

Encoding phylogenetic trees in terms of weighted quartets

Stefan Grünewald

CAS-MPG Partner Institute for Computational Biology,
Shanghai Institutes for Biological Sciences,
Shanghai, P.R.China.

Katharina T. Huber

School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ, UK.

Vincent Moulton

School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ, UK.

Charles Semple

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

January 15, 2007

Corresponding Author:

Prof. Vincent Moulton,
School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ, UK.

Fax: +44 1603 593345

E-mail: vincent.moulton@cmp.uea.ac.uk

Abstract

One of the main problems in phylogenetics is to develop systematic methods for constructing evolutionary or phylogenetic trees. For a set of species X , an *edge-weighted phylogenetic X -tree* or *phylogenetic tree* is a (graph theoretical) tree with leaf set X and no degree 2 vertices, together with a map assigning a non-negative length to each edge of the tree. Within phylogenetics, several methods have been proposed for constructing such trees that work by trying to piece together *quartet trees* on X , i.e. phylogenetic trees each having four leaves in X . Hence, it is of interest to characterise when a collection of quartet trees corresponds to a (unique) phylogenetic tree. Recently, Dress and Erdős provided such a characterisation for *binary* phylogenetic trees, that is, phylogenetic trees all of whose internal vertices have degree 3. Here we provide a new characterisation for arbitrary phylogenetic trees.

1 Introduction

Phylogenetics is the field concerned with the construction and analysis of evolutionary trees and networks to describe and understand the evolution of species, populations and individuals. One of the main problems in phylogenetics is to develop systematic methods for constructing evolutionary or phylogenetic trees. Stated more formally, given a set of species X , this problem is equivalent to constructing an *edge-weighted phylogenetic X -tree* (or *phylogenetic tree* for short), that is, a (graph theoretical) tree with leaf set X and no degree 2 vertices, together with a map assigning a non-negative real length to each edge of the tree. For example, see Fig. 1 for a phylogenetic tree with leaf set $X = \{\text{gibbon, orangutan, human, chimp, gorilla}\}$ representing the currently accepted evolutionary scenario for the apes.

Various methods have been proposed for constructing such trees, motivated in part by the increasing availability of molecular sequence data (see e.g. [10, 11]). For example, one such class of methods — commonly known as *supertree* methods — is based on the idea of building up trees from smaller trees. This approach is currently quite popular due to its application within present efforts to construct the Tree of Life (see e.g. [14] and <http://www.phylo.org/>). It has the advantage that it may be used in case incomplete data is available and/or when more costly tree building techniques are preferred (such as e.g. likelihood based approaches) which can only be performed for small subsets of the species in question. However, in general the combinatorial problem of piecing together trees on subsets of a set into a larger tree on the complete set is difficult, and several different strategies have been presented for its solution (see e.g. [5] for a recent overview).

In terms of supertrees, *quartet trees*, that is, phylogenetic trees having 4

leaves, can be regarded as the smallest building block from which phylogenetic trees may be constructed. Indeed, several methods have been proposed for constructing phylogenetic trees from quartet trees, such as Tree-puzzling [13], Addquart [3], quartet cleaning [4], dynamic programming [2], and linear programming [15] to name a few (see [10, Chapter 12] for a more detailed overview). Although it is straight-forward to induce a collection of quartet-trees from a phylogenetic tree (see Fig. 1), the reverse process — that is, trying to piece together a collection of quartet trees to form a phylogenetic tree inducing that collection — is not. For the purposes of constructing phylogenetic trees from quartet trees it is thus a fundamental question to characterise when a collection of quartet trees corresponds precisely to a phylogenetic tree.

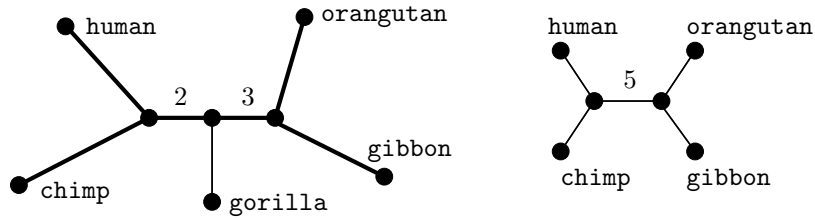


Figure 1: A phylogenetic X -tree \mathcal{T} (left) with leaf set $X = \{\text{gibbon, orangutan, human, chimp, gorilla}\}$ and internal edge weights 2 and 3, together with the quartet tree (right) **human chimp|orangutan gibbon** induced by \mathcal{T} , as indicated by the bold edges in \mathcal{T} .

For *unweighted* phylogenetic trees, i.e. phylogenetic trees in which just the combinatorial structure or *topology* of the tree is taken into account and edge weights are ignored, Colonius and Schulze gave a characterisation in 1977 for when a collection of unweighted quartet trees is induced by a (necessarily unique) unweighted phylogenetic tree [7, 8] (see also [1] for an alternative characterisation). However, it was not until much more recently in 2003 that Dress and Erdős gave an analogous result in the weighted setting for *binary* phylogenetic trees (trees in which every internal vertex has degree 3) [9]. The main result of this paper, which we now present in more formal terms, provides such a characterisation for arbitrary phylogenetic trees.

Let $\mathcal{Q}(X)$ denote the set of *quartets* on X , that is, the set of bipartitions of the form $\{\{a, b\}, \{c, d\}\}$, with $a, b, c, d \in X$ distinct, which we also denote by $ab|cd$. A *weighted quartet* is an element of $q \in \mathcal{Q}(X)$ together with a weight $\mu(q)$ in the non-negative reals, $\mathbb{R}_{\geq 0}$. Weighted quartets correspond to quartet trees with pendant edge-weights suppressed (e.g. in Fig. 1 the quartet tree pictured corresponds to the quartet **human chimp|orangutan gibbon** with weight 5). Now, given a weight for each quartet on X , that is, a map $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$, consider the following conditions:

(T1) For all $a, b, c, d \in X$, at least two of $\mu(ab|cd)$, $\mu(ac|bd)$, and $\mu(ad|bc)$ are equal to 0.

(T2) For all $x \in X - \{a, b, c, d\}$, if $\mu(ab|cd) > 0$, then either

$$\mu(ab|cx) > 0 \text{ and } \mu(ab|dx) > 0$$

or

$$\mu(ax|cd) > 0 \text{ and } \mu(bx|cd) > 0.$$

(T3) For all $a, b, c, d, e \in X$, if $\mu(ab|cd) > \mu(ab|ce) > 0$, then

$$\mu(ae|cd) = \mu(ab|cd) - \mu(ab|ce).$$

(T4) For all $a, b, c, d, e \in X$, if $\mu(ab|cd) > 0$ and $\mu(bc|de) > 0$, then

$$\mu(ab|de) = \mu(ab|cd) + \mu(bc|de).$$

For a phylogenetic tree \mathcal{T} with leaf set X , define the map

$$\mu_{\mathcal{T}} : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}, ab|cd \mapsto \mu_{\mathcal{T}}(ab|cd),$$

which takes each element $ab|cd \in \mathcal{Q}(X)$ to the length $\mu_{\mathcal{T}}(ab|cd)$ of the path in \mathcal{T} connecting the path between a and b and the path between c and d in case the latter 2 paths are vertex disjoint and 0 else. Then we shall prove the following result:

Theorem 1 *Let $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ be a map. Then $\mu = \mu_{\mathcal{T}}$ for some edge-weighted phylogenetic X -tree \mathcal{T} if and only if μ satisfies conditions (T1)-(T4). Moreover, if such a tree exists, then, up to phylogenetic X -tree isomorphism and the weights of the pendant edges, \mathcal{T} is unique.*

Note that for unweighted phylogenetic trees, conditions (T1) and (T2) are essentially equivalent to the conditions presented in [1, Proposition 2] for characterising when a collection of quartets corresponds to a phylogenetic tree. Also, as indicated above, in [9, Theorem 1.1] an analogous result to Theorem 1 is proven for *binary* phylogenetic trees. However, there appears to be no obvious way to generalise the arguments used in [9] to non-binary trees. This necessitated the new line of reasoning that we present in the proof of Theorem 1.

Clearly, given an arbitrary map $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ each of conditions (T1)-(T4) can be checked in polynomial time as a function of $|X|$. Thus Theorem 1 can also be used to provide a polynomial-time algorithm for deciding if μ corresponds to a tree or not. Furthermore, when this is the case one can use the polynomial-time supertree algorithm ‘‘BUILD’’ [12] and the approach described in [12, Proposition 6.4.4] to obtain a phylogenetic tree \mathcal{T} such that $\mu = \mu_{\mathcal{T}}$.

Note also that various approaches are proposed in [9, Section 4] for constructing a *binary* phylogenetic tree corresponding to a map $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$, some of which are implemented in the quartet-based tree building method presented in [15]. Using our new characterisation it should be possible to extend this method to the non-binary setting.

The rest of the paper is organised as follows. In Section 2, we show that conditions (T1)–(T4) are independent and prove Theorem 1, and in Section 3 we prove an analogue of this theorem (Theorem 2) for rooted phylogenetic trees. Throughout the paper, X denotes a finite set, and the notation and terminology follows [12].

2 The Main Result

We begin this section by noting that conditions (T1)–(T4) are independent (see Table 1).

Condition	X	μ
(T1)	$\{a, b, c, d\}$	$\mu(ab cd) = \mu(ac bd) = 1, \mu(ad bc) = 0$
(T2)	$\{a, b, c, d, e\}$	$\mu(ab cd) = 1, \text{ else } \mu = 0$
(T3)	$\{a, b, c, d, e\}$	$\mu(ab cd) = 3, \mu(ae cd) = \mu(be cd) = 1, \text{ else } \mu = 0$
(T4)	$\{a, b, c, d, e\}$	$\mu(ab cd) = \mu(ac de) = \mu(ab de) = \mu(ab ce) = \mu(bc de) = 1, \text{ else } \mu = 0$

Table 1: The independence of conditions (T1)–(T4). For each row, all conditions hold except for that given in column one for the set X in column two and the map $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ given in column three. Note that in row one $|X| = 4$, but if $|X| \geq 5$ then it is straight-forward to show that (T2) and (T4) imply (T1).

We now show that properties (T1)–(T4) imply another property, which we call (T5), that will be of use in the proof of Theorem 1 below.

Lemma 1 *Let $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ be a map that satisfies properties (T1)–(T2). Then the following property holds too:*

(T5) *For all $a, b, c, d, e \in X$,*

$$\mu(ab|cd) \geq \min\{\mu(ab|ce), \mu(ab|de)\}.$$

Proof: Suppose that properties (T1)–(T4) hold but that (T5) does not hold. Then there exist five elements $a, b, c, d, e \in X$ with

$$(1) \quad \mu(ab|cd) < \min\{\mu(ab|ce), \mu(ab|de)\}.$$

We claim first that

$$\mu(ab|cd) > 0.$$

To see this, assume that $\mu(ab|cd) = 0$. Then (1) implies that $\mu(ab|ce) > 0$ and $\mu(ab|de) > 0$. Applying (T2) to $\mu(ab|ce) > 0$ and noting that $\mu(ab|cd) = 0$, we obtain $\mu(bd|ce) > 0$. Similarly, applying (T2) to $\mu(ab|de) > 0$, we also obtain $\mu(bc|de) > 0$; a contradiction in view of (T1). Hence $\mu(ab|cd) > 0$ as claimed. Using (1), $\mu(ab|ce) > \mu(ab|cd) > 0$ follows. Hence, by (T3), we have $\mu(ad|ce) = \mu(ab|ce) - \mu(ab|cd) > 0$. Since $\mu(ab|cd) > 0$, and therefore $\mu(ad|bc) = 0$ by (T1), we obtain $\mu(bd|ce) > 0$ by applying (T2) to $\mu(ad|ce)$. Hence, by (T4),

$$\mu(ab|de) + \mu(bd|ce) = \mu(ab|ce) = \mu(ab|cd) + \mu(bd|ce),$$

and so $\mu(ab|de) = \mu(ab|cd)$, contradicting (1). ■

To prove Theorem 1, we will require some new notation and a well-known result concerning phylogenetic trees. A *split* of X is a bipartition $\{A, B\}$ of X , denoted $A|B$, and a set of splits is called a *split system*. A split $A|B$ with either $|A| = 1$ or $|B| = 1$ is called a *trivial split*. A split $A|B$ *displays* a quartet $ab|cd$ if either $a, b \in A$ and $c, d \in B$, or $a, b \in B$ and $c, d \in A$.

Splits arise naturally from phylogenetic trees. In particular, given a phylogenetic tree \mathcal{T} with leaf set X , each edge e of \mathcal{T} induces a split of X as follows: If V_1 and V_2 are the vertex sets of the two components of $\mathcal{T} \setminus e$, then $(V_1 \cap X)|(V_2 \cap X)$ is a split of X . We denote the collection of splits of X induced by the edges of \mathcal{T} by $\Sigma(\mathcal{T})$. Moreover, we say that a split system Σ is *compatible* if there is a phylogenetic \mathcal{T} such that $\Sigma = \Sigma(\mathcal{T})$.

Checking compatibility of split systems is straight-forward. In particular, call two splits $A|B$ and $A'|B'$ of X *pairwise compatible* if at least one of the intersections

$$A \cap A', A \cap B', B \cap A', \text{ and } B \cap B'$$

is empty. Then the Split-Equivalence Theorem [12, Theorem 3.1.4], originally proven in [6], implies that Σ is a split system of X containing all trivial splits on X , then there is a phylogenetic tree \mathcal{T} with leaf set X with $\Sigma = \Sigma(\mathcal{T})$ if and only if any pair of splits in Σ is compatible. Moreover, if such a phylogenetic tree exists, then, up to isomorphism, \mathcal{T} is unique.

We now prove Theorem 1:

Proof: First suppose that \mathcal{T} is an edge-weighted phylogenetic X -tree. Clearly, $\mu_{\mathcal{T}}$ satisfies (T1). To see that $\mu_{\mathcal{T}}$ satisfies (T2) suppose $a, b, c, d \in X$ with $\mu_{\mathcal{T}}(ab|cd) > 0$. Then $\Sigma(\mathcal{T})$ contains a split $\sigma = A|B$ that displays $ab|cd$. Without loss of generality, we may assume that $a, b \in A$ and $c, d \in B$. Let

$x \in X - \{a, b, c, d\}$. Now either $x \in A$ or $x \in B$. If $x \in A$, then σ displays $ax|cd$ and $bx|cd$, and so $\mu_{\mathcal{T}}(ax|cd) > 0$ and $\mu_{\mathcal{T}}(bx|cd) > 0$. On the other hand, if $x \in B$, then σ displays $ab|cx$ and $ab|dx$, and so $\mu_{\mathcal{T}}(ab|cx) > 0$ and $\mu_{\mathcal{T}}(ab|dx) > 0$. Hence $\mu_{\mathcal{T}}$ satisfies (T2).

To show that $\mu_{\mathcal{T}}$ satisfies (T3), suppose $a, b, c, d, e \in X$ with $\mu_{\mathcal{T}}(ab|cd) > \mu_{\mathcal{T}}(ab|ce) > 0$. Since $\mu_{\mathcal{T}}(ab|ce) > 0$, $\Sigma(\mathcal{T})$ contains a split $\sigma = A|B$ that displays $ab|ce$. Without loss of generality, we may assume that $a, b \in A$ and $c, e \in B$. Furthermore, as $\mu_{\mathcal{T}}(ab|cd) > \mu_{\mathcal{T}}(ab|ce)$, $\Sigma(\mathcal{T})$ contains a split $\sigma' = A'|B'$ with $a, b, e \in A'$ and $c, d \in B'$. Then $d \in B$ follows from the pairwise compatibility of σ and σ' . Moreover, since $\Sigma(\mathcal{T})$ is compatible, we have that every split in $\Sigma(\mathcal{T})$ that displays $ab|ce$ also displays $ab|cd$ and that a split in $\Sigma(\mathcal{T})$ displays $ae|cd$ if and only if it displays $ab|cd$ but not $ab|ce$. This implies $\mu_{\mathcal{T}}(ae|cd) = \mu_{\mathcal{T}}(ab|cd) - \mu_{\mathcal{T}}(ab|ce)$. Hence, $\mu_{\mathcal{T}}$ satisfies (T3).

Lastly, to see that $\mu_{\mathcal{T}}$ satisfies (T4), suppose that there exist elements $a, b, c, d, e \in X$ with $\mu_{\mathcal{T}}(ab|cd) > 0$ and $\mu_{\mathcal{T}}(bc|de) > 0$. Then, it is easily seen that $\mu_{\mathcal{T}}(ab|de) > 0$ and, in particular, the length of the path in \mathcal{T} separating the path from a to b and the path from d to e is equal to $\mu_{\mathcal{T}}(ab|cd) + \mu_{\mathcal{T}}(bc|de)$. Hence $\mu_{\mathcal{T}}$ satisfies (T4).

Now suppose that $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ satisfies (T1)-(T4). We prove the converse of the theorem by induction on the cardinality of the support $\{q \in \mathcal{Q}(X) : \mu(q) > 0\}$ of μ . Note that if this cardinality is zero, then $\mu(q) = 0$ for all $q \in \mathcal{Q}(X)$. Hence, by choosing \mathcal{T} to be the phylogenetic tree with leaf set X having no interior edges we have $\mu = \mu_{\mathcal{T}}$.

So, suppose $\mu = \mu_{\mathcal{T}}$ holds for some edge-weighted phylogenetic X -tree \mathcal{T} whenever the corresponding support has cardinality smaller than $|\{q \in \mathcal{Q}(X) : \mu(q) > 0\}|$. Note that this immediately implies that there exists a quartet $q \in \mathcal{Q}(X)$ with $\mu(q) > 0$.

Let $ss'|tt'$ be a quartet of minimal positive weight. Let A, B be disjoint subsets of X such that $s, s' \in A$, $t, t' \in B$, $\mu(a_1a_2|b_1b_2) > 0$ for all $a_1, a_2 \in A$ and $b_1, b_2 \in B$, and $|A| + |B|$ is maximal. We claim that $A|B$ is a split of X . To see this claim, which is fundamental to the inductive step of the proof, suppose that A and B are subsets of X that satisfy the assumptions of the claim but $A|B$ is not a split of X . Then there is an element $x \in X - (A \cup B)$. Furthermore, because of the maximality condition on $|A| + |B|$, there exist (not necessarily distinct) elements $a_1, a_2, a_3 \in A$ and $b_1, b_2, b_3 \in B$ with $|\{a_1, a_2, b_1, b_2\}| = 4$ such that

$$\mu(a_1a_2|b_3x) = 0 \text{ and } \mu(a_3x|b_1b_2) = 0.$$

Since, by (T5),

$$\mu(a_1a_2|b_3x) \geq \min\{\mu(a_1a_2|b_3b), \mu(a_1a_2|xb)\}$$

for all $b \in B - \{b_3\}$, it follows that

$$\mu(a_1a_2|bx) = 0$$

holds for all $b \in B$. Similarly,

$$\mu(ax|b_1b_2) = 0$$

for all $a \in A$. With $a = a_1$ and $b = b_1$, this implies that

$$\mu(a_1a_2|b_1x) = 0 \text{ and } \mu(a_1x|b_1b_2) = 0,$$

contradicting the fact that $\mu(a_1a_2|b_1b_2) > 0$ and that (T2) holds. Hence $A|B$ is a split of X , as claimed.

Now choose subsets A and B of X as in the claim of the last paragraph, and let $\sigma = A|B$. We next show that the map $\mu' : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ defined by setting, for all quartets $x_1x_2|y_1y_2 \in \mathcal{Q}(X)$,

$$\mu'(x_1x_2|y_1y_2) = \begin{cases} \mu(x_1x_2|y_1y_2) - \mu(ss'|tt') & \text{if } \sigma \text{ displays } x_1x_2|y_1y_2, \\ \mu(x_1x_2|y_1y_2) & \text{else,} \end{cases}$$

satisfies properties (T1)-(T4). As μ satisfies (T1), μ' satisfies (T1). We next show that μ' satisfies (T2). Suppose there exist elements $a, b, c, d \in X$ with $\mu'(ab|cd) > 0$ but (T2) is not satisfied. Then, for some $x \in X - \{a, b, c, d\}$, $i \in \{c, d\}$, and $j \in \{a, b\}$, we have that

$$(2) \quad \mu'(ab|ix) = \mu'(jx|cd) = 0.$$

As $\mu'(ab|cd) > 0$, it follows that $\mu(ab|cd) > 0$ and so, as μ satisfies (T2), either

$$\mu(ab|cx) > 0 \text{ and } \mu(ab|dx) > 0$$

or

$$\mu(ax|cd) > 0 \text{ and } \mu(bx|cd) > 0.$$

Without loss of generality, we may assume $\mu(ab|cx) > 0$ and $\mu(ab|dx) > 0$. It now follows from (2) and the definition of μ' that either σ displays $ab|cx$ and $\mu(ab|cx) = \mu(ss'|tt')$ or σ displays $ab|dx$ and $\mu(ab|dx) = \mu(ss'|tt')$. We next obtain a contradiction in the case σ displays $ab|cx$ and $\mu(ab|cx) = \mu(ss'|tt')$. The argument in case σ displays $ab|dx$ and $\mu(ab|dx) = \mu(ss'|tt')$ is similar and omitted.

Assume that σ displays $ab|cx$ and $\mu(ab|cx) = \mu(ss'|tt')$. Without loss of generality, we may assume that $a, b \in A$ and $c, x \in B$. As σ is a split of X either $d \in A$ or $d \in B$. If $d \in A$, then σ displays $ad|cx$ and so $\mu(ad|cx) > 0$. Since μ satisfies (T4) and $\mu(ab|cd) > 0$,

$$\mu(ab|cx) = \mu(ab|cd) + \mu(ad|cx) > \mu(ss'|tt')$$

follows which is impossible. Thus $d \in B$. But this cannot hold either since the fact that σ displays $ab|cd$ combined with $\mu'(ab|cd) > 0 = \mu'(ab|cx)$ implies $\mu(ab|cd) > \mu(ab|cx) > 0$. Consequently, since μ satisfies (T3), $\mu(ax|cd) = \mu(ab|cd) - \mu(ab|cx) > 0$. Hence,

$$\mu(ax|cd) > 0,$$

and so, as μ satisfies (T2), either

$$\mu(ax|bc) > 0 \text{ or } \mu(bx|cd) > 0$$

must hold. Yet the former of the two inequalities cannot hold in view of $\mu(ab|cx) > 0$ and the fact that μ satisfies (T1). And the latter cannot hold since $a, b \in A$ and $c, d, x \in B$, implies that for all $i \in \{a, b\}$ the quartet $ix|cd$ is not displayed by σ . Consequently, for all $i \in \{a, b\}$, we have $\mu(ix|cd) = \mu'(ix|cd)$ and, by (2), there exists some $j \in \{a, b\}$ so that even $\mu(jx|cd) = \mu'(jx|cd) = 0$. Since $\mu(ax|cd) > 0$ it follows that $j = b$ and so $\mu(bx|cd) = 0$. This concludes the proof that μ' satisfies (T2).

We next show that μ' satisfies (T3). Suppose that there exist elements $a, b, c, d, e \in X$ with $\mu'(ab|cd) > \mu'(ab|ce) > 0$ but

$$(3) \quad \mu'(ae|cd) \neq \mu'(ab|cd) - \mu'(ab|ce).$$

First we assume $\mu(ab|cd) \leq \mu(ab|ce)$ which implies that σ displays $ab|ce$ but not $ab|cd$. Hence, σ displays $ad|ce$ and, in view of (T4), we have

$$\mu(ab|ce) = \mu(ab|cd) + \mu(ad|ce).$$

Since σ displays $ad|ce$, we have

$$\mu'(ab|ce) \geq \mu(ab|cd) = \mu'(ab|cd),$$

a contradiction. Therefore, we have $\mu(ab|cd) > \mu(ab|ce) > 0$ and, by applying (T3), we get

$$\mu(ae|cd) = \mu(ab|cd) - \mu(ab|ce).$$

To obtain the required contradiction, we next analyse the relationship between σ and the quartets in $\mathcal{Q} = \{ae|cd, ab|cd, ab|ce\}$. If $ae|cd$ is displayed by σ then, since σ is a split of X , either b is contained in the same subset of σ as a, e or not. Consequently, σ also displays either $ab|cd$ or $ae|bc$. In the latter case $\mu(ae|bc) > 0$ follows which is impossible in view of $\mu(ae|bc) > 0$ and (T1). Thus σ must display $ab|cd$. Similarly, one can show that if σ displays $ab|ce$ then the only other element in \mathcal{Q} that is displayed by σ is $ab|cd$. Since if σ displays $ab|cd$ only one of the other quartets in \mathcal{Q} can also be displayed by σ it follows that either no quartet in \mathcal{Q} is displayed by σ , both $ae|cd$ and $ab|cd$ are displayed by σ , or both $ab|cd$ and $ab|ce$ are displayed by σ . In all three cases we obtain a contradiction to (3). Thus μ' satisfies (T3).

Lastly, suppose that μ' does not satisfy (T4). Then there exist elements $a, b, c, d, e \in X$ with $\mu'(ab|cd) > 0$ and $\mu'(bc|de) > 0$ but $\mu'(ab|de) \neq \mu'(ab|cd) + \mu'(bc|de)$. It suffices to consider two cases:

- (i) $\mu(ab|de) \neq \mu'(ab|de)$, $\mu(ab|cd) = \mu'(ab|cd)$, and $\mu(bc|de) = \mu'(bc|de)$; and
- (ii) $\mu(ab|de) = \mu'(ab|de)$, and either $\mu(ab|cd) \neq \mu'(ab|cd)$ or $\mu(bc|de) \neq \mu'(bc|de)$.

If case (i) holds, σ displays $ab|de$. Without loss of generality, we may assume that $a, b \in A$ and $d, e \in B$. Since σ is a split of X , either $c \in A$ or $c \in B$. If $c \in A$, then σ displays the quartet $bc|de$, and so $\mu(bc|de) \neq \mu'(bc|de)$; a contradiction. A similar argument also shows that $c \notin B$. Consider (ii). Since σ cannot simultaneously display both $ab|cd$ and $bc|de$, we may assume without loss of generality that

$$\mu(ab|cd) \neq \mu'(ab|cd) \text{ and } \mu(bc|de) = \mu'(bc|de).$$

Then $ab|cd$ is displayed by σ . Again without loss of generality, we may assume that $a, b \in A$ and $c, d \in B$. Since σ is a split of X , either $e \in A$ or $e \in B$. If $e \in A$, then σ displays $ab|de$, and so $\mu(ab|de) \neq \mu'(ab|de)$; a contradiction. If $e \in B$, then $be|cd$ is displayed by σ . Thus $\mu(be|cd) > 0$, and therefore $\mu(bc|de) = 0$ by (T1). But then

$$0 = \mu(bc|de) = \mu'(bc|de) > 0;$$

a contradiction. It now follows that μ' satisfies (T4).

Since $|\{q \in Q(X) : \mu'(q) > 0\}| < |\{q \in Q(X) : \mu(q) > 0\}|$ and μ' satisfies (T1)–(T4), it follows by the induction hypothesis, that $\mu' = \mu_{\mathcal{T}'}$ for some edge-weighted phylogenetic X -tree \mathcal{T}' . Now σ is not in $\Sigma(\mathcal{T}')$ since $\mu'(ss'|tt') = 0$ but $ss'|tt'$ is displayed by σ . Furthermore, $\Sigma(\mathcal{T}') \cup \{\sigma\}$ is compatible; for otherwise, by the above consequence of the Split Equivalence Theorem, there are two quartets, $ab|cd$ and $ac|bd$ say, with $\mu(ab|cd) > 0$ and $\mu(ac|bd) > 0$, contradicting the fact that μ satisfies (T1). It is now easily seen that the edge-weighted phylogenetic X -tree \mathcal{T} with $\Sigma(\mathcal{T}) = \Sigma(\mathcal{T}') \cup \{\sigma\}$ and weights $\mu_{\mathcal{T}}(S) = \mu_{\mathcal{T}'}(S)$, for all $S \in \Sigma(\mathcal{T}')$ and $\mu(\sigma) = \mu(ss'|tt')$ has the property that $\mu = \mu_{\mathcal{T}}$.

The uniqueness statement in the theorem follows in view of the fact that the set of quartets

$$\bigcup_{A|B \in \Sigma(\mathcal{T})} \{ab|cd : a, b \in A \text{ and } c, d \in B\}$$

uniquely determines the topology of \mathcal{T} ([1, Proposition 2] — see also [12, Corollary 6.3.8]). This completes the proof of the theorem. ■

Remark 1 *An alternative proof for the existence of some edge-weighted phylogenetic X -tree \mathcal{T} with $\mu = \mu_{\mathcal{T}}$ in case μ satisfies conditions (T1)–(T4) can be obtained using the characterisation for when a collection of quartets is induced by a (necessarily unique) unweighted phylogenetic tree given in [1, Proposition 2]. In particular, if $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ is such a map, then using (T1),*

(T2) and [1, Proposition 2] it can be immediately deduced that there is an unweighted phylogenetic tree \mathcal{T} that induces precisely the quartets in the support of μ . Condition (T3) can then be used to show that for any pair of quartets $q = ab|cd, q' = a'b'|c'd'$ induced by \mathcal{T} that distinguish some edge e of \mathcal{T} (i.e. there are no degree 2 vertices in the restriction of \mathcal{T} to $\{a, b, c, d\}$ and $\{a', b', c', d'\}$, respectively) the equality $\mu(q) = \mu(q')$ holds. Subsequently, assigning to each edge e of \mathcal{T} the weight $\mu(q)$ of any quartet q that distinguishes e , (T4) can then be used to show that the weighting induced by (the now edge-weighted tree) \mathcal{T} on any quartet q is equal to $\mu(q)$.

3 Rooted Trees

In this section, we establish the analogue of Theorem 1 for rooted phylogenetic X -trees. This analogue is stated as Theorem 2. We begin with some definitions and a result concerning rooted phylogenetic trees.

A *rooted phylogenetic X -tree* \mathcal{T} is a rooted tree with no degree-two vertices except possibly the root which has degree at least two, whose leaf set is X . The rooted analogue of a quartet – which corresponds to a rooted phylogenetic tree with three leaves – is a *rooted triple*, that is, a split $A|B$ of a set Y with $|Y| = 3$ with either $|A| = 1$ or $|B| = 1$. We will use the convention that for any rooted triple $A|B$ the set to the left of “|” is of size 2. We denote the rooted triple $\{a, b\}|\{c\}$ by $ab|c$. For the set X , we denote the set of all rooted triples $ab|c$, where $a, b, c, \in X$, by $\mathcal{R}(X)$.

Associated with each vertex u of \mathcal{T} is a *cluster* A of X , that is a proper subset of X . In particular, viewing the edges of \mathcal{T} as arcs directed away from the root, the cluster corresponding to u is the subset of X that contains precisely the elements of X that can be reached from u on a directed path. We denote the set of clusters of \mathcal{T} by $\mathcal{H}(\mathcal{T})$. Note that this definition of $\mathcal{H}(\mathcal{T})$ slightly differs from the definition given in [12].

As with compatibility of split systems, it is straight-forward to check when a collection of clusters corresponds to a rooted phylogenetic X -tree. In particular, as a consequence of Split Equivalence Theorem (see [12] for details), it can be shown that if \mathcal{H} is a collection of clusters of X containing all singletons, then there is a rooted phylogenetic X -tree \mathcal{T} such that $\mathcal{H} = \mathcal{H}(\mathcal{T})$ if and only if, for all $A, B \in \mathcal{H}$,

$$A \cap B \in \{\emptyset, A, B\}.$$

Moreover, if such a rooted phylogenetic X -tree exists, then, up to isomorphism, \mathcal{T} is unique.

For a rooted phylogenetic X -tree \mathcal{T} with each edge weighted by a non-negative real number, let $\lambda_{\mathcal{T}} : \mathcal{R}(X) \rightarrow \mathbb{R}_{\geq 0}$ denote the map that is obtained

by setting $\lambda_{\mathcal{T}}(ab|c)$ be the length of the path in \mathcal{T} that joins the path between a and b , and the path between c and the root of \mathcal{T} in case both paths are vertex disjoint and 0 otherwise.

We now prove the analogue of Theorem 1 for rooted phylogenetic trees.

Theorem 2 *Let $\lambda : \mathcal{R}(X) \rightarrow \mathbb{R}_{\geq 0}$ be a map and let z be an element not in X . Then $\lambda = \lambda_{\mathcal{T}}$ for some rooted, edge-weighted phylogenetic X -tree \mathcal{T} if and only if the map $\mu : \mathcal{Q}(X \cup \{z\}) \rightarrow \mathbb{R}_{\geq 0}$ defined by*

$$\mu(ab|cd) = \begin{cases} \lambda(ab|c) & \text{if } d = z; \\ \min\{\lambda(ab|c), \lambda(ab|d)\} + \min\{\lambda(cd|a), \lambda(cd|b)\} & \text{otherwise,} \end{cases}$$

satisfies (T1)-(T4). Moreover, if such a rooted edge-weighted phylogenetic X -tree exists, then, up to isomorphism and weights of the pendant edges, \mathcal{T} is unique.

Proof: We begin the proof with some preliminaries. Given a collection \mathcal{H} of clusters and a weighting $\omega : \mathcal{H} \rightarrow \mathbb{R}_{> 0}$, define a map $\lambda_{\mathcal{H}} : \mathcal{R}(X) \rightarrow \mathbb{R}_{\geq 0}$ by setting, for $ab|c \in \mathcal{R}(X)$,

$$\lambda_{\mathcal{H}}(ab|c) = \lambda_{(\mathcal{H}, \omega)}(ab|c) = \sum_{A \in \mathcal{H}, a, b \in A, c \in X - A} \omega(A).$$

Defining for a rooted, edge-weighted phylogenetic X -tree \mathcal{T} , the weight $\omega(A)$ of any cluster $A \in \mathcal{H}(\mathcal{T})$ to be the weight of the incoming edge of the unique vertex u of \mathcal{T} from which precisely the elements of A can be reached on a directed path starting at u , we have $\lambda_{\mathcal{T}} = \lambda_{\mathcal{H}(\mathcal{T})}$. In a similar fashion, given a split system Σ on X with weight function $\omega : \Sigma \rightarrow \mathbb{R}_{> 0}$, if we define a map $\mu_{\Sigma} : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ by setting, for $q \in \mathcal{Q}(X)$,

$$\mu_{\Sigma}(q) = \mu_{(\Sigma, \omega)}(q) = \sum_{\sigma \in \Sigma, q \text{ is displayed by } \sigma} \omega(\sigma).$$

then given an edge-weighted phylogenetic X -tree \mathcal{T} , we have $\mu_{\mathcal{T}} = \mu_{\Sigma(\mathcal{T})}$.

Now, suppose $\lambda : \mathcal{R}(X) \rightarrow \mathbb{R}_{\geq 0}$ is a map, z is an element not in X , and that the map μ as defined in the theorem satisfies (T1)-(T4). Then, by Theorem 1 and the last observation, there is an edge-weighted phylogenetic $(X \cup \{z\})$ -tree \mathcal{T}_z with $\mu = \mu_{\Sigma(\mathcal{T}_z)}$. Let \mathcal{T} be the rooted edge-weighted phylogenetic X -tree obtained from \mathcal{T}_z by rooting it at the unique vertex adjacent to z , and then deleting z and its incident edge. Label the root of \mathcal{T} by ρ . We claim that $\lambda = \lambda_{\mathcal{T}}$.

Let $a, b, c \in X$ and suppose that $w = \lambda(ab|c)$. Then $\mu(ab|cz) = w$, and so the length of the path P in \mathcal{T}_z that joins the path from a to b and the path from

c to z is w . Since P is also the path in \mathcal{T} that joins the path from a to b and the path from c to ρ , it follows that $\lambda_{\mathcal{T}}(ab|c) = w$. The claim now follows.

For the converse, suppose that $\lambda = \lambda_{\mathcal{T}}$ for some rooted edge-weighted phylogenetic X -tree \mathcal{T} , and let μ be as defined in the statement of the theorem. Now let \mathcal{T}_z be the (unrooted) edge-weighted phylogenetic $(X \cup \{z\})$ -tree that is obtained from \mathcal{T} by attaching a vertex labelled z via a new pendant edge to the root and assigning weight 1 to it and then viewing the resulting tree as an unrooted edge-weighted phylogenetic $(X \cup \{z\})$ -tree. We show that $\mu = \mu_{\mathcal{T}_z}$.

Let $a, b, c, d \in X \cup \{z\}$ and suppose that $w = \mu_{\Sigma(\mathcal{T}_z)}(ab|cd)$. It suffices to show that $\mu(ab|cd) = w$.

If, up to permuting elements, $d = z$, then $w = \lambda_{\mathcal{H}(\mathcal{T})}(ab|c)$ and so, $\lambda(ab|c) = w$. By definition, this implies that $\mu(ab|cd) = w$. Now assume that none of the elements a, b, c , and d is z . If $w = 0$, then there are no edges separating the path from a to b and the path from c to d in \mathcal{T}_z and hence also in \mathcal{T} . This implies that either $\lambda_{\mathcal{H}(\mathcal{T})}(ab|c) = 0$ or $\lambda_{\mathcal{H}(\mathcal{T})}(ab|d) = 0$ in \mathcal{T} and that either $\lambda_{\mathcal{H}(\mathcal{T})}(cd|a) = 0$ or $\lambda_{\mathcal{H}(\mathcal{T})}(cd|b) = 0$ in \mathcal{T} . As $\lambda = \lambda_{\mathcal{T}}$, it follows by definition that $\mu(ab|cd) = 0$. Thus we may assume that $w > 0$. Up to permuting elements, we may further assume that the path in \mathcal{T}_z from z to either c or d does not intersect the path from a to b . There are now two cases to consider depending upon where the path P from z initially meets the minimal subtree \mathcal{S} of \mathcal{T}_z connecting a, b, c , and d :

- (i) P does not initially meet \mathcal{S} on the path from c to d ; and
- (ii) P initially meets \mathcal{S} on the path from c to d .

In case (i), we have that $\lambda_{\mathcal{H}(\mathcal{T})}(ab|c) = \lambda_{\mathcal{H}(\mathcal{T})}(ab|d)$, and $\lambda_{\mathcal{H}(\mathcal{T})}(cd|a) = \lambda_{\mathcal{H}(\mathcal{T})}(cd|b)$. Therefore, it follows that

$$\begin{aligned} \mu_{\Sigma(\mathcal{T}_z)}(ab|cd) &= \lambda_{\mathcal{H}(\mathcal{T})}(ab|c) + \lambda_{\mathcal{H}(\mathcal{T})}(cd|a) \\ &= \lambda(ab|c) + \lambda(cd|a) \\ &= \mu(ab|cd). \end{aligned}$$

The proof for case (ii) is similar and omitted. The first part of the theorem now follows by Theorem 1. Furthermore, the uniqueness part holds as the set

$$\bigcup_{A \in \mathcal{H}(\mathcal{T})} \{ab|c : a, b \in A \text{ and } c \in X - A\}$$

of rooted triples uniquely determines the topology of \mathcal{T} . Together with their weights the elements in the above set therefore uniquely determine \mathcal{T} , ignoring of course the length of the pendant edges (see [12, Theorem 6.4.1]). ■

Acknowledgement: We thank the anonymous referees for their helpful remarks. Huber and Moulton thank the Department of Mathematics and Statistics, University of Canterbury, New Zealand for hosting them during the preliminary stages of this work, during which time Moulton was supported by a University of Canterbury Erskin Fellowship. Grünewald thanks the Allan Wilson Centre for Molecular Ecology and Evolution and the Department of Mathematics and Statistics, University of Canterbury where he made most of his contribution to this work. Semple was supported by the New Zealand Marsden Fund (UOC310).

References

- [1] H. -J. Bandelt, A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Adv. Appl. Math.*, **7** (1986) 309–343.
- [2] A. Ben-Dor, B. Chor, D. Graur, R. Ophir, D. Pelleg, Constructing phylogenies from quartets: elucidation of eutherian superordinal relationships, *J. Comp. Biol.*, **5** (1998) 377–390.
- [3] V. Berry, O. Gascuel, Inferring evolutionary trees with strong combinatorial evidence, *Theoretical Computer Science*, **240** (2000) 271–298.
- [4] V. Berry, T. Jiang, P. Kearney, M. Li, T. Wareham, Quartet cleaning: improved algorithms and simulations, *Proceedings of the 7th European symposium on algorithm, (ESA99)* (1999) 313–324.
- [5] O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life; Computational Biology* **4**, Kluwer Academic Publishers, Dordrecht, The Netherlands (2004).
- [6] P. Buneman, 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), Edinburgh University Press, Edinburgh, pp. 387–395 (1971).
- [7] H. Colonius, H.H. Schultze, Trees constructed from empirical relations, *Braunschweiger Berichte aus dem Institut fuer Psychologie*, **1** (1977).
- [8] H. Colonius, H. H. Schultze, Tree structure from proximity data, *British J. Math. Statist. Psych.*, **34** (1981) 167–180.
- [9] A. Dress, P. Erdős, X -trees and weighted quartet systems, *Ann. of Combin.*, **7** (2003) 155–169.
- [10] J. Felsenstein, *Inferring phylogenies*, Sinauer Associates, Inc. (2004).
- [11] M. Salemi, A. -M. Vandamme (Ed.s), *The Phylogenetic Handbook*, Cambridge University Press (2003).

- [12] C. Semple, M. Steel, *Phylogenetics*, Oxford University Press, 2003.
- [13] K. Strimmer, A. von Haeseler, Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies, *Mol. Biol. Evol.*, **13** (1996) 964–969.
- [14] A. M. Sugden, B. R. Jasny, E. Culotta, E. Pennisi Charting the evolutionary history of life, *Science*, **300 (5626)** (2003) 1691.
- [15] J. Weyer-Menkhoff, C. Devauchelle, A. Grossmann, S. Grünwald, Integer linear programming as a tool for constructing trees from quartet data, *Computational Biology and Chemistry*, **29** (2005) 196–203.