FOUR CHARACTERS SUFFICE

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.

Il a été récemment montré que seulement cinq caractéres (des fonctions sur un ensemble fini X) suffisent pour définir de maniere convexe un arbre binair avec X comme ensemble de feuilles. Dans cet article nous montrerons que la meilleure solution possible est quatre caractéres (étant donné que trois caractéres, en general, ne sont pas suffisant).

Keywords: phylogenetic tree, convexly define, semi-dyadic closure, character compatibility

1. INTRODUCTION

The field of *phylogenetics* compares observable characteristics of (biological) species in order to reconstruct and analyse their evolutionary history. Generally this history is represented by a tree, with leaves labelled by the species. If each of the comparisons between the species involves 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-labelled trees. This equivalence was described by Peter Buneman in his classic paper [4] from (1971). More recently there has been considerable interest, both from computer scientists and mathematicians, in extending these results to data in which there may be many character states - so called 'multi-state characters' [1, 7, 8, 9]. Recent whole genome data has given rise to extensive data sets of multi-state 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 multi-state 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 [9] recently showed that just *five* such characters suffice, regardless of the number of species (we describe this result more precisely in Section 4). 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 tantalising question of whether this five character result could be improved to four characters was left as a posed problem [9, Problem 6.2], as the methods used in that paper did not seem to readily apply.

In this note we employ a different approach, and show that four characters are indeed sufficient, a result that is optimal since three characters do not, in general, suffice to determine a tree [9]. In particular, we describe an edge-coloring of a tree using four colors based on the Klein 4-group $\mathbb{Z}_2 \times \mathbb{Z}_2$, which induces characters in the same way as the edge coloration using five colors in [9]. 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 note is as follows. In Section 2, we introduce some terminology for characters and trees. Next, in Section 3 we describe an edge-coloring of a tree which will be used to produce a set of at most four characters which determine that tree. Finally, in Section 4, we state our main result that four characters suffice to completely reconstruct a tree (Theorem 1) and give a sketch of its proof.

2. Characters and Quartet Trees

Throughout the note, X denotes a non-empty finite set and $n = |X|$. A *phylogenetic tree (on X)* is a tree $\mathcal{T}$ that has X as its set of labelled leaves and interior vertices that are unlabelled 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 smallest 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 1). 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,

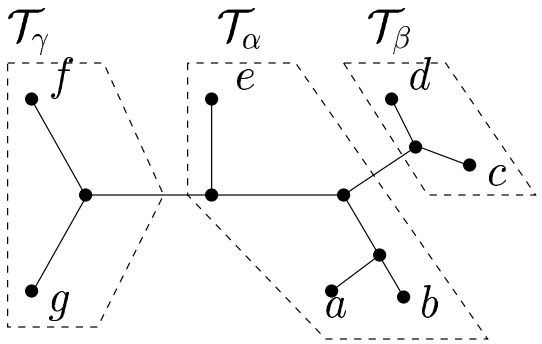

FIGURE 1. For $X = \{a, b, c, d, e, f, g\}$ and $C = \{\alpha, \beta, \gamma\}$ the character $\chi : X \to C$ with $\chi^{-1}(\alpha) = \{a, b, e\}$, $\chi^{-1}(\beta) = \{c, d\}$ and $\chi^{-1}(\gamma) = \{f, g\}$ is convex on the phylogenetic tree depicted in the figure.

$\mathcal{T}$ 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 [9] and [10].

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 $\{k, l\}$ labelling 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 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 $\mathcal{T}$.

We conclude this section by relating 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 components of $\mathcal{T}\backslash\mathcal{E}$. In addition, for any collection $\mathcal{C}$ of characters on X, let

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

3. Handy Edge-Colorings

Suppose that $\mathcal{T}$ is a trivalent phylogenetic tree on X. An *edge-coloring* of $\mathcal{T}$ is an assignment of colors to the edges of $\mathcal{T}$ so that two adjacent edges are assigned different colors. We now describe a method for edge-coloring a trivalent phylogenetic tree $\mathcal{T}$ on X with four colors R, R', L, L' that will be used later to define a set of at most four characters that convexly define $\mathcal{T}$. This edge-coloring can be viewed as being based on the Klein 4-group $\mathbb{Z}_2 \times \mathbb{Z}_2$ (somewhat analogous to the edge-coloring in [9] based on $\mathbb{Z}_5$) though it is more convenient to picture it in the way that we now describe.

Choose any leaf r of $\mathcal{T}$ and regard $\mathcal{T}$ as a rooted tree with r as its root. 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 path starting at r and ending at v, arbitrarily color the two edges incident with v and not contained in this path 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 $\mathcal{T}$. In Figure 2 we picture the tree in Figure 1 together with a handy edge-coloring.

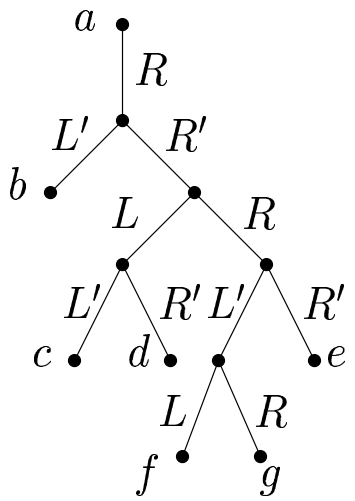

Figure 2. A tree with a handy edge-coloring.

Now, given a handy edge-coloring of $\mathcal{T}$, we describe how to associate a quartet tree in $\mathcal{Q}(\mathcal{T})$ to each interior edge of $\mathcal{T}$ (see Figure 3). Assume $e = uv$ is an interior edge of $\mathcal{T}$, where u, v are vertices of $\mathcal{T}$ and u is the vertex in e closest to r. Suppose that e is colored by R (we will consider the cases where e is colored by L, R' or L' below). Then the last edge of the path from r to u is colored either by (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 of the path starting at v that has first edge colored R' and all subsequent edges colored alternately by L and L'; b is the last vertex of the path that starts at v and has edges colored alternately by L' and L; c is the last vertex of the path that starts at u and has edges colored alternately by L and L'; d is the last vertex of the path that starts at u, 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 path that starts at u, has first two edges colored L' and R', respectively, and all subsequent edges colored alternately by L and L'.

In case the edge $e = uv$ 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 $e = uv$ 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 used to define the quartets are colored by at most three colors.

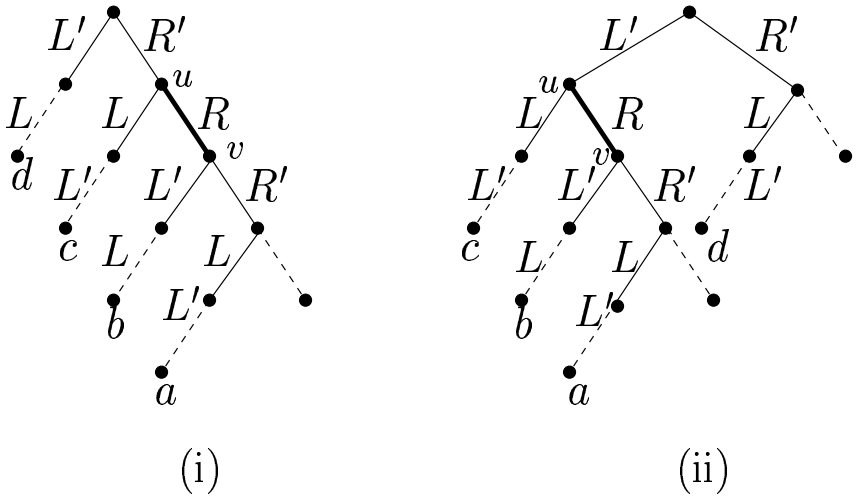

FIGURE 3. The figure depicts the two cases for associating a quartet tree $ab|cd$ to an interior edge e of $\mathcal{T}$, here in bold, that is labelled by R.

4. Handy Edge-Colorings Convexly Define Trees

Suppose that we are given a handy edge-coloring of a trivalent phylogenetic tree $\mathcal{T}$ on X. 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 a character on X for which $\pi(\chi_F) = \pi_F$. We denote by $\mathcal{C}(\mathcal{T})$ the set of (at most) 4 characters induced by this edge-coloring.

The main result from [9] states 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' by 'four'.

Theorem 1. *Suppose that $\mathcal{T}$ is a trivalent phylogenetic tree on X. Then the (at most) four characters in $\mathcal{C}(\mathcal{T})$ convexly define $\mathcal{T}$.*

We conclude this note by providing a sketch proof for this theorem. We require a new concept. For $\mathcal{Q}$ a set of quartet trees, let $\mathrm{scl}_2(\mathcal{Q})$ be the *semi-dyadic closure* of $\mathcal{Q}$ (cf. [3, 5, 6, 11]) 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 semi-dyadic closure gives us the following key link between characters and quartet-trees.

Lemma 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 $\mathcal{T}$.*

We now state our main technical tool.

Theorem 2. *Suppose that $\mathcal{T}$ is a trivalent phylogenetic tree on X with a handy edge-coloring. Then*

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

Now, note that each character in $\mathcal{C}(\mathcal{T})$ is convex on $\mathcal{T}$ and also

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

Thus, since $\mathrm{scl}_2(\mathcal{Q}_0(\mathcal{T})) = \mathcal{Q}(\mathcal{T})$, it follows immediately from Lemma 1 that $\mathcal{C}(\mathcal{T})$ convexly defines $\mathcal{T}$.

Acknowledgements K.T.H. and V.M. thank The Swedish Research Council (VR) and M.S. thanks the New Zealand Marsden Fund. All

authors thank The Swedish Foundation for International Cooperation in Research and Education (STINT).

References

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

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

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

[4] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the archaeological and historical sciences* (ed. F. R. Hodson, D. G. Kendall, and P. Tautu), pp.387–395. Edinburgh University Press.

[5] Colonius, H. and Schulze, H. H. (1981). Tree structures for proximity data. *British Journal of Mathematical and Statistical Psychology*, **34**, 167–180.

[6] Dekker, M. C. H. (1986). Reconstruction methods for derivation trees. Unpublished Masters thesis. Vrije Universiteit, Amsterdam, Netherlands.

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

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

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

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

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

Department of Biometry and Informatics, Swedish University of Agricultural Sciences, Box 7013, 750 07 Uppsala, Sweden, and, The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden

E-mail address: katharina.huber@bi.slu.se

Communicating author; The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden

E-mail address: vincent.moulton@lcb.uu.se

Biomathematics Research Centre, University of Canterbury, Box 4800, Christchurch, New Zealand

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