TREE RECONSTRUCTION FROM MULTI-STATE CHARACTERS

CHARLES SEMPLE AND MIKE STEEL

Abstract. In evolutionary biology, a *character* is a function χ from a set X of present-day species into a finite set of states. Suppose the species in X have evolved according to a bifurcating tree $\mathcal{T}$. Biologists would like to use characters to infer this tree. Assume that χ is the result of an evolutionary process on $\mathcal{T}$ that has not involved reverse or parallel transitions, such characters are called *homoplasy-free.* In this case, χ provides direct combinatorial information about the underlying evolutionary tree $\mathcal{T}$ for X. We consider the question of how many homoplasy-free characters are required so that $\mathcal{T}$ can be correctly reconstructed. We first establish lower bounds showing that, when the number of states is bounded, the number of homoplasy-free characters required to reconstruct $\mathcal{T}$ grows (at least) linearly with the size of X. In contrast, our main result shows that, when the state space is sufficiently large, every bifurcating tree can be uniquely determined by just five homoplasy-free characters. We briefly describe the relevance of this result for some new types of genomic data, and for the amalgamation of evolutionary trees.

1. Introduction

A central problem in systematic biology is the construction of bifurcating trees to represent the evolutionary history of a collection of present-day species. The data that biologists use for this task are functions defined on the set of species. In this paper, we are concerned with a particularly useful class of such functions, ones whose evolution has been "homoplasy-free". We investigate the question of how many such characters are required to correctly reconstruct a bifurcating tree. In this section, we state our main result, Theorem 1.1, after introducing some necessary concepts and definitions.

Throughout the paper, X denotes a non-empty finite set. A *phylogenetic tree* $\mathcal{T}$ *(for X)* is a tree that has X as its set of leaves and whose interior vertices 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 regarded as equivalent if the identity map on X induces a graph isomorphism on the underlying trees. Thus, up to equivalence, there are precisely three trivalent phylogenetic trees for a set X of size 4. In biology, phylogenetic trees are widely used to represent evolutionary relationships for a set X of present-day species.

Date: 17 May 2001.

Key words and phrases. tree, character compatibility, phylogeny, chordal graph.

We thank the New Zealand Marsden Fund (UOC-MIS-005) and the New Zealand Foundation for Research, Science and Technology (UOCX0012) for supporting this research.

A (qualitative or discrete) *character on* X is a function χ from X into a set C of *character states.* If $|\chi(X)| = r$, then χ is an *r–state character.* In biology, characters describe attributes of the species under consideration and are the data that biologists use to reconstruct phylogenetic trees. Characters can be morphological (for example, wings versus no-wings), biochemical, physiological, behavioural, embryological, or genetic (for example, the nucleotide at a particular DNA sequence position, or the order of certain genes on a chromosome).

Let $\mathcal{T}$ be a phylogenetic tree for X, and let χ be a character from X into a set C of character states. 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. Then χ is *convex on* $\mathcal{T}$ if the subtrees in $\{\mathcal{T}_\alpha : \alpha \in \chi(X)\}$ are pairwise vertex disjoint. Convexity has a fundamental biological interpretation that we will describe in the next section.

To illustrate these concepts, let X be the set $\{1, 2, \ldots, 7\}$ and let C be the set $\{\alpha, \beta, \gamma, \delta\}$ of character states. Let $\chi : X \to C$ be the character defined by $\chi(1) = \chi(2) = \alpha$, $\chi(3) = \chi(5) = \beta$, $\chi(4) = \gamma$, and $\chi(6) = \chi(7) = \delta$. Then χ is convex on the phylogenetic tree $\mathcal{T}$ shown in Fig. 1 since the minimal subtrees $\mathcal{T}_\alpha$, $\mathcal{T}_\beta$, $\mathcal{T}_\gamma$, and $\mathcal{T}_\delta$ contained within the dashed boundaries are pairwise vertex disjoint.

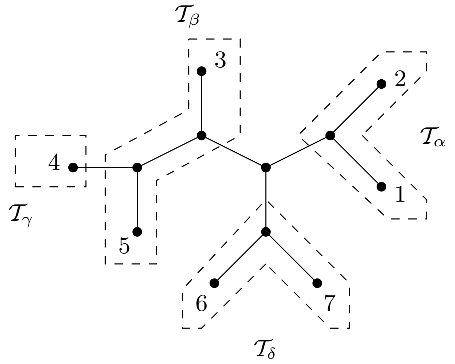


Figure 1

A collection $\mathcal{C}$ of characters 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 with this property, we say that $\mathcal{C}$ *convexly defines* $\mathcal{T}$. Observe that if $\mathcal{C}$ convexly defines a phylogenetic tree $\mathcal{T}$, then $\mathcal{T}$ must be trivalent, for otherwise, we could easily construct a different tree on which all the characters in $\mathcal{C}$ are convex.

The problem of determining whether a collection of characters is compatible is called the *character compatibility problem* or *perfect phylogeny problem.* In general, this problem is NP-complete [10], but it has a polynomial-time solution if a bound is placed either on the number of characters [5] or on the maximum number of states associated to any character [1, 4]. Buneman [2] and Meacham [6] indicated, and Steel [10] formally proved a graph-theoretic characterization for when a collection $\mathcal{C}$ of characters on X are compatible. This characterization is based on a certain intersection graph that is associated with $\mathcal{C}$. Furthermore, as a straightforward consequence of the main result in [9], we have an analogous characterization for when $\mathcal{C}$ convexly defines a phylogenetic tree.

In this paper, we are interested in determining the number of characters required to convexly define a trivalent phylogenetic tree $\mathcal{T}$ for X. We show that if the number of states for each character is bounded, then the number of characters required must grow at least linearly with the size of X. In contrast, the main result of this paper shows that if no such restriction is imposed, we require no more than five characters on X to convexly define $\mathcal{T}$. In particular, we prove the following theorem.

Theorem 1.1. *Every trivalent phylogenetic tree can be convexly defined by at most five characters.*

The paper is organized as follows. In the next section, we outline the biological background to the above definitions. Section 3 states two compatibility results that will be needed in the proof of Theorem 1.1. In Section 4, we establish a lower bound on the number of characters required to convexly define a trivalent phylogenetic tree if the number of character-states for each character is bounded. Section 5 contains the proof of Theorem 1.1, and we make some remarks on its optimality in Section 6. Section 7 briefly describes some possible applications.

Throughout this paper, given a phylogenetic tree $\mathcal{T}$ for X and a subset A of X, we let $\mathcal{T}(A)$ denote the set of vertices in the minimal subtree of $\mathcal{T}$ that contains the leaves in A. Also, for a character $\chi : X \to C$, we let $\pi(\chi)$ denote the partition of X corresponding to $\{\chi^{-1}(\alpha) : \alpha \in C\}$. Furthermore, we call a character χ on X *trivial* if at most one block of $\pi(\chi)$ has size at least two since, in this case, χ is convex on every phylogenetic tree for X.

2. Relevance of Convexity to Biology

There is a simple biological rationale for regarding convexity as a fundamental concept, which we discuss in this section.

Let $\mathcal{T}$ be a phylogenetic tree, and suppose that we subdivide an edge of $\mathcal{T}$ to create a degree–2 vertex ρ that we distinguish. We call the distinguished vertex ρ the *root vertex* of $\mathcal{T}$, and refer to the resulting tree, denoted $\mathcal{T}^{+\rho}$, as a *rooted phylogenetic tree (for X)*.

Phylogenetic trees (and their rooted counterparts) provide a convenient representation for evolutionary relationships in biology. In particular, for a rooted phylogenetic tree $\mathcal{T}^{+\rho}$ for X, we view the edges of $\mathcal{T}^{+\rho}$ as being directed away from the root ρ, and then 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 $\mathcal{T}$ correspond to other hypothetical ancestral species that are descended from the ancestral species at ρ.

Let $\mathcal{T}^{+\rho}$ be a rooted phylogenetic tree for X, and suppose each extant and ancestral species has an associated character state lying in some set C of character states. In this way, we can regard the character state as also "evolving" from

ρ towards the elements of X on $\mathcal{T}^{+\rho}$. This leads to a concept of evolutionary "innovation", namely, that each time a species changes its character state, the new state it acquires is arising for the first time in the tree. There are several equivalent ways to formalize this concept, one of which is the following. 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 will say that c is *homoplasy-free.*

(i) Suppose that $v_1, \ldots, v_k$ is a path in $\mathcal{T}^{+\rho}$ directed away from the root ρ. If, for some $i \in \{2, \ldots, k-1\}$,
$$c(v_1) = c(v_k) \neq c(v_i),$$
c is said to exhibit a *reverse transition.* Informally, this corresponds to a new state arising, but then reverting to an earlier state.

(ii) Suppose that $v_1, \ldots, v_k$, and $w_1, \ldots, w_l$ are paths in $\mathcal{T}^{+\rho}$ directed away from the root ρ, and that $v_1 = w_1$. If
$$c(v_k) = c(w_l) \neq c(v_1),$$
c is said to exhibit a *convergent transition.* Informally, this corresponds to the same state arising in different parts of the tree.

We now explain the connection between these biologically-motivated concepts and convexity. To do this, we use the following lemma whose straightforward proof is omitted. For a phylogenetic tree $\mathcal{T}$, let $V(\mathcal{T})$ denote the set of vertices of $\mathcal{T}$.

Lemma 2.1. *Let χ be a character on X, taking values in a set C, and let $\mathcal{T}$ be a phylogenetic tree for X. Then χ is convex on $\mathcal{T}$ if and only if there is a function $\overline{\chi} : V(\mathcal{T}) \rightarrow C$ satisfying the following properties:*

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

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

□

Let $\mathcal{T}^{+\rho}$ be a rooted phylogenetic tree for X, 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 tree $\mathcal{T}$. If we restrict 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 $x \in X$. This character is precisely the character describing the states assigned to the set of present-day species X. Now, if c is homoplasy-free, then χ is convex on $\mathcal{T}$ since $\overline{\chi} : V(\mathcal{T}) \rightarrow C$ defined by $\overline{\chi}(u) = c(u)$, for all $u \in V(\mathcal{T})$, satisfies conditions (C1) and (C2).

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

It is important to note, however, that even if c is not homoplasy-free on a rooted phylogenetic tree $\mathcal{T}^{+\rho}$, it is still entirely possible that the associated character χ may be convex on $\mathcal{T}$.

3. Compatibility of Characters

We now turn our attention to establishing the results in Sections 4 and 5.

An edge $\{u, v\}$ of a phylogenetic tree $\mathcal{T}$ is said to be *distinguished* by a character χ that is convex on $\mathcal{T}$ if, for every mapping $\overline{\chi}$ satisfying (C1) and (C2) in Lemma 2.1, $\overline{\chi}(u) \neq \overline{\chi}(v)$. The next proposition from [10] gives a necessary condition for a collection of characters to convexly define a phylogenetic tree.

Proposition 3.1. *If a collection $\mathcal{C}$ of characters on X convexly defines a trivalent phylogenetic tree $\mathcal{T}$, then every interior edge of $\mathcal{T}$ is distinguished by a character in $\mathcal{C}$.* □

The converse of Proposition 3.1 does not hold in general. However, as a straightforward consequence of the main result in [9], a characterization for when a collection $\mathcal{C}$ of characters on X convexly defines a phylogenetic tree for X can be obtained by supplementing the necessary condition in Proposition 3.1 with a graph-theoretic property based on a certain intersection graph associated with $\mathcal{C}$. Using this characterization, one obtains immediately the next theorem.

For a collection $\mathcal{C}$ of characters, the *intersection graph of* $\mathcal{C}$, denoted $\mathrm{int}(\mathcal{C})$, is the graph that has vertex set $\{(A, \chi) : A \in \pi(\chi) \text{ and } \chi \in \mathcal{C}\}$ and has an edge between (A, χ) and (A', χ') precisely when $A \cap A' \neq \emptyset$. Observe that the existence of such an edge automatically ensures that $\chi \neq \chi'$. A graph G is said to be *chordal* (sometimes called *triangulated*) if every vertex induced subgraph of G that is a cycle has at most three edges.

Theorem 3.2. *Let $\mathcal{C}$ be a collection of characters on X. Then $\mathcal{C}$ convexly defines a trivalent phylogenetic tree if the following two conditions are satisfied:*

(i) *there is a trivalent phylogenetic tree $\mathcal{T}$ for X on which each character in $\mathcal{C}$ is convex, and every interior edge of $\mathcal{T}$ is distinguished by a character in $\mathcal{C}$; and*
(ii) $\mathrm{int}(\mathcal{C})$ *is chordal.*

□

4. Lower Bounds

We begin with a technical result that characterizes when an edge of a phylogenetic tree is distinguished by a character. The straightforward proof is omitted.

Lemma 4.1. *Let $\mathcal{T}$ be a phylogenetic tree for X, and let χ be a character that is convex on $\mathcal{T}$. An interior edge $\{u, v\}$ of $\mathcal{T}$ is distinguished by χ if and only if there are elements x, x', y, and y' in X such that the following two properties hold:*

(i) *the path in $\mathcal{T}$ connecting x and x' contains u but not v, while the path in $\mathcal{T}$ connecting y and y' contains v but not u; and*

(ii)
$$\chi(x) = \chi(x') \neq \chi(y) = \chi(y').$$

□

The following proposition sets lower bounds on the number of characters required to convexly define a phylogenetic tree on n leaves. In particular, it shows that if the number of states is bounded, then the number of characters required must grow at least linearly with n.

Proposition 4.2. *Let $\mathcal{C}$ be a collection $\{\chi_1, \ldots, \chi_k\}$ of characters on a set X of size n and, for all $i \in \{1, \ldots, k\}$, let n_i denote the number of blocks of the partition $\pi(\chi_i)$ that contain at least two elements. Suppose that $\mathcal{C}$ convexly defines a trivalent phylogenetic tree $\mathcal{T}$ for X. Then*

$$\sum_{i=1}^{k}(n_i - 1) \geq n - 3. \tag{1}$$

In particular, if each character in $\mathcal{C}$ has at most r states and $\mathcal{C}$ convexly defines $\mathcal{T}$, then $k \geq \frac{n-3}{r-1}$.

Proof. For each $i \in \{1, 2, \ldots, k\}$, let G_i be the graph that is obtained from $\mathcal{T}$ by doing the following sequence of operations:

(i) for each $A \in \pi(\chi_i)$, contract every edge in the minimal subtree of $\mathcal{T}$ that contains the leaves in A and label the resulting vertex A; and then
(ii) delete all vertices of the resulting graph that are not labelled by an element in $\{A : A \in \pi(\chi_i) \text{ and } |A| \geq 2\}$.

For all i, let V_i and E_i denote the vertex set and edge set of G_i, respectively. Since χ_i is convex on $\mathcal{T}$,

$$|V_i| = |\{A : A \in \pi(\chi_i) \text{ and } |A| \geq 2\}| = n_i,$$

for all i. Also, since every component of G_i is a tree,

$$|E_i| \leq n_i - 1, \tag{2}$$

for all i. Clearly, for all i, the elements of E_i correspond to precisely the interior edges of $\mathcal{T}$ that are distinguished by χ_i. By Proposition 3.1 and the assumption that $\mathcal{C}$ convexly defines $\mathcal{T}$, each interior edge of $\mathcal{T}$ must be distinguished by some character in $\mathcal{C}$. Therefore, by Lemma 4.1, there is a surjective map from $\bigcup_{i=1}^{k} E_i$ into the set of interior edges of $\mathcal{T}$. Now $\mathcal{T}$ has exactly $n-3$ interior edges as $\mathcal{T}$ is trivalent, and so

$$n - 3 \leq |\bigcup_{i=1}^{k} E_i| \leq \sum_{i=1}^{k} |E_i| \leq \sum_{i=1}^{k}(n_i - 1),$$

where the last inequality follows from (2). This establishes (1). Moreover, by inserting the bound $n_i \leq r$, for all i, in inequality (1), we get $k(r-1) \geq n-3$ to complete the proof of the proposition. □

Note that for distinct, non-trivial 2–state characters, (1) becomes $k \geq n-3$, and one can subsequently show that, in this case, $k = n-3$.

Proposition 4.2 suggests that if we wish to minimize k, then we should make r large. However, one should be careful not to make r too large, for in the extreme case, where each character has n states, each character is trivial, and no collection of such characters can convexly define a phylogenetic tree. One way to obtain an insight into this trade-off is to consider the proportion of trivalent phylogenetic trees for X on which a given character χ on X is convex. The following elegant result is from [3].

Theorem 4.3. *Let χ be an r–state character on X, and let $a_1, a_2, \ldots, a_r$ denote the size of the blocks of $\pi(\chi)$. Then the proportion p of trivalent phylogenetic trees for X on which χ is convex is exactly*

$$p = \frac{1}{B(n-r+2)} \prod_{i=1}^{r} B(a_i+1), \tag{3}$$

where $n = |X|$ and $B(m) = 1 \times 3 \times 5 \times \cdots \times (2m-5)$ is the number of trivalent phylogenetic trees trees on a set of size m. □

If the blocks of $\pi(\chi)$ all have the same size, then (3) is minimized for fixed n and r. In this setting, the graph of $-\log(p)$ as a function of r is shown for $n = 120$ in Fig. 2 for the integer divisors of n. Thus, for $n = 120$, the type of character that is most informative, in terms of minimizing the number of trivalent phylogenetic trees on which the character is convex, is a 24–state character, with each state assigned to five species.

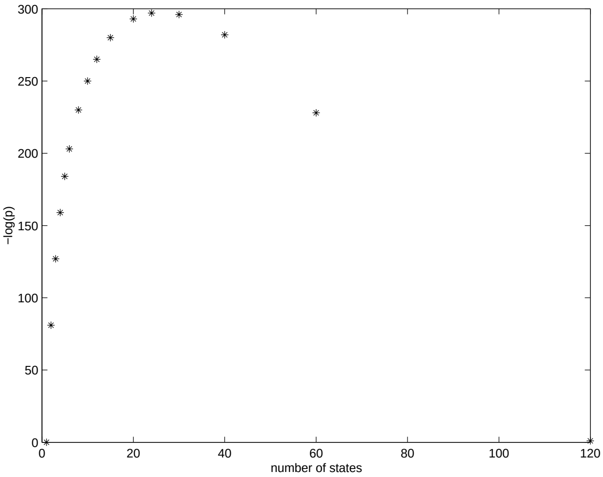


Figure 2. Distribution of trivalent phylogenetic trees on 120 leaves by the number of characters states.

5. Proof of Theorem 1.1

A *proper edge colouring* of a graph G is an assignment of colours to the edges of G so that any two adjacent edges are assigned different colours. In showing that every trivalent phylogenetic tree $\mathcal{T}$ can be defined by at most five characters, we initially construct a particular type of proper edge colouring on $\mathcal{T}$. We describe this construction first, and then establish a lemma that will be essential in the proof of Theorem 1.1.

$\mathbb{Z}_5$–edge colouring. Let T be a trivalent tree, and let $\mathbb{Z}_5$ denote the cyclic group of the set $\{0, 1, 2, 3, 4\}$ of elements under addition modulo 5. Select any leaf l of T, and direct all edges of T away from l. We will colour the resulting arcs of T with elements of $\mathbb{Z}_5$. This assignment of elements of $\mathbb{Z}_5$ is performed recursively as follows. Assign 0 to the edge of T incident with l. For each arc (u, v) of T that has been assigned an element a of $\mathbb{Z}_5$, but for which the two outgoing arcs from v have not yet been assigned an element of $\mathbb{Z}_5$, assign one of the outgoing arcs the element $a - 1$ and the other outgoing arc the element $a + 1$. The resulting assignment of elements of $\mathbb{Z}_5$ to the edges of T is called a $\mathbb{Z}_5$*–edge colouring* of T. Since, for each $a \in \mathbb{Z}_5$, the elements a, $a - 1$, and $a + 1$ are all distinct, no two adjacent edges are assigned the same element and so such an edge colouring is proper. An example of a $\mathbb{Z}_5$–edge colouring of a trivalent tree is shown in Fig. 3. In this example, we selected the leaf $l = x_1$ as our initial vertex.

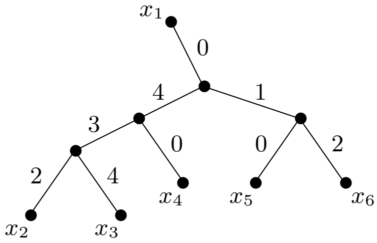


Figure 3. A $\mathbb{Z}_5$–edge colouring.

Lemma 5.1. *Let a and b be elements of $\mathbb{Z}_5$. Then every $\mathbb{Z}_5$–edge colouring of a trivalent tree T has the property that, for every vertex v of T, there is a path from v to a leaf of T that does not contain an edge coloured either a or b.*

Proof. Fix a $\mathbb{Z}_5$–edge colouring of T, and observe that, for each interior vertex v', the difference between the elements assigned to the outgoing arcs from v' is either 2 or -2. We use this observation in the proof.

If v is a leaf, then we are done. Therefore we may assume that v is an interior vertex of T. We divide the proof into two cases:

(I) $a - b \notin \{-2, 2\}$; and
(II) $a - b \in \{-2, 2\}$.

In Case (I), by the observation above, we can construct a desired path from v to a leaf of T as follows: at the last vertex in the path so far constructed, choose the next edge in the path to be an outgoing arc that is not coloured by an element of $\{a,b\}$; continue this process until the last edge added is a pendant edge of T.

Now consider Case (II). Let c denote the unique element of $\mathbb{Z}_5$ such that $\{c-1,c+1\}=\{a,b\}$. Note that if v' is an interior vertex of T and the two outgoing arcs from v' are coloured by elements of $\{a,b\}$, then the incoming arc to v' must be coloured c. Let (v,w_1) and (v,w_2) denote the two outgoing arcs from v. First, suppose that either (v,w_1) or (v,w_2) is not coloured by an element of $\{a,b\}$. A desired path P, beginning at v, can be constructed as follows: at the last vertex in the path so far constructed, choose the next edge in P to be an outgoing arc that is not coloured by an element of $\{a,b,c\}$; continue this process until the last edge added is a pendant edge of T. Clearly, this process can only fail to reach a leaf of T if, at some interior vertex in the construction, each of the outgoing arcs are coloured by elements of (i) $\{a,b\}$, (ii) $\{a,c\}$, and (iii) $\{b,c\}$. Since $a-c$ is an element of $\{-1,1\}$ and not an element of $\{-2,2\}$, there is no vertex of T whose outgoing arcs are coloured a and c. Thus (ii), and similarly (iii), cannot occur. Furthermore, since the construction of P does not choose an arc coloured c, (i) cannot occur. Hence P has the desired property.

Now suppose that both (v,w_1) and (v,w_2) are coloured by elements of $\{a,b\}$. Let (w,v) denote the incoming arc to v. Then (w,v) must be coloured c and the other outgoing arc from w is not coloured by an element of $\{a,b\}$. To obtain a desired path P from v to a leaf of T, simply choose $\{v,w\}$ as the first edge in P, and then construct a (directed) path from w to a leaf using the processed described in the last paragraph. This completes the proof of the lemma. □

We remark here that Lemma 5.1 does not hold for proper edge colourings of trivalent trees using p colours, where $p\leq 4$. The only non-trivial case to check is when $p=4$, and this can be settled by considering cases. Nevertheless, the colouring of the edges of a trivalent tree using the approach of a $\mathbb{Z}_5$–edge colouring applies equally to the cyclic group $\mathbb{Z}_p$ (the set $\{0,1,\dots,p-1\}$ of elements under addition modulo p) for all positive integers p, and so it is instructive to see where the proof breaks down in the cases $p\leq 4$. When $p=2$, the colouring is not proper since $a+1=a-1$ in $\mathbb{Z}_2$. When $p=3$, the condition $\{-1,1\}\cap\{-2,2\}=\emptyset$ used in Case (II) of the proof does not hold and, when $p=4$, the element c defined at the start of Case (II) is not unique in $\mathbb{Z}_4$.

We now prove the main result of this paper.

Proof of Theorem 1.1. Construct a $\mathbb{Z}_5$–edge colouring for $\mathcal{T}$, and let S denote the set of elements of $\mathbb{Z}_5$ that are assigned to at least one edge of $\mathcal{T}$. For each element a in S, let $\sim_a$ denote the equivalence relation on X defined by $x\sim_a y$ if the path in $\mathcal{T}$ from x to y contains no edge that is assigned colour a. For each $a\in S$, let π_a denote the partition of X that arises from the equivalence classes of $\sim_a$, and let χ_a denote a character on X so that $\pi(\chi_a)=\pi_a$. Let $\mathcal{C}=\{\chi_a : a\in S\}$. We claim that this set of at most five characters convexly defines $\mathcal{T}$. To prove this claim, it suffices, by Theorem 3.2, to show that $\mathcal{C}$ satisfies the following three properties:

(i) each character in $\mathcal{C}$ is convex on $\mathcal{T}$;

(ii) each interior edge of $\mathcal{T}$ is distinguished by a character in $\mathcal{C}$; and
(iii) $\mathrm{int}(\mathcal{C})$ is chordal.

By the way that each character in $\mathcal{C}$ is defined, (i) immediately holds. To show that (ii) holds, suppose that $e = \{u, v\}$ is an interior edge of $\mathcal{T}$. Let e_1 and e_2 be the edges of $\mathcal{T}$ incident with u, and let e_3 and e_4 be the edges of $\mathcal{T}$ incident with v. Let a denote the colour assigned to e. Clearly, for each $i \in \{1, 2\}$, there is a path from u to a leaf, x_i say, of $\mathcal{T}$ that contains e_i and no edge coloured a. Consequently, $x_1 \sim_a x_2$. Similarly, there are leaves x_3 and x_4 of $\mathcal{T}$ such that $x_3 \sim_a x_4$ and the path from x_3 to x_4 in $\mathcal{T}$ contains v but not u. Furthermore, x_2 is not equivalent to x_3 under $\sim_a$ since the path between x_2 and x_3 contains the edge e and this is coloured a. Therefore, by Lemma 4.1, e is distinguished by χ_a. It follows that (ii) holds.

Lastly, we show that (iii) holds. Let G be the intersection graph that has the same vertex set as $\mathrm{int}(\mathcal{C})$, namely, $\{(A, \chi) : A \in \pi(\chi) \text{ and } \chi \in \mathcal{C}\}$, and has an edge between (A, χ) and (A', χ') precisely if $\mathcal{T}(A) \cap \mathcal{T}(A') \neq \emptyset$. Since G is an intersection graph of subtrees of a tree, it follows by a result in [2] that G is a chordal graph. We complete the proof that (iii) holds, and hence that the theorem holds, by showing that G is identical to $\mathrm{int}(\mathcal{C})$.

If $A \cap A' \neq \emptyset$, then $\mathcal{T}(A) \cap \mathcal{T}(A') \neq \emptyset$, and so every edge in $\mathrm{int}(\mathcal{C})$ is also an edge in G. Now suppose that $\{(A, \chi_a), (B, \chi_b)\}$ is an edge in G, where a and b are elements of $\mathbb{Z}_5$. Let v be a vertex in $\mathcal{T}(A) \cap \mathcal{T}(B)$. Observe that if $u \in \mathcal{T}(A)$, and y is any leaf of $\mathcal{T}$ for which the (unique) path in $\mathcal{T}$ from u to y does not contain an edge coloured a, then $y \in A$. A similar observation holds for $\mathcal{T}(B)$ and b. Now, by Lemma 5.1, there is path from v to a leaf, x say, of $\mathcal{T}$ that does not contain an edge coloured by an element of $\{a, b\}$. It follows that $x \in A$ and $x \in B$, in particular, $A \cap B$ is non-empty. Hence $\{(A, \chi_a), (B, \chi_b)\}$ is an edge of $\mathrm{int}(\mathcal{C})$, and so $\mathrm{int}(\mathcal{C})$ is identical to G. □

6. Remarks

In this section, we make some remarks as to the optimality of Theorem 1.1, which was proved in the last section.

Firstly, there is an infinite family of trivalent phylogenetic trees that can be convexly defined by just two characters. Consider the class of *caterpillar* phylogenetic trees for X shown in Fig. 4, where $x_1, \ldots, x_n$ is a permutation of the elements of X. Up to equivalence, there are $n!/8$ such phylogenetic trees for a given set X of size $n > 3$.

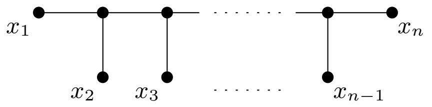


Figure 4. A caterpillar phylogenetic tree.

Now let $\mathcal{T}$ be a caterpillar phylogenetic tree for X. If n is even, let χ_1 and χ_2 be two characters on X such that

$$\pi(\chi_1) = \{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, x_6\}, \ldots, \{x_{n-1}, x_n\}\},$$

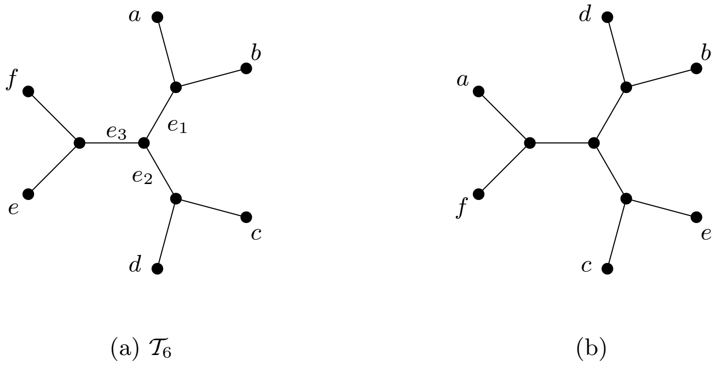
and

$$\pi(\chi_2) = \{\{x_1\}, \{x_2, x_3\}, \{x_4, x_5\}, \{x_6, x_7\}, \ldots, \{x_{n-2}, x_{n-1}\}, \{x_n\}\}.$$

Then $\mathcal{C}_1 = \{\chi_1, \chi_2\}$ convexly defines $\mathcal{T}$. To see this, observe that both χ_1 and χ_2 are convex on $\mathcal{T}$, and that every interior edge of $\mathcal{T}$ is distinguished by a character in $\mathcal{C}_1$. Also, as $\text{int}(\mathcal{C}_1)$ is a path, $\text{int}(\mathcal{C}_1)$ is chordal. Consequently, by Theorem 3.2, $\mathcal{C}_1$ convexly defines $\mathcal{T}$. If n is odd, one can similarly construct two characters on X that together convexly define $\mathcal{T}$.

Next we introduce a slightly stronger concept than "convexly defines". An X–tree is a tree $T = (V, E)$ together with a map $\phi : X \to V$ such that every vertex v in $V - \phi(X)$ has degree at least three. X–trees are a natural and mathematically useful generalization of phylogenetic trees for X; in the case an X–tree is a phylogenetic tree for X, ϕ identifies X with the leaves of T. The concept of convexity extends naturally to X–trees. In particular, a character χ on X is *convex on an X–tree* $\mathcal{T}$ if the subtrees in $\{\mathcal{T}_\alpha : \alpha \in \chi(X)\}$ are pairwise vertex disjoint, where $\mathcal{T}_\alpha$ is the minimal subtree of $\mathcal{T}$ containing the vertices in $\phi(\chi^{-1}(\alpha))$. If an X–tree $\mathcal{T}$ is the only X–tree on which a collection $\mathcal{C}$ of characters is convex, we say that $\mathcal{C}$ *strongly defines* $\mathcal{T}$, in which case, $\mathcal{T}$ must be a trivalent phylogenetic tree.

If a collection $\mathcal{C}$ of characters strongly defines $\mathcal{T}$, then $\mathcal{C}$ convexly defines $\mathcal{T}$, but the converse may not hold. However, one can extend the proof of Theorem 1.1 to show that every trivalent phylogenetic tree can be strongly defined by at most five characters. The next proposition and the remark that immediately follows it, shows that at least four characters are required to convexly define, and hence strongly define, every trivalent phylogenetic tree. Let $\mathcal{T}_6$ denote the phylogenetic tree shown in Fig. 5(a).



(a) $\mathcal{T}_6$ (b)

FIGURE 5

Proposition 6.1. *The phylogenetic tree $\mathcal{T}_6$ cannot be strongly defined by at most three characters.*

Proof. Suppose, to the contrary, that $\mathcal{C}$ is such a collection of characters. Then, as $\mathcal{T}_6$ contains three interior edges each of which is incident with the other two, it follows by Proposition 3.1 that $|\mathcal{C}| = 3$. In particular, each character in $\mathcal{C}$ distinguishes exactly one of the interior edges e_1, e_2, and e_3. Let χ_1, χ_2, and χ_3 denote the corresponding character in $\mathcal{C}$, respectively. By Lemma 4.1, this implies, for all $i \in \{1, 2, 3\}$, that at least two of the character states of χ_i is of size at least two. Furthermore, it is easily seen that, as $\mathcal{C}$ strongly defines $\mathcal{T}$, each of the pendant edges of $\mathcal{T}$ must also be distinguished by a character in $\mathcal{C}$. Thus each singleton of $\{a, b, c, d, e, f\}$ is a character state of some member of $\mathcal{C}$. It now follows that each character in $\mathcal{C}$ is a 4–state character consisting of two states of size 2 and two states of size 1.

Without loss of generality, we may assume that $\pi(\chi_1) = \{\{a, b\}, \{c, e\}, \{d\}, \{f\}\}$, and that $\{c, d\}$ and $\{e, f\}$ are elements of $\pi(\chi_2)$ and $\pi(\chi_3)$, respectively. It is easily deduced that either

(i) $\pi(\chi_2) = \{\{c, d\}, \{a, f\}, \{e\}, \{b\}\}$ and $\pi(\chi_3) = \{\{e, f\}, \{b, d\}, \{c\}, \{a\}\}$, or
(ii) $\pi(\chi_2) = \{\{c, d\}, \{b, f\}, \{e\}, \{a\}\}$ and $\pi(\chi_3) = \{\{e, f\}, \{a, d\}, \{c\}, \{b\}\}$.

Up to symmetry, these two cases are identical. Thus we may assume that χ_2 and χ_3 are as in (i). But then there are two distinct phylogenetic trees for $\{a, b, c, d, e, f\}$ on which every character in $\mathcal{C}$ is convex, namely, $\mathcal{T}_6$ and the phylogenetic tree shown in Figure 5(b). We conclude that $\mathcal{T}_6$ cannot be strongly defined by at most three characters. □

Despite Proposition 6.1, $\mathcal{T}_6$ can be convexly defined by three characters. However, not all trivalent phylogenetic trees can be convexly defined by three characters. For example, one can show that the trivalent phylogenetic tree obtained from $\mathcal{T}_6$ by joining a pair of new leaf vertices to each leaf vertex, thus creating a tree with twelve leaves, cannot be convexly defined using just three characters. In fact, based on the constructive process described in this example, one can show that if at least k characters are required to convexly define a trivalent phylogenetic tree $\mathcal{T}$, then there is a trivalent phylogenetic tree, constructed from $\mathcal{T}$, that requires at least k characters for it to be strongly defined. An interesting problem that remains is the following:

Problem 6.2. *Determine whether every trivalent phylogenetic tree can be convexly defined by four characters.* □

7. Applications

Recently, there has been considerable interest in the use of homoplasy-free multistate characters in genetics for phylogenetic inference. This is largely due to the analysis of new types of genomic data (SINEs, LINEs, and gene order data - see for example [7]). In contrast to the more traditional aligned nucleotide sequence data, where one has only 4–state characters, these new types of data typically have a very large state space. It has been argued qualitatively that, if such a character evolves under a Markov process, then it should stand a high chance of being homoplasy-free, since the probability that it has reverted to a previous state in its evolution should be small. If this is the case, our main result (Theorem 1.1) suggests that it may be possible to reconstruct large trees from a relatively small number of such characters. In practise, this number would no doubt be more than 5, but perhaps only in the order of tens rather than thousands as required for 4–state characters. Furthermore, for a bounded size set $\mathcal{C}$ of characters, there is a polynomial-time algorithm for determining if $\mathcal{C}$ is compatible, and, if so, constructing a phylogenetic tree on which all the characters in $\mathcal{C}$ are convex [5].

In this section, we attempt to quantify some informal arguments that have been presented in the biological literature by presenting explicit lower bounds on the probability that a character that evolves under a (large state) Markov process will be homoplasy free on the underlying tree.

Suppose a character evolves according to a Markov process on a rooted phylogenetic tree $\mathcal{T}^{+\rho}$ for X. Thus, at each vertex v of $\mathcal{T}^{+\rho}$, we have an associated random variable $\xi(v)$ taking values in some state space C. The Markov assumption is that, for each arc (u, v) of $\mathcal{T}^{+\rho}$, conditional on the value of $\xi(u)$, the value of $\xi(v)$ is independent of the ξ values at all other vertices that are not descendants of v (where a vertex w is a descendant of v if v lies on the path from ρ to w).

Proposition 7.1. *Given a Markov process on a rooted phylogenetic tree $\mathcal{T}^{+\rho}$, let $p(\mathcal{T})$ denote the probability that the resulting randomly-generated character χ is convex on $\mathcal{T}$. Suppose that, for each arc (u, v) of $\mathcal{T}^{+\rho}$ and each pair α, β of distinct states in C, the conditional probability that $\xi(v) = \beta$ given that $\xi(u) = \alpha$ is at most $p_{\max}$. Then*

$$p(\mathcal{T}) \geq 1 - (2n - 3)(n - 1)p_{\max}$$

where $n = |X|$.

Proof. Let $v_1, \ldots, v_t$ denote any total ordering of the vertices of $\mathcal{T}^{+\rho}$ that is consistent with the partial order induced by directing all edges of $\mathcal{T}^{+\rho}$ away from ρ. That is, if v_i lies on the path between ρ and v_j then $i < j$. Note that $v_1 = \rho$, $t \leq 2n - 1$, and $\mathcal{T}^{+\rho}$ has $t - 1$ edges. For each vertex v_j of $\mathcal{T}^{+\rho}$, other than ρ, let $v_{a(j)}$ denote the vertex that is immediately ancestral to v_j in $\mathcal{T}^{+\rho}$ under the total ordering. Then ξ is homoplasy-free, and so $\chi = \xi|X$ is convex on $\mathcal{T}$, provided the sequence $\xi(v_1), \xi(v_2), \ldots, \xi(v_t)$ satisfies the condition that, if $\xi(v_j) = \xi(v_i)$ for $i < j$, then

$$\xi(v_{a(j)}) = \xi(v_j).$$

For $j \geq 3$, let H_j be the event that $\xi(v_j)$ differs from $\xi(v_{a(j)})$ but takes the same value as $\xi(v_i)$ for some $i < j$. In view of the previous paragraph, a sufficient condition for χ to be convex on $\mathcal{T}$ is that none of the events $H_3, \ldots, H_t$ occur. Thus, by the Bonferroni inequality,

$$p(\mathcal{T}) \geq 1 - \mathbb{P}(\bigcup_{j=3}^{t} H_j) \geq 1 - \sum_{j=3}^{t} \mathbb{P}(H_j). \tag{4}$$

Furthermore, for $j \geq 3$,

$$\mathbb{P}(H_j) \leq (j - 2)p_{\max}, \tag{5}$$

since there are at most $j-2$ states that $\xi(v_j)$ can take in order for H_j to occur, and for each such state the probability that $\xi(v_j)$ takes this value is at most $p_{\max}$. Combining (4) and (5) and using the identity $\sum_{j=3}^{t}(j-2) = \frac{1}{2}(t-1)(t-2)$ (with t equal to its maximal value $2n-1$) the result now follows.

□

For many Markov models on a large state space, even when some of the details and underlying parameters of the model are unknown, it should be possible to place upper bounds on $p_{\max}$. In such a case the previous proposition can provide a reasonable lower bound to the probability that a resulting character is convex on the underlying tree $\mathcal{T}$. As an example, consider a simple model of gene order rearrangement, where the order of a sequence of N genes on a chromosome are altered by inversions of randomly selected blocks of consecutive (unsigned) genes. Suppose the start and end points of such inversions are uniformly chosen from the set $1, \ldots, N$. Then it can be shown that $p_{\max} \leq \frac{2}{N(N-1)}$ and so, for example, for $N = 100$ and $n = 10$, Proposition 7.1 gives a moderately high bound (0.97) on the probability that a generated character is convex on $\mathcal{T}$.

A second, more specialized applications of Theorem 1.1 in phylogenetics is to "supertree" construction. Given a collection of phylogenetic trees that classify overlapping sets of species, the supertree approach attempts to produces a parent phylogenetic tree that classifies the union of the sets of species in the input trees. Currently, the most popular method is the "MRP" (matrix representation with parsimony) approach. In this approach, one recodes each input tree by a set of binary characters (one for each interior edge of the tree) and then applies a method called maximum parsimony to the resulting set of characters [8]. One problem with this approach is that the number of species (leaves) in an input trees affects the number of characters it contributes to the analysis, leading to concerns about a "size bias" effect. Our result shows that this might be avoided by recoding each tree by a fixed number of convex multi-state characters.

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), 1216–1224.

[2] P. Buneman, A characterisation of rigid circuit graphs, *Discrete Math.* **9** (1974), 205–212.

[3] M. Carter, M. D. Hendy, D. Penny, L. A. Székely, and N. C. Wormald, On the distribution of lengths of evolutionary trees, *SIAM J. Discr. Math.* **3** (1990), 38–47.

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

[5] F. R. McMorris, T. Warnow, and T. Wimer, Triangulating vertex coloured graphs, *SIAM J. Discr. Math.* **7** (1994), 296–306.

[6] C. A. Meacham, Theoretical and computational considerations of the compatibility of qualitative taxonomic characters, *in* "Numerical Taxonomy", NATO ASI Series Vol. G1, Springer-Verlag, 1983.

[7] A. Rokas and P. W. H. Holland, Rare genomic changes as a tool for phylogenetics, *Trends in Ecology and Evolution* **15**(11) (2000), 454–459.

[8] M. J. Sanderson, A. Purvis, and C. Henze, Phylogenetic supertrees: assembling the trees of life, *Trends in Ecology and Evolution* **13** (1998), 105–109.

[9] C. Semple and M. Steel, A characterization for a set of partial partitions to define an X–tree, *Discrete Mathematics*, in press.

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

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

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