CYCLIC PERMUTATIONS AND EVOLUTIONARY TREES

CHARLES SEMPLE AND MIKE STEEL

ABSTRACT. Given a tree $\mathcal{T}$ with leaf set X, there are certain ways of arranging the elements of X in a circular order so that $\mathcal{T}$ can be embedded in the plane and 'preserve' this ordering. We investigate some new combinatorial properties of these 'circular orderings'. We then use these properties to establish two results concerning dissimilarity maps on X that are induced by edge-weighted trees with leaf set X.

1. INTRODUCTION

A *phylogenetic X-tree* $\mathcal{T}$ is a tree that has X as its set of leaves and whose interior vertices are of degree at least three. Figure 1 shows a phylogenetic tree with $\{1, 2, \ldots, 7\}$ as its set of leaves. In evolutionary biology, phylogenetic X-trees are widely used to represent the ancestral relationships of a set X of present-day species (for further details, see [12, 14]).

A *dissimilarity map (on X)* is a function $\delta : X \times X \to \mathbb{R}$ such that, for all $x, y \in X$, $\delta(x, x) = 0$ and $\delta(x, y) = \delta(y, x)$. In evolutionary biology, such a map might measure the genetic difference between two species. For an arbitrary dissimilarity map δ on X, a classical problem in classification is to determine if there is a phylogenetic X-tree $\mathcal{T}$ and a real-valued weighting of the edges of $\mathcal{T}$ so that, for all $x, y \in X$, the sum of the weights of the edges of $\mathcal{T}$ in the path connecting x and y is equal to $\delta(x, y)$. If such a phylogenetic X-tree and edge weighting w exists, where w is non-negative, δ is said to be a *tree metric*. The problem of recognizing and characterizing tree metrics has a well known solution that dates back more than 30 years (see [3, 5, 13, 17]).

In this paper, we prove two new results on tree metrics. The first result is a novel description of the total sum of the edge weights of a real-valued edge-weighted phylogenetic tree. The second result is an explicit convergence result for the 'minimum length tree reconstruction method'. Typically, an arbitrary dissimilarity map δ on X is not a tree metric. However, one would still like to construct an edge-weighted phylogenetic X-tree from δ. The minimum length tree reconstruction method is one such method. Both of these results are derived by considering, for a phylogenetic X-tree $\mathcal{T}$, cyclic permutations of X that provide a 'circular ordering for $\mathcal{T}$'.

Date: 17 December 2002.

Key words and phrases. phylogenetic tree, cyclic permutation, circular ordering, dissimilarity map, tree metric.

We thank the New Zealand Marsden Fund (UOC-MIS-005) for supporting this research.

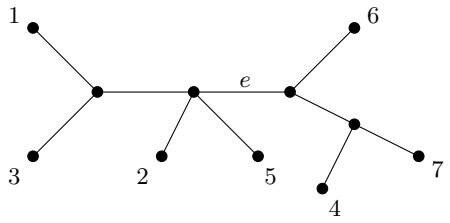

FIGURE 1. A phylogenetic tree.

The results in our paper are complementary to, though quite different from, the investigation into 'circular orderings' by [8] and [7]. The former of these papers establishes an equivalence between circular orderings for a phylogenetic X-tree and another class of cyclic permutations of X (called 'Yushmanov orderings'), from which algorithms are then derived. The authors of [7] use circular orderings to develop an approach for reconstructing phylogenetic X-trees from dissimilarity maps on X based on the 'travelling salesman' problem.

The purpose of our paper is twofold. Firstly, to establish some new combinatorial properties of circular orderings and, secondly, to show how circular orderings can be used to derive results on tree metrics. The latter is done by using these combinatorial properties to prove the two tree metric results mentioned in the last paragraph.

Unless stated otherwise, the phylogenetic terminology in this paper follows Semple and Steel [12]. Also, throughout this paper, X denotes a finite set. The paper is organized as follows. The central concept is the notion of a circular ordering for a phylogenetic tree. We describe this in the next section as well as stating some well known results on phylogenetic trees. Section 3 establishes some new combinatorial properties of circular orderings and phylogenetic trees. These properties are used in Section 4 to prove our two results on tree metrics.

2. Preliminaries

For a phylogenetic tree $\mathcal{T}$, we denote the set of interior vertices and the set of interior edges of $\mathcal{T}$ by $\mathring{V}(\mathcal{T})$ and $\mathring{E}(\mathcal{T})$, respectively. If every interior vertex of $\mathcal{T}$ has degree three, $\mathcal{T}$ is a *trivalent* phylogenetic tree (in [12], a trivalent phylogenetic tree is called a binary phylogenetic tree). The following lemma dates back to Schröder [11].

Lemma 2.1.

(i) *A trivalent phylogenetic tree with n leaves has $2n-3$ edges and $n-2$ interior vertices.*
(ii) *For a fixed set X of size at least three, the number of trivalent phylogenetic X-trees is*

$$\frac{(2n-4)!}{(n-2)!2^{n-2}} = 1 \times 3 \times 5 \times \cdots \times (2n-5),$$

where $n = |X|$.

Two phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ are regarded as *equivalent* if the identity map on X induces a graph isomorphism between $\mathcal{T}_1$ and $\mathcal{T}_2$, in which case we write $\mathcal{T}_1 \cong \mathcal{T}_2$. Thus, up to equivalence, there are precisely three trivalent phylogenetic trees for a set X of size four.

An X-*split* is a partition of X into two non-empty sets. We denote the X-split whose blocks are A and B by $A|B$. Associated with every phylogenetic X-tree $\mathcal{T}$ is a particular collection of X-splits. This collection consists of those X-splits $A|B$ that are induced by the components of the graph resulting from the deletion of a single edge e of $\mathcal{T}$. We say that the X-split $A|B$ *corresponds to* e and let $\Sigma(\mathcal{T})$ denote the set of X-splits that correspond to the edges of $\mathcal{T}$. For example, referring to Fig. 1, $\{1,2,3,5\}|\{4,6,7\}$ is the split of $\mathcal{T}$ corresponding to e. As part of a characterization of a certain type of collection of splits, Buneman [2] proved the following result.

Proposition 2.2. *Let* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *be phylogenetic* X*-trees. Then* $\Sigma(\mathcal{T}_1) = \Sigma(\mathcal{T}_2)$ *if and only if* $\mathcal{T}_1 \cong \mathcal{T}_2$.

Let $\pi = (x_1, x_2, \ldots, x_n)$ be a cyclic permutation of X. For all $1 \leq i \leq j \leq n$, let $A_{ij} = \{x_k : i \leq k \leq j\}$ and let $\Sigma^\circ(\pi)$ denote the set

$$\Sigma^\circ(\pi) = \{A_{ij}|X - A_{ij} : 1 \leq i \leq j \leq n-1\}$$

of X-splits. Arranging the elements $x_1, x_2, \ldots, x_n$ clockwise in a circle in the plane, we may view $\Sigma^\circ(\pi)$ as the set of X-splits that can be obtained by separating these elements according to which side of a line segment in the plane they lie on. Consequently, $|\Sigma^\circ(\pi)| = \binom{n}{2}$. A collection Σ of X-splits is said to be *circular* if $\Sigma \subseteq \Sigma^\circ(\pi)$ for some cyclic permutation π of X. In case $\Sigma(\mathcal{T}) \subseteq \Sigma^\circ(\pi)$ for some phylogenetic X-tree $\mathcal{T}$, we say that π provides a *circular ordering* for $\mathcal{T}$. For example, $(1,6,7,4,5,2,3)$ is a circular ordering for the phylogenetic tree shown in Fig. 1, but $(1,6,7,2,3,4,5)$ is not such an ordering. Throughout the paper, for a cyclic permutation $\pi = (x_1, x_2, \ldots, x_n)$, we will adopt the convention that $x_{n+1} = x_1$.

3. Circular Orderings and Phylogenetic Trees

In this section, we establish some properties of circular orderings and phylogenetic trees. Let $\mathcal{T}$ be a phylogenetic X-tree. For all vertices v of $\mathcal{T}$, let $d(v)$ denote the degree of v and, for all distinct $x, y \in X$, let $I(\mathcal{T}; x, y)$ denote the set of interior vertices of $\mathcal{T}$ in the path connecting x and y.

Proposition 3.1.

(i) *Let* $\mathcal{T}$ *be a phylogenetic* X*-tree with at least one interior vertex. Then the number of distinct circular orderings for* $\mathcal{T}$ *is*

$$\prod_{v \in \mathring{V}(\mathcal{T})} (d(v) - 1)!$$

Furthermore, for all distinct elements $x, y \in X$, the proportion of these circular orderings for which y immediately follows x is

$$\prod_{v \in I(\mathcal{T};x,y)} (d(v)-1)^{-1}.$$

(ii) *Let π be a cyclic permutation of X and let $|X| = n$. Suppose that $n \geq 3$. Then the number of trivalent phylogenetic X-trees for which π is a circular ordering equals the (Catalan) number*

$$\frac{1}{n-1}\binom{2n-4}{n-2}.$$

Proof. To prove both parts of (i), it suffices to show that, for all (not necessarily distinct) elements $x, y \in X$, the number of circular orderings for $\mathcal{T}$ in which y immediately follows x is

$$\prod_{v \in O(\mathcal{T};x,y)} (d(v)-1)! \prod_{v \in I(\mathcal{T};x,y)} (d(v)-2)!, \tag{1}$$

where $O(\mathcal{T}; x, y)$ denotes the set of interior vertices of $\mathcal{T}$ not in the path connecting x and y. In the case $x = y$, the set $I(\mathcal{T}; x, y)$ is empty and the condition 'y immediately follows x' is redundant. The proof of (1) is by induction on the number of interior vertices of $\mathcal{T}$.

Let x and y be elements of X. If $|\mathring{V}(\mathcal{T})| = 1$, then the number of circular orderings for $\mathcal{T}$ in which y immediately follows x is the number of cyclic permutations of X in which y immediately follows x. The number of such cyclic permutations is $(n-1)!$ if $x = y$ and $(n-2)!$ if $x \neq y$. It follows that if $|\mathring{V}(\mathcal{T})| = 1$, then (1) holds.

Now suppose that $|\mathring{V}(\mathcal{T})| = k$, where $k \geq 2$, and that (1) holds for all ordered pairs of leaves of all phylogenetic trees with $k-1$ interior vertices. Let u be an interior vertex of $\mathcal{T}$ that is adjacent to exactly one other interior vertex and let X' denote the subset of X whose elements are precisely the elements of X adjacent to u. Now $\mathcal{T}$ has at least two such vertices, so, by making an appropriate choice for u, we may assume that $x \notin X'$.

Choose an element z in X' such that, if y is an element of X', z is chosen to be y. Let $\mathcal{T}'$ be the phylogenetic tree obtained from $\mathcal{T}$ by deleting the elements of $X' - z$ and suppressing u. Since $\mathcal{T}'$ is a phylogenetic tree and $|\mathring{V}(\mathcal{T}')| = |\mathring{V}(\mathcal{T})| - 1$, it follows by our induction assumption that the number of circular orderings for $\mathcal{T}'$ in which y immediately follows x is

$$\prod_{v \in O(\mathcal{T}';x,y)} (d(v)-1)! \prod_{v \in I(\mathcal{T}';x,y)} (d(v)-2)! \tag{2}$$

Now $X'|(X - X')$ is an X-split of $\mathcal{T}$ and so, for every circular ordering for $\mathcal{T}$ in which y immediately follows x, the elements of X' in this ordering are consecutive. Furthermore, the only proper subsets of X' that are blocks of an X-split of $\mathcal{T}$ are singletons. By our choice of u, there are just two (distinct) cases to consider:

(I) either $x = y$ or $z \neq y$; or

(II) $z = y$.

In (I), u is not in the path of $\mathcal{T}$ connecting x and y while, in (II), u is in the path of $\mathcal{T}$ connecting x and y. If (I) holds, then, for every circular ordering for $\mathcal{T}'$, we can replace z with any ordering of the elements of X' to obtain a circular ordering for $\mathcal{T}$ in which y immediately follows x. Furthermore, if (II) holds, then, for every circular ordering for $\mathcal{T}'$ in which y immediately follows x, we can replace y with any ordering of the elements in X' with y as the first element to obtain a circular ordering for $\mathcal{T}$ in which y immediately follows x. Moreover, all such circular orderings for $\mathcal{T}$ can be obtained in precisely one of these two ways as the deletion of $X' - z$ from any such ordering provides a circular ordering for $\mathcal{T}'$ in which y immediately follows x. Since any two circular orderings obtained in this way are distinct and $|X'| = d(u) - 1$, it follows by (2) that the number of circular orderings for $\mathcal{T}$ in which y immediately follows x is

$$|X'|! \prod_{v \in O(\mathcal{T}';x,y)} (d(v)-1)! \prod_{v \in I(\mathcal{T}';x,y)} (d(v)-2)! = \prod_{v \in O(\mathcal{T};x,y)} (d(v)-1)! \prod_{v \in I(\mathcal{T};x,y)} (d(v)-2)!$$

and

$$(|X'|-1)! \prod_{v \in O(\mathcal{T}';x,y)} (d(v)-1)! \prod_{v \in I(\mathcal{T}';x,y)} (d(v)-2)! = \prod_{v \in O(\mathcal{T};x,y)} (d(v)-1)! \prod_{v \in I(\mathcal{T};x,y)} (d(v)-2)!$$

in cases (I) and (II), respectively. Thus (1) holds, completing the proof of (i).

To prove (ii), let $C(n)$ denote the set of pairs $(\mathcal{T}, \pi)$, where $\mathcal{T}$ is a trivalent phylogenetic X-tree and π is a circular ordering for $\mathcal{T}$. We will count $C(n)$ in two ways. By Lemma 2.1(ii), the number of choices for $\mathcal{T}$ is $\frac{(2n-4)!}{(n-2)!2^{n-2}}$ and, by Lemma 2.1(i) and Proposition 3.1(i), for each $\mathcal{T}$, the number of choices for π is 2^{n-2}. Alternatively, we may count $C(n)$ by noting that the number of distinct cyclic permutations of X is exactly $(n-1)!$ and, for each such cyclic permutation π, the number of phylogenetic X-trees $\mathcal{T}$ for which $(\mathcal{T}, \pi) \in C(n)$ is precisely the number we want. Equating these two counts of $C(n)$ and then rearranging gives the desired result. □

An illustration of the system of paths described in the statement of Theorem 3.2 is shown in Fig. 2, where $(1, 6, 7, 4, 5, 2, 3)$ is the associated cyclic permutation.

Theorem 3.2. *Let $\pi = (x_1, x_2, \ldots, x_n)$ be a cyclic permutation of X and let $\mathcal{T}$ be a phylogenetic X-tree. For all $i \in \{1, 2, \ldots, n\}$, let P_i denote the path in $\mathcal{T}$ from x_i to x_{i+1}. Then*

(i) *Every pendant edge of $\mathcal{T}$ occurs in exactly two of the paths $P_1, P_2, \ldots, P_n$.*
(ii) *Every interior edge of $\mathcal{T}$ occurs in a positive and even number of the paths $P_1, P_2, \ldots, P_n$.*

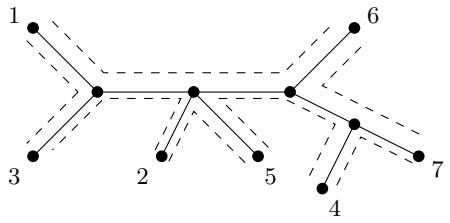

FIGURE 2. A set of paths (dashed lines) for $(1, 6, 7, 4, 5, 2, 3)$.

(iii) *π is a circular ordering for $\mathcal{T}$ if and only if every interior edge of $\mathcal{T}$ occurs in exactly two of the paths $P_1, P_2, \ldots, P_n$.*

Proof. Part (i) is an immediate consequence of the fact that, for all i, the element x_i occurs in exactly two of the pairs $(x_1, x_2), (x_2, x_3), \ldots, (x_n, x_1)$.

To prove (ii), let e be an interior edge of $\mathcal{T}$ and let $A|B$ be the X-split of $\mathcal{T}$ corresponding to e. Without loss of generality, we may assume that $x_1 \in A$. Then there is an element x_i of A such that x_{i+1} is an element of B, in which case e is an edge in the path P_i. Furthermore, there is an element x_j of B such that x_{j+1} is an element of A, in which case e is an edge in the path P_j, where $P_i \neq P_j$. Hence e occurs in at least two of the paths $P_1, P_2, \ldots, P_n$. Furthermore, by extending this argument, it is easily seen that the number of such paths is even. This completes the proof of (ii).

We next prove (iii). Suppose that every interior edge of $\mathcal{T}$ occurs in exactly two of the paths $P_1, P_2, \ldots, P_n$. The proof of the sufficient part of (iii) is by induction on the size of X. Evidently, if $n \leq 3$, then π is a circular ordering for $\mathcal{T}$. Now assume that $n \geq 4$ and that this direction holds for all phylogenetic trees with $n-1$ leaves. Let $\mathcal{T}'$ be the phylogenetic tree obtained from $\mathcal{T}$ by deleting x_n and suppressing any resulting degree-two vertex. Then, as every interior edge of $\mathcal{T}$ occurs in exactly two of the paths $P_1, P_2, \ldots, P_n$, every interior edge of $\mathcal{T}'$ occurs in exactly two of the paths $P_1, P_2, \ldots, P_{n-2}, P'_{n-1}$, where P'_{n-1} is the path in $\mathcal{T}'$ from x_{n-1} to x_1. Therefore, by our induction assumption, $\pi' = (x_1, x_2, \ldots, x_{n-1})$ is a circular ordering for $\mathcal{T}'$ and so $\Sigma(\mathcal{T}') \subseteq \Sigma^\circ(\pi')$.

Now consider $\mathcal{T}$. Let σ be an element of $\Sigma(\mathcal{T})$. We complete the sufficient direction of (ii) by showing that σ is an element of $\Sigma^\circ(\pi)$. If $\sigma = \{x_n\}|(X - x_n)$, then $\sigma \in \Sigma^\circ(\pi)$. Thus assume that $\sigma \neq \{x_n\}|(X - x_n)$. Then there is an interior edge e of $\mathcal{T}$ corresponding to σ and an element $z \in X - \{x_1, x_{n-1}, x_n\}$ that is in the same block of σ as x_n. Let σ' denote the $(X - x_n)$-split obtained from σ by removing x_n from the appropriate block. Clearly, σ' is an element of $\Sigma(\mathcal{T}')$ and, in particular, an element of $\Sigma^\circ(\pi')$. It now follows that $\sigma \in \Sigma^\circ(\pi)$ unless x_{n-1} and x_1 are both in the block of σ not containing x_n. But then, as $z \in \{x_2, x_3, \ldots, x_{n-2}\}$, at least four of the paths $P_1, P_2, \ldots, P_n$ contain e. This contradicts the initial assumption that every interior edge of $\mathcal{T}$ occurs in exactly two of the paths $P_1, P_2, \ldots, P_n$. Hence π is a circular ordering for $\mathcal{T}$.

For the converse of (iii), suppose that π is a circular ordering for $\mathcal{T}$, but there is an interior edge e of $\mathcal{T}$ that occurs in at least four of the paths $P_1, P_2, \ldots, P_n$. Let Q_1, Q_2, and Q_3 denote the first three such paths. Let a and b denote the initial and terminal vertices of Q_1, respectively, and let c and d denote the initial and terminal vertices of Q_3, respectively. Then a, b, c, and d are all distinct, and e induces an X-split of $\mathcal{T}$ in which a and c are in one block, and b and d are in the other block. Since π is a circular ordering for $\mathcal{T}$, it follows that a and c (as well as b and d) are adjacent in the cyclic permutation $\pi|\{a, b, c, d\}$. However, $\pi|\{a, b, c, d\} = (a, b, c, d)$; a contradiction. This completes the proof of (iii) and the theorem. □

Let S be a non-empty subset of X. For a phylogenetic X-tree $\mathcal{T}$, let $\mathcal{T}|S$ denote the phylogenetic S-tree for which

$$\Sigma(\mathcal{T}|S) = \{A \cap S | B \cap S : A|B \in \Sigma(\mathcal{T}) \text{ and } A \cap S, B \cap S \neq \emptyset\}.$$

Furthermore, for a cyclic permutation π of X, let $\pi|S$ denote the cyclic permutation of S obtained by restricting π to S.

The straightforward proof of the next lemma is omitted. For a phylogenetic tree $\mathcal{T}$, we denote the set of circular orderings for $\mathcal{T}$ by $\mathrm{o}(\mathcal{T})$

Lemma 3.3. *If $\mathcal{T}$ is a phylogenetic X-tree and S is a non-empty subset of X, then*

$$\mathrm{o}(\mathcal{T}|S) \supseteq \{\pi|S : \pi \in \mathrm{o}(\mathcal{T})\}.$$

Although not needed for this paper, we note that the converse of Lemma 3.3 also holds, in particular, $\mathrm{o}(\mathcal{T}|S) = \{\pi|S : \pi \in \mathrm{o}(\mathcal{T})\}$. However, the proof is less straightforward.

Proposition 3.4 allows us to use subsets of X of size four to analyse circular orderings of phylogenetic X-trees.

Theorem 3.4. *Let $\pi = (x_1, x_2, \ldots, x_n)$ be a cyclic permutation of X and let $\mathcal{T}$ be a phylogenetic X-tree. Then π is a circular ordering for $\mathcal{T}$ if and only if, for all subsets S of X of size four, $\pi|S$ is a circular ordering for $\mathcal{T}|S$.*

Proof. If π is a circular ordering for $\mathcal{T}$, then, by Lemma 3.3, $\pi|S$ is a circular ordering for $\mathcal{T}|S$.

Now suppose that π is not a circular ordering for $\mathcal{T}$. Then $\mathcal{T}$ must contain at least one interior edge. For all $i \in \{1, 2, \ldots, n\}$, let P_i denote the path in $\mathcal{T}$ from x_i to x_{i+1}. Since π is not a circular ordering for $\mathcal{T}$, it follows by Theorem 3.2(ii) and (iii) that there is an interior edge e of $\mathcal{T}$ that occurs in (at least) three of these paths. Let Q_1, Q_2, and Q_3 denote the first three such paths. Let a and b denote the initial and terminal vertices of Q_1, respectively, and let c and d denote the initial and terminal vertices of Q_3, respectively. As a, b, c, and d are distinct, $\pi|\{a, b, c, d\} = (a, b, c, d)$ and $\{a, c\}|\{b, d\}$ is a split of $\mathcal{T}|\{a, b, c, d\}$. But (a, b, c, d) is not a circular ordering for $\mathcal{T}|\{a, b, c, d\}$. This completes the proof of Theorem 3.4. □

For phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, we write $\mathcal{T} \leq \mathcal{T}'$ precisely if $\Sigma(\mathcal{T}) \subseteq \Sigma(\mathcal{T}')$. It is easily verified that $\leq$ induces a partial order on the set of phylogenetic X-trees.

Corollary 3.5. *Let $\mathcal{T}$ and $\mathcal{T}'$ be phylogenetic X-trees. Then*

(i) *$\mathcal{T} \leq \mathcal{T}'$ if and only if $\mathrm{o}(\mathcal{T}') \subseteq \mathrm{o}(\mathcal{T})$.*
(ii) *$\mathrm{o}(\mathcal{T}) = \mathrm{o}(\mathcal{T}')$ if and only if $\mathcal{T} \cong \mathcal{T}'$.*

Proof. By [12, Theorem 6.3.5], $\mathcal{T} \leq \mathcal{T}'$ if and only if, for all subsets S of X size four, $\mathcal{T}|S \leq \mathcal{T}'|S$. Also, it is readily checked that for all subsets S of X of size four, $\mathcal{T}|S \leq \mathcal{T}'|S$ if and only if $\mathrm{o}(\mathcal{T}'|S) \subseteq \mathrm{o}(\mathcal{T}|S)$. Combining these two characterizations, we deduce that $\mathcal{T} \leq \mathcal{T}'$ if and only if $\mathrm{o}(\mathcal{T}'|S) \subseteq \mathrm{o}(\mathcal{T}|S)$ for all subsets S of X of size four. Now, by Theorem 3.4, we have $\mathrm{o}(\mathcal{T}'|S) \subseteq \mathrm{o}(\mathcal{T}|S)$ for all subsets S of X of size four if and only if $\mathrm{o}(\mathcal{T}') \subseteq \mathrm{o}(\mathcal{T})$. Part (i) of the corollary now follows. Part (ii) is an immediate consequence of (i) and Proposition 2.2. □

4. Application to Tree Metrics

In this section, we apply circular orderings to the study of tree metrics. We show how the theory developed in the last section provides a convenient tool for establishing two new results concerning tree metrics, neither of which explicitly mentions circular orderings.

Let $\mathcal{T}$ be a phylogenetic X-tree and suppose that the edges of $\mathcal{T}$ have real-valued weights assigned by a function $w : E(\mathcal{T}) \to \mathbb{R}$. For all $x, y \in X$, let $P(\mathcal{T}; x, y)$ denote the set of edges of $\mathcal{T}$ in the path connecting vertices x and y. Define the map $d_{(\mathcal{T};w)} : X \times X \to \mathbb{R}$ by setting, for all $x, y \in X$,

$$d_{(\mathcal{T};w)}(x,y) = \begin{cases} \sum_{e \in P(\mathcal{T};x,y)} w(e), & \text{if } x \neq y, \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$l(\mathcal{T}, w) = \sum_{e \in E(\mathcal{T})} w(e).$$

We call $l(\mathcal{T}, w)$ the *total edge weight of $\mathcal{T}$*.

The proof of Lemma 4.1 is a direct consequence of Theorem 3.2. Part (i) of this lemma is a classical and well-known result, for example, see [6, 8, 16].

Lemma 4.1. *Let $\mathcal{T}$ be a phylogenetic X-tree and let $\pi = (x_1, x_2, \ldots, x_n)$ be a cyclic permutation of X. Let $w : E(\mathcal{T}) \to \mathbb{R}$ be an edge weighting of $\mathcal{T}$ and let $d = d_{(\mathcal{T};w)}$.*

(i) *If π is a circular ordering for $\mathcal{T}$, then*

$$l(\mathcal{T}, w) = \frac{1}{2} \sum_{i=1}^{n} d(x_i, x_{i+1}).$$

(ii) *Suppose that w is strictly positive on all edges of $\mathcal{T}$ and let*

$$w_{\min} = \min\{w(e) : e \in \mathring{E}(\mathcal{T})\}.$$

Then π is a circular ordering for $\mathcal{T}$ if and only if

$$l(\mathcal{T}, w) > \frac{1}{2}\sum_{i=1}^{n} d(x_i, x_{i+1}) - w_{\min}.$$

Recently, Pauplin [10] described an elegant representation of the total edge weight of any trivalent phylogenetic tree $\mathcal{T}$ with real-valued edge weighting w as a linear function of the $d_{(\mathcal{T};w)}(x, y)$ values. The first of our two results extends this representation to arbitrary phylogenetic trees, using an approach that explains the slightly mysterious coefficients appearing in the representation given in [10]. Essentially, our proof reveals that, for all distinct $x, y \in X$, the coefficient of $d_{(\mathcal{T};w)}(x, y)$ is the proportion of circular orderings for $\mathcal{T}$ in which y immediately follows x.

Let $\lambda : X \times X \to \mathbb{R}^{\geq 0}$ be the dissimilarity map on X defined, for all $x, y \in X$, in terms of the degrees of the interior vertices of $\mathcal{T}$ as follows:

$$\lambda(x, y) = \begin{cases} \prod_{v \in I(\mathcal{T};x,y)} (d(v) - 1)^{-1}, & \text{if } x \neq y, \\ 0, & \text{if } x = y. \end{cases}$$

Theorem 4.2. *Let $\mathcal{T}$ be a phylogenetic X-tree, $w : E(\mathcal{T}) \to \mathbb{R}$ be an edge weighting of $\mathcal{T}$, and $d = d_{(\mathcal{T};w)}$. Then*

$$l(\mathcal{T}, w) = \sum_{\{x,y\} \subseteq X} \lambda(x, y) d(x, y).$$

Proof. By Lemma 4.1,

$$l(\mathcal{T}, w) = \frac{1}{|\mathrm{o}(\mathcal{T})|} \sum_{(x_1,\ldots,x_n) \in \mathrm{o}(\mathcal{T})} \left[\frac{1}{2} \sum_{i=1}^{n} d(x_i, x_{i+1}) \right].$$

However, we may rewrite the right-hand side of this last equation as

$$\frac{1}{2} \frac{1}{|\mathrm{o}(\mathcal{T})|} \sum_{(x,y):x,y \in X} n_{\mathcal{T}}(x, y) d(x, y),$$

where $n_{\mathcal{T}}(x, y)$ is the number of circular orderings for $\mathcal{T}$ in which y immediately follows x. By Proposition 3.1(i), $\frac{n_{\mathcal{T}}(x,y)}{|\mathrm{o}(\mathcal{T})|} = \lambda(x, y)$ for all distinct $x, y \in X$. Thus

$$l(\mathcal{T}, w) = \frac{1}{2} \sum_{(x,y):x,y \in X} \lambda(x, y) d(x, y)$$

and the result now follows. □

We now turn to our second application, which concerns the reconstruction of a phylogenetic tree from a dissimilarity map δ. This is a central problem in molecular systematics (see, for example, [14]). In case δ is a tree metric, say $\delta = d_{(\mathcal{T};w)}$, it is straightforward to recover $\mathcal{T}$ from δ by standard methods. However, dissimilarity maps derived from data are generally some perturbation of—but not exactly equal

to—a tree metric. An important theoretical question, that is central to the statistical analysis of tree reconstruction methods, is how 'close' a dissimilarity map δ needs to be to a tree metric $d_{(\mathcal{T};w)}$ in order to ensure that a particular tree reconstruction method will recover $\mathcal{T}$ from δ. For certain tree reconstruction methods, it is relatively easy to answer this question; see, for example, [4, 9]. But for other methods, such as the popular 'neighbour-joining' method, the solution appears to require some intricate arguments. We next apply some of our results on circular orderings to investigate this question for one of the earliest methods proposed for reconstructing phylogenetic trees from dissimilarity maps.

For a dissimilarity map δ on X and a phylogenetic X-tree $\mathcal{T}$, we say that a positive edge weighting $w : E(\mathcal{T}) \to \mathbb{R}^{>0}$ of $\mathcal{T}$ is *admissible* for δ if $d_{(\mathcal{T};w)}(x, y) \geq \delta(x, y)$ for all $x, y \in X$. Furthermore, given a dissimilarity map δ on X the *minimum length tree reconstruction method* returns a phylogenetic X-tree that minimizes the total edge weight $l(\mathcal{T}, w)$ over all admissible edge weightings w for δ of all phylogenetic X-trees $\mathcal{T}$.

Theorem 4.3 shows that if a dissimilarity map δ is 'close enough' to one that is induced by a trivalent phylogenetic tree $\mathcal{T}$, then the minimum length tree reconstruction method applied to δ will return $\mathcal{T}$. Although the minimum length tree reconstruction method dates back 25 years (see [15]) and is one of the original techniques for reconstructing phylogenetic trees from dissimilarity maps, Theorem 4.3 is the first explicit convergence result for this method. For two dissimiliarity maps δ and δ' on X, the l_∞-metric is defined as

$$\|\delta - \delta'\|_\infty = \max\{|\delta(x, y) - \delta'(x, y)| : x, y \in X\}.$$

Theorem 4.3. *Let δ be a dissimilarity map on X and let $\mathcal{T}$ be a trivalent phylogenetic X-tree. Let w be a positive, real-valued edge weighting of $\mathcal{T}$ and set $d = d_{(\mathcal{T};w)}$. If*

$$\|d - \delta\|_\infty < \frac{1}{n} w_{\min},$$

where $n = |X|$ and $w_{\min} = \min\{w(e) : e \in \mathring{E}(\mathcal{T})\}$, then the minimum length tree reconstruction method applied to δ returns $\mathcal{T}$.

Proof. Clearly, the theorem holds if $|X| \leq 3$, so assume that $|X| \geq 4$. Then $\mathcal{T}$ has at least one interior edge. Let $\epsilon = \frac{1}{n} w_{\min}$, and let $w_1 : E(\mathcal{T}) \to \mathbb{R}$ be an edge weighting of $\mathcal{T}$ that agrees with w on $\mathring{E}(\mathcal{T})$ and, for all $e \in E(\mathcal{T}) - \mathring{E}(\mathcal{T})$, we have $w_1(e) = w(e) + \frac{1}{2}\epsilon$. Let $d_1 = d_{(\mathcal{T};w_1)}$. Then $d_1(x, y) \geq \delta(x, y)$ for all $x, y \in X$ and so w_1 is an admissible edge weighting of $\mathcal{T}$ for δ. Furthermore,

$$l(\mathcal{T}, w_1) = l(\mathcal{T}, w) + \frac{1}{2} w_{\min}. \tag{3}$$

Now suppose that $\mathcal{T}'$ is a phylogenetic X-tree that is different to $\mathcal{T}$. Since $\mathcal{T}$ is trivalent, $\mathcal{T} \not\leq \mathcal{T}'$ and so, by Corollary 3.5(i), there exists a cyclic permutation $(x_1, x_2, \ldots, x_n)$ in $\mathrm{o}(\mathcal{T}') - \mathrm{o}(\mathcal{T})$. Let w'be an admissible edge weighting of $\mathcal{T}'$ for δ and let $d' = d_{(\mathcal{T}';w')}$. Then, by Lemma 4.1(i),

$$l(\mathcal{T}', w') = \frac{1}{2}\sum_{i=1}^{n} d'(x_i, x_{i+1}) \geq \frac{1}{2}\sum_{i=1}^{n} \delta(x_i, x_{i+1}) > \frac{1}{2}\sum_{i=1}^{n} \big[d(x_i, x_{i+1}) - \epsilon\big]. \tag{4}$$

Moreover, since $(x_1, x_2, \ldots, x_n)$ is not a circular ordering for $\mathcal{T}$, it follows by Lemma 4.1(ii) that

$$\frac{1}{2}\sum_{i=1}^{n} d(x_i, x_{i+1}) \geq l(\mathcal{T}, w) + w_{\min}. \tag{5}$$

Combining (3), (4), and (5), we deduce that

$$l(\mathcal{T}', w') > l(\mathcal{T}, w_1)$$

and so the minimum length tree reconstruction method applied to δ does indeed return $\mathcal{T}$. □

Concluding remarks. The two results we have described here illustrate how circular orderings can be a convenient vehicle for deriving results on tree metrics. A remaining question is whether Theorem 4.3 can be improved. In particular, the condition

$$\|d - \delta\|_\infty < \frac{1}{n} w_{\min}, \tag{6}$$

involves n on the right-hand side, while the analogous conditions for some other tree reconstruction methods do not involve n (see [1, 4, 9]). It would be interesting to know whether (6) can be improved to remove (or weaken) this dependence on n.

References

[1] K. Atteson, The performance of neighbor-joining methods of phylogenetic reconstruction, *Algorithmica* **25** (1999), 251-278.

[2] P. Buneman, The recovery of trees from measures of dissimilarity, in "Mathematics in the archaeological and historical sciences," Edinburgh University Press, 1971.

[3] P. Buneman, A note on the metric property of trees, *J. Combinatorial Theory Ser. B* **17** (1974), 48-50.

[4] P. L. Erdös, L. A. Székely, M. Steel, and T. Warnow, A few logs suffice to build (almost) all trees (II), *Theoret. Comput. Sci.* **221** (1999), 77-118.

[5] S. L. Hakimi and A. N. Patrinos, The distance matrix of a graph and its tree realization, *Quart. Appl. Math.* **30** (1972), 255-269.

[6] M. D. Hendy, Minimality of trees constructed from dissimilarity data, *Ars Combin.* **17** (1984), 203-222.

[7] C. Korostensky and G. H. Gonnet, Using traveling salesman problem algorithms for evolutionary tree construction, *Bioinformatics* **16** (2000), 619-627.

[8] V. Makarenkov and B. Leclerc, Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees, in "Mathematical hierarchies and biology," DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 37, American Mathematical Society, 1997.

[9] V. Moulton and M. Steel, Retractions of finite distance functions onto tree metrics, *Discrete Appl. Math.* **91** (1999), 215-233.

[10] Y. Pauplin, Direct calculation of a tree length using a distance matrix, *J. Mol. Evol.* **51** (2000), 41-47.

[11] E. Schröder, Vier combinatorische probleme, *Zeit. Math. Physik* **15** (1870), 361-376.

[12] C. Semple and M. Steel, "Phylogenetics," Oxford University Press, Oxford, 2003.

[13] J. M. S. Simões-Pereira, A note on the tree realizability of a distance matrix, *J. Combininatorial Theory* **6** (1969), 303-310.

[14] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis, Phylogenetic inference, in "Molecular Systematics," 2nd edn, Sinauer, Sunderland, USA, 1996.

[15] M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer, Additive evolutionary trees, *J. Theoret. Biol.* **64** (1977), 199-213.

[16] S. V. Yushmanov, Construction of a tree with p leaves from $2p-3$ elements of its distance matrix (Russian), *Matematicheskie Zametki* **35** (1984), 877-887.
[17] K. A. Zaretskii, Constructing trees from the set of distances between pendant vertices, *Uspehi Matematiceskih Nauk* **20** (1965), 90-92.

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: c.semple@math.canterbury.ac.nz, m.steel@math.canterbury.ac.nz