Identifying X-Trees with Few Characters

Magnus Bordewich[1], Charles Semple[2] and Mike Steel[2*]

[1] Department of Computer Science
Durham University,
Durham DH1 3LE, United Kingdom
m.j.r.bordewich@durham.ac.uk

[2] Department of Mathematics and Statistics
University of Canterbury
Christchurch, New Zealand
c.semple@math.canterbury.ac.nz, m.steel@math.canterbury.ac.nz

Submitted: Jan 12, 2006; Accepted: Sep 14, 2006; Published: Sep 28, 2006
Mathematics Subject Classification: 92B15, 92B10, 05C05

Abstract

Previous work has shown the perhaps surprising result that, for any binary phylogenetic tree $\mathcal{T}$, there is a set of four characters that define $\mathcal{T}$. Here we deal with the general case, where $\mathcal{T}$ is an arbitrary X-tree. We show that if d is the maximum degree of any vertex in $\mathcal{T}$, then the minimum number of characters that identify $\mathcal{T}$ is $\log_2 d$ (up to a small multiplicative constant).

1 Introduction

For a finite set X, an *X-tree* $\mathcal{T} = (T; \phi)$ is an ordered pair consisting of a tree T, with vertex set V say, and a map $\phi : X \to V$ with the property that, for all $v \in V$ with degree at most two, $v \in \phi(X)$. X-trees are commonly referred to as *semi-labelled trees.* An X-tree is *binary* if every interior vertex has degree three. An X-tree is *phylogenetic* if ϕ is a bijection from X to the leaf set of T. For example, in Fig. 1, $\mathcal{T}_1$ and $\mathcal{T}_2$ are both X-trees, where $\mathcal{T}_2$ is also phylogenetic. In evolutionary biology, semi-labelled trees are used to represent the ancestral history of a collection X of species. Moreover, it has recently been recognised that their rooted counterparts have important practical applications [2, 4]. The data that is used to reconstruct such trees are functions on subsets of X. In biology,

*The first author was supported by the New Zealand Institute of Mathematics and its Applications funded programme *Phylogenetic Genomics*. The second and third authors were supported by the New Zealand Marsden Fund (UOC310). This work was done while the first author was a Postdoctoral Fellow at the University of Canterbury.

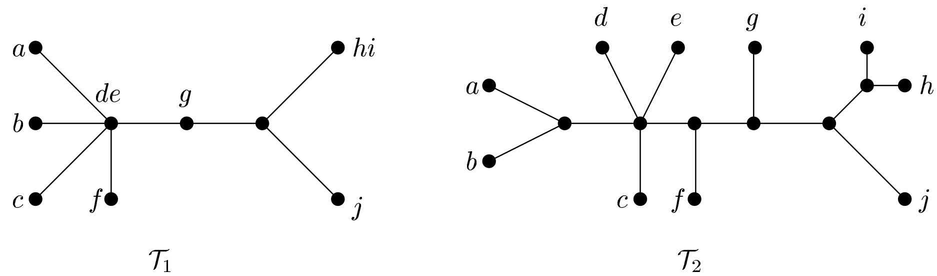

Figure 1: An X-tree $\mathcal{T}_1$ and a phylogenetic X-tree $\mathcal{T}_2$, which is a refinement of $\mathcal{T}_1$, where $X = \{a, b, c, d, e, f, g, h, i\}$.

these functions are commonly known as *characters*. In this paper, we are interested in characters whose evolution has been "homoplasy-free", and in the question of how many characters are needed to reconstruct an X-tree. This question has been investigated previously in two papers.

Semple and Steel [5] showed that, for any binary phylogenetic X-tree $\mathcal{T}$, there exists a collection $\mathcal{C}$ of at most five characters that *defines* $\mathcal{T}$; that is, $\mathcal{T}$ is the only phylogenetic X-tree (up to isomorphism) that "displays" $\mathcal{C}$. Huber *et al.* [3] sharpened this result by showing that there is always a collection of at most four characters that defines $\mathcal{T}$. Since it is not always possible to define a binary phylogenetic tree with three characters (see [5]), this last result is the best possible.

In practice, "definitiveness" is a very restrictive notion, and only applies to binary phylogenetic trees. One that is more useful and generalises this notion is "identifiability". A collection $\mathcal{C}$ of characters *identifies* an X-tree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{C}$ and all X-trees that display $\mathcal{C}$ are "refinements" of $\mathcal{T}$ (see [5]). In this paper, we investigate this latter notion and consider the question of how many characters are needed to identify an arbitrary X-tree. The results in this paper are strikingly different to the results in the two earlier papers. However, the four character result mentioned in the previous paragraph turns out to be an immediate consequence of the main result of this paper. The rest of this section formally describes this result.

For an X-tree $\mathcal{T}$, the set X is called the *label set* of $\mathcal{T}$ and is denoted $\mathcal{L}(\mathcal{T})$. Furthermore, if v is a vertex of T, then $\phi^{-1}(v)$ is the *label set* of v, and the elements of this set are the elements of X *labelling* v.

A *character on* X is a partition of a subset of X, where we typically denote the partition $\{A_1, A_2, \ldots, A_k\}$ by $A_1|A_2|\cdots|A_k$. If a character $\chi = A|B$ has only two parts in the partition, then χ is a *two-state* character. Let χ be a character on X and let $\mathcal{T} = (T; \phi)$ be an X-tree. We say that $\mathcal{T}$ *displays* χ if there is a subset E of edges of T such that, for all blocks A and B in χ, $\phi(A)$ and $\phi(B)$ are subsets of the vertex sets of different components of the graph obtained from T by deleting the edges in E. This notion of displays captures the biological notion of characters evolving in a homoplasy-free way. Extending the examples of X-trees shown in Fig. 1, let $\chi = \{ac\}|\{fi\}$ be a character on X. (For brevity of notation, in the remainder of this paper we shall omit the braces

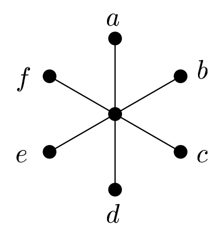

Figure 2: The 6-star.

from this notation when the meaning is clear, hence write $\chi = ac|fi$.) Then $\mathcal{T}_2$ displays χ, but $\mathcal{T}_1$ does not display χ. More generally, $\mathcal{T}$ *displays* a collection $\mathcal{C}$ of characters on X if $\mathcal{T}$ displays each character in $\mathcal{C}$.

There is interest not only in whether an X-tree $\mathcal{T}$ displays a collection $\mathcal{C}$ of characters, but also in whether it is the only X-tree that displays $\mathcal{C}$; that is $\mathcal{C}$ *defines* $\mathcal{T}$; in which case, $\mathcal{T}$ is a binary phylogenetic X-tree. However, a closely related notion, and one that is more general and more useful in practice, is that of identifiability. Associated with each edge e of an X-tree $\mathcal{T} = (T;\phi)$ is an *X-split*; that is, a bipartition of X into the label sets of the two connected components of $\mathcal{T}\backslash e = (T\backslash e, \phi)$. An X-tree $\mathcal{T}'$ is a *refinement* of $\mathcal{T}$ if every X-split of $\mathcal{T}$ is an X-split of $\mathcal{T}'$. Graphically speaking, $\mathcal{T}'$ is a refinement of $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by contracting edges and amalgamating the label sets. In Fig. 1, $\mathcal{T}_2$ is a refinement of $\mathcal{T}_1$. An X-tree $\mathcal{T}$ *displays* an X'-tree $\mathcal{T}'$ if $X' \subseteq X$, and the subtree of $\mathcal{T}$ induced by the vertices labelled with elements of X' is a refinement of $\mathcal{T}'$.

We say that $\mathcal{C}$ *identifies* an X-tree $\mathcal{T}$, if $\mathcal{T}$ displays $\mathcal{C}$ and every X-tree $\mathcal{T}'$ that displays $\mathcal{C}$ is a refinement of $\mathcal{T}$. Observe that if $\mathcal{T}$ is a binary phylogenetic tree, then $\mathcal{C}$ identifies $\mathcal{T}$ if and only if $\mathcal{C}$ defines $\mathcal{T}$. Moreover, if $\mathcal{T}$ is not a binary phylogenetic tree, then no set of characters defines $\mathcal{T}$, although $\mathcal{T}$ can be identified. A characterisation of what it means for a collection of characters to identify an X-tree has recently been given in terms of chordal graphs [1].

Example 1.1. Let $X = \{a, b, c, d, e, f\}$ and let $\mathcal{T}$ be the phylogenetic X-tree shown in Fig. 2. The collection

$$\mathcal{C} = \big\{a|b|c|def, a|bcf|d|e, ace|b|d|f, abd|c|e|f\big\}$$

of characters on X identifies $\mathcal{T}$. In other words, not only does $\mathcal{T}$ display $\mathcal{C}$, but every X-tree that displays $\mathcal{C}$ is a refinement of $\mathcal{T}$.

To see that $\mathcal{C}$ does indeed identify $\mathcal{T}$, let $\mathcal{T}'$ be an X-tree that displays $\mathcal{C}$. We will show that $\mathcal{T}'$ is a refinement of $\mathcal{T}$. First observe that, for every pair of elements in X, there is a character in $\mathcal{C}$ in which this pair are in separate blocks. This implies that no vertex of $\mathcal{T}'$ has a label set with more than one element of X in it. We next show that $\mathcal{T}'$ is phylogenetic (that is, leaf labelled).

It follows from the first two characters in $\mathcal{C}$ that $\mathcal{T}'$ displays the characters $a|def$ and $a|bcf$. As the intersection of the last two blocks in each of these characters is non-empty, this implies that $\mathcal{T}'$ must also display the character $a|bcdef$. A similar check shows that $\mathcal{T}'$ must display each of the characters $b|acdef$, $c|abdef$, $d|abcef$, $e|abcdf$, and $f|abcde$. This means that $\mathcal{T}'$ is phylogenetic. Since every phylogenetic X-tree is a refinement of $\mathcal{T}$, we now deduce that $\mathcal{T}'$ is a refinement of $\mathcal{T}$, and conclude that $\mathcal{C}$ identifies $\mathcal{T}$. $\square$

As stated earlier, it is shown in [3] that, for any binary phylogenetic tree $\mathcal{T}$, there is a collection of at most four characters that defines $\mathcal{T}$. In this paper, we deal with the general case in which $\mathcal{T}$ is an arbitrary X-tree and consider the minimal size of a collection of characters that identifies $\mathcal{T}$. In particular, we establish the following analogue of the four character result.

Theorem 1.2. *Let X be a finite set, let k be a positive integer, and let $\mathcal{T}$ be an X-tree. Suppose that the maximum degree of any vertex in $\mathcal{T}$ is d.*

(i) If $k = 4\lceil \log_2(d-2)\rceil + 4$, then there is a collection of k characters that identifies $\mathcal{T}$.

(ii) If $k < \log_2 d$, then there is no collection of k characters that identifies $\mathcal{T}$.

The proof of Theorem 1.2 is constructive and hence, given $\mathcal{T}$, a set of characters of size $k = 4\lceil \log_2(d-2)\rceil + 4$ may be found efficiently. Observe that the four character result is an immediate consequence of (i) in Theorem 1.2, indeed the following slightly stronger corollary holds.

Corollary 1.3. *Let X be a finite set and let $\mathcal{T}$ be a binary X-tree. Then there is a collection of four characters that identifies $\mathcal{T}$.*

For an arbitrary X-tree $\mathcal{T}$ in which the maximum degree of a vertex is d, Theorem 1.2 says that the minimum number of characters that identify $\mathcal{T}$ is (roughly) between $\log_2 d$ and $4\log_2 d$. Some range will always be required, since for a given maximum degree, more characters are required to identify some X-trees than others; for example, the 3-star requires only three characters, while some other binary X-trees require four (see [5]).

Throughout the paper, the notation and terminology mostly follows Semple and Steel [6]. The set X will always be a finite set, and for an X-tree $\mathcal{T} = (T;\phi)$, we will often refer to the vertices and edges of T as the vertices and edges of $\mathcal{T}$ provided no ambiguity arises. Let $\psi : A \to B$ be a map and let $b \in B$, we will frequently use $\psi^{-1}(b)$ to denote the (possibly empty) subset of A whose elements are mapped to b under ψ. For any graph G on vertex set V, and any map $\phi : X \to V$, we define the *induced character χ of G* to be the partition of X induced by the connected components of G.

2 Proof of Theorem 1.2

In this section, we prove Theorem 1.2. We begin by showing that the lower bound on the number of characters required to identify an X-tree must grow at least logarithmically with the size of its maximum vertex degree. To establish this result, we first prove the following lemma. An X-tree $\mathcal{T} = (T;\phi)$ is a *d-star* if T is a star tree with d leaves and ϕ is one-to-one. Observe that the interior vertex of T may or may not be labelled; in the latter case, $\mathcal{T}$ is a phylogenetic tree. A 6-star with no interior label is shown in Fig. 2.

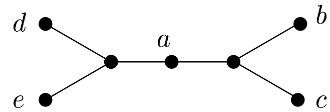

Figure 3: An X-tree displaying $a|b|c|de,\ a|bc|d|e,\ ace|b|d,\ abd|e|c$.

Lemma 2.1. *Let $\mathcal{T}$ be a d-star and let k be a positive integer such that $k < \log_2 d$. Then no set of k characters identifies $\mathcal{T}$.*

Proof. Here we show that the result holds if $\mathcal{T}$ has no interior label. The proof that the result holds when $\mathcal{T}$ has an interior label is similar and omitted. Let X denote the label set of $\mathcal{T}$, and let $\mathcal{C} = \{\chi_1, \ldots, \chi_k\}$ be a collection of characters that identifies $\mathcal{T}$. We will show that $k \geq \log_2 |X|$. For each character $\chi_i \in \mathcal{C}$, the partial partition of X given by χ_i has at most one block with more than one label, for otherwise the $|X|$-star does not display χ_i. For each $i \in \{1, 2, \ldots, k\}$ for which χ_i has such a block, let B_i be the set of elements in this block. For each element $a \in X$, since every X-tree that displays $\mathcal{C}$ must contain the X-split $a|(X - a)$, we have

$$\bigcup_{i: a \notin B_i} B_i = X - a.$$

The number of distinct unions of the blocks B_i is at most 2^k, and this must be at least the number of labels $|X|$. It follows that $k \geq \log_2 d$ as required. $\square$

Remark. Despite the above lemma, it is interesting to note that the number of characters required to identify the d-star is not monotonic in d. We showed in Example 1.1 that we could identify the 6-star with the four characters

$$a|b|c|def,\ a|bcf|d|e,\ ace|b|d|f,\ abd|e|c|f.$$

It would be intuitive to assume that by removing f from each of these characters the resulting four characters identify the 5-star obtained from this particular 6-star by deleting the vertex labelled f (and its incident edge). However, this is not the case as the X-tree shown in Fig. 3 displays each of the characters

$$a|b|c|de,\ a|bc|d|e,\ ace|b|d,\ abd|e|c,$$

but this X-tree is not a refinement of this 5-star. Indeed, it is simple to show, by exhaustive arguments, that no set of four characters identifies the 5-star.

Proof of Theorem 1.2(ii). Let $\mathcal{T} = (T; \phi)$, let v be a vertex of $\mathcal{T}$ whose degree is d, and let $e_1, e_2, \ldots, e_d$ denote the edges of $\mathcal{T}$ incident with v. Let $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d$ denote the subtrees of $\mathcal{T}$ attached to v via $e_1, e_2, \ldots, e_d$, respectively. Furthermore, let $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_d$ denote the label sets of $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d$. Now suppose that $\mathcal{C}$ is a collection of k characters that identifies $\mathcal{T}$. Since $\mathcal{T}$ displays $\mathcal{C}$, for each χ in $\mathcal{C}$, at most one block contains elements in both $\mathcal{L}_i$ and $\phi^{-1}(v)$ for some i, or elements in both $\mathcal{L}_i$ and $\mathcal{L}_j$ for some distinct i and j. This fact is used freely in the rest of the proof. Let $\mathcal{C}'$ be the collection of characters obtained from $\mathcal{C}$ by replacing each character $\chi = A_1|A_2|\ldots|A_m$ in $\mathcal{C}$ with a character χ' formed as follows.

(i) Firstly, for each block A_j of χ define

$$A'_j = \begin{cases} \{l_i : A_j \cap \mathcal{L}_i \neq \emptyset,\ 1 \leq i \leq d\} & \text{if } A_j \cap \phi^{-1}(v) = \emptyset, \\ \{l_i : A_j \cap \mathcal{L}_i \neq \emptyset,\ 1 \leq i \leq d\} \cup \{z\} & \text{if } A_j \cap \phi^{-1}(v) \neq \emptyset. \end{cases}$$

(ii) Secondly, remove repeated blocks by forming the set

$$M' = \{j : 1 \leq j \leq m,\ \nexists\, j' < j \text{ such that } A'_{j'} = A'_j\}.$$

(iii) Lastly, if there is a block containing at least two distinct elements, then remove each single-element block that contains one of these elements. That is, form

$$M'' = \{j : j \in M',\ \text{if } |A'_j| = 1 \text{ then } \nexists\, j' \neq j \text{ such that } A'_j \subset A'_{j'}\}.$$

Now the character χ' is given by the partial partition $\{A'_j : j \in M''\}$. If $\phi^{-1}(v)$ is empty, then let $\mathcal{T}'$ be the d-star on $\{l_1, l_2, \ldots, l_d\}$, while if $\phi^{-1}(v)$ is non-empty, then let $\mathcal{T}'$ be the d-star on $\{l_1, l_2, \ldots, l_d, z\}$ in which the interior vertex is labelled z.

Now consider $\mathcal{C}'$. Clearly, $\mathcal{T}'$ displays $\mathcal{C}'$. We next show that $\mathcal{C}'$ identifies $\mathcal{T}'$. Suppose that this is not the case. Then there exists a semi-labelled tree $\mathcal{T}'' = (T''; \phi'')$ that displays $\mathcal{C}'$ and it is not a refinement of $\mathcal{T}'$. Let $\mathcal{T}^+$ be the X-tree obtained from $\mathcal{T}''$ by adjoining each of $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d$ to T'' at the vertex of T'' labelled by $l_1, l_2, \ldots, l_d$, respectively, and then labelling the vertex of T'' labelled by z with $\phi^{-1}(v)$. Since $\mathcal{T}''$ displays $\mathcal{C}'$, it is easily checked that $\mathcal{T}^+$ displays $\mathcal{C}$. But, as $\mathcal{T}''$ is not a refinement of $\mathcal{T}'$, it is easily seen that $\mathcal{T}^+$ is not a refinement of $\mathcal{T}$, contradicting the identifiability of $\mathcal{C}$. Hence $\mathcal{C}'$ identifies $\mathcal{T}'$. Since $|\mathcal{C}'| \leq |\mathcal{C}|$, it now follows that we have constructed a collection of at most k characters that identifies a d-star. But $k < \log_2 d$. This contradiction to Lemma 2.1 completes the proof. $\square$

With the lower bound in Theorem 1.2 established, we now turn to the upper bound (Theorem 1.2(i)). To this end, let $\mathcal{T} = (T; \phi)$ be an X-tree with maximum vertex degree d. Let s be an integer such that $\binom{s}{\lceil s/2 \rceil} \geq d - 2$. We will eventually show that there is a collection of at most $2s + 2$ characters that identifies $\mathcal{T}$.

We begin by defining a collection of two-state characters on X based on $\mathcal{T}$. Let Q be the collection of subsets of $\{1, 2, \ldots, s\}$ of size $\lceil s/2 \rceil$, and let q_0 denote the element $\{1, 2, \ldots, \lceil s/2 \rceil\}$ in Q. Let p be a fixed element that is not in Q. In what follows, q_0 and p play central roles. A labelling of the edges of $\mathcal{T}$ by the elements of $Q \cup \{p\}$ is *good* if there is a leaf ρ of $\mathcal{T}$ such that, for each vertex $v \in V(\mathcal{T})$, the edges incident with v that are not on the path from v to ρ have distinct labels, and

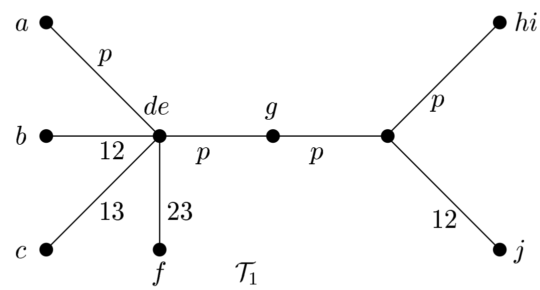

Figure 4: A good labelling of the X-tree $\mathcal{T}_1$, where ρ is the leaf labelled hi.

(i) if there is one such edge, then it is labelled p, while

(ii) if there are at least two such edges, then one is labelled p and one is labelled q_0.

Note that, by distinguishing a leaf ρ of $\mathcal{T}$ and recursively labelling the edges of $\mathcal{T}$ beginning with the edge incident with ρ in the appropriate way, it is straightforward to construct a good labelling for $\mathcal{T}$. To illustrate these ideas, recall the semi-labelled tree $\mathcal{T}_1$ shown in Fig. 1. Since the maximum vertex degree is five, we take $s = 3$, and so $Q = \{\{1,2\},\{1,3\},\{2,3\}\}$. Choosing ρ to be the leaf labelled hi, a good labelling of $\mathcal{T}_1$ is shown in Fig. 4.

Now suppose that we have a good labelling of $\mathcal{T}$ that is induced by a leaf ρ. For descriptive purposes regard the edges of $\mathcal{T}$ to be directed in such a way that each edge points away from ρ (*i.e.* edge (v, w) is directed from v to w if v is on the path from ρ to w). For each $v \in V(\mathcal{T})$, we associate two subsets of X, denoted $p(v)$ and $q_0(v)$, as follows. First consider the path in $\mathcal{T}$ that starts at v and follows the edges labelled p away from ρ. Since every non-leaf vertex has an edge coming out labelled p, this path extends all the way to a leaf of $\mathcal{T}$. Set $p(v) = \phi^{-1}(w)$, where w is the first vertex in this path that is labelled by an element of X. Now consider the path in $\mathcal{T}$ that starts at v and follows the edges labelled q_0 away from ρ. Since every vertex of degree at least three has an edge coming out labelled q_0, this path either extends all the way to a leaf of $\mathcal{T}$ or to a degree-two vertex of $\mathcal{T}$ that is labelled. Set $q_0(v) = \phi^{-1}(w')$, where w' is the first vertex in this path that is labelled by an element of X. Note that, if v is labelled, then $p(v) = q_0(v) = \phi^{-1}(v)$. Furthermore, if W is a subset of the vertices of T, then let $p(W)$ and $q_0(W)$ denote the sets $\{p(w) : w \in W\}$ and $\{q_0(w) : w \in W\}$, respectively.

Using the good labelling of $\mathcal{T}$, we are now ready to define a two-state character $\chi_{\mathcal{T}}(e)$ for each edge e of $\mathcal{T}$. Suppose that $e = (v, w)$, where v is on the path from ρ to w. Let u be the parent vertex of v (unless $v = \rho$) and let V be the set of children of v not including w. Let W be the set of children of w. Lastly, let u_p be the child of u such that (u, u_p) is labelled p and, provided u has at least two children, let u_{q_0} be the child of u such that (u, u_{q_0}) is labelled q_0. This set-up is illustrated in Fig. 5. We define $\chi_{\mathcal{T}}(e)$ as follows (where $l(e)$ is the label of the edge e):

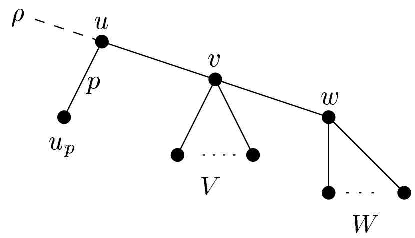

Figure 5: The vertices surrounding the edge (v, w).

$$\chi_{\mathcal{T}}(e) = \begin{cases} p(V)p(u)\phi^{-1}(v) \ \mid \ p(W)\phi^{-1}(w), & \text{if } l(v,w) \neq p, l(u,v) \neq p; \\ p(V)p(u_{q_0})\phi^{-1}(v) \ \mid \ p(W)\phi^{-1}(w), & \text{if } l(v,w) \neq p, l(u,v) = p; \\ q_0(V)q_0(u)\phi^{-1}(v) \ \mid \ q_0(W)\phi^{-1}(w), & \text{if } l(v,w) = p, l(u,v) \neq q_0; \\ q_0(V)q_0(u_p)\phi^{-1}(v) \ \mid \ q_0(W)\phi^{-1}(w), & \text{if } l(v,w) = p, l(u,v) = q_0, \end{cases}$$

where $AB|CD$ denotes the two-state character that induces the partition $\{A \cup B, C \cup D\}$. We denote the collection $\{\chi_{\mathcal{T}}(e) : e \in \mathcal{T}\}$ of two-state characters by $\mathcal{C}_{\mathcal{T}}$.

Continuing the example with $\mathcal{T}_1$ and the good labelling shown in Fig. 4, we have

$$\mathcal{C}_{\mathcal{T}_1} = \{a|bcdefg, b|acdefg, c|abdefg, f|abcdeg, abcdef|gj, deg|hij, gj|hi, j|ghi\},$$

where the associated edges are taken in order from top to bottom and left to right.

Lemma 2.2. *Let $\mathcal{T}$ be an X-tree, and suppose that we have a good labelling of $\mathcal{T}$. Then the set $\mathcal{C}_{\mathcal{T}}$ of characters identifies $\mathcal{T}$.*

Proof. The proof is by induction on the number of vertices of $\mathcal{T}$. If $\mathcal{T}$ consists of a single vertex, then the lemma holds trivially. Furthermore, if $\mathcal{T}$ consists of two vertices, then $\mathcal{T}$ has exactly one edge and it is clear that the single character in $\mathcal{C}_{\mathcal{T}}$ identifies $\mathcal{T}$.

Now suppose that the lemma holds for all X-trees with fewer vertices than $\mathcal{T}$ and suppose that $\mathcal{T} = (T; \phi)$ has n vertices, where $n \geq 3$. Under the good labelling of $\mathcal{T}$, let ℓ be a leaf of $\mathcal{T}$ that is at maximum distance from ρ. Let $\mathcal{T} - \ell$ be obtained from $\mathcal{T}$ by deleting ℓ and its incident edge. Let w be the parent vertex of ℓ in $\mathcal{T}$, and let v be the parent of w. Let W be the set of vertices that are children of w, including ℓ. Observe that W is a set of leaves, as ℓ is at maximum distance from ρ. Let $\mathcal{T}'$ be an X-tree that displays $\mathcal{C}_{\mathcal{T}}$. The proof is partitioned into three cases depending upon the structure and labelling of $\mathcal{T}$. In each case, we will show that $\mathcal{T}'$ refines $\mathcal{T}$, thus establishing the lemma.
Case 1. $\mathcal{T} - \ell$ is a semi-labelled tree.

Without loss of generality, choose ℓ so that the good labelling of $\mathcal{T}$ induces a good labelling of $\mathcal{T} - \ell$. Since $\mathcal{T} - \ell$ is a semi-labelled tree on $n-1$ vertices, it follows by the inductive hypothesis that the set $\mathcal{C}_{\mathcal{T}-\ell}$ of characters identifies $\mathcal{T} - \ell$. Comparing each edge e of $\mathcal{T} - \ell$ with its counterpart in $\mathcal{T}$, $\chi_{\mathcal{T}-\ell}(e)$ is a sub-character of $\chi_{\mathcal{T}}(e)$ (that is, $\chi_{\mathcal{T}-\ell}(e)$ can be obtained from $\chi_{\mathcal{T}}(e)$ by deleting the element $\phi^{-1}(\ell)$ if it occurs). Therefore $\mathcal{T}'$

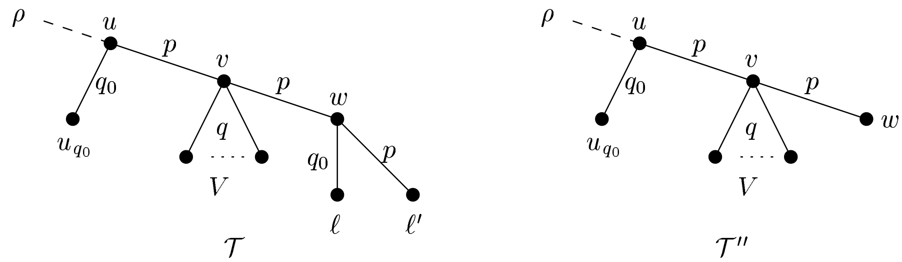

Figure 6: The structure of $\mathcal{T}$ and $\mathcal{T}''$ in Case 2. The labelling of $\mathcal{T}''$ is induced by $\mathcal{T}$, except for w which is unlabelled in $\mathcal{T}$, and labelled $\phi^{-1}(\ell')$ in $\mathcal{T}''$.

displays $\mathcal{C}_{\mathcal{T}-\ell}$, and so $\mathcal{T}'$ also displays $\mathcal{T}-\ell$. Since $\mathcal{T}'$ displays both $\mathcal{T}-\ell$ and $\chi_{\mathcal{T}}(v,w)$, this forces

$$\phi^{-1}(W \cup w) \mid (X - \phi^{-1}(W \cup w))$$

to be an X-split of $\mathcal{T}'$. For some $x \in [X - \phi^{-1}(W \cup w)]$, the character $\chi_{\mathcal{T}}(w,\ell)$ is

$$\phi^{-1}(\ell)|\phi^{-1}(W-\ell)\phi^{-1}(w)x.$$

It now follows that $\phi^{-1}(\ell)$ must label a subtree of $\mathcal{T}'$ which has no other labels. Since $\mathcal{T}'$ displays $\mathcal{T}-\ell$ and since $\phi^{-1}(\ell)$ can be contracted to a leaf on the correct side of the edge (v,w), we deduce that $\mathcal{T}'$ is a refinement of $\mathcal{T}$.

Case 2. $\mathcal{T}-\ell$ is not a semi-labelled tree, and the edge (v,w) is labelled p.

Since $\mathcal{T}-\ell$ is not a semi-labelled tree, it follows that W consists of ℓ and exactly one other leaf ℓ' say, and w is unlabelled. Without loss of generality, we may assume that the edge (w,ℓ) is labelled q_0 (otherwise, we consider ℓ'). Let $\mathcal{T}''$ be the tree obtained from $\mathcal{T}$ by deleting each of ℓ and ℓ' and their incident edges, and then labelling w by $\phi^{-1}(\ell')$. Comparing each edge e of $\mathcal{T}''$ with its counterpart in $\mathcal{T}$, $\chi_{\mathcal{T}''}(e) = \chi_{\mathcal{T}}(e)$ provided $\chi_{\mathcal{T}}(e)$ does not contain $\phi^{-1}(\ell)$. Since the edges (v,w) and (w,ℓ) are labelled p and q_0, respectively, $\chi_{\mathcal{T}}(e)$ contains $\phi^{-1}(\ell)$ only if e is incident with v and labelled p. Now the character $\chi_{\mathcal{T}''}(v,w)$ is the sub-character of $\chi_{\mathcal{T}}(v,w)$ obtained by deleting $\phi^{-1}(\ell)$. The only other edge incident with v and possibly labelled p is the edge (u,v) on the path from ρ to v (see Fig. 6). Suppose (u,v) is labelled p and that $\chi_{\mathcal{T}''}(u,v) = A|B\phi^{-1}(\ell')$. Then $\chi_{\mathcal{T}}(u,v) = A|B\phi^{-1}(\ell)$ and, for some $a \in A$, $\chi_{\mathcal{T}}(v,w) = aB|\phi^{-1}(\{\ell,\ell'\})$. (In fact, $a = q_0(u)$; recalling that if u has no edge labelled q_0, then it must be of degree two, and so $q_0(u) = \phi^{-1}(u)$.) Any X-tree displaying the latter two characters must also display the character $\chi_{\mathcal{T}''}(u,v)$. Hence, for all edges e in $\mathcal{T}''$, any X-tree displaying $\mathcal{C}_{\mathcal{T}}$ also displays $\chi_{\mathcal{T}''}(e)$. We conclude that $\mathcal{T}'$ must display $\mathcal{C}_{\mathcal{T}''}$, and therefore display $\mathcal{T}''$. A similar argument to that used in Case 1 now shows that $\mathcal{T}'$ is a refinement of $\mathcal{T}$.

Case 3. $\mathcal{T}-\ell$ is not a semi-labelled tree, and the edge (v,w) is not labelled p.

As in Case 2, W consists of ℓ and exactly one other leaf ℓ', and w is not labelled. Without loss of generality, we may assume that the edge (w,ℓ) is labelled p (otherwise, we consider ℓ'). Defining $\mathcal{T}''$ as in Case 2 and comparing each edge e of $\mathcal{T}''$ with its counterpart in $\mathcal{T}$, $\chi_{\mathcal{T}''}(e) = \chi_{\mathcal{T}}(e)$ provided $\chi_{\mathcal{T}}(e)$ does not contain $\phi^{-1}(\ell)$. Now $\chi_{\mathcal{T}}(e)$

contains $\phi^{-1}(\ell)$ only if e is incident with v and is not labelled p. Again, the character $\chi_{\mathcal{T}''}(v,w)$ is the sub-character of $\chi_{\mathcal{T}}(v,w)$ obtained by deleting $\phi^{-1}(\ell)$. Let e' be any other edge incident with v and not labelled p. Then, for some $A, B \subseteq X$, we have $\chi_{\mathcal{T}''}(e') = A|B\phi^{-1}(\ell')$. But then, $\chi_{\mathcal{T}}(e') = A|B\phi^{-1}(\ell)$ and $\chi_{\mathcal{T}}(v,w) = aB|\phi^{-1}(\{\ell,\ell'\})$ for some $a \in A$. (In this case, $a = p(u')$ where u' is the end-vertex of e' which is not v.) Any X-tree displaying the latter two characters must also display the character $\chi_{\mathcal{T}''}(e')$. As in the previous case, we again conclude that $\mathcal{T}'$ displays $\mathcal{C}_{\mathcal{T}''}$, and therefore displays $\mathcal{T}''$. Again, a similar argument to that used in Case 1 now shows that $\mathcal{T}'$ is a refinement of $\mathcal{T}$. This completes the proof of the lemma. □

Given an X-tree $\mathcal{T}$, Lemma 2.2 shows that there exists a set of $|E(\mathcal{T})|$ characters that identifies $\mathcal{T}$. In particular, $\mathcal{C}_{\mathcal{T}}$ is such a set. Using $\mathcal{C}_{\mathcal{T}}$, we next demonstrate a set $\mathcal{C}'_{\mathcal{T}}$ (consisting of at most $2s+2$ characters) that is displayed by $\mathcal{T}$ and has the property that any X-tree displaying $\mathcal{C}'_{\mathcal{T}}$ must also display $\mathcal{C}_{\mathcal{T}}$. Since $\mathcal{C}_{\mathcal{T}}$ identifies $\mathcal{T}$, it will follow that $\mathcal{C}'_{\mathcal{T}}$ also identifies $\mathcal{T}$. To define $\mathcal{C}'_{\mathcal{T}}$, we say that, for any $F \subseteq E(\mathcal{T})$, the *character associated with* F is the character induced by the graph $\mathcal{T} - F$. Starting with a good labelling of $\mathcal{T}$, let P^o (resp. P^e) be the set of edges of $\mathcal{T}$ labelled p that end at an odd (resp. even) distance from ρ. For $1 \le i \le s$, let Q_i^o (resp. Q_i^e) be the set of edges of $\mathcal{T}$ that end at an odd (resp. even) distance from ρ and are labelled by a set $q \in Q$ such that $i \in q$. Set $\mathcal{C}'_{\mathcal{T}}$ to be the union of the characters associated with P^o, P^e, Q_i^o, and Q_i^e for $1 \le i \le s$. Since each of these characters are induced by subgraphs of $\mathcal{T}$, it is immediate that $\mathcal{T}$ displays $\mathcal{C}'_{\mathcal{T}}$.

In the ongoing example, the set $\mathcal{C}'_{\mathcal{T}_1}$ consists of $P^e = a|bcdefg|hij$, $P^o = abcdef|gj|hi$, $Q_1^e = adefghi|b|c|j$, $Q_2^o = acdeghi|b|f|j$, and $Q_3^e = abdeghij|c|f$. The characters Q_1^o, Q_2^o, and Q_3^o are *null* characters in this case, that is they contain a single block and any tree therefore displays them.

Lemma 2.3. *Let $\mathcal{T}$ be an X-tree, and suppose that we have a good labelling of $\mathcal{T}$. Let $\mathcal{C}'_{\mathcal{T}}$ be the set of characters induced by P^o, P^e, Q_i^o, and Q_i^e for $1 \le i \le s$. If $\mathcal{T}'$ is an X-tree that displays $\mathcal{C}'_{\mathcal{T}}$, then $\mathcal{T}'$ also displays $\mathcal{C}_{\mathcal{T}}$.*

Proof. Let $e = (v,w)$ be an edge of $\mathcal{T}$, such that v is on the path from ρ to w. Let u be the parent vertex of v, and V be the set of children of v not including w. Let W be the set of children of w. Lastly, let u_p be the child of u such that (u,u_p) is labelled p, and u_{q_0} be the child of u such that (u,u_{q_0}) is labelled q_0.

We establish the lemma by showing that each of the characters in $\mathcal{C}_{\mathcal{T}}$ is displayed by any X-tree that displays $\mathcal{C}'_{\mathcal{T}}$. If $l(v,w) = p$, then it follows from the definition that $\chi_{\mathcal{T}}(e)$ is a sub-character of either the character associated with P^o or the character associated with P^e.

Suppose that $l(v,w) \neq p$. There are two cases to consider: (i) $l(u,v) \neq p$ or (ii) $l(u,v) = p$. Furthermore, each of these cases is divided into two sub-cases depending upon whether w is at an odd or even distance from ρ.

To prove (i), first assume that $l(u,v) \neq p$ and w is at an odd distance from ρ. Here

$$\chi_{\mathcal{T}}(e) = p(V)p(u)\phi^{-1}(v) \mid p(W)\phi^{-1}(w).$$

Let v_p be the vertex in V such that the edge (v, v_p) is labelled p. Let q_w be the label of (v, w) and, for each $v' \in V - v_p$, let $q_{v'}$ be the label of (v, v'). Since $q_w - q_{v'} \neq \emptyset$, there exists an element, $i_{v'}$ say, in $q_w - q_{v'}$. For each v', the character associated with $Q^o_{i_{v'}}$ has the sub-character

$$\chi_{v'} = p(v')p(v_p)p(u)\phi^{-1}(v) \mid p(W)\phi^{-1}(w).$$

A routine check now shows that any tree displaying the set of characters $\{\chi_{v'} : v' \in V\}$ must also display $\chi_{\mathcal{T}}(e)$. Hence any X-tree that displays $\mathcal{C}'_{\mathcal{T}}$ also displays the character $\chi_{\mathcal{T}}(e)$. The sub-case that w is an even distance from ρ is proved similarly.

Case (ii) is established using a similar argument to that used in (i). The details are omitted. This completes the proof of the lemma. □

Theorem 1.2(i) now follows from Lemmas 2.2 and 2.3.

Proof of Theorem 1.2(i). Let $\mathcal{T}$ be an X-tree with maximum vertex degree d. Let s be an integer such that $\binom{s}{\lceil \frac{s}{2} \rceil} \geq d - 2$. Then, by Lemmas 2.2 and 2.3, there exists a set of $2s + 2$ characters that identifies $\mathcal{T}$.

Part (i) of Theorem 1.2 is now established by setting $t = \lceil \log_2(d - 2) \rceil$, noting that

$$\binom{2t+1}{t+1} = \frac{(2t+1) \cdot 2t \cdot (2t-1) \cdots t}{(t+1) \cdot t \cdot (t-1) \cdots 1} > 2^t \geq d - 2,$$

and substituting $s = 2t + 1$. This completes the proof of Theorem 1.2. □

References

[1] Bordewich, M., Huber, K., Semple, C.: Identifying phylogenetic trees. Discrete Mathematics 300, 30-43 (2005).

[2] Daniel, P., Semple, C.: Supertree algorithms for nested taxa. In: O. Bininda-Emonds: Phylogenetic supertrees: combining information to reveal the Tree of Life (Computational Biology Series) Dordrecht: Kluwer, pp. 151-171, 2004.

[3] Huber, K., Moulton, V., Steel, M.: Four characters suffice to convexly define a phylogenetic tree. SIAM Journal on Discrete Mathematics 18, 835-843 (2005).

[4] Page, R. D. M.: Taxonomy, Supertrees, and the Tree of Life. In: O. Bininda-Emonds: Phylogenetic supertrees: combining information to reveal the Tree of Life (Computational Biology Series) Dordrecht: Kluwer, pp. 247-265, 2004.

[5] Semple, C., Steel, M.: Tree reconstruction from multi-state characters. Advances in Applied Mathematics 28, 169-184 (2002).

[6] Semple, C., Steel, M.: Phylogenetics, Oxford University Press, 2003.