SIAM J. DISCRETE MATH.
Vol. 29, No. 2, pp. 835–853

DEFINING A PHYLOGENETIC TREE WITH THE MINIMUM NUMBER OF r-STATE CHARACTERS*

MAGNUS BORDEWICH† AND CHARLES SEMPLE‡

Abstract. Semple and Steel (2002) showed that if $\mathcal{T}$ is a phylogenetic X-tree and $\mathcal{C}$ is a collection of r-state characters that defines $\mathcal{T}$, then $|\mathcal{C}| \geq \lceil (n-3)/(r-1) \rceil$, where $n = |X|$. In this paper, we show that, provided n is sufficiently large, this lower bound is sharp. Furthermore, we show that, for all $n \geq 13$, there exists a collection of 4-state characters of size $\lceil (n-3)/3 \rceil$ that defines $\mathcal{T}$, but there is a phylogenetic X-tree with $n = 12$ which is not defined by any set of 3 characters.

Key words. phylogenetic tree, r-state character, chordal graph

AMS subject classification. 05B35

DOI. 10.1137/130924469

1. Introduction.

A central task in evolutionary biology is the reconstruction of phylogenetic (evolutionary) trees. Such trees represent the ancestral history of a collection of present-day species. In biology, characters describe attributes of the species under consideration and are the typical data used for reconstructing phylogenetic trees. Characters can be morphological or genetic, such as the nucleotide at a certain position on a DNA sequence. A natural question to ask is how many characters are required to recover the correct phylogenetic tree? More precisely, given an arbitrary phylogenetic tree $\mathcal{T}$, how small can a collection $\mathcal{C}$ of characters be so that $\mathcal{T}$ is the only phylogenetic tree consistent with $\mathcal{C}$? If each character is allowed an unbounded number of "states," Semple and Steel [8] showed that $|\mathcal{C}| \leq 5$. Huber, Moulton, and Steel [6] improved this upper bound to $|\mathcal{C}| \leq 4$ and this result is sharp. However, in practice, characters with an unbounded number of states are unrealistic. In this paper, we derive the analogous sharp result for characters with a bounded number of states.

Throughout the paper, X is always a finite set with $|X| \geq 3$. A *phylogenetic X-tree* $\mathcal{T}$ is an unrooted tree with no degree-two vertices and whose leaf set is X. In addition, $\mathcal{T}$ is *binary* if every interior vertex has degree three. A *character* χ *on* X is a function from X into a set of character states. If $|\chi(X)| \leq r$, then χ is an *r-state character*. For example, if χ is the character that assigns one of the four nucleotides at a certain position on a DNA sequence, then χ is a 4-state character.

Let $\mathcal{T}$ be a phylogenetic X-tree and let χ be a character on X into a set C of character states. For each $\alpha \in \chi(X)$, let $\mathcal{T}(\alpha)$ denote the minimal subtree of $\mathcal{T}$ whose leaf set is $\chi^{-1}(\alpha)$. We say that χ is *convex* on $\mathcal{T}$ if the subtrees in $\{\mathcal{T}(\alpha) : \alpha \in C\}$ are vertex disjoint. More generally, for a collection $\mathcal{C}$ of characters on X, we say $\mathcal{C}$ is *convex* on $\mathcal{T}$ if each character in $\mathcal{C}$ is convex on $\mathcal{T}$. A collection $\mathcal{C}$ of characters on X *defines* a phylogenetic X-tree $\mathcal{T}$ if, up to isomorphism, $\mathcal{T}$ is the only phylogenetic X-tree for which $\mathcal{C}$ is convex, in which case, $\mathcal{T}$ is necessarily binary. The relevance of convexity to biology is discussed at the end of this section.

*Received by the editors June 10, 2013; accepted for publication (in revised form) January 28, 2015; published electronically April 21, 2015.
http://www.siam.org/journals/sidma/29-2/92446.html
†School of Engineering Computer Sciences, Durham University, Durham DH1 3LE, United Kingdom (m.j.r.bordewich@durham.ac.uk).
‡Biomathematics Research Centre, School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand (charles.semple@canterbury.ac.nz). This author was supported by the Allan Wilson Centre for Molecular Ecology and Evolution and the New Zealand Marsden Fund.

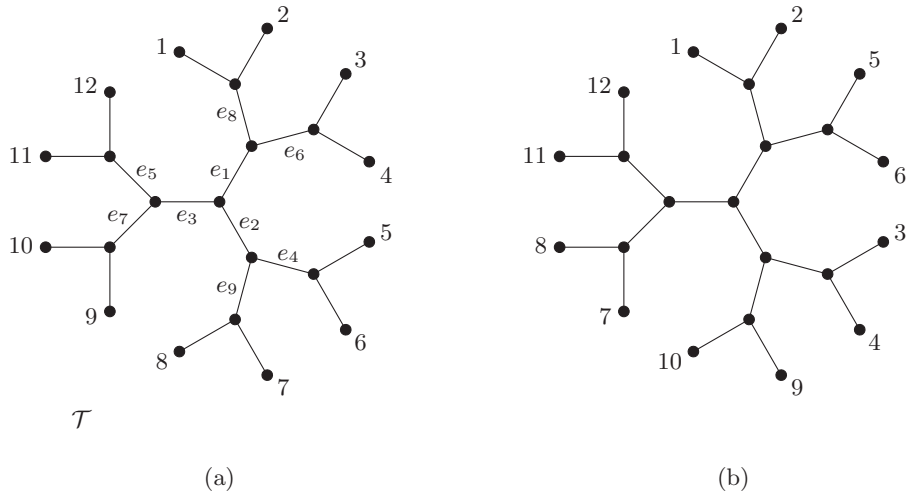

FIG. 1. (a) *A binary phylogenetic tree* $\mathcal{T}$ *on* $\{1, 2, \ldots, 12\}$ *that is not defined by a collection of three* 4*-state characters.* (b) *A binary phylogenetic tree on* $\{1, 2, \ldots, 12\}$ *for which the characters displayed by* $\{e_1, e_4, e_7\}$, $\{e_2, e_5, e_8\}$, *and* $\{e_3, e_6, e_9\}$ *are also convex.*

To illustrate, consider the binary phylogenetic X-tree $\mathcal{T}$ shown in Figure 1(a), where $X = \{1, 2, \ldots, 12\}$, and the 4-state character $\chi : X \to \{\alpha, \beta, \gamma, \delta\}$ defined by $\chi(1) = \chi(2) = \chi(3) = \chi(4) = \alpha$, $\chi(5) = \chi(6) = \beta$, $\chi(7) = \chi(8) = \chi(11) = \chi(12) = \gamma$, and $\chi(9) = \chi(10) = \delta$. Ignoring the edge labels for now, it is easily checked that $\mathcal{T}(\alpha)$, $\mathcal{T}(\beta)$, $\mathcal{T}(\gamma)$, and $\mathcal{T}(\delta)$ are vertex disjoint, and so χ is convex on $\mathcal{T}$.

Semple and Steel [8] showed that if $\mathcal{T}$ is a binary phylogenetic X-tree and $\mathcal{C}$ is a collection of r-state characters that defines $\mathcal{T}$, then

$$|\mathcal{C}| \geq \left\lceil \frac{n-3}{r-1} \right\rceil,$$

where $n = |X|$. In this paper, we show that, provided n is large enough, this lower bound on the size of $\mathcal{C}$ is sharp. In particular, we establish the following theorem.

THEOREM 1.1. *Let* r *be a positive integer exceeding one. Then there exists a positive integer* n_r *such that, for all binary phylogenetic* X*-trees* $\mathcal{T}$ *with* $n = |X| \geq n_r$*, there is a collection* $\mathcal{C}$ *of* r*-state characters of size*

$$|\mathcal{C}| = \left\lceil \frac{n-3}{r-1} \right\rceil$$

that defines $\mathcal{T}$*.*

When $r = 2$, Theorem 1.1 reduces to a result of Buneman [2]. Here $n_2 = 3$.

In addition to establishing Theorem 1.1 for all $r \geq 2$, we derive the exact result for when $r = 4$.

THEOREM 1.2. *Let* $\mathcal{T}$ *be a binary phylogenetic* X*-tree and let* $n = |X|$*. If* $n \geq 13$*, then there is a collection* $\mathcal{C}$ *of* 4*-state characters of size*

$$|\mathcal{C}| = \left\lceil \frac{n-3}{3} \right\rceil$$

that defines $\mathcal{T}$*. Moreover, if* $n = 12$*, then there is a binary phylogenetic* X*-tree that is not defined by any collection of* $3 = \lceil \frac{12-3}{3} \rceil$ *characters.*

Throughout the paper, notation and terminology follow Semple and Steel [10]. The paper is organized as follows. Section 2 contains some preliminaries. In section 3 we establish two lemmas, both of which are used in the proofs of the main results. Theorem 1.1 is proved in section 4, while Theorem 1.2 is proved in section 5.

Relevance of convexity to biology. Let $\mathcal{T}$ be a phylogenetic X-tree, and suppose that we subdivide an edge of $\mathcal{T}$ to create a degree-two vertex ρ. We call ρ the *root vertex* of $\mathcal{T}$, and refer to the resulting tree, denoted $\mathcal{T}^{+\rho}$, as a *rooted phylogenetic X-tree*.

Phylogenetic trees (and their rooted counterparts) provide a convenient representation of evolutionary relationships in biology. In particular, viewing the edges of $\mathcal{T}^{+\rho}$ as directed away from the root ρ, we regard $\mathcal{T}^{+\rho}$ as describing the evolution of the set X of extant species from an ancestral species at ρ. The remaining interior vertices of $\mathcal{T}^{+\rho}$ correspond to other hypothetical ancestral species descended from the species at ρ.

Now suppose each extant and ancestral species has an associated character state lying in some set C of character states. In this way, we regard the character state as also "evolving" from ρ towards the species in X. This leads to a concept of evolutionary "innovation," namely, that each time a species changes its character state, the new state it aquires is arising for the first time in the tree. Formalizing this concept, let c be the map from the vertices of $\mathcal{T}^{+\rho}$ into C so that $c(v)$ is equal to the character state assigned to vertex v. Then the "innovation" concept corresponds to the requirement that neither of the following two events occur, in which case we say that c is *homoplasy-free*.

(i) If $v_1v_2\cdots v_k$ is a path in $\mathcal{T}^{+\rho}$ directed away from the root ρ and, for some $i \in \{2, 3, \ldots, k-1\}$,

$$c(v_1) = c(v_k) \neq c(v_i),$$

then c exhibits a *reverse transition*. Informally, this corresponds to a new state arising, but then reverting to an earlier state.

(ii) If $v_1v_2\cdots v_k$ and $w_1w_2\cdots w_l$ are distinct directed paths in $\mathcal{T}^{+\rho}$ directed away from the root ρ with $v_1 = w_1$, $v_2 \neq w_2$, and

$$c(v_k) = c(w_l) \neq c(v_1),$$

then c exhibits a *convergent transition*. Informally, this corresponds to the same state arising in different parts of the tree.

Reverse and convergent transitions do happen in biology, but such events are considered relatively rare.

To explain the connection between these biologically motivated concepts and convexity, we use the following lemma, whose straightforward proof is omitted.

LEMMA 1.3. *Let χ be a character on X, taking values in a set C, and let $\mathcal{T}$ be a phylogenetic X-tree. Then χ is convex on $\mathcal{T}$ if and only if there is a function $\bar{\chi} : V(\mathcal{T}) \to C$ satisfying the following properties:*

(C1) *$\bar{\chi}|X = \chi$; and*

(C2) *if $\alpha \in C$, then the subgraph of $\mathcal{T}$ induced by $\{v \in V(\mathcal{T}) : \bar{\chi}(v) = \alpha\}$ is connected.*

Let $\mathcal{T}^{+\rho}$ be a rooted phylogenetic X-tree, and suppose that each vertex v of $\mathcal{T}^{+\rho}$ has an associated character state $c(v)$ that is an element of a set C of character states. Consider the associated phylogenetic X-tree $\mathcal{T}$. Restricting our attention to

the values that c takes at the leaves of $\mathcal{T}$, we obtain an induced character χ on X by setting $\chi(x) = c(x)$ for all $x \in X$. If c is homoplasy-free, then χ is convex on $\mathcal{T}$ since $\bar{\chi} : V(\mathcal{T}) \to C$ defined by $\bar{\chi}(u) = c(u)$, for all $u \in V(\mathcal{T})$, satisfies (C1) and (C2).

On the other hand, if a character χ_1 is convex on a phylogenetic X-tree $\mathcal{T}_1$ with a corresponding function $\bar{\chi_1} : V(\mathcal{T}_1) \to C$ satisfying (C1) and (C2), then, for all choices of a root ρ, we can extend $\bar{\chi_1}$ to a map from $V(\mathcal{T}_1) \cup \{\rho\}$ to C that is homoplasy-free.

Note that if c is not homoplasy-free on a rooted phylogenetic tree $\mathcal{T}^{+\rho}$, it is still possible that the associated character may be convex on $\mathcal{T}$.

2. Preliminaries. We begin by generalizing phylogenetic X-trees to X-trees. An *X-tree* $\mathcal{T} = (T; \phi)$ is an ordered pair consisting of a tree T and a mapping ϕ from X to the vertex set V of T with the property that if $v \in V$ and v has degree at most two, then $v \in \phi(X)$. Now, let χ be a character on X and let $\mathcal{T} = (T; \phi)$ be an X-tree. We denote the partition of X induced by χ by $\pi(\chi)$, that is,

$$\pi(\chi) = \{\chi^{-1}(\alpha_i) : \alpha_i \in \chi(X)\}.$$

Generalizing the notion of convexity to X-trees, we say χ is *convex* on $\mathcal{T}$ if, for all $\alpha \in \chi(X)$, the subtrees in $\{\mathcal{T}(\alpha) : \alpha \in \mathcal{C}\}$ are vertex disjoint, where $\mathcal{T}(\alpha)$ denotes the minimal subtree of $\mathcal{T}$ connecting the vertices in $\chi^{-1}(\alpha)$. An equivalent definition is that χ is convex on $\mathcal{T}$ if there is a subset F of edges of $\mathcal{T}$ whose deletion from $\mathcal{T}$ gives a graph with the property that, for all $A, B \in \pi(\chi)$, there are two components where $\phi(A)$ is a subset of the vertex set of one component and $\phi(B)$ is a subset of the vertex set of the other component. In this case, we say that χ *is displayed by* F. More generally, a collection $\mathcal{C}$ of characters on X is *convex* on $\mathcal{T}$ if each character in $\mathcal{C}$ is convex on $\mathcal{T}$, in which case, $\mathcal{C}$ is *compatible*. For a compatible collection $\mathcal{C}$ of characters on X, we say that $\mathcal{C}$ *infers* a character χ if χ is convex on every X-tree on which $\mathcal{C}$ is convex.

Let e be an edge of an X-tree $\mathcal{T} = (T; \phi)$, and let V_1 and V_2 be the vertex sets of the components of $T \backslash e$. Then the bipartition $\{\phi^{-1}(V_1), \phi^{-1}(V_2)\}$ of X, denoted σ_e, is an *X-split of* $\mathcal{T}$. If $\mathcal{T}$ is a phylogenetic tree and e is an interior edge, we refer to σ_e as a *nontrivial X-split of* $\mathcal{T}$. The following theorem is well known and is simply a rephrasing of the previously mentioned result of Buneman [2] in the language of X-splits.

THEOREM 2.1. *Let $\mathcal{T}$ be a binary phylogenetic X-tree. The collection Σ of nontrivial X-splits of $\mathcal{T}$ defines $\mathcal{T}$, that is, up to isomorphism, $\mathcal{T}$ is the only phylogenetic X-tree whose collection of X-splits contains Σ.*

An X-tree $\mathcal{T}'$ is a *refinement* of an X-tree $\mathcal{T}$ if each X-split of $\mathcal{T}$ is an X-split of $\mathcal{T}'$. A collection $\mathcal{C}$ of characters *identifies* an X-tree $\mathcal{T}$ if $\mathcal{C}$ is convex on $\mathcal{T}$ and every X-tree on which $\mathcal{C}$ is convex is a refinement of $\mathcal{T}$. Note that if $\mathcal{C}$ identifies a binary phylogenetic tree $\mathcal{T}$, then $\mathcal{C}$ defines $\mathcal{T}$. Furthermore, a compatible collection $\mathcal{C}$ of characters on X *infers* an X-split $\{X_1, X_2\}$ if $\{X_1, X_2\}$ is an X-split on every X-tree on which $\mathcal{C}$ is convex.

Distinguishing and strongly distinguishing. Let $\mathcal{T}$ be a binary phylogenetic X-tree and let $\mathcal{C}$ be a collection of characters on X. If $e = \{u_1, u_2\}$ is an interior edge of $\mathcal{T}$, then e is *distinguished* by $\mathcal{C}$ if there is a character χ in $\mathcal{C}$ with $A_1, A_2 \in \pi(\chi)$ such that u_1 but not u_2 is contained in the minimal subtree of $\mathcal{T}$ connecting elements in A_1, and u_2 but not u_1 is contained in the minimal subtree connecting elements in A_2. We say $\mathcal{T}$ is *distinguished* by $\mathcal{C}$ if every interior edge of $\mathcal{T}$ is distinguished by a character in $\mathcal{C}$. As we shall see below, the notion of distinguish has basic links with defining. However, for identifying, we need a stronger notion.

Let $\mathcal{T} = (T; \phi)$ be an X-tree and let $e = \{u_1, u_2\}$ be an edge of $\mathcal{T}$. Then e is *strongly distinguished* by a character χ on X if there exist $A_1, A_2 \in \pi(\chi)$ such that, for each $i \in \{1, 2\}$, the following hold:

(i) $\phi(A_i)$ is a subset of the vertex set of the component of $T\backslash e$ containing u_i;
(ii) the vertex set of each component of $T\backslash u_i$, except for the one containing the other end vertex of e, contains an element of $\phi(A_i)$; and
(iii) $\phi^{-1}(u_i)$ is a subset of A_i.

We say $\mathcal{T}$ is *strongly distinguished* by a collection $\mathcal{C}$ of characters if every edge of $\mathcal{T}$ is strongly distinguished by some character in $\mathcal{C}$.

Intersection graphs. Let $\mathcal{C}$ be a collection of characters on X' and let $\mathcal{T} = (T; \phi)$ be an X'-tree. If $X \subseteq X'$, the minimal subtree of $\mathcal{T}$ connecting the vertices in $\phi(X)$ is denoted by $\mathcal{T}(X)$. We next define two graphs each of which has vertex set

$$V(\mathcal{C}) = \bigcup_{\chi \in \mathcal{C}} \{(\chi, A) : A \in \pi(\chi)\}.$$

(I) The *partition intersection graph of* $\mathcal{C}$, denoted $\text{int}(\mathcal{C})$, is the graph with vertex set $V(\mathcal{C})$ and an edge joining (χ_1, A) and (χ_2, B) if $A \cap B$ is nonempty.
(II) The *subtree intersection graph of* $\mathcal{T}$ *induced by* $\mathcal{C}$, denoted $\text{int}(\mathcal{C}, \mathcal{T})$, is the graph with vertex set $V(\mathcal{C})$ and an edge joining (χ_1, A) and (χ_2, B) if $\chi_1 \neq \chi_2$ and $\mathcal{T}(A) \cap \mathcal{T}(B)$ is nonempty.

It is well known that if a graph G is the subtree intersection graph of subtrees of a tree, then G is chordal [3, 4, 11]. Thus, in (II), the intersection graph $\text{int}(\mathcal{C}, \mathcal{T})$ is chordal.

A graph is *chordal* if every cycle with at least four vertices has an edge connecting two nonconsecutive vertices. For a collection $\mathcal{C}$ of characters on X, a chordal graph G is a *restricted chordal completion*, also called a *proper triangulation*, of $\text{int}(\mathcal{C})$ if G can be obtained from $\text{int}(\mathcal{C})$ by adding only edges joining vertices whose first components are distinct. If G is a restricted chordal completion of $\text{int}(\mathcal{C})$, but there is no restricted chordal completion G' of $\text{int}(\mathcal{C})$ in which $E(G)$ is a proper subset of $E(G')$, we say that G has *no nontrivial restricted chordal completions*.

Past results. A number of results have been established equating the compatibility of a collection $\mathcal{C}$ of characters with the intersection graph $\text{int}(\mathcal{C})$ of $\mathcal{C}$. For example, Buneman [3] showed that $\mathcal{C}$ is compatible if and only if $\text{int}(\mathcal{C})$ has a restricted chordal completion. Furthermore, if $\mathcal{C}$ is a collection of 2-state characters, then $\mathcal{C}$ is compatible if and only if $\text{int}(\mathcal{C})$ is chordal [10], and more recently Lam, Gusfield, and Sridhar [7] have characterized the compatibility of a collection of 3-state characters $\mathcal{C}$ in terms of $\text{int}(\mathcal{C})$. The intersection graph of a collection of characters, and restricted chordal completions of it, have also been used in algorithmic approaches to determining compatibility; see, for example, Gysel and Gusfield [5].

The following results will be used to prove Theorem 1.1 and 1.2. The first result is one direction of the main result in [9]. For a collection $\mathcal{C}$ of characters on X, a restricted chordal completion G of $\text{int}(\mathcal{C})$ is *minimal* if, for every nonempty subset F of $E(G) - E(\text{int}(\mathcal{C}))$, the graph $G\backslash F$ is not chordal.

THEOREM 2.2. *Let $\mathcal{T}$ be a binary phylogenetic X-tree and let $\mathcal{C}$ be a collection of characters on X. Then $\mathcal{C}$ defines $\mathcal{T}$ if*

(i) *each character is convex on $\mathcal{T}$ and every edge of $\mathcal{T}$ is distinguished by a character in $\mathcal{C}$, and*
(ii) *there is a unique minimal restricted chordal completion of* $\text{int}(\mathcal{C})$.

For a collection $\mathcal{C}$ of characters on X, let $\mathcal{G}(\mathcal{C})$ denote the set

$$\mathcal{G}(\mathcal{C}) = \{G : \text{there is an } X\text{-tree } \mathcal{T} \text{ in which } \mathcal{C} \text{ is convex and } G = \text{int}(\mathcal{C}, \mathcal{T})\}$$

of chordal graphs. Note that $\mathcal{G}(\mathcal{C})$ is a subset of the collection of restricted chordal completions of $\text{int}(\mathcal{C})$. We obtain a partial order $\leq$ on $\mathcal{G}(\mathcal{C})$ by setting $G_1 \leq G_2$ if $E(G_1) \subseteq E(G_2)$ for all $G_1, G_2 \in \mathcal{G}(\mathcal{C})$. The next theorem is one direction of the main result in [1].

THEOREM 2.3. *Let $\mathcal{T}$ be an X-tree and let $\mathcal{C}$ be a collection of characters on X. Then $\mathcal{C}$ identifies $\mathcal{T}$ if*

(i) *each character in $\mathcal{C}$ is convex on $\mathcal{T}$ and every edge of $\mathcal{T}$ is strongly distinguished by a character in $\mathcal{C}$, and*
(ii) *there is a unique maximal element in $\mathcal{G}(\mathcal{C})$.*

In reference to Theorem 2.3, if $\text{int}(\mathcal{C})$ has no nontrivial restricted chordal completions, then there is precisely one graph in $\mathcal{G}(\mathcal{C})$, namely, $\text{int}(\mathcal{C})$, in which case, this is the unique maximal element in $\mathcal{G}(\mathcal{C})$. In particular, we have the following corollary, of which we make frequent use.

COROLLARY 2.4. *Let $\mathcal{T}$ be an X-tree and let $\mathcal{C}$ be a collection of characters on X. Then $\mathcal{C}$ identifies $\mathcal{T}$ if*

(i) *each character in $\mathcal{C}$ is convex on $\mathcal{T}$ and every edge of $\mathcal{T}$ is strongly distinguished by a character in $\mathcal{C}$, and*
(ii) $\text{int}(\mathcal{C})$ *has no nontrivial restricted chordal completions.*

3. Two lemmas. In this section, we prove two lemmas. Both lemmas are used in the proofs of Theorems 1.1 and 1.2.

Let $\mathcal{T} = (T; \phi)$ be an X-tree, let $e = \{u, v\}$ be an edge of $\mathcal{T}$, and let $\mathcal{T}/e$ be the X-tree obtained from $\mathcal{T}$ by *contracting* e, that is, letting w denote the identified vertex in T/e, the X-tree $(T/e; \phi')$, where, for all $y \in X$,

$$\phi'(y) = \begin{cases} \phi(y) & \text{if } \phi(y) \notin \{u, v\}, \\ w & \text{otherwise.} \end{cases}$$

Furthermore, let F be a subset of edges of T. Let $V_1, V_2, \ldots, V_k$ denote the vertex sets of the components of $T \backslash F$. The *partition of X displayed by F* is the partition

$$\{\phi^{-1}(V_i) : i \in \{1, 2, \ldots, k\}\}.$$

LEMMA 3.1. *Let $r \geq 2$ and let $\mathcal{T}$ be a phylogenetic X-tree. Furthermore suppose that $\mathcal{T}$ has a path containing (in order) $2r-2$ interior edges $e_1, e_2, \ldots, e_{2r-2}$. Let $\{X_1, X_2, \ldots, X_{2r-1}\}$ be the partition of X displayed by $E' = \{e_1, e_2, \ldots, e_{2r-2}\}$, where, for all $i \in \{1, 2, \ldots, 2r-2\}$, the edge e_i is the only edge in E' in the minimal subtree of $\mathcal{T}$ connecting the elements in $X_i \cup X_{i+1}$. Then any two r-state characters χ_1 and χ_2 with*

$$\pi(\chi_1) = \{X_1, X_2 \cup X_3, X_4 \cup X_5, \ldots, X_{2r-2} \cup X_{2r-1}\}$$

and

$$\pi(\chi_2) = \{X_1 \cup X_2, X_3 \cup X_4, \ldots, X_{2r-3} \cup X_{2r-2}, X_{2r-1}\}$$

infer the X-splits $\sigma_{e_1}, \sigma_{e_2}, \ldots, \sigma_{e_{2r-2}}$.

Proof. Let $\mathcal{T}'$ be the X-tree obtained from $\mathcal{T}$ by contracting each of the edges in $E(\mathcal{T}) - \{e_1, e_2, \ldots, e_{2r-2}\}$. Let e_i be an edge of $\mathcal{T}'$. First suppose e_i is not a pendant edge of $\mathcal{T}'$. Then, for some $j \in \{1, 2\}$, there is a character χ_j such that $X_{i-1} \cup X_i \in \pi(\chi_j)$ and $X_{i+1} \cup X_{i+2} \in \pi(\chi_j)$. Using $X_{i-1} \cup X_i$ and $X_{i+1} \cup X_{i+2}$, it is easily checked that χ_j strongly distinguishes e_i. Now suppose that e_i is a pendant edge of $\mathcal{T}'$. Then e_i is either e_1 or e_{2r-2}. If it is e_1, then χ_1 strongly distinguishes e_i using X_1 and $X_2 \cup X_3$, while if it is e_{2r-2}, then χ_2 strongly distinguishes e_i using $X_{2r-3} \cup X_{2r-2}$ and X_{2r-1}. Now consider the partition intersection graph $\mathrm{int}(\{\chi_1, \chi_2\})$. A routine check shows that $\mathrm{int}(\{\chi_1, \chi_2\})$ is a path in which every second vertex has the same first coordinate. It follows that $\mathrm{int}(\{\chi_1, \chi_2\})$ has no nontrivial restricted chordal completions and so, by Corollary 2.4, χ_1 and χ_2 identify $\mathcal{T}'$. In particular, χ_1 and χ_2 infer the X-splits $\sigma_{e_1}, \sigma_{e_2}, \ldots, \sigma_{e_{2r-2}}$. □

Let e be an edge of an X-tree $\mathcal{T} = (T; \phi)$, and let V_1 and V_2 be the vertex sets of the components of $T \backslash e$. Let χ_e denote the character $\chi_e : X \to \{\alpha_e, \beta_e\}$ defined, for all $y \in X$, by

$$\chi_e(y) = \begin{cases} \alpha_e & \text{if } y \in \phi^{-1}(V_1), \\ \beta_e & \text{otherwise.} \end{cases}$$

LEMMA 3.2. *Let $\mathcal{T} = (T; \phi)$ be an X-tree and let χ be a character on X that is convex on $\mathcal{T}$, where $\pi(\chi) = \{Y_1, Y_2, \ldots, Y_r\}$, where $r \geq 2$. Let $\{f_1, f_2, \ldots, f_{r-1}\}$ be a set of edges that displays χ. Let $E' = \{e_1, e_2, \ldots, e_s\}$ be a subset of edges of $\mathcal{T}$ with $E' \cap \{f_1, f_2, \ldots, f_{r-1}\}$ empty satisfying the following two properties:*

(i) *for all distinct $i, j \in \{1, 2, \ldots, r-1\}$, there is an interior edge $e \in E'$ on the path from an end vertex of f_i to an end vertex of f_j; and*

(ii) *for each $e = \{u, v\} \in E'$, there is a path from u (resp., v) to a vertex w of $\mathcal{T}$ avoiding v (resp., u) and $f_1, f_2, \ldots, f_{r-1}$, and $\phi^{-1}(w)$ is nonempty.*

Then the collection

$$\{\chi, \chi_{e_1}, \chi_{e_2}, \ldots, \chi_{e_s}\}$$

of characters on X infers each of the X-splits $\sigma_{f_1}, \sigma_{f_2}, \ldots, \sigma_{f_{r-1}}$.

Proof. Let $\mathcal{T}' = (T; \phi')$ be the X-tree obtained from $\mathcal{T}$ by contracting each edge not in

$$E' \cup \{f_1, f_2, \ldots, f_{r-1}\},$$

and consider the collection

$$\mathcal{C} = \{\chi, \chi_{e_1}, \chi_{e_2}, \ldots, \chi_{e_s}\}$$

of characters on X. We next show that $\mathcal{C}$ identifies $\mathcal{T}'$.

Clearly, $\mathcal{C}$ is convex on $\mathcal{T}'$. Now let e be an edge of $\mathcal{T}'$. If $e \in E'$, then χ_e strongly distinguishes e. Suppose that $e = f_i = \{u_1, u_2\}$ for some $i \in \{1, 2, \ldots, r-1\}$. Since E' satisfies (i), every edge in $\mathcal{T}'$ adjacent to f_i is in E'. Together with (ii), this implies that there exist $Z_1, Z_2 \in \pi(\chi)$ such that, for each $i \in \{1, 2\}$, we have the following: $\phi'(Z_i)$ is a subset of the vertex set of the component of $\mathcal{T}' \backslash f_i$ containing u_i; the vertex set of each component of $\mathcal{T}' \backslash u_i$ (except for the one containing the other end vertex of f_i) contains an element of $\phi'(Z_i)$; and $\phi^{-1}(u_i)$ is a subset of Z_i. Thus χ strongly distinguishes f_i.

Consider $\mathrm{int}(\mathcal{C})$. We will show that $\mathrm{int}(\mathcal{C})$ has no nontrivial restricted chordal completions. Let e and e' be distinct elements in E', and let $\pi(\chi_e) = \{A, B\}$ and

$\pi(\chi_{e'}) = \{A', B'\}$. Now exactly one of the intersections $A \cap A'$, $A \cap B'$, $B \cap A'$, and $B \cap B'$ is empty. Without loss of generality, we may assume that $A \cap A'$ is empty. Then $\{(\chi_e, A), (\chi_{e'}, A')\}$ is not an edge in $\text{int}(\mathcal{C})$. Indeed, no nontrivial restricted chordal completion contains this edge; otherwise $(\chi_e, A), (\chi_{e'}, A'), (\chi_e, B), (\chi_{e'}, B')$ are the vertices of a 4-cycle in such a completion. But it is not possible for two nonconsecutive vertices in this cycle to be joined by an edge as they have the same first component. Hence any nontrivial restricted chordal completion of $\text{int}(\mathcal{C})$ must contain an edge of the form $\{(\chi, Y_i), (\chi_{e_j}, A_j)\}$, where $Y_i \cap A_j$ is empty. Suppose there exists such a completion G containing such an edge. Since E' satisfies (i) and (ii), there is a $Y_k \in \pi(\chi)$ with $Y_k \cap A_j$ and $Y_k \cap B_j$ both nonempty, where $\pi(\chi_{e_j}) = \{A_j, B_j\}$. Note that, as $Y_k \cap A_j$ is nonempty, Y_i and Y_k are distinct. Furthermore, as $Y_i \cap A_j$ is empty, $Y_i \cap B_j$ is nonempty. Therefore $(\chi, Y_i), (\chi_{e_j}, A_j), (\chi, Y_k), (\chi_{e_j}, B_j)$ are the vertices of a 4-cycle of G. But again it is not possible for two nonconsecutive vertices in this cycle to be joined by an edge; a contradiction. Thus there is no nontrivial restricted chordal completion of $\text{int}(\mathcal{C})$. Therefore, by Corollary 2.4, $\mathcal{C}$ identifies $\mathcal{T}'$, which in turn implies that every X-tree on which $\mathcal{C}$ is convex is a refinement of $\mathcal{T}'$. In particular, $\sigma_{f_1}, \sigma_{f_2}, \ldots, \sigma_{f_{r-1}}$ are X-splits of such an X-tree. This completes the proof of the lemma. □

4. Proof of Theorem 1.1. This section consists of the proof of Theorem 1.1.

Proof of Theorem 1.1. We assume that $|X| > n_0$, where n_0 will be chosen sufficiently large so that we can choose our initial characters without worrying about the topology of $\mathcal{T}$. We select a set E_0 of interior edges $e_1, e_2, \ldots, e_k$ on a path in $\mathcal{T}$ with the properties that

(i) k is a multiple of $2r-2$ and is at least $4r^5 + r$, and

(ii) each of the components of $\mathcal{T} \backslash E_0$ contains at least $\lceil \log r \rceil$ vertices.

By applying Lemma 3.1 k times, there is a set $\mathcal{C}_0$ of $\frac{k}{r-1}$ r-state characters which infers the X-splits $\sigma_{e_1}, \sigma_{e_2}, \ldots, \sigma_{e_k}$.

We now proceed iteratively to construct the remaining characters. Set $i = 1$. In step i, select a set F_i of $r-1$ interior edges of $\mathcal{T}$ such that each edge in F_i is in a different component of $\mathcal{T} \backslash E_{i-1}$. Define χ_i to be a character for which $\pi(\chi_i)$ is the partition of X displayed by F_i. Let $E_i = E_{i-1} \cup F_i$. By induction and Lemma 3.2, the collection $\mathcal{C}_i = \mathcal{C}_{i-1} \cup \{\chi_i\}$ infers each of the X-splits in $\{\sigma_e : e \in E_i\}$. Now, increase i by 1 and repeat.

In order to ensure that we do not exhaust the supply of edges in distinct components too early, we shall always select the edges in F_i to be in the $r-1$ largest components (measured by the number of interior edges of $\mathcal{T}$) of $\mathcal{T} \backslash E_{i-1}$, and to select the interior edge within each such component that as closely as possible results in two subsequent components of equal size on its deletion. By doing so, we will show that we can continue the process until the final step, l say, in which there may be less than $r-1$ interior edges of $\mathcal{T}$ not yet selected. However, these edges will be in distinct components of $\mathcal{T} \backslash E_{l-1}$ and we generate the final character χ_l, for which $\pi(\chi_l)$ is the partition of X displayed by the remaining interior edges of $\mathcal{T}$ not yet selected. We end with a set $\mathcal{C}_l$ of r-state characters that infers every nontrivial X-split of $\mathcal{T}$ and therefore, by Theorem 2.1, defines $\mathcal{T}$, where $l = \lceil \frac{n-3}{r-1} \rceil$.

Let $i \geq 1$. Our first claim is that if a component $\mathcal{T}'$ of $\mathcal{T} \backslash E_{i-1}$ has $m \geq 4$ interior edges of $\mathcal{T}$, then we can select such an interior edge e of $\mathcal{T}'$ for which each of the two components, $\mathcal{T}_1'$ and $\mathcal{T}_2'$ say, of $\mathcal{T}' \backslash e$ has greater than $m/10$ interior edges of $\mathcal{T}$, and with one component having at least $m/3$ interior edges of $\mathcal{T}$. To see this, choose e that maximizes the minimum number of interior edges of $\mathcal{T}_1'$ and $\mathcal{T}_2'$. For the purposes of

obtaining a contradiction, we may assume, without loss of generality, $\mathcal{T}_1'$ has at most $m/10$ interior edges of $\mathcal{T}$, and so $\mathcal{T}_2'$ has at least $9m/10-1$ interior edges of $\mathcal{T}$. Now $\mathcal{T}_2'$ has at least one interior edge of $\mathcal{T}$ adjacent to e. If $\mathcal{T}_2'$ has exactly one such edge f, then the components of $\mathcal{T}'\backslash f$ each have (strictly) more interior edges of $\mathcal{T}$ than $\mathcal{T}_1'$, contradicting the choice of e. Therefore, $\mathcal{T}_2'$ has two such edges, f_1 and f_2 say. Let $f \in \{f_1, f_2\}$ be the edge such that the component of $\mathcal{T}_2'\backslash\{f_1, f_2\}$ pendant to f has the most interior edges of $\mathcal{T}$. This component must have at least $\lceil 9m/20 - 3/2\rceil$ interior edges of $\mathcal{T}$ which is greater than $m/10$ for $m \geq 4$. Again, the components of $\mathcal{T}'\backslash f$ each have (strictly) more interior edges of $\mathcal{T}$ than $\mathcal{T}_1'$, contradicting the choice of e. For the second part of the claim, it follows by the pigeonhole principle that at least one of the two components must have at least $(m-1)/2 > m/3$ interior edges of $\mathcal{T}$.

Our second claim is that if we have performed t iterations and the largest remaining component has

$$m_t \geq 4 \cdot 9^{\lceil \log r \rceil}$$

interior edges of $\mathcal{T}$, then, for all $k \leq \min\{2^t, r-1\}$, the kth largest component (measured by the number of interior edges of $\mathcal{T}$) has at least

$$\frac{m_t}{9^{\lceil \log k \rceil}}$$

interior edges of $\mathcal{T}$. We prove this claim by induction on t. For $t = 0$, the claim trivially holds since $k \leq 1$ and the largest component has at least m_0 interior edges of $\mathcal{T}$. Now suppose that the claim holds for the first t iterations. The largest component after $t+1$ iterations arose either from a component that was divided in the $(t+1)$th iteration, in which case, by the first claim, $m_{t+1} < 9m_t/10$, or was not divided in the $(t+1)$th iteration, in which case, it was at most the rth largest component at the start of that iteration. In the first case, for $k \leq \min\{2^{t+1}, r-1\}$, there were at least $k/2$ components of size at least

$$\frac{m_t}{9^{\lceil \log(k/2) \rceil}} = \frac{m_t}{9^{\lceil \log k \rceil - 1}}$$

at the start of the $(t+1)$th iteration. By the first claim, each of these components generate two components of size at least

$$\frac{m_t}{10 \cdot 9^{\lceil \log k \rceil - 1}} > \frac{m_{t+1}}{9^{\lceil \log k \rceil}},$$

and so the inductive step holds. In the second case, for $k \leq \min\{2^{t+1}, r-1\}$, there were at least k components of size at least m_{t+1} at the start of the $(t+1)$th iteration and, by the first claim, each of these generated a component with more than $m_{t+1}/3$ interior edges of $\mathcal{T}$. For $k \geq 2$, this is at least

$$\frac{m_{t+1}}{9^{\lceil \log k \rceil}}.$$

Furthermore, for $k = 1$, the claim holds trivially. This completes the proof of the second claim.

We now fill in the argument outlined earlier. There are at least $4r^5 + r > (r-1)\lceil \log r \rceil$ components in $\mathcal{T}\backslash E_0$ with at least $\lceil \log r \rceil$ interior edges of $\mathcal{T}$. Hence, we do the first $\lceil \log r \rceil$ iterations without ever selecting an edge in a component which

only has one interior edge of $\mathcal{T}$. The endgame is reached when we first have to select an edge from a component consisting of only one interior edge of $\mathcal{T}$. From here, the number of components starts decreasing. At this point, the number of components is still at least $4r^5 + r$. Since the $(r-1)$th largest component has 1 interior edge of $\mathcal{T}$, it follows by the second claim that

$$m_t < 4 \cdot 9^{\lceil \log r \rceil} < 4r^4.$$

Furthermore, as the $(r-1)$th largest component has 1 interior edge of $\mathcal{T}$, there are at least $4r^5$ components that have exactly one interior edge of $\mathcal{T}$. After at most $4r^4$ further iterations, there will be no components with more than 1 interior edge of $\mathcal{T}$ remaining and, for each of these iterations, we will have always selected a set of $r-1$ interior edges of $\mathcal{T}$ from distinct components since there are sufficient components consisting of at least one such edge. We can now select the remaining interior edges of $\mathcal{T}$ in sets of $r-1$ until we are left with the final set of size at most $r-1$. Because we always selected a full set of $r-1$ interior edges of $\mathcal{T}$, apart from possibly the final set, and there are $n-3$ interior edges in total, our resulting collection of characters has size $\lceil \frac{n-3}{r-1} \rceil$. This completes the proof of the theorem. □

5. Four-state characters. In this section, we prove Theorem 1.2. We begin with an example to show that when $n = 12$ there is a binary phylogenetic X-tree that is not defined by three 4-state characters.

Consider the binary phylogenetic tree $\mathcal{T}$ shown in Figure 1(a). Suppose for contradiction that three 4-state characters define $\mathcal{T}$. Since these characters define $\mathcal{T}$, each interior edge of $\mathcal{T}$ must be distinguished by one of the characters, and so no character is displayed by a subset of edges containing two adjacent edges. Thus we may assume by symmetry that these characters are displayed, respectively, by the subsets $\{e_1, e_4, e_7\}$, $\{e_2, e_5, e_8\}$, and $\{e_3, e_6, e_9\}$ of edges of $\mathcal{T}$. In particular, the partitions of $\{1, 2, \ldots, 12\}$, namely,

$$\big\{\{1,2,3,4\},\{5,6\},\{7,8,11,12\},\{9,10\}\big\},$$
$$\big\{\{1,2\},\{3,4,9,10\},\{5,6,7,8\},\{11,12\}\big\},$$

and

$$\big\{\{1,2,5,6\},\{3,4\},\{7,8\},\{9,10,11,12\}\big\}$$

induced by the characters define $\mathcal{T}$. But the same collection of characters is also convex on the binary phylogenetic tree shown in Figure 1(b); a contradiction. Thus, for $n = 12$, Theorem 1.2 does not hold.

The rest of the proof of Theorem 1.2 is by induction on n. For this induction, it is the base case that requires the most work. The base case consists of directly establishing the result for $n = 13, 14$ (Corollary 5.9) and $n = 15$ (Corollary 5.8). We begin with several lemmas which will eventually be used to establish these corollaries. Once the base case is established, Lemma 5.10 deals with binary phylogenetic trees of a special structure (*caterpillar-like* trees) and, in all remaining binary phylogenetic trees, we identify three leaves that may be removed to give an easy inductive step.

LEMMA 5.1. *Let $\mathcal{C}$ be a compatible collection $\{\chi_1, \chi_2, \ldots, \chi_k\}$ of characters, and suppose that G is a restricted chordal completion of* $\mathrm{int}(\mathcal{C})$. *If, for each i and j, the subgraph of G induced by those vertices whose first coordinate is either χ_i or χ_j is a tree, then there is no nontrivial restricted chordal completion of G.*

Proof. Suppose that G has the desired tree property for each pair of characters in $\mathcal{C}$, but there is a nontrivial restricted chordal completion G' of G. Let e be an edge of G' that is not an edge of G. Then e joins two vertices v_i and v_j whose first coordinates are distinct, say χ_i and χ_j. By the tree property in G, there is a path in G' from v_i to v_j avoiding e whose first coordinates alternate between χ_i and χ_j. In particular, there is a cycle in G' with at least four vertices whose first coordinates alternate between χ_i and χ_j. But then, for each four successive vertices in the cycle, there is no edge connecting two nonconsecutive vertices, so G' is not chordal. This contradiction completes the proof of the lemma. □

Let $\mathcal{T}'$ be a phylogenetic X'-tree and let $\mathcal{T}$ be a phylogenetic X-tree with $X \subseteq X'$. The next lemma will show that a set of characters that defines $\mathcal{T}$ can be extended to a set of characters on X' that infer the same structure. We say $\mathcal{T}$ is a *restriction* of $\mathcal{T}'$ if $\mathcal{T}$ can be obtained from the minimal subtree of $\mathcal{T}'$ connecting the elements in X by suppressing degree-two vertices. A bipartition $\{A', B'\}$ *extends* another bipartition $\{A, B\}$ if, for some choice of A and B, we have $A \subseteq A'$ and $B \subseteq B'$. Observe that $\mathcal{T}$ is a restriction of $\mathcal{T}'$ if and only if, for all interior edges e in $\mathcal{T}$, there is an edge f in $\mathcal{T}'$ such that the bipartition corresponding to f in $\mathcal{T}'$ extends that corresponding to e in $\mathcal{T}$.

Suppose that $\mathcal{T}$ is a restriction of $\mathcal{T}'$. A subset F of interior edges of $\mathcal{T}'$ is $\mathcal{T}$-*representable* if, for each interior edge e in $\mathcal{T}$, there is precisely one edge f in F such that σ_f extends σ_e. Now let $\mathcal{C}$ be a collection of characters that are convex on $\mathcal{T}$. Let χ be a character in $\mathcal{C}$, and suppose that χ is displayed by the subset E_χ of edges in $\mathcal{T}$. For each $\chi \in \mathcal{C}$, let χ_F be a character on X' displayed in $\mathcal{T}'$ by the subset

$$\{f : \sigma_f \text{ extends } \sigma_e,\ e \in E_\chi,\ f \in F\}$$

of F. Furthermore, let $\mathcal{C}_F = \{\chi_F : \chi \in \mathcal{C}\}$.

LEMMA 5.2. *Let $\mathcal{T}'$ be a phylogenetic X'-tree and let $\mathcal{T}$ be a binary phylogenetic X-tree with $X \subseteq X'$. Suppose $\mathcal{T}$ is a restriction of $\mathcal{T}'$. Let F be a $\mathcal{T}$-representable subset of interior edges of $\mathcal{T}'$. If $\mathcal{C}$ is a collection of characters on X that defines $\mathcal{T}$, then the collection $\mathcal{C}_F$ of characters on X' infers each of the X'-splits of $\mathcal{T}'$ induced by the edges in F.*

Proof. Let $\mathcal{S}$ be the phylogenetic tree obtained from $\mathcal{T}'$ by contracting each of its interior edges not in F. We will show that $\mathcal{C}_F$ identifies $\mathcal{S}$, thus showing that $\mathcal{C}_F$ infers each of the X'-splits of $\mathcal{T}'$ induced by the edges in F. Suppose that there is a phylogenetic X'-tree $\mathcal{T}_1$ that is not a refinement of $\mathcal{S}$ but $\mathcal{C}_F$ is convex on $\mathcal{T}_1$. Since $\mathcal{C}_F$ is convex on $\mathcal{T}_1$, and so $\mathcal{T}_1|X \cong \mathcal{T} \cong \mathcal{S}|X$, it follows that there is an element z in $X' - X$ such that $\{A \cup \{z\}, B\}$ is an $(X \cup \{z\})$-split of $\mathcal{S}|(X \cup \{z\})$ but $\{A, B \cup \{z\})$ is an $(X \cup \{z\})$-split of $\mathcal{T}_1|(X \cup \{z\})$. Here X is the disjoint union of A and B. Now, consider $\mathcal{S}|(X \cup \{z\})$, and let e be the edge of $\mathcal{S}|(X \cup \{z\})$ such that $\sigma_e = \{A \cup \{z\}, B\}$. Since $\mathcal{C}$ defines $\mathcal{T}$, there is a character χ in $\mathcal{C}$ with $A_1, B_1 \in \pi(\chi)$, where $A_1 \subseteq A$ and $B_1 \subseteq B$, and $a_1, a_2 \in A_1$ and $b_1, b_2 \in B_1$ such that the path in $\mathcal{S}|(X \cup \{z\})$ from a_1 to a_2 passes through one end vertex of e, while the path in $\mathcal{S}|(X \cup \{z\})$ from b_1 to b_2 passes through the other end vertex of e. Thus the character χ_F in $\mathcal{C}_F$ corresponding to χ has the property that there are parts $A_F, B_F \in \pi(\chi_F)$ with $A_1 \cup \{z\} \subseteq A_F$ and $B_1 \subseteq B_F$. But then, as $\mathcal{S}|X$ is isomorphic to $\mathcal{T}_1|X$, it follows that χ_F is not convex on $\mathcal{T}_1|(X \cup \{z\})$, and therefore not convex on $\mathcal{T}_1$. It follows from this contradiction that $\mathcal{C}_F$ identifies $\mathcal{S}$. □

An *internal pseudopath of length t* in a phylogenetic tree $\mathcal{T}$ is a set of t interior edges that lie on a path in $\mathcal{T}$. In particular, the interior edges need not be consecutive.

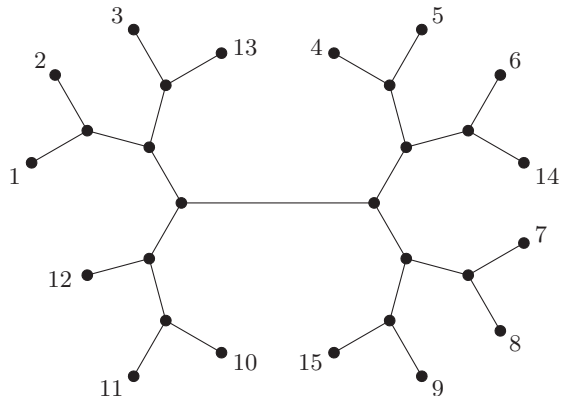

FIG. 2. *The unique binary phylogenetic tree on* $\{1, 2, \ldots, 15\}$ *with no internal pseudopath of length* 6.

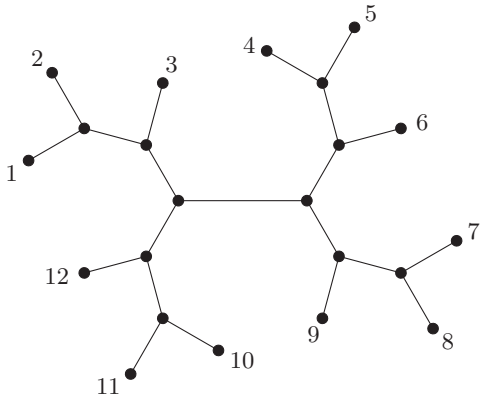

FIG. 3. *A binary phylogenetic tree on* $\{1, 2, \ldots, 12\}$.

LEMMA 5.3. *Let* $\mathcal{T}$ *be a phylogenetic* X*-tree with* $X = \{1, 2, \ldots, 15\}$ *and no internal pseudopath of length* 6. *Then there is a collection of four* 4*-state characters that defines* $\mathcal{T}$.

Proof. By attempting to construct such a tree, it is easily checked that there is no binary phylogenetic tree with 15 leaves whose maximum length internal path is at most 4. Furthermore, another check shows that, up to isomorphism, the binary phylogenetic tree, $\mathcal{T}'$ say, shown in Figure 2 is the only binary phylogenetic tree on $\{1, 2, \ldots, 15\}$ whose maximum length internal path is 5. Consider the phylogenetic tree $\mathcal{T}$ on $\{1, 2, \ldots, 12\}$ shown in Figure 3 and a collection $\mathcal{C} = \{\chi_1, \chi_2, \chi_3\}$ of 4-state characters on $\{1, 2, \ldots, 12\}$, where

$$\pi(\chi_1) = \big\{\{1, 2\}, \{3, 12\}, \{4, 5, 6, 7, 8, 9\}, \{10, 11\}\big\},$$
$$\pi(\chi_2) = \big\{\{1, 2, 3\}, \{4, 5\}, \{6, 10, 11, 12\}, \{7, 8, 9\}\big\},$$

and

$$\pi(\chi_3) = \big\{\{1, 2, 3, 9\}, \{4, 5, 6\}, \{7, 8\}, \{10, 11, 12\}\big\}.$$

Observe that $\mathcal{C}$ is convex on $\mathcal{T}$. The intersection graph $\mathrm{int}(\mathcal{C})$ is not chordal but, as $(\chi_1, \{3, 12\}), (\chi_3, \{1, 2, 3, 9\}), (\chi_1, \{4, 5, 6, 7, 8, 9\}), (\chi_2, \{6, 10, 11, 12\})$ is a cycle in

$\text{int}(\mathcal{C})$, every restricted choral completion of $\text{int}(\mathcal{C})$ must include an edge joining $(\chi_2, \{6, 10, 11, 12\})$ and $(\chi_3, \{1, 2, 3, 9\})$. Let G denote $\text{int}(\mathcal{C})$ with this additional edge. Now G is the intersection graph $\text{int}(\mathcal{C}, \mathcal{T})$ and so it is chordal. Thus G is the unique minimal restricted chordal completion of $\text{int}(\mathcal{C})$. Hence, as $\mathcal{T}$ is distinguished by $\mathcal{C}$, it follows by Theorem 2.2 that $\mathcal{C}$ defines $\mathcal{T}$. Since $\mathcal{T}$ is a restriction of $\mathcal{T}'$, it follows by Lemma 5.2 that any collection F of $\mathcal{T}$-representable edges infers the X-splits of $\mathcal{T}'$ induced by the edges of F. It now follows by Lemma 3.2 that a collection of four 4-state characters $\{\chi_1', \chi_2', \chi_3', \chi_4'\}$ on $\{1, 2, \ldots, 15\}$, where

$$\pi(\chi_1') = \big\{\{1, 2\}, \{3, 12, 13\}, \{4, 5, 6, 7, 8, 9, 14, 15\}, \{10, 11\}\big\},$$
$$\pi(\chi_2') = \big\{\{1, 2, 3, 13\}, \{4, 5\}, \{6, 10, 11, 12, 14\}, \{7, 8, 9, 15\}\big\},$$
$$\pi(\chi_3') = \big\{\{1, 2, 3, 9, 13, 15\}, \{4, 5, 6, 14\}, \{7, 8\}, \{10, 11, 12\}\big\},$$

and

$$\pi(\chi_4') = \big\{\{1, 2, 4, 5, 7, 8, 10, 11, 12\}, \{3, 13\}, \{6, 14\}, \{9, 15\}\big\},$$

defines $\mathcal{T}'$. This completes the proof of the lemma. □

Let F be a subset of edges of a phylogenetic tree $\mathcal{T}$ and let $C_1, C_2, \ldots, C_k$ denote the components of $\mathcal{T} \backslash F$. For all i, let E_i denote the set of interior edges of $\mathcal{T}$ in C_i. Note that, for some i, the set E_i may be empty. We say that F *separates* the interior edges of $\mathcal{T}$ not in F into sets $E_1, E_2, \ldots, E_k$.

LEMMA 5.4. *Let $\mathcal{T}$ be a binary phylogenetic X-tree with* 15 *leaves and an internal pseudopath P of length* 6. *If P separates the interior edges of $\mathcal{T}$ not in P into sets of size at most two, then there is a collection of four* 4*-state characters that defines $\mathcal{T}$.*

Proof. Suppose that P separates the interior edges of $\mathcal{T}$ not in P into sets of size at most two. In order, let $e_1, e_2, \ldots, e_6$ denote the edges of P. Let $\{X_1, X_2, \ldots, X_7\}$ be the partition of X displayed by $\{e_1, e_2, \ldots, e_6\}$, where, for all $i \in \{1, 2, \ldots, 6\}$, the edge e_i is the only edge in the set in the minimal subtree of $\mathcal{T}$ connecting the elements in $X_i \cup X_{i+1}$. By Lemma 3.1, any two 4-state characters χ_1 and χ_2 with

$$\pi(\chi_1) = \{X_1, X_2 \cup X_3, X_4 \cup X_5, X_6 \cup X_7\}$$

and

$$\pi(\chi_2) = \{X_1 \cup X_2, X_3 \cup X_4, X_5 \cup X_6, X_7\}$$

infer the X-splits of $\mathcal{T}$ induced by $e_1, e_2, \ldots, e_6$. We next describe two further characters which, together with χ_1 and χ_2, infer the remaining X-splits of $\mathcal{T}$ induced by its interior edges.

Let $E_1, E_2, \ldots, E_7$ denote the sets of interior edges of $\mathcal{T}$ separated by $\{e_1, e_2, \ldots, e_6\}$. By our initial assumption, $|E_i| \leq 2$ for all i. Now select edges f_1, f_2, and f_3 in $E_1 \cup E_2 \cup \cdots \cup E_7$ such that no two are from the same E_i and an edge is selected from each E_i of size 2. Let χ_3 denote any character displayed by $\{f_1, f_2, f_3\}$. By Lemma 3.2, it follows that χ_1, χ_2, and χ_3 infer the X-splits of $\mathcal{T}$ induced by f_1, f_2 and f_3. Now select the remaining interior edges of $\mathcal{T}$, say f_4, f_5, and f_6 in $E_1 \cup E_2 \cup \cdots \cup E_7 - \{f_1, f_2, f_3\}$, and let χ_4 denote any character displayed by $\{f_4, f_5, f_6\}$. Since $|E_i| \leq 2$ for all i, it follows by another application of Lemma 3.2 that χ_1, χ_2, χ_3, and χ_4 infer the X-splits of $\mathcal{T}$ induced by f_4, f_5, and f_6. Hence, as all nontrivial X-splits of $\mathcal{T}$ are inferred by the collection $\{\chi_1, \chi_2, \chi_3, \chi_4\}$, we have our desired collection of four 4-state characters. □

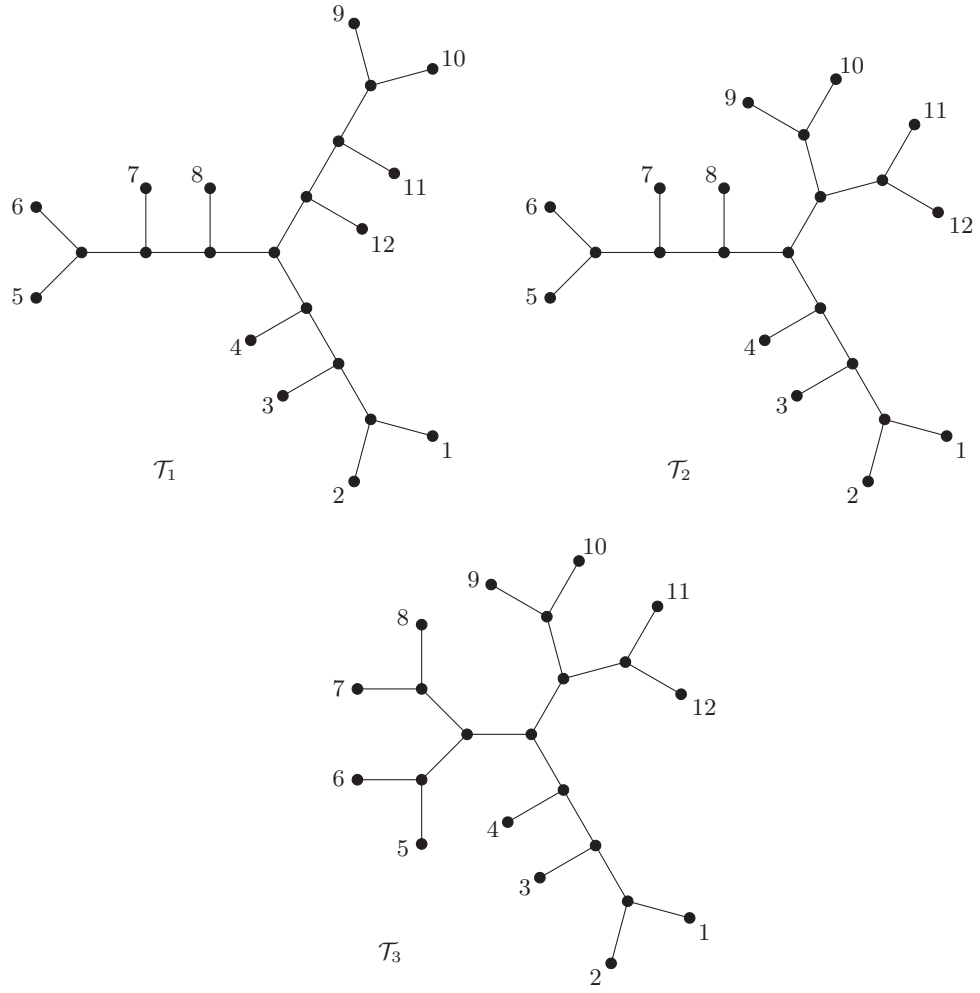

FIG. 4. *The phylogenetic trees $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$.*

LEMMA 5.5. *Let $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ be the three phylogenetic X-trees, where $X = \{1, 2, \ldots, 12\}$, shown in Figure 4. Then, for each $s \in \{1, 2, 3\}$, there is a collection of three 4-state characters that defines $\mathcal{T}_s$.*

Proof. Let $\mathcal{C}_1 = \{\chi_{11}, \chi_{12}, \chi_{13}\}$ be a collection of characters on $\{1, 2, \ldots, 12\}$, where

$$\pi(\chi_{11}) = \big\{\{1,2\},\{3,4,12\},\{5,6,7,8\},\{9,10,11\}\big\},$$
$$\pi(\chi_{12}) = \big\{\{1,2,3\},\{4,7,8\},\{5,6\},\{9,10,11,12\}\big\},$$
$$\pi(\chi_{13}) = \big\{\{1,2,3,4\},\{8,11,12\},\{5,6,7\},\{9,10\}\big\},$$

$\mathcal{C}_2 = \{\chi_{21}, \chi_{22}, \chi_{23}\}$ be a collection of characters on $\{1, 2, \ldots, 12\}$, where

$$\pi(\chi_{21}) = \big\{\{1,2\},\{3,4,9,10\},\{5,6,7,8\},\{11,12\}\big\},$$
$$\pi(\chi_{22}) = \big\{\{1,2,3\},\{4,7,8\},\{5,6\},\{9,10,11,12\}\big\},$$
$$\pi(\chi_{23}) = \big\{\{1,2,3,4\},\{8,11,12\},\{5,6,7\},\{9,10\}\big\},$$

and $\mathcal{C}_3 = \{\chi_{31}, \chi_{32}, \chi_{33}\}$ be a collection of characters on $\{1, 2, \ldots, 12\}$, where

$$\pi(\chi_{31}) = \big\{\{1,2\},\{3,4,9,10\},\{5,6,7,8\},\{11,12\}\big\},$$
$$\pi(\chi_{32}) = \big\{\{1,2,3\},\{4,7,8\},\{5,6\},\{9,10,11,12\}\big\},$$
$$\pi(\chi_{33}) = \big\{\{1,2,3,4\},\{5,6,11,12\},\{7,8\},\{9,10\}\big\}.$$

For all $s \in \{1,2,3\}$, it is easily checked that $\mathcal{C}_s$ is convex on $\mathcal{T}_s$ and distinguished by $\mathcal{C}_s$.

The intersection graph $\mathrm{int}(\mathcal{C}_1)$ is chordal and so it is the unique minimal restricted chordal completion of $\mathrm{int}(\mathcal{C}_1)$. Therefore, by Theorem 2.2, $\mathcal{C}_1$ defines $\mathcal{T}_1$. Now $\mathrm{int}(\mathcal{C}_2)$ is not chordal, but any restricted chordal completion of $\mathrm{int}(\mathcal{C}_2)$ must include the edge joining $(\chi_{21}, \{3,4,9,10\})$ and $(\chi_{23}, \{8,11,12\})$. The graph $\mathrm{int}(\mathcal{C}_2)$ together with this edge is chordal and so it is the unique minimal restricted chordal completion of $\mathrm{int}(\mathcal{C}_2)$. Thus, by Theorem 2.2, $\mathcal{C}_2$ defines $\mathcal{T}_2$. For $\mathcal{C}_3$, the situation is similar to that for $\mathcal{C}_2$ except that, for any restricted chordal completion of $\mathrm{int}(\mathcal{C}_3)$, two specific edges and not one must be included. These edges join $(\chi_{31}, \{3,4,9,10\})$ and $(\chi_{33}, \{5,6,11,12\})$, and join $(\chi_{32}, \{4,7,8\})$ and $(\chi_{33}, \{5,6,11,12\})$. This completes the proof of the lemma. □

LEMMA 5.6. *Let $\mathcal{T}$ be a binary phylogenetic X-tree with 15 leaves and an internal pseudopath P of length 6. If P separates the internal edges of $\mathcal{T}$ not in P into sets of size at most 3, then there is a collection of four 4-state characters that defines $\mathcal{T}$.*

Proof. Suppose that P separates the interior edges of $\mathcal{T}$ not in P into sets of size at most three. In order, let $e_1, e_2, \ldots, e_6$ denote the edges of P. Let $E_1, E_2, \ldots, E_7$ denote the sets of interior edges of $\mathcal{T}$ separated by $\{e_1, e_2, \ldots, e_6\}$. By Lemma 5.4, we may assume that $|E_i| = 3$ for some i. First suppose that E_i is the only set of size three. If E_i contains an edge that can extend P to an internal pseudopath of length 7, then it is easily checked we can choose another internal path of length 6 and that this path has the property of separating the interior edges of $\mathcal{T}$ not in it into sets of size at most two. In this instance, by Lemma 5.4, there is a collection of four 4-state characters that defines $\mathcal{T}$. Thus we may assume that there is no such edge in E_i. By symmetry, we may assume that $i \in \{1,2,3,4\}$. If (i) $i = 1$, (ii) $i = 2$ and $|E_1| \leq 1$, or (iii) $i = 3$ and $|E_1| = |E_2| = 0$, then it is easily seen that we can choose another internal path of length 6 whose first two edges are in E_i with the property that it separates the internal edges not in it into sets of size at most two. If (i) $i = 2$ and $|E_1| = 2$, (ii) $i = 3$ and $|E_1 \cup E_2| \in \{1,2\}$, or (iii) $i = 4$, then, for some $s \in \{1,2,3\}$, the binary phylogenetic tree $\mathcal{T}_s$ in Figure 4 is a restriction of $\mathcal{T}$ up to labeling. Moreover, it is easily seen that a $\mathcal{T}_s$-representable subset F of edges can be chosen so that F separates the three interior edges of $\mathcal{T}$ not in F into singletons. Note that if $i = 3$, then $|E_1 \cup E_2| \leq 2$; otherwise, $|E_2| = 3$ or we can extend P. It now follows by Lemmas 3.2, 5.2, and 5.5 that there is a collection of four 4-state characters that defines $\mathcal{T}$.

Now suppose that there are distinct i and j such that $|E_i| = |E_j| = 3$. Without loss of generality, we may assume that $i < j$. If either E_i or E_j contains an edge that can extend P, then $\mathcal{T}$ has an internal pseudopath of length 6 separating the interior edges of $\mathcal{T}$ not in it into sets of size at most two apart from one possible set which has size at most three. By Lemma 5.4 and the argument in the previous paragraph, we may assume that neither E_i nor E_j has such an edge. Furthermore, by symmetry, we may assume that $i \in \{1,2,3,4\}$. If $|i - j| \geq 3$, then it is easily checked that we can choose another internal pseudopath of length 6 whose first and last edges are in $E_i \cup E_j$ with the property that it separates the interior edges not in it into sets in

which at most one set has size three and all other sets have size at most two. By Lemma 5.4 and the argument in the previous paragraph, there is a collection of four 4-state characters that defines $\mathcal{T}$. If $|i-j| \in \{1,2\}$, then, unless $\{i,j\} = \{3,5\}$, there is some $s \in \{1,2,3\}$ such that $\mathcal{T}_s$ in Figure 4 is a restriction of $\mathcal{T}$ up to labeling. Furthermore, there is a $\mathcal{T}_s$-representable subset F of edges that can be chosen so that F separates the three interior edges of $\mathcal{T}$ not in F into singletons. By Lemmas 3.2, 5.2, and 5.5, there is a collection of four 4-state characters that defines $\mathcal{T}$. In the exceptional case, we can choose an internal pseudopath of length 6 whose first and last edges are in $E_3 \cup E_5$ with the property that it separates the remaining internal edges of $\mathcal{T}$ into sets of size at most two, in which case, by Lemma 5.4, there is a collection of four 4-state characters that defines $\mathcal{T}$. □

LEMMA 5.7. *Let $\mathcal{T}$ be a binary phylogenetic X-tree with 15 leaves and an internal pseudopath P of length 6. If P separates the interior edges of $\mathcal{T}$ not in P into sets one of which has size at least four, then there is a collection of four 4-state characters that defines $\mathcal{T}$.*

Proof. Suppose that P separates the interior edges of $\mathcal{T}$ not in P into sets one of which has size at least four. In order, let $e_1, e_2, \ldots, e_6$ denote the edges of P and let $E_1, E_2, \ldots, E_7$ be the sets of interior edges not in P separated by $\{e_1, e_2, \ldots, e_6\}$. For some i, we have $|E_i| \in \{4,5,6\}$. By symmetry, we may assume that $i \in \{1,2,3,4\}$. First suppose that $|E_i| = 4$. If E_i contains an edge that can extend P to an internal pseudopath of length 7, then we can choose another internal pseudopath of $\mathcal{T}$ of length 6 with the property that it separates the interior edges of $\mathcal{T}$ not in it into sets of size at most three, in which case, by Lemma 5.6, there is a collection of four 4-state characters that defines $\mathcal{T}$. Thus we may assume that there is no such edge in E_i. If (i) $i \in \{1,2\}$ or (ii) $i = 3$ and $|E_1 \cup E_2| \leq 1$, then we can choose another internal pseudopath of length 6 whose first edge is in E_i with the property that it separates the interior edges not in it into sets of size at most three. By Lemma 5.6, there is a set of four 4-state characters that defines $\mathcal{T}$. If (i) $i = 3$ and $|E_1 \cup E_2| = 2$, or (ii) $i = 4$, then, for some $s \in \{1,2,3\}$, the binary phylogenetic tree $\mathcal{T}_s$ in Figure 4 is a restriction of $\mathcal{T}$ up to labeling. Furthermore, a $\mathcal{T}_s$-representable subset F of edges can be chosen so that F separates the three interior edges not in F into singletons. It now follows by Lemmas 3.2, 5.2, and 5.5 that there is a collection of four 4-state characters that defines $\mathcal{T}$.

Now suppose that $|E_i| \in \{5,6\}$. Then, regardless of i, we can choose another internal pseudopath of length 6 whose first edge is in E_i with the property that it separates the interior edges not in it into sets of size at most four. By Lemma 5.6 and the argument in the last paragraph, there is a set of four 4-state characters that defines $\mathcal{T}$. □

The following corollary is an immediate consequence of Lemmas 5.4, 5.6, and 5.7.

COROLLARY 5.8. *Let $\mathcal{T}$ be a binary phylogenetic X-tree with 15 leaves and an internal pseudopath of length 6. Then there is a collection of four 4-state characters that defines $\mathcal{T}$.*

COROLLARY 5.9. *Let $\mathcal{T}$ be a binary phylogenetic X-tree with 13 or 14 leaves. Then there is a collection of four 4-state characters that defines $\mathcal{T}$.*

Proof. We prove the corollary for when $\mathcal{T}$ is a binary phylogenetic tree with 13 leaves. The analogous proof for when $\mathcal{T}$ is a binary phylogenetic tree with 14 leaves is simpler and omitted. Let $\mathcal{T}$ be a binary phylogenetic X-tree with $X = \{1,2,\ldots,13\}$. We next extend $\mathcal{T}$ to a binary phylogenetic X'-tree, where $X' = \{1,2,\ldots,15\}$ as follows. For each nontrivial X-split of $\mathcal{T}$, add the elements 14 and 15 to the cell of the X-split containing 13. The resulting collection of X'-splits together with the

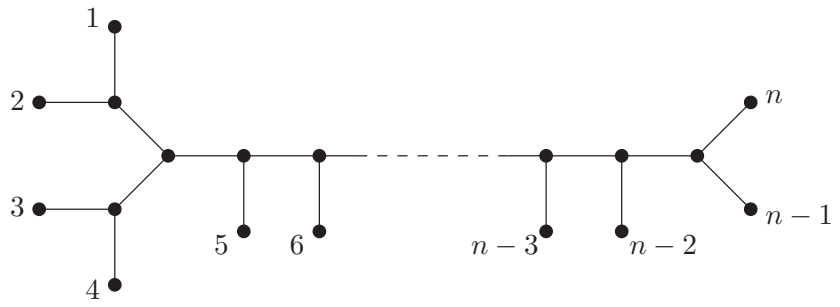

FIG. 5. *A caterpillar-like tree with exactly three cherries* $\{1,2\}$, $\{3,4\}$, *and* $\{n-1,n\}$.

X'-split $\{\{14,15\},\{1,2,\ldots,13\}\}$ is the collection of nontrivial X'-splits of a binary phylogenetic X'-tree $\mathcal{T}'$. By Corollary 5.8, there is a collection $\mathcal{C}'$ of four 4-state characters that defines $\mathcal{T}'$. Now, amongst the characters in $\mathcal{C}'$, there is a character χ_1' in which $\{13,14,15\} \in \pi(\chi_1')$ and a character χ_2' in which $\{14,15\} \in \pi(\chi_2')$. Deleting $\{13,14,15\}$ from $\pi(\chi_1')$, deleting $\{14,15\}$ from $\pi(\chi_2')$, and deleting the elements 14 and 15 from the partitions induced by the remaining two characters in $\mathcal{C}'$, gives the partitions of a collection $\mathcal{C}$ of four 4-state characters on $\{1,2,\ldots,13\}$. Note that, except for χ_2', each of the characters in $\mathcal{C}'$ has the property of mapping the elements 14 and 15 to the same state in which 13 is mapped. Clearly, $\mathcal{C}$ is convex on $\mathcal{T}$. Furthermore, $\mathcal{C}$ defines $\mathcal{T}$. If not, then there is another binary phylogenetic X-tree on which $\mathcal{C}$ is convex. By extending this tree to a binary phylogenetic X'-tree in the same way that we extended $\mathcal{T}$ to $\mathcal{T}'$, we obtained a binary phylogenetic X'-tree on which $\mathcal{C}'$ is convex but distinct to $\mathcal{T}'$; a contradiction. This completes the proof of the corollary. □

Let $\mathcal{T}$ be a binary phylogenetic X-tree. A *cherry* is a 2-element subset $\{x,y\}$ of X with the property that x and y have a common neighbor. A binary phylogenetic tree is *caterpillar-like* if it does not have three cherries each two of which are separated by at least three interior edges. It is easily checked that there are exactly three types of caterpillar-like trees. The first type has precisely two cherries, the second type has precisely three cherries, two of which are separated by only two interior edges, and the third type has precisely four cherries, in two pairs each separated by only two interior edges. An illustration of a caterpillar-like tree with exactly three cherries is shown in Figure 5.

LEMMA 5.10. *Let $\mathcal{T}$ be a binary phylogenetic X-tree that is caterpillar-like with n leaves, where $n \geq 11$. Then there is a collection of 4-state characters of size $\lceil \frac{n-3}{3} \rceil$ that defines $\mathcal{T}$.*

Proof. Let P be a longest internal path of $\mathcal{T}$. Since $n \geq 11$, it is easily checked that P has at least 6 internal edges. In fact, the exact number of internal edges in P is $n-k$, where $k \in \{3,4,5\}$, depending upon whether $\mathcal{T}$ is the first, second, or third type of caterpillar-like tree. We next partition the edges of P into three nonempty parts, each part consisting of consecutive edges in P. If we view these parts as being ordered, the first part contains the first edge of P and the third part contains the last edge of P. If $n-k \equiv 0 \pmod 3$, then each part consists of $\frac{n-k}{3}$ edges. If $n-k \equiv 1 \pmod 3$, then the first part consists of $\lfloor \frac{n-k}{3} \rfloor + 1$ edges and the other two parts each consist of $\lfloor \frac{n-k}{3} \rfloor$ edges. If $n-k \equiv 2 \pmod 3$, then the first and second parts each consist of $\lfloor \frac{n-k}{3} \rfloor + 1$ edges and the third part consists of $\lfloor \frac{n-k}{3} \rfloor$ edges.

Using the above partition of P, we next construct a collection of 4-state characters of size $\lfloor \frac{n-k}{3} \rfloor$ that will infer all but at most four of the nontrivial X-splits of $\mathcal{T}$. View each part as an ordered set with the ordering consistent with the order of the edges in P. For all $i \in \{1, 2, \ldots, \lfloor \frac{n-k}{3} \rfloor\}$, define χ_i to be a 4-state character on $\{1, 2, \ldots, n\}$ displayed (collectively) by the ith edge in each part. By Lemma 3.1, χ_1 and χ_2 infer the X-splits of $\mathcal{T}$ induced by the edges displaying χ_1 and χ_2. Moreover, for each $j \in \{3, 4, \ldots, \lfloor \frac{n-k}{3} \rfloor\}$, it follows by Lemma 3.2 that $\{\chi_1, \chi_2, \chi_j\}$ infers the X-splits of $\mathcal{T}$ induced by the edges displaying χ_j. Thus, except for at most two edges, the collection

$$\mathcal{C} = \left\{\chi_1, \chi_2, \ldots, \chi_{\lfloor \frac{n-k}{3} \rfloor}\right\}$$

infers each of the X-splits of $\mathcal{T}$ induced by the edges of P.

Now there are at most two interior edges of $\mathcal{T}$ not in P. Therefore there are at most four interior edges of $\mathcal{T}$ whose X-splits are not inferred by $\mathcal{C}$. Let F denote the set consisting of these interior edges of $\mathcal{T}$. By the way in which P was partitioned, no two of the edges in F are adjacent. If $|F| = 0$, in which case $\mathcal{T}$ is the first type of caterpillar-like tree, $\mathcal{C}$ is an appropriate collection of 4-state characters that defines $\mathcal{T}$. If $|F| \in \{1, 2, 3\}$, then define $\chi_{\lfloor \frac{n-k}{3} \rfloor + 1}$ to be a character on $\{1, 2, \ldots, n\}$ displayed by the edges in F. It follows by Lemma 3.2 that $\mathcal{C} \cup \{\chi_{\lfloor \frac{n-k}{3} \rfloor + 1}\}$ infers each of the X-splits induced by the edges in F and thus defines $\mathcal{T}$. Lastly, if $|F| = 4$, then define $\chi_{\lfloor \frac{n-k}{3} \rfloor + 1}$ to be a character on $\{1, 2, \ldots, n\}$ displayed by three of the edges in F and $\chi_{\lfloor \frac{n-k}{3} \rfloor + 2}$ to be a character on $\{1, 2, \ldots, n\}$ displayed by the remaining edge in F. By two applications of Lemma 3.2, $\mathcal{C} \cup \{\chi_{\lfloor \frac{n-k}{3} \rfloor + 1}, \chi_{\lfloor \frac{n-k}{3} \rfloor + 2}\}$ infers each of the X-splits induced by the edges in F and thus defines $\mathcal{T}$. This completes the proof of the lemma. □

Proof of Theorem 1.2. It remains to show that, for all $n \geq 13$, if $\mathcal{T}$ is a phylogenetic X-tree on $\{1, 2, \ldots, n\}$, then there is a collection of 4-state characters of size $\lceil \frac{n-3}{3} \rceil$ that defines $\mathcal{T}$. The proof is by induction on n. If $n \in \{13, 14, 15\}$, then the theorem holds by Corollaries 5.8 and 5.9. Now suppose that $n \geq 16$ and that the theorem holds for all binary phylogenetic trees on $\{1, 2, \ldots, n-3\}$. By Lemma 5.10, we may assume that $\mathcal{T}$ is not a caterpillar-like tree. Hence $\mathcal{T}$ has three pairs of cherries in which each pair is separated by at least three interior edges. Without loss of generality, we may assume that these cherries are $\{n-5, n-2\}$, $\{n-4, n-1\}$, and $\{n-3, n\}$. Let $\mathcal{T}_{n-3}$ denote the binary phylogenetic tree obtained from $\mathcal{T}$ by deleting the leaves $n-2$, $n-1$, and n, and suppressing the resulting degree-two vertices. By induction, there is a collection $\mathcal{C}_{n-3}$ of 4-state characters of size $\lfloor \frac{n-6}{3} \rfloor$ that defines $\mathcal{T}_{n-3}$. Now $\mathcal{T}_{n-3}$ is a restriction of $\mathcal{T}$, and so there is a collection F of interior edges of $\mathcal{T}$ that is $\mathcal{T}_{n-3}$-representable. By Lemma 5.2, $(\mathcal{C}_{n-3})_F$ is a collection of characters on $\{1, 2, \ldots, n\}$ that infer the X-splits of $\mathcal{T}$ induced by the edges in F. Let χ be a character on $\{1, 2, \ldots, n\}$ in which

$$\pi(\chi) = \big\{\{n-5, n-2\}, \{n-4, n-1\}, \{n-3, n\}, \{1, 2, \ldots, n-6\}\big\}$$

and let $\mathcal{C}$ be the collection $(\mathcal{C}_{n-3})_F \cup \{\chi\}$ of 4-state characters on $\{1, 2, \ldots, n\}$. Since any two of the cherries $\{n-5, n-2\}$, $\{n-4, n-1\}$, and $\{n-3, n\}$ are separated by at least three interior edges of $\mathcal{T}$, it follows by Lemma 3.2 that $\mathcal{C}$ defines $\mathcal{T}$. Moreover, $\mathcal{C}$ consists of $\lfloor \frac{n-3}{3} \rfloor$ 4-state characters. This completes the proof of the theorem. □

We end with a remark regarding where the argument breaks down for when $n \leq 11$. The proof of Theorem 1.2 uses induction on n, where the base case consists

of establishing the result for three consecutive values of n. For a fixed $n \leq 9$, there is no choice whereby the theorem holds for n, $n+1$, and $n+2$. The reason is that, for $5 \leq n \leq 9$, we simply don't have enough characters to define every binary phylogenetic tree with n leaves. For example, when $n = 9$, we have $\lceil \frac{n-3}{4-1} \rceil = 2$. But any binary phylogenetic tree with an interior vertex incident with three interior edges cannot be defined by two characters as these two characters won't distinguish all three of these edges.

Acknowledgment. We thank the referees for their comments and careful reading of the paper.

REFERENCES

[1] M. Bordewich, K. T. Huber, and C. Semple, *Identifying phylogenetic trees*, Discrete Math., 300 (2005), pp. 30–43.
[2] P. Buneman, *The recovery of trees from measures of dissimilarity*, in Mathematics in the Archaeological and Historical Sciences, E. R. Hodson, D. G. Kendall, P. Tautu, eds., Edinburgh University Press, Edinburgh, 1971, pp. 387–395.
[3] P. Buneman, *A characterisation of rigid circuit graphs*, Discrete Math., 9 (1974), pp. 205–212.
[4] F. Gavril, *The intersection graphs of subtrees in trees are exactly the chordal graphs*, J. Combin. Theory Ser. B, 16 (1974), pp. 47–56.
[5] R. Gysel and D. Gusfield, *Extensions and improvements to the chordal graph approach to the multi-state perfect phylogeny problem*, in Bioinformatics Research and Applications (ISBRA 2010), Lecture Notes in Bioinform. 6053, Springer, Berlin, 2010, pp. 52–60.
[6] K. T. Huber, V. Moulton, and M. Steel, *Four characters suffice to convexly define a phylogenetic tree*, SIAM J. Discrete Math., 18 (2005), pp. 835–843.
[7] F. Lam, D. Gusfield, and S. Sridhar, *Generalizing the splits equivalence theorem and four gamete condition: Perfect phylogeny on three-state characters*, SIAM J. Discrete Math., 25 (2011), pp. 1144–1175.
[8] C. Semple and M. Steel, *Tree reconstruction from multi-state characters*, Adv. Appl. Math., 28 (2002), pp. 169–184.
[9] C. Semple and M. Steel, *A characterisation for a set of partial partitions to define an X-tree*, Discrete Math., 247 (2002), pp. 169–186.
[10] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, Oxford, 2003.
[11] J. R. Walter, *Representations of chordal graphs as subtrees of a tree*, J. Graph Theory, 2 (1978), pp. 265–267.