SIAM J. DISCRETE MATH.
Vol. 18, No. 4, pp. 835–843

FOUR CHARACTERS SUFFICE TO CONVEXLY DEFINE A PHYLOGENETIC TREE*

KATHARINA T. HUBER†, VINCENT MOULTON†, AND MIKE STEEL‡

Abstract. It was recently shown that just five characters (functions on a finite set X) suffice to convexly define a trivalent tree with leaf set X. Here we show that four characters suffice which, since three characters are not enough in general, is the best possible.

Key words. phylogenetic tree, X-tree, convexly define, display, semidyadic closure, character compatibility

AMS subject classifications. 92B15, 92B10, 05C05

DOI. 10.1137/S0895480102416696

1. Introduction.

The field of *phylogenetics* compares observable characteristics of (biological) species in order to reconstruct and analyze their evolutionary history. Generally this history is represented by a tree, with leaves labeled by the species. If each of the comparisons between the species involve just two possible character states (for example, "wings" vs. "no-wings") and each state has evolved only once, then there is a direct equivalence between such data and leaf-labeled trees. This equivalence was described by Peter Buneman in his classic paper [4] from 1971. More recently there has been considerable interest, from both computer scientists and mathematicians, in extending these results to data in which there may be many character states—so-called "multistate characters" [1], [7], [8], [10]. Recent whole genome data has given rise to extensive data sets of multistate characters, often with a large number of states (such as those obtained by comparing gene order between species).

This leads to the natural question of how many multistate characters are required to completely determine an underlying evolutionary tree, under the assumption that each state has evolved just once. In a surprising result, the authors of [10] recently showed that just *five* such characters suffice, regardless of the number of species (we describe this result more precisely in section 5). Their result applied a graph-theoretic argument involving chordal graphs to a specific edge-coloring of trees based on the cyclic group of order 5. However, the tantalizing question of whether this five character result could be improved to four characters was left as a posed problem [10, Problem 6.2], as the methods used in that paper did not seem to readily apply.

In this paper we employ a different approach, and show that four characters are indeed sufficient, a result that is optimal since three characters are not sufficient to completely define all trees [10]. We reproduce the tree topology in [10] that illustrates four as lower bound for Figure 1. In particular, we describe an edge-coloring of a tree

*Received by the editors October 29, 2002; accepted for publication (in revised form) August 31, 2004; published electronically May 20, 2005. This research was supported by The Swedish Foundation for International Cooperation in Research and Education (STINT).

http://www.siam.org/journals/sidma/18-4/41669.html

†School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK (katharina.huber@cmp.uea.ac.uk, vincent.moulton@cmp.uea.ac.uk). The research of these authors presented in this paper was supported by The Linnaeus Centre for Bioinformatics, Uppsala University, Sweden, where the results were established, and The Swedish Research Council (VR).

‡Biomathematics Research Centre, University of Canterbury, Box 4800, Christchurch, New Zealand (m.steel@math.canterbury.ac.nz). The research of this author was supported by the New Zealand Marsden Fund.

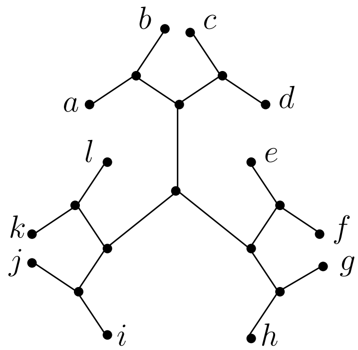

FIG. 1. *The figure depicts an evolutionary tree on the set* $X = \{a, b, \ldots, l\}$ *that cannot be convexly defined by three characters (see* [10] *for more details).*

using four colors, which induces characters in the same way as the edge coloration using five colors in [10]. To establish that the induced characters can be used to completely reconstruct the tree, we consider a set of small subtrees (each with four leaves) that are generated by the edge-coloring, and show that these subtrees determine the tree. This then allows us to establish that the characters induced by the edge-coloring determine the underlying tree.

The structure of this paper is as follows. In section 2, we introduce some terminology for trees and describe a closure operation on subtrees. Next, in section 3, we describe an edge-coloring of trees that produces subtrees on which this closure operation is applied. In section 4, we establish our main technical tool (Theorem 4.1), and in section 5, we use this result to show that four characters suffice to completely reconstruct a tree (Theorem 5.2).

2. Quartet trees and semidyadic closure. Throughout the paper, X denotes a nonempty finite set and $n = |X|$. A *phylogenetic tree (on X)* is a tree $\mathcal{T}$ that has X as its set of labeled leaves and interior vertices that are unlabeled and of degree at least three. If each interior vertex has degree exactly three, we say that $\mathcal{T}$ is *trivalent*. Two phylogenetic trees for X are *isomorphic* if the identity map on X induces a graph isomorphism on the underlying tree.

A (qualitative or discrete) *character on X* is a function χ from X into a set C of *character states*. Suppose that $\mathcal{T}$ is a phylogenetic tree on X, and let $\chi : X \to C$ be a character on X. For each state α in $\chi(X)$, let $\mathcal{T}_\alpha$ denote the minimal subtree of $\mathcal{T}$ containing the leaves that are assigned state α by χ. We say that χ is *convex on $\mathcal{T}$* if the subtrees in $\{\mathcal{T}_\alpha \mid \alpha \in \chi(X)\}$ are pairwise disjoint (see Figure 2). A collection of characters $\mathcal{C}$ on X is *compatible* if there is a phylogenetic tree $\mathcal{T}$ such that each character in $\mathcal{C}$ is convex on $\mathcal{T}$. If, in addition, $\mathcal{T}$ is the only phylogenetic tree on X with this property, we say that $\mathcal{C}$ *convexly defines* $\mathcal{T}$. The biological relevance of these concepts is explained further in [10] and [11].

We call a trivalent phylogenetic tree on a 4-set a *quartet tree*. If $\mathcal{T}$ is a quartet tree on the set $\{i, j, k, l\}$ and removal of the interior edge e of $\mathcal{T}$ results in the sets $\{i, j\}$ and $\{k, l\}$ labeling the different components of $\mathcal{T}\backslash\{e\}$, then we denote $\mathcal{T}$ by $ij|kl$. Now, given a phylogenetic tree $\mathcal{T}$ on X and a subset Y of X, let $\mathcal{T}|Y$ denote the minimal subtree of $\mathcal{T}$ that connects the leaves in Y, in which any resulting degree 2 vertices are suppressed. In particular, $\mathcal{T}|Y$ is a trivalent phylogenetic tree on Y and we say that $\mathcal{T}$ *displays* $\mathcal{T}|Y$. Given a collection $\mathcal{Q}$ of quartet trees, we say that a

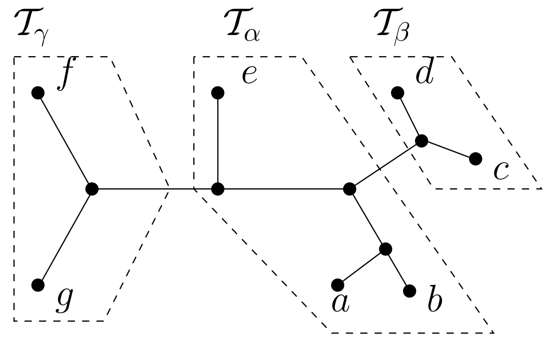

FIG. 2. *For $X = \{a, b, c, d, e, f, g\}$ and $C = \{\alpha, \beta, \gamma\}$, the character $\chi : X \to C$ with $\chi^{-1}(\alpha) = \{a, b, e\}$, $\chi^{-1}(\beta) = \{c, d\}$ and $\chi^{-1}(\gamma) = \{f, g\}$ is convex on the phylogenetic tree depicted in the figure.*

phylogenetic tree $\mathcal{T}$ *displays* $\mathcal{Q}$ precisely if $\mathcal{T}$ displays each quartet tree in $\mathcal{Q}$. For a trivalent phylogenetic tree $\mathcal{T}$ on X, let $\mathcal{Q}(\mathcal{T}) = \{\mathcal{T}|Y : Y \subseteq X, |Y| = 4\}$ be the set of all $\binom{n}{4}$ quartet trees displayed by $\mathcal{T}$.

For $\mathcal{Q}$ a set of quartet trees, let $\mathrm{scl}_2(\mathcal{Q})$ be the *semidyadic closure* of $\mathcal{Q}$, that is, the minimal set of quartet trees that contains $\mathcal{Q}$ and for which we have

$$ab|cd, ac|de \in \mathrm{scl}_2(\mathcal{Q}) \Rightarrow ab|ce, ab|de, bc|de \in \mathrm{scl}_2(\mathcal{Q}).$$

The following lemma summarizes some straightforward properties of the semidyadic closure that are part of the folklore (see [2], [5], [6], and [12]).

LEMMA 2.1. *For any set $\mathcal{Q}$ of quartet trees and any subsets $A, B \subseteq \mathcal{Q}$,*

(i) $A \subseteq \mathrm{scl}_2(A)$,
(ii) $A \subseteq B \Rightarrow \mathrm{scl}_2(A) \subseteq \mathrm{scl}_2(B)$,
(iii) $\mathrm{scl}_2(\mathrm{scl}_2(A)) = \mathrm{scl}_2(A)$,
(iv) $\mathrm{scl}_2(A \cup B) = \mathrm{scl}_2(\mathrm{scl}_2(A) \cup B)$.
(v) *If $\mathcal{Q} = \mathcal{Q}(\mathcal{T})$ for some trivalent phylogenetic tree $\mathcal{T}$, then* $\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q}$.

We recall one further useful property of the semidyadic closure that will be of use later. Suppose i, j is a *cherry* (a pair of leaves that are adjacent to a common vertex) of a trivalent phylogenetic $\mathcal{T}$ and select leaves u, v as shown in Figure 3(a). Let $\mathcal{T}' = \mathcal{T}|(X - \{j\})$ be the tree as shown in Figure 3(b). Then $\mathcal{T}$ is the only phylogenetic tree that displays both $\mathcal{T}'$ and $ij|uv$ and so, by [3, Lemma 3], we have the following result.

LEMMA 2.2. *For a trivalent phylogenetic tree $\mathcal{T}'$ and quartet tree $ij|uv$ as described,*

$$\mathrm{scl}_2(\mathcal{Q}(\mathcal{T}') \cup \{ij|uv\}) = \mathcal{Q}(\mathcal{T}).$$

For a set $\mathcal{Q}$ of quartet trees let $\mathrm{co}(\mathcal{Q})$ be the set of phylogenetic trees on X (up to isomorphism) that display $\mathcal{Q}$. We close this section with a lemma that summarizes an easily established property of $\mathrm{co}(\mathcal{Q})$.

LEMMA 2.3. *If $\mathcal{Q}$ is a set of quartet trees and* $\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$ *for some trivalent phylogenetic tree $\mathcal{T}$, then* $\mathrm{co}(\mathcal{Q}) = \{\mathcal{T}\}$.

3. Quartet trees from handy edge-colorings. An *edge-coloring* of a graph is an assignment of colors to the edges of the graph so that two adjacent edges are assigned different colors. We begin this section by giving a method for edge-coloring

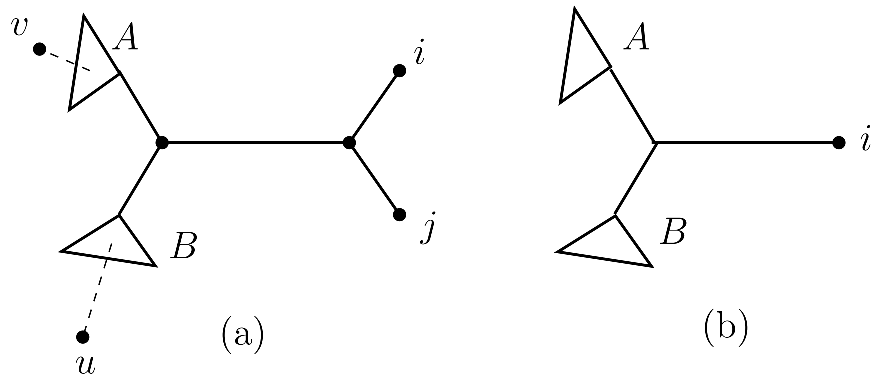

FIG. 3.

a trivalent phylogentic tree $\mathcal{T}$ on X with four colors R, R', L, L'. This edge-coloring is similar to the edge-coloring in [10] based on five colors.

Choose any leaf r of $\mathcal{T}$ and regard $\mathcal{T}$ as a rooted directed tree with r as its root and all edges directed away from r. Color the edge containing r by R. Given any vertex v of $\mathcal{T}$ with degree 3 that is at the end of an even (respectively, odd) length edge path starting at r and ending at v, arbitrarily color the two edges coming out of v by L and R (respectively, L' and R'). This gives an edge-coloring of $\mathcal{T}$ by the colors R, R', L, L', and we call any edge-coloring produced in this way a *handy edge-coloring* of $\mathcal{T}$.

Now, given a handy edge-coloring of $\mathcal{T}$, we describe how to associate a quartet tree with leaves in X to each interior edge of $\mathcal{T}$ (see Figure 4). Assume $e = (u, v)$ is an interior edge of $\mathcal{T}$ colored by R (we will consider the cases where e is colored by L, R', or L' below). The edge coming into u is colored by either (i) R' or (ii) L'. In case (i), we associate the quartet tree $ab|cd$ to edge e as follows: a is the last vertex in the directed path that starts at v and has first edge colored R' and all subsequent edges colored alternately by L and L'; b is the last vertex of the directed path that starts at v and has edges colored alternately by L' and L; c is the last vertex of the directed path that starts at u and has edges colored alternately by L and L'; d is the last vertex of the undirected path that starts at u and has first edge colored R' and all subsequent edges colored alternately by L' and L. In case (ii), a, b, c are all obtained in the same way and d is the last vertex of the undirected path that starts at u, has first two edges colored L' and R', respectively, and has all subsequent edges colored alternately by L and L'.

In case the edge $e = (u, v)$ is labeled by R', the quartet tree $ab|cd$ is obtained in a similar way by following the four distinct paths whose first vertices are either u or v and whose last edges are alternately colored using only the colors L and L'. In case the edge $e = (u, v)$ is labeled by either L or L', a similar procedure is followed in which colors L and R and L' and R' are interchanged so that, in particular, the quartet tree $ab|cd$ is obtained by following the four distinct paths whose first vertices are either u or v and whose last edges are alternately colored using only the colors R and R'.

We denote the collection of $n - 3$ quartet trees obtained in this way by $\mathcal{Q}_0(\mathcal{T})$. Note that in all cases the paths obtained are colored always by at most three colors. Whenever we picture a phylogenetic tree with a handy edge-coloring, we always regard

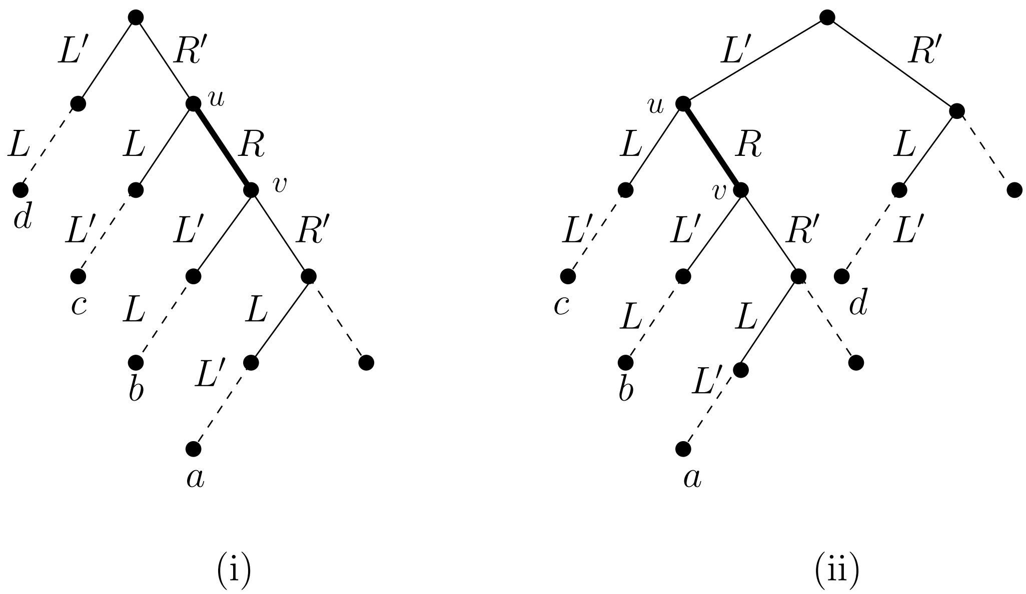

FIG. 4. *The figure depicts the two cases for associating a quartet tree $ab|cd$ to an interior edge e of $\mathcal{T}$, here in bold, that is labeled by R.*

edges below a particular vertex to be colored with R or R' when they are on the right or L or L' when they are on the left.

4. $\mathcal{Q}_0(\mathcal{T})$ determines $\mathcal{Q}(\mathcal{T})$ via semidyadic closure. Suppose that $\mathcal{T}$ is a trivalent phylogenetic tree on X with a handy edge-coloring. In the next section we describe (at most) four characters that convexly define $\mathcal{T}$ and come from the handy edge-coloring of $\mathcal{T}$. The proof that these four characters convexly define $\mathcal{T}$ is based on the following result.

THEOREM 4.1. *Suppose that $\mathcal{T}$ is a trivalent phylogenetic tree on X. Then*

$$\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})) = \mathcal{Q}(\mathcal{T}).$$

Proof. We use induction on n. It is easily checked that the result holds when $n = 4$, since in this case $\mathcal{Q}_0(\mathcal{T}) = \mathcal{Q}(\mathcal{T}) = \{\mathcal{T}\}$.

Suppose the theorem holds for any trivalent phylogenetic tree on X with strictly less than $n \geq 5$ leaves. Suppose also that $\mathcal{T}$ is a trivalent phylogenetic tree on X with n leaves. Select a cherry i, j whose central vertex is at maximal edge distance from the reference leaf. If we now consider the handy edge-coloring of $\mathcal{T}$, then there are four cases (plus their mirror images) for the local tree structure around the cherry i, j, as depicted in Figure 5.

Note that in case (b) we could have instead selected the cherry k, l and this produces (the mirror image of) case (a) so we can "transform" case (b) into (a). It thus suffices to consider only cases (a), (c), and (d). For these cases, let $\mathcal{T}' = \mathcal{T}|(X - \{j\})$. Note that the edge-coloring of $\mathcal{T}$ induces a valid handy edge-coloring of $\mathcal{T}'$, where the color assigned to the edge containing i is the same as that assigned to the edge in $\mathcal{T}$ adjacent to the cherry i, j.

First consider cases (a) and (c). It is straightforward to check using the definition of a handy edge-coloring that the only interior edge of $\mathcal{T}$ yielding a quartet tree in

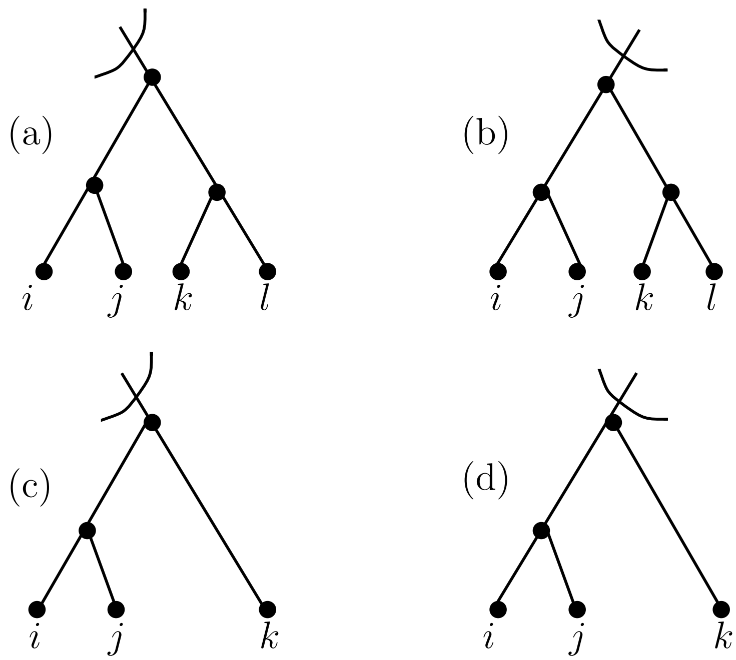

FIG. 5.

$\mathcal{Q}_0(\mathcal{T})$ that contains j is the interior edge that is adjacent to the cherry i, j. Moreover, every interior edge of $\mathcal{T}'$ corresponds to an interior edge of $\mathcal{T}$ and each of these edges gives rise to the same quartet tree in $\mathcal{Q}_0(\mathcal{T}')$ as it does in $\mathcal{Q}_0(\mathcal{T})$. From these observations it easily follows that

$$\mathcal{Q}_0(\mathcal{T}') = \mathcal{Q}_0(\mathcal{T}) - \{ij|kx\} \tag{1}$$

for some $x \in X$ (and with $x \neq l$ in case (a)).

Now, by the induction hypothesis applied to $\mathcal{T}'$,

$$\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T}')) = \mathcal{Q}(\mathcal{T}')$$

and by Lemma 2.1 (iv) and Lemma 2.2,

$$\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T}') \cup \{ij|kx\}) = \mathcal{Q}(\mathcal{T}).$$

Thus, by (1),

$$\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})) = \mathcal{Q}(\mathcal{T}),$$

and so the induction step is established for cases (a) and (c).

Thus it suffices to consider now just case (d). The edge e coming into the cherry i, j induces the quartet tree $ij|ku \in \mathcal{Q}_0(\mathcal{T})$ and the edge e' incident to e but not containing k induces the quartet tree $jk|uv \in \mathcal{Q}_0(\mathcal{T})$, for some pair of leaves $u, v \in X$ (see Figure 6).

Thus,

$$\mathrm{scl}_2(\{ij|ku, jk|uv\}) \subseteq \mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})).$$

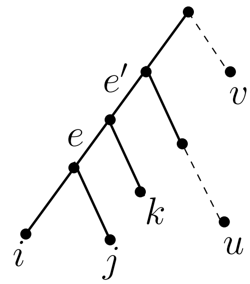

FIG. 6.

But $ik|uv \in \mathrm{scl}_2(\{ij|ku, jk|uv\})$ and so

$$ik|uv \in \mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})). \tag{2}$$

Now, it is straightforward to check using the definition of a handy edge-coloring that the only interior edges of $\mathcal{T}$ yielding quartet trees in $\mathcal{Q}_0(\mathcal{T})$ that contain j are the edges e and e'. Moreover, every interior edge of $\mathcal{T}'$ corresponds to an interior edge of $\mathcal{T}$ and each of these gives rise to the same quartet tree in $\mathcal{Q}_0(\mathcal{T}')$ as it does in $\mathcal{Q}_0(\mathcal{T})$ except e', which gives rise to $ik|uv$ in $\mathcal{Q}_0(\mathcal{T}')$. From these observations it easily follows that

$$\mathcal{Q}_0(\mathcal{T}') = (\mathcal{Q}_0(\mathcal{T}) - \{ij|ku, jk|uv\}) \cup \{ik|uv\}. \tag{3}$$

Combining (2), (3), and Lemma 2.1 (parts (i), (ii), and (iii)) we have

$$\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T}') \cup \{ik|uv\}) \subseteq \mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})). \tag{4}$$

On the other hand, if we apply Lemma 2.2, the induction hypothesis for $\mathcal{T}'$, and Lemma 2.1 (iv), we obtain (respectively) the following three equalities:

$$\begin{aligned}\mathcal{Q}(\mathcal{T}) &= \mathrm{scl}_2(\mathcal{Q}(\mathcal{T}') \cup \{ik|uv\})\\ &= \mathrm{scl}_2(\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T}')) \cup \{ik|uv\})\\ &= \mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T}') \cup \{ik|uv\}).\end{aligned}$$

Combining these equalities with (4) gives

$$\mathcal{Q}(\mathcal{T}) \subseteq \mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})).$$

However, this implies $\mathcal{Q}(\mathcal{T}) = \mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T}))$ in view of $\mathcal{Q}_0(\mathcal{T}) \subseteq \mathcal{Q}(\mathcal{T})$ and using Lemma 2.1 (parts (ii) and (v)). This establishes the induction step and thereby completes the proof of Theorem 4.1. $\square$

5. Handy edge-colorings convexly define trees. We now relate characters and quartet trees. Given a character $\chi : X \to C$ on X, we denote by $\pi(\chi)$ the partition $\{\chi^{-1}(\alpha) : \alpha \in C\}$ of X. Suppose that $\mathcal{T}$ is a phylogenetic tree on X and that $\mathcal{C}$ is a set of characters on X. We say that $\mathcal{T}$ *displays* $\mathcal{C}$ if each character in $\mathcal{C}$ is convex on $\mathcal{T}$. Note that $\mathcal{T}$ displays $\mathcal{C}$ precisely if for each $\chi \in \mathcal{C}$ there exists some set $\mathcal{E}$ of edges of $\mathcal{T}$ such that, for all distinct $A, B \in \pi(\chi)$, A and B are subsets of different connected components of $\mathcal{T} \backslash \mathcal{E}$.

For any collection $\mathcal{C}$ of characters on X, let

$$\mathcal{Q}(\mathcal{C}) = \{ij|kl : \text{ there exists some } \chi \in \mathcal{C} \text{ and some } A, B \in \pi(\chi) \text{ such that } i, j \in A \text{ and } k, l \in B\}.$$

LEMMA 5.1. *Let $\mathcal{C}$ be a collection of characters on X, and suppose that $\mathcal{T}$ is a trivalent phylogenetic tree that displays $\mathcal{C}$. If there exists some $\mathcal{Q}_1 \subseteq \mathcal{Q}(\mathcal{C})$ with $\mathrm{scl}_2(\mathcal{Q}_1) = \mathcal{Q}(\mathcal{T})$, then $\mathcal{C}$ convexly defines $\mathcal{T}$.*

Proof. Note that Lemma 2.1 (ii) gives $\mathrm{scl}_2(\mathcal{Q}_1) \subseteq \mathrm{scl}_2(\mathcal{Q}(\mathcal{C}))$. Thus,

$$\mathcal{Q}(\mathcal{T}) \subseteq \mathrm{scl}_2(\mathcal{Q}(\mathcal{C})). \tag{5}$$

On the other hand, since each character in $\mathcal{C}$ is convex on $\mathcal{T}$, we have $\mathcal{Q}(\mathcal{C}) \subseteq \mathcal{Q}(\mathcal{T})$ and so

$$\mathrm{scl}_2(\mathcal{Q}(\mathcal{C})) \subseteq \mathcal{Q}(\mathcal{T}), \tag{6}$$

by Lemma 2.1 (parts (ii), (iii), and (v)). Combining (5) and (6) gives $\mathrm{scl}_2(\mathcal{Q}(\mathcal{C})) = \mathcal{Q}(\mathcal{T})$, and so, by Lemma 2.3 we have $\mathrm{co}(\mathcal{Q}(\mathcal{C})) = \{\mathcal{T}\}$. But from [12, Proposition 2(1)], if $\mathrm{co}(\mathcal{Q}(\mathcal{C})) = \{\mathcal{T}\}$, then $\mathcal{C}$ convexly defines $\mathcal{T}$. This completes the proof. □

We now specialize to a set of (at most four) characters that are induced by any handy edge-coloring of a trivalent phylogenetic tree $\mathcal{T}$ on X and show that these characters convexly define $\mathcal{T}$.

Suppose that we are given a handy edge-coloring of $\mathcal{T}$. To each color $F \in \{L, L', R, R'\}$ that is assigned to at least one edge of $\mathcal{T}$, we associate a character on X in the following way. Denote by $\sim_F$ the equivalence relation on X defined by $x \sim_F y$ $(x, y \in X)$ if the path in $\mathcal{T}$ from x to y does not contain an edge that is assigned color F. Let π_F denote the partition of X that arises from the equivalence classes of $\sim_F$ and let χ_F denote the character on X for which $\pi(\chi_F) = \pi_F$. We denote by $\mathcal{C}(\mathcal{T})$ the (at most) four characters induced by this edge-coloring.

The main result from [10] is that, for any trivalent phylogenetic tree $\mathcal{T}$ on X, there exists a set $\mathcal{C}$ of at most five characters on X, such that $\mathcal{T}$ is the only phylogenetic tree on X that displays $\mathcal{C}$. The following theorem shows that, by taking $\mathcal{C} = \mathcal{C}(\mathcal{T})$, we can improve the result by replacing "five" with "four."

THEOREM 5.2. *Suppose that $\mathcal{T}$ is a trivalent phylogenetic tree on X. Then the (at most) four characters in $\mathcal{C}(\mathcal{T})$ convexly define $\mathcal{T}$.*

Proof. First note that each character in $\mathcal{C}(\mathcal{T})$ is convex on $\mathcal{T}$. Note also that since $\mathcal{Q}_0(\mathcal{T})$ is the set of quartet trees corresponding to the handy edge-coloring of $\mathcal{T}$, we have

$$\mathcal{Q}_0(\mathcal{T}) \subseteq \mathcal{Q}(\mathcal{C}(\mathcal{T})).$$

Also, by Theorem 4.1, $\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})) = \mathcal{Q}(\mathcal{T})$. Thus, since $\mathcal{T}$ displays $\mathcal{C}(\mathcal{T})$ we may apply Lemma 5.1 to deduce that $\mathcal{C}(\mathcal{T})$ convexly defines $\mathcal{T}$. □

Note that the proof of this result shows how to construct $\mathcal{T}$ from $\mathcal{C}(\mathcal{T})$ in polynomial time using the semidyadic closure operation. Alternatively, since $|\mathcal{Q}_0(\mathcal{T})| = |X| - 3$ the "split-closure" approach described by Semple and Steel [9] would also apply. It can also be shown that $\mathcal{C}(\mathcal{T})$ "strongly" defines $\mathcal{T}$ in the sense of [10].

Acknowledgment. The authors thank Charles Semple for some helpful comments on an earlier version of this manuscript.

REFERENCES

[1] R. Agarwala and D. Fernández-Baca, *A polynomial-time algorithm for the phylogeny problem when the number of character states is fixed*, SIAM J. Comput., 23 (1994), pp. 1216–1224.

[2] H.-J. Bandelt and A. W. M. Dress, *Reconstructing the shape of a tree from observed dissimilarity data*, Adv. Appl. Math., 7 (1986), pp. 309–343.

[3] S. Böcker, D. Bryant, A. W. M. Dress, and M. A. Steel, *Algorithmic aspects of tree amalgamation*, J. Algorithms, 37 (2000), pp. 522–537.

[4] P. Buneman, *The recovery of trees from measures of dissimilarity*, in Mathematics in the Archaeological and Historical Sciences, F. R. Hodson, D. G. Kendall, and P. Tautu, eds., Edinburgh University Press, Edinburgh, 1971, pp. 387–395.

[5] H. Colonius and H. H. Schulze, *Tree structures for proximity data*, British J. Math. Statist. Psych., 34 (1981), pp. 167–180.

[6] M. C. H. Dekker, *Reconstruction Methods for Derivation Trees*, unpublished Masters thesis, Vrije Universiteit, Amsterdam, Netherlands, 1986.

[7] S. Kannan and T. Warnow, *A fast algorithm for the computation and enumeration of perfect phylogenies*, SIAM J. Comput., 26 (1997), pp. 1749–1763.

[8] F. R. McMorris, T. J. Warnow, and T. Wimer, *Triangulating vertex-colored graphs*, SIAM J. Discrete Math., 7 (1994), pp. 296–306.

[9] C. Semple and M. Steel, *Tree reconstruction via a closure operation on partial splits*, in Proceedings of journées ouvertes: Biologie, informatique et mathématique, Lecture Notes in Comput. Sci. 2066, O. Gascuel and M.-F. Sagot, eds., Springer-Verlag, Berlin, 2001, pp. 126–134.

[10] C. Semple and M. Steel, *Tree reconstruction from multi-state characters*, Adv. Appl. Math., 28 (2002), pp. 169–184.

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

[12] M. Steel, *The complexity of reconstructing trees from qualitative characters and subtrees*, J. Classification, 9 (1992), pp. 91–116.