } T \text{ and } T'\}$$

and

$$h_t(T, T') = \min\{h(N) : N \text{ is a temporal tree-child network on } X \text{ that displays } T \text{ and } T'\},$$

and refer to  $h_{tc}(T, T')$  and  $h_t(T, T')$  as the *tree-child hybrid number* and the *temporal tree-child hybrid number*, respectively, of  $T$  and  $T'$ . The following lemma was essentially established as part of the proof of [?, Theorem 2] and formally noted in [?]. It shows that the hybrid number of two phylogenetic trees is equal to their tree-child hybrid number.



**Lemma 1.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then  $h(T, T') = h_{tc}(T, T')$ .*

We next define a concept that is closely related to agreement forests. Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, and let  $F = \{T_\rho, T_1, T_2, \dots, T_k\}$  be an agreement forest for  $T$  and  $T'$ . We say that  $F$  is *acyclic* if the directed graph with vertex set  $F$  and for which  $(T_i, T_j)$  with  $i, j \in \{\rho, 1, 2, \dots, k\}$  and  $i \neq j$  is an edge precisely if the root of  $T(L(T_i))$  is an ancestor of the root of  $T(L(T_j))$ , or the root of  $T'(L(T_i))$  is an ancestor of the root of  $T'(L(T_j))$  has no directed cycles. Note that such an agreement forest always exists. Now, let  $F = \{T_\rho, T_1, T_2, \dots, T_k\}$  be an acyclic agreement forest for  $T$  and  $T'$ . Similar to agreement forests, we say that  $F$  is a *maximum acyclic agreement forest* for  $T$  and  $T'$  if  $k$  is minimum over all acyclic agreement forests for  $T$  and  $T'$ . We denote this minimum number by  $m_a(T, T')$ . The relevance of acyclic agreement forests is the next theorem established in [?, Theorem 2].

**Theorem 2.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then  $h(T, T') = m_a(T, T')$ .*

Similar to computing the rSPR distance, Bordewich and Semple [?] showed that computing the hybrid number for two phylogenetic trees is NP-hard. Moreover, by Lemma ??, it immediately follows that computing the tree-child hybrid number is also NP-hard.

**Notational remark.** In what follows, we sometimes compare the  $\Theta$  labelings of two phylogenetic trees where  $\Theta \in \{\text{HOP}, \text{OLA}, \text{P2V}\}$ . In this case, we write  $f_\Theta^T$  instead of  $f_\Theta$  to make a direct reference to a phylogenetic tree  $T$  with  $n$  leaves. For  $\Theta = \text{P2V}$ , we write  $f_{\text{P2V}}^T$  to refer to the P2V labeling of  $T_n \simeq T$ .

### 3.1 Bounding by the rSPR distance

In this section, we focus on the relationship between the order-dependent measures OLA, P2V, and HOP, and the rSPR distance. We begin by showing that given two phylogenetic  $X$ -trees  $T$  and  $T'$ , the measure  $d_{\text{OLA}}^*(T, T')$  is bounded from above by a function that is linear in the rSPR distance  $d_{\text{rSPR}}(T, T')$  of  $T$  and  $T'$ .

**Theorem 3.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then, there exists an ordering  $\sigma$  such that*

$$d_{\text{OLA}}^\sigma(T, T') \leq 28 \cdot d_{\text{rSPR}}(T, T').$$

*In particular,  $d_{\text{OLA}}^*(T, T') \leq 28 \cdot d_{\text{rSPR}}(T, T')$ .*

In order to prove this statement, we first show that two tree reduction rules can be used to reduce the size of the label set of a pair of phylogenetic  $X$ -trees while preserving the OLA distance with respect to  $\sigma$  between them. These two rules coincide with the two tree reduction rules used by [?] for the rSPR distance.

Let  $T$  be a phylogenetic  $X$ -tree, and let  $C = (x_1, x_2, \dots, x_l)$  be a sequence of elements in  $X$  with  $l \geq 3$ . We say that  $C$  is a *chain* of  $T$  if the parent of  $x_1$  coincides with the parent of  $x_2$  or the parent of  $x_2$  is the parent of the parent of  $x_1$ , and, for each  $i \in \{3, 4, \dots, l\}$ , the parent of  $x_i$  is the parent of the parent of  $x_{i-1}$ . In what follows, we sometimes abuse notation and write  $L(C)$  to denote the set  $\{x_1, x_2, \dots, x_l\}$ .

Now, let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Each of the following reductions applied to  $T$  and  $T'$  results in two new phylogenetic trees  $S$  and  $S'$  with fewer leaves.

- **Subtree reduction:** Let  $P$  be a maximal common pendant subtree of  $T$  and  $T'$  with at least two leaves. Obtain  $S$  and  $S'$  from  $T$  and  $T'$ , respectively, by replacing  $P$  with a single leaf with a new label that is not in  $X$ . Thus, a subtree reduction replaces a common pendant subtree with a single leaf.
- **Chain reduction:** Let  $C = (x_1, x_2, \dots, x_l)$  be a maximal common chain of  $T$  and  $T'$  with  $l \geq 4$ . Obtain  $S$  and  $S'$  from  $T|(X - \{x_4, x_5, \dots, x_l\})$  and  $T'|(X - \{x_4, x_5, \dots, x_l\})$  respectively, by replacing leaf labels  $x_1, x_2$ , and  $x_3$  with three new labels that are not in  $X$ . Thus, a chain reduction replaces a common chain of length at least four with a common chain of length three.

We next describe a 2-step construction to obtain two phylogenetic trees from  $T$  and  $T'$  by repeated applications of the subtree and chain reductions. For  $m \geq 0$ , let  $P_1, P_2, \dots, P_m$  be distinct maximal pendant subtrees with at least two leaves that are common to  $T$  and  $T'$ . Observe that  $L(P_i) \cap L(P_j) = \emptyset$  for two distinct elements  $i, j \in \{1, 2, \dots, m\}$ . If  $m = 0$ , then set  $S = T$  and  $S' = T'$  and, otherwise, obtain two phylogenetic  $X'$ -trees  $S$  and  $S'$  from  $T$  and  $T'$ , respectively, by applying the subtree reduction to each  $P_i$  such that each  $P_i$  is replaced with a single leaf labeled  $p_i$ , where  $p_i \notin X$ . Set

$$X' = \left(X - \bigcup_{i=1}^m L(P_i)\right) \cup \{p_1, p_2, \dots, p_m\}.$$

Next, for  $m' \geq 0$ , let  $C_1, C_2, \dots, C_{m'}$  be distinct maximal chains of length at least four that are common to  $S$  and  $S'$ . Again, by the maximality of each such chain, there exists no element in  $X'$  that is a leaf of  $C_i$  and  $C_j$  for two distinct elements  $i, j \in \{1, 2, \dots, m'\}$ . If  $m' = 0$ , set  $R = S$  and  $R' = S'$  and, otherwise, obtain two phylogenetic  $X''$ -trees  $R$  and  $R'$  from  $S$  and  $S'$ , respectively, by applying the chain reduction to each  $C_i = (x'_1, x'_2, \dots, x'_l)$  such that  $x'_1, x'_2, x'_3$  is replaced with  $c_i^1, c_i^2, c_i^3$ , respectively. Set

$$X'' = \left(X' - \bigcup_{i=1}^{m'} L(C_i)\right) \cup \bigcup_{i=1}^{m'} \{c_i^1, c_i^2, c_i^3\}.$$

If at least one subtree or one chain reduction has been applied in the process of obtaining  $R$  and  $R'$  from  $T$  and  $T'$ , respectively, then we refer to  $R$  and  $R'$  as a *reduced tree pair* with respect to  $T$  and  $T'$ . Moreover, if  $R$  and  $R'$  cannot be further reduced under the subtree or chain reduction, we refer to  $R$  and  $R'$  as a *fully reduced tree pair* with respect to  $T$  and  $T'$ . Observe that a fully reduced tree pair with respect to  $T$  and  $T'$  can be obtained by applying the above 2-step process so that all maximal pendant subtrees that are common to  $T$  and  $T'$  are reduced under the subtree reduction and, then, all maximal chains that are common to  $S$  and  $S'$  are reduced under the chain reduction.

Now let  $\sigma''$  be an ordering on  $X''$  such that, for each  $i \in \{1, 2, \dots, m'\}$ , the elements  $c_i^1, c_i^2$ , and  $c_i^3$  have consecutive ranks under  $\sigma''$  such that  $\sigma''(c_i^1) < \sigma''(c_i^2) < \sigma''(c_i^3)$ . Starting with  $\sigma''$ , obtain an ordering  $\sigma'$  on  $X'$  such that, for each  $C_i = (x'_1, x'_2, \dots, x'_l)$  with  $i \in \{1, 2, \dots, m'\}$ , the elements in  $L(C_i)$  have consecutive ranks under  $\sigma'$  with  $\sigma'(x'_1) < \sigma'(x'_2) < \dots < \sigma'(x'_l)$  and at most one of the following holds for any two distinct elements  $y$  and  $y'$  in  $X''$  with  $\sigma''(y) < \sigma''(y')$ :

1. If  $y$  and  $y'$  are both elements in  $X'$ , then  $\sigma'(y) < \sigma'(y')$ .
2. If  $y \in \{c_i^1, c_i^2, c_i^3\}$  and  $y' \in \{c_j^1, c_j^2, c_j^3\}$  with  $i \neq j$ , then  $\sigma'(z) < \sigma'(z')$  for each pair  $z$  and  $z'$  of elements with  $z \in L(C_i)$  and  $z' \in L(C_j)$ .
3. If  $y \in X'$  and  $y' \in \{c_i^1, c_i^2, c_i^3\}$ , then  $\sigma'(y) < \sigma'(z)$  for each  $z \in L(C_i)$ .
4. If  $y \in \{c_i^1, c_i^2, c_i^3\}$  and  $y' \in X'$ , then  $\sigma'(z) < \sigma'(y')$  for each  $z \in L(C_i)$ .

Lastly, obtain an ordering  $\sigma$  on  $X$  such that, for each  $P_i$  with  $i \in \{1, 2, \dots, m\}$ , the elements in  $L(P_i)$  have consecutive ranks under  $\sigma$  and at most one of the following holds for any two distinct elements  $y$  and  $y'$  in  $X'$  with  $\sigma'(y) < \sigma'(y')$ :

1. If  $y$  and  $y'$  are both elements in  $X$ , then  $\sigma(y) < \sigma(y')$ .
2. If  $y = p_i$  and  $y' = p_j$  with  $i \neq j$ , then  $\sigma(z) < \sigma(z')$  for each pair  $z$  and  $z'$  of elements with  $z \in L(P_i)$  and  $z' \in L(P_j)$ .
3. If  $y \in X$  and  $y' = p_i$ , then  $\sigma'(y) < \sigma'(z)$  for each  $z \in L(P_i)$ .
4. If  $y = p_i$  and  $y' \in X$ , then  $\sigma'(z) < \sigma'(y')$  for each  $z \in L(P_i)$ .

We refer to  $\sigma$  (resp.  $\sigma'$ ) as a *reduction preserving ordering* on  $X$  (resp.  $X'$ ) relative to  $R$  and  $R'$ .

**Lemma 2.** Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Let  $U$  and  $U'$  be two phylogenetic  $X'$ -trees that can be obtained from  $T$  and  $T'$ , respectively, by a single application of the subtree or chain reduction. Let  $\sigma'$  be an ordering on  $X'$  such that, if  $T$  and  $T'$  have a maximal common chain  $C$  that has been reduced to the chain  $(c^1, c^2, c^3)$  in  $U$  and  $U'$ , then  $c^1$ ,  $c^2$ , and  $c^3$  have consecutive ranks with  $\sigma'(c^1) < \sigma'(c^2) < \sigma'(c^3)$ . Then,

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{OLA}}^{\sigma'}(S, S'),$$

where  $\sigma$  is a reduction preserving ordering on  $X$  relative to  $S$  and  $S'$ .

*Proof.* Let  $T, T', U, U'$ ,  $\sigma$ , and  $\sigma'$  be as stated in the lemma. Let  $\mathbf{v}_U^{\sigma'} = [u_1, u_2, \dots, u_{|X'|}]$  and  $\mathbf{v}_{U'}^{\sigma'} = [u'_1, u'_2, \dots, u'_{|X'|}]$  be the OLA vectors of  $U$  and  $U'$  under  $\sigma'$ . By assumption, the Hamming distance between  $\mathbf{v}_U^{\sigma'}$  and  $\mathbf{v}_{U'}^{\sigma'}$  equals  $d_{\text{OLA}}^{\sigma'}(U, U')$ .

Now, first consider the case that  $U$  and  $U'$  have been obtained from  $T$  and  $T'$  by a single application of the subtree reduction. In this case, let  $P$  denote the common maximal pendant subtree of  $T$  and  $T'$ , and let  $p \in X' - X$  denote the unique leaf present in  $U$  and  $U'$  but not in  $T$  and  $T'$ . Further, suppose that  $\sigma'(p) = i$  for some  $i \in \{1, \dots, |X'|\}$ . In particular, suppose that  $u_i$  and  $u'_i$  are the elements associated with  $p$  in  $\mathbf{v}_U^{\sigma'}$  and  $\mathbf{v}_{U'}^{\sigma'}$ , respectively. Since  $\sigma$  is a reduction preserving ordering on  $X$  relative to  $U$  and  $U'$ , we now claim that we can obtain the OLA vectors for  $T$  and  $T'$  from the OLA vectors of  $U$  and  $U'$  by setting

$$\begin{aligned} \mathbf{v}_T^\sigma &= [u_1, u_2, \dots, u_{i-1}, u_i, v_2, \dots, v_{|L(P)|}, \tilde{u}_{i+1}, \dots, \tilde{u}_{|X'|}] \quad \text{and} \\ \mathbf{v}_{T'}^\sigma &= [u'_1, u'_2, \dots, u'_{i-1}, u'_i, v_2, \dots, v_{|L(P)|}, \tilde{u}'_{i+1}, \dots, \tilde{u}'_{|X'|}], \end{aligned}$$

where, the Hamming distance between  $[\tilde{u}_{i+1}, \tilde{u}_{i+2}, \dots, \tilde{u}_{|X'|}]$  and  $[\tilde{u}'_{i+1}, \tilde{u}'_{i+2}, \dots, \tilde{u}'_{|X'|}]$  equals that of  $[u_{i+1}, u_{i+2}, \dots, u_{|X'|}]$  and  $[u'_{i+1}, u'_{i+2}, \dots, u'_{|X'|}]$ . This is due to the following facts:

- (i) All elements of  $X$  whose ranks under  $\sigma'$  are less than  $\sigma'(p)$  are considered first and in the same order under  $\sigma'$  and  $\sigma$  when iteratively building  $U$  and  $U'$ , respectively  $T$  and  $T'$ , to obtain the OLA vectors. Thus, the first  $i - 1$  coordinates of  $\mathbf{v}_U^{\sigma'}$  and  $\mathbf{v}_{T'}^\sigma$  (resp.  $\mathbf{v}_{U'}^{\sigma'}$  and  $\mathbf{v}_T^\sigma$ ) coincide.
- (ii) Now, consider the elements of  $X \cap L(P)$  and let  $x_p$  the leaf with minimal rank under  $\sigma$  in this set, i.e.,  $\sigma(x_p) = \min_{x \in L(P)} \sigma(x)$ . Since  $P$  is a common pendant subtree of  $T$  and  $T'$  and the elements of  $L(P)$  have consecutive ranks under  $\sigma$ , it follows from Algorithm ?? that the vector coordinates for  $x' \in X \cap L(P) - \{x_p\}$  are identical in  $T$  and  $T'$  and they correspond to positions  $i+2, i+3, \dots, i+|L(P)|-1$  of  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$ , respectively. Now, the element associated with  $x_p$  may differ between  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$ ; however, it clearly coincides with the element associated with  $p$ , namely  $u_i$  (resp.  $u'_i$ ) in  $\mathbf{v}_U^{\sigma'}$  (resp.  $\mathbf{v}_{U'}^{\sigma'}$ ).
- (iii) Finally, all elements of  $X$  whose ranks under  $\sigma'$  are greater than  $\sigma'(p)$  are considered last and in the same order under  $\sigma'$  and  $\sigma$  when iteratively building  $U$  and  $U'$  (resp.  $T$  and  $T'$ ) to obtain the OLA vectors. Note that for each vertex  $v$  of  $T$  (resp.  $T'$ ) that is introduced after the elements in  $L(P)$  are added, we have  $f_{\text{OLA}}^T(v) = f_{\text{OLA}}^U(u) + |L(P)| - 1$  and  $f_{\text{OLA}}^{T'}(v) = f_{\text{OLA}}^{U'}(u) + |L(P)| - 1$ , where  $u$  is the vertex of  $U$  (resp.  $U'$ ) corresponding to  $v$  in  $T$  (resp.  $T'$ ). Since the elements of  $X$  whose ranks under  $\sigma'$  are greater than  $\sigma'(p)$  are clearly not added as siblings to vertices of  $P$ , we can conclude that  $u_j = u'_j$  if and only if  $\tilde{u}_j = \tilde{u}'_j$  for each  $j \in \{i+1, i+2, \dots, |X'|\}$ . This implies that the Hamming distance between  $[\tilde{u}_{i+1}, \tilde{u}_{i+2}, \dots, \tilde{u}_{|X'|}]$  and  $[\tilde{u}'_{i+1}, \tilde{u}'_{i+2}, \dots, \tilde{u}'_{|X'|}]$  equals that of  $[u_{i+1}, u_{i+2}, \dots, u_{|X'|}]$  and  $[u'_{i+1}, u'_{i+2}, \dots, u'_{|X'|}]$ .

In summary, the Hamming distance between  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$  equals the Hamming distance of  $\mathbf{v}_U^{\sigma'}$  and  $\mathbf{v}_{U'}^{\sigma'}$ . Thus, when  $U$  and  $U'$  have been obtained from  $T$  and  $T'$  by a single application of the subtree reduction,  $d_{\text{OLA}}^\sigma(T, T') = d_{\text{OLA}}^{\sigma'}(U, U')$  as claimed.

Next, consider the case that  $U$  and  $U'$  have been obtained from  $T$  and  $T'$  by a single application of the chain reduction. In this case, let  $C = (x_1, x_2, x_3, \dots, x_{|L(C)|})$  denote the common chain of  $T$  and  $T'$ , and let

$c^1, c^2, c^3 \in X' - X$  denote the leaves present in  $U$  and  $U'$  but not in  $T$  and  $T'$ . Let  $i \in \{1, \dots, |X'| - 2\}$  be such that  $u_i, u_{i+1}, u_{i+2}$  (resp.  $u'_i, u'_{i+1}, u'_{i+2}$ ) are the elements associated with  $c^1, c^2, c^3$  in  $\mathbf{v}_U^{\sigma'}$  (resp.  $\mathbf{v}_{U'}^{\sigma'}$ ). Since  $\sigma$  is a reduction preserving ordering on  $X$  relative to  $U$  and  $U'$ , we now claim that we can obtain the OLA vectors for  $T$  and  $T'$  from the OLA vectors of  $U$  and  $U'$  by setting

$$\begin{aligned}\mathbf{v}_T^\sigma &= [u_1, u_2, \dots, u_{i-1}, u_i, u_{i+1}, u_{i+2}, v_4, \dots, v_{|L(C)|}, \tilde{u}_{i+3}, \dots, \tilde{u}_{|X'|}] \quad \text{and} \\ \mathbf{v}_{T'}^\sigma &= [u'_1, u'_2, \dots, u'_{i-1}, u'_i, u'_{i+1}, u'_{i+2}, v_4, \dots, v_{|L(C)|}, \tilde{u}'_{i+3}, \dots, \tilde{u}'_{|X'|}],\end{aligned}$$

where, the Hamming distance between  $[\tilde{u}_{i+3}, \tilde{u}_{i+4}, \dots, \tilde{u}_{|X'|}]$  and  $[\tilde{u}'_{i+3}, \tilde{u}'_{i+4}, \dots, \tilde{u}'_{|X'|}]$  equals that of  $[u_{i+3}, u_{i+4}, \dots, u_{|X'|}]$  and  $[u'_{i+3}, u'_{i+4}, \dots, u'_{|X'|}]$ . The reasoning is similar as in the previous case. In particular, the arguments presented in (i) and (iii) above for the elements of  $X - L(C)$  immediately carry over. Thus, it remains to consider the elements of  $X \cap L(C)$ . Since  $C = (x_1, x_2, x_3, \dots, x_{|L(C)|})$  is a common chain of  $T$  and  $T'$  and, by construction, the elements of  $L(C)$  have consecutive ranks under  $\sigma$  with  $\sigma(x_1) < \sigma(x_2) < \dots < \sigma(x_{|L(C)|})$ , it follows from Algorithm ?? that the elements of  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$ , associated with  $x_4, x_5, \dots, x_{|L(C)|}$  coincide (specifically, the element associated with  $x_j$  is  $-\sigma(x_{j-1})$  for  $j \in \{4, 5, \dots, |L(C)|\}$ ) in both vectors. Furthermore, the elements associated with  $x_1, x_2, x_3$  may differ between  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$ ; however, they coincide with the elements associated with  $c^1, c^2, c^3$  in  $\mathbf{v}_U^{\sigma'}$  (resp.  $\mathbf{v}_{U'}^{\sigma'}$ ), namely  $u_i, u_{i+1}, u_{i+2}$  (resp.  $u'_i, u'_{i+1}, u'_{i+2}$ ).

This shows that the Hamming distance between  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$  equals the Hamming distance of  $\mathbf{v}_U^{\sigma'}$  and  $\mathbf{v}_{U'}^{\sigma'}$ , thereby completing the proof.  $\square$

We now recall Lemma 3.3 from [? ]:

**Lemma 3.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Let  $R$  and  $R'$  be a fully reduced tree pair with respect to  $T$  and  $T'$ , and let  $X'$  be the label set of  $R$ . Then*

$$|X'| \leq 28 \cdot d_{\text{rSPR}}(T, T').$$

*Proof of Theorem ??.* Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, let  $S$  and  $S'$  be two phylogenetic  $X'$ -trees obtained from  $T$  and  $T'$  by reducing all maximal common pendant subtrees, and let  $R$  and  $R'$  be two phylogenetic  $X''$ -trees obtained from  $S$  and  $S'$  by reducing all maximal common chains. By Lemma ??,  $|X''| \leq 28 \cdot d_{\text{rSPR}}(T, T')$ .

Let  $\sigma''$  be an ordering on  $X''$  such that, if  $S$  and  $S'$  have a maximal common chain  $C$  that has been reduced to the chain  $(c^1, c^2, c^3)$  in  $R$  and  $R'$ , then  $c^1, c^2$ , and  $c^3$  have consecutive ranks with  $\sigma''(c^1) < \sigma''(c^2) < \sigma''(c^3)$ . Further, let  $\sigma'$  and  $\sigma$  be orderings on  $X'$  and  $X$ , respectively, satisfying the conditions stated above, such that  $\sigma$  (resp.  $\sigma'$ ) is a reduction preserving ordering on  $X$  (resp.  $X'$ ) relative to  $R$  and  $R'$ .

We now repeatedly apply Lemma ?? with respect to chain reductions to conclude that

$$d_{\text{OLA}}^{\sigma'}(S, S') = d_{\text{OLA}}^{\sigma''}(R, R').$$

Next, we repeatedly apply Lemma ?? with respect to subtree reductions to conclude that

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{OLA}}^{\sigma'}(S, S').$$

In summary, this implies that based on an ordering  $\sigma''$  on  $X''$ , for a reduction preserving ordering  $\sigma$  on  $X$  relative to  $R$  and  $R'$ , we have

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{OLA}}^{\sigma''}(R, R').$$

Now, clearly  $d_{\text{OLA}}^{\sigma''}(R, R') \leq |X''|$ , implying that

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{OLA}}^{\sigma''}(R, R') \leq |X''| \leq 28 \cdot d_{\text{rSPR}}(T, T'),$$

which completes the proof.  $\square$

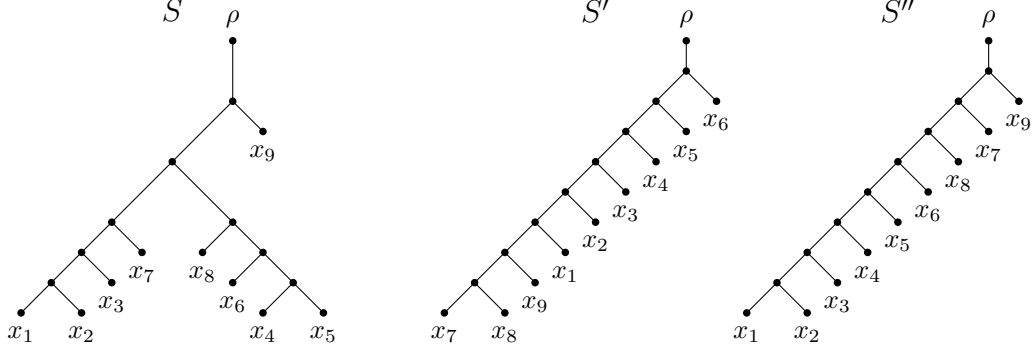


Figure 5: Two phylogenetic  $X$ -trees  $S$  and  $S'$  with  $d_{\text{OLA}}^*(S, S') \geq 4$  (verified via exhaustive enumeration of all orderings), whereas  $d_{\text{rSPR}}(S, S') = 3$ . Moreover,  $d_{\text{OLA}}^*(S, S'') = 1$  and  $d_{\text{OLA}}^*(S', S'') = 2$ , implying that the OLA measure is not a distance since the triangle quality is violated for  $S$ ,  $S'$ , and  $S''$ .

We end this section with three remarks on the relationship between the rSPR distance and the OLA and P2V measures, and an open problem raised in [? ]. First, although Theorem ?? shows that  $d_{\text{OLA}}^*(T, T')$  is bounded above by a function that is linear in  $d_{\text{rSPR}}(T, T')$  for any two phylogenetic trees  $T$  and  $T'$ , there also exist pairs  $S$  and  $S'$  of phylogenetic trees such that  $d_{\text{rSPR}}(S, S') < d_{\text{OLA}}^*(S, S')$ . For example, Figure ?? shows such a tree pair for which  $3 = d_{\text{rSPR}}(S, S') < d_{\text{OLA}}^*(S, S') = 4$ .

Second, for two phylogenetic  $X$ -trees  $T$  and  $T'$ , Richman et al. [? , page 10] ask whether or not it is possible to bound the difference

$$\left| d_{\text{OLA}}^\sigma(T, T') - d_{\text{OLA}}^{\sigma'}(T, T') \right|,$$

where  $\sigma$  and  $\sigma'$  are two orderings on  $X$ . We answer this question negatively by providing, for all  $n \geq 1$ , two phylogenetic trees on  $n + 1$  leaves for which the difference in OLA distances with respect to two distinct orderings can be as large as  $n - 2$ . Consider the two phylogenetic  $X$ -trees  $T$  and  $T'$  with  $|X| = n + 1$  as shown in Figure ?? . Recall, that we have  $d_{\text{OLA}}^\sigma(T, T') \leq n - 1$  for any ordering  $\sigma$  on  $X$ . Let  $\sigma_1$  be the ordering on  $X$  with

$$\sigma_1(x_n) < \sigma_1(x_{n-1}) < \cdots < \sigma_1(x_1) < \sigma_1(y),$$

and let  $\sigma_{n-1}$  be the ordering on  $X$  with

$$\sigma_{n-1}(x_n) < \sigma_{n-1}(x_{n-1}) < \sigma_{n-1}(y) < \sigma_{n-1}(x_{n-2}) < \sigma_{n-1}(x_{n-3}) < \cdots < \sigma_{n-1}(x_1).$$

It is straightforward to check that  $d_{\text{OLA}}^{\sigma_1}(T, T') = 1$  and  $d_{\text{OLA}}^{\sigma_{n-1}}(T, T') = n - 1$ . In general, for each  $i \in \{1, 2, \dots, n - 1\}$ , there exists an ordering  $\sigma_i$  on  $X$  such that  $d_{\text{OLA}}^{\sigma_i}(T, T') = i$ . More precisely,  $\sigma_i$  is the unique ordering such that

1.  $\sigma_i(y) = n + 1 - (i - 1)$  and
2.  $\sigma_i(x_n) < \sigma_i(x_{n-1}) < \cdots < \sigma_i(x_1)$ .

Hence, for two arbitrary orderings  $\sigma$  and  $\sigma'$  on  $X$ , we have

$$\left| d_{\text{OLA}}^\sigma(T, T') - d_{\text{OLA}}^{\sigma'}(T, T') \right| \leq n - 2,$$

and this bound is sharp.

Third, we briefly turn to the P2V measure of two phylogenetic trees  $T$  and  $T'$  and show that, similar to OLA, there is no clear relationship between  $d_{\text{rSPR}}(T, T')$  and  $d_{\text{P2V}}^*(T, T')$  because  $d_{\text{P2V}}^*(T, T')$  can be strictly

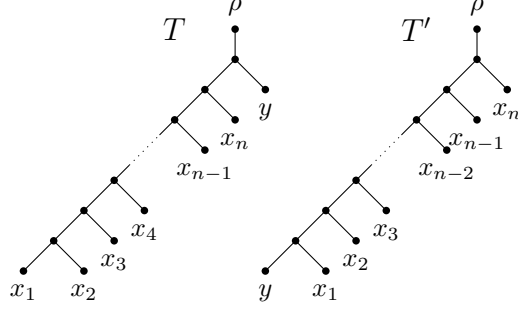


Figure 6: Two phylogenetic  $X$ -trees  $T$  and  $T'$  with  $|X| = n + 1$  for which the OLA distance differs by as much as  $n - 2$ , depending on the ordering.

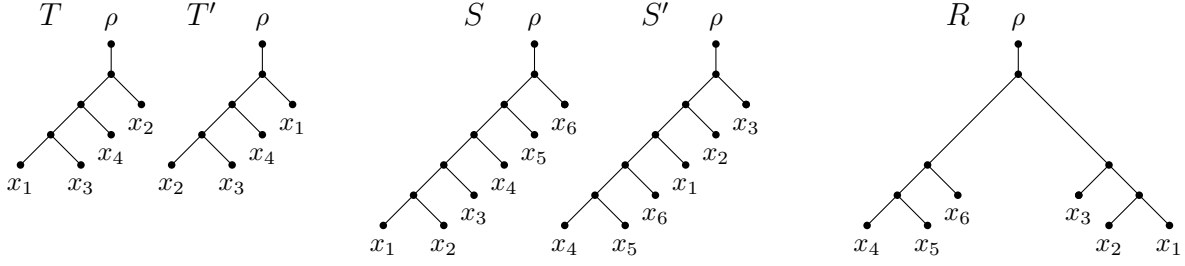


Figure 7: Left: Two phylogenetic  $X$ -trees  $T$  and  $T'$  with  $d_{\text{rSPR}}(T, T') = 2$  and  $d_{\text{P2V}}^*(T, T') = 1$ . Middle: Two phylogenetic  $X$ -trees  $S$  and  $S'$  with  $d_{\text{rSPR}}(S, S') = 2$  and  $d_{\text{P2V}}^*(S, S') = 3$ . Middle and right: Three phylogenetic  $X$ -trees  $S$ ,  $S'$ , and  $R$  with  $h(S, S') = d_{\text{HOP}}^*(S, S') = 3$ ,  $h(S, R) = d_{\text{HOP}}^*(S, R) = 1$ , and  $h(S', R) = d_{\text{HOP}}^*(S', R) = 1$ , showing that the HOP measure is not a distance. Similarly, as  $d_{\text{P2V}}^*(S, R) = d_{\text{P2V}}^*(S', R) = 1$ , whereas  $d_{\text{P2V}}^*(S, S') = 3$ , the P2V measure is also not a distance.

greater or strictly smaller than  $d_{\text{rSPR}}(T, T')$ . Figure ?? shows an example of two phylogenetic trees  $T$  and  $T'$  with  $2 = d_{\text{rSPR}}(T, T') > d_{\text{P2V}}^*(T, T') = 1$ , but also shows an example of two phylogenetic trees  $S$  and  $S'$  with  $2 = d_{\text{rSPR}}(S, S') < d_{\text{P2V}}^*(S, S') = 3$ .

Despite the somewhat negative results presented in this section, we will see in the next section that there is a direct relationship between  $h(T, T')$  and  $d_{\text{HOP}}^*(T, T')$  for any two phylogenetic trees  $T$  and  $T'$  on the same label set. Indeed, we will show that both measures are equivalent.

### 3.2 Bounding by the hybrid number

In this section, we relate each of the three order-dependent measures to the hybrid number. Specifically, for two phylogenetic  $X$ -trees  $T$  and  $T'$ , we establish that  $d_{\text{HOP}}^*(T, T')$  and  $h(T, T')$  are equivalent, which also proves the conjecture of [?, Remark 3, p. 9] concerning the computational complexity of computing  $d_{\text{HOP}}^*(T, T')$ . Further, we show that  $d_{\text{OLA}}^*(T, T')$  and  $d_{\text{P2V}}^*(T, T')$  are bounded above by the hybrid number. However, unlike the HOP measure, there are phylogenetic trees for which the hybrid number is strictly larger than the OLA and P2V measures.

**Theorem 4.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then, for each  $\Theta \in \{\text{HOP}, \text{OLA}, \text{P2V}\}$ ,*

$$d_{\Theta}^*(T, T') \leq h(T, T').$$

We note that the hybrid number is an upper bound on the rSPR distance between two phylogenetic

$X$ -trees and that both measures can differ by up to  $n - \lceil 2\sqrt{n} \rceil$  [?] where  $n = |X|$ . As such, Theorem ?? does not further strengthen the results presented in Section ??.

Now to prove Theorem ??, we require a few additional lemmas. The next two lemmas show that, if two phylogenetic trees  $T$  and  $T'$  with label set  $X$  have a common pendant subtree  $S$ , then the HOP and OLA labelings of  $S$  in  $T$  and  $T'$  are identical for any fixed ordering on  $X$ .

**Lemma 4.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, where  $|X| = n$ , and let  $\sigma$  be an ordering of the elements in  $X$ . Suppose that  $T$  and  $T'$  have a common pendant subtree  $S$ . Then, the HOP labeling of  $S$  under  $\sigma$  is identical for  $T$  and  $T'$ .*

*Proof.* The proof is by induction on  $n$ . If  $n = 1$ , then the result follows immediately.

Now assume that the result holds for all pairs of phylogenetic trees that have at most  $n - 1$  leaves. Let  $x$  be the element in  $X$  such that  $\sigma(x) = n$ . Let  $T_1 = T|(X - \{x\})$  and let  $T'_1 = T'|(X - \{x\})$ . If  $x \in L(S)$ , set  $S_1 = S|(X - \{x\})$  and, otherwise, set  $S_1 = S$ . Since  $S$  is a common pendant subtree of  $T$  and  $T'$ , it follows that  $S_1$  is a common pendant subtree of  $T_1$  and  $T'_1$ . Moreover, by the induction assumption, the HOP labeling of  $S_1$  under  $\sigma_{-x}$  is identical for  $T_1$  and  $T'_1$ . Let  $e$  and  $e'$  be the unique edge in  $T_1$  and  $T'_1$ , respectively, such that  $T$  can be obtained from  $T_1$  by subdividing  $e$  with a new vertex  $u$  and adding the new edge  $(u, x)$  and, similarly,  $T'$  can be obtained from  $T'_1$  by subdividing  $e'$  with a new vertex  $u'$  and adding the new edge  $(u', x)$ . Then, the HOP labeling of  $T$  and  $T'$  can be obtained from that of  $T_1$  and  $T'_1$ , respectively, by setting  $f_{\text{HOP}}^T(x) = f_{\text{HOP}}^{T'}(x) = \sigma(x)$  and  $f_{\text{HOP}}^T(u) = f_{\text{HOP}}^{T'}(u') = \sigma(x)$ , thereby implying that the HOP labeling of  $S$  under  $\sigma$  is identical in  $T$  and  $T'$ .  $\square$

**Lemma 5.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, where  $|X| = n$ , and let  $\sigma$  be an ordering of the elements in  $X$ . Furthermore, let  $\mathbf{v}$  and  $\mathbf{v}'$  be the OLA vectors for  $T$  and  $T'$ , respectively, under  $\sigma$ . Suppose that  $T$  and  $T'$  have a common pendant subtree  $S$ . Then, the OLA labeling of  $S$  under  $\sigma$  is identical for  $T$  and  $T'$ . Moreover,  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(S)$  have Hamming distance at most 1.*

*Proof.* We establish both parts using induction on  $n$ . If  $n = 1$ , then the result follows since  $T \simeq T'$ .

Now assume that  $n > 1$  and that the result holds for all pairs of phylogenetic  $X$ -trees that have at most  $n - 1$  leaves. Since the lemma holds whenever  $|L(S)| = 1$ , we may assume for the remainder of the proof that  $|L(S)| \geq 2$ . Let  $x$  be the element in  $X$  such that  $\sigma(x) = n$ . Let  $T_1 = T|(X - \{x\})$ , let  $T'_1 = T'|(X - \{x\})$ , and let  $S_1 = S|(L(S) - \{x\})$ . Observe that  $S_1 = S$  if  $x \notin L(S)$ . Since  $S$  is a common pendant subtree of  $T$  and  $T'$ , it follows that  $S_1$  is a common pendant subtree of  $T_1$  and  $T'_1$ . Let  $\mathbf{v}_1 = [v_1, v_2, \dots, v_{n-1}]$  and  $\mathbf{v}'_1 = [v'_1, v'_2, \dots, v'_{n-1}]$  be the OLA vectors for  $T_1$  and  $T'_1$ , respectively, under  $\sigma_{-x}$ . Then the following two statements follow from the induction assumption. First, the OLA labeling of  $S_1$  is identical for  $T_1$  and  $T'_1$  under  $\sigma_{-x}$ . Second,  $\mathbf{v}_1$  and  $\mathbf{v}'_1$  restricted to the elements in  $L(S_1)$  have Hamming distance at most 1. Now, let  $e = (u, w)$  and  $e' = (u', w')$  be the unique edge in  $T_1$  and  $T'_1$ , respectively, such that  $T$  can be obtained from  $T_1$  by subdividing  $e$  with a new vertex  $v$  and adding the new edge  $(v, x)$  and, similarly,  $T'$  can be obtained from  $T'_1$  by subdividing  $e'$  with a new vertex  $v'$  and adding the new edge  $(v', x)$ . We can view the vertex set of  $T$  as the union of the vertex set of  $T_1$  and  $\{v, x\}$  and, similarly for  $T'$ . The OLA labeling of  $T$  and  $T'$  under  $\sigma$  can be obtained from that of  $T_1$  and  $T'_1$ , respectively, under  $\sigma_{-x}$  by setting  $f_{\text{OLA}}^T(x) = f_{\text{OLA}}^{T'}(x) = n$ , setting  $f_{\text{OLA}}^T(v) = f_{\text{OLA}}^{T'}(v') = -\sigma(x)$  and, for each vertex  $t$  (resp.  $t'$ ) in  $T_1$  (resp.  $T'_1$ ), setting  $f_{\text{OLA}}^T(t) = f_{\text{OLA}}^{T_1}(t)$  (resp.  $f_{\text{OLA}}^{T'}(t') = f_{\text{OLA}}^{T'_1}(t')$ ). Since  $S$  is a common pendant subtree of  $T$  and  $T'$  and the OLA labeling of  $S_1$  under  $\sigma_{-x}$  is identical for  $T_1$  and  $T'_1$ , it follows that the OLA labeling of  $S$  under  $\sigma$  is also identical in  $T$  and  $T'$ . Now, by definition of  $\mathbf{v}$  and  $\mathbf{v}'$ , recall that  $\mathbf{v} = [v_1, v_2, \dots, v_{n-1}, v_n]$  and  $\mathbf{v}' = [v'_1, v'_2, \dots, v'_{n-1}, v'_n]$ . Furthermore, since the coordinate is determined by the sibling,  $v_n = f_{\text{OLA}}^T(w)$  and  $v'_n = f_{\text{OLA}}^{T'}(w')$ . If  $x \notin L(S)$ , then it immediately follows that  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(S)$  also have Hamming distance at most 1. On the other hand, if  $x \in L(S)$ , then  $f_{\text{OLA}}^T(w) = f_{\text{OLA}}^{T'}(w')$  since the OLA labeling of  $S$  under  $\sigma$  is identical in  $T$  and  $T'$ . Thus, as  $\mathbf{v}_1$  and  $\mathbf{v}'_1$  restricted to the elements in  $L(S_1)$  have Hamming distance at most 1, it again follows that  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(S)$  also have Hamming distance at most 1.  $\square$

The analogue of Lemma ?? for P2V does not hold. However, the following weaker version holds.

**Lemma 6.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, where  $|X| = n$ . Suppose that  $T$  and  $T'$  have a common pendant subtree  $S$ . Let  $\sigma$  be an ordering of the elements in  $X$  such that the elements in  $X - L(S)$  are bijectively mapped to the elements in  $\{1, 2, \dots, n - |L(S)|\}$ . Let  $\mathbf{v}$  and  $\mathbf{v}'$  be the P2V vectors for  $T$  and  $T'$ , respectively, under  $\sigma$ . Then the P2V labeling of  $S$  under  $\sigma$  is identical for  $T$  and  $T'$ . Moreover,  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(S)$  have Hamming distance at most 1.*

*Proof.* We first show that the P2V labeling of  $S$  under  $\sigma$  is identical for  $T$  and  $T'$ . By the choice of  $\sigma$ , there exists an ordering,  $t_1, t_2, \dots, t_{|L(S)|-1}$ , on the internal vertices of  $T|L(S)$  such that  $f_{\text{P2V}}^T(t_i) = n + i$  for each  $i \in \{1, 2, \dots, |L(S)| - 1\}$ . Since  $S$  is a common pendant subtree of  $T$  and  $T'$  it is now straightforward to check that the P2V labeling of  $S$  under  $\sigma$  is identical for  $T$  and  $T'$ .

To complete the proof, we show by induction on  $|L(S)|$  that  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(S)$  have Hamming distance at most 1. If  $|L(S)| = 1$ , then the result clearly follows.

Now assume that  $|L(S)| > 1$  and that the result holds for all pairs of phylogenetic trees that have a common pendant subtree whose size is strictly less than  $|L(S)|$ . Let  $x$  be the element in  $X$  such that  $\sigma(x) = n$ . By the choice of  $\sigma$ , we have  $x \in L(S)$ . Let  $T_1 = T|(X - \{x\})$ , let  $T'_1 = T'|(X - \{x\})$ , and let  $S_1 = S|(L(S) - \{x\})$ . Since  $S$  is a common pendant subtree of  $T$  and  $T'$  and  $|L(S) - \{x\}| \geq 1$ , it follows that  $S_1$  is a common pendant subtree of  $T_1$  and  $T'_1$ . Let  $\mathbf{v}_1 = [v_1, v_2, \dots, v_{n-1}]$  and  $\mathbf{v}'_1 = [v'_1, v'_2, \dots, v'_{n-1}]$  be the P2V vectors for  $T_1$  and  $T'_1$ , respectively, under  $\sigma_{-x}$ . Then, by the induction assumption, we have that  $\mathbf{v}_1$  and  $\mathbf{v}'_1$  restricted to the elements in  $L(S_1)$  have Hamming distance at most 1. Now, let  $e = (u, w)$  and  $e' = (u', w')$  be the unique edge in  $T_1$  and  $T'_1$ , respectively, such that  $T$  can be obtained from  $T_1$  by subdividing  $e$  with a new vertex  $v$  and adding the new edge  $(v, x)$  and, similarly,  $T'$  can be obtained from  $T'_1$  by subdividing  $e'$  with a new vertex  $u'$  and adding the new edge  $(u', x)$ . Then  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  and  $\mathbf{v}' = [v'_1, v'_2, \dots, v'_n]$  with  $v_n = f_{\text{P2V}}^T(w)$  and  $v'_n = f_{\text{P2V}}^{T'}(w')$ . Since  $w$  and  $w'$  is a vertex of  $T|L(S)$  and  $T'|L(S)$ , respectively, and the P2V labeling of  $S$  is identical for  $T$  and  $T'$  under  $\sigma$ , we have  $v_n = v'_n = f_{\text{P2V}}^T(w)$ . Hence, as  $\mathbf{v}_1$  and  $\mathbf{v}'_1$  restricted to the elements in  $L(S_1)$  have Hamming distance at most 1, it now follows that  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(S)$  also have Hamming distance at most 1.  $\square$

We are now in the position to prove Theorem ??.

*Proof of Theorem ??.* The proof is by induction on  $n$ . If  $n \leq 2$ , then  $T$  is isomorphic to  $T'$ , and the statement immediately holds.

Suppose that  $n \geq 3$  and that the statement holds for all pairs of phylogenetic trees with at most  $n - 1$  leaves. Let  $F = \{T_\rho, T_1, T_2, \dots, T_k\}$  be an acyclic agreement forest for  $T$  and  $T'$ . Since  $F$  is acyclic, there is a tree in  $F$  that is pendant in  $T$  and  $T'$ . Without loss of generality, we may assume that this tree is  $T_k$ . Let  $n_k = |L(T_k)|$ , let  $T_1 = T|(X - L(T_k))$ , and let  $T'_1 = T'|(X - L(T_k))$ . Furthermore, let  $F_1 = \{T_\rho, T_1, T_2, \dots, T_{k-1}\}$ . Since  $F$  is an acyclic agreement forest for  $T$  and  $T'$ , it follows that  $F_1$  is an acyclic agreement forest for  $T_1$  and  $T'_1$ . By the induction assumption and Theorem ??, there is an ordering  $\sigma_1$  on  $X - L(T_k)$  such that

$$d_{\Theta}^*(T_1, T'_1) \leq d_{\Theta}^{\sigma_1}(T_1, T'_1) \leq k - 1. \quad (1)$$

Let  $\sigma$  be an ordering on  $X$  such that  $\sigma(x) = \sigma_1(x)$  for each  $x \in X - L(T_k)$ . Then  $\sigma(y) > n - n_k$  for each  $y \in L(T_k)$ . Let  $p$  and  $p'$  denote the parent of the root of  $T_k$  in  $T$  and  $T'$ , respectively.

We next complete the induction step for  $\Theta = \text{HOP}$ . Since  $T_k$  is a pendant subtree of  $T$  and  $T'$  it follows that  $p$  and  $p'$  have the same label which, by construction, is  $n - n_k + 1$ . Moreover, by Lemma ??, the HOP labeling of the vertices of  $T$  and  $T'$  that correspond to  $T_k$  under  $\sigma$  are identical. Hence each element in  $\{n - n_k + 2, n - n_k + 3, \dots, n\}$  is contained in a longest common subsequence of the HOP sequences for  $T$  and  $T'$ . It now follows that

$$d_{\text{HOP}}^*(T, T') \leq d_{\text{HOP}}^{\sigma}(T, T') \leq d_{\text{HOP}}^{\sigma_1}(T_1, T'_1) + 1 \leq (k - 1) + 1 = k,$$



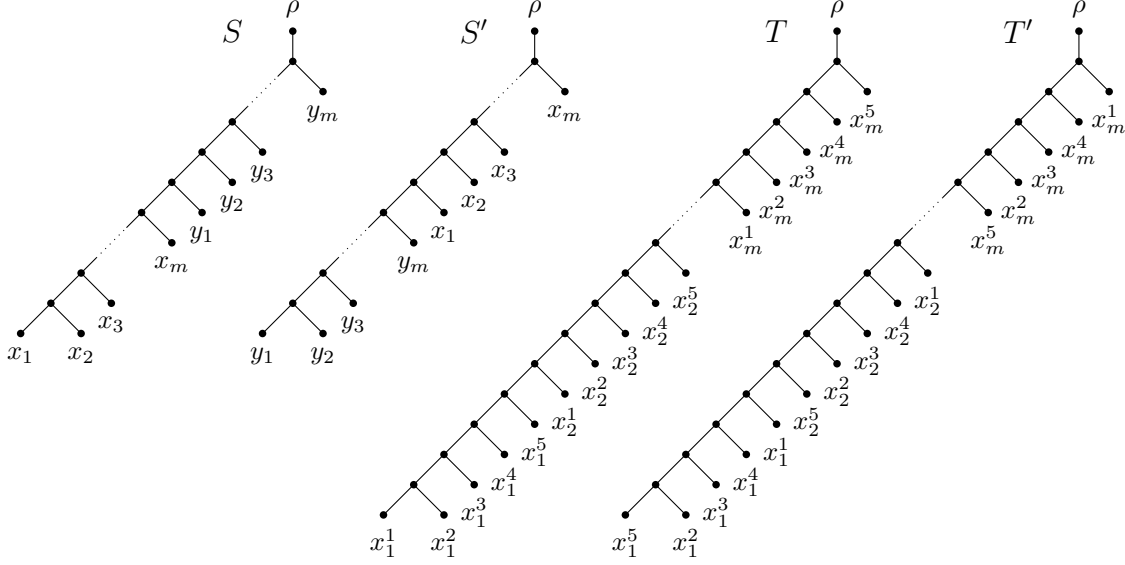


Figure 8: Left: Two phylogenetic trees  $S$  and  $S'$  on  $2m$  leaves. For  $m \geq 3$ , it follows that  $d_{\text{OLA}}^*(S, S') < h(S, S')$  since  $h(S, S') = m$  and  $d_{\text{OLA}}^*(S, S') \leq 2$ . Right: Two phylogenetic trees  $T$  and  $T'$  on  $5m$  leaves. For  $m \geq 1$ , it follows that  $d_{\text{P2V}}^*(T, T') < h(T, T')$  since  $h(T, T') = 2m$  and  $d_{\text{P2V}}^*(T, T') \leq m$ . For further details, see text.

where the second inequality is strict if  $n - n_k + 1$  is an element of a longest common subsequence of the HOP sequences for  $T$  and  $T'$ . Choosing  $F$  to be a maximum acyclic agreement forest establishes the lemma for  $\Theta = \text{HOP}$ .

Next let  $\Theta \in \{\text{OLA}, \text{P2V}\}$ , and let  $\mathbf{v}_1 = [v_1, v_2, \dots, v_{n-n_k}]$  and  $\mathbf{v}'_1 = [v'_1, v'_2, \dots, v'_{n-n_k}]$  be the  $\Theta$  vectors for  $T_1$  and  $T'_1$ , respectively, under  $\sigma_1$ . It follows from Equation (??) that the Hamming distance between  $\mathbf{v}_1$  and  $\mathbf{v}'_1$  is at most  $k - 1$ . Moreover, by the choice of  $\sigma$ , the  $\Theta$  vectors for  $T$  and  $T'$  are  $\mathbf{v} = [v_1, v_2, \dots, v_{n-n_k}, v_{n-n_k+1}, \dots, v_n]$  and  $\mathbf{v}' = [v'_1, v'_2, \dots, v'_{n-n_k}, v'_{n-n_k+1}, \dots, v'_n]$ . Hence, by Lemma ?? for  $\Theta = \text{OLA}$  and Lemma ?? for  $\Theta = \text{P2V}$ ,  $\mathbf{v}$  and  $\mathbf{v}'$  restricted to  $L(T_k)$  have Hamming distance at most 1. In turn, this implies that

$$d_{\Theta}^*(T, T') \leq d_{\Theta}^{\sigma}(T, T') = d_{\Theta}^{\sigma_1}(T_1, T'_1) + 1 \leq (k - 1) + 1 = k$$

Again choosing  $F$  to be a maximum acyclic agreement forest establishes the lemma for  $\Theta \in \{\text{OLA}, \text{P2V}\}$ .  $\square$

To see that the inequality of Theorem ?? can be strict for trees of arbitrary size under OLA and P2V, Figure ?? shows two phylogenetic trees  $S$  and  $S'$  on  $2m$  leaves such that, for each  $m \geq 3$ ,  $d_{\text{OLA}}^*(S, S') < h(S, S')$ , and also shows two phylogenetic trees  $T$  and  $T'$  on  $5m$  leaves such that, for each  $m \geq 1$ ,  $d_{\text{P2V}}^*(T, T') < h(T, T')$ . First, consider the ordering

$$\sigma(x_1) < \sigma(x_2) < \dots < \sigma(x_m) < \sigma(y_1) < \sigma(y_2) < \dots < \sigma(y_m)$$

on the label set of  $S$  and  $S'$ . Then

$$\begin{aligned} \mathbf{v} &= [0, 1, -2, -3, \dots, -(m-1), -m, -(m+1), -(m+2), -(m+3), \dots, -(2m-1)], \\ \mathbf{v}' &= [0, 1, -2, -3, \dots, -(m-1), 1, m+1, -(m+2), -(m+3), \dots, -(2m-1)], \end{aligned}$$

where  $\mathbf{v}$  and  $\mathbf{v}'$  is the OLA vector for  $S$  and  $S'$ , respectively, under  $\sigma$  and, thus,  $d_{\text{OLA}}^*(S, S') \leq d_{\text{OLA}}^{\sigma}(S, S') = 2$ . On the other hand, a maximum acyclic agreement forest for  $S$  and  $S'$  has size  $m + 1$  and, thus, by

Theorem ??, we have  $h(S, S') = m$ . Second, consider the ordering

$$\sigma(x_1^1) < \sigma(x_1^2) < \sigma(x_1^3) < \sigma(x_1^4) < \sigma(x_1^5) < \dots < \sigma(x_m^1) < \sigma(x_m^2) < \sigma(x_m^3) < \sigma(x_m^4) < \sigma(x_m^5)$$

on the label set of  $T$  and  $T'$ . Then

$$\begin{aligned} \mathbf{w} &= [0, 1, 1, 5, 7, 10, 12, 13, 15, 17, \dots, \\ &\quad 10(m-1), 10(m-1)+2, 10(m-1)+3, 10(m-1)+5, 10(m-1)+7], \\ \mathbf{w}' &= [0, 1, 2, 5, 7, 10, 11, 13, 15, 17, \dots, \\ &\quad 10(m-1), 10(m-1)+1, 10(m-1)+3, 10(m-1)+5, 10(m-1)+7], \end{aligned}$$

where  $\mathbf{w}$  and  $\mathbf{w}'$  is the P2V vector for  $T$  and  $T'$ , respectively, under  $\sigma$  and, thus,  $d_{\text{P2V}}^*(T, T') \leq d_{\text{P2V}}^\sigma(T, T') = m$ . On the other hand, a maximum acyclic agreement forest for  $S$  and  $S'$  has size  $2m+1$  and, thus, by Theorem ??, we have  $h(T, T') = 2m$ .

We now establish the equivalence of  $d_{\text{HOP}}^*(T, T')$  and  $h(T, T')$  for two phylogenetic trees  $T$  and  $T'$ :

**Theorem 5.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then*

$$d_{\text{HOP}}^*(T, T') = h(T, T').$$

In Theorem ??, we already established that the HOP measure  $d_{\text{HOP}}^*(T, T')$  for two phylogenetic trees  $T$  and  $T'$  is bounded above by  $h(T, T')$ . The next lemma shows that  $d_{\text{HOP}}^*(T, T')$  is also bounded below by  $h(T, T')$ .

**Lemma 7.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, where  $|X| = n$ . Then*

$$d_{\text{HOP}}^*(T, T') \geq h(T, T').$$

*Proof.* Let  $\sigma$  be an ordering on  $X$ . Consider the following process: Let  $\mathbf{v}_T^\sigma$  and  $\mathbf{v}_{T'}^\sigma$  be the HOP vectors for  $T$  and  $T'$ , respectively. Let  $\text{LCS}$  denote  $\text{LCS}(\mathbf{v}_T^\sigma, \mathbf{v}_{T'}^\sigma)$ . Set  $i = 1$ . For each element on the path from the non-leaf vertex  $v$  with  $f_{\text{HOP}}(v) = i$  to the element  $x \in X$  with  $\sigma(x) = f_{\text{HOP}}(x) = i$  in  $T$  that is not in  $\text{LCS}$ , delete the outgoing edge not on this path. Increment  $i$  by 1 and repeat. Continue this process until  $i = n + 1$ . Let  $F = \{T_\rho, T_1, T_2, \dots, T_k\}$  denote the resulting forest. Note that  $k = n - |\text{LCS}|$ . Apply this same process to  $T'$  to get the forest  $F' = \{T'_\rho, T'_1, T'_2, \dots, T'_k\}$ .

We complete the proof by showing that  $F$  is an acyclic agreement forest for  $T$  and  $T'$ . To this end, we use induction on  $n$  to first show that  $F$  is isomorphic to  $F'$ , that is,  $F$  is an agreement forest for  $T$  and  $T'$  before establishing that  $F$  is also acyclic. If  $n \in \{1, 2\}$ , then  $T$  is isomorphic to  $T'$ , and so  $F$  is an acyclic agreement forest for  $T$  and  $T'$ .

Now suppose that the statement holds for all pairs of phylogenetic  $X$ -trees with  $|X| \leq n-1$ , where  $n \geq 3$ . Let  $x$  be the element in  $X$  with  $\sigma(x) = n$ , and recall that we use  $\sigma_{-x}$  to denote the ordering on  $X - \{x\}$  such that  $\sigma_{-x}(y) = \sigma(y)$  for each element  $y \in X - \{x\}$ . Let  $T_1 = T|(X - \{x\})$  and  $T'_1 = T'|(X - \{x\})$ , and let  $\text{LCS}_1$  denote  $\text{LCS}(\mathbf{v}_{T_1}^{\sigma'}, \mathbf{v}_{T'_1}^{\sigma'})$ , and  $\mathbf{v}_{T_1}^{\sigma'}$  and  $\mathbf{v}_{T'_1}^{\sigma'}$  are the respective HOP vectors of  $T_1$  and  $T'_1$  under  $\sigma' = \sigma_{-x}$ . Let  $F_1 = \{S_\rho, S_1, S_2, \dots, S_l\}$  denote the forest obtained from  $T_1$  by applying the process that is described in the first paragraph of the proof. By the induction assumption,  $F_1$  is an acyclic agreement forest for  $T_1$  and  $T'_1$ .

Before continuing, we make the following observations: (i)  $\text{LCS}_1 = \text{LCS}|(X - \{x\})$ , where  $\text{LCS}|(X - \{x\})$  refers to the vector obtained from  $\text{LCS}$  by deleting the coordinates corresponding to  $x$  (i.e., by deleting the coordinates  $\sigma(x)$  and  $\sigma(x)$ ), and (ii) the parent of  $x$  is labeled by  $n$  for each of  $T$  and  $T'$ . Thus, (iii)  $F_1$  is isomorphic to  $F|(X - \{x\}) = \{T_\rho|(X - \{x\}), T_1|(X - \{x\}), \dots, T_k|(X - \{x\})\}$  and  $F_1$  is isomorphic to  $F'|(X - \{x\}) = \{T'_\rho|(X - \{x\}), T'_1|(X - \{x\}), \dots, T'_k|(X - \{x\})\}$ .

We break the remainder of the proof into two cases depending on whether  $n \in \text{LCS}$ . Let  $i$  be the element of  $\{1, 2, \dots, n-1\}$  such that the subsequence  $[i, \dots, \underline{i}]$  of  $\mathbf{v}_T^\sigma$  contains  $n$ , and let  $i'$  be the element of  $\{1, 2, \dots, n-1\}$  such that the subsequence  $[i', \dots, \underline{i}']$  of  $\mathbf{v}_{T'}^\sigma$  contains  $n$ . To ease reading, we use  $j$  to refer to the internal vertex  $v$  with  $f_{\text{HOP}}(v) = j$ , and  $\underline{j}$  to refer to the element  $x \in X$  with  $\sigma(x) = j$ .

First assume that  $n \notin \text{LCS}$ . Then the outgoing edge of the internal vertex of  $T$  labeled  $n$  not on the path from  $i$  to  $\underline{i}$  is cut. Similarly, the outgoing edge of the internal vertex of  $T'$  labeled  $n$  not on the path from  $i'$  to  $\underline{i}'$  is cut. By (ii) and (iii),  $F$  is an agreement forest for  $T$  and  $T'$  and, as  $n$  is an isolated vertex,  $F$  is an acyclic agreement forest for  $T$  and  $T'$ .

Second assume that  $n \in \text{LCS}$ . Then  $i = i'$  and the outgoing edge of the internal vertex labeled  $n$  in  $T$  (resp.  $T'$ ) not on the path from  $i$  to  $i'$  is not cut. Thus,  $\underline{n}$  is in the same component as  $\underline{i}$  in  $F$  and  $F'$ . Let  $T_i$  and  $T'_i$  denote these components in  $F$  and  $F'$ , respectively. We next show that  $T_i$  is isomorphic to  $T'_i$ , thereby showing, by (iii), that  $F$  is an agreement forest for  $T$  and  $T'$ . Keeping the corresponding internal labels of  $T$  and  $T'$ , consider the paths in  $T_i$  and  $T'_i$  from each of their roots to  $\underline{i}$ . By construction, the vertex labels on these paths are identical. Furthermore, for each of these vertices, if its label is  $j$ , then this vertex is the least common ancestor of  $\underline{j}$  and  $\underline{i}$  in  $T_i$  and  $T'_i$ . Hence,  $T_i$  is isomorphic to  $T'_i$ , and so  $F$  is an agreement forest for  $T$  and  $T'$ . Lastly, we show that  $F$  is acyclic. To this end, let  $G$  denote the directed graph associated with  $F$  as detailed in the definition of an acyclic agreement forest. For each vertex in  $G$ , label it with the least leaf value in the corresponding component. Assume that  $(i, j)$  is a directed edge in  $G$ . Then  $i < j$ . To see this, we may assume without loss of generality that the root of  $T_i$  is an ancestor of the root of  $T_j$  in  $T$ . Since  $i$  is the least value in  $L(T_i)$ , the outgoing edge of the vertex labeled  $i$  in  $T$  that is on the path from  $i$  to  $\underline{i}$  is cut. Similarly, the outgoing edge of the vertex labeled  $j$  in  $T$  that is on the path from  $j$  to  $\underline{j}$  is cut. But then, for the root of  $T_i$  to be an ancestor of the root of  $T_j$  in  $T$ , there must be a path from  $i$  to  $\underline{j}$  that contains  $j$  and each of these two cut edges. In turn, this implies  $i < j$ . Returning to  $G$ , as  $\sigma$  is an ordering on  $X$ , it now follows that  $G$  contains no (directed) cycles. Hence  $F$  is an acyclic agreement forest for  $T$  and  $T'$ . Since  $F$  is an acyclic agreement forest for  $T$  and  $T'$ , Theorem ?? implies that  $h(T, T') \leq n - |\text{LCS}|$ , and choosing  $\sigma$  to be an ordering realizing  $d_{\text{HOP}}^*(T, T')$  completes the proof.  $\square$

Finally, to establish Theorem ??, we combine Theorem ?? and Lemma ??. Since computing the hybrid number for two phylogenetic trees is NP-hard [? ], the next corollary is an immediate consequence of Theorem ?? and answers the open problem in [? , page 10, Remark 2].

**Corollary 1.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Computing  $d_{\text{HOP}}^*(T, T')$  is NP-hard.*

Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, and let  $\sigma$  be an ordering on  $X$ . In the language of this paper, Zhang et al. [? ] used certain shortest common supersequences of the HOP vectors of  $T$  and  $T'$  under  $\sigma$  to construct a tree-child network that embeds  $T$  and  $T'$ . Furthermore, in Proposition 3 of the supplementary material of [? ], the authors show that repeating the construction process for each ordering on  $X$  can be used to compute  $h_{tc}(T, T')$ . Hence, since  $h(T, T') = h_{tc}(T, T')$  by Lemma ??, it follows from Theorem ?? that computing  $d_{\text{HOP}}^*(T, T')$  is equivalent to the shortest common supersequence approach of [? ].

## 4 Bounding order-dependent measures by cherry-picking sequences

In this section, we provide sufficient conditions when all three distances  $d_{\text{OLA}}^\sigma(T, T')$ ,  $d_{\text{P2V}}^\sigma(T, T')$ , and  $d_{\text{HOP}}^\sigma(T, T')$  are equal and show a connection to the temporal tree-child hybrid number. Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, and let  $S = (x_1, x_2, \dots, x_n)$  be a sequence of the elements in  $X$ . We call  $S$  a *cherry-picking sequence* for  $T$  precisely if each  $x_i$  with  $i \in \{1, 2, \dots, n-1\}$  labels a leaf of a cherry in  $T|(X - \{x_1, x_2, \dots, x_{i-1}\})$ . Furthermore, if  $S$  is a cherry-picking sequence for  $T$  and  $T'$ , then we say that  $S$  is a *common cherry-picking sequence* for  $T$  and  $T'$ .

Let  $S = (x_1, x_2, \dots, x_n)$  be a common cherry-picking sequence of two phylogenetic  $X$ -trees  $T$  and  $T'$ . We say that  $x_i$  with  $i \in \{1, 2, \dots, n\}$  *agrees* if  $i = n$ , or  $x_i$  is a leaf of a cherry  $\{x_i, y\}$  in  $T|(X - \{x_1, x_2, \dots, x_{i-1}\})$  and  $\{x_i, y\}$  is also a cherry in  $T'|(X - \{x_1, x_2, \dots, x_{i-1}\})$ ; otherwise, we say that  $x_i$  *disagrees*. The *weight* of  $S$ ,  $wt(S)$ , is

$$wt(S) = |\{x_i \in S : i \in \{1, 2, \dots, n-1\} \text{ and } x_i \text{ disagrees}\}|.$$

Lastly, we call  $S$  a *minimum common cherry-picking sequence* for  $T$  and  $T'$  if  $wt(S)$  is minimized over all common cherry-picking sequences for  $T$  and  $T'$ . This minimum number is denoted by  $s(T, T')$ . The next theorem is established in [?, Theorem 2], where it is also shown that if  $T$  and  $T'$  have no common cherry-picking sequence, then there is no temporal tree-child network that displays  $T$  and  $T'$  [?, Theorem 1].

**Theorem 6.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees. Then  $h_t(T, T') = s(T, T')$ .*

Consider two phylogenetic trees  $T$  and  $T'$  that have a common cherry-picking sequence (a “natural ordering” in the terminology of [? ])  $S = (x_1, x_2, \dots, x_n)$ . Throughout this section, we refer to the ordering  $\sigma$  with  $\sigma(x_i) = n - i + 1$  for each  $i \in \{1, 2, \dots, n\}$  as the ordering on  $X$  that is *induced* by  $S$ . The main result of this section is the following theorem.

**Theorem 7.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, and suppose that  $S = (x_1, x_2, \dots, x_n)$  is a common cherry-picking sequence for  $T$  and  $T'$ . Let  $\sigma$  be the ordering on  $X$  that is induced by  $S$ . Then,*

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{HOP}}^\sigma(T, T') = d_{\text{P2V}}^\sigma(T, T') = wt(S).$$

We start with a lemma about OLA and P2V vectors. The ordering associated with a cherry-picking sequence corresponds to a tree-growing process that, at every step, attaches the new leaf to a pendant edge to create a cherry. As such, the resulting vector consists of labels solely from leaves (and not internal vertices).

**Lemma 8.** *Let  $T$  be a phylogenetic  $X$ -tree with  $|X| = n$ , and let  $S = (x_1, x_2, \dots, x_n)$  be a cherry-picking sequence for  $T$ . Further, let  $\sigma$  be the ordering on  $X$  that is induced by  $S$ . Let  $\mathbf{u} = [u_1, u_2, \dots, u_n]$  be the OLA vector of  $T$  under  $\sigma$ , and let  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  be the P2V vector of  $T$  under  $\sigma$ . Then  $\mathbf{u} = \mathbf{v}$ , i.e., the OLA and P2V vectors of  $T$  are identical under  $\sigma$ .*

*Proof.* The proof is by induction on  $n$ . If  $n = 1$ , then  $\mathbf{u} = \mathbf{v} = [0]$ , and if  $n = 2$ , then  $\mathbf{u} = \mathbf{v} = [0, 1]$ , and the statement immediately holds.

Suppose that  $n \geq 3$  and that the statement holds for all phylogenetic trees with at most  $n - 1$  leaves. Let  $T$  be a phylogenetic  $X$ -tree with  $|X| = n$  and cherry-picking sequence  $S = (x_1, x_2, \dots, x_n)$ . By assumption,  $\sigma(x_1) = n$ . Further,  $x_1$  is part of a cherry, say  $\{c, x_1\}$ , in  $T$ . Let  $T_1 = T|(X - \{x_1\})$ . By the induction assumption, for the OLA and P2V vectors, say  $\mathbf{u}' = [u'_1, u'_2, \dots, u'_{n-1}]$  and  $\mathbf{v}' = [v'_1, v'_2, \dots, v'_{n-1}]$ , of  $T_1$ , we have  $\mathbf{u}' = \mathbf{v}'$ . Since  $\{c, x_1\}$  is a cherry in  $T$ , the OLA and P2V vectors, say  $\mathbf{u}$  and  $\mathbf{v}$ , for  $T$  under  $\sigma$  can be obtained from  $\mathbf{u}'$  and  $\mathbf{v}'$  by setting  $\mathbf{u} = [u'_1, u'_2, \dots, u'_{n-1}, \sigma(c)]$  and  $\mathbf{v} = [v'_1, v'_2, \dots, v'_{n-1}, \sigma(c)]$ , where  $\sigma(c)$  is the rank of  $c$  under  $\sigma$ . Since by the induction assumption  $u'_i = v'_i$  for each  $i \in \{1, 2, \dots, n - 1\}$ , clearly  $\mathbf{u} = \mathbf{v}$ , which completes the proof.  $\square$

While the coordinates in the OLA and P2V vectors of a tree disagree when a leaf is attached to an internal edge in the tree-generating process, the coordinates coincide when a leaf is attached to a pendant edge. Hence a consequence of Lemma ?? is the next corollary.

**Corollary 2.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees with  $|X| = n$ . Suppose that  $S = (x_1, x_2, \dots, x_n)$  is a common cherry-picking sequence for  $T$  and  $T'$ . Then there exists an ordering  $\sigma$  on  $X$  such that*

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{P2V}}^\sigma(T, T').$$

We next show that if  $S$  is a common cherry-picking sequence of two phylogenetic  $X$ -trees  $T$  and  $T'$ , and  $\sigma$  is the ordering on  $X$  induced by  $S$ , then the OLA and HOP distances between  $T$  and  $T'$  under  $\sigma$  are the same and equal to  $wt(S)$ .

**Lemma 9.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees with  $|X| = n$ . Suppose that  $S = (x_1, x_2, \dots, x_n)$  is a common cherry-picking sequence for  $T$  and  $T'$  of weight  $wt(S)$ . Let  $\sigma$  be the ordering on  $X$  that is induced by  $S$ . Then*

$$d_{\text{OLA}}^\sigma(T, T') = d_{\text{HOP}}^\sigma(T, T') = wt(S).$$

*Proof.* The proof is by induction on  $n$ . If  $n \in \{1, 2\}$ , then  $T$  and  $T'$  are isomorphic, so  $d_{\text{OLA}}^\sigma(T, T') = d_{\text{HOP}}^\sigma(T, T') = 0$ . Further,  $wt(S) = 0$ .

Now assume that  $n > 2$  and that the result holds for all pairs of phylogenetic  $X$ -trees with a common cherry-picking sequence that have at most  $n - 1$  leaves.

Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees, where  $|X| = n$ , that have a common cherry-picking sequence  $S = (x_1, x_2, \dots, x_n)$  and let  $\sigma$  be the ordering on  $X$  that is induced by  $S$ . By assumption,  $\sigma(x_1) = n$ . Let  $T_1 = T|(X - \{x_1\})$ , and let  $T'_1 = T'|(X - \{x_1\})$ . Let  $\mathbf{u}_1 = [u_1, u_2, \dots, u_{n-1}]$  and  $\mathbf{u}'_1 = [u'_1, u'_2, \dots, u'_{n-1}]$  be the OLA vectors for  $T_1$  and  $T'_1$  under  $\sigma_{-x_1}$ , and let  $\mathbf{v}^1 = [1, \mathbf{v}_1^1, \underline{1}, \mathbf{v}_2^1, \underline{v_2}, \dots, \mathbf{v}_{n-2}^1, \underline{v_{n-2}}, \underline{v_{n-1}}]$  and  $\tilde{\mathbf{v}}^1 = [1, \tilde{\mathbf{v}}_1^1, \underline{1}, \tilde{\mathbf{v}}_2^1, \underline{\tilde{v}_2}, \dots, \tilde{\mathbf{v}}_{n-2}^1, \underline{\tilde{v}_{n-2}}, \underline{\tilde{v}_{n-1}}]$  be the HOP vectors for  $T_1$  and  $T'_1$  under  $\sigma_{-x_1}$ . Let  $S_1 = (x_2, x_3, \dots, x_n)$  be the common cherry-picking sequence for  $T_1$  and  $T'_1$  obtained from  $S$  by omitting  $x_1$ . By the induction assumption,

$$d_{\text{OLA}}^{\sigma_{-x_1}}(T_1, T'_1) = d_{\text{HOP}}^{\sigma_{-x_1}}(T_1, T'_1) = wt(S_1). \quad (2)$$

Now, by assumption  $x_1$  is in a cherry,  $\{c, x_1\}$  say, in  $T$ , and in a cherry,  $\{c', x_1\}$  say, in  $T'$ . As in the proof of Lemma ??, this implies that the OLA vectors, say  $\mathbf{u}$  and  $\mathbf{u}'$  for  $T$  and  $T'$  under  $\sigma$  can be obtained from  $\mathbf{u}_1$  and  $\mathbf{u}'_1$  by setting  $\mathbf{u} = [u_1, u_2, \dots, u_{n-1}, \sigma(c)]$  and  $\mathbf{u}' = [u'_1, u'_2, \dots, u'_{n-1}, \sigma(c')]$ . Further, the HOP labeling of  $T$  and  $T'$  is obtained from the HOP labeling for  $T_1$  and  $T'_1$  by assigning label  $n$  to the vertex that is introduced when adjoining leaf  $x_1$  to  $T_1$  (resp.  $T'_1$ ), while keeping all other labels unchanged. Referring to Algorithm ?? (Lines 15–17) this implies that for  $T$ , the vertex labeled  $n$  is the last element in  $\mathbf{v}(P_{\sigma(c)})$ , and for  $T'$ , it is the last element in  $\mathbf{v}(P_{\sigma(c')})$ . Thus, the HOP vectors, say  $\mathbf{v}$  and  $\mathbf{v}'$ , for  $T$  and  $T'$  under  $\sigma$ , can be obtained from  $\mathbf{v}^1$  and  $\tilde{\mathbf{v}}^1$  by inserting (the first occurrence of) element  $n$  immediately before the element  $\sigma(c)$  (resp.  $\sigma(c')$ ) and appending (the second occurrence of) element  $n$  at the end of the resulting vectors.

We now distinguish two cases:

- (i) If  $c = c'$ ,  $\{c, x_1\} = \{c', x_1\}$  is a common cherry of  $T$  and  $T'$ , and thus,  $wt(S) = wt(S_1)$ . Moreover,  $d_{\text{OLA}}^{\sigma_{-x_1}}(T_1, T'_1) = d_{\text{OLA}}^\sigma(T, T')$ . Furthermore, we have  $|\text{LCS}(\mathbf{v}_{\sigma(c)}, \mathbf{v}'_{\sigma(c)})| = |\text{LCS}(\mathbf{v}_{\sigma_{-x_1}(c)}^1, \tilde{\mathbf{v}}_{\sigma_{-x_1}(c)}^1)| + 1$  and  $|\text{LCS}(\mathbf{v}_i, \mathbf{v}'_i)| = |\text{LCS}(\mathbf{v}_i^1, \tilde{\mathbf{v}}_i^1)|$  for all  $i \in \{1, 2, \dots, n-1\} - \{\sigma(c)\}$ . In other words,  $\text{Sim}_{\text{HOP}}^\sigma(T, T') = \text{Sim}_{\text{HOP}}^{\sigma_{-x_1}}(T_1, T'_1) + 1$ . Thus,

$$d_{\text{HOP}}^\sigma(T, T') = n - \text{Sim}_{\text{HOP}}^\sigma(T, T') = n - 1 - \text{Sim}_{\text{HOP}}^{\sigma_{-x_1}}(T_1, T'_1) = d_{\text{HOP}}^{\sigma_{-x_1}}(T_1, T'_1).$$

The statement now follows immediately from Equation (??).

- (ii) If  $c \neq c'$ , then  $\{c, x_1\}$  is a cherry in  $T$ , while  $\{c', x_1\}$  is a cherry in  $T'$ . For the weights of the common cherry-picking sequence  $S$  of  $T$  and  $T'$  (resp.  $S_1$  of  $T_1$  and  $T'_1$ ) this implies that  $wt(S) = wt(S_1) + 1$ . By construction of the OLA vector (see Line 12 of Algorithm ??),  $c \neq c'$  implies the last positions of the vector differ. We have  $d_{\text{OLA}}^\sigma(T, T') = d_{\text{OLA}}^{\sigma_{-x_1}}(T_1, T'_1) + 1$ . Finally, for HOP, since the (first occurrence of) element  $n$  is inserted immediately before  $\sigma(c)$  in  $\mathbf{v}^1$  to obtain  $\mathbf{v}$ , while it is inserted immediately before  $\sigma(c')$  in  $\tilde{\mathbf{v}}^1$  to obtain  $\mathbf{v}'$  (and since  $c \neq c'$ , we have  $\sigma(c) \neq \sigma(c')$ ), we have  $|\text{LCS}(\mathbf{v}_i, \mathbf{v}'_i)| = |\text{LCS}(\mathbf{v}_i^1, \tilde{\mathbf{v}}_i^1)|$

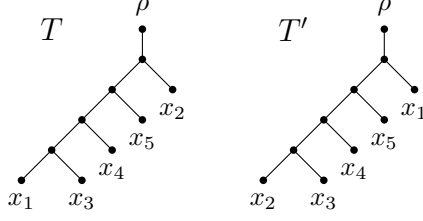


Figure 9: Two phylogenetic  $X$ -trees  $T$  and  $T'$  with precisely two common cherry-picking sequences,  $S = (x_3, x_4, x_5, x_1, x_2)$  and  $S' = (x_3, x_4, x_5, x_2, x_1)$ .

for all  $i \in \{1, 2, \dots, n-1\}$ , i.e., no LCS changes its length. This implies that  $\text{Sim}_{\text{HOP}}^\sigma(T, T') = \text{Sim}_{\text{HOP}}^{\sigma-x_1}(T_1, T'_1)$  and

$$\begin{aligned} d_{\text{HOP}}^\sigma(T, T') &= n - \text{Sim}_{\text{HOP}}^\sigma(T, T') = n - \text{Sim}_{\text{HOP}}^{\sigma-x_1}(T_1, T'_1) = n - 1 - \text{Sim}_{\text{HOP}}^{\sigma-x_1}(T_1, T'_1) + 1 \\ &= d_{\text{HOP}}^{\sigma-x_1}(T_1, T'_1) + 1. \end{aligned}$$

The statement now again immediately follows from Equation (??). This completes the proof.  $\square$

With Lemmas ?? and ??, we can now establish Theorem ??:

*Proof of Theorem ??.* The first equality follows immediately from Corollary ??, and the second and third equality from Lemma ??:

$$d_{\text{P2V}}^\sigma(T, T') = d_{\text{OLA}}^\sigma(T, T') = d_{\text{HOP}}^\sigma(T, T') = \text{wt}(S).$$

$\square$

Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees that have a common cherry-picking sequence. If  $\sigma$  is an ordering on  $X$  that is induced by a cherry-picking sequence that is common to  $T$  and  $T'$ , then we refer to  $\sigma$  as a *cherry-picking sequence (CPS) ordering* on  $X$ . Now for each  $\Theta \in \{\text{HOP}, \text{OLA}, \text{P2V}\}$ , define

$$d_\Theta^{\text{CPS}}(T, T') = \min_\sigma d_\Theta^\sigma(T, T'),$$

where the minimum is taken over all CPS orderings  $\sigma$  on  $X$ . The next corollary follows from Theorems ?? and ??.

**Corollary 3.** *Let  $T$  and  $T'$  be two phylogenetic  $X$ -trees with a common cherry-picking sequence. Then*

$$d_{\text{OLA}}^{\text{CPS}}(T, T') = d_{\text{HOP}}^{\text{CPS}}(T, T') = d_{\text{P2V}}^{\text{CPS}}(T, T') = h_t(T, T').$$

It is worth nothing that for two phylogenetic trees  $T$  and  $T'$  and any  $\Theta \in \{\text{HOP}, \text{OLA}, \text{P2V}\}$ , the two measures  $d_\Theta^*(T, T')$  and  $d_\Theta^{\text{CPS}}(T, T')$  are not necessarily equal. To see this, consider the two trees that are shown in Figure ???. Their only two common cherry-picking sequences are  $S = (x_3, x_4, x_5, x_1, x_2)$  and  $S' = (x_3, x_4, x_5, x_2, x_1)$  with  $\text{wt}(S) = \text{wt}(S') = 3$ . However, the ordering  $\sigma$  on  $X$  with  $\sigma(x_3) < \sigma(x_4) < \sigma(x_5) < \sigma(x_1) < \sigma(x_2)$ , is not a CPS ordering on  $X$ . It is straightforward to check that  $d_\Theta^*(T, T') \leq d_\Theta^\sigma(T, T') = 2$ .

## 5 Conclusion and open problems

Our paper explores the relationship between three novel dissimilarity measures on phylogenetic trees (each of which, given a fixed ordering of the leaf set, takes polynomial time to compute) and the popular, but computationally hard, rSPR distance. While there is little relationship between these measures and rSPR, we show that direct relationships exist when we compare these measures with the hybrid number and temporal hybrid number. We end the paper with some open problems concerning these intriguing classes of measure.

### Capturing the rSPR distance

The first problem is to develop an order-dependent measure that is efficient to compute but, when minimized over all orders, is equivalent to the rSPR distance. For the order-dependent measures developed thus far, the ordering of the leaves introduces additional structure that allows for polynomial running time to compute the distances. For example, the HOP ordering provides a clever way to decompose the tree into paths, and for each leaf, the path between it and the root can be viewed as a backbone where leaves later in the ordering are attached. This decomposition captures the agreement forests induced by rSPR moves, albeit well-behaved rSPR moves that do not move a component with a lower minimal element to one with a higher minimal element. Theorem ?? shows that this decomposition precludes cycles in the agreement forests and equates HOP to the hybrid number. Is it possible to keep the efficient running time that the order provides but capture more general rSPR moves?

**Open Problem 1.** *Is there an order-dependent measure  $\varphi$  that captures rSPR? That is, for two phylogenetic  $X$ -trees  $T$  and  $T'$ , if we minimize across all orderings of  $X$*

$$d_{\varphi}^*(T, T') = d_{\text{rSPR}}(T, T').$$

*but, for fixed ordering  $\sigma$  on  $X$ , the value  $d_{\varphi}^{\sigma}(T, T')$  can be computed in polynomial time.*

### Sharpening the upper bound for OLA

In Theorem ?? we proved that for all phylogenetic  $X$ -trees  $T$  and  $T'$ , there exists an ordering  $\sigma$  on  $X$  for which  $d_{\text{OLA}}^{\sigma}(T, T')$  is bounded from above by  $28 \cdot d_{\text{rSPR}}(T, T')$ . Can this upper bound be lowered?

**Open Problem 2.** *Does there exist a constant  $c < 28$  such that for any two phylogenetic  $X$ -trees  $T$  and  $T'$ , there exists an ordering  $\sigma$  such that*

$$d_{\text{OLA}}^{\sigma}(T, T') \leq c \cdot d_{\text{rSPR}}(T, T')?$$

### Relating the three order-dependent measures

In this paper, we have focused on relating the three order-dependent measures OLA, P2V, and HOP to well-established concepts for phylogenetic  $X$ -trees, including the rSPR distance, the hybrid number, and cherry-picking sequences. However, an interesting future direction is to relate the three order-dependent measures to each other. For instance, one could ask how the minimal measures  $d_{\text{OLA}}^*(T, T')$ ,  $d_{\text{P2V}}^*(T, T')$ , and  $d_{\text{HOP}}^*(T, T')$  are related for two given phylogenetic  $X$ -trees  $T$  and  $T'$ . Is one always smaller than the other? What can be said about the orderings  $\sigma$  that induce the minimum? Are they the same for  $d_{\text{OLA}}^*(T, T')$ ,  $d_{\text{P2V}}^*(T, T')$ , and  $d_{\text{HOP}}^*(T, T')$ ?

To give a flavor of these types of questions, we consider the relationship between the OLA and HOP measures. In the following, let  $T$  be a phylogenetic  $X$ -tree with at least two leaves. We refer to  $T$  as a *caterpillar* if we can order  $X$ , say  $x_1, x_2, \dots, x_n$ , such that  $x_1$  and  $x_2$  have the same parent and, for

each  $i \in \{3, 4, \dots, n\}$ , the parent of  $x_i$  is the parent of  $x_{i-1}$ . If  $T$  is a caterpillar, then we denote it by  $(x_1, x_2, x_3, \dots, x_n)$  or, equivalently,  $(x_2, x_1, x_3, \dots, x_n)$ . As an example, Figure ?? shows the two caterpillars  $T = (x_1, x_3, x_4, x_5, x_2)$  and  $T' = (x_2, x_3, x_4, x_5, x_1)$ .

First consider the two caterpillars

$$S = (x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_m) \text{ and } S' = (y_1, y_2, \dots, y_m, x_1, x_2, \dots, x_m)$$

on  $2m$  leaves as depicted in Figure ?. Here, we have  $d_{\text{OLA}}^*(S, S') = d_{\text{SPR}}(S, S') = 2$  and  $d_{\text{HOP}}^*(S, S') = m$ . For OLA, an ordering  $\sigma$  on  $X$  that realizes  $d_{\text{OLA}}^*(S, S')$  is, for instance, given by

$$\sigma(x_1) < \sigma(x_2) < \dots < \sigma(x_m) < \sigma(y_1) < \sigma(y_2) < \dots < \sigma(y_m).$$

The same ordering  $\sigma$  also realizes  $d_{\text{HOP}}^*(S, S')$ . In summary, the ordering  $\sigma$  given above realizes both  $d_{\text{OLA}}^*(S, S')$  and  $d_{\text{HOP}}^*(S, S')$ . Furthermore, we have  $d_{\text{OLA}}^*(S, S') < d_{\text{HOP}}^*(S, S')$ . In particular,  $d_{\text{OLA}}^*(S, S') = 2$  is a constant, whereas  $d_{\text{HOP}}^*(S, S') = m$  depends on  $m$ .

Next, consider the two caterpillars

$$T = (x_1, x_2, y_1, x_3, y_2, \dots, y_{m-3}, x_{m-1}, y_{m-2}, x_m, y_{m-1})$$

and

$$T' = (x_1, x_m, y_1, x_{m-1}, y_2, \dots, y_{m-3}, x_3, y_{m-2}, x_2, y_{m-1})$$

on  $2m - 1$  leaves. Here, we have  $d_{\text{HOP}}^*(T, T') = h(T, T') = m - 1$ . An ordering  $\sigma$  on  $X$  realizing  $d_{\text{HOP}}^*(T, T')$  is given by

$$\sigma(x_1) < \sigma(x_2) < \dots < \sigma(x_m) < \sigma(y_1) < \dots < \sigma(y_{m-1}).$$

Under this ordering  $\sigma$ , we have  $d_{\text{OLA}}^\sigma(T, T') = 2m - 3 > h(T, T')$ . By Theorem ??,  $\sigma$  clearly does not realize  $d_{\text{OLA}}^*(T, T')$ , indicating that an ordering that realizes the minimum for one of the three-order dependent measures (here,  $d_{\text{HOP}}^*(T, T')$ ) does not necessarily realize the minimum for the others (here,  $d_{\text{OLA}}^*(T, T')$ ).

We thus end by posing the following broad open problem.

**Open Problem 3.** *Given two rooted phylogenetic  $X$ -trees  $T$  and  $T'$  and an ordering  $\sigma$  on  $X$ , what can be said about the relationship of  $d_{\text{OLA}}^\sigma(T, T')$ ,  $d_{\text{P2V}}^\sigma(T, T')$ , and  $d_{\text{HOP}}^\sigma(T, T')$ ? Further, what can be said about the relationship of  $d_{\text{OLA}}^*(T, T')$ ,  $d_{\text{P2V}}^*(T, T')$ , and  $d_{\text{HOP}}^*(T, T')$  and the orderings realizing them?*

## 6 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1929284 while the authors were in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the Theory, Methods, and Applications of Quantitative Phylogenomics semester program. The first and third authors thank the New Zealand Marsden Fund for their financial support.

## References

- [1] BARONI, M., GRÜNEWALD, S., MOULTON, V., AND SEMPLE, C. Bounding the number of hybridisation events for a consistent evolutionary history. *Journal of Mathematical Biology* 51 (2005), 171–182.
- [2] BORDEWICH, M., AND SEMPLE, C. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics* 8 (2005), 409–423.



- [3] BORDEWICH, M., AND SEMPLE, C. Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics* 155 (2007), 914–928.
- [4] CHAUVE, C., COLIJN, C., AND ZHANG, L. A vector representation for phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences* 380, 20240226 (2025).
- [5] HUMPHRIES, P. J., LINZ, S., AND SEMPLE, C. Cherry picking: A characterization of the temporal hybridization number for a set of phylogenies. *Bulletin of Mathematical Biology* 75 (2013), 1879–1890.
- [6] HUMPHRIES, P. J., AND SEMPLE, C. Note on the hybridization number and subtree distance in phylogenetics. *Applied Mathematics Letters* 22 (2009), 611–615.
- [7] KUHNER, M. K., AND YAMATO, J. Practical performance of tree comparison metrics. *Systematic Biology* 64 (2015), 205–214.
- [8] PENN, M. J., SCHEIDWASSER, N., KHURANA, M. P., DUCHÊNE, D. A., DONNELLY, C. A., AND BHATT, S. Phylo2Vec: A vector representation for binary trees. *Systematic Biology* 74 (2025), 250–266.
- [9] PENN, M. J., SCHEIDWASSER, N., PENN, J., DONNELLY, C. A., DUCHÊNE, D. A., AND BHATT, S. Leaping through tree space: Continuous phylogenetic inference for rooted and unrooted trees. *Genome Biology and Evolution* 15 (2023), evad213.
- [10] RICHMAN, H., ZHANG, C., AND MATSEN, F. A. Vector encoding of phylogenetic trees by ordered leaf attachment. *arXiv:2503.10169* (2025).
- [11] SCHEIDWASSER, N., NAG, A., PENN, M. J., JAKOB, A., ANDERSEN, F. M., KHURANA, M. P., SETIAWAN, L., GORDON, M., DUCHÊNE, D. A., AND BHATT, S. phylo2vec: A library for vector-based phylogenetic tree manipulation. *arXiv:2506.19490* (2025).
- [12] SEMPLE, C., AND STEEL, M. *Phylogenetics*. Oxford University Press, 2003.
- [13] ST. JOHN, K. The shape of phylogenetic treespace. *Systematic Biology* 66 (2017), e83–e94.
- [14] ZHANG, L., ABHARI, N., COLIJN, C., AND WU, Y. A fast and scalable method for inferring phylogenetic networks from trees by aligning lineage taxon strings. *Genome Research* 33 (2023), 1053–1060.