A parsimony-based metric for phylogenetic trees

Vincent Moulton\textsuperscript{a}, Taoyang Wu\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a}School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

Abstract

In evolutionary biology various metrics have been defined and studied for comparing phylogenetic trees. Such metrics are used, for example, to compare competing evolutionary hypotheses or to help organize algorithms that search for optimal trees. Here we introduce a new metric $d_p$ on the collection of binary phylogenetic trees each labelled by the same set of species. The metric is based on the so-called parsimony score, an important concept in phylogenetics that is commonly used to construct phylogenetic trees. Our main results include a characterization of the unit neighborhood of a tree in the $d_p$ metric, and an explicit formula for its diameter, that is, a formula for the maximum possible value of $d_p$ over all possible pairs of trees labelled by the same set of species. We also show that $d_p$ is closely related to the well-known tree bisection and reconnection (TBR) and subtree prune and regraft (SPR) distances, a connection which will hopefully provide a useful new approach to understanding properties of these and related metrics.

Keywords: metric, phylogenetic trees, parsimony score, tree operations, unit neighborhood, diameter

2000 MSC: 05C05, 05C99, 92B10.

1. Introduction

In evolutionary biology, researchers are often faced with the problem of comparing two evolutionary or phylogenetic trees on a given set of species. This problem commonly arises because there are various methods to construct such trees, and these often give different solutions which then need to be compared. In addition, some of the methods for constructing phylogenetic trees rely on searching through the set of all possible trees, and it can be useful to compare trees so as to efficiently organize such searches (see, e.g. [20]). For these reasons various metrics have been developed for comparing phylogenetic trees, see e.g. [1, 2, 4, 9, 15, 16, 18, 19]. These metrics have different properties which can make them more (or less!) useful depending on the situation in which they are to be used. For example, the so-called Robinson-Foulds metric [16] can give a quick way to compare trees, but is somewhat coarse in identifying details, whereas other metrics, such as the quartet-distance [9], can pick out more fine detail but can be more difficult to work with computationally.

In this paper, we introduce a new tree metric which is based on the concept of parsimony. To define this metric we first need to recall some concepts in phylogenetics (cf. [17]). Let $X$ be a finite set, corresponding to a set of species. A character on $X$ is a surjective map $\chi$ from $X$ into another finite set $C$. In biology, characters are commonly morphological (e.g. a species in $X$ either

\*Corresponding author (Tel: +44 1603 59 2954 Fax: +44 1603 593345)

\*\*Email addresses: vincent.moulton@cmp.uea.ac.uk (VM), taoyang.wu@uea.ac.uk (TW)
has fins or not) or genetic (e.g. the nucleotide in some position of the DNA for a species in \( X \) is A, T, C or G). Now, given such a character \( \chi \), and a phylogenetic tree \( T = (V, E) \) on \( X \) (i.e. a graph-theoretical tree with vertex set \( V \), edge set \( E \) and leaf-set \( X \), such that every interior vertex has degree three), an extension \( \overline{\chi} \) of \( \chi \) to \( T \) is a map \( \overline{\chi} : V \to C \) with \( \chi(x) = \overline{\chi}(x) \) for all \( x \in X \). The changing number of \( \overline{\chi} \) is the cardinality of the set \( \Delta(\overline{\chi}) \) consisting of all edges \( \{u, v\} \) in \( T \) with \( \overline{\chi}(u) \neq \overline{\chi}(v) \). The extension \( \chi \) is optimal if it has the minimum changing number over all possible extensions of \( \chi \) to \( T \), and the parsimony score \( l(T, \chi) \) of \( \chi \) on \( T \) is defined as the changing number of an optimal extension of \( \chi \) to \( T \). Note that in phylogenetics it is common practice to look for a phylogenetic tree that minimizes the sum of the parsimony scores over a given set of characters (see, e.g. [17, Chapter 5]) as such a tree is considered to represent a simplest explanation for how present-day species might have evolved.

Now, given two phylogenetic trees \( T \) and \( T' \) on \( X \), we define

\[
d_p(T, T') = \max_{\chi \in \Xi(X)} |l(T, \chi) - l(T', \chi)|,
\]

where \( \Xi(X) \) denotes the set of all characters on \( X \). In other words, \( d_p(T, T') \) is the largest difference in the parsimony scores for the trees \( T \) and \( T' \) over all possible characters on \( X \). Note that, by definition, \( d_p(T, T') = d_p(T', T) \) and that \( d_p(T, T') \geq 0 \) with equality holding if and only if \( T = T' \). Moreover, if \( T'' \) is also a phylogenetic tree on \( X \), and \( \chi^* \) is a character on \( X \) with \( d_p(T, T') = |l(T, \chi^*) - l(T', \chi^*)| \), then we have

\[
d_p(T, T') = |l(T, \chi^*) - l(T', \chi^*)| \\
\leq |l(T, \chi^*) - l(T'' ,\chi^*)| + |l(T'' ,\chi^*) - l(T', \chi^*)| \\
\leq d_p(T, T'') + d_p(T'', T'),
\]

where the first inequality follows from the triangle inequality, and the second from the definition of \( d_p \). In other words, the function \( d_p \) is a metric on the set of all phylogenetic trees on \( X \).

In this paper, we investigate some properties of this new metric. In particular, after presenting some preliminaries in the next section, in Section 3 we first show that the definition of \( d_p \) can be reformulated in terms of a special type of character. In Sections 4 and 5 we then prove some technical results which allow us to prove the main theorems in the paper. These include a way to bound the \( d_p \) distance between two trees that can be decomposed in a certain fashion (Proposition 4.1), and a result that gives the exact value of \( d_p \) between two particular phylogenetic trees (Theorem 5.1).

In Sections 6 – 8 we go on to prove the main results of this paper. More specifically, in Section 6, we characterize the set of trees that are \( d_p \) distance 1 from a given tree (Theorem 6.4), in Section 7 we show that \( d_p \) is very closely related to the so-called TBR and SPR metrics on phylogenetic trees (Theorem 7.1) and, finally, in Section 8 we give an exact formula for the diameter of \( d_p \), that is the maximum value that \( d_p \) can take over all possible pairs of trees (Theorem 8.1). Together, these last two results allow us to very slightly improve upon a lower bound for the diameter of the TBR distance given in [10].

Before proceeding, it is worth mentioning that in this paper we do not explicitly deal with the problem of computing the value of \( d_p \) for some given pair of trees. It is known that the problem of computing the TBR or SPR distance is NP-hard and also fixed parameter tractable [2, 3, 5, 13]. It would be interesting to know whether or not this is also the case for \( d_p \). In relation to answering this question, note there has recently been a great deal of work presented on structural properties of \( d_{\text{tbr}} \) and \( d_{\text{spr}} \) (see e.g. [7, 8, 10]). In light of the close connection between these metrics and \( d_p \) given in Theorem 7.1, it could be of some interest to see whether some of these results might
be extended to $d_e$, especially as this could hopefully provide new insights into some open problems concerning $d_{\text{TBR}}$ and $d_{\text{SPR}}$, such as the one stated at the very end of this paper.

2. Preliminaries

We begin by recalling some basic definitions concerning phylogenetic trees. We refer the reader to [17] for a more detailed exposition of the concepts mentioned here and in the introduction. From now on we will assume that $X$ is a finite set with $n = |X| \geq 4$, unless stated otherwise. For brevity we will denote any partition $\{A_1, A_2, \ldots, A_p\}$, $p \geq 2$, of $X$ by $A_1|A_2|\ldots|A_p$ (so, in particular, the order of the $A_i$ in this last expression is not important). In addition, we shall sometimes also denote a character $\chi : X \to C = \{\alpha_1, \ldots, \alpha_p\}$, $p \geq 2$, by the partition $\chi^{-1}(\alpha_1)|\chi^{-1}(\alpha_2)|\ldots|\chi^{-1}(\alpha_p)$, or by $\chi^{-1}(\alpha_1)|\chi^{-1}(\alpha_2)|\ldots|\chi^{-1}(\alpha_{p-1})|\chi^{-1}(\alpha_p)$. Finally, for simplicity, we often write a set $A = \{a_1, \ldots, a_k\}$ within a partition as $a_1a_2\ldots a_k$. For example, if a character $\chi$ corresponds to the partition $\{x_1, x_3\}|\{x_2, x_5\}|\{x_4, x_6\}$ of $X = \{x_1, \ldots, x_6\}$, then we can instead write it as $x_1x_3|x_2x_5|x_4x_6$, or $x_1x_3|x_2x_5|\chi^{-1}(\alpha_{p-1})|$ when $X$ is clear from the context.

**Phylogenetic trees** Let $T = (V, E)$ be a tree, that is, a graph $T = (V, E)$ with vertex set $V$ and edge set $E$. Leaves of $T$ are vertices with degree 1; non-leaf vertices are called interior vertices of $T$. The tree $T$ is binary if every interior vertex has degree three. A cherry of $T$ is a pair $\{v, w\} \subseteq V$ of leaves of $T$ that are adjacent to the same interior vertex of $T$.

Now suppose that $T$ is a phylogenetic tree on $X$ (i.e. a binary tree with leaf set $X$). Given a subset $X'$ of $X$, we let $T(X')$ denote the tree consisting of the minimum subtree of $T$ that connects all of the vertices in $X'$. Furthermore, the restriction of $T$ to $X'$, denoted by $T|_{X'}$, is the phylogenetic tree on $X'$ obtained from $T(X')$ by contracting all vertices of degree-two. A split $A|B$ ($= B|A$) of $T$ is a bipartition $\{A, B\}$ of the leaf set $X$ of $T$ such that $T(A)$ and $T(B)$ are vertex disjoint subtrees of $T$. Note that given a phylogenetic tree $T$ on $X$ and a non-trivial split $A|B$ in $T$ (i.e. a split $A|B$ with $|A| \geq 2$ and $|B| \geq 2$), there exists a unique edge $e$ of $T$ whose removal yields the trees $T(A)$ and $T(B)$. In this case we also say that the split $A|B$ is induced by the edge $e$.

A phylogenetic tree $T$ is called a caterpillar if every one of its interior vertices is adjacent to some leaf. Note that this is equivalent to $T$ containing precisely two cherries. For simplicity, we therefore denote the fact that $T$ is a caterpillar with two cherries $\{c_1, c_2\}$ and $\{c_3, c_4\}$ by writing $T$ as $c_1c_2 : x_1x_3 \cdots x_{n-4} : c_3c_4$, where $\{c_1, c_2\}, \{c_3, c_4\} \subseteq X$ and the remaining leaves $x_1, x_2, \ldots, x_{n-4}$ of $T$ are listed in the obvious way (e.g. the phylogenetic tree $T$ in Fig. 1 is the caterpillar $[1 \ 2 : 3 \ 4 : 5 \ 6]$). In particular, when $|X| = 4$, $T$ will be written as $c_1c_2 :: c_3c_4$. Note that the order of $c_1$ and $c_2$, as well as that of $c_3$ and $c_4$, in this coding scheme is not important, while the ordering of the elements $x_i$ is essential. A pair of leaves $x$ and $y$ in a caterpillar $T$ is called a sibling if the path connecting $x$ and $y$ in $T$ contains exactly three edges.

**Tree operations** We now introduce two tree operations. A TBR (tree bissection and reconnection) operation on $T$ involves deleting some edge $e$ from $T$ (bisection), and subsequently inserting a new edge so that the resulting tree $T'$ is distinct from $T$ (reconnection). Since we require $T'$ to be binary, it is necessary to subdivide an edge in one (in the case that the other component is an isolated labelled vertex) or both components created in the bisection stage before inserting the new edge. An example is given in Fig. 1. In addition, such a TBR operation is called an SPR (subtree prune and regraft) operation if one further constraint is satisfied: For the split $X_1|X_2$ induced by the edge $e$ specified in the bisection stage, there exists a subset $A := X' \cup \{x\}$ with $X' \subseteq \{X_1, X_2\}$ and $x \in X \setminus X'$ so that $T|_{A} = T'|_{A}$. Note that a more restrictive NNI (nearest neighbor interchange) operation may also be defined, but we will not consider this operation or related concepts in this
paper as its connection with TBR and SPR operations is well documented in the literature (see e.g. [2]).

Given two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ with the same leaf set, the TBR distance $d_{\text{tbr}}(\mathcal{T}, \mathcal{T}')$ between $\mathcal{T}$ and $\mathcal{T}'$ is defined as the minimum number of TBR operations that is needed to be applied one-by-one to change $\mathcal{T}$ into $\mathcal{T}'$. The SPR distance $d_{\text{spr}}(\mathcal{T}, \mathcal{T}')$ is defined in a similar manner. Note that $d_{\text{spr}}(\mathcal{T}, \mathcal{T}') \geq d_{\text{tbr}}(\mathcal{T}, \mathcal{T}')$ clearly holds.

We now recall a useful way to reformulate the TBR distance between two trees. Given two phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ on $X$, an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_2$ is a partition $\mathcal{F} = \{X_1, \ldots, X_k\}$ of $X$ such that (i) for $i = 1, 2$, the trees in $\{\mathcal{T}_i(X_1), \ldots, \mathcal{T}_i(X_k)\}$ are vertex disjoint subtrees of $\mathcal{T}_i$, and (ii) $\mathcal{T}_1|_{X_i}$ is isomorphic to $\mathcal{T}_2|_{X_i}$ for $1 \leq i \leq k$. The size of $\mathcal{F}$ is defined as $|\mathcal{F}|$. If $|\mathcal{F}| \leq |\mathcal{F}'|$ holds for all agreement forests $\mathcal{F}'$ for $\mathcal{T}_1$ and $\mathcal{T}_2$, then $\mathcal{F}$ is called a maximum-agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_2$, and we define $\text{MAF}(\mathcal{T}_1, \mathcal{T}_2) = |\mathcal{F}| - 1$ (that is, $\text{MAF}(\mathcal{T}_1, \mathcal{T}_2)$ is the size of a maximum-agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_2$ minus 1). In [2] it is shown that

$$d_{\text{tbr}}(\mathcal{T}, \mathcal{T}') = \text{MAF}(\mathcal{T}, \mathcal{T}')$$

(1)

holds for any pair $\mathcal{T}, \mathcal{T}'$ of phylogenetic trees on $X$.

![Figure 1: Two trees $\mathcal{T}, \mathcal{T}'$ that are one TBR operation apart.](image)

Figure 1: Two trees $\mathcal{T}, \mathcal{T}'$ that are one TBR operation apart. In particular, the tree $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by deleting the edge $e$ in $\mathcal{T}$ and reinserting the two edge $f$ between the edges in $\mathcal{T}$ incident with leaves 2 and 5, respectively.

**Parsimony scores** We now recall some facts concerning parsimony scores. In [6, Lemma 1] (based on [7, Lemma 5.1]), it is shown that if $\mathcal{T}$ and $\mathcal{T}'$ are two phylogenetic trees on $X$ with $d_{\text{tbr}}(\mathcal{T}, \mathcal{T}') \leq 1$, and $\chi$ is any character on $X$, then $l(\mathcal{T}', \chi) \leq l(\mathcal{T}, \chi) + 1$. This immediately implies the following relationship between $d_{\text{tbr}}$.

**Lemma 2.1.** If $\mathcal{T}, \mathcal{T}'$ are two phylogenetic trees on $X$, then $d_{\text{tbr}}(\mathcal{T}, \mathcal{T}') \leq d_{\text{tbr}}(\mathcal{T}, \mathcal{T}')$.

**Proof.** Suppose $d_{\text{tbr}}(\mathcal{T}, \mathcal{T}') = k$ for some $k \geq 1$. Then there is a sequence $\mathcal{T}_0 = \mathcal{T}', \mathcal{T}_1, \ldots, \mathcal{T}_k = \mathcal{T}$ of phylogenetic trees on $X$ so that $d_{\text{tbr}}(\mathcal{T}_{i-1}, \mathcal{T}_i) = 1$ for all $1 \leq i \leq k$. But, by the aforementioned [6, Lemma 1], $|l(\mathcal{T}_{i-1}, \chi) - l(\mathcal{T}_i, \chi)| \leq 1$ holds for every character $\chi$ on $X$. Therefore, $d_{\text{tbr}}(\mathcal{T}_{i-1}, \mathcal{T}_i) \leq 1$ for all $1 \leq i \leq k$ and so the lemma follows by applying the triangle inequality.

If $\chi_1$ and $\chi_2$ are both characters on $X$, then $\chi_1$ is a refinement of $\chi_2$ if for each element $\alpha$ in the image of $\chi_1$, we have $\chi_2(x) = \chi_2(y)$ for all $x, y \in \chi_1^{-1}(\alpha)$. The following fact is well known and easy to prove.

**Lemma 2.2.** If $\chi_1$ and $\chi_2$ are two characters on $X$ such that $\chi_1$ is a refinement of $\chi_2$, then, for every phylogenetic tree $\mathcal{T}$ on $X$, we have $l(\mathcal{T}, \chi_2) \leq l(\mathcal{T}, \chi_1)$. 

\[\square\]
We conclude this section by recalling a useful tool for computing parsimony scores that was introduced in [11]. Given a character $\chi$ on $X$ and a phylogenetic tree $T$ on $X$ an Erdős-Székely (ES) path system (for $\chi$ on $T$) is a collection of directed paths $P(x, y)$ from leaf $x$ to leaf $y$ in $T$, $x, y \in X$, such that (i) for each path $P(x, y)$ in this collection, $\chi(x) \neq \chi(y)$, and (ii) if $P(x, y)$ and $P(x', y')$ are two paths in this collection that have some edges in common, then $P(x, y)$ and $P(x', y')$ traverse these edges in the same direction and $\chi(y) \neq \chi(y')$.

**Theorem 2.3.** [11] Let $\chi$ be a character on $X$. Then for any phylogenetic tree $T$ on $X$, $l(T, \chi)$ is equal to the maximum size of an ES-path system for $\chi$ on $T$. □

3. A connection with convexity

In this section, we provide an alternative formulation for the metric $d_{p}$ in terms of the following concept that naturally arises when considering phylogenetic trees (cf. [17]). Given a phylogenetic tree $T$ on $X$, a character $\chi : X \to C$ is convex on $T$ if the score $h(T, \chi)$ of $\chi$ on $T$, defined as

$$h(T, \chi) = l(T, \chi) - |C| + 1,$$

is equal to zero. Note that this condition is equivalent to there being an extension $\chi$ of $\chi$ to $T$ such that for each $\alpha \in C$, the subgraph of $T$ induced by $\{v \in V : \chi(v) = \alpha\}$ is connected [17]. For later use, the cardinality of $\chi$, denoted by $|\chi|$, is defined as the number of elements contained in $C$.

Now, given two phylogenetic trees $T$ and $T'$ on $X$, let

$$\rho_{T}(T') = \max\{h(T', \chi) : \chi \text{ is a convex character on } T\},$$

and

$$\rho(T, T') = \max\{\rho_{T}(T'), \rho_{T'}(T)\}.$$  

Note that $\rho_{T}(T')$ is not necessarily equal to $\rho_{T'}(T)$ (see e.g. Fig. 2). However, in case $T$ and $T'$ have the same tree topology, that is, they are isomorphic as unlabeled trees, these two quantities must be equal.

Our aim is to prove that $\rho(T, T')$ equals $d_{p}(T_{1}, T_{2})$. To this end, we first show that the computation of $\rho_{T}(T')$ can be restricted to proper characters, that is, those characters $\chi : X \to C$ such that $|\chi^{-1}(\alpha)| \geq 2$ holds for each $\alpha \in C$.

**Lemma 3.1.** Suppose that $T$ and $T'$ are two phylogenetic trees on $X$. Then

$$\rho_{T}(T') = \max\{h(T', \chi) : \chi \text{ is a proper convex character on } T\}.$$  

**Proof.** For this proof, an element $\alpha$ in the image of a character $\chi$ is called trivial if $|\chi^{-1}(\alpha)| = 1$. Assume that $\rho_{T}(T') = h(T', \chi)$ holds for a convex character $\chi$ on $T$ with $k \geq 1$ trivial elements. Then it suffices to construct a convex character $\chi^{*}$ on $T$ such that $\chi^{*}$ has at most $k - 1$ trivial elements and $\rho_{T}(T') = h(T', \chi^{*})$.

To this end, let $\alpha$ be a trivial element of $\chi$ and let $x$ be the unique element in $X$ with $\chi(x) = \alpha$. Denote the vertex in $T$ that is adjacent to $x$ by $u$. Then there exists an optimal extension $\chi$ of $\chi$ to $T$ such that $\chi(u) \neq \alpha$. Now consider the character $\chi^{*}$ defined by $\chi^{*}(y) = \chi(u)$ for $y = x$, and $\chi^{*}(y) = \chi(y)$ otherwise. Then $\chi^{*}$ has at most $k - 1$ trivial elements. In addition, we have $l(T', \chi^{*}) \geq l(T', \chi) - 1$ and so $h(T', \chi^{*}) \geq h(T', \chi)$. Hence $h(T', \chi^{*}) \geq \rho_{T}(T')$ and so $h(T', \chi^{*}) = \rho_{T}(T')$, as required. □

We now prove a useful fact concerning refinements.
Lemma 3.2. Suppose that $\chi : X \to \mathcal{C}$ is a character and that $\mathcal{T}$ is a phylogenetic tree on $X$ with $l(\mathcal{T}, \chi) > |\mathcal{C}| - 1$. Then there exists a character $\chi'$ on $X$ such that (i) $\chi'$ is a refinement of $\chi$ and $|\chi'| = |\chi| + 1$, and (ii) $l(\mathcal{T}, \chi) = l(\mathcal{T}, \chi')$. In particular, there exists a refinement $\chi'$ of $\chi$ such that $l(\mathcal{T}, \chi') = l(\mathcal{T}, \chi)$ and $\chi'$ is convex on $\mathcal{T}$ (that is, $h(\mathcal{T}, \chi') = 0$).

Proof. Fix an optimal extension $\overline{\chi}$ of $\chi$ and denote the set of vertices in $\mathcal{T}$ by $V$. Since $\chi$ is not convex, there exists a subset $U$ of $V$ such that (i) all members of $U$ are mapped to the same element by $\overline{\chi}$, denoted by $\overline{\chi}(U)$, (ii) for each vertex $v$ in $V \setminus U$ that is adjacent to a vertex in $U$, $\overline{\chi}(v) \neq \overline{\chi}(U)$, (iii) $U \cap X \neq \emptyset$, and (iv) there exists some leaf $x \in X \setminus U$ such that $\chi(x) = \chi(U)$.

Now consider some $\epsilon$ not in $\mathcal{C}$, and let $\chi' : X \to \mathcal{C} \cup \{\epsilon\}$ be the map defined by $\chi'(x) = \epsilon$ for $x \in U \cap X$, and $\chi'(x) = \chi(x)$ otherwise. Then clearly $\chi'$ is a character on $X$ such that $\chi'$ is a refinement of $\chi$ and $|\chi'| = |\chi| + 1$. Therefore it only remains to show that $l(\mathcal{T}, \chi) = l(\mathcal{T}, \chi')$ holds.

To this end, first note that by Lemma 2.2 we have $l(\mathcal{T}, \chi) \leq l(\mathcal{T}, \chi')$. To see that the reverse inequality holds, consider the extension $\overline{\chi}'$ of $\chi'$ defined by $\overline{\chi}'(v) = \epsilon$ for $v \in U$, and $\overline{\chi}'(v) = \overline{\chi}(v)$ for $v \in V \setminus U$. Then, since $U$ satisfies (i) and (ii) above, we can conclude $l(\mathcal{T}, \chi') \leq |\Delta(\overline{\chi}')| = |\Delta(\overline{\chi})| = l(\mathcal{T}, \chi)$, as required.

We now show that $\rho$ and $d_v$ are equal.

Theorem 3.3. Suppose that $\mathcal{T}_1$ and $\mathcal{T}_2$ are two phylogenetic trees on $X$. Then

$$d_v(\mathcal{T}_1, \mathcal{T}_2) = \rho(\mathcal{T}_1, \mathcal{T}_2).$$

Proof. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be as in the statement of the theorem. It immediately follows by the definition of $\rho$ that $d_v(\mathcal{T}_1, \mathcal{T}_2) \geq \rho(\mathcal{T}_1, \mathcal{T}_2)$ holds. To see that the reverse inequality holds first note that, switching $\mathcal{T}_1$ and $\mathcal{T}_2$ if necessary, we may assume

$$\max_{\chi \in \Xi(X)} |l(\mathcal{T}_1, \chi) - l(\mathcal{T}_2, \chi)| = l(\mathcal{T}_1, \chi') - l(\mathcal{T}_2, \chi') = h(\mathcal{T}_1, \chi') - h(\mathcal{T}_2, \chi')$$

holds for some character $\chi'$ on $X$.

If $h(\mathcal{T}_2, \chi') = 0$, then we are done. Otherwise, by Lemma 3.2, there exists a character $\chi''$ so that $\chi''$ is a refinement of $\chi'$, $l(\mathcal{T}_2, \chi') = l(\mathcal{T}_2, \chi'')$ and $h(\mathcal{T}_2, \chi'') = 0$. In addition, we must have $l(\mathcal{T}_1, \chi'') = l(\mathcal{T}_1, \chi')$ as otherwise Lemma 2.2 would lead to a contradiction. Therefore, in summary, we have

$$\max_{\chi \in \Xi(X)} |l(\mathcal{T}_1, \chi) - l(\mathcal{T}_2, \chi)| = l(\mathcal{T}_1, \chi'') - l(\mathcal{T}_2, \chi'') = h(\mathcal{T}_1, \chi'') - h(\mathcal{T}_2, \chi'') \leq \rho(\mathcal{T}_2, \mathcal{T}_1) \leq \rho(\mathcal{T}_1, \mathcal{T}_2),$$

which completes the proof.

For later use, we present the following lemma concerning the $d_v$ metric on phylogenetic trees with five leaves.


Figure 3: The trees $T[A]$ and $T[B]$ for the split $A|B = \{1, 2, 3\}|\{4, 5, 6\}$ of the leaf set of the tree $T_1$ in Fig. 1.

**Lemma 3.4.** Suppose that $T_1$ and $T_2$ are two distinct phylogenetic trees on $X$. If $|X| = 5$, then $d_\rho(T_1, T_2) = 1$.

**Proof.** Let $X = \{x_1, \ldots, x_5\}$. Relabeling if necessarily, we may assume that the two cherries of $T_1$ are $\{x_1, x_2\}$ and $\{x_3, x_4\}$. That is, $T_1$ is a caterpillar and $T_1 = [x_1 x_2 : x_5 : x_3 x_4]$. By Lemma 2.1, we may further assume that $d_{tbr}(T_1, T_2) > 1$ as otherwise the lemma clearly holds. Since $|X| = 5$, this implies $d_{tbr}(T_1, T_2) = 2$. Swapping $x_3$ and $x_4$ if necessary, we can assume that $T_2 = [x_1 x_3 : x_5 : x_2 x_4]$. Therefore, $\chi_1 = x_1 x_3 x_2 x_4 x_5$ and $\chi_2 = x_1 x_3 x_5 x_2 x_4$ are the only two proper convex characters on $T_2$. Together with $h(T_1, \chi_1) = h(T_1, \chi_2) = 1$, using Lemma 3.1 it follows that $\rho_{T_2}(T_1) = 1$. In addition, a similar argument leads to $\rho_{T_2}(T_2) = 1$. Hence by Theorem 3.3

$$d_\rho(T_1, T_2) = \rho(T_1, T_2) = \max(\rho_{T_1}(T_2), \rho_{T_2}(T_1)) = 1,$$

as required. $\square$

The following corollary, albeit simple, will be useful since it shows that the distance induced by $d_\rho$ on subtrees cannot increase.

**Corollary 3.5.** Suppose that $T_1$ and $T_2$ are two phylogenetic trees on $X$, and $Y$ is a subset of $X$. Then $\rho_{T_1|Y}(T_2|Y) \leq \rho_{T_1}(T_2)$. In particular, by Theorem 3.3, $d_\rho(T_1|Y, T_2|Y) \leq d_\rho(T_1, T_2)$.

**Proof.** It clearly suffices to prove the corollary for the special case $Y = X \setminus \{x\}$ for some $x \in X$. Let $\chi$ be a character on $Y$ such that $\chi$ is convex on $T_1|Y$, and $h(T_2|Y, \chi) = \rho_{T_1|Y}(T_2|Y)$ holds. Consider the character $\chi^*$ on $X$ that is obtained from $\chi$ by mapping $x$ to a new symbol not in the image of $\chi$. Then $\chi^*$ is convex on $T_1$ and $h(T|Y, \chi) = h(T, \chi^*)$. Thus $\rho_{T_1|Y}(T_2|Y) \leq \rho_{T_1}(T_2)$, as required. $\square$

4. Tree decompositions

In this section we prove two results which allow us to restrict our attention to certain special subtrees when computing $d_\rho$ for two trees that have a non-trivial split in common. To this end, given a phylogenetic tree $T$ on $X$ and an edge $e$ of $T$ corresponding to a non-trivial split $A|B$ of $T$, let $T[A]$ and $T[B]$ be the two trees obtained from $T$ by dividing $e$ into three new edges through adding two new vertices $r_A$ and $r_B$, and deleting the edge $\{r_A, r_B\}$ as illustrated in Fig. 3. In particular, $T[A]$ and $T[B]$ are phylogenetic trees on $A \cup \{r_A\}$ and $B \cup \{r_B\}$, respectively.

**Proposition 4.1.** Suppose that $T_1, T_2$ are two phylogenetic trees on $X$. If they contain a common non-trivial split $A|B$, then

$$d_\rho(T_1, T_2) \leq d_\rho(T_1[A], T_2[A]) + d_\rho(T_1[B], T_2[B]).$$

(2)
Proof. Let \( \mathcal{T}_1, \mathcal{T}_2 \) be as in the statement of the theorem. Then \( \mathcal{T}_1[A] \) and \( \mathcal{T}_2[A] \) are two phylogenetic trees on \( A \cup \{ r_A \} \), and \( \mathcal{T}_1[B] \) and \( \mathcal{T}_2[B] \) are two phylogenetic trees on \( B \cup \{ r_B \} \). Let \( e_1 = \{ u_1, v_1 \} \) and \( e_2 = \{ u_2, v_2 \} \) be the edges in \( \mathcal{T}_1, \mathcal{T}_2 \) corresponding to \( A[B] \), respectively, with \( u_i \) being the vertex in \( \mathcal{T}_i(A) \), \( i = 1, 2 \). Note that by Theorem 3.3, it suffices to show

\[
\rho_{\mathcal{T}_1}(\mathcal{T}_2) \leq \rho_{\mathcal{T}_1[A]}(\mathcal{T}_2[A]) + \rho_{\mathcal{T}_1[B]}(\mathcal{T}_2[B]).
\]

To this end, let \( \chi \) be a character on \( X \) with \( h(\mathcal{T}_1, \chi) = 0 \) and \( h(\mathcal{T}_2, \chi) = \rho_{\mathcal{T}_1}(\mathcal{T}_2) \). In addition, fix an optimal extension \( \bar{\chi} \) of \( \chi \) to \( \mathcal{T}_1 \). Then the character \( \chi_A \) on \( A \cup \{ r_A \} \) defined by putting \( \chi_A(a) = \chi(a) \) if \( a \in A \) and \( \chi_A(r_A) = \bar{\chi}(v_1) \) is convex on \( \mathcal{T}_1[A] \), and the character \( \chi_B \) on \( B \cup \{ r_B \} \) defined by putting \( \chi_B(b) = \chi(b) \) if \( b \in B \) and \( \chi_B(r_B) = \bar{\chi}(u_1) \) is convex on \( \mathcal{T}_1[B] \). Therefore, it suffices to show

\[
h(\mathcal{T}_2, \chi) \leq h(\mathcal{T}_2[A], \chi_A) + h(\mathcal{T}_2[B], \chi_B). \tag{3}
\]

Now, fix an optimal extension \( \bar{\chi}_A \) of \( \chi_A \) to \( \mathcal{T}_2[A] \), and \( \bar{\chi}_B \) of \( \chi_B \) to \( \mathcal{T}_2[B] \). Consider the extension \( \bar{\chi} \) of \( \chi \) to \( \mathcal{T}_2 \) defined by putting \( \bar{\chi}(w) = \bar{\chi}_A(w) \) for vertex \( w \) in \( \mathcal{T}_2(A) \), and \( \bar{\chi}(w) = \bar{\chi}_B(w) \) for vertex \( w \) in \( \mathcal{T}_2(B) \). Note that, since \( \chi \) is convex on \( \mathcal{T}_1 \), \( \chi(A) \cap \chi(B) \) can contain at most one element.

If \( \chi(A) \cap \chi(B) \) contains exactly one element, which we denote by \( \alpha \), then as \( \bar{\chi} \) is optimal, we have \( \bar{\chi}(u_1) = \bar{\chi}(v_1) = \alpha \), and so \( \chi_A(r_A) = \alpha \) and \( \chi_B(r_B) = \alpha \). Therefore, we have

\[
l(\mathcal{T}_2, \chi) \leq |\Delta(\bar{\chi})| = |\Delta(\bar{\chi}_A)| + |\Delta(\bar{\chi}_B)| = l(\mathcal{T}_2[A], \chi_A) + l(\mathcal{T}_2[B], \chi_B),
\]

and hence

\[
h(\mathcal{T}_2, \chi) + |\chi| - 1 \leq l(\mathcal{T}_2[A], \chi_A) + l(\mathcal{T}_2[B], \chi_B) + |\chi_A| + |\chi_B| - 2.
\]

But \( |\chi| = |\chi_A| + |\chi_B| - 1 \), and so Eq. (3) holds.

If, on the other hand, \( \chi(A) \cap \chi(B) = \emptyset \), then as \( \bar{\chi} \) is optimal, we have \( \bar{\chi}(u_1) \in \chi(A) \) and \( \bar{\chi}(v_1) \in \chi(B) \). Since \( \chi_A(r_A) = \bar{\chi}(v_1) \notin \chi(A) \), we can assume \( \bar{\chi}_A(u_2) \neq \bar{\chi}_A(r_A) \). Similarly, we can assume \( \bar{\chi}_B(v_2) \neq \bar{\chi}_B(r_B) \). Hence

\[
l(\mathcal{T}_2, \chi) + 1 \leq |\Delta(\bar{\chi})| + 1 = |\Delta(\bar{\chi}_A)| + |\Delta(\bar{\chi}_B)| = l(\mathcal{T}_2[A], \chi_A) + l(\mathcal{T}_2[B], \chi_B).
\]

Therefore, as \( |\chi(A)| + 1 = |\chi_A| \) and \( |\chi(B)| + 1 = |\chi_B| \), Eq. (3) holds, as required. \( \square \)

Note that the strict inequality may hold in Eq. (2). For example, for the two trees \( \mathcal{T}_1 \) and \( \mathcal{T}_2 \) in Fig. 1 and the split \( A[B] = \{ 1, 2, 3 \}|\{ 4, 5, 6 \} \) of \( \mathcal{T}_1 \) and \( \mathcal{T}_2 \), we have

\[
1 = d_r(\mathcal{T}_1, \mathcal{T}_2) < d_r(\mathcal{T}_1[A], \mathcal{T}_2[A]) + d_r(\mathcal{T}_1[B], \mathcal{T}_2[B]) = 2.
\]

However, for the special case where \( \mathcal{T}_1[B] \) is isomorphic to \( \mathcal{T}_2[B] \), Corollary 3.5 and Proposition 4.1 immediately imply that the equality must always hold in Eq. (2) (note that the same property is shown to hold for \( d_{\text{thr}} \) and \( d_{\text{spr}} \) in [2, Theorem 3.4]), i.e. we have:

**Theorem 4.2.** Suppose that \( \mathcal{T}_1, \mathcal{T}_2 \) are phylogenetic trees on \( X \) that have a common non-trivial split \( A[B] \) such that \( \mathcal{T}_1[B] = \mathcal{T}_2[B] \). Then \( d_r(\mathcal{T}_1, \mathcal{T}_2) = d_r(\mathcal{T}_1[A], \mathcal{T}_2[A]) \). \( \square \)

### 5. \( d_p \) on caterpillars

In this section, we shall prove the following rather technical theorem concerning the \( d_p \) distance between two caterpillars. This result will be key in understanding unit neighborhoods of \( d_p \) in the next section.
Theorem 5.1. Let $|X| \geq 7$ and suppose that $T_1$ and $T_2$ are two caterpillars on $X$ with $d_{\text{mr}}(T_1, T_2) = 2$ and that have no cherry in common. Then $d_p(T_1, T_2) = 2$.

Before proving this theorem we present a useful lemma.

Lemma 5.2. Suppose $|X| \geq 7$ and let $\chi : X \to C$ be a character with $|C| \leq 3$. If $T$ is a caterpillar on $X$ with cherries $\{y_1, z_1\}$ and $\{y_2, z_2\}$ such that (i) $\chi(y_i) \neq \chi(z_i)$ for $i = 1, 2$ and (ii) for each element $\alpha \in C$, $\chi(x) = \alpha$ for some $x \in X \setminus \{y_1, y_2, z_1, z_2\}$, then $h(T, \chi) \geq 2$.

Proof. If $|C| = 2$, then Condition (ii) implies that there exist two elements $x_1$ and $x_2$ in $X \setminus \{y_1, y_2, z_1, z_2\}$ with $\chi(x_1) \neq \chi(x_2)$. Therefore $\{P(y_1, x_1), P(y_2, x_2), P(x_1, x_2)\}$ is an ES-path system for $\chi$ on $T$ by Condition (i). Hence $h(T, \chi) \geq 2$.

If $|C| = 3$, then Condition (ii) implies that there exists a subset $\{x_1, x_2, x_3\} \subseteq X \setminus \{y_1, y_2, z_1, z_2\}$ so that $\chi(x_i) \neq \chi(x_j)$ for $1 \leq i < j \leq 3$. Therefore, by Condition (i), $\{P(z_1, x_1), P(z_2, x_2), P(x_1, x_2), P(x_1, x_3)\}$ is an ES-path system for $\chi$ on $T$. Hence $h(T, \chi) \geq 2$. $\square$

Proof of Theorem 5.1: Let $T_1$ and $T_2$ be as in the statement of the theorem. By Lemma 2.1 we have $d_p(T_1, T_2) \leq 2$, and so it suffices to show that $d_p(T_1, T_2) \geq 2$ holds. To this end, for this proof we shall say that a character on $X$ is good if it is convex on $T_1$ and $h(T_2, \chi) \geq 2$, or it is convex on $T_2$ and $h(T_1, \chi) \geq 2$. In particular, note that if such a character exists then $d_p(T_1, T_2) \geq 2$ by Theorem 3.3, and so the existence of a good character proves the theorem.

Now, denote the two cherries in $T_1$ by $C_1$ and $C_2$ and put $\text{ch}(T_1) = C_1 \cup C_2$. Similarly, let $\text{ch}(T_2)$ be the union of the two cherries of $T_2$. Since $d_{\text{mr}}(T_1, T_2) = 2$, there is a maximum-agreement forest $F = \{X_1, X_2, X_3\}$ between $T_1$ and $T_2$. Without loss of generality assume $X_1 \cap \text{ch}(T_1) \neq \emptyset$. We first consider a special case.

Claim 1: If $|X_2| = |X_3| = 1$, then $d_p(T_1, T_2) \geq 2$.

Proof of Claim: Denote the element contained in $X_2$ by $y$, and the one contained in $X_3$ by $z$. In addition, let $T_1|_{X_1} = [c_1 c_2 : x_1 \cdots x_k : c_3 c_4]$ for some $k \geq 1$, where $c_i$ ($1 \leq i \leq 4$) and $x_j$ ($1 \leq j \leq k$) are elements in $X$. We can assume without loss of generality that $|\text{ch}(T_1) \cap \{y, z\}| \leq |\text{ch}(T_2) \cap \{y, z\}|$ holds. We now consider three cases.

Case I: $|\text{ch}(T_1) \cap \{y, z\}| = 0$. Since this implies that neither $y$ nor $z$ is contained in $\text{ch}(T_1)$, we have $C_1 = \{c_1, c_2\}$ and $C_2 = \{c_3, c_4\}$. Hence neither $C_1$ nor $C_2$ is a cherry in $T_2$. So, by symmetry, the cherries of $T_2$ are $\{c_1, y\}$ and $\{z, c_3\}$. Therefore, since $\chi$ is convex on $T_1$ and $h(T_2, \chi) \geq 2$ by Lemma 5.2, it follows that the character $\chi = c_1 c_2 c_3 c_4$ is good.

Case II: $|\text{ch}(T_1) \cap \{y, z\}| = 1$. Without loss of generality, assume $y \in \text{ch}(T_1)$ and $z \notin \text{ch}(T_1)$. By symmetry, we can also assume $C_1 = \{c_1, y\}$ and $C_2 = \{c_3, c_4\}$. Thus $T_1$ can be obtained from $T_1' := T_1|_{X_1 \cup X_2}$ by attaching the element $z$ to some edge $e_i$ with $0 \leq i \leq k + 1$ (see Fig. 4). Now, since $|\text{ch}(T_1) \cap \{y, z\}| \leq |\text{ch}(T_2) \cap \{y, z\}|$, either $y$ or $z$ is contained in $\text{ch}(T_1)$, and so it suffices to consider the following three subcases.

(II-1): $\text{ch}(T_2)$ contains $y$ but not $z$. This implies that $T_2$ can be obtained from $T_2' := T_2|_{X_1 \cup X_2}$ by attaching $z$ to some edge $f_j$ with $0 \leq j \leq k + 1$ (see Fig. 4). Now, note first that if $z$ is attached to $e_0$ in $T_1'$, then $c_1 y z$ is a good character by Lemma 5.2. Similarly, if $z$ is attached to $f_{k+1}$ in $T_2'$, then $c_3 y z$ is a good character. Moreover, if $z$ is attached to $e_{k+1}$ in $T_1'$, then $z$ must be attached to $f_j$ in $T_2'$ for some $0 \leq j < k$, which implies that $c_3 c_4 z$ is a good character. Similarly, if $z$ is
attached to $f_0$ in $T'_2$ then $c_1c_2z$ is a good character. Therefore, we only need to consider the case where $z$ is attached to $e_i$ for $1 \leq i \leq k$, and to $f_j$ for $1 \leq j \leq k$. Moreover, we may assume that $z$ is not attached to $e_1$, as otherwise $c_1c_2x_1$ is a good character, and $z$ is not attached to $f_k$, as otherwise $c_3c_4x_k$ is a good character. Therefore, it only remains to consider when $z$ is attached to $e_i$ for $1 < i \leq k$, and to $f_j$ for $1 \leq j < k$.

To this end, put $T'_i := T'_i|_{X_i \cup X_3}$ for $i = 1, 2$. Then $T'_1 \neq T'_2$, and so there must exist an element $x \in \{x_1, \ldots, x_k\}$ such that either (a) $d_{T'_1}(c_1, x) < d_{T'_1}(c_1, z)$ and $d_{T'_2}(c_1, x) > d_{T'_2}(c_1, z)$ or (b) $d_{T'_1}(c_1, x) > d_{T'_1}(c_1, z)$ and $d_{T'_2}(c_1, x) < d_{T'_2}(c_1, z)$. Put $X' = \{c_1, c_2, c_3, x, y, z\}$. Then in case (a) we have $\mathcal{T}_1|_{X'} = \{c_1y : c_2x_{\eta} : c_3c_4\}$ and $\mathcal{T}_2|_{X'} = \{c_1x : zxc_4 : c_3y\}$. Therefore, $d_p(\mathcal{T}_1, \mathcal{T}_2) \geq d_p(\mathcal{T}_1|_{X}, \mathcal{T}_2|_{X}) \geq 2$, where the first inequality follows from Corollary 3.5 and the second is obtained by considering the character $\chi = c_3c_4z|c_1c_2yx$ on $X'$. To see that $d_p(\mathcal{T}_1, \mathcal{T}_2) \geq 2$ must hold in case (b) is similar.

Figure 4: Two trees $T'_1, T'_2$ considered in the proof of Theorem 5.1.

(II-2): $\text{ch}(\mathcal{T}_2)$ contains $z$ but not $y$. By switching $c_3$ and $c_4$ if necessary, we may assume that the two cherries in $\mathcal{T}_2$ are $\{c_1, c_2\}$ and $\{c_3, z\}$. Hence the character $\chi = c_1y|c_3c_4|$ is good by Lemma 5.2.

(II-3): $\text{ch}(\mathcal{T}_2)$ contains both $z$ and $y$. Note first that we may assume $\{c_3, c_4\} \cap \text{ch}(\mathcal{T}_2) \neq \emptyset$. Indeed, if neither $c_3$ nor $c_4$ is contained in $\text{ch}(\mathcal{T}_2)$, then the two cherries in $\mathcal{T}_2$ must be $\{y, z\}$ and $\{c_1, c_2\}$. Thus, switching $c_3$ and $c_4$ if necessary, we may assume that $c_3$ is a sibling of $y$ in $\mathcal{T}_2$, which implies that $c_3yz$ is a good character.

Now, without loss of generality, assume $c_3 \in \text{ch}(\mathcal{T}_2)$. Note that if $\{c_3, z\}$ is a cherry in $\mathcal{T}_2$, then the other cherry of $\mathcal{T}_2$ must be $\{c_2, y\}$, and hence $\chi = c_1y|c_3c_4|$ is a good character. Therefore, we can assume that $\{c_3, y\}$ is a cherry in $\mathcal{T}_2$. Denote the other cherry in $\mathcal{T}_2$ by $C'$. Then $C'$ is either equal to $\{c_1, z\}$ or $\{c_2\}$.

If $C' = \{c_1, z\}$, then $z$ is not attached to $e_0$ in $\mathcal{T}_1'$ as otherwise $\{X_1 \cup X_3, X_2\}$ is a maximum agreement forest of $\{\mathcal{T}_1, \mathcal{T}_2\}$. On the other hand, if $z$ is attached to $e_i$ for some $i \geq 1$, then $c_1c_2z$ is a good character. Therefore, we may assume that $z$ is attached to $e_1$ in $\mathcal{T}_1'$. This implies $\chi = c_1yc_2|c_3c_4|$ is a convex character on $T_1$, and hence $\chi$ is a good character by Lemma 5.2.

If $C' \neq \{c_1, z\}$, then $C' = \{c_2, z\}$, and the proof is similar. In particular, we may assume that $z$ is not attached to $e_0$ in $\mathcal{T}_1'$ as otherwise $c_1yz$ is a good character. In addition, if $z$ is attached to $e_i$ for some $i \geq 1$, then $c_1c_2z$ is a good character. Therefore we may assume that $z$ is attached to $e_1$ in $\mathcal{T}_1'$. This implies that $\chi = c_1yc_2|c_3c_4|$ is a good character.

Case III: $\{y, z\} \subset \text{ch}(\mathcal{T}_1)$. Since $|\text{ch}(\mathcal{T}_1) \cap \{y, z\}| \leq |\text{ch}(\mathcal{T}_2) \cap \{y, z\}|$, we must have $\{y, z\} \subset \text{ch}(\mathcal{T}_2)$. We now consider two subcases.

(III-1): $\{y, z\}$ is a cherry in $\mathcal{T}_1$ or $\mathcal{T}_2$. Without loss of generality, assume $\{y, z\}$ is a cherry in $\mathcal{T}_1$. By symmetry, we may assume $\mathcal{T}_1 = [zy : c_1c_2x_1 \cdots x_k : c_3c_4]$. Since $\{y, z\} \subset \text{ch}(\mathcal{T}_2)$ but $\{y, z\}$ is not a cherry of $\mathcal{T}_2$ then – switching $y$ and $z$ and $c_3$ and $c_4$ if necessary – we can assume $C'_1 := \{c_3, z\}$ is a cherry in $\mathcal{T}_2$, and that the other cherry $C'_2$ in $\mathcal{T}_2$ must contain $y$ and one of the elements in $\{c_1, c_2\}$ (indeed, we must have $C'_2 = \{c_2, y\}$, as otherwise $C'_2 = \{c_1, y\}$ and so $\{X_1 \cup X_2, X_3\}$ is a maximum-agreement forest). But then $C'_1 := \{c_3, z\}$ and $C'_2 = \{c_2, y\}$ which implies $c_1zy$ – is a good character.
(III-2): \{y,z\} is not a cherry in \(T_1\) or \(T_2\). By symmetry, we may assume \(T_1 = [c_1 y : c_2 x_1 \cdots x_k c_4 : c_3 z]\). Note that we can further assume \(c_1 \not\in \text{ch}(T_1)\), as otherwise \(\{c_1, z\}\) is a cherry in \(T_2\), and hence \(c_1 c_2 z\)-, being convex on \(T_2\), is a good character. Similarly, we may assume \(c_3 \not\in \text{ch}(T_1)\) as otherwise \(\{c_3, y\}\) is a cherry in \(T_2\), and hence \(c_3 c_4 y\)- is a good character.

Therefore, we must have \(\text{ch}(T_2) = \{c_2, c_4, y, z\}\). Denoting the cherry in \(T_2\) containing \(y\) by \(C'_1\), and the cherry containing \(z\) by \(C'_2\), we have either \(C'_1 = \{c_2, y\}\) and \(C'_2 = \{c_4, z\}\), or \(C'_1 = \{c_4, y\}\) and \(C'_2 = \{c_2, z\}\). In both cases, the character \(\chi = c_1 y | c_3 z\)- is good. This completes the proof of Claim 1.

Now, since \(|\text{ch}(T_1)| = 4\) we can assume \(|\text{ch}(T_1) \cap X_1| \geq 2\). Moreover, in view of Claim 1, from now on we can assume without loss of generality that \(\text{ch}(T_1) \not\subseteq X_1\) and \(C_1 \subseteq X_1\) (and so \(|\text{ch}(T_1) \cap X_1| \leq 3\)). We now show that if \(T_1\) and \(T_2\) have the split \(X_1|(X \setminus X_1)\) in common then the theorem must hold.

Claim 2: If \(X_1|(X \setminus X_1)\) is a common split of \(T_1\) and \(T_2\), then \(d_p(T_1, T_2) \geq 2\).

Proof of Claim: For simplicity, put \(A = X_1\), \(B = X_2 \cup X_3\) and \(u = r_A\). Then \(T_1[A]\) and \(T_2[A]\) are two caterpillars on \(A \cup \{u\}\). Note first that \(C_1\) is a cherry in \(T_1[A]\), but not a cherry in \(T_2[A]\), as otherwise \(C_1\) would also be a cherry in \(T_2\), a contradiction. Hence \(d_p(T_1[A], T_2[A]) \geq 1\).

Since \(d_p(T_1[A], T_2[A]) \geq 1\), there exists a subset \(A' := \{x_1, x_2, x_3\} \subseteq A\) such that, for \(A^* := A' \cup \{u\}\), \(T_1[A]^* \neq T_2[A]^*\). In particular, by relabeling the elements \(x_i\) if necessary, we can assume \(T_1[A]^* = [ux_1 :: x_2 x_3]\) and \(T_2[A]^* = [ux_2 :: x_1 x_3]\). In addition, since \(d_p(T_1|B, T_2|B) \geq 1\) (as \(\{A, B\}\) is not a maximum-agreement forest), there exists a subset \(B' := \{y_1, y_2, y_3, y_4\} \subseteq B\) such that \(T_1|B' \neq T_2|B'\). In particular, by relabeling the elements \(y_i\) if necessary, we can assume \(T_1|B' = [y_1 y_2 :: y_3 y_4]\) and \(T_2|B' = [y_1 y_3 :: y_2 y_4]\).

It follows that, for \(X' = A' \cup B'\), we may assume without loss of generality that \(T_1|X'\) is obtained by attaching \(u\) in \(T_1[A]^*\) to the edge incident to \(y_1\) in \(T_1|B'\). In other words, we have \(T_1|X' = [x_2 x_3 : x_1 y_1 y_2 :: y_3 y_4]\).

Now, for \(1 \leq i \leq 4\), let \(e_i\) be the pendant edge incident to \(y_i\) in \(T_2|B'\), and denote the tree obtained by attaching \(u\) in \(T_1[A]^*\) to edge \(e_i\) by \(T'_i\). Consider the character \(\chi = x_2 x_3 y_3 y_4\)-. This is convex on \(T_2\). On the other hand, since \(T'_i\) contains a cherry \(\{x_1, x_3\}\) and the other cherry of \(T'_i\) is either \(\{y_2, y_4\}\) or \(\{y_1, y_3\}\), we have \(h(T'_i, \chi) \geq 2\) by Lemma 5.2, and hence \(d_p(T_1|X', T'_i) \geq 2\), for all \(1 \leq i \leq 4\). But then, as \(T_2|X' \in \{T'_1, \ldots, T'_4\}\), we have \(d_p(T_1, T_2) \geq d_p(T_1|X', T_2|X') \geq 2\), where the second inequality follows by Corollary 3.5. This completes the proof of Claim 2.

We now continue with the proof of the theorem, under the assumptions made up to the statement of Claim 2. Without loss of generality we assume \(\text{ch}(T_1) \cap X_2 \neq \emptyset\) from now on. The remainder of the proof is divided into two claims (Claims 3 and 4) according to whether or not \(\text{ch}(T_1)\) intersects \(X_3\). From now on we denote the two leaves contained in \(C_1\) by \(c_1\) and \(c_2\).

Claim 3: If \(X_3 \cap \text{ch}(T_1) \neq \emptyset\), then \(d_p(T_1, T_2) \geq 2\).

Proof of Claim: By Claim 1, without loss of generality we may assume \(|X_2| > 1\) and \(X_3 = \{z\}\) some \(z \in X\). Hence the cherry \(C_2\) in \(T_1\) not equal to \(C_1\) is equal to \(\{z, y_0\}\) for some \(y_0 \in X_2\). We now consider three cases.

(I): \(|X_2| = |X| - 3\). Then \(X_1 = \{c_1, c_2\} = C_1\) and \(T_1 = [c_1 c_2 : y_1 \cdots y_k : y_0 z]\) with \(X_2 = \{y_1, \ldots, y_k, y_0\}\). For some \(k \geq 3\). Let \(e_i, i \in \{0, 1, 2, k\}\), be the pendant edge in \(T_2|X_2 = [y_1 y_2 : \ldots \]
Claim 4: If $X_3 \cap \text{ch}(T_1) = \emptyset$, then $d_p(T_1, T_2) \geq 2$.

Proof of Claim: Let $X_3 = \{z_1, \ldots, z_t\}$ for some $t \geq 1$, and let $C_2 = \{c_3, c_4\}$ be the cherry in $T_1$ that is different from $C_1$. In view of Claim 1 we can assume $C_2 \subseteq X_2$. Since $T_1$ and $T_2$ share no common cherry, either $C_1 \subsetneq X_1$ or $C_2 \subsetneq X_2$ must hold, thus we may assume without loss of generality that $X_1 = \{c_1, c_2, x_1, \ldots, x_k\}$ for some $k \geq 1$.

Suppose first that $X_2 = \{c_3, c_4\}$. Then $c_3$ and $c_4$ is a pair of siblings in $T_2$, and so $X_1|(X_2 \cup X_3)$ is a split of $T_2$. By Claim 2, we may assume $X_1|(X_2 \cup X_3)$ is not a split of $T_1$. Thus $X_3 = \{z_1\}$ and the unique sibling of $c_3$ in $T_1$, which we denote by $x_k$, belongs to $X_1$. Now, one cherry in $T_2$ must contain $x_k$, and the other one is either equal to $\{c_3, z_1\}$ or $\{c_4, z_1\}$. Therefore, the character $\chi = c_3c_4x_k$ is good.

So, suppose $X_2 = \{c_3, c_4, y_1, \ldots, y_m\}$ for some $m \geq 1$. Since $C_1 \subseteq X_1$ and $C_2 \subseteq X_2$, by Claim 2 we can assume that $X_1|(X \setminus X_1)$ is a split of $T_1$ but not of $T_2$, and $X_2|(X \setminus X_2)$ is a split of $T_2$ but not of $T_1$. Hence $t = 1$ and so $T_1$ is obtained by attaching leaf $z$ to one of interior edges in $T_1|X_1 \cup X_2 = \{c_1c_2: x_1 \cdots x_k y_{m-1} \cdots y_1 : c_3c_4\}$. Since $z \in \text{ch}(T_2)$ implies $\emptyset \neq \text{ch}(T_2) \cap X_1$ for $1 \leq i \leq 3$, by Claim 3 we have $z \not\in \text{ch}(T_2)$. Thus, without loss of generality, we may further assume
that \( \mathcal{T}_2|_{X_1 \cup X_2} \) is obtained by connecting the pendant edge incident to \( c_1 \) in \( \mathcal{T}_2|_{X_1} \) to the pendant edge incident to \( c_3 \) in \( \mathcal{T}_2|_{X_2} \).

Consider the set \( A = \{c_1, c_2, x_k, z, y_m, c_3, c_4\} \). Then since \( \mathcal{T}_1 \) contains the split \( X_1|(X \setminus X_1) \) but not the split \( X_2|(X \setminus X_2) \), we must have \( \mathcal{T}_1|_{A} = \{c_1c_2 : zx_ky_m : c_3c_4\} \). Hence the character
\[
\chi = c_1c_2z|x_ky_mc_3c_4 \text{ on } A \text{ is convex on } \mathcal{T}_1|_{A}.
\]
On the other hand, \( \mathcal{T}_2 \) contains the split \( X_2|(X \setminus X_2) \) but not the split \( X_1|(X \setminus X_1) \). Therefore, as \( c_1 \) and \( c_3 \) are siblings in \( \mathcal{T}_2 \), and thus also in \( \mathcal{T}_2|_{A} \), it follows that \( \{P(c_2, x_m), P(c_1, c_3), P(z, c_4)\} \) is an ES-path system on \( \mathcal{T}_2|_{A} \). Therefore \( d_p(\mathcal{T}_1|_{A}, \mathcal{T}_2|_{A}) \geq 2 \), and so \( d_p(\mathcal{T}_1, \mathcal{T}_2) \geq 2 \), as required. This completes the proof of Claim 4 and also the theorem. \( \square \)

6. Unit neighborhoods in \( d_p \)

In this section, we present a result which can be used to characterize the set of trees that are at distance one from a given phylogenetic tree \( \mathcal{T} \) in the \( d_p \) metric (see Theorem 6.4). This set is also known as the \textit{unit neighborhood} of \( \mathcal{T} \). It can be helpful to understand such neighborhoods as they often form the basis for algorithms that search through phylogenetic trees to find some tree that optimizes some criterion, such as parsimony or likelihood (see e.g. [12, 20]).

We begin by considering a way to “reduce” a pair of trees which have a special subtree in common. To this end, suppose that \( (\mathcal{T}_1, \mathcal{T}_2) \) is a pair of distinct phylogenetic trees on \( X \) that contain a common non-trivial split \( A|B \) of \( X \) such that \( \mathcal{T}_1[B] = \mathcal{T}_2[B] \) holds. Then \( \mathcal{T}_i[A], \ i = 1, 2, \) can be regarded as the tree obtained from \( \mathcal{T}_i \) by replacing the subtree \( \mathcal{T}_i(B) \) by a single leaf \( r_A \) (see Fig. 5 for an illustration). In this case we shall say that \( (\mathcal{T}_1[A], \mathcal{T}_2[A]) \) is obtained from \( (\mathcal{T}_1, \mathcal{T}_2) \) by a \textit{subtree reduction} (cf. Rule 1 in [2, Section 3]). In addition, given an arbitrary pair \( (\mathcal{T}_1, \mathcal{T}_2) \) of distinct phylogenetic trees on \( X \), we shall say that \( (\tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_2) \) is a \textit{primitive pair} for \( (\mathcal{T}_1, \mathcal{T}_2) \) if it can be obtained from \( (\mathcal{T}_1, \mathcal{T}_2) \) by a sequence of subtree reductions and no further subtree reduction can be applied to \( (\tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_2) \) (or, equivalently, \( \tilde{\mathcal{T}}_1 \) and \( \tilde{\mathcal{T}}_2 \) share no common cherry). Note that it is straight-forward to see that only one primitive pair can be associated to \( (\mathcal{T}_1, \mathcal{T}_2) \) (in other words, if \( (\mathcal{T}_1^*, \mathcal{T}_2^*) \) is also a primitive pair for \( (\mathcal{T}_1, \mathcal{T}_2) \), then \( \tilde{\mathcal{T}}_i \) and \( \tilde{\mathcal{T}}_i^* \) are isomorphic as phylogenetic trees, \( i = 1, 2 \)).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{subtree_reduction.png}
\caption{An illustration of subtree reduction. The pair \((\mathcal{T}_1[A], \mathcal{T}_2[A])\) is obtained from \((\mathcal{T}_1, \mathcal{T}_2)\) by a subtree reduction with respect to the common subtree on leaf set \( B \).}
\end{figure}

We first show that the concept of primitive pair is closely related to the three metrics considered in this paper.

**Lemma 6.1.** Suppose that \( (\mathcal{T}_1, \mathcal{T}_2) \) is a pair of phylogenetic trees on \( X \). Then we have
\[
d_p(\mathcal{T}_1, \mathcal{T}_2) = d_p(\tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_2), \quad d_{tbr}(\mathcal{T}_1, \mathcal{T}_2) = d_{tbr}(\tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_2) \quad \text{and} \quad d_{spr}(\mathcal{T}_1, \mathcal{T}_2) = d_{spr}(\tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_2).
\]
Proof. Let \((T_1', T_2')\) be a pair of phylogenetic trees obtained from \((T_1, T_2)\) by a subtree reduction. Then by Theorem 4.2 we have \(d_p(T_1, T_2) = d_p(T_1', T_2')\). Since \((\tilde{T}_1, \tilde{T}_2)\) can be obtained from \((T_1, T_2)\) by a finite number of subtree reductions, it follows that \(d_p(T_1, T_2) = d_p(\tilde{T}_1, \tilde{T}_2)\), as required. The statements for \(d_{\text{mir}}\) and \(d_{\text{spr}}\) can be established using a similar argument because these two distances are also preserved by subtree reductions (cf. [2, Theorem 3.4]).

Now we present a result which shows that caterpillars naturally arise when considering trees which are \(d_p\) distance 1 apart.

**Lemma 6.2.** Suppose that \((T_1, T_2)\) is a pair of phylogenetic trees on \(X\) with \(d_p(T_1, T_2) = 1\). Then \(\tilde{T}_1\) and \(\tilde{T}_2\) are both caterpillars.

**Proof.** It suffices to show that if \(T_1\) and \(T_2\) are phylogenetic trees on \(X\) with \(d_p(T_1, T_2) = 1\) which have no common cherry, then \(T_1\) and \(T_2\) are both caterpillars.

Suppose this were not the case. Without loss of generality, we can assume that \(T_1\) contains three cherries, say \(\{c_1, c_2\}, \{c_3, c_4\}, \{c_5, c_6\}\), or, in other words, putting \(A := \{c_1, c_2, c_3, c_4, c_5, c_6\}\), \(T_1|_A\) is a phylogenetic tree with six leaves and three cherries. We now consider four cases.

(I): \(T_2|_A\) has three cherries in common with \(T_1|_A\). Then \(T_1|_A = T_2|_A\) and there must exist three elements in \(X\), say \(x_1, x_2, x_3\), such that \(x_i\) is adjacent to a vertex in the path \(P(c_{2i-1}, c_{2i})\) in \(T_2\) for all \(1 \leq i \leq 3\). Let \(B = A \cup \{x_1, x_2, x_3\}\). Then, by symmetry, we may assume that \(\{c_1, x_1\}, \{c_3, x_2\}, \{c_5, x_3\}\) are the three cherries contained in \(T_2|_B\). Consider the character \(\chi = c_1 c_2 c_3 c_4 c_5 c_6 x_1 x_2 x_3\). Then \(\chi\) is convex on \(T_2|_B\), and \(\{P(c_1, x_1), P(c_2, c_4), P(c_2, c_6), P(c_3, x_2), P(c_5, x_3)\}\) is an ES-path system. Hence by Theorem 2.3, we have

\[
d_p(T_1, T_2) \geq d_p(T_1|_B, T_2|_B) \geq h(T_2|_B, \chi) = l(T_2|_B, \chi) - |\chi| + 1 \geq 2,
\]

a contradiction.

(II): \(T_2|_A\) has two cherries in common with \(T_1|_A\). Without loss of generality, we may assume that these two cherries are \(\{c_1, c_2\}\) and \(\{c_3, c_4\}\). Then \(T_2|_A\) contains exactly two cherries. By switching \(c_5\) and \(c_6\) if necessary, we may assume that \(T_2|_A\) contains the split \(c_1 c_2 c_3 c_4 c_5 c_6\). In addition, using an argument similar to that used in Case (I), we may assume that there exist \(x_1, x_2 \in X\) such that \(\{c_1, x_1\}\) and \(\{c_3, x_2\}\) are the only two cherries contained in \(T_2|_B\), where \(B := A \cup \{x_1, x_2\}\). Consider the character \(\chi = c_1 x_1 c_3 x_2\). Then \(\chi\) is convex character on \(T_2|_B\), while \(\{P(c_1, c_2), P(c_3, c_4), P(c_5, x_1), P(c_5, x_2)\}\) is an ES-path system on \(T_1|_B\). As in Case (I) this leads to a contradiction.

(III): \(T_2|_A\) has one cherry in common with \(T_1|_A\). Without loss of generality, we may assume this cherry is \(\{c_1, c_2\}\). Then by symmetry we can assume \(T_2|_A\) contains another cherry \(\{c_3, c_5\}\). By an argument similar to that in Case (I), we may assume that there is an element \(x \in X\) such that \(\{c_1, x\}\) is a cherry contained in \(T_2|_B\), where \(B := A \cup \{x\}\). Consider the character \(\chi = c_1 x c_3 c_5 c_2 c_4 c_6\). Then \(\chi\) is a convex character on \(T_2|_B\), while \(\{P(c_1, c_2), P(c_3, c_4), P(c_5, x)\}\) is an ES-path system on \(T_1|_B\). As in Case (I) this leads to a contradiction.

(IV): \(T_2|_A\) and \(T_1|_A\) have no cherries in common. By symmetry we can assume that \(\{c_1, c_3\}\) and \(\{c_2, c_5\}\) are two cherries contained in \(T_2|_A\). Consider the character \(\chi = c_1 c_3 c_2 c_5 c_4 c_6\). Then \(\chi\) is convex character on \(T_2|_B\), while \(\{P(c_1, c_2), P(c_1, c_6), P(c_3, c_4), P(c_3, c_5)\}\) is an ES-path system on \(T_1|_B\). As in Case (I) this leads to a contradiction, which completes the proof of the lemma. \(\square\)
Using Theorem 5.1, the main result from the last section, we now show that if \(|X| \geq 6\) and two caterpillars \(T_1\) and \(T_2\) have no cherry in common, then they are \(d_p\) distance 1 apart if and only if they differ by precisely one TBR operation.

**Proposition 6.3.** Suppose \(|X| \geq 6\). Let \(T_1\) and \(T_2\) be two caterpillars on \(X\) that have no cherry in common. Then \(d_p(T_1, T_2) = 1\) if and only if \(d_{TBR}(T_1, T_2) = 1\).

**Proof.** Let \(T_1\) and \(T_2\) be two caterpillars as given in the theorem. By Lemma 2.1, it suffices to show that if \(d_p(T_1, T_2) = 1\) then \(d_{TBR}(T_1, T_2) = 1\). To this end, we shall show that if \(d_{TBR}(T_1, T_2) \geq 2\) then \(d_p(T_1, T_2) \geq 2\). We do this by induction on \(|X|\).

Suppose \(|X| = 6\). Assume \(T_1 = [c_1c_2 : c_3c_6 : c_3c_4]\). Let \(A\) be the split of \(X\) in \(T_2\) with \(|A| = 3\). Switching \(A\) and \(B\) if necessary, we may assume that \(|A \cap \{c_1, c_2, c_5\}| = 2\). Indeed, clearly we can assume \(|A \cap \{c_1, c_2, c_5\}| \geq 2\) and if \(|A \cap \{c_1, c_2, c_5\}| = 3\), then \(d_{TBR}(T_1, T_2) = 1\), a contradiction. By symmetry, we may further assume \(c_1 \in A\) and \(c_4 \notin A\).

Let \(C_1\) be the cherry of \(T_2\) that is contained in \(A\), and \(C_2\) the one contained in \(B\). If \(c_2 \in C_2\), then \(c_2 \notin A\) implies that either (i) \(A = \{c_1, c_2, c_6\}\) or (ii) \(A = \{c_1, c_2, c_3\}\). If (i) holds then, by symmetry, we may assume \(T_2 = [c_1c_6 : c_2c_4 : c_3c_5]\), and hence \(d_p(T_1, T_2) \geq 2\) in view of the character \(\chi = c_1c_2c_5|c_3c_4c_6\). If (ii) holds, then \(c_5 \in C_2\) as, otherwise \(d_{TBR}(T_1, T_2) = 1\) would hold, a contradiction. Hence we can again conclude that \(d_p(T_1, T_2) \geq 2\) by considering the character \(\chi = c_1c_2c_5|c_3c_4c_6\).

So suppose \(c_2 \in B\). Since for the character \(\chi = c_1c_3c_5|c_2c_4c_6\) we have \(h(T_1, \chi) = 2\), we can assume that \(c_3 \notin A\), as otherwise the proposition clearly follows. Therefore, we have \(A = \{c_1, c_5, c_6\}\). Hence \(B = \{c_2, c_3, c_4\}\), and so \(c_2 \in C_2\). In addition, we must have \(c_6 \in C_1\) as otherwise we would have \(d_{TBR}(T_1, T_2) = 1\), a contradiction. Hence, by considering character \(\chi = c_1c_2c_5|c_3c_4c_6\) we can again conclude that \(d_p(T_1, T_2) \geq 2\) holds. This completes the proof of the base case.

Now assume \(|X| = n \geq 7\) and the result holds for all pairs of caterpillars with \(n - 1\) leaves that do not share a common cherry. Note that if \(d_{TBR}(T_1, T_2) = 2\), then \(d_p(T_1, T_2) = 2\) by Theorem 5.1. Therefore, we may further assume \(d_{TBR}(T_1, T_2) \geq 3\). Fix a cherry \(\{x_1, x_2\}\) of \(T_1\) and denote the sibling of \(x_1\) in \(T_1\) by \(x_3\). Let \(x := x_1\) if \(\{x_1, x_3\}\) is a cherry of \(T_2\), and \(x := x_2\) otherwise. Put \(X' = X \setminus \{x\}\), and consider the tree \(T'_i = T_i|X',\ i = 1, 2\). Then by construction, \(T'_1\) and \(T'_2\) are a pair of caterpillars on \(X'\) with \(|X'| = n - 1\) which share no common cherry, and \(d_{TBR}(T'_1, T'_2) \geq d_{TBR}(T_1, T_2) - 1 \geq 2\). Therefore by induction, we have \(d_p(T'_1, T'_2) \geq 2\). By Corollary 3.5 we conclude that \(d_p(T_1, T_2) \geq 2\) holds, which completes the induction step, and the proof of the proposition.

Using the last proposition, we can now characterize pairs of trees that are at \(d_p\) distance 1 from one another.

**Theorem 6.4.** Suppose that \(T_1, T_2\) are phylogenetic trees on \(X\). Then \(d_p(T_1, T_2) = 1\) if and only if \(T_1, T_2\) are both caterpillars and either (i) \(d_{TBR}(T_1, T_2) = 1\), or (ii) \(d_{TBR}(T_1, T_2) = 2\) and \(|T_1| = |T_2| = 5\).

**Proof.** To establish the “only if” direction, suppose that \(T_1, T_2\) are two phylogenetic trees on \(X\) with \(d_p(T_1, T_2) = 1\). By Lemma 6.2 \(T_1\) and \(T_2\) are caterpillars. Hence, by Proposition 6.3, if \(|T_1| = |T_2| \geq 6\), then \(T_1\) and \(T_2\) satisfy (i). The theorem now follows since it is straight-forward to check that, for \(T'_1\) and \(T'_2\) two arbitrary distinct trees on a set \(Y\), if \(|Y| = 4\) then \(d_{TBR}(T'_1, T'_2) = 1\), and if \(|Y| = 5\) then \(d_{TBR}(T'_1, T'_2) \leq 2\).

To establish the other direction, suppose that \(T_1, T_2\) are two phylogenetic trees on \(X\) such that the pair \(T_1, T_2\) satisfies either (i) or (ii) as stated in the theorem. By Lemma 2.1 and Lemma 3.4, we have \(d_p(T_1, T_2) = 1\) in both cases. Therefore using Lemma 6.1, we conclude that \(d_p(T_1, T_2) = d_{TBR}(T_1, T_2) = 1\), as required.

\(\square\)
In [2, Lemma 2.7] it is shown that if \( \mathcal{T}_1, \mathcal{T}_2 \) are phylogenetic trees on \( X \) then \( d_{\text{spr}}(\mathcal{T}_1, \mathcal{T}_2) \leq 2d_{\text{tbr}}(\mathcal{T}_1, \mathcal{T}_2) \). Moreover, it is easy to check that any two phylogenetic trees with five leaves differ by at most two SPR operations. Thus, by Lemma 6.1 and Theorem 6.4 we have the following result concerning the SPR and TBR distance between two trees that are at \( d_p \) distance 1 from one another.

**Corollary 6.5.** Suppose that \( \mathcal{T}_1 \) and \( \mathcal{T}_2 \) are two phylogenetic trees on \( X \) with \( d_p(\mathcal{T}_1, \mathcal{T}_2) = 1 \). Then \( d_{\text{tbr}}(\mathcal{T}_1, \mathcal{T}_2) \leq d_{\text{spr}}(\mathcal{T}_1, \mathcal{T}_2) \leq 2 \).

Using Theorem 6.4 and the formula for the size of the unit neighborhood in \( d_{\text{tbr}} \) (see [14, Theorem 3.6]), we can also obtain a formula for the size of the unit neighborhood \( N_p(\mathcal{T}) := \{ \mathcal{T}' : d_p(\mathcal{T}, \mathcal{T}') = 1 \} \). For \( i = 2, 3 \), let \( V_i(\mathcal{T}) \) be the set of nodes with degree \( i \) in the tree obtained from \( \mathcal{T} \) by removing all the leaves and their incident edges.

**Corollary 6.6.** If \( \mathcal{T} \) is a phylogenetic tree on \( X \) with \( n = |X| \geq 4 \), then

\[
|N_p(\mathcal{T})| = 4\left( \sum |A| \cdot |B| \right) - 4(n - 2)(n - 3) + 2|V_2(\mathcal{T})| + 6|V_3(\mathcal{T})|,
\]

where the sum is taken over all non-trivial splits \( A|B \) of \( \mathcal{T} \).

**Proof.** Let \( N_{\text{tbr}}(\mathcal{T}) \) be the set consisting of those trees \( \mathcal{T}_1 \) with \( d_{\text{tbr}}(\mathcal{T}, \mathcal{T}_1) = 1 \). Then by Theorem 6.4 we have \( N_{\text{tbr}}(\mathcal{T}) \subseteq N_p(\mathcal{T}) \). In addition, for each \( \mathcal{T}_1 \in N_p(\mathcal{T}) - N_{\text{tbr}}(\mathcal{T}) \), both trees in the primitive pair \( (\mathcal{T}', \mathcal{T}_1) \) for \( (\mathcal{T}, \mathcal{T}_1) \) are trees with five leaves and \( d_{\text{tbr}}(\mathcal{T}', \mathcal{T}_1) = 2 \) holds. Since there are precisely two trees that are two TBR operations away from a phylogenetic tree with five leaves, we conclude that a vertex \( v \) in \( V_i(\mathcal{T}) \) (\( i = 2, 3 \)) will contribute a subset \( \varphi(v) \) of \( N_p(\mathcal{T}) - N_{\text{tbr}}(\mathcal{T}) \) with \( |\varphi(v)| = i(i - 1) \), and \( \varphi(v) \cap \varphi(v') = \emptyset \) for each \( v' \in V_2(\mathcal{T}) \cup V_3(\mathcal{T}) - \{v\} \). This implies \( |N_p(\mathcal{T}) - N_{\text{tbr}}(\mathcal{T})| = 2|V_2(\mathcal{T})| + 6|V_3(\mathcal{T})| \). The corollary follows since \( |N_{\text{tbr}}(\mathcal{T})| = 4\left( \sum |A| \cdot |B| \right) - 4(n - 2)(n - 3) \) holds by [14, Theorem 3.6].

Note that it is known (see [14]) that caterpillars and “complete trees” have the maximum and minimum sized neighborhoods relative to \( d_{\text{tbr}} \), respectively. It would be interesting to find out which trees have these properties relative to \( d_p \).

### 7. A connection with SPR and TBR metrics

In this section we prove the following theorem which provides a close connection between the metrics \( d_p \), \( d_{\text{spr}} \) and \( d_{\text{tbr}} \).

**Theorem 7.1.** Suppose that \( \mathcal{T} \) and \( \mathcal{T}' \) are two phylogenetic trees on \( X \). Then

\[
d_p(\mathcal{T}, \mathcal{T}') \leq d_{\text{tbr}}(\mathcal{T}, \mathcal{T}') \leq d_{\text{spr}}(\mathcal{T}, \mathcal{T}') \leq 2d_p(\mathcal{T}, \mathcal{T}').
\]

(4)

In addition, if \( |X| = 4m, m \geq 2 \), then there exist phylogenetic trees \( \mathcal{T}_1^m \) and \( \mathcal{T}_2^m \) on \( X \) such that

\[
\frac{d_{\text{tbr}}(\mathcal{T}_1^m, \mathcal{T}_2^m)}{d_p(\mathcal{T}_1^m, \mathcal{T}_2^m)} = 2 - \frac{1}{m}.
\]

(5)

**Proof.** The first inequality in Eq. (4) holds by Lemma 2.1 and the second one clearly holds by definition. To see that the third inequality holds, first note that we may assume \( d_p(\mathcal{T}, \mathcal{T}') \geq 1 \). Let \( t = d_p(\mathcal{T}, \mathcal{T}') \). Then, by definition, there exists a sequence of phylogenetic trees \( \mathcal{T}_0 := \mathcal{T}, \mathcal{T}_1, \ldots, \mathcal{T}_t := \mathcal{T} \) on \( X \) so that \( d_p(\mathcal{T}_i-1, \mathcal{T}_i) = 1 \) holds for \( 1 \leq i \leq t \). By Corollary 6.5, \( d_{\text{spr}}(\mathcal{T}_i-1, \mathcal{T}_i) \leq 2 \) for \( 1 \leq i \leq t \), and hence \( d_{\text{spr}}(\mathcal{T}, \mathcal{T}') \leq 2t \). Therefore \( d_{\text{spr}}(\mathcal{T}, \mathcal{T}') \leq 2d_p(\mathcal{T}, \mathcal{T}') \).
We now show that there exist trees $T_1^m$ and $T_2^m$ on $X$ for which Eq. (5) holds. For $m \geq 2$, let $T^m_1$ and $T^m_2$ be two phylogenetic trees on $X = \{1, 2, \ldots, 4m\}$ as depicted in Fig. 6. Let $A_i = \{4i - 3, 4i - 2, 4i - 1, 4i\}$ and $B_i = A_i - \{4i\}$, $1 \leq i \leq m$.

We first show that $d_p(T_1^m, T_2^m) = m$ holds. Indeed, consider the character $\chi = \{1, 2\}|\{3, 4\}|\{4m - 3, 4m - 2\}|\{4m - 1, 4m\}$ on $X$. Then $\chi$ is convex on $T_1^m$ and

$$\bigcup_{1 \leq i \leq m} \{P(4i - 2, 4i), P(4i - 3, 4i - 1)\}$$

is an ES-path system for $\chi$ on $T_2^m$. Hence $d_p(T_1^m, T_2^m) \geq m$ holds by Theorem 2.3. On the other hand, for each $2 \leq i \leq m, A_i\{X \setminus A_i\}$ is a split of both $T^m_1$ and $T^m_2$. Let $X' = \{1, 2, 3, 4, a_2, \ldots, a_m\}$ and consider the two caterpillars $T_1' = [12 : a_2 \cdots a_m : 34]$ and $T_2' = [13 : a_2 \cdots a_m : 24]$ on $X'$. Then by recursively applying Proposition 4.1 we have

$$d_p(T_1^m, T_2^m) \leq d_p(T_1', T_2') + \sum_{i=2}^{m} d_p(T_1^m[A_i], T_2^m[A_i]) \leq m.$$ 

Hence $d_p(T_1^m, T_2^m) = m$ as claimed.

Now, since $\{B_1, A_1 \setminus B_1, \ldots, B_m, A_m \setminus B_m\}$ is an agreement forest for $(T^m_1, T^m_2)$, we have $\text{MAF}(T^m_1, T^m_2) \leq 2m - 1$. Hence it remains to show that $\text{MAF}(T^m_1, T^m_2) \geq 2m - 1$ holds since by Eq. (1) we would then have $d_{\text{TRI}}(T^m_1, T^m_2) = \text{MAF}(T^m_1, T^m_2) = 2m - 1 = 2d_p(T^m_1, T^m_2) - 1$.

We shall show $\text{MAF}(T^m_1, T^m_2) \geq 2m - 1$ by induction. The base case $m = 2$ is clear. Now assume $\text{MAF}(T_1^{m-1}, T_2^{m-1}) \geq 2m - 3$ holds for some $m > 2$. Let $F = \{X_1, \ldots, X_k\}$ be a maximum agreement forest for $(T_1^m, T_2^m)$. Then by definition there exists at most one $X_i \in F$ such that $X_i \cap A_m \neq \emptyset$ and $X_i \setminus A_m \neq \emptyset$. We therefore consider two cases.

**Case I:** For each block $X_i$ in $F$, either $X_i \subset A_m$ or $X_i \cap A_m = \emptyset$. Let $I$ be the set consisting of those indices $i \in \{1, \ldots, k\}$ with $X_i \subset A_m$. Then $|I| \geq 2$ since $T^m_1[A_m] \neq T^m_2[A_m]$. Now consider the set $F'$ obtained from $F$ by deleting all blocks $X_i$ with $i \in I$. Then $F'$ is an agreement forest for $(T_1^{m-1}, T_2^{m-1})$ with $|F'| \leq |F| - 2$. Therefore by induction we have

$$\text{MAF}(T_1^m, T_2^m) = |F| \geq |F'| + 2 \geq \text{MAF}(T_1^{m-1}, T_2^{m-1}) + 2 \geq 2m - 1,$$

as required.

**Case II:** There exists a unique block, say $X_1$, in $F$ such that $X_1 \cap A_m \neq \emptyset$ and $X_1 \setminus A_m \neq \emptyset$, while for each $i \in \{2, \ldots, k\}$, we have either $X_i \subset A_m$ or $X_i \cap A_m = \emptyset$. Let $I$ be the set consisting of those indices $i \in \{1, \ldots, k\}$ with $X_i \subset A_m$. Fix an element $x$ in $X_1 \setminus A_m$ and consider the set $A' = X_1 \cap (\{x\} \cup A_m)$. Then $T^m_1|_{A'} = T^m_2|_{A'}$ implies $|X_1 \cap A_m| \leq 2$, and hence $|I| \geq 2$. Let $F'$ be
obtained from $\mathcal{F}$ by deleting all blocks $X_i$ with $i \in I$ and replacing $X_1$ by $X_1 \setminus A_m$. Then $\mathcal{F}'$ is an agreement forest for $(\mathcal{T}_1^{m-1}, \mathcal{T}_2^{m-1})$ with $|\mathcal{F}'| \leq |\mathcal{F}| - 2$. Using a similar argument as in Case I it follows that $\text{MAF}(\mathcal{T}_1^{m}, \mathcal{T}_2^{m}) \geq 2m - 1$, as required.

Clearly for the trees $\mathcal{T}_1^{m}, \mathcal{T}_2^{m}, m \geq 2$, constructed in the proof of the last theorem, we have

$$(2 - \frac{1}{m}) \leq \frac{d_{\text{spr}}(\mathcal{T}_1^{m}, \mathcal{T}_2^{m})}{d_p(\mathcal{T}_1^{m}, \mathcal{T}_2^{m})} \leq 2.$$ 

We conjecture that the second inequality is in fact an equality. More generally, it would be interesting to know whether for any $X$ with $|X| \geq 5$ there exist phylogenetic trees $\mathcal{T}, \mathcal{T}'$ on $X$ such that $d_{\text{spr}}(\mathcal{T}, \mathcal{T}') = 2d_p(\mathcal{T}, \mathcal{T}')$ holds.

8. The diameter of $d_p$

Given a metric $d$ on the set of phylogenetic trees on $X$, the diameter of $d$, denoted by $\text{diam}(X, d)$, is the maximum value of $d(\mathcal{T}, \mathcal{T}')$ over all pairs of phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on $X$. The exact values of $\text{diam}(X, d_{\text{ypn}})$ and $\text{diam}(X, d_{\text{spr}})$ are still unknown although some recent progress has been made on upper and lower bounds for these quantities (see e.g. [10]). Therefore, the following result and its corollary are of interest (recall that for any real number $x \in \mathbb{R}$, $\lceil x \rceil$ is the largest integer not greater than $x$ while $\lfloor x \rfloor$ is the smallest integer not smaller than $x$):

**Theorem 8.1.** Suppose $|X| = n \geq 4$ and let $k = \lfloor n \rfloor$. Then

$$\text{diam}(X, d_p) = \begin{cases} n - 2k + 1 & \text{if } n = k^2, \\ n - 2k & \text{otherwise.} \end{cases}$$

**Proof.** Let $X = \{x_1, x_2, \ldots, x_n\}$. In addition, put $\psi(n) = 1$ for $n = k^2$ and $\psi(n) = 0$ otherwise. Then we want to show $\text{diam}(X, d_p) = n - 2k + \psi(n)$. To do this, by Theorem 3.3 it suffices to show that the maximum value of $\rho_{\mathcal{T}}(\mathcal{T})$ over all phylogenetics trees $\mathcal{T}$ and $\mathcal{T}'$ on $X$ is $n - 2k + \psi(n)$.

To this end, first suppose that $\mathcal{T}$ and $\mathcal{T}'$ are two arbitrary phylogenetics trees on $X$. Let $\chi : X \to \mathcal{C}$ be a character on $X$ such that $\chi$ is convex on $\mathcal{T}'$ and $h(\mathcal{T}, \chi) = \rho_{\mathcal{T}'}(\mathcal{T})$ holds. Let $r$ be the number of elements in $\mathcal{C}$ and choose some $\alpha \in \mathcal{C}$ so that $|\chi^{-1}(\alpha)| \geq |\chi^{-1}(\alpha')|$ holds for all $\alpha' \in \mathcal{C}$. Then $|\chi^{-1}(\alpha)| \geq n/r$. Now, consider the extension $\overline{\chi}$ of $\chi$ to $\mathcal{T}$ defined by putting $\overline{\chi}(v) = \alpha$ for every interior vertex $v$ in $\mathcal{T}$. Then we have

$$l(\mathcal{T}, \chi) \leq \Delta(\overline{\chi}) = n - |\chi^{-1}(\alpha)| \leq n(1 - 1/r),$$

and hence

$$h(\mathcal{T}, \chi) \leq n(1 - \frac{1}{r}) - (r - 1) = (n + 1) - (r + \frac{n}{r}).$$

Note that $r + \frac{n}{r} \geq 2\sqrt{n}$ where the equality holds if and only if $r = \frac{n}{r}$, that is, $n = k^2$ and $r = k$. In other words, we have $[r + \frac{n}{r}] \geq 2k + 1 - \psi(n)$. This implies

$$\rho_{\mathcal{T}'}(\mathcal{T}) = h(\mathcal{T}, \chi) \leq \lfloor (n + 1) - (r + \frac{n}{r}) \rfloor = (n + 1) - \lceil r + \frac{n}{r} \rceil \leq n - 2k + \psi(n).$$

In light of this last inequality, it only remains to show that there exists a pair of phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on $X$ with $\rho_{\mathcal{T}'}(\mathcal{T}) = n - 2k + \psi(n)$. Since for any character $\chi$ on $X$ there always
exists a phylogenetic tree $T'$ on $X$ with $h(T', \chi) = 0$, it suffices to construct a phylogenetic tree $T$ and a character $\chi$ on $X$ with $h(T, \chi) = n - 2k + \psi(n)$.

To do this, let $T$ be the caterpillar $[x_1x_2 : x_3 \cdots x_{n-2} : x_{n-1}x_n]$. Now consider the character $\chi : X \to \{0, 1, \ldots, k - 1\}$ defined by $\chi(x_j) = j \mod k$ for $1 \leq j \leq n$. Then it remains to show

$$l(T, \chi) = n - k - 1 + \psi(n),$$

as this implies $h(T, \chi) = l(T, \chi) - k + 1 = n - 2k + \psi(n)$.

To establish (6), consider first the extension $\chi$ of $\chi$ defined by $\chi(v) = 1$ for every interior vertex $v$ of $T$. Then we have $\Delta(\chi) = n - k$ if $n = k^2$ and $\Delta(\chi) = n - k - 1$ otherwise. That is, we have

$$l(T, \chi) \leq \Delta(\chi) = n - k - 1 + \psi(n). \quad (7)$$

Conversely, for $1 \leq i < k$, put $A_i = \{(i - 1)k + 1, (i - 1)k + 2, \ldots, ik\}$, $T_i = T|_{A_i}$, $A_k = X_n \setminus \cup_{1 \leq i < k} A_i$ and $T_k = T|_{A_k}$. Then the restriction $\chi_i$ of $\chi$ to $A_i$ is a character on $A_i$, $1 \leq i \leq k$. Note that we have $l(T_i, \chi_i) = k - 1$ for $1 \leq i < k$, and $l(T_k, \chi_k) = (k - 1)$ if $n = k^2$ and $l(T_k, \chi_k) = (k - 1) + (n - k^2 - 1)$ otherwise. Now consider an optimal extension $\chi$ of $\chi$ to $T$. Since the restriction of $\chi$ to the tree $T_i$ ($1 \leq i \leq k$) – which we denote by $\chi|T_i$ – is also an extension of $\chi_i$ on $T_i$, we conclude

$$l(T, \chi) = \Delta(\chi) \geq \sum_{i=1}^{k} \Delta(\chi|T_i)$$

$$\geq \sum_{i=1}^{k} l(T_i, \chi_i)$$

$$= (k - 1)k + (n - k^2 + \psi(n))$$

$$= n - k + \psi(n).$$

Together with (7), this establishes (6), and hence completes the proof of the theorem. \hfill \Box

By Theorem 7.1 we immediately have:

**Corollary 8.2.** Suppose $|X| = n \geq 4$ and let $k = \lfloor n \rfloor$. Then

$$\text{diam}(X, d_{sPR}) \geq \text{diam}(X, d_{tBR}) \geq \begin{cases} n - 2\sqrt{n} + 1 & \text{if } n = k^2, \\ n - 2k & \text{otherwise}. \end{cases}$$

Note that to date the best known lower bound for $\text{diam}(X, d_{tBR})$ is $n - 2\lfloor \sqrt{n} \rfloor + 1$ (cf. [10]), and so the last result represents a rather modest improvement on this bound. More importantly, since we have taken a quite different approach to the one taken in [10], we hope that our new approach might ultimately lead to a way to determine an exact formula for $\text{diam}(X, d_{tBR})$ and $\text{diam}(X, d_{sPR})$.

**Acknowledgments:** We thank Peter Erdös, Mike Steel and László Székely for useful discussions. We also thank the anonymous referee for helpful comments. TW also thanks the Marie Curie Fellowship HUBI MTKD-CT-2006-042794 for supporting his visit to Alfréd Rényi Institute of Mathematics, where part of this work was done.

**References**


