Noname manuscript No.
(will be inserted by the editor)

Neighborhoods of trees in circular orderings

Sarah Baskowski · Vincent Moulton · Andreas Spillner · Taoyang Wu

November 17, 2014

Abstract In phylogenetics a common strategy used to construct an evolutionary tree for a set of species X is to search in the space of all such trees for one that optimizes some given score function (such as the minimum evolution, parsimony or likelihood score). As this can be computationally intensive, it was recently proposed to restrict such searches to the set of all those trees that are compatible with some circular ordering of the set X. To inform the design of efficient algorithms to perform such searches, it is therefore of interest to find bounds for the number of trees compatible with a fixed ordering in the neighborhood of a tree that is determined by certain tree operations commonly used to search for trees: the nearest neighbor interchange (NNI), the subtree prune and regraft (SPR) and the tree bisection and reconnection (TBR) operations. We show that the size of such a neighborhood of a binary tree associated to the NNI operation is independent of the tree's topology, but that this is not the case for the SPR and TBR operations. We also give tight upper and lower bounds for the size of the neighborhood of a binary tree for the SPR and TBR operations and characterize those trees for which these bounds are attained.

Keywords phylogenetic tree · nearest neighbor interchange · subtree prune and regraft · tree bisection and reconnection · NeighborNet · circular ordering

S. Baskowski
The Genome Analysis Centre, Norwich, United Kingdom,
E-mail: S.Bastkowski@uea.ac.uk

V. Moulton (🖂)
School of Computing Sciences, University of East Anglia, Norwich, United Kingdom
E-mail: vincent.moulton@cmp.uea.ac.uk

A. Spillner
Department of Mathematics and Computer Science, University of Greifswald, Germany.
E-mail: mail@andreas-spillner.de

T. Wu (🖂)
School of Computing Sciences, University of East Anglia, Norwich, United Kingdom
E-mail: taoyang.wu@uea.ac.uk

1 Introduction

In evolutionary biology, phylogenetic trees are often constructed so as to represent and understand the evolution of some set of species. Formally speaking, given a set X of species, a *phylogenetic tree* on X is a graph-theoretical tree with leaf set X that has no degree two vertices (see e.g. Figure 1(a)). Several methods have been developed to construct phylogenetic trees (see e.g. Lemey et al, 2009). A strategy often used for this purpose is to systematically search through the space of all phylogenetic trees with fixed leaf set X for one that optimizes some pre-specified criterion or *score* (for example, parsimony, minimum evolution or likelihood, cf. Felsenstein (2004, Chapter 4)). As tree-space is hyper-exponentially large, such searches often rely on locally optimizing the score in the *neighborhood* of a tree and, starting from some fixed tree, repeatedly choosing some tree in a neighborhood with a better score until no such tree can be found (see e.g. Whelan and Money, 2010). These neighborhoods are often defined by making a single tree modification or *operation*, the most commonly used operations being the *nearest neighbor interchange* (NNI), the *subtree prune and regraft* (SPR) and the *tree bisection and reconnection* (TBR) operations (Kubatko, 2008). Several results have been proven concerning properties of tree neighborhoods for these operations, including formulae for their size which can be useful for analyzing the run-time of search algorithms (see e.g. Allen and Steel, 2001; Humphries and Wu, 2013; Li et al, 1996).

Recently, following an approach suggested in Bryant (1996, 1997), it was proposed to restrict the search for optimal trees by utilizing circular orderings of the set X so as to reduce the time required for searching all of tree-space (Bastkowski et al, 2014). More specifically, a search based on dynamic programming was developed to find a tree amongst all those trees compatible with a fixed circular ordering of X that optimizes the so-called minimum evolution criterion. The orderings were obtained from phylogenetic networks constructed using the NeighborNet algorithm (Bryant and Moulton, 2004) (cf. Figure 1(b)). Moreover, it was found that the trees obtained in this way compared favorably with those obtained by using FastME (Desper and Gascuel, 2002), a leading program for searching for minimum evolution trees in the whole of tree-space.

In light of these findings, it could be useful to find ways to efficiently search for trees in circular orderings that maximize alternative scores, such as parsimony or likelihood. Note that the set of trees compatible with a fixed circular ordering has close links with the set of triangulations of a convex polygon (Sleator et al, 1988; De Loera et al, 2010), and that in Semple and Steel (2004) it is shown to be exponentially large. Moreover, in contrast to minimum evolution, the complexity of computing a tree in a circular ordering optimizing the parsimony or likelihood score is currently unknown. Even so, in principle local searches could still be performed, and so it is of interest to obtain bounds on the size of neighborhoods obtained when restricting the aforementioned operations to those trees that are compatible with some fixed circular ordering of X, which we call *circular neighborhoods* for short. Indeed, as well as providing bounds for run-times of local search algorithms using circular neighborhoods, the structural results we derive to analyze these neighborhoods could be useful for designing such algorithms. We now summarize the contents of the rest of the paper.

In the next section, we begin by reviewing some definitions and results concerning tree operations and circular orderings. Then, after deriving some useful characterizations for circular neighborhoods relative to the SPR and TBR operations in Section 3, we present formulae for the number of trees in the circular neighborhoods

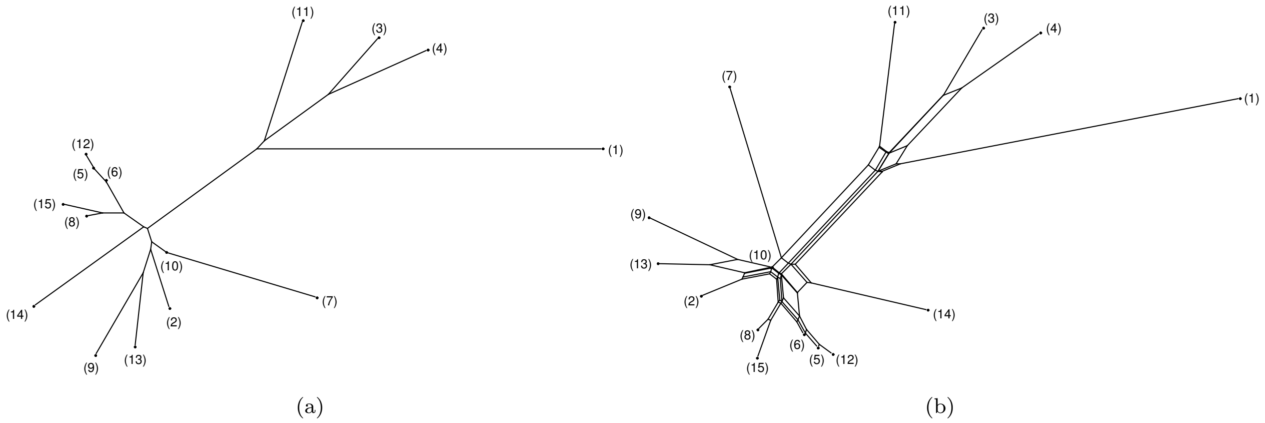

Fig. 1 (a) A phylogenetic tree $\mathcal{T}$ on $X = \{1, 2, 3, \ldots, 15\}$ relating 15 plant species from the genus *Solanum* (cf. Table 1 in the Appendix). The tree $\mathcal{T}$ was generated from aligned chloroplast DNA sequences presented in Tepe et al (2011) with the program FastME (Desper and Gascuel, 2002) by searching the space of all phylogenetic trees on X for one that optimizes the minimum evolution score using the NNI operation. (b) A phylogenetic network generated from the same sequence alignment by NeighborNet. The corresponding circular ordering $\pi = (1, 14, 12, 5, 6, 15, 8, 2, 13, 9, 10, 7, 11, 3, 4)$ of X is obtained by reading off the labels in clockwise order around the network. The tree $\mathcal{T}$ is compatible with the ordering π.

induced by the NNI (Theorem 1), SPR and TBR operations (Theorem 3) in Section 4. Note that formulae for the size of a neighborhood of a tree in the space of *all* phylogenetic trees are given in Robinson and Foulds (1981); Allen and Steel (2001); Humphries and Wu (2013). In particular, in tree-space the size of the NNI and SPR neighborhoods of a tree do not depend on the tree's topology (Robinson and Foulds, 1981; Allen and Steel, 2001), whereas the size of the TBR neighborhood does (Humphries and Wu, 2013).

Interestingly, as a consequence of our results, it follows that the size of a circular neighborhood of a tree does not depend on the tree's topology for the NNI operation, but it does for both the SPR and TBR operations. Thus it is of interest to characterize those trees whose induced circular SPR and TBR neighborhoods have maximum or minimum size, which we do in Section 5 (Theorems 4 and 5, respectively). In Section 6 we present characterizations for the set of bipartitions or *splits* of X that can be obtained from all trees in a circular neighborhood by removing an appropriate edge (Theorem 7 and Theorem 10). These sets were introduced and studied for full tree-space in Bryant (2004). In addition, we give a formula for the number of those trees in a circular neighborhood for which removing an edge yields a given split (Proposition 2). We conclude in the last section with a discussion of our results and some open problems.

2 Preliminaries

We begin by recalling some basic definitions concerning phylogenetic trees and related concepts (cf. Semple and Steel, 2003). From now on we will assume that X is a finite set with $|X| = n \geq 4$, unless stated otherwise.

Trees

Let $T = (V, E)$ be a *tree*, that is, an acyclic, connected graph with vertex set V and edge set E in which elements of E are 2-element subsets of V. A vertex of T is called a *leaf* if it has degree 1, and an *interior* *vertex* otherwise. Similarly, an edge of T is called a *pendant edge* if it is incident to a leaf, and an *interior edge* otherwise. The set of interior edges of T is denoted by $E^o(T)$. Unless explicitly stated otherwise, all trees considered here are *binary*, that is, all of their interior vertices have degree three. Given two vertices u and v in a tree T, the unique path between u and v is denoted by $P(u,v) = P_T(u,v)$. The set of edges contained in $P(u,v)$ is denoted by $E(P(u,v))$, and the set of interior edges by $E^o(P(u,v))$. The *length* of the path $P(u,v)$ is defined as the number of edges in $E(P(u,v))$. In addition, for two edges e_1 and e_2 in T, let $P(e_1,e_2) = P_T(e_1,e_2)$ be the unique shortest path in T that contains both e_1 and e_2. A *cherry* in T is a pair of leaves that are adjacent to the same interior vertex. A *caterpillar* is a binary tree in which each interior vertex is adjacent to some leaf. Clearly, a binary tree with more than three leaves is a caterpillar if and only if it contains precisely two cherries.

A *phylogenetic tree* $\mathcal{T}$ on X is a binary tree with leaf set X. Given a subset Y of X, we denote by $\mathcal{T}(Y)$ the smallest subtree of $\mathcal{T}$ (with any degree 2 vertices suppressed) that contains all leaves in Y. A *split of* X is a bipartition of X into two disjoint, non-empty subsets. These two subsets are referred to as the *blocks* of the split. A split with blocks A and B is denoted by $A|B$ $(= B|A)$. Every edge e of a phylogenetic tree $\mathcal{T}$ on X *induces* a unique split $S_e = S_e(\mathcal{T})$ of X into subsets A_e and B_e, that is, the subtrees $\mathcal{T}(A_e)$ and $\mathcal{T}(B_e)$ are precisely those obtained by removing e from $\mathcal{T}$. The set of all splits induced by the edges of $\mathcal{T}$ is denoted by $\Sigma(\mathcal{T})$. Moreover, two splits $S_1 = A_1|B_1$ and $S_2 = A_2|B_2$ of X are *compatible* if one of the intersections $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, $B_1 \cap A_1$ is empty, and *incompatible* otherwise. Note that, for any phylogenetic tree $\mathcal{T}$, the splits in $\Sigma(\mathcal{T})$ are necessarily pairwise compatible (Semple and Steel, 2003).

Operations

We recall the definitions of the three tree operations mentioned in the introduction, starting with TBR. Each TBR operation on a phylogenetic tree $\mathcal{T}$ is described in terms of a TBR *triplet*, that is, a triplet $(e_1, e_2; f)$ of edges in $\mathcal{T}$ such that (i) f is an edge in $P(e_1,e_2)$, (ii) $e_1 \cap e_2 = \emptyset$ and (iii) $|f \cap e_i| \in \{0,2\}$ for $i = 1,2$ (cf. Figure 2(a)). Note that $(e_1,e_2;f)$ and $(e_2,e_1;f)$ are considered as the same TBR triplet and we denote by $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ the set of all TBR triplets in $\mathcal{T}$. Now, for each triplet $\theta = (e_1,e_2;f)$ in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$, the tree $\theta(\mathcal{T})$, that is, the tree obtained by applying the corresponding TBR operation to $\mathcal{T}$, is constructed as follows. First both edges e_i, $i \in \{1,2\}$, are subdivided by a new vertex u_i into two edges e_i' and e_i'' such that the path $P(e_1'', e_2'')$ contains the edges e_1' and e_2' (cf. Figure 2(b)). Then, putting $h = e_i'$ if $f = e_i$ for some $i \in \{1,2\}$ and $h = f$ otherwise, edge h is removed. Finally, a new edge between u_1 and u_2 is added and any remaining vertices of degree 2 are suppressed to obtain a binary tree again (cf. Figure 2(c)). Note that the definition of a TBR operation above is tailored towards our purposes here, but is easily checked to be equivalent to those in the literature (e.g. in Allen and Steel, 2001; Humphries and Wu, 2013). Also note that Condition (iii) ensures that each TBR operation as previously defined corresponds to precisely one triplet in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$.

SPR and NNI operations are special cases of TBR operations. We call a TBR triplet $(e_1,e_2;f)$ in a phylogenetic tree $\mathcal{T}$ an SPR *triplet* if $f \in \{e_1,e_2\}$ and we call it an NNI *triplet* if, in addition, the path $P(e_1,e_2)$ contains precisely three edges. Moreover, we denote by $\mathcal{O}_{\text{SPR}}(\mathcal{T})$ and $\mathcal{O}_{\text{NNI}}(\mathcal{T})$ the set of all SPR triplets and all NNI triplets, respectively, in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$. Note that, by definition, we have $\mathcal{O}_{\text{NNI}}(\mathcal{T}) \subseteq \mathcal{O}_{\text{SPR}}(\mathcal{T}) \subseteq \mathcal{O}_{\text{TBR}}(\mathcal{T})$ and,

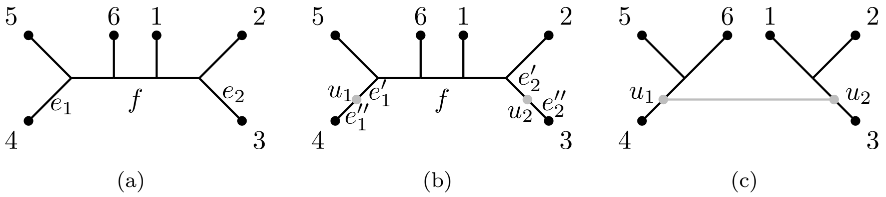

Fig. 2 (a) A phylogenetic tree $\mathcal{T}$ on $X = \{1, 2, \dots, 6\}$ with a TBR triplet $\theta = (e_1, e_2; f)$. (b) Edge e_i, $i \in \{1, 2\}$, is subdivided into two edges e_i' and e_i'' by a new vertex u_i (gray). (c) A new edge (gray) is added between u_1 and u_2, edge f is removed and the resulting vertices of degree 2 are suppressed to obtain the tree $\theta(\mathcal{T})$.

in general, these inclusions are strict. For example, the TBR operation depicted in Figure 2 is not an SPR operation. Also note that it was shown in Robinson (1971) that for two arbitrary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on the same leaf set X there exists a sequence of NNI operations that transform $\mathcal{T}$ into $\mathcal{T}'$. Therefore, for any operation OP $\in \{\text{NNI}, \text{SPR}, \text{TBR}\}$, the *distance* $d_{\text{OP}}(\mathcal{T}, \mathcal{T}')$ between $\mathcal{T}$ and $\mathcal{T}'$ defined as the minimum number of operations of type OP that suffice to transform $\mathcal{T}$ into $\mathcal{T}'$ is finite. It follows immediately from the definition that we have $d_{\text{NNI}}(\mathcal{T}, \mathcal{T}') \geq d_{\text{SPR}}(\mathcal{T}, \mathcal{T}') \geq d_{\text{TBR}}(\mathcal{T}, \mathcal{T}')$. Note that all of these distances are in fact metrics on tree-space and they have been studied in the literature (e.g. Li et al, 1996; Ding et al, 2011; Gordon et al, 2013).

Circular orderings

Here, we recall some definitions related to circular orderings, mainly following the ones used in Semple and Steel (2004). Let $\pi = (x_1, x_2, \dots, x_n)$ be a circular ordering of X. For a non-empty subset Y of X, let $\pi(Y)$ denote the ordering of Y obtained by restricting π to Y. Moreover, defining, for all $1 \leq i \leq j \leq n-1$, the subsets $X_{i,j} := \{x_k : i \leq k \leq j\}$ and $X_{i,j}^c := X - X_{i,j}$, we let $\Sigma^o(\pi) = \{X_{i,j} \,|\, X_{i,j}^c : 1 \leq i \leq j \leq n-1\}$ denote the set of all splits of X that are *compatible* with π. A collection Σ of splits on X is said to be *circular* (with respect to π) if $\Sigma \subseteq \Sigma^o(\pi)$ and, for any phylogenetic tree $\mathcal{T}$ on X with $\Sigma(\mathcal{T}) \subseteq \Sigma^o(\pi)$, we say that π is a *circular ordering* for $\mathcal{T}$. A *circular phylogenetic tree* $(\mathcal{T}, \pi)$ on X is a pair consisting of a phylogenetic tree $\mathcal{T}$ on X and a circular ordering π for $\mathcal{T}$. Intuitively speaking, a circular phylogenetic tree is a phylogenetic tree embedded in the plane. For any ordering π of X, the set of phylogenetic trees for which π is a circular ordering is denoted by $\mathscr{T}_\pi$. Note that, for any ordering π of X, $|\mathscr{T}_\pi|$ equals the Catalan number $\frac{1}{n-1}\binom{2n-4}{n-2}$ (Semple and Steel, 2004, Proposition 3.1) and, for later use, we also recall the following result.

Theorem 1 *(Semple and Steel, 2004, Theorem 3.4) Let $\pi = (x_1, \dots, x_n)$ be an ordering of X and let $\mathcal{T}$ be a phylogenetic tree on X. Then π is a circular ordering for $\mathcal{T}$ if and only if, for all subsets $Y \subseteq X$ of size four, $\pi(Y)$ is a circular ordering for $\mathcal{T}(Y)$.*

In the following we will sometimes be interested in circular phylogenetic trees that only differ by relabeling their leaves. To make this more precise, let $\mathcal{T}$ be a phylogenetic tree on X and $\pi = (x_1, x_2, \dots, x_n)$ a circular ordering of X. Given a bijection κ from X to X, we denote by $\kappa(\mathcal{T})$ the phylogenetic tree obtained from $\mathcal{T}$ by relabeling the leaves of $\mathcal{T}$ using the map κ. In addition, we define two circular phylogenetic trees $(\mathcal{T}, \pi)$ and $(\mathcal{T}', \pi')$ to be *relabeling-equivalent*, denoted by $(\mathcal{T}, \pi) \sim_{re} (\mathcal{T}', \pi')$, if there exists a bijection κ from X to

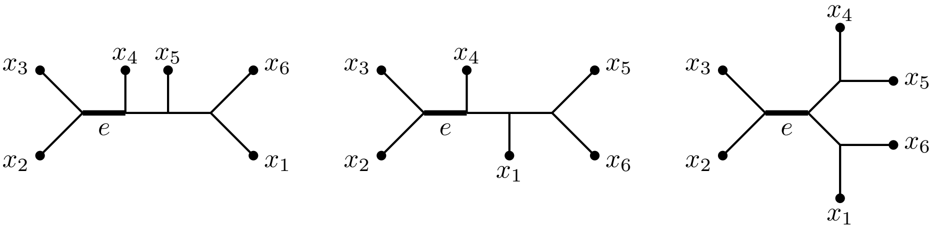

Fig. 3 Representatives of the three different $\sim_{re}$-equivalence classes of circular phylogenetic trees on $X = \{x_1, \ldots, x_6\}$. The ordering $(x_1, x_2, x_3, x_4, x_5, x_6)$ is a circular ordering for any one of these phylogenetic trees and the canonical index pair for the bold edge e is $(1, 3)$ in each tree.

X such that $\mathcal{T}' = \kappa(\mathcal{T})$, and $\pi' = (\kappa(x_1), \kappa(x_2), \ldots, \kappa(x_n))$. Note that, for $n = 4$, 5, 6, 7 and 8 the number of $\sim_{re}$-equivalence classes of circular phylogenetic trees on X are 1, 1, 3, 5 and 13, respectively (cf. Figure 3 for $n = 6$ and the figures in the Appendix for $n = 7, 8$.).

We end this section with introducing two indices on circular phylogenetic trees that will be used later. Let $(\mathcal{T}, \pi)$ be a circular phylogenetic tree on X. For any $1 \leq i \leq n$, the path $P_i^* = P_{i,\pi}^* = P_{\mathcal{T}}(x_i, x_{i+1})$ is called a *canonical path* in $\mathcal{T}$. Here, and in the remainder of this paper, we will use the convention that $x_{n+1} = x_1$. Note that every edge e of $\mathcal{T}$ is contained in precisely two canonical paths (Semple and Steel, 2004, Theorem 3.2) and we call the unique pair (i, j) with $e \in E(P_{i,\pi}^*) \cap E(P_{j,\pi}^*)$ the *canonical index pair* for e (cf. Figure 3). Putting $P_{e,\pi}^+ = P_{\mathcal{T}}(x_i, x_{j+1})$ and $P_{e,\pi}^- = P_{\mathcal{T}}(x_{i+1}, x_j)$, we define the *$\alpha$-index*

$$\alpha(\mathcal{T}, \pi) := \sum_{e \in E(\mathcal{T})} \left(|E(P_{e,\pi}^+)| + |E(P_{e,\pi}^-)|\right)$$

and the *β-index*

$$\beta(\mathcal{T}, \pi) := \sum_{e \in E^o(\mathcal{T})} \Big(|E(P_{e,\pi}^+)| - 2\Big) \cdot \Big(|E(P_{e,\pi}^-)| - 2\Big)$$

of $(\mathcal{T}, \pi)$. Note that, for any two circular phylogenetic trees $(\mathcal{T}, \pi)$ and $(\mathcal{T}', \pi')$ with $(\mathcal{T}, \pi) \sim_{re} (\mathcal{T}', \pi')$, the equalities $\alpha(\mathcal{T}, \pi) = \alpha(\mathcal{T}', \pi')$ and $\beta(\mathcal{T}, \pi) = \beta(\mathcal{T}', \pi')$ hold. Moreover, in view of $\min\{|E(P_{e,\pi}^+)|, |E(P_{e,\pi}^-)|\} \geq 2$ for all $e \in E^o(\mathcal{T})$, we have $\beta(\mathcal{T}, \pi) \geq 0$.

3 Circular tree operations

In this section, we prove some key results that will help us to derive formulae for the size of circular neighborhoods. More specifically, for any circular ordering $\pi = (x_1, \ldots, x_n)$ and any $\mathcal{T} \in \mathscr{T}_\pi$, we define the set $\mathcal{O}_{\text{TBR}}^\pi(\mathcal{T}) = \{\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T}) : \theta(\mathcal{T}) \in \mathscr{T}_\pi\}$ consisting of those TBR operations that preserve π. The elements of $\mathcal{O}_{\text{TBR}}^\pi(\mathcal{T})$ will be referred to as *circular* TBR *operations* on $\mathcal{T}$ (with respect to π). Similarly, we consider the set $\mathcal{O}_{\text{SPR}}^\pi(\mathcal{T})$ of *circular* SPR *operations* and the set $\mathcal{O}_{\text{NNI}}^\pi(\mathcal{T})$ of *circular* NNI *operations* on $\mathcal{T}$. The following theorem gives a concise characterization of circular TBR operations.

Theorem 2 *Let $(\mathcal{T}, \pi)$ be a circular phylogenetic tree on X and $\theta = (e_1, e_2; f) \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$. Then we have $\theta \in \mathcal{O}_{\text{TBR}}^\pi(\mathcal{T})$ if and only if $\{e_1, e_2\} \subseteq E(P_{f,\pi}^+) \cup E(P_{f,\pi}^-) \cup \{f\}$.*

Proof For simplicity, let $\mathcal{T}' = \theta(\mathcal{T})$. First assume that θ is not circular on $\mathcal{T}$. Then, by Theorem 1, there exists a 4-element subset $Y = \{x_1, x_2, x_3, x_4\} \subseteq X$ such that $\pi(Y)$ is not a circular ordering for $\mathcal{T}'(Y)$.

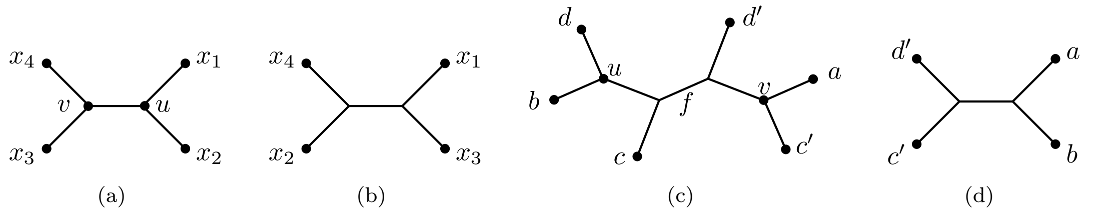

Fig. 4 The constructions used in the proof of Theorem 2. (a) The phylogenetic tree $\mathcal{T}(Y)$ for $Y = \{x_1, x_2, x_3, x_4\}$. (b) The tree $\mathcal{T}'(Y) = [\theta(\mathcal{T})](Y)$. (c) The phylogenetic tree $\mathcal{T}(Z)$ for $Z = \{a, b, c, c', d, d'\}$. (d) The phylogenetic tree $\mathcal{T}'(\{a, b, c', d'\})$.

Without loss of generality we assume that $\pi(Y) = (x_1, x_2, x_3, x_4)$ and that $\mathcal{T}(Y)$ and $\mathcal{T}'(Y)$ are as depicted in Figures 4(a) and (b), respectively. Let u and v denote the two endpoints of the path consisting of the edges in $E(P_{\mathcal{T}}(x_1, x_3)) \cap E(P_{\mathcal{T}}(x_2, x_4))$ such that u is contained in $P_{\mathcal{T}}(x_1, x_2)$ (see Figure 4(a)). Note that if $f \in E(P_{\mathcal{T}}(u, v))$, then by $S_f(\mathcal{T}) \in \Sigma(\mathcal{T}')$ we have $\{x_1, x_2\}|\{x_3, x_4\} \in \Sigma(\mathcal{T}'(Y))$, a contradiction. This leads to $f \notin E(P_{\mathcal{T}}(u, v))$, and hence we may assume without loss of generality that $f \in E(P_{\mathcal{T}}(x_1, u)$. By swapping e_1 and e_2 if necessary, we can further assume that $e_1 \in E(P_{\mathcal{T}}(x_1, u))$. Together with $\{x_1, x_3\}|\{x_2, x_4\} \in \Sigma(\mathcal{T}'(Y))$, this implies $e_2 \in E(P_{\mathcal{T}}(x_3, v)$, and therefore, we must have $e_2 \notin E(P^+_{f,\pi}) \cup E(P^-_{f,\pi}) \cup \{f\}$, as required.

To establish the other implication, assume that $e_2 \notin E(P^+_{f,\pi}) \cup E(P^-_{f,\pi}) \cup \{f\}$. We consider a subset $Z = \{a, b, c, c', d, d'\} \subseteq X$ with $P_{\mathcal{T}}(c, d) = P^+_{f,\pi}$, $P_{\mathcal{T}}(c', d') = P^-_{f,\pi}$ and $\{e_1, e_2\} \subseteq P_{\mathcal{T}}(a, b)$ and assume without loss of generality that $\pi(Z) = (a, c', c, b, d, d')$. The phylogenetic tree $\mathcal{T}(Z)$ is depicted in Figure 4(c). Let u and v denote the endpoints of the path in $\mathcal{T}$ that consists of the edges in $E(P_{\mathcal{T}}(a, b)) \cap E(P_{\mathcal{T}}(c', d))$. In view of $e_2 \notin E(P^+_{f,\pi}) \cup E(P^-_{f,\pi}) \cup \{f\}$ we assume without loss of generality that $e_2 \in E(P_{\mathcal{T}}(a, v))$. Note that this implies a, c', d' are pairwise distinct. Moreover, we must either have $e_1 \in E(P_{\mathcal{T}}(b, u))$ or $e_1 \in E(P^+_{f,\pi})$ (note that, in the latter case, elements b and d coincide). The position of e_1 and e_2 in $(\mathcal{T}, \pi)$ implies that the phylogenetic tree $\mathcal{T}'(\{a, b, c', d'\})$ must be as depicted in Figure 4(d). But $\pi(\{a, b, c', d'\}) = (a, c', b, d')$ is clearly not a circular ordering for $\mathcal{T}'(\{a, b, c', d'\})$ and, therefore, again by Theorem 1, π is not a circular ordering for $\mathcal{T}'$.

As a direct consequence of Theorem 2, we also obtain the following characterization of circular SPR operations.

Corollary 1 *Let $(\mathcal{T}, \pi)$ be circular phylogenetic tree on X and $\theta = (e_1, e_2; f) \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$. Then we have $\theta \in \mathcal{O}^{\pi}_{\text{SPR}}(\mathcal{T})$ if and only if $\{f\} \subset \{e_1, e_2\} \subseteq E(P^+_{f,\pi}) \cup E(P^-_{f,\pi}) \cup \{f\}$.*

4 The size of tree operation neighborhoods

We now turn to deriving formulae for the size of a circular neighborhood of a tree relative to NNI, SPR and TBR operations. Using the notation from the previous sections, the TBR *neighborhood* of a phylogenetic tree $\mathcal{T}$ is the set $N_{\text{TBR}}(\mathcal{T}) = \{\theta(\mathcal{T}) : \theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})\}$ consisting of all trees that are precisely one TBR operation away from $\mathcal{T}$. The NNI *neighborhood* $N_{\text{NNI}}(\mathcal{T})$ and the SPR *neighborhood* $N_{\text{SPR}}(\mathcal{T})$ of $\mathcal{T}$ are defined analogously; we clearly have $N_{\text{NNI}}(\mathcal{T}) \subseteq N_{\text{SPR}}(\mathcal{T}) \subseteq N_{\text{TBR}}(\mathcal{T})$. Moreover, it is known that, for any phylogenetic tree $\mathcal{T}$ on X with $|X| = n$, we have $|N_{\text{NNI}}(\mathcal{T})| = 2n - 6$ (Robinson, 1971), $|N_{\text{SPR}}(\mathcal{T})| = 2(n - 3)(2n - 7)$ (Allen and Steel,

2001) and

$$|N_{\text{TBR}}(\mathcal{T})| = -(4n-2)(n-3) + 4 \sum_{A|B \in \Sigma^*(\mathcal{T})} |A| \cdot |B|,$$

where $\Sigma^*(\mathcal{T}) = \{A|B \in \Sigma(\mathcal{T}) : |A| \geq 2, |B| \geq 2\}$ (Humphries and Wu, 2013). In particular, it follows that $|N_{\text{TBR}}(\mathcal{T})|$ depends on $\mathcal{T}$ while $|N_{\text{NNI}}(\mathcal{T})|$ and $|N_{\text{SPR}}(\mathcal{T})|$ do not.

Now, defining, for any OP $\in \{\text{NNI}, \text{SPR}, \text{TBR}\}$ and any circular phylogenetic tree $(\mathcal{T}, \pi)$, the *circular neighborhood* $N^{\pi}_{\text{OP}}(\mathcal{T}) = N_{\text{OP}}(\mathcal{T}) \cap \mathscr{T}_{\pi}$ of $\mathcal{T}$ with respect to π, we derive formulae for the size of $N^{\pi}_{\text{OP}}(\mathcal{T})$. We start with the NNI operation. The following result is a direct consequence of well-known properties of triangulations of a convex polygon (see, e.g. De Loera et al, 2010). For the convenience of the reader, we include a short self-contained proof.

Lemma 1 *Let $(\mathcal{T}, \pi)$ be a circular phylogenetic tree on X with $n = |X| \geq 4$. Then we have $|N^{\pi}_{\text{NNI}}(\mathcal{T})| = n-3$ and, for each tree $\mathcal{T}' \in N^{\pi}_{\text{NNI}}(\mathcal{T})$, there are precisely four distinct operations θ in $\mathcal{O}^{\pi}_{\text{NNI}}(\mathcal{T})$ with $\theta(\mathcal{T}) = \mathcal{T}'$.*

Proof First note that, for each NNI operation $\theta = (e_1, e_2; f) \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$, there exists a unique edge in $E(P(e_1, e_2)) - \{e_1, e_2\}$, which we will denote by $\kappa(\theta)$, and this edge is necessarily an interior edge of $\mathcal{T}$. Since $\Sigma(\mathcal{T}) - \Sigma(\theta(\mathcal{T}))$ contains precisely the split $S_{\kappa(\theta)}(\mathcal{T})$, we know that if $\kappa(\theta) \neq \kappa(\theta')$ for two NNI operations θ and θ', then also $\theta(\mathcal{T}) \neq \theta'(\mathcal{T})$. In view of the fact that, for each interior edge e in $\mathcal{T}$, there are precisely four NNI operations $\theta \in \mathcal{O}^{\pi}_{\text{NNI}}(\mathcal{T})$ so that $\kappa(\theta) = e$, this establishes the second assertion in the lemma. Using, in addition, the fact that there are $n-3$ interior edges in a phylogenetic tree with n leaves, we obtain $|N^{\pi}_{\text{NNI}}(\mathcal{T})| = n-3$, as required.

Note that this result implies that $|N^{\pi}_{\text{NNI}}(\mathcal{T})|$ does not depend on $\mathcal{T}$, just as with NNI neighborhoods in tree-space. As we shall see this is not the case for SPR and TBR neighborhoods. To this end, we first present a useful technical lemma.

Lemma 2 *For any* OP $\in \{\text{SPR}, \text{TBR}\}$ *and any circular phylogenetic tree $(\mathcal{T}, \pi)$ on X with $n = |X| \geq 4$, we have $|N^{\pi}_{\text{OP}}(\mathcal{T})| = |\mathcal{O}^{\pi}_{\text{OP}}(\mathcal{T})| - 3(n-3)$.*

Proof Let OP $\in \{\text{SPR}, \text{TBR}\}$ and $\theta, \theta' \in \mathcal{O}^{\pi}_{\text{OP}}(\mathcal{T})$ be two distinct operations with $\theta(\mathcal{T}) = \theta'(\mathcal{T})$. Then by Lemma 3.1 in Humphries and Wu (2013), we must have $\theta, \theta' \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$. Together with $\theta(\mathcal{T}) = \theta(\mathcal{T}') \in \mathscr{T}_{\pi}$, it follows that $\theta, \theta' \in \mathcal{O}^{\pi}_{\text{NNI}}(\mathcal{T})$ and, therefore, $|N^{\pi}_{\text{OP}}(\mathcal{T})| = |\mathcal{O}^{\pi}_{\text{OP}}(\mathcal{T})| - 3|N^{\pi}_{\text{NNI}}(\mathcal{T})| = |\mathcal{O}^{\pi}_{\text{OP}}(\mathcal{T})| - 3(n-3)$, where the last equality follows from Lemma 1.

Now, based on Lemma 2, we derive formulae for the size of the circular SPR and circular TBR neighborhoods in terms of the α- and β-indices defined in Section 2.

Theorem 3 *Given a circular phylogenetic tree $(\mathcal{T}, \pi)$ on X with $n = |X| \geq 4$, we have*

$$|N^{\pi}_{\text{SPR}}(\mathcal{T})| = \alpha(\mathcal{T}, \pi) - 9n + 21 \quad \text{and} \quad |N^{\pi}_{\text{TBR}}(\mathcal{T})| = \alpha(\mathcal{T}, \pi) + \beta(\mathcal{T}, \pi) - 9n + 21.$$

Proof First we consider the circular SPR neighborhood of $(\mathcal{T}, \pi)$. Recall that, for each $(e_1, e_2; f) \in \mathcal{O}^{\pi}_{\text{SPR}}(\mathcal{T})$, we must have $f \in \{e_1, e_2\}$. Therefore, swapping e_1 and e_2 if necessary, we can assume that $e_1 = f$. We distinguish between the two cases that f is a pendant and that f is an interior edge.

If f is a pendant edge we can assume without loss of generality that $|E(P^+_{f,\pi})| \geq 2$ and $|E(P^-_{f,\pi})| = 0$. Then, by Theorem 2, $(e_1, e_2; f) \in \mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})$ if and only if $e_2 \in E(P^+_{f,\pi})$ and $e_2 \cap f = \emptyset$. Since there are $|E(P^+_{f,\pi})| - 2$ edges in $E(P^+_{f,\pi})$ that are not incident with f, the number of operations in $\mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})$ whose last coordinate is f is $|E(P^+_{e,\pi})| + |E(P^-_{e,\pi})| - 2$.

Similarly, if f is an interior edge then e_2 can be any edge contained in $E(P^+_{f,\pi}) \cup E(P^-_{f,\pi})$ that is not incident with f. Clearly, the number of edges satisfying this condition and, thus, also the number of operations in $\mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})$ whose last coordinate is f is $|E(P^+_{f,\pi})| + |E(P^-_{f,\pi})| - 4$.

In summary, since there are $n-3$ interior edges and n pendant edges in $\mathcal{T}$, we have

$$|\mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})| = -4(n-3) - 2n + \sum_{e \in E(\mathcal{T})} (|E(P^+_{f,\pi})| + |E(P^-_{f,\pi})|) = \alpha(\mathcal{T}, \pi) - 6n + 12,$$

which, together with Lemma 2, implies $|N^\pi_{\text{SPR}}(\mathcal{T})| = \alpha(\mathcal{T}, \pi) - 9n + 21$.

Next we consider the circular TBR neighborhood of $(\mathcal{T}, \pi)$. We first count the operations $\theta = (e_1, e_2; f) \in \mathcal{O}^\pi_{\text{TBR}}(\mathcal{T}) - \mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})$. Note that, for any such operation, the last coordinate f must be an interior edge on the path $P(e_1, e_2)$. Now, for each interior edge f in $\mathcal{T}$, by Theorem 2 and swapping e_1 and e_2 if necessary, we have $(e_1, e_2; f) \in \mathcal{O}^\pi_{\text{TBR}}(\mathcal{T}) - \mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})$ if and only if $e_1 \in P^+_{f,\pi}$, $e_2 \in P^-_{f,\pi}$ and $e_1 \cap f = \emptyset = e_2 \cap f$. Clearly, the number of pairs of edges e_1 and e_2 satisfying this condition is $(|E(P^+_{f,\pi})| - 2) \cdot (|E(P^-_{f,\pi})| - 2)$. In other words, this is also the number of operations in $\mathcal{O}^\pi_{\text{TBR}}(\mathcal{T}) - \mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})$ whose last coordinate is f, implying that $|\mathcal{O}^\pi_{\text{TBR}}(\mathcal{T}) - \mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})| = \beta(\mathcal{T}, \pi)$. Thus, in view of $|\mathcal{O}^\pi_{\text{SPR}}(\mathcal{T})| = \alpha(\mathcal{T}, \pi) - 6n + 12$, we obtain $|\mathcal{O}^\pi_{\text{TBR}}(\mathcal{T})| = \alpha(\mathcal{T}, \pi) + \beta(\mathcal{T}, \pi) - 6n + 12$ which, together with Lemma 2, yields $|N^\pi_{\text{TBR}}(\mathcal{T})| = \alpha(\mathcal{T}, \pi) + \beta(\mathcal{T}, \pi) - 9n + 21$.

5 Tight bounds for the size of SPR and TBR neighborhoods

In the last section, we saw that the size of the circular neighborhoods $N^\pi_{\text{SPR}}(\mathcal{T})$ and $N^\pi_{\text{TBR}}(\mathcal{T})$ depend on the topology of the tree $\mathcal{T}$. In this section, we establish bounds for the size of these neighborhoods and characterize those circular phylogenetic trees for which they are tight. We begin by presenting an approach to calculating the α-index. To this end, for a circular phylogenetic tree $(\mathcal{T}, \pi)$ on X we define

$$\alpha^*(T, \pi) = \sum_{i=1}^{n} |E(P^*_i)|^2,$$

where $\{P^*_i\}_{1 \leq i \leq n}$ is the set of canonical paths in $\mathcal{T}$.

Lemma 3 *Given a circular phylogenetic tree $(\mathcal{T}, \pi)$ on X, we have $\alpha(T, \pi) = \alpha^*(T, \pi) - 4n + 6$.*

Proof Let $\{e_1, \ldots, e_m\}$ be the edge set of $\mathcal{T}$, where $m = 2n - 3$. Consider the following index encoding the incidence between edges and canonical paths

$$\delta_{i,k} = \begin{cases} 1, & \text{if } e_k \in E(P^*_i), \\ 0, & \text{otherwise}, \end{cases}$$

for $1 \leq i \leq n$ and $1 \leq k \leq m$. Since each edge e occurs in exactly two of the canonical paths in $\{P^*_1, P^*_2, \ldots, P^*_n\}$, we have $\sum_{i=1}^n \delta_{i,k} = 2$ for each k. On the other hand, we know that $\sum_{k=1}^m \delta_{i,k} = |E(P^*_i)|$ holds for $1 \leq i \leq n$.

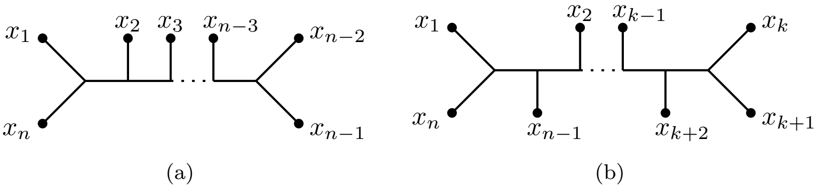

Fig. 5 Two circular caterpillars in $\mathscr{T}_\pi$ with $\pi = (x_1, \dots, x_n)$ for an even integer $n \geq 6$: (a) a skew caterpillar and (b) a centipede (where $k = \lfloor n/2 \rfloor$).

Now, fix some $k \in \{1, \dots, m\}$ and let (j, j') be the canonical index pair for e_k. Then we have

$$|E(P^+_{e_k,\pi})| + |E(P^-_{e_k,\pi})| = |E(P^*_j)| - 1 + |E(P^*_{j'})| - 1 = \sum_{i=1}^{n} \delta_{i,k}(|E(P^*_i)| - 1).$$

Here the first equality follows from $E(P^+_{e_k,\pi}) \cup E(P^-_{e_k,\pi}) = \big(E(P^*_j) \cup E(P^*_{j'})\big) \backslash \{e_k\}$, and the second one from $\delta_{i,k} = 1$ if and only if $i \in \{j, j'\}$. Therefore, we have

$$\begin{aligned}\alpha(\mathcal{T}, \pi) &= \sum_{k=1}^{m}\sum_{i=1}^{n} \delta_{i,k}(|E(P^*_i)| - 1) = \sum_{i=1}^{n}\sum_{k=1}^{m} \delta_{i,k}(|E(P^*_i)| - 1) = \sum_{i=1}^{n} |E(P^*_i)|(|E(P^*_i)| - 1)\\ &= \alpha^*(T, \pi) - \sum_{i=1}^{n} |E(P^*_i)| = \alpha^*(T, \pi) - \sum_{i=1}^{n}\sum_{k=1}^{m} \delta_{i,k} = \alpha^*(T, \pi) - \sum_{k=1}^{m}\sum_{i=1}^{n} \delta_{i,k}\\ &= \alpha^*(T, \pi) - 2m = \alpha^*(T, \pi) - 4n + 6,\end{aligned}$$

as required.

To characterize the circular phylogenetic trees that minimize and maximize the size of the circular SPR and TBR neighborhood, respectively, we introduce the following two families of circular caterpillars. A circular phylogenetic tree $(\mathcal{T}, \pi)$ is a *skew caterpillar* if $\mathcal{T}$ is a caterpillar with at least six leaves and $(\mathcal{T}, \pi)$ contains a (necessarily unique) canonical path of length $n - 1$, while $(\mathcal{T}, \pi)$ is a *centipede* if $\mathcal{T}$ is a caterpillar with at least six leaves and $(\mathcal{T}, \pi)$ contains $n - 4$ canonical paths of length 4 (see Figure 5 for an illustration). Note that a skew caterpillar contains exactly two canonical paths of length 2 and $n - 3$ canonical paths of length 3, while a centipede contains precisely two canonical paths of length 2 and two canonical paths of length 3. This implies that no circular phylogenetic tree on X could be both a skew caterpillar and a centipede for $|X| \geq 6$.

We now present tight bounds on the size of a circular SPR neighborhood.

Theorem 4 *Given a circular phylogenetic tree $(\mathcal{T}, \pi)$ on X with $n = |X| \geq 4$, we have*

$$3n - 11 \leq |N^\pi_{\text{SPR}}(\mathcal{T})| \leq (n - 1)^2 - 4n + 8. \tag{1}$$

In addition, for $n \geq 7$ the lower bound is attained if and only if $(\mathcal{T}, \pi)$ is a centipede and the upper bound is attained if and only if $(\mathcal{T}, \pi)$ is a skew caterpillar.

Proof By Lemma 3 and Theorem 3, we have

$$|N^\pi_{\text{SPR}}(\mathcal{T})| = \alpha^*(\mathcal{T}, \pi) - 13n + 27.$$

Therefore it suffices to show that

$$26 + 16(n - 4) \leq \alpha^*(\mathcal{T}, \pi) \leq 8 + 9(n - 3) + (n - 1)^2 \tag{2}$$

holds and that for $n \geq 7$, the lower bound is attained precisely when $(\mathcal{T}, \pi)$ is a centipede, and the upper bound is attained precisely when $(\mathcal{T}, \pi)$ is a skew caterpillar. We will prove this by induction on $n = |X|$.

It is straightforward to see that Eq. (2) holds for the cases $4 \leq n \leq 6$ by checking all possible $\sim_{re}$-equivalence classes of circular phylogenetic trees on X. When $n = 7$, there are five $\sim_{re}$-equivalence classes of circular trees $(\mathcal{T}, \pi)$ on X (see Figure 8 in the Appendix), and it is straightforward to check that $74 \leq \alpha^*(\mathcal{T}, \pi) \leq 80$ always holds, with the lower bound being attained precisely when $(\mathcal{T}, \pi)$ is a centipede, and the upper bound precisely when $(\mathcal{T}, \pi)$ is a skew caterpillar.

Now assume that $n > 7$ and that the result holds for $n - 1$. Without loss of generality, we may assume that $\{x_1, x_n\}$ is a cherry of $\mathcal{T}$, that is, $|E(P_n^*)| = 2$. Let $X' = X - \{x_n\}$, $\pi' = (x_1, \dots, x_{n-1})$ and $\mathcal{T}'$ be the tree obtained from $\mathcal{T}$ by removing leaf x_n and suppressing the resulting degree-two vertex. Let e'_i be the pendant edge incident to x_i in $\mathcal{T}'$. Denoting the canonical paths of $(\mathcal{T}', \pi')$ by $P'_1, \dots, P'_{n-1}$, we have $|E(P'_i)| = |E(P_i^*)| - 1$ for $i \in \{1, n-1\}$ and $|E(P'_i)| = |E(P_i^*)|$ for $1 < i < n - 1$. This implies

$$\begin{aligned}
\alpha^*(\mathcal{T}, \pi) &= |E(P_1^*)|^2 + |E(P_{n-1}^*)|^2 + |E(P_n^*)|^2 + \sum_{1<i<n-1} |E(P_i^*)|^2 \\
&= (|E(P'_1)| + 1)^2 + (|E(P'_{n-1})| + 1)^2 + 2^2 + \sum_{1<i<n-1} |E(P'_i)|^2 \\
&= \alpha^*(\mathcal{T}', \pi') + 2|E(P'_1)| + 2|E(P'_{n-1})| + 6. \qquad (3)
\end{aligned}$$

Noting that P'_1 is the path between x_1 and x_2 in $\mathcal{T}'$, and P'_{n-1} the path between x_{n-1} and x_1, we have

$$5 \leq |E(P'_1)| + |E(P'_{n-1})| \leq n. \qquad (4)$$

To see that the first inequality holds, note that $\{x_{n-1}, x_1\}$ and $\{x_1, x_2\}$ cannot be both cherries in $\mathcal{T}'$, and hence either P'_1 or P'_{n-1} contains at least three edges while the other one contains at least two edges. To see that the second inequality holds, note that $E(P'_1) \cap E(P'_{n-1})$ contains exactly the pendant edge e'_1, and $(E(P'_1) \cup E(P'_{n-1})) \subseteq (\{e'_1, e'_2, e'_{n-1}\} \cup E^o(\mathcal{T}'))$, where $|E^o(\mathcal{T}')| = n - 4$. Combining (3) and Eq. (4), we know that Eq. (2) also holds for n.

Clearly, if $(\mathcal{T}, \pi)$ is a centipede, then we have $\alpha^*(\mathcal{T}, \pi) = 26 + 16(n - 4)$. Conversely, if $\alpha^*(\mathcal{T}, \pi) = 26 + 16(n - 4)$, then by Eq. (3) and Eq. (4) we know that

$$\begin{aligned}
\alpha^*(\mathcal{T}', \pi') &= \alpha^*(\mathcal{T}, \pi) - 6 - 2(|E(P'_1)| + |E(P'_{n-1})|) \\
&\leq 26 + 16(n - 4) - 6 - 2 \cdot 5 \\
&= 26 + 16(n - 5),
\end{aligned}$$

with equality holding if and only if $|E(P'_1)| + |E(P'_{n-1})| = 5$. Together with the induction assumption, this yields $\alpha^*(\mathcal{T}', \pi') = 26 + 16(n - 5)$ and that $(\mathcal{T}', \pi')$ is a centipede. Therefore, we can conclude that $(\mathcal{T}, \pi)$ contains exactly two canonical paths of size 2, two canonical paths of size 3, and $n - 4$ paths of size 4. In other words, $(\mathcal{T}, \pi)$ is a centipede, as required.

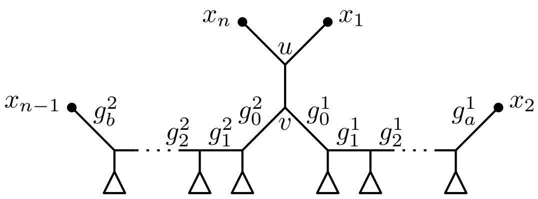

Fig. 6 The tree $\mathcal{T}$ used in the proof of Proposition 1.

On the other hand, if $(\mathcal{T}, \pi)$ is a skew caterpillar, then $\alpha^*(\mathcal{T}, \pi) = 8 + 9(n-3) + (n-1)^2$. Conversely, if $\alpha^*(\mathcal{T}, \pi) = 8 + 9(n-3) + (n-1)^2$, then by Eq. (3) and Eq. (4) we know that

$$\begin{aligned}\alpha^*(\mathcal{T}', \pi') &= \alpha^*(\mathcal{T}, \pi) - 6 - 2(|E(P'_1)| + |E(P'_{n-1})|) \\ &\geq 8 + 9(n-3) + (n-1)^2 - 6 - 2n \\ &= 8 + 9(n-4) + (n-2)^2,\end{aligned}$$

where the quality holds if and only if $|E(P'_1)| + |E(P'_{n-1})| = n$. Together with the induction assumption, this implies that $(\mathcal{T}', \pi')$ is a skew caterpillar, and hence $(\mathcal{T}, \pi)$ contains a canonical path of size $n-1$. Therefore $(\mathcal{T}, \pi)$ is a skew caterpillar, which completes the induction step, and hence also the proof of the theorem.

Next we turn to circular TBR neighborhoods. We begin with the β-index. Note that we have $\beta(\mathcal{T}, \pi) = 0$ for every tree $\mathcal{T}$ with four or five leaves. For $n = 6$, we have $\beta(\mathcal{T}, \pi) = 1$ when $\mathcal{T}$ is a centipede or skew caterpillar, and 0 otherwise. For larger n, we have the following result.

Proposition 1 *Given a circular phylogenetic tree $(\mathcal{T}, \pi)$ on X with $n = |X| \geq 7$, we have*

$$(n-5) \leq \beta(\mathcal{T}, \pi) \leq \frac{(n-5)(n-4)(n-3)}{6}. \tag{5}$$

In addition, for $n \geq 8$ the minimum is attained if and only if $(\mathcal{T}, \pi)$ is a centipede and the maximum is attained if and only if $(\mathcal{T}, \pi)$ is a skew caterpillar.

Proof We shall establish the lemma by induction on n. The result can be seen to hold for $n = 7, 8$ by checking all possible $\sim_{re}$-equivalence classes of circular phylogenetic trees on seven and eight leaves (see Figure 8 and Figure 9 in the Appendix for the list of these classes).

So, assume $n \geq 9$ and that the result holds for $n-1$. Since the β-index depends only on the $\sim_{re}$-equivalence class of a circular phylogenetic tree, by relabeling the leaves if necessary, we may assume that $\{x_1, x_n\}$ is a cherry of $\mathcal{T}$. Let u be the interior vertex incident to both x_1 and x_n, and $\{u, v\}$ the edge inducing the split $\{x_1, x_n\}|X - \{x_1, x_n\}$. In addition, denote the edges in the path from v to x_2 consecutively by g_0^1 to g_a^1, and those in the path from v to x_{n-1} by g_0^2 to g_b^2, (see Figure 6 for an illustration). Without loss of generality, we further assume that $a \geq b$ as the case $b < a$ can be established in a similar way. Note that this implies $a \geq 1$. On the other hand, we have $a + b \leq n - 4$ as $\mathcal{T}$ contains at most $n-3$ interior edges.

Consider the path $\widetilde{P}$ between x_2 and x_{n-1}. We have $E(\widetilde{P}) = \{g_a^1, \ldots, g_0^1, g_0^2, \ldots, g_b^2\}$ and $E^o(\widetilde{P}) = E(\widetilde{P}) - \{g_a^1, g_b^2\}$. Also, for each edge $e \in E^o(\widetilde{P})$, let P_e^0 be the unique path in $\{P_{e,\pi}^+, P_{e,\pi}^-\}$ that does not contain the edge $\{u, v\}$ and P_e^1 the other one. Let $X' = X - \{x_n\}$, $\pi' = (x_1, \ldots, x_{n-1})$ and $\mathcal{T}'$ be the tree obtained from

$\mathcal{T}$ by removing leaf x_n and suppressing the resulting degree two vertex. Then, by construction, we have

$$\beta(\mathcal{T},\pi) - \beta(\mathcal{T}',\pi') = \sum_{e\in E^o(\widetilde{P})} (|E(P_e^0)| - 2) = -2(a+b) + \sum_{e\in E^o(\widetilde{P})} |E(P_e^0)|. \tag{6}$$

Here the first equality holds because for each edge e in $E^o(\mathcal{T}) - (E^o(\widetilde{P}) \cup \{\{u,v\}\})$, neither $P^+_{e,\pi}$ nor $P^-_{e,\pi}$ contains $\{u,v\}$, and it contributes the same to the sum $\beta(\mathcal{T},\pi)$ and $\beta(\mathcal{T}',\pi')$. On the other hand, an edge e in $E^o(\widetilde{P})$ contributes $(|E(P_e^0)|-2)(|E(P_e^1)|-2)$ to $\beta(\mathcal{T},\pi)$, but $(|E(P_e^0)|-2)(|E(P_e^1)|-3)$ to $\beta(\mathcal{T}',\pi')$.

Since $a \geq 1$, we know that $|E(\widetilde{P})| \geq 3$ holds and hence $|E^o(\widetilde{P})| \geq 1$. Moreover, for each $e \in E^o(\widetilde{P})$, we have $|E(P_e^0)| \geq 2$. In addition, since $n > 6$, there exists at least one interior edge e in $E^o(\widetilde{P})$ with $|P_e^0| \geq 3$. Therefore we have $\beta(\mathcal{T},\pi) - \beta(\mathcal{T}',\pi') \geq 1$, and it is straightforward to check that equality holds when $(\mathcal{T},\pi)$ is a centipede. Using the induction assumption, we can thus conclude that $\beta(\mathcal{T},\pi) = (n-5)$ if $(\mathcal{T},\pi)$ is a centipede. Conversely if $\beta(\mathcal{T},\pi) = (n-5)$ then $\beta(\mathcal{T},\pi) - \beta(\mathcal{T}',\pi') \geq 1$ implies that $\beta(\mathcal{T}',\pi) = (n-6)$. By the induction assumption, we know that $\mathcal{T}'$ is a centipede. This implies that $\mathcal{T}$ is also a centipede, because replacing one vertex in a centipede on $n-1$ leaves leads to either a centipede or a caterpillar with higher β-index.

We now consider the $\sim_{re}$-equivalence class with maximum β-index. Let

$$\varphi(a,b) := \sum_{1\leq i\leq a} i + \sum_{1\leq j\leq b} j.$$

Since g_i^1 is contained in $P^0_{g_j^1}$ if and only if $i > j$, we know that for each edge g_i^1 with $i \geq 1$, the number of edges e in $E^o(\widetilde{P})$ such that g_i^1 is contained in P_e^0 is equal to i. Therefore, g_i^1 contributes precisely i to the sum $\sum_{e\in E^o(\widetilde{P})} |E(P_e^0)|$. Similarly, g_i^2 contributes i. Let $F = E(\mathcal{T}) - \{\{u,x_1\},\{u,x_n\},\{u,v\}\} \cup E(\widetilde{P})\}$. Then each edge $f \in F$ contributes at most 1 to the sum because there exists at most one $e \in E^o(\widetilde{P})$ with $f \in E(P_e^0)$. Noting that $|F| = (2n-3) - 3 - |E(\widetilde{P})|$, we conclude that

$$\begin{aligned}
\beta(\mathcal{T},\pi) - \beta(\mathcal{T}',\pi') &\leq -2(a+b) + \varphi(a,b) + (2n-6) - (a+b+2) && (7)\\
&= \varphi(a,b) + 2(n-4) - 3(a+b) \\
&\leq (1+\cdots+(a+b)) + (2n-4) - 3(a+b) && (8)\\
&= 2(n-4) + \frac{(a+b)(a+b-5)}{2} \\
&\leq (n-4)(n-5)/2. && (9)
\end{aligned}$$

Moreover, equality holds in (7) if and only if for each edge f in F, there exists exactly one $e \in E^o(\widetilde{P})$ with f contained in P_e^0. In addition, equality holds in (8) precisely when $b = 0$, and in (9) if and only if $a+b = n-4$. In other words, $\beta(\mathcal{T},\pi) - \beta(\mathcal{T}',\pi') = (n-4)(n-5)/2$ if and only if $\mathcal{T}$ is a skew caterpillar. Together with the induction assumption, it therefore follows that $\beta(\mathcal{T},\pi) \leq \frac{(n-3)(n-4)(n-5)}{6}$, in which equality holds if and only if $(\mathcal{T},\pi)$ is a skew caterpillar. This completes the proof of the induction step, and hence also the proposition.

We conclude this section rephrasing the last result in terms of the size of circular TBR neighborhoods.

Theorem 5 *Given a circular phylogenetic tree $(\mathcal{T},\pi)$ on X with $n = |X| \geq 7$, then we have*

$$4(n-4) \leq |N^\pi_{\mathrm{TBR}}(\mathcal{T})| \leq (n^3 - 6n^2 + 11n - 6)/6, \tag{10}$$

where the minimum is attained if and only if $(\mathcal{T},\pi)$ *is a centipede and the maximum is attained if and only if* $(\mathcal{T},\pi)$ *is a skew caterpillar.*

Proof By Theorem 3, it follows that $|N^{\pi}_{\text{TBR}}(\mathcal{T})| = |N^{\pi}_{\text{SPR}}(\mathcal{T})| + \beta(\mathcal{T},\pi)$. Together with Theorem 4 and Proposition 1, the theorem follows.

6 Splits in tree neighborhoods

In Section 4 we studied trees contained in the neighborhoods of circular operations. In this section we will investigate the splits induced by these trees. The idea of considering splits in the neighborhood of a tree in the tree-space was proposed by Bryant (2004) as an approach to improving the efficiency of searches in tree-space.

In order to state our results in a more general setting, we begin with extending the definition of neighborhood. Given a distance d on the set of phylogenetic trees on X and a positive integer r, the *r-neighborhood* of $\mathcal{T}$ with respect to d is the set $N_d(\mathcal{T},r)$ consisting of all phylogenetic trees $\mathcal{T}'$ on X with $0 < d(\mathcal{T},\mathcal{T}') \leq r$ and the *split neighborhood* of $\mathcal{T}$, denoted by $S_d(\mathcal{T},r)$, is the set of splits appearing in at least one of the trees in $N_d(\mathcal{T},r)$. Besides the three distances $d_{\text{NNI}}, d_{\text{SPR}}$ and d_{TBR} induced by the corresponding tree operations, another commonly used distance is the *Robinson-Foulds* distance d_{RF}, defined as

$$d_{\text{RF}}(\mathcal{T},\mathcal{T}') = \frac{1}{2}|\Sigma(\mathcal{T}) - \Sigma(\mathcal{T}')| + \frac{1}{2}|\Sigma(\mathcal{T}) - \Sigma(\mathcal{T}')|$$

for any two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on X (Robinson and Foulds, 1981). Note that the distance d_{RF} is bounded above by d_{NNI}, that is, we have $d_{\text{RF}}(\mathcal{T},\mathcal{T}') \leq d_{\text{NNI}}(\mathcal{T},\mathcal{T}')$. Moreover, $d_{\text{RF}}(\mathcal{T},\mathcal{T}') = 1$ if and only if $d_{\text{NNI}}(\mathcal{T},\mathcal{T}') = 1$. To simplify the notation, for $\text{OP} \in \{\text{NNI}, \text{SPR}, \text{TBR}, \text{RF}\}$, we will use the abbreviation $N_{\text{OP}}(\mathcal{T},r)$ for $N_{d_{\text{OP}}}(\mathcal{T},r)$, and $S_{\text{OP}}(\mathcal{T},r)$ for $S_{d_{\text{OP}}}(\mathcal{T},r)$.

Given two circular phylogenetic trees $(\mathcal{T},\pi)$ and $(\mathcal{T}',\pi)$ in $\mathscr{T}_\pi$, let $d^{\pi}_{\text{NNI}}(\mathcal{T},\mathcal{T}')$ be the minimum number of circular NNI operations that suffice to transform $(\mathcal{T},\pi)$ into $(\mathcal{T}',\pi)$, that is, $d^{\pi}_{\text{NNI}}(\mathcal{T},\mathcal{T}') = r$ if r is the smallest non-negative number such that there exists a sequence of circular trees $(\mathcal{T}_0,\pi),\dots,(\mathcal{T}_r,\pi)$ with $\mathcal{T}_0 = \mathcal{T}$, $\mathcal{T}_r = \mathcal{T}'$, and $\mathcal{T}_k \in N^{\pi}_{\text{NNI}}(\mathcal{T}_{k-1})$ for $1 \leq k \leq r$. The distances d^{π}_{SPR} and d^{π}_{TBR} on $\mathscr{T}_\pi$ are defined in a similar way. Clearly, we have $d_{\text{OP}}(\mathcal{T},\mathcal{T}') \leq d^{\pi}_{\text{OP}}(\mathcal{T},\mathcal{T}')$ for all $\mathcal{T},\mathcal{T}' \in \mathscr{T}_\pi$ and for all $\text{OP} \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$. But to our best knowledge, it remains open whether equality always holds.

Now, the r-neighborhood of a circular tree $(\mathcal{T},\pi)$ with respect to the NNI operation is defined as

$$N^{\pi}_{\text{NNI}}(\mathcal{T},r) := \{\mathcal{T}' \in \mathscr{T}_\pi \,:\, d^{\pi}_{\text{NNI}}(\mathcal{T},\mathcal{T}') \leq r\},$$

and the splits neighborhood is defined as

$$S^{\pi}_{\text{NNI}}(\mathcal{T},r) := \{A|B \,:\, A|B \in \Sigma(\mathcal{T}') \;\text{ for some }\; \mathcal{T}' \in N^{\pi}_{\text{NNI}}(\mathcal{T},r)\}.$$

The neighborhoods N^{π}_{SPR} and N^{π}_{TBR}, and split neighborhoods S^{π}_{SPR} and S^{π}_{TBR} are defined in a similar manner.

6.1 The NNI split neighborhood

We begin with split neighborhoods for circular NNI operations. Given a split $A|B \in \Sigma(X)$, an edge e in a phylogenetic tree $\mathcal{T}$ on X is *conflicting with* $A|B$ if the split $S_e(\mathcal{T})$ induced by e is incompatible with $A|B$. In addition, an internal vertex v in $\mathcal{T}$ is *conflicting with* $A|B$ if every edge incident with v conflicts with $A|B$. Using these concepts, we recall the following characterization of the split neighborhoods for the Robinson-Foulds and NNI distances.

Theorem 6 *(Bryant, 2004, Theorem 3.3 and 4.1) Let $\mathcal{T}$ be a phylogenetic tree on X. Considering a split $A|B$ in $\Sigma(X)$ and denoting the set of edges and vertices of $\mathcal{T}$ conflicting with $A|B$ respectively by E' and V', then $A|B$ is contained in $S_{\text{RF}}(\mathcal{T}, r)$ for some $r > 0$ if and only if $|E'| \leq r$, and $A|B$ is contained in $S_{\text{NNI}}(\mathcal{T}, r)$ if and only if $|E'| + |V'| \leq r$.*

To obtain a characterization of the split neighborhoods of the NNI operation for circular trees, we will also need the following technical lemma.

Lemma 4 *Given a circular tree $(\mathcal{T}, \pi)$ and a split $A|B$ in $\Sigma^o(\pi)$, there exists no vertex in $\mathcal{T}$ that conflicts with $A|B$. In particular, the edges of $\mathcal{T}$ conflicting with $A|B$ form a path in $\mathcal{T}$.*

Proof Let v be an arbitrary interior vertex v in $\mathcal{T}$. First we show that v is not conflicting with $A|B$. To this end, let u_1, u_2, u_3 denote the three vertices adjacent to v and let $A_i|B_i$ denote the split associated to edge $\{v, u_i\}$, $i \in \{1, 2, 3\}$. Without loss of generality we can assume that there exist $1 \leq j < k < n$ with $A_1 = \{x_1, x_2, \ldots, x_j\}$, $A_2 = \{x_{j+1}, x_{j+2}, \ldots, x_k\}$ and $A_3 = \{x_{k+1}, x_{k+2}, \ldots, x_n\}$. Moreover, again without loss of generality, we assume that $A = \{x_l, x_{l+1}, \ldots, x_{l'}\}$ for some $1 \leq l \leq l' < n$. It is easy to verify that there must exist some $i \in \{1, 2, 3\}$ with $A \cap A_i = \emptyset$ or $B \cap A_i = \emptyset$, implying that $A|B$ is not conflicting with $\{v, u_i\}$.

Now, by Bryant (2004, Lemma 3.1), the edges of $\mathcal{T}$ conflicting with $A|B$ form a connected subgraph of $\mathcal{T}$. Since there exists no vertex in $\mathcal{T}$ that conflicts with $A|B$, we know this connected subgraph does not contain any vertex with degree three, and hence it must be a path.

By the last lemma, we can establish the following characterization of the splits in the neighborhoods of circular NNI operation.

Theorem 7 *Let $(\mathcal{T}, \pi)$ be a circular tree. Considering a split $A|B \in \Sigma^o(\pi)$ and denoting the set of edges of $\mathcal{T}$ conflicting with $A|B$ by E', then $A|B$ is contained in $S^{\pi}_{\text{NNI}}(\mathcal{T}, r)$ for some $r > 0$ if and only if $|E'| \leq r$.*

Proof Fix a split $A|B \in \Sigma^o(\pi) \cap S^{\pi}_{\text{NNI}}(\mathcal{T}, r)$. Denote the set of edges and vertices conflicting with $A|B$ by E' and V', respectively. Since $S^{\pi}_{\text{NNI}}(\mathcal{T}, r) \subseteq S_{\text{NNI}}(\mathcal{T}, r)$, by Theorem 6 we know that $|E'| + |V'| \leq r$. On the other hand, by Lemma 4 we know $|V'| = 0$, and hence we have $|E'| \leq r$, as required.

To establish the other direction, suppose that the set E' of edges that are conflicting with $A|B$ satisfies $|E'| \leq r$. Choose an edge $\{u, v\}$ of E' such that u is adjacent to no other edges in E', and denote the four clusters incident with the edge $\{u, v\}$ by A_1, A_2, B_1, B_2 (see $\mathcal{T}$ in Figure 7 for an illustration). Then the splits $A_1|(X - A_1)$ and $A_2|(X - A_2)$ are compatible with $A|B$ while $(A_1 \cup A_2)|(B_1 \cup B_2)$ is not compatible with $A|B$. Switching A and B if necessary, we may assume $A_1 \subseteq A$. This leads to $A_2 \subseteq B$.

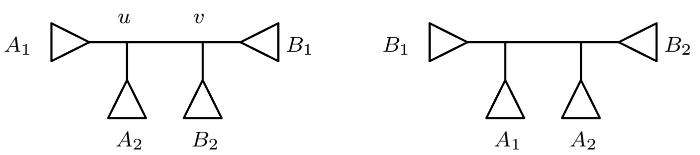

Fig. 7 Two trees used in the proof of Theorem 7: $\mathcal{T}$ (left) and $\mathcal{T}'$ (right).

By Lemma 4, the vertex v is not conflicting with $A|B$, and hence $A|B$ is compatible with at least one of the splits $B_1|(X-B_1)$ and $B_2|(X-B_2)$. Without loss of generality, we assume that $A|B$ is compatible with $B_2|(X-B_2)$. This implies $B_2 \subseteq B$ as otherwise we must have $B_1 \subseteq A$, a contradiction. Therefore $A_2, B_2 \subseteq B$. Now consider the tree $\mathcal{T}'$ obtained from $\mathcal{T}$ by one circular NNI operation as depicted in Figure 7. Then $\mathcal{T}'$ contains one less conflicting edge. Therefore, repeating the process at most $|E'|$ times leads to a tree $\mathcal{T}^*$ in $\mathscr{T}_\pi$ with $A|B \in \Sigma(\mathcal{T}^*)$ and $d^\pi_{\text{NNI}}(\mathcal{T},\mathcal{T}^*) \le r$, as required.

The last theorem enables us to obtain another characterization of splits in the circular NNI operations.

Corollary 2 *Given a circular phylogenetic tree $(\mathcal{T},\pi)$ on X, we have*

$$S^\pi_{\text{NNI}}(\mathcal{T},r) = S_{\text{NNI}}(\mathcal{T},r)\bigcap \Sigma^o(\pi) = S_{\text{RF}}(\mathcal{T},r)\bigcap \Sigma^o(\pi)$$

for each positive integer r.

Proof Since $d_{\text{RF}}(\mathcal{T}_1,\mathcal{T}_2) \le d_{\text{NNI}}(\mathcal{T}_1,\mathcal{T}_2) \le d^\pi_{\text{NNI}}(\mathcal{T}_1,\mathcal{T}_2)$ holds for two arbitrary trees $\mathcal{T}_1,\mathcal{T}_2$ in $\mathscr{T}_\pi$, we have $N^\pi_{\text{NNI}}(\mathcal{T},r) \subseteq N_{\text{NNI}}(\mathcal{T},r) \subseteq N_{\text{RF}}(\mathcal{T},r)$ and hence

$$S^\pi_{\text{NNI}}(\mathcal{T},r) \subseteq S_{\text{NNI}}(\mathcal{T},r)\bigcap \Sigma^o(\pi) \subseteq S_{\text{RF}}(\mathcal{T},r)\bigcap \Sigma^o(\pi).$$

It remains to show that $S_{\text{RF}}(\mathcal{T},r)\bigcap \Sigma^o(\pi) \subseteq S^\pi_{\text{NNI}}(\mathcal{T},r)$. To this end, fix a split $A|B$ in $S_{\text{RF}}(\mathcal{T},r)\bigcap \Sigma^o(\pi)$ and denote the set of those edges in $\mathcal{T}$ conflicting with $A|B$ by E'. Then by Theorem 6 we have $|E'| \le r$. Using Theorem 7, we have $A|B \in S^\pi_{\text{NNI}}(\mathcal{T},r)$, as required.

Given a phylogenetic tree $\mathcal{T}$ on X and a positive number k, let $\text{INT}_k(\mathcal{T})$ be the number of paths containing precisely k edges in the tree obtained from $\mathcal{T}$ by removing all its leaves and pendant edges. Note that this is related to the *Whitney number* of $\mathcal{T}$ (Jamison, 1987). We end this subsection with a formula relating the number of splits contained in $S^\pi_{\text{NNI}}(\mathcal{T},r)$ to $\text{INT}_k(\mathcal{T})$.

Theorem 8 *Given a circular phylogenetic tree $(\mathcal{T},\pi)$ on X, we have*

$$|S^\pi_{\text{NNI}}(\mathcal{T},r)| = \sum_{k=1}^{r} \text{INT}_k(\mathcal{T}).$$

In particular, we have $|S^\pi_{\text{NNI}}(\mathcal{T},r)| = |S^{\pi'}_{\text{NNI}}(\mathcal{T},r)|$ for every circular ordering π' with $\mathcal{T} \in \mathscr{T}_{\pi'}$

Proof This follows from Lemma 4 and Theorem 7.

Note that by the last theorem, it is straightforward to see that for a circular tree $(\mathcal{T},\pi)$ on X, we have $|S^\pi_{\text{NNI}}(\mathcal{T},r)| \le |S^\pi_{\text{NNI}}(\mathcal{T}',r)|$ for all $\mathcal{T}' \in \mathscr{T}_\pi$ and $1 \le r \le |X|$ if and only if $\mathcal{T}$ is a caterpillar. But it remains open to characterize the trees whose split neighborhoods have minimum size relative to circular NNI operations.

6.2 The SPR and TBR split neighborhoods

Now we proceed to characterizing splits contained in the circular SPR and TBR neighborhoods. To this end, we will use a result obtained in Bryant (2004) relating the split neighborhoods for d_{SPR} and d_{TBR} to the parsimony length of a character, which we now recall. A *binary character* for X is a function $\chi : X \to \{0,1\}$, and an *extension* of χ on a phylogenetic tree $\mathcal{T}$ on X is a function $\hat{\chi} : V(\mathcal{T}) \to \{0,1\}$ such that $\hat{\chi}(x) = \chi(x)$ for each leaf x of $\mathcal{T}$. The *length* of $\hat{\chi}$, denoted $\hat{l}_{\mathcal{T}}(\hat{\chi})$, is the number of edges $\{u,v\} \in E(\mathcal{T})$ for which $\hat{\chi}(u) \neq \hat{\chi}(v)$. The *parsimony length* of χ, denoted $l_{\mathcal{T}}(\chi)$, is the minimum of $\hat{l}_{\mathcal{T}}(\hat{\chi})$ over all extensions $\hat{\chi}$ of χ. Given a subset A of X, we let χ_A denote the binary character that maps x to 1 for $x \in A$, and 0 otherwise. Note that for each split $A|B$ of X, we have $l_{\mathcal{T}}(\chi_A) = l_{\mathcal{T}}(\chi_B)$.

Theorem 9 *(Bryant, 2004, Theorem 5.2) Let $\mathcal{T}$ be a phylogenetic tree $\mathcal{T}$ on X. Given a split $A|B$ of X and a positive integer r, the following three statements are equivalent:* (i) *$A|B \in S_{\text{SPR}}(\mathcal{T}, r)$;* (ii) *$A|B \in S_{\text{TBR}}(\mathcal{T}, r)$;* (iii) *$l_{\mathcal{T}}(\chi_A) \leq r+1$.*

This theorem is of interest since it shows that the split neighborhoods for SPR operations are the same as those for TBR operations, although the tree neighborhoods for these two families of operations are generally different. As the following theorem shows, the same holds for the circular neighborhoods for these operations.

Theorem 10 *Let $(\mathcal{T}, \pi)$ be a circular phylogenetic tree. Given a split $A|B \in \Sigma^o(\pi)$ and a positive number r, the following three statements are equivalent:* (i) *$A|B \in S^{\pi}_{\text{SPR}}(\mathcal{T}, r)$;* (ii) *$A|B \in S^{\pi}_{\text{TBR}}(\mathcal{T}, r)$;* (iii) *$l_{\mathcal{T}}(\chi_A) \leq r+1$.*

Proof Since $N^{\pi}_{\text{SPR}}(\mathcal{T}, r) \subseteq N^{\pi}_{\text{TBR}}(\mathcal{T}, r)$, we know that (i) implies (ii). On the other hand, by Theorem 9 we also know that (ii) implies (iii). Therefore it remains to show (iii) implies (i).

Fix a split $A|B$ with $l_{\mathcal{T}}(\chi_{A|B}) \leq r+1$. For simplicity, we may assume $\pi = (x_1, x_2, \ldots, x_n)$ and $A = \{x_1, \ldots, x_k\}$ for some $1 \leq k < n$. Since (iii) holds, we have $l_{\mathcal{T}}(\chi_A) = s+1$ for some nonnegative number $s \leq r$. If $s = 0$ then the assertion holds because $l_{\mathcal{T}}(\chi_A) = 1$ if and only if $A|B \in \Sigma(\mathcal{T})$. Otherwise, consider a minimum length extension $\hat{\chi}$ of χ_A and let $ch(\hat{\chi})$ be the set of edges $\{u,v\}$ in $\mathcal{T}$ with $\hat{\chi}(u) \neq \hat{\chi}(v)$.

For each $1 \leq i \leq n$ with $i \notin \{k, n\}$, since $\chi_A(x_i) = \chi_A(x_{i+1})$, the canonical path P^*_i between x_i and x_{i+1} in $\mathcal{T}$ contains either no edges in $ch(\hat{\chi})$, or at least two edges in $ch(\hat{\chi})$. Noting that each edge in $ch(\hat{\chi})$ is contained in exactly two canonical paths in $\{P^*_1, P^*_2, \ldots, P^*_n\}$, we can fix some $1 \leq i \leq n$ so that P^*_i contains at least two distinct edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ in $ch(\hat{\chi})$. Since e_1 and e_2 are both contained in $ch(\hat{\chi})$, we must have $e_1 \cap e_2 = \emptyset$ as otherwise $\hat{\chi}$ is not a minimal extension. This implies that $\theta = (e_1, e_2; e_1)$ encodes an SPR operation. Moreover, because $e_2 \in E(P^+_{e_1,\pi}) \cup E(P^-_{e_1,\pi})$, by Theorem 2 we know $\theta \in \mathcal{O}^{\pi}_{\text{SPR}}(\mathcal{T})$.

Switching the labels if necessary, we may assume u_2, v_1 are contained in the shortest path between u_1 and v_2 in $\mathcal{T}$. Denoting the two vertices incident with v_1 other than u_1 by v_3 and v_4, we have $\hat{\chi}(v_3) = \hat{\chi}(v_4) = \hat{\chi}(v_1)$ because $\hat{\chi}(v_1) \neq \hat{\chi}(u_1)$ and $\hat{\chi}$ is a minimal extension of a binary character. Note that the tree $\mathcal{T}' = \theta(\mathcal{T})$ is obtained from $\mathcal{T}$ inserting a new vertex u on edge e_2, deleting the vertex v_1 and the three edges incident with it, adding two edges (u, u_1) and (v_3, v_4). Consider the extension of χ on $\mathcal{T}'$ defined by putting $\chi^*(u) = \hat{\chi}(u_1)$, and $\chi^*(w) = \hat{\chi}(w)$ for all $w \neq u_1$ in $V(\mathcal{T}')$. Since $\hat{\chi}(u_2) \neq \hat{\chi}(v_2)$, we have either $\chi^*(u) = \chi^*(u_2)$ or $\chi^*(u) = \chi^*(v_2)$. Therefore we can conclude that $l_{\mathcal{T}'}(\chi^*) = s$ and hence $l_{\mathcal{T}'}(\chi_A) \leq s$. Repeating the process we will obtain a circular phylogenetic tree $(\mathcal{T}'', \pi)$ such that $A|B \in \Sigma(\mathcal{T}'')$ and $d^{\pi}_{\text{SPR}}(\mathcal{T}, \mathcal{T}'') \leq s$, as required.

The last theorem leads to the following result relating splits contained in the circular SPR and TBR neighborhoods to those splits in the SPR and TBR neighborhoods.

Corollary 3 *Given a circular ordering π and a tree $\mathcal{T} \in \mathscr{T}_\pi$, we have*

$$S^{\pi}_{\text{SPR}}(\mathcal{T}, r) = S_{\text{SPR}}(\mathcal{T}, r) \bigcap \Sigma^o(\pi) = S^{\pi}_{\text{TBR}}(\mathcal{T}, r) = S_{\text{TBR}}(\mathcal{T}, r) \bigcap \Sigma^o(\pi)$$

for each positive integer r.

Proof Since Theorem 9 implies $S_{\text{SPR}}(\mathcal{T}, r) = S_{\text{TBR}}(\mathcal{T}, r)$, and Theorem 10 implies $S^{\pi}_{\text{SPR}}(\mathcal{T}, r) = S^{\pi}_{\text{TBR}}(\mathcal{T}, r)$, it remains to show $S^{\pi}_{\text{SPR}}(\mathcal{T}, r) = S_{\text{SPR}}(\mathcal{T}, r) \bigcap \Sigma^o(\pi)$.

To this end, note first that $N^{\pi}_{\text{SPR}}(\mathcal{T}, \mathcal{T}') \subseteq N_{\text{SPR}}(\mathcal{T}, \mathcal{T}')$ implies $S^{\pi}_{\text{SPR}}(\mathcal{T}, r) \subseteq S_{\text{SPR}}(\mathcal{T}, r) \bigcap \Sigma^o(\pi)$. On the other hand, for an arbitrary split $A|B \in S_{\text{SPR}}(\mathcal{T}, r) \bigcap \Sigma^o(\pi)$, we have $l_{\mathcal{T}}(\chi_{A|B}) \leq r + 1$ in view of Theorem 9. Together with Theorem 10, this implies $A|B \in S^{\pi}_{\text{SPR}}(\mathcal{T}, r)$ and hence $S_{\text{SPR}}(\mathcal{T}, r) \bigcap \Sigma^o(\pi) \subseteq S^{\pi}_{\text{SPR}}(\mathcal{T}, r)$, as required.

It is shown in Bryant (2004, Corollary 5.3) that $|S_{\text{SPR}}(\mathcal{T}, r)|$ (and hence also $|S_{\text{TBR}}(\mathcal{T}, r)|$) is determined solely by n and r. In other words, it does not depend on the phylogenetic tree $\mathcal{T}$. However, the following example shows that this is not the case for $|S^{\pi}_{\text{SPR}}(\mathcal{T}, r)|$, which depends on both $\mathcal{T}$ and π.

Let $\pi = (x_1, \ldots, x_n)$ for $n \geq 6$ and put $k = \lfloor n/2 \rfloor$. Consider the skewed caterpillar $\mathcal{T}_1$ and the centipede $\mathcal{T}_2$ in $\mathscr{T}_\pi$ for which P^*_n and P^*_k are the only canonical paths of size two (see Figure 5 for an illustration). Note that $l_{\mathcal{T}_1}(\chi_S) \leq 2$ for all $S \in \Sigma^o(\pi)$ and hence by Theorem 10 we have

$$|S^{\pi}_{\text{SPR}}(\mathcal{T}_1, 1)| = |\Sigma^o(\pi)| = \binom{n}{2}.$$

On the other hand, for the split $A|B = \{x_1, \ldots, x_k\}|\{x_{k+1}, \ldots, x_n\}$ we have $l_{\mathcal{T}_2}(\chi_{A|B}) = k > 2$, and hence $|S_{\text{SPR}}(\mathcal{T}_1, 1)| \neq |S_{\text{SPR}}(\mathcal{T}_2, 1)|$.

This example also shows that for a circular tree $(\mathcal{T}, \pi)$ on X that is $\sim_{re}$-equivalent to a skew caterpillar, we have $|S^{\pi}_{\text{NNI}}(\mathcal{T}, r)| = |S^{\pi}_{\text{NNI}}(\mathcal{T}, 1)| = |\Sigma^o(\pi)| = \binom{n}{2}$. Therefore, skew caterpillars maximize the size of their circular SPR and TBR neighborhoods. However, it remains open to characterize the trees that minimize them.

6.3 Counting trees in a neighborhood containing a split

Interestingly, using Theorem 10, it is possible to solve a 'dual' problem to counting splits in a neighborhood of a tree. In particular, we can count the number of trees in $\mathscr{T}_\pi$ that are within SPR or TBR distance r of a tree containing a given split $A|B \in \Sigma^o(\pi)$. To this end, we need the following technical lemma.

Lemma 5 *Given a circular ordering π and a split $A|B \in \Sigma^o(\pi)$ with $|A| = a$ and $|B| = b$, the number of trees in $\mathscr{T}_\pi$ on which $\chi_{A|B}$ has parsimony length k equals*

$$2^{k-1} \sum_{\substack{a_1+\cdots+a_k=a,\\ b_1+\cdots+b_k=b,\\ a_i>0,\quad b_j>0}} \prod_{1 \leq i \leq k} (2a_i - 3) \prod_{1 \leq j \leq k} (2b_j - 3)$$

Proof Without loss of generality, we may assume $A = \{x_1, \ldots, x_a\}$. Let $\mathscr{T}_\pi$ denote the set of trees in $\mathscr{T}_\pi$ on which $\chi_{A|B}$ has parsimony length k by $\mathscr{T}_\pi^k$. Then each tree $\mathcal{T}$ in $\mathscr{T}_\pi^k$ can be constructed by the following four steps.

(i) Decompose A into k disjoint non-empty 'consecutive' subsets $A_1, \ldots, A_k$, that is, for each $1 \leq t \leq k$, $A_t = X_{i,j}$ for some $1 \leq i \leq j \leq a$. Similarly, we decompose B into k disjoint non-empty 'consecutive' subsets $B_1, \ldots, B_k$.

(ii) For each subset $Y \in \{A_1, \ldots, A_k, B_1, \ldots, B_k\}$, construct a phylogenetic tree $\mathcal{T}_Y$ on Y.

(iii) Consider a circular ordering $\pi' = (1, 2, \ldots, 2k)$ on $Z := \{1, 2, \ldots, 2k\}$, and construct a tree $\mathcal{T}'$ on $\mathscr{T}_{\pi'}$ such that $l_{\mathcal{T}'}(\chi_S) = k$ for the split $S = \{1, \ldots, k\}|\{k+1, \ldots, 2k\}$.

(iv) Finally, a tree $\mathcal{T}$ is obtained from $\mathcal{T}'$ by replacing each leaf i in $\mathcal{T}'$ with the tree $\mathcal{T}_{A_i}$ if $1 \leq i \leq k$, and $\mathcal{T}_{B_{j-k}}$ if $k+1 \leq j \leq 2k$.

Since different choices in Step (i)-(iii) give different trees $\mathcal{T}$ in $\mathscr{T}_\pi^k$, the lemma follows from the fact that there are exactly

$$\sum_{\substack{a_1+\cdots+a_k=a,\\ b_1+\cdots+b_k=b,\\ a_i>0,\quad b_j>0}} \prod_{1\leq i\leq k} (2a_i - 3) \prod_{1\leq j\leq k} (2b_j - 3)$$

different choices in Step (i) and (ii), and 2^{k-1} different choices in Step (iii).

Together with Theorem 10, the last lemma gives the following result, which is an analogue of Bryant (2004, Corollary 5.4).

Proposition 2 *Given a circular ordering π and a split $A|B \in \Sigma^o(\pi)$ with $|A| = a$ and $|B| = b$, the number of trees in $\mathscr{T}_\pi$ that are within* SPR *or* TBR *distance r of a tree containing $A|B$ equals*

$$\sum_{k=1}^{r+1} 2^{k-1} \Big(\sum_{\substack{a_1+\cdots+a_k=a,\\ b_1+\cdots+b_k=b,\\ a_i>0,\quad b_j>0}} \prod_{1\leq i\leq k} (2a_i - 3) \prod_{1\leq j\leq k} (2b_j - 3) \Big).$$

7 Discussion

In this paper we have studied circular neighborhoods of a tree induced by three commonly used tree rearrangement operations: NNI, SPR and TBR, motivated in part by the need to find efficient approaches to search for phylogenetic trees. Using an observation that relates circular neighborhoods to these operations (Theorem 2), we have derived bounds for the number of trees in circular neighborhoods, as well as giving characterizations for the trees whose circular neighborhoods attain these bounds. In addition, we have obtained various results concerning the splits induced by trees in circular neighborhoods.

Our theoretical results provide some useful insights into the efficiency of searching for trees. For instance, consider a local search based on the SPR operation. By the formula for the size of an SPR neighborhood in tree-space given in Allen and Steel (2001, p.4), the number of comparisons required to find a local optimum in a neighbourhood of a tree with n leaves is bounded above by $2(n-3)(2n-7)$. On the other hand, Theorem 1 above implies that this number is between $3n - 11$ and $(n-1)^2 - 4n + 8$ when restricted to the set of trees in a circular ordering. In consequence, the potential speed-up resulting from using a circular ordering in an

SPR-based local search is between 4 and $4n/3$ for each local move. This could be significant if the number of moves required to reach a locally optimal tree is large, although it should be noted that an optimal tree for the whole of tree space may not always be captured within some precomputed circular ordering.

Besides the problems mentioned in the last section, there are several directions that might be followed in future work. For instance, given two arbitrary trees $\mathcal{T}$ and $\mathcal{T}'$ that are both compatible with some circular ordering π, it could be of interest to study the minimal number $d^{\pi}_{\text{OP}}(\mathcal{T}, \mathcal{T}')$ of operations $\text{OP} \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$ required to transform to $\mathcal{T}$ to $\mathcal{T}'$ whilst remaining in the same ordering. Indeed, this would provide insights into the number of moves required to reach a local optimal tree as mentioned above. For $\text{OP} = \text{NNI}$ the distance d^{π}_{NNI} is equivalent to the rotation distance between triangulations, and it is known that computing this distance is fixed-parameter tractable (Cleary and St. John, 2009; Luccio et al, 2010), although it remains an open question as to whether its computation in general is NP-complete. Moreover, it has been known for some time that $d^{\pi}_{\text{NNI}}(\mathcal{T}\ \mathcal{T}') \leq 2n - 10$ holds for any pair of trees $\mathcal{T}, \mathcal{T}'$ with n leaves, $n \geq 12$ (Sleator et al, 1988), and recently it has been shown that this bound is actually tight by Pournin (2014). For the other two distances d^{π}_{SPR} and d^{π}_{TBR}, less is known and it could be interesting to see whether analagous results hold. In regards to this, clearly $d^{\pi}_{\text{OP}}(\mathcal{T}, \mathcal{T}') \geq d_{\text{OP}}(\mathcal{T}, \mathcal{T}')$ holds for the distance d_{OP} defined in Section 2, but does equality always hold? Note that Corollaries 2 and 3 provide some evidence that support this possibility.

In a related direction, in Gordon et al (2013) it is shown that tree-space can be transversed by NNI operations without repetition, that is, the trees in tree-space can be arranged in a sequence that contains each tree exactly once such that two consecutive trees differ by a single NNI move. It could be interesting to see whether an analagous result holds for the trees that respect a circular ordering. More generally, it could also be worthwhile to investigate whether our findings might be applied to other problems that have been addressed using tree operations, such as tree reconciliation (see, e.g. Bansal and Eulenstein, 2008) and terraces in tree-space (Sanderson et al, 2011).

Acknoeledgements Parts of the results presented here were included in the PhD dissertation of Sarah Baskowski, University of East Anglia, 2014. We also thank two anonymous referees for their constructive suggestions.

References

Allen B, Steel M (2001) Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics 5:1–15

Bansal MS, Eulenstein O (2008) An ω (nˆ 2/log n) speed-up of tbr heuristics for the gene-duplication problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 5(4):514–524

Bastkowski S, Spillner A, Moulton V (2014) Fishing for minimum evolution trees with Neighbor-Nets. Information Processing Letters 114:13–18

Bryant D (1996) Hunting for trees in binary character sets: efficient algorithms for extraction, enumeration and optimization. Journal of Computational Biology 3:275–288

Bryant D (1997) Building trees, hunting for trees and comparing trees. PhD thesis, University of Canterbury, NZ

Bryant D (2004) The splits in the neighborhood of a tree. Annals of Combinatorics 8:1–11

Bryant D, Moulton V (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution 21:255–265

Cleary S, St John K (2009) Rotation distance is fixed-parameter tractable. Information Processing Letters 109:918–922

De Loera J, Rambau J, Santos F (2010) Triangulations: structures for algorithms and applications, Algorithms and Computation in Mathematics, vol 25. Springer

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. Journal of Computational Biology 9:687–705

Ding Y, Grunewald S, Humphries P (2011) On agreement forests. Journal of Combinatorial Theory, Series A 118:2059–2065

Felsenstein J (2004) Inferring Phylogenies. Sinauer Associates Inc.

Gordon K, Ford E, St John K (2013) Hamiltonian walks of phylogenetic treespaces. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 10(4):1076–1079

Humphries P, Wu T (2013) On the neighborhoods of trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10:721–728

Jamison R (1987) Alternating whitney sums and matchings in trees. Discrete Mathematics 67:177–189

Kubatko L (2008) Inference of phylogenetic trees. In: Friedman A (ed) Tutorials in Mathematical Biosciences IV: Evolution and Ecology, Lecture Notes in Mathematics, vol 1922, Springer, pp 1–38

Lemey P, Salemi M, Vandamme AM (2009) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press

Li M, Tromp J, Zhang L (1996) On the nearest neighbor interchange distance between evolutionary trees. Journal of Theoretical Biology 182:463–467

Luccio F, Enriquez AM, Pagli L (2010) Lower bounds on the rotation distance of binary trees. Information Processing Letters 110(21):934–938

Pournin L (2014) The diameter of associahedra. Advances in Mathematics 259:13–42

Robinson D (1971) Comparison of labeled trees with valency three. Journal of Combinatorial Theory, Series B 11:105–119

Robinson D, Foulds L (1981) Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147

Sanderson MJ, McMahon MM, Steel M (2011) Terraces in phylogenetic tree space. Science 333(6041):448–450

Semple C, Steel M (2003) Phylogenetics. Oxford University Press

Semple C, Steel M (2004) Cyclic permutations and evolutionary trees. Advances in Applied Mathematics 32:669–680

Sleator D, Tarjan R, Thurston W (1988) Rotation distance, triangulations and hyperbolic geometry. Journal of the American Mathematical Society 1:647–681

Tepe E, Farruggia F, Bohs L (2011) A 10-gene phylogeny of *Solanum* section *Herpystichum* (Solanaceae) and a comparison of phylogenetic methods. American Journal of Botany 98:1356–1365

Whelan S, Money D (2010) The prevalence of multifurcations in tree-space and their implications for tree-search. Molecular Biology and Evolution 27:2674–2677

Appendix

	species	accession		species	accession		species	accession
1	*S. caripense*	GQ221590	6	*S. evolvulifolium*	GQ221591	11	*S. brevifolium*	GQ221589
2	*S. dalibardiforme*	HQ856066	7	*S. anceps*	GQ221568	12	*S. loxophyllum*	HQ856058
3	*S. bulbocastanum*	DQ180444	8	*S. pacificum*	HQ856065	13	*S. pentaphyllum*	HQ856061
4	*S. lycopersicum*	DQ180450	9	*S. phaseoloides*	GQ221592	14	*S. trifolium*	HQ856067
5	*S. crassinervium*	HQ856062	10	*S. limoncochanese*	HQ856063	15	*S. dolichorhachis*	HQ856057

Table 1 The names of the *Solanum* species related by the phylogenetic tree in Figure 1 and GenBank accessions of the corresponding sequences.

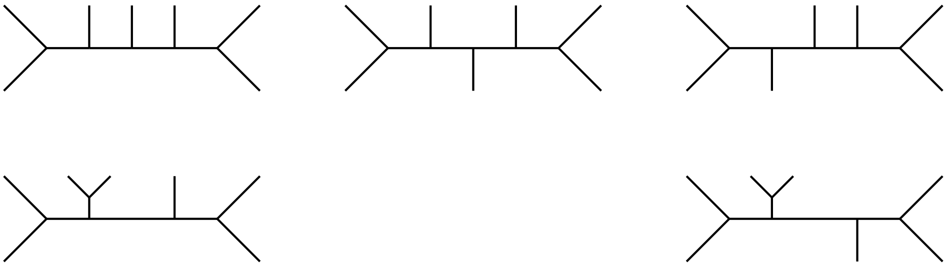

Fig. 8 Representatives of the five different $\sim_{re}$-equivalence classes of circular phylogenetic trees on seven leaves. Here the labels of the leaves are omitted for simplicity.

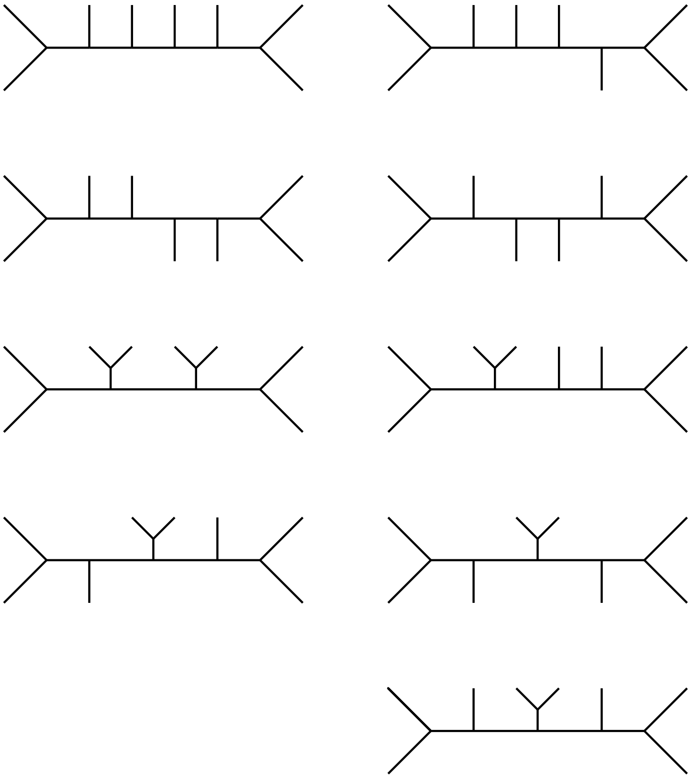

Fig. 9 Representatives of the thirteen different $\sim_{re}$-equivalence classes of circular phylogenetic trees on eight leaves. Here the labels of the leaves are omitted for simplicity.