Phylogenetic trees defined by at most three characters

Katharina T. Huber[a] Simone Linz[b] Vincent Moulton[a]
Charles Semple[c]

Submitted: Nov 16, 2023; Accepted: XXX; Published: XXX

Abstract

In evolutionary biology, phylogenetic trees are commonly inferred from a set of characters (partitions) of a collection of biological entities (e.g., species or individuals in a population). Such characters naturally arise from molecular sequences or morphological data. Interestingly, it has been known for some time that any binary phylogenetic tree can be (convexly) defined by a set of at most four characters, and that there are binary phylogenetic trees for which three characters are not enough. Thus, it is of interest to characterise those phylogenetic trees that are defined by a set of at most three characters. In this paper, we provide such a characterisation, in particular proving that a binary phylogenetic tree T is defined by a set of at most three characters precisely if T has no internal subtree isomorphic to a certain tree.

Mathematics Subject Classifications: 05C05, 92D15

1 Introduction

In evolutionary biology, phylogenetic trees are typically inferred from alignments of molecular sequence data like DNA or protein sequences [6]. Each row of such an alignment represents a biological entity (e.g., a species or an individual in a population) and each column is referred to as a *character*. In mathematical terms, each character is simply a partition of the set of the biological entities in question. If a character has only two states that, for example, indicate the presence or absence of a biological feature, then the character is called binary and corresponds to a bipartition. More frequently, however, biologists analyse data sets that consist of multistate characters, where a character can take on two or more states.

[a] School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom (k.huber@uea.ac.uk, v.moulton@uea.ac.uk).
[b] School of Computer Science, University of Auckland, Auckland 1142, New Zealand (s.linz@auckland.ac.nz).
[c] School of Mathematics and Statistics, University of Canterbury, Christchurch 8140, New Zealand (charles.semple@canterbury.ac.nz).

https://doi.org/10.37236/00000

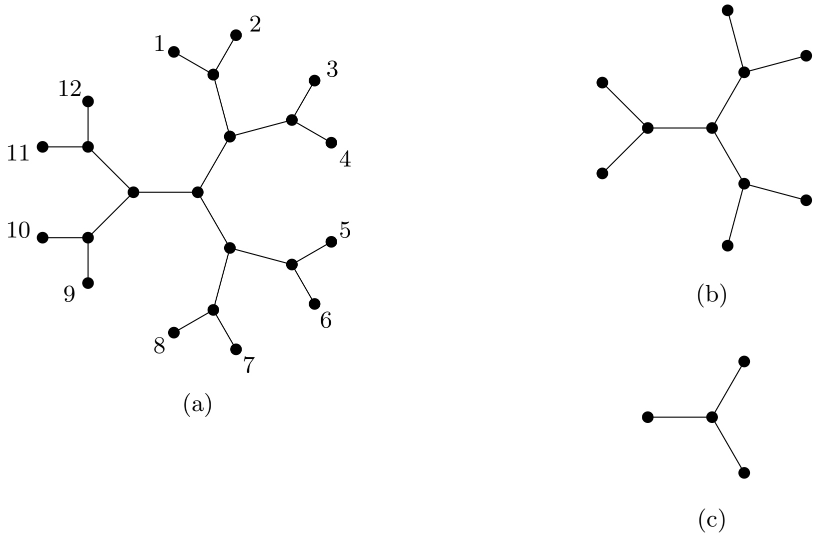


Figure 1: (a) A phylogenetic tree T with leaf set $X = \{1, 2, \ldots, 12\}$ that is not defined by any set of at most three characters. (b) The snowflake. (c) The 3-star. The snowflake is isomorphic (in the usual graphical sense) to the maximal internal subtree of T.

A fundamental question in the study of character evolution is whether or not a collection $\mathcal{C}$ of characters is *compatible* [10, Chapter 4]. Biologically speaking, compatibility of $\mathcal{C}$ indicates that there exists a phylogenetic tree T (i.e., an unrooted tree without degree-two vertices whose set X of leaves corresponds to the biological entities) on which each character χ in $\mathcal{C}$ evolves without any so-called parallel or reverse transitions. This implies that each character state of χ only evolves once on T, in which case $\mathcal{C}$ is *convex* on T. Stated another way, for each χ in $\mathcal{C}$ the subtrees of T spanned by the elements in each of the parts of χ are pairwise vertex disjoint.

If $\mathcal{C}$ is a collection of binary characters, the Splits Equivalence Theorem [3] can be used to decide if $\mathcal{C}$ is compatible. Moreover, an elegant graph-theoretic result that is based on chordalisations of the so-called partition intersection graph $\mathrm{Int}(\mathcal{C})$ of $\mathcal{C}$ (formally defined in Section 2) characterises when collections of multistate characters are compatible [4, 7, 11]. Based on this characterisation, it was further shown in [8] that there exists a certain type of chordalisation of $\mathrm{Int}(\mathcal{C})$ that is unique precisely if $\mathcal{C}$ *defines* a phylogenetic tree T, that is, $\mathcal{C}$ is convex on T and any other phylogenetic tree on which $\mathcal{C}$ is convex is isomorphic to T.

This last result begs the following question: How many characters are needed to define a given binary phylogenetic tree (i.e., a phylogenetic tree in which every vertex has degree one or three), when the number of character states is unbounded? Surprisingly, Semple and Steel [9] showed that five characters suffice, a bound that was subsequently sharpened to four by Huber et al. [5]. Moreover, as shown in [9], four is a tight upper bound since there exist binary phylogenetic trees that are not defined by three characters. Indeed, it turns out that the smallest such tree has twelve leaves and is shown in Figure 1(a) (see Lemma 7). Since the collection of binary phylogenetic trees defined by two characters is well understood (see, for example, [10, Chapter 4.8, Exercise 10] and below), in this paper we provide an answer to the following problem: Characterise those binary phylogenetic trees that are defined by a set of at most three characters.

The main result of this paper (Theorem 1) gives a solution to this problem in terms of forbidden subtrees in the form of the 6-leaf tree in Figure 1(b) which is sometimes called the *snowflake*. To state it, an *internal edge* of a tree T is a non-pendant edge and an *internal subtree* of T is a subtree whose edges are all internal.

Theorem 1. *Let T be a binary phylogenetic tree. Then T is defined by a set of at most three characters if and only if T has no internal subtree isomorphic to the snowflake.*

The analogous result for binary phylogenetic trees defined by a set of at most two characters is given by the following theorem, an immediate consequence of a result stated in [10] (see Theorem 6). Up to isomorphism, we refer to the unique tree with four vertices, three of which are leaves, as the 3-*star*. An illustration of the 3-star is shown in Figure 1(c).

Theorem 2. *Let T be a binary phylogenetic tree. Then T is defined by a set of at most two characters if and only if T has no internal subtree isomorphic to the* 3*-star.*

Consisting of two main ingredients, the proof of Theorem 1 essentially works as follows. First, we define three operations. Two of these operations, which we collectively call cherry modifications, extend a binary phylogenetic tree by attaching either one or two new leaves to a cherry, where a cherry refers to two leaves that are adjacent to the same internal vertex. The third operation, which we call a cherry union, amalgamates two binary phylogenetic trees across two cherries with a leaf in common. Second, we analyse sets of three characters that arise from certain edge-colourings of binary phylogenetic trees called internal 3-colourings. Using these concepts and extending the concept of the partition intersection graph of a set of characters to the partition intersection graph of an internal 3-colouring, we consider how the partition intersection graph arising from an internal 3-colouring behaves relative to the aforementioned operations. In particular, for proving the necessary direction of Theorem 1, we show that if a binary phylogenetic tree T has an internal subtree isomorphic to the snowflake, then, up to isomorphism, T can be obtained from the binary phylogenetic tree shown in Figure 1(a) by applying a sequence of cherry modifications, where each modification results in a binary phylogenetic tree not defined by a set of at most three characters (see Theorem 15). Conversely, for the sufficient direction of Theorem 1 (see Theorem 19), we inductively show that if a binary phylogenetic tree T does not have an internal subtree isomorphic to the snowflake, then either it is a special type of binary phylogenetic tree or it is the cherry union of two binary phylogenetic trees each of which is defined by a set of at most three characters. In both cases, it will follow that T is defined by a set of at most three characters.

The rest of this paper is organised as follows. In the next section, we consider sets of characters that define a binary phylogenetic tree, and state some useful results concerning such sets and their relationship with partition intersection graphs from [9] and [11]. In Section 3, we show how to define binary phylogenetic trees using internal edge-colourings. In particular, we essentially show that any binary phylogenetic tree with at least six

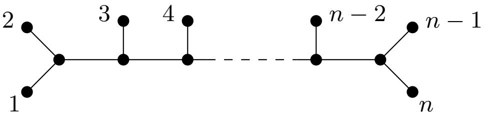


Figure 2: A caterpillar with leaf set $[n]$, where $n \geqslant 3$.

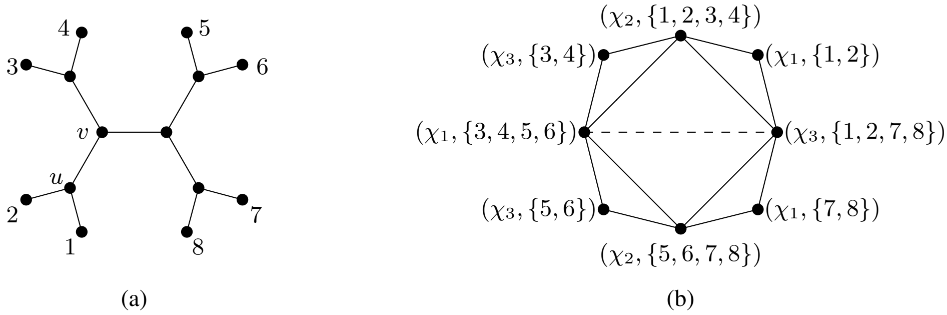


Figure 3: (a) A binary phylogenetic X-tree, where $X = [8]$. (b) The partition intersection graph $\text{Int}(\mathcal{C})$ of $\mathcal{C} = \{\chi_1, \chi_2, \chi_3\}$ (solid edges) and a restricted chordal completion of $\text{Int}(\mathcal{C})$ (solid and dashed edges), where $\chi_1 = \{\{1,2\},\{3,4,5,6\},\{7,8\}\}$, $\chi_2 = \{\{1,2,3,4\},\{5,6,7,8\}\}$, and $\chi_3 = \{\{1,2,7,8\},\{3,4\},\{5,6\}\}$.

leaves is defined by a set of three characters if and only if it can be defined by an internal 3-colouring (Proposition 12). In Section 4, we establish the necessary direction of Theorem 1. This relies on the two types of cherry modifications. The sufficient direction of Theorem 1 is established in Section 5 and relies on the cherry union of two phylogenetic trees. The paper concludes with a brief discussion in Section 6.

2 Preliminaries

Throughout the paper, X denotes a finite set with $|X| \geqslant 3$ and, for any positive integer k, we set $[k] = \{1, 2, \dots, k\}$. Furthermore, for a graph G, the vertex and edge sets of G are denoted by $V(G)$ and $E(G)$, respectively. For concepts from phylogenetics, we shall mainly use the terminology given in [10].

Phylogenetic trees. A *phylogenetic (X-)tree* T is a tree with leaf set X and no vertices of degree two. In addition, T is *binary* if every *internal vertex* (i.e., non-leaf vertex) of T has degree three. A binary phylogenetic tree is a *caterpillar* if every internal vertex is adjacent to a leaf. Such a phylogenetic tree is shown in Figure 2. Note that any binary phylogenetic X-tree with $3 \leqslant |X| \leqslant 5$ is a caterpillar.

For a binary phylogenetic X-tree T, a pair (x, y) of distinct leaves $x, y \in X$ is a *cherry* of T if x and y are adjacent to a common vertex. Note that the order of x and y in (x, y) does not matter. Also, note that every binary phylogenetic tree has at least one cherry (see, for example, [10, Proposition 1.2.5]). In Figure 3(a), $(3, 4)$ is a cherry. Furthermore, for a non-empty subset $A \subseteq X$, we let $T(A)$ denote the minimal subtree of T connecting the leaves in A.

Characters. A *character on X* is a partition of X, that is, a collection of non-empty subsets of X (or *parts*) whose pairwise intersections are empty and whose union is X. We say that a character χ on X is *convex* on a phylogenetic X-tree T if $T(A)$ and $T(B)$ are vertex disjoint for every distinct $A, B \in \chi$, and that a set $\mathcal{C}$ of characters is *convex* on T if every character in $\mathcal{C}$ is convex on T. Furthermore, $\mathcal{C}$ is *compatible* if there is a phylogenetic tree on which $\mathcal{C}$ is convex. A set $\mathcal{C}$ of characters on X *defines* T if $\mathcal{C}$ is convex on T and any phylogenetic X-tree T' that shares this property with T is *isomorphic* to T (that is, there is a graph isomorphism between T and T' whose restriction to X is the identity). Note that if $\mathcal{C}$ defines T, then T is necessarily binary.

Given a set $\mathcal{C}$ of characters that is convex on a binary phylogenetic tree T, we say that $\chi \in \mathcal{C}$ *distinguishes* an internal edge $e = \{u, v\}$ of T if there exist distinct $A, B \in \chi$ and distinct elements $x, y \in A$ and $w, z \in B$, such that u but not v lies on the path in T between x and y, and v but not u lies on the path in T between w and z. In addition, we say that T is *distinguished by* $\mathcal{C}$ if every internal edge of T is distinguished by some character in $\mathcal{C}$. To illustrate, consider the collection $\mathcal{C} = \{\chi_1, \chi_2, \chi_3\}$ of characters on $X = [8]$, where $\chi_1 = \{\{1, 2\}, \{3, 4, 5, 6\}, \{7, 8\}\}$, $\chi_2 = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}\}$, and $\chi_3 = \{\{1, 2, 7, 8\}, \{3, 4\}, \{5, 6\}\}$. It is easily seen that $\mathcal{C}$ is convex on the binary phylogenetic X-tree T shown in Figure 3(a). Furthermore, $\mathcal{C}$ distinguishes T. For example, the edge $\{u, v\}$ is distinguished by χ_1.

The next lemma is well known but never explicitly stated.

Lemma 3. *Let T be a binary phylogenetic X-tree, and let $\mathcal{C}$ be a set of characters on X that distinguishes T. Then no two incident internal edges of T are distinguished by the same character in $\mathcal{C}$.*

Proof. Let $e = \{u, v\}$ and $f = \{v, w\}$ be internal edges of T. Since e is distinguished by $\mathcal{C}$, there is a character χ in $\mathcal{C}$ and states A and B in χ such that $T(A)$ contains u but not v and $T(B)$ contains v but not u. But then $T(B)$ contains f and, in particular w, and so f is distinguished by a character in $\mathcal{C}$ that is not χ. □

Partition intersection graphs. Given a set $\mathcal{C}$ of characters, we let $\mathrm{Int}(\mathcal{C})$ denote the *partition intersection graph of* $\mathcal{C}$, that is, the graph with vertex set

$$\{(\chi, A) : \chi \in \mathcal{C} \text{ and } A \in \chi\}$$

and edge set

$$\{\{(\chi, A), (\chi', B)\} : A \cap B \neq \emptyset\}.$$

Note that, necessarily, if (χ, A) and (χ', B) are joined by an edge, then $\chi \neq \chi'$.

A graph is *chordal* if every cycle with at least four vertices has an edge connecting two nonconsecutive vertices. A *restricted chordal completion* G of $\mathrm{Int}(\mathcal{C})$ is a chordal graph that is obtained from $\mathrm{Int}(\mathcal{C})$ by adding only edges that join vertices whose first components are distinct. We refer to the edges of G not in $\mathrm{Int}(\mathcal{C})$ as *completion edges.* Furthermore, G is *minimal* if the deletion of any completion edge of G results in a graph that is not chordal. Continuing the example above, the partition intersection graph $\mathrm{Int}(\mathcal{C})$ of $\mathcal{C} = \{\chi_1, \chi_2, \chi_3\}$

is shown in Figure 3(b) (solid edges), and a restricted chordal completion of $\mathrm{Int}(\mathcal{C})$ is shown in the same figure (solid and dashed edges).

The next two results, established in [11, Proposition 3] and [8, Theorem 1.2], respectively, will be key in what follows. More specifically, they characterise sets of characters that are convex on a phylogenetic tree and sets that define a phylogenetic tree.

Theorem 4. *Let $\mathcal{C}$ be a set of characters on X. Then $\mathcal{C}$ is convex on a phylogenetic X-tree if and only if* $\mathrm{Int}(\mathcal{C})$ *has a restricted chordal completion.*

Theorem 5. *Let T be a binary phylogenetic X-tree, and let $\mathcal{C}$ be a set of characters on X. Then $\mathcal{C}$ defines T if and only if*

(i) *$\mathcal{C}$ is convex on T and $\mathcal{C}$ distinguishes T, and*

(ii) $\mathrm{Int}(\mathcal{C})$ *has a unique minimal restricted chordal completion.*

Because of their frequency of use, Theorems 4 and 5 will be often used without reference in Sections 4 and 5.

Theorem 2 is an immediate consequence of the next theorem. It follows from [10, Chapter 4.8, Exercise 10]. However, for completeness, we include a proof.

Theorem 6. *Let T be a binary phylogenetic tree. Then T is defined by a set of at most two characters if and only if T is a caterpillar.*

Proof. Let X denote the leaf set of T. Using Lemma 3, it is easily checked that T is defined by a set of at most one character if and only if $|X| \in \{3, 4\}$. Thus we may assume that $|X| \geqslant 5$. If T is defined by a set of two characters, then, by Lemma 3, T has no internal vertex incident with three internal edges. Thus every internal vertex of T is adjacent to a leaf, and so T is a caterpillar.

Conversely, suppose that T is a caterpillar. Without loss of generality, we may assume that the leaf set of T is $[n]$ and that its leaves are labelled as shown in Figure 2. Say n is even, and consider the set $\{\chi_1, \chi_2\}$ of characters where

$$\chi_1 = \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \ldots, \{n-1, n\}\}$$

and

$$\chi_2 = \{\{1, 2, 3\}, \{4, 5\}, \{6, 7\}, \ldots, \{n-2, n-1, n\}\}.$$

Now $\{\chi_1, \chi_2\}$ is convex on T and distinguishes T. Furthermore, $\mathrm{Int}(\{\chi_1, \chi_2\})$ is a path, and so $\mathrm{Int}(\{\chi_1, \chi_2\})$ is chordal. In particular, $\mathrm{Int}(\{\chi_1, \chi_2\})$ has a unique restricted chordal completion, namely itself. Hence, by Theorem 5, $\{\chi_1, \chi_2\}$ defines T. A similar argument holds if n is odd. Thus if T is a caterpillar, then T is defined by two characters, completing the proof of the theorem. □

We end this section with three lemmas. The first is mentioned in [9, p. 182], and established in [2, Section 5].

Lemma 7. *The binary phylogenetic tree shown in Figure 1(a) is not defined by a set of at most three characters.*

Lemma 8. *Let $\mathcal{C}$ be a compatible collection of characters on X, and let G be a minimal restricted chordal completion of* $\operatorname{Int}(\mathcal{C})$*. If e is a completion edge of G, then e joins two vertices of a vertex-induced cycle of* $\operatorname{Int}(\mathcal{C})$ *with at least four vertices.*

Proof. Let E' be the subset of completion edges of G whose end vertices do not join two vertices in the same vertex-induced cycle of $\operatorname{Int}(\mathcal{C})$ of size at least four, and suppose that E' is nonempty. Then, as G is a minimal restricted chordal completion of $\operatorname{Int}(\mathcal{C})$, the graph $G\backslash E'$ is not chordal. But then there is a vertex-induced cycle C' of $G\backslash E'$ with at least four vertices all of which are in the same vertex-induced cycle of $\operatorname{Int}(\mathcal{C})$. This implies that C' is a vertex-induced cycle of G, a contradiction. This completes the proof of the lemma. □

The third lemma shows that a set of characters that defines a binary phylogenetic tree T and contains a character such that one of its parts is a singleton (i.e., has cardinality one) can be slightly modified so that the resulting set of characters still defines T, but has one less singleton.

Lemma 9. *Let $\mathcal{C}$ be a set of characters on X, and suppose that $\mathcal{C}$ defines a binary phylogenetic X-tree T. Let χ be a character in $\mathcal{C}$, and suppose that $A \in \chi$ with $|A| = 1$. Then there exists some $B \in \chi - \{A\}$ such that $\mathcal{C}' = (\mathcal{C} - \{\chi\}) \cup \{\chi'\}$ defines T, where*

$$\chi' = (\chi - \{A, B\}) \cup \{A \cup B\}.$$

Proof. Let $\chi = \{A_1, A_2, \ldots, A_k\}$ where $k \geqslant 2$, and suppose that $A = A_i$ for some $i \in [k]$. Since $\mathcal{C}$ defines T, it follows that χ is convex on T. Therefore there exists some $j \in [k]-\{i\}$ and a path in T from the leaf in A_i to a vertex in $T(A_j)$ whose edges are all contained in the set

$$E(T) - \big(E(T(A_1)) \cup E(T(A_2)) \cup \cdots \cup E(T(A_k))\big).$$

Let $\chi' = (\chi - \{A_i, A_j\}) \cup \{A_i \cup A_j\}$ and $\mathcal{C}' = (\mathcal{C} - \{\chi\}) \cup \{\chi'\}$. By construction, χ' is convex on T, and so $\mathcal{C}'$ is convex on T. We next show that $\mathcal{C}'$ defines T.

If $\mathcal{C}'$ does not define T, then there is a binary phylogenetic X-tree T' on which $\mathcal{C}'$ is convex and T' is not isomorphic to T. But then, as $\mathcal{C}'$ is convex on T', it follows that $\mathcal{C}$ is also convex on T', contradicting the fact that $\mathcal{C}$ defines T. Thus $\mathcal{C}'$ defines T, and so setting $B = A_j$ completes the proof of the lemma. □

3 Internal k-Colourings

Let k be a positive integer, and let T be a phylogenetic X-tree. A *k-assignment* γ is a map $\gamma : E^0(T) \rightarrow [k]$, where $E^0(T)$ is the set of internal edges of T. An *internal k-colouring* of T is a k-assignment γ with $\gamma(E^0(T)) = [k]$ such that every pair of adjacent internal edges in T are assigned different elements in $[k]$. For convenience, we view the elements in $[k]$

as colours. For an internal k-colouring γ and $c \in [k]$, we let $\pi(\gamma, c)$ denote the character on X that is obtained by removing all internal edges from T that are assigned colour c under γ and taking the collection of subsets of X that are contained within each of the resulting connected components. In addition, we set

$$\Pi(\gamma) = \{\pi(\gamma, c) : c \in [k]\}$$

and let $\text{Int}(\gamma)$ denote the partition intersection graph of $\Pi(\gamma)$. Clearly, $\Pi(\gamma)$ is convex on T and T is distinguished by $\Pi(\gamma)$. We shall say that a binary phylogenetic X-tree is *defined by an internal k-colouring* γ if it is defined by $\Pi(\gamma)$. Note that a character on X can contain a singleton whereas a character induced by an internal k-colouring, for some k, cannot.

The purpose of the next lemma is to clarify under which conditions a collection $\mathcal{C}$ of characters that defines a binary phylogenetic tree T equates to a collection $\Pi(\gamma)$ of characters that is induced by an internal k-colouring γ of T.

Lemma 10. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 4$, and let $\mathcal{C}$ be a set of characters on X with cardinality k, where $k \in [3]$, that is convex on T and also distinguishes T. Suppose that no character in $\mathcal{C}$ contains a singleton, that every internal edge in T is distinguished by exactly one character in $\mathcal{C}$, and that every character in $\mathcal{C}$ distinguishes at least one internal edge of T. Then there exists an internal k-colouring $\gamma : E^0(T) \to [k]$ such that $\mathcal{C} = \Pi(\gamma)$.*

Proof. Let $\mathcal{C} = \{\chi_1, \ldots, \chi_k\}$, where $k \in [3]$. Consider the k-assignment $\gamma : E^0(T) \to [k]$ that takes each internal edge of T to colour i if χ_i distinguishes that edge. Note that γ is well-defined since $|X| \geqslant 4$ implies that $E^0(T) \neq \emptyset$, and every element of $E^0(T)$ is distinguished by exactly one character in $\mathcal{C}$. Since $\mathcal{C}$ distinguishes T, it follows by Lemma 3 that any two internal edges in T that are incident with the same vertex are assigned different colours under γ. Moreover, since every character in $\mathcal{C}$ distinguishes at least one internal edge of T, it follows that $\gamma(E^0(T)) = [k]$. Hence γ is an internal k-colouring of T. To complete the proof of the lemma, we show that $\mathcal{C} = \Pi(\gamma)$.

For all i, let π_i denote $\pi(\gamma, c_i)$. Suppose that $\mathcal{C} \neq \Pi(\gamma)$. Since, for all i, we have that π_i is obtained by deleting all internal edges of T that are distinguished by χ_i, the definition of $\Pi(\gamma)$ implies that there must exist some $\pi_j \in \Pi(\gamma)$ such that χ_j refines π_j (i.e., there is some $A \in \chi_j$ and some $B \in \pi_j$ such that $A \subsetneq B$). Let A_1 and A_2 be distinct non-empty subsets of B such that $A_1, A_2 \in \chi_j$, and let P be the shortest path in T connecting $T(A_1)$ and $T(A_2)$. Since $|A_1|, |A_2| \geqslant 2$, the path P consists of internal edges of T. Furthermore, as A_1 and A_2 are subsets of B, no edge in P is distinguished by χ_j. To see this, if there is such an edge, then, by construction, A_1 and A_2 are in different parts of π_j, in which case either $A_1 \notin B$ or $A_2 \notin B$, a contradiction. Thus P has at least two edges; otherwise, P consists of a single edge distinguished by χ_j. Let u be the vertex of P that is adjacent to a vertex in $T(A_1)$ but is not in $T(A_1)$, and let $\{u, v\}$ be the edge of T incident with u but not in P. If u is incident with three internal edges of T, then, as no edge in P is distinguished by χ_j, the edge $\{u, v\}$ is distinguished by χ_j. But then χ_j is not convex on T, a contradiction. It follows that v is a leaf, in which case v appears as a singleton in

χ_j; otherwise, the edge of P incident with u and a vertex in $T(A_1)$ is distinguished by χ_j. This last contradiction implies that $\mathcal{C} = \Pi(\gamma)$. □

Clearly, a binary phylogenetic X-tree T is defined by an internal 1-colouring if and only if $|X| = 4$. The next proposition is an immediate consequence of Theorem 6 and Lemma 10.

Proposition 11. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 5$. Then T is defined by an internal* 2*-colouring if and only if T is a caterpillar.*

We now turn our attention to internal 3-colourings and establish the main result of this section. In view of Proposition 11, we focus on binary phylogenetic trees that are not caterpillars, that is, binary phylogenetic trees that have an internal subtree isomorphic to the 3-star. The next proposition is used several times in Sections 4 and 5.

Proposition 12. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 6$, and suppose that T is not a caterpillar. Then T is defined by a set of three characters on X if and only if T is defined by an internal* 3*-colouring of T.*

Proof. If T is defined by an internal 3-colouring γ of T, then T is defined by the set $\Pi(\gamma)$ of characters on X and this set has size three. To prove the converse, that is, if T is defined by a set $\mathcal{C}$ of three characters on X, then T is defined by an internal 3-colouring of T, we shall freely use Theorem 5.

Let $\mathcal{C}$ be a set of three characters on X that defines T. Note that, by repeated application of Lemma 9, we may assume that $\mathcal{C}$ contains no character that contains a singleton. Furthermore, since $\mathcal{C}$ defines T, every internal edge of T is distinguished by some character in $\mathcal{C}$ and, since T is not a caterpillar, every character in $\mathcal{C}$ must distinguish at least one internal edge of T. Let E_2 denote the set of internal edges of T distinguished by exactly two characters. Observe that, as $|\mathcal{C}| = 3$ and T has at least two internal edges, no internal edge of T is distinguished by exactly three characters in $\mathcal{C}$.

If $E_2 = \emptyset$, then, by Lemma 10, the characters in $\mathcal{C}$ induce an internal 3-colouring of T, and so the converse holds in this case. Therefore assume that $E_2 \neq \emptyset$ and let $e = \{u, v\}$ be an edge in E_2. If either u or v is not adjacent to a leaf of T, then, as $|\mathcal{C}| = 3$, we have that e is distinguished by exactly one character in $\mathcal{C}$, a contradiction. So assume that u and v are adjacent to leaves x and y, respectively. We next construct from $\mathcal{C}$ a set $\mathcal{C}'$ of three characters that defines T such that e is distinguished by exactly one character in $\mathcal{C}'$, the number of internal edges distinguished by exactly two characters in $\mathcal{C}'$ is $|E_2| - 1$, and no character in $\mathcal{C}'$ contains a singleton. This will complete the proof of the converse since, by repeatedly applying this construction to reduce the number of internal edges distinguished by exactly two characters, we eventually obtain a set $\mathcal{C}^*$ of three characters on X such that each internal edge of T is distinguished by exactly one character in $\mathcal{C}^*$, no character in $\mathcal{C}^*$ contains a singleton, and each character in $\mathcal{C}^*$ distinguishes some internal edge of T.

Let $\mathcal{C} = \{\chi_1, \chi_2, \chi_3\}$, and suppose that e is distinguished by χ_1 and χ_2. Let e_1 denote the edge incident with u that is neither e nor $\{u, x\}$, and let e_2 denote the edge incident with v that is neither e nor $\{v, y\}$. There are two cases to consider depending on whether (i) either e_1 or e_2 is pendant, and (ii) neither e_1 nor e_2 is pendant. In what follows, we prove (ii). The proof for (i) is similar, but more straightforward, and is omitted. Now consider (ii). Since $\mathcal{C}$ distinguishes every internal edge of T, it follows by Lemma 3 that e_1, as well as e_2, is distinguished by exactly one character, namely χ_3. Let A_1 and A_2 be the parts in χ_1 so that $x \in A_1$ and $y \in A_2$. Let $\chi_1' = (\chi_1 - \{A_1, A_2\}) \cup \{A_1 \cup A_2\}$ and let $\mathcal{C}' = (\mathcal{C} - \{\chi_1\}) \cup \{\chi_1'\}$. We now show that $\mathcal{C}'$ defines T.

As $\mathcal{C}$ is convex on T and distinguishes every internal edge of T, it follows that $\mathcal{C}'$ is also convex on T and distinguishes every internal edge of T. Consider the partition intersection graph $\mathrm{Int}(\mathcal{C})$ of $\mathcal{C}$, and let G be the unique minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C})$. By [8, Proposition 4.1], $\mathrm{Int}(\mathcal{C})$ is connected, and so G is connected. Furthermore $(\chi_3, \{x, y\})$ is a cut vertex of $\mathrm{Int}(\mathcal{C})$ and so, by Lemma 8, $(\chi_3, \{x, y\})$ is a cut vertex of G.

Now, consider the partition intersection graph $\mathrm{Int}(\mathcal{C}')$ of $\mathcal{C}'$. This graph is obtained from $\mathrm{Int}(\mathcal{C})$ by identifying the vertices (χ_1, A_1) and (χ_1, A_2), labelling the identified vertex as $(\chi_1', A_1 \cup A_2)$, deleting one of the two resulting parallel edges joining $(\chi_3, \{x, y\})$ and $(\chi_1', A_1 \cup A_2)$, and relabelling all other vertices of the form (χ_1, A_i), where $i \not\in \{1, 2\}$ with (χ_1', A_i). Observe that, as $(\chi_3, \{x, y\})$ is a cut vertex of $\mathrm{Int}(\mathcal{C})$,

$$\{(\chi_3, \{x, y\}), (\chi_1', A_1 \cup A_2)\}$$

is a vertex-cut of $\mathrm{Int}(\mathcal{C}')$. Moreover, as $\mathrm{Int}(\mathcal{C})$ is connected, $\mathrm{Int}(\mathcal{C}')$ is connected.

Let V_x (respectively, V_y) be the subset of $V(\mathrm{Int}(\mathcal{C})) - \{(\chi_3, \{x, y\})\}$ with the property that a vertex v is in V_x (respectively, V_y) if there is a path from v to (χ_1, A_1) (respectively, (χ_1, A_2)) in $\mathrm{Int}(\mathcal{C})$ avoiding $(\chi_3, \{x, y\})$. Note that $V_x \neq \emptyset$ because $(\chi_1, A_1) \in V_x$, and that $V_y \neq \emptyset$ because $(\chi_1, A_2) \in V_y$. Similarly, let V_x' (respectively, V_y') denote the subset of $V(\mathrm{Int}(\mathcal{C}'))$ obtained from V_x (respectively, V_y) by deleting (χ_1, A_1) (respectively, (χ_1, A_2)) and replacing (χ_1, A_i), where $i \not\in \{1, 2\}$, with (χ_1', A_i).

Let G' denote the graph obtained from $\mathrm{Int}(\mathcal{C}')$ by joining two vertices in $V(\mathrm{Int}(\mathcal{C}'))$ with an edge precisely if the corresponding vertices of $\mathrm{Int}(\mathcal{C})$ are joined by a completion edge in G. More precisely, G' is constructed from $\mathrm{Int}(\mathcal{C}')$ as follows:

(1) if vertices of the form (χ_2, B) and (χ_3, C) are joined by a completion edge in G, then the same vertices are joined by an edge in G';

(2) if a vertex of the form (χ_1, A_i) and a vertex t are joined by a completion edge in G, where $i \not\in \{1, 2\}$, then (χ_1', A_i) and t are joined by an edge in G'; and

(3) if a vertex of the form (χ_1, A_i) and a vertex t are joined by a completion edge in G, where $i \in \{1, 2\}$, then $(\chi_1', A_1 \cup A_2)$ and t are joined by an edge in G'.

Note that, as $(\chi_3, \{x, y\})$ is a cut vertex of G, it follows from the construction of G' that

$$\{(\chi_3, \{x, y\}), (\chi_1', A_1 \cup A_2)\}$$

is a vertex-cut of G'.

We now show that G' is a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$. Assume that C is a vertex-induced cycle of G' with at least four vertices. Then, as G is a restricted chordal completion of $\mathrm{Int}(\mathcal{C})$, and $\{(\chi_3, \{x,y\})\}$ and $\{(\chi_3, \{x,y\}), (\chi_1', A_1 \cup A_2)\}$ are vertex-cuts of G and G', respectively, it follows that $(\chi_3, \{x,y\})$ and $(\chi_1', A_1 \cup A_2)$ as well as a vertex in V_x' and a vertex in V_y' are vertices in C. But, since $\{(\chi_3, \{x,y\}), (\chi_1', A_1 \cup A_2)\}$ is an edge in $\mathrm{Int}(\mathcal{C}')$ and thus in G, this is a contradiction as C is a vertex-induced cycle of G' with at least four vertices. It follows that G' is a restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$. Furthermore, a similar argument shows that if G' is not a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$, then G is not a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C})$, a contradiction. Thus G' is a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$.

To see that G' is the unique minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$, suppose that G_1' is also a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$. Since $\{(\chi_3, \{x,y\}), (\chi_1', A_1 \cup A_2)\}$ is a vertex-cut of $\mathrm{Int}(\mathcal{C}')$ and the two vertices in this vertex-cut are joined by an edge, it follows by Lemma 8 that no completion edge in G_1' joins a vertex in V_x' to a vertex in V_y'. Now let G_1 be the graph obtained from G_1' by reversing the construction described in (1)–(3) above. That is, G_1 is obtained from G_1' as follows:

(1)$'$ if vertices of the form (χ_2, B) and (χ_3, C) are joined by a completion edge in G_1', then the same vertices are joined by an edge in G_1;

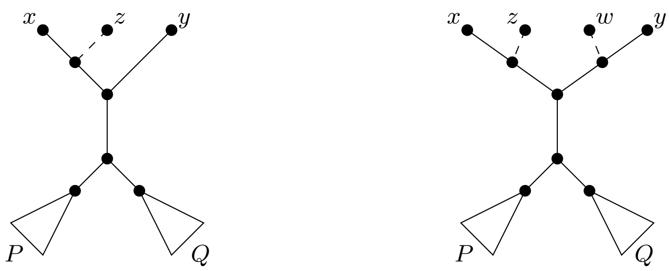
(2)$'$ if a vertex of the form (χ_1', A_i) and a vertex t are joined by a completion edge in G_1', where $i \not\in \{1,2\}$, then (χ_1, A_i) and t are joined by an edge in G_1; and

(3)$'$ if $(\chi_1', A_1 \cup A_2)$ and a vertex t in V_x' (resp. V_y') are joined by a completion edge in G_1', then (χ_1, A_1) (resp. (χ_1, A_2)) and t are joined by an edge in G_1.

Arguing as before, it is easily checked that G_1 is a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C})$. By the uniqueness of G, it follows that G_1 is isomorphic to G. Since the constructions given by (1)–(3) and (1)$'$–(3)$'$ undo each other, this implies that G_1' is isomorphic to G'. It follows that G' is the unique minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C}')$. Hence $\mathcal{C}'$ defines T and $\mathcal{C}'$ has the desired properties, in particular, the number of internal edges distinguished by exactly two characters in $\mathcal{C}'$ is $|E_2| - 1$ and no character in $\mathcal{C}'$ contains a singleton. □

4 Proof of Theorem 1: Necessary Direction

We begin the necessary direction of the proof of Theorem 1 by describing two operations that extend a binary phylogenetic tree. Let T be a binary phylogenetic X-tree, where $|X| \geqslant 3$, and let (x,y) be a cherry of T. The first operation, called a *fork modification*, adds a new leaf $z \not\in X$ to (x,y) as shown in Figure 4(a) to obtain a binary phylogenetic $(X \cup \{z\})$-tree T'. Formally, T' is obtained from T by subdividing the pendant edge incident with either x or y, say x, and then adjoining a new leaf $z \not\in X$ to T by adding an edge joining z and the subdivision vertex. The second operation, called a *balanced modification*, adds two new leaves $w, z \not\in X$ to (x,y) as shown in Figure 4(b) to obtain



(a) (b)

Figure 4: An illustration of (a) a fork modification and (b) a balanced modification, where P and Q denote the leaf sets of the corresponding subtrees.

a binary phylogenetic $(X \cup \{w, z\})$-tree T''. More precisely, T'' is obtained from T by subdividing each of the pendant edges incident with x and y, and then adjoining new leaves $w, z \notin X$ to T by adding an edge joining one of the leaves, say z, to the subdivision vertex on the pendant edge incident with x and an edge joining w to the subdivision vertex on the pendant edge incident with y. Collectively, we refer to fork and balanced modifications as *cherry modifications*. Although not explicitly needed for the paper, it is straightforward to show that if T is a binary phylogenetic X-tree, where $|X| \geqslant 3$, then, up to isomorphism, T can be obtained from a binary phylogenetic tree on three leaves by a sequence of cherry modifications.

The next lemma shows that a fork modification preserves the property of being defined by at most three characters. For ease of reading, if γ is an internal k-colouring of a binary phylogenetic tree, we write the vertices of $\mathrm{Int}(\gamma)$ in the form (c, A) instead of $(\pi(\gamma, c), A)$.

Lemma 13. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 3$, and let T' be a binary phylogenetic tree obtained from T by a fork modification. Then T is defined by a set of at most three characters if and only if T' is defined by a set of at most three characters.*

Proof. Let (x, y) be a cherry of T. We may assume that T' is obtained from T as shown in Figure 4(a). Thus (x, z) is a cherry of T' and the leaf set of T' is $X \cup \{z\}$. By Theorem 6, the lemma holds if T is a caterpillar as T' is also a caterpillar. So we may assume that T has an internal vertex incident with three internal edges and $|X| \geqslant 6$. There are two cases to consider depending on whether (i) exactly one of P and Q has size one and (ii) $|P|, |Q| \geqslant 2$. We will establish the lemma for (ii). The proof for (i) is similar and omitted.

Suppose that $|P|, |Q| \geqslant 2$, and T is defined by three characters. Then, by Proposition 12, T is defined by an internal 3-colouring γ. Let $\{c_1, c_2, c_3\}$ be the codomain of γ and let γ' be an internal 3-colouring of T' that extends γ, that is, γ' has the property that if $e \in E^0(T)$, then $\gamma'(e) = \gamma(e)$. Since $|P|, |Q| \geqslant 2$, we may assume without loss of generality that

$$(c_1, \{x, y\}),\ (c_2, \{x, y\} \cup P'),\ (c_3, \{x, y\} \cup Q')$$

are vertices of $\mathrm{Int}(\gamma)$, where P' and Q' are non-empty subsets of P and Q, respectively, while

$$(c_1, \{x, y, z\}),\ (c_2, \{x, z\}),\ (c_2, \{y\} \cup P'),\ (c_3, \{x, y, z\} \cup Q')$$

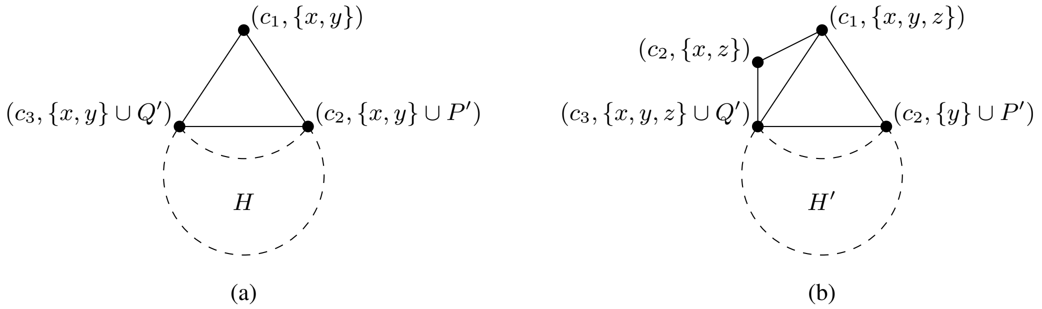


Figure 5: Illustrations of (a) $\mathrm{Int}(\gamma)$ and (b) $\mathrm{Int}(\gamma')$ in the proof of Lemma 13, where H and H' represent the parts of $\mathrm{Int}(\gamma)$ and $\mathrm{Int}(\gamma')$ not explicitly shown.

are vertices of $\mathrm{Int}(\gamma')$. The partition intersection graphs of $\Pi(\gamma)$ and $\Pi(\gamma')$ are illustrated in Figure 5. It is now easily checked that $\mathrm{Int}(\gamma')$ can be constructed from $\mathrm{Int}(\gamma)$ by relabelling $(c_1, \{x, y\})$ as $(c_1, \{x, y, z\})$, $(c_2, \{x, y\} \cup P')$ as $(c_2, \{y\} \cup P')$, and $(c_3, \{x, y\} \cup Q')$ as $(c_3, \{x, y, z\} \cup Q')$, and adding a new vertex $(c_2, \{x, z\})$ adjacent to precisely $(c_1, \{x, y, z\})$ and $(c_3, \{x, y, z\} \cup Q')$.

Since $\Pi(\gamma')$ is convex on T', the partition intersection graph $\mathrm{Int}(\gamma')$ has a restricted chordal completion by Theorem 4. If G' is a minimal restricted chordal completion of $\mathrm{Int}(\gamma')$, then, by Lemma 8, no completion edge of G' is incident with $(c_2, \{x, z\})$. Therefore if $\mathrm{Int}(\gamma')$ has two distinct minimal restricted chordal completions, then, by the above construction, $\mathrm{Int}(\gamma)$ has two distinct minimal restricted chordal completions, a contradiction as $\Pi(\gamma)$ defines T. Thus $\mathrm{Int}(\gamma')$ has a unique minimal restricted chordal completion. Since $\Pi(\gamma')$ distinguishes T', it follows that γ' defines T', and so T' is defined by three characters. The proof that if T' is defined by three characters, then T is defined by three characters is the same argument but in reverse. This completes the proof of the lemma. □

In contrast to a fork modification, a balanced modification does not necessarily preserve the property of being defined by a set of at most three characters. However, it does preserve the property of not being defined by a set of at most three characters.

Lemma 14. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 3$, and let T' be a binary phylogenetic tree obtained from T by a balanced modification. If T is not defined by a set of at most three characters, then T' is not defined by a set of at most three characters.*

Proof. Suppose that T is not defined by at most three characters. Then T is not a caterpillar and so T, and therefore T', has an internal vertex incident with three internal edges and $|X| \geqslant 6$. Let (x, y) be a cherry of T. We may assume that T' is obtained from T as shown in Figure 4(b). Thus (x, z) and (w, y) are cherries of T' and the leaf set of T' is $X \cup \{w, z\}$ and, as $|X| \geqslant 6$, either $|P| \geqslant 2$ or $|Q| \geqslant 2$. There are two cases: (i) exactly one of P and Q has size one and (ii) $|P|, |Q| \geqslant 2$. We will establish the lemma for (ii). The proof of (i) is similar and omitted.

Suppose that $|P|, |Q| \geqslant 2$ and, for the purpose of obtaining a contradiction, that T' is defined by three characters. By Proposition 12, T' is defined by an internal 3-colouring

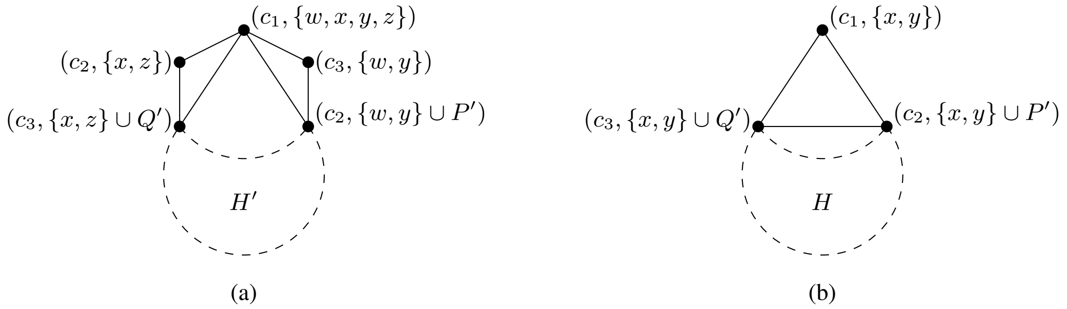

$(c_1, \{w, x, y, z\})$
$(c_2, \{x, z\})$
$(c_3, \{w, y\})$
$(c_3, \{x, z\} \cup Q')$
$(c_2, \{w, y\} \cup P')$
H'
(a)
$(c_1, \{x, y\})$
$(c_3, \{x, y\} \cup Q')$
$(c_2, \{x, y\} \cup P')$
H
(b)

Figure 6: Illustrations of (a) $\mathrm{Int}(\gamma')$ and (b) $\mathrm{Int}(\gamma)$ in the proof of Lemma 14, where H' and H represent the parts of $\mathrm{Int}(\gamma')$ and $\mathrm{Int}(\gamma)$ not explicitly shown.

γ'. Let $\{c_1, c_2, c_3\}$ be the codomain of γ' and let γ be the restriction of γ' to the edges of T. That is, for each $e \in E^0(T)$, we have $\gamma(e) = \gamma'(e)$. Since $|P|, |Q| \geqslant 2$, we may assume without loss of generality that

$$(c_1, \{w, x, y, z\}),\ (c_2, \{x, z\}),\ (c_2, \{w, y\} \cup P'),\ (c_3, \{w, y\}),\ (c_3, \{x, z\} \cup Q')$$

are vertices of $\mathrm{Int}(\gamma')$, where P' and Q' are non-empty subsets of P and Q, respectively, while

$$(c_1, \{x, y\}),\ (c_2, \{x, y\} \cup P'),\ (c_3, \{x, y\} \cup Q')$$

are vertices of $\mathrm{Int}(\gamma)$. The partition intersection graphs of $\Pi(\gamma')$ and $\Pi(\gamma)$ are illustrated in Figure 6. A routine check shows that $\mathrm{Int}(\gamma)$ can be constructed from $\mathrm{Int}(\gamma')$ by deleting $(c_2, \{x, z\})$ and $(c_3, \{w, y\})$, relabelling $(c_2, \{w, y\} \cup P')$ as $(c_2, \{x, y\} \cup P')$ and $(c_3, \{x, z\} \cup Q')$ as $(c_3, \{x, y\} \cup Q')$, and joining $(c_2, \{x, y\} \cup P')$ and $(c_3, \{x, y\} \cup Q')$ with an edge.

Since γ' defines T', the partition intersection graph $\mathrm{Int}(\gamma')$ has a unique minimal restricted chordal completion G'. As $\mathrm{Int}(\gamma')$ has no vertex-induced cycles of size at least four containing either $(c_2, \{x, z\})$ or $(c_3, \{w, y\})$, it follows by Lemma 8 that no completion edge of G' is incident with these vertices. Furthermore,

$$\{(c_2, \{w, y\} \cup P'), (c_3, \{x, z\} \cup Q')\}$$

is a completion edge of G'. To see this, let G be a minimal restricted chordal completion of $\mathrm{Int}(\gamma)$. Since γ is convex on T, such a graph exists. Let E_1 denote the set of completion edges of G and let E_1' denote the collection of 2-element subsets of $V(\mathrm{Int}(\gamma'))$ obtained from E_1 by replacing $(c_2, \{x, y\} \cup P')$ with $(c_2, \{w, y\} \cup P')$ and $(c_3, \{x, y\} \cup Q')$ with $(c_3, \{x, z\} \cup Q')$. Since G is a minimal restricted chordal completion of $\mathrm{Int}(\gamma)$, it follows by the above construction that the graph obtained from $\mathrm{Int}(\gamma')$ by adding the edges in E_1' as well as the edge $\{(c_2, \{w, y\} \cup P'), (c_3, \{x, z\} \cup Q')\}$ to $\mathrm{Int}(\gamma')$ is a minimal restricted chordal completion G_1' of $\mathrm{Int}(\gamma')$. If $\{(c_2, \{w, y\} \cup P'), (c_3, \{x, z\} \cup Q')\}$ is not a completion edge of G', then, as this edge is a completion edge of G_1', it follows that $\mathrm{Int}(\gamma')$ has two distinct minimal restricted chordal completions. This is a contradiction as γ' defines T'. It now follows that, as $\mathrm{Int}(\gamma')$ has a unique minimal restricted chordal completion, $\mathrm{Int}(\gamma)$ has a unique minimal restricted chordal completion. Since $\Pi(\gamma)$ is convex on T and distinguishes T, we deduce that γ defines T, and so T is defined by three characters. This last contradiction completes the proof of the lemma. □

The next theorem is the necessary direction of Theorem 1.

Theorem 15. *Let T be a binary phylogenetic X-tree, and suppose that T has an internal subtree isomorphic to the snowflake. Then, up to isomorphism, T can be obtained from the phylogenetic tree in Figure 1(a) by a sequence of cherry modifications. In particular, T is not defined by a set of at most three characters.*

Proof. Since T has an internal subtree isomorphic to the snowflake, $|X| \geqslant 12$. The proof is by induction on $n = |X|$. If $n = 12$, then T is isomorphic to the phylogenetic tree shown in Figure 1(a) and so, by Lemma 7, the theorem holds.

Now suppose that the theorem holds for all binary phylogenetic trees with at most $n - 1 \geqslant 12$ leaves. Let S be an internal subtree of T isomorphic to the snowflake, and let v be the unique vertex of S that is at distance two from each of its leaves. Now let u_1 be an internal vertex of T at maximum distance from v. Observe that u_1 is adjacent to two leaves, say x and z. Let u denote the unique internal vertex of T adjacent to u_1, and let u_2 denote the vertex of T adjacent to u, that is not u_1, but at the same distance from v as u_1. If u_2 is a leaf, label this vertex y. Otherwise, u_2 is adjacent to two leaves, say y and w.

Let T' be the binary phylogenetic tree such that T is obtained from T' by applying either a fork modification to the cherry (x, y) or a balanced modification to the cherry (x, y). Since $n \geqslant 13$, it follows by the choice of u_1 that neither u_1 nor u_2 are vertices in S. Thus T' has an internal subtree isomorphic to the snowflake. Therefore, by induction, T' can be obtained from the binary phylogenetic tree shown in Figure 1(a) by a sequence of cherry modifications and, moreover, T' is not defined by three characters. Theorem 15 now follows by Lemmas 13 and 14. □

5 Proof of Theorem 1: Sufficient Direction

In this section, we complete the proof of Theorem 1 by proving the sufficient direction. To do this, we first introduce the operation of cherry union. Let T_1 and T_2 be two binary phylogenetic trees with leaf sets X_1 and X_2, respectively, such that $|X_1|, |X_2| \geqslant 4$ and $X_1 \cap X_2 = \{x\}$. In addition, suppose that T_1 has a cherry (x, y_1) and T_2 has a cherry (x, y_2). Let T be the binary phylogenetic tree with leaf set $(X_1 \cup X_2) - \{y_1, y_2\}$ that is obtained from T_1 and T_2 by deleting the leaf x in exactly one of T_1 and T_2, identifying the vertices y_1 and y_2, and suppressing the two resulting degree-2 vertices. We say that T is the *cherry union* of T_1 and T_2, and denote T by $T_1 \Box T_2$. An illustration of a cherry union is shown in Figure 7.

The next lemma shows that a cherry union preserves the property of being defined by a set of at most three characters. As in the last section, if γ is an internal k-colouring of a binary phylogenetic tree, we write the vertices of $\mathrm{Int}(\gamma)$ in the form (c, A) instead of $(\pi(\gamma, c), A)$.

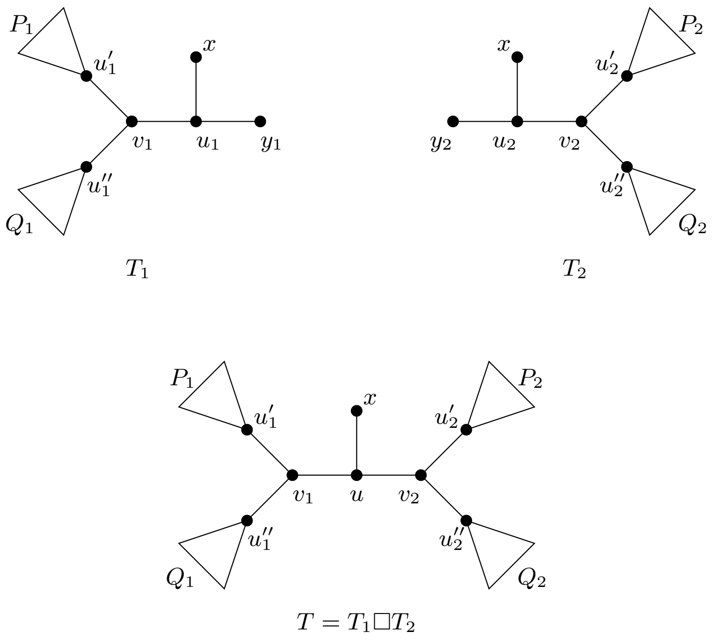


Figure 7: The cherry union T of two binary phylogenetic trees T_1 and T_2, where the leaf set of T_1 is $P_1 \cup Q_1 \cup \{x, y_1\}$, the leaf set of T_2 is $P_2 \cup Q_2 \cup \{x, y_2\}$, and the leaf set of T is $P_1 \cup Q_1 \cup P_2 \cup Q_2 \cup \{x\}$.

Lemma 16. *Let T_1 and T_2 be two binary phylogenetic trees with leaf sets X_1 and X_2, respectively, where $|X_1|, |X_2| \geqslant 4$. Suppose that $X_1 \cap X_2 = \{x\}$, and (x, y_1) and (x, y_2) are cherries in T_1 and T_2, respectively. If T_1 and T_2 are each defined by a set of at most three characters, then $T_1 \square T_2$ is defined by a set of at most three characters.*

Proof. We may assume that T_1, T_2, and $T_1 \square T_2$ are as shown in Figure 7. Suppose that T_1 and T_2 are each defined by a set of at most three characters. If one of T_1 and T_2, say T_2 is a caterpillar, then it is easily seen that, up to leaf labels, $T_1 \square T_2$ is obtained from T_1 by a sequence of fork modifications and so, by Lemma 13, $T_1 \square T_2$ is defined by most three characters. Thus we may assume that neither T_1 nor T_2 is a caterpillar. By Proposition 12, T_1 and T_2 are defined by internal 3-colourings γ_1 and γ_2, respectively. Without loss of generality, we may assume that the codomain of γ_1 and the codomain of γ_2 is $\{c_1, c_2, c_3\}$. Moreover, by recolouring if necessary, we may assume that if u_1 and u_2 denote the vertices of T_1 and T_2 adjacent to x and y_1, and adjacent to x and y_2, respectively, then the (unique) internal edges of T_1 and T_2 incident with u_1 and u_2 are assigned different colours.

Since neither T_1 nor T_2 is a caterpillar, $|P_1|, |P_2| \geqslant 2$. There are three cases to consider: (i) $|Q_1| = 1 = |Q_2|$, (ii) exactly one of Q_1 and Q_2 has size one, and (iii) $|Q_1|, |Q_2| \geqslant 2$. We establish the lemma for (iii). The proofs of (i) and (ii) are similar and omitted.

Suppose that (iii) holds. Then, without loss of generality, we may assume that $(c_1, \{x, y_1\})$, $(c_2, \{x, y_1\} \cup P_1')$, and $(c_3, \{x, y_1\} \cup Q_1')$ are vertices of $\mathrm{Int}(\gamma_1)$, where P_1' and Q_1' are non-empty subsets of P_1 and Q_1, respectively. Similarly, $(c_2, \{x, y_2\})$, $(c_1, \{x, y_2\} \cup P_2')$,

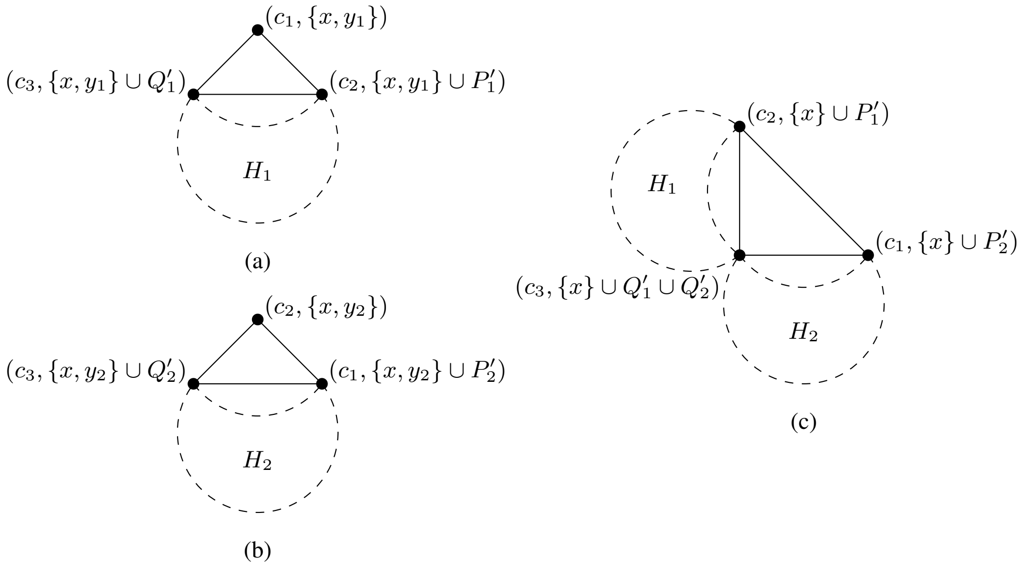

$(c_1, \{x, y_1\})$
$(c_3, \{x, y_1\} \cup Q_1')$
$(c_2, \{x, y_1\} \cup P_1')$
H_1
(a)
$(c_2, \{x, y_2\})$
$(c_3, \{x, y_2\} \cup Q_2')$
$(c_1, \{x, y_2\} \cup P_2')$
H_2
(b)
$(c_2, \{x\} \cup P_1')$
H_1
$(c_1, \{x\} \cup P_2')$
$(c_3, \{x\} \cup Q_1' \cup Q_2')$
H_2
(c)

Figure 8: Illustrations of (a) $\mathrm{Int}(\gamma_1)$, (b) $\mathrm{Int}(\gamma_2)$, and (c) $\mathrm{Int}(\gamma)$ in the proof of Lemma 16, where H_1 and H_2 represent the parts of $\mathrm{Int}(\gamma_1)$ and $\mathrm{Int}(\gamma_2)$ not explicitly shown.

and $(c_3, \{x, y_2\} \cup Q_2')$ are vertices of $\mathrm{Int}(\gamma_2)$, where P_2' and Q_2' are non-empty subsets of P_2 and Q_2, respectively. Illustrations of $\mathrm{Int}(\gamma_1)$ and $\mathrm{Int}(\gamma_2)$ are shown in Figure 8. Let γ be the internal 3-colouring of $T = T_1 \Box T_2$ induced by γ_1 and γ_2. Thus, in reference to Figure 7, $\{u, v_1\}$ and $\{u_2'', v_2\}$ are coloured c_1, $\{u_1'', v_1\}$ and $\{u, v_2\}$ are coloured c_2, and $\{u_1', v_1\}$ and $\{u_2', v_2\}$ are coloured c_3. Since γ is an internal 3-colouring of T, it follows that $\Pi(\gamma)$ is convex on T and distinguishes T.

Now consider $\mathrm{Int}(\gamma)$, and observe that $(c_1, \{x\} \cup P_2')$, $(c_2, \{x\} \cup P_1')$, and $(c_3, \{x\} \cup Q_1' \cup Q_2')$ are vertices of $\mathrm{Int}(\gamma)$. Also, as $\Pi(\gamma)$ is convex on T, we have in addition to these three vertices, for all $i \in \{1, 2, 3\}$, that (c_i, D) is a vertex of $\mathrm{Int}(\gamma)$ if and only if either (c_i, D) is a vertex of $\mathrm{Int}(\gamma_1)$ or (c_i, D) is a vertex of $\mathrm{Int}(\gamma_2)$. It is now easily checked that $\mathrm{Int}(\gamma)$ can be constructed from $\mathrm{Int}(\gamma_1)$ and $\mathrm{Int}(\gamma_2)$ by identifying the vertices $(c_1, \{x, y_1\})$ and $(c_1, \{x, y_2\} \cup P_2')$, $(c_2, \{x, y_1\} \cup P_1')$ and $(c_2, \{x, y_2\})$, and $(c_3, \{x, y_1\} \cup Q_1')$ and $(c_3, \{x, y_2\} \cup Q_2')$ together with the corresponding edges, and then relabelling the identified vertices as $(c_1, \{x\} \cup P_2')$, $(c_2, \{x\} \cup P_1')$, and $(c_3, \{x\} \cup Q_1' \cup Q_2')$, respectively. An illustration of $\mathrm{Int}(\gamma)$ is shown in Figure 8.

Let G be a minimal restricted chordal completion of $\mathrm{Int}(\gamma)$. Since $(c_1, \{x\} \cup P_2')$, $(c_2, \{x\} \cup P_1')$, and $(c_3, \{x\} \cup Q_1' \cup Q_2')$ is a 3-clique of $\mathrm{Int}(\gamma)$, it follows by the above construction that if C is a vertex-induced cycle of $\mathrm{Int}(\gamma)$ with at least four vertices, then, modulo replacing $(c_2, \{x\} \cup P_1')$ with $(c_2, \{x, y_1\} \cup P_1')$ and $(c_3, \{x\} \cup Q_1' \cup Q_2')$ with $(c_3, \{x, y_1\} \cup Q_1')$, or $(c_1, \{x\} \cup P_2')$ with $(c_1, \{x, y_2\} \cup P_2')$ and $(c_3, \{x\} \cup Q_1' \cup Q_2')$ with $(c_3, \{x, y_2\} \cup Q_2')$, the cycle C is a vertex-induced cycle of either $\mathrm{Int}(\gamma_1)$ or $\mathrm{Int}(\gamma_2)$ with at least four vertices. Thus, by Lemma 8 we deduce that if $\mathrm{Int}(\gamma)$ has two minimal restricted chordal completions, then either $\mathrm{Int}(\gamma_1)$ or $\mathrm{Int}(\gamma_2)$ has two minimal restricted chordal completions, contradicting that γ_1 and γ_2 define T_1 and T_2, respectively. Hence

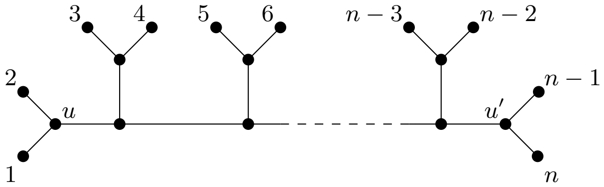


Figure 9: A cherried caterpillar with leaf set $[n]$, where $n \geqslant 4$.

G is the unique minimal restricted chordal completion of $\mathrm{Int}(\gamma)$. Since γ is convex on T and distinguishes T, the internal 3-colouring γ defines T. This completes the proof of the lemma. □

A binary phylogenetic X-tree T is a *cherried caterpillar* if either $|X| = 4$ or T can be obtained from a caterpillar by replacing each leaf with a pair of leaves in a cherry, that is, for each pendant edge of a caterpillar, subdividing it and adjoining a new leaf by adding an edge joining the new leaf and the subdivision vertex. A cherried caterpillar is illustrated in Figure 9.

Lemma 17. *Let T be a cherried caterpillar. Then T is defined by a set of at most three characters.*

Proof. Without loss of generality, we may assume that the leaf set of T is $[n]$ and that its leaves are labelled as shown in Figure 9. If $n = 4$, then the lemma trivially holds. So assume that $|X| \geqslant 6$. Let u and u' be the internal vertices of T adjacent to the leaves 1 and 2, and adjacent to the leaves $n-1$ and n, respectively. Now consider an internal 3-colouring γ of T that assigns each edge on the path joining u and u' one of two colours, say c_1 and c_2, and assigns all remaining internal edges the third colour, say c_3. Let $\{\chi_1, \chi_2, \chi_3\}$ denote the set of characters on $[n]$ induced by γ, where χ_i is the character induced by c_i for all $i \in \{1, 2, 3\}$. If $|X| = 6$, then $\mathrm{Int}(\mathcal{C})$ is chordal, and so it has a unique minimal restricted chordal completion. Furthermore, if $|X| \geqslant 8$, then it is easily checked that $\mathrm{Int}(\mathcal{C})$ has a unique vertex-induced cycle C with at least four vertices. Moreover, C has a unique vertex whose first coordinate is χ_3. The second coordinate of this vertex is $\{1, 2, n-1, n\}$.

Let G be a minimal restricted chordal completion of $\mathrm{Int}(\mathcal{C})$. If there is a completion edge of G joining a vertex in C whose first coordinate is χ_1 and a vertex in C whose first coordinate is χ_2, then G contains a cycle C' with at least four vertices all of which are vertices of C whose first coordinates are either χ_1 or χ_2. But by considering the subgraph of G induced by the vertices of C', it is easily checked that this subgraph is not chordal, a contradiction. Therefore all of the completion edges of G are incident with the unique vertex in C whose first coordinate is χ_3. It now follows that $\mathrm{Int}(\gamma)$ has a unique minimal restricted chordal completion. Since $\mathcal{C}$ is convex on T and $\mathcal{C}$ distinguishes T, the lemma follows by Theorem 5. □

Lemma 18. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 4$, and suppose that T has no internal subtree isomorphic to the snowflake. Then either T is a cherried caterpillar or T has a leaf not in a cherry.*

Proof. If T has a leaf not in a cherry, then the lemma holds, so suppose that every leaf of T is in a cherry. If $|X| = 4$, then T is a cherried caterpillar, so we may assume that $|X| \geqslant 6$. Let u and u' be internal vertices of T such that, amongst all pairs of internal vertices, the length of the path P connecting u and u' is maximised. Let $P = u, v_1, v_2, \ldots, v_k, u'$, where $k \geqslant 1$ as $|X| \geqslant 6$. By maximality, u is adjacent to two leaves and u' is adjacent to two leaves. For all $i \in \{1, 2, \ldots, k\}$, let w_i denote the vertex of T adjacent to v_i that is not on P. Since every leaf of T is in a cherry, w_i is not a leaf, and so w_i is an internal vertex of T for all i. Consider w_1. Since the length of the path connecting w_1 and u' is the same as the length of P, it follows by the maximality of P that w_1 adjacent to two leaves. Similarly, w_k is also adjacent to two leaves. If $k \in \{1, 2\}$, this implies that T is a cherried caterpillar. So assume that $k \geqslant 3$. Now consider w_j, where $j \in \{2, 3, \ldots, k-1\}$. If w_j is not adjacent to two leaves, then, as every leaf of T is in a cherry, w_j is adjacent to two internal vertices w'_j and w''_j, neither of which is v_j. But then the internal subtree of T induced by the vertices in

$$\{w'_j, w''_j, w_j, v_j, v_{j-1}, w_{j-1}, v_{j-2}, v_{j+1}, w_{j+1}, v_{j+2}\},$$

where $v_{j-2} = u$ if $j = 2$ and $v_{j+2} = u'$ if $j = k-1$, is isomorphic to the snowflake, a contradiction. Thus, for all j, we have that w_j is adjacent to two leaves. In particular, T is a cherried caterpillar. This completes the proof of the lemma. □

The next theorem proves the sufficient direction of Theorem 1, thereby completing its proof.

Theorem 19. *Let T be a binary phylogenetic X-tree, where $|X| \geqslant 4$, and suppose that T has no internal subtree isomorphic to the snowflake. Then T is defined by a set of at most three characters.*

Proof. The proof is by induction on $n = |X|$. If $n = 4$, the theorem trivially holds. Suppose that $n \geqslant 5$, and that the theorem holds for all binary phylogenetic trees whose leaf sets have size at most $n - 1 \geqslant 4$. If T is a cherried caterpillar, then, by Lemma 17, T is defined by a set of at most three characters. Therefore assume that T is not a cherried caterpillar. Then, by Lemma 18, T has a leaf, x say, not in a cherry. Let v be the internal vertex of T adjacent to x, and let u_1 and u_2 be the internal vertices of T adjacent to v. Let F_1 denote the forest obtained from T by deleting the two edges incident with u_2 that are not $\{v, u_2\}$, and let T_1 denote the binary phylogenetic X_1-tree obtained from the component of F_1 containing u_1 by relabelling u_2, now a leaf, as z_1, where $z_1 \notin X$. Similarly, let F_2 denote the forest obtained from T by deleting the two edges incident with u_1 that are not $\{v, u_1\}$, and let T_2 denote the binary phylogenetic X_2-tree obtained from the component of F_2 containing u_2 by relabelling u_1, now a leaf, as z_2, where $z_2 \notin X \cup \{z_1\}$. Observe that $4 \leqslant |X_1|, |X_2| \leqslant n - 1$, $X_1 \cap X_2 = \{x\}$ and, more particularly, T is isomorphic to $T_1 \square T_2$.

Since T has no internal subtree isomorphic to the snowflake, it is easily checked that T_1 has no internal subtree isomorphic to the snowflake. Thus, by the induction assumption, T_1 is defined by three characters. Similarly, T_2 is also defined by three characters. As T is isomorphic to $T_1 \square T_2$, Theorem 19 now follows by Lemma 16. □

6 Discussion

Since each binary phylogenetic tree is defined by at most four characters [5], the results presented in this paper partitions the set of all binary phylogenetic X-trees, where $|X| \geqslant 5$, into three classes. In particular, those binary phylogenetic X-trees that are defined by a set of exactly two, those defined by a set of three but no less than three, and those defined by a set of four but no less than four characters. Moreover, given a binary phylogenetic X-tree T, we can decide which class T is contained in in time that is linear in $|X|$. To see this, let $\mathcal{C}_2(X)$, $\mathcal{C}_3(X)$, and $\mathcal{C}_4(X)$ denote these three classes of binary phylogenetic trees, respectively, and note that a binary phylogenetic X-tree has $2|X| - 2$ vertices [10, Proposition 2.1.3]. Now, as T is contained in $\mathcal{C}_2(X)$ if and only if T is a caterpillar tree by Theorem 6, it is clear that containment in $\mathcal{C}_2(X)$ can be checked in $O(|X|)$ time. Furthermore, if T is not a caterpillar, then we can again check in $O(|X|)$ time if T is contained in $\mathcal{C}_3(X)$ or $\mathcal{C}_4(X)$ using Theorem 1 by simply considering, for each internal vertex u, the internal vertices of T at distance at most two from u, and checking whether or not they induce a snowflake.

There remain some interesting questions and investigations. For example, given a binary phylogenetic X-tree T in $\mathcal{C}_3(X)$ or $\mathcal{C}_4(X)$, can we determine (up to the natural notion of equivalence) the number of ways that T can be defined by three or four characters, respectively? Also, it would be interesting to compute

$$\lim_{|X|\to\infty} \frac{|\mathcal{C}_2(X) \cup \mathcal{C}_3(X)|}{|\mathcal{C}_4(X)|}.$$

This could be of practical interest since, if this limit is 0, it would imply that, as the size of X grows, almost all binary phylogenetic X-trees are in $\mathcal{C}_4(X)$.

Finally, the notion of defining a binary phylogenetic tree can be generalised to the weaker notion of "identifiability" [1]. In particular, an *X-tree* is an ordered pair $(T; \phi)$ consisting of a tree T with vertex set, V say, and a map $\phi : X \to V$ such that if $v \in V$ has degree at most two, then $v \in \phi(X)$. A phylogenetic X-tree is an X-tree in which ϕ is a bijection from X to the set of leaves of T. Intuitively, an X-tree can be obtained by contracting some edges of a binary phylogenetic X-tree [10]. A collection of characters on X *identifies* an X-tree $(T; \phi)$ if $\mathcal{C}$ is convex on T (analogous to that described in this paper) and all other X-trees on which $\mathcal{C}$ is convex are "refinements" of T.

In [2], it is proven that if d is the maximum degree of any vertex in an X-tree $(T; \phi)$ and k is a positive integer, then, in case

$$k = 4\lceil \log_2(d-2) \rceil + 4,$$

there is a collection of k characters that identifies $(T; \phi)$ and, in case $k < \log_2 d$, there is no collection of k characters that identifies T. Bearing this in mind, it would be interesting to investigate whether the results in the present paper for binary phylogenetic trees can be extended in some way to X-trees.

Acknowledgements

KTH and VM would like to thank The Royal Society in the context of its International Exchanges Scheme for support and also the University of Canterbury for hosting them during a brief visit. SL and CS were supported by the New Zealand Marsden Fund. All authors would like to thank the Institute for Mathematical Sciences, National University of Singapore where some of the research was partially completed while they were visiting in 2023.

References

[1] M. Bordewich, K. T. Huber, and C. Semple. Identifying phylogenetic trees. *Discrete Mathematics*, 300: 30–43, 2005.

[2] M. Bordewich and C. Semple. Defining a phylogenetic tree with the minimum number of r-state characters, *SIAM Journal on Discrete Mathematics*, 29: 835–853, 2015.

[3] P. Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences.* Edinburgh University Press, 1971.

[4] P. Buneman. A characterisation of rigid circuit graphs. *Discrete Mathematics*, 9: 205–212, 1974.

[5] K. T. Huber, V. Moulton, and M. Steel. Four characters suffice to convexly define a phylogenetic tree. *SIAM Journal on Discrete Mathematics*, 18: 835–843, 2005.

[6] P. Lemey, M. Salemi, and A.-M. Vandamme. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Cambridge University Press, 2009.

[7] C. A. Meacham. Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In *Numerical Taxonomy*, volume G1 of *NATO ASI Series*, pages 304–314. Springer-Verlag, Berlin, 1983.

[8] C. Semple and M. Steel. A characterization for a set of partial partitions to define an X-tree. *Discrete Mathematics*, 247: 169–186, 2002.

[9] C. Semple and M. Steel. Tree reconstruction from multi-state characters. *Advances in Applied Mathematics*, 28: 169–184, 2002.

[10] C. Semple and M. Steel. *Phylogenetics*, Oxford University Press, Oxford, 2003.

[11] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9: 91–116, 1992.