Annals of Combinatorics 11 (2007) 143-160
0218-0006/07/020143-18
DOI 10.1007/s00026-007-0311-4

Annals of Combinatorics

Phylogenetic Diversity Over an Abelian Group

Andreas Dress[1] and Mike Steel[2]

[1]Department for Combinatorics and Geometry, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P.R. China
dress@mis.mpg.de

[2]Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand
m.steel@math.canterbury.ac.nz

Received July 13, 2005

AMS Subject Classification: 05C05, 92D15

Abstract. There is a natural way to associate to any tree T with leaf set X, and with edges weighted by elements from an abelian group G, a map from the power set of X into G — simply add the elements on the edges that connect the leaves in that subset. This map has been well-studied in the case where G has no elements of order 2 (particularly when G is the additive group of real numbers) and, for this setting, subsets of leaves of size *two* play a crucial role. However, the existence and uniqueness results in that setting do not extend to arbitrary abelian groups. We study this more general problem here, and by working instead with both, *pairs and triples* of leaves, we obtain analogous existence and uniqueness results. Some particular results for elementary abelian 2-groups are also described.

Keywords: X-trees, split systems, abelian groups, group-valued distances, phylogenetic diversity

1. Introduction

In this paper we study the reconstruction of trees from path distances (and 3-way distances) when the edges of the tree are weighted by elements from an abelian group. This extends earlier work from [2] on tree reconstruction from path distances when the underlying abelian group has no elements of order 2. The corresponding uniqueness and existence theorems no longer hold for general abelian groups. However, by moving from path distances to three-way distances, we derive corresponding existence and uniqueness results. Some special applications to elementary abelian 2-groups (of relevance to binary and DNA sequences) are also mentioned. The arguments rely heavily on the theory of symbolic ultrametric representations from [4]. We begin by introducing some notation that will be used throughout this paper.

1.1. Preliminaries

Let X denote a nonempty finite set, let $\mathcal{P}(X)$ denote the power set of X, let $\mathcal{S}(X)$ denote the set of *X-splits*, i.e., the set consisting of all subsets $S = \{A, B\}$ of $\mathcal{P}(X)$ of cardinality 2 containing two disjoint subsets A, B of X for which $A \cup B = X$ holds (including the *trivial* X-split $\{X, \emptyset\}$), and let $\mathcal{S}^*(X)$ denote the set of nontrivial *X-splits*, i.e., the set consisting of all splits $S \in \mathcal{S}(X)$ with $S \neq \{X, \emptyset\}$. For a subset Y of X and an X-split $S = \{A, B\}$, let S_Y denote the *induced* Y-split defined by $S_Y := \{A \cap Y, B \cap Y\}$, and put

$$\mathcal{S}^*(X\colon Y) := \{S \in \mathcal{S}(X)\colon S_Y \in \mathcal{S}^*(Y)\}.$$

Further, given any (additive) abelian group $\mathcal{G} = (\mathcal{G}, +)$ with identity element $0_{\mathcal{G}}$, write

- $\mathcal{P}(X \mid \mathcal{G})$ for the group of *$\mathcal{G}-$valued set systems* (over X), i.e., the group of all maps from $\mathcal{P}(X)$ into $\mathcal{G}$,
- $\mathcal{S}(X \mid \mathcal{G})$ for the group of all maps from $\mathcal{S}(X)$ into $\mathcal{G}$,
- and $\mathcal{S}^*(X \mid \mathcal{G})$ for the group of all maps from $\mathcal{S}^*(X)$ into $\mathcal{G}$.

Next, given any integer $s \in \mathbb{N} := \{1, 2, \ldots\}$, write

- $\binom{X}{s} \subseteq \mathcal{P}(X)$ for the collection of subsets of X of size s,
- $\binom{X}{\leq s}$ for the collection of nonempty subsets of X of size at most s,
- and $\mathcal{P}(X, s \mid \mathcal{G})$ for the group of all maps from $\binom{X}{\leq s}$ into $\mathcal{G}$.

Further, given any map D in $\mathcal{P}(X|\mathcal{G})$ or in $\mathcal{P}(X, s|\mathcal{G})$, any two integers $r, k \in \mathbb{N}$ with $k \leq r \leq s$, and any r elements $x_1, \ldots, x_r$ in X, write

- $D^{(r)}$ for the restriction of D to $\binom{X}{\leq r}$,
- $D(x_1 \cdots x_r)$ for the value $D(\{x_1, \ldots, x_r\})$ of D at the set $\{x_1, \ldots, x_r\}$,
- and $D(x_1 \cdots x_r\colon k)$ for the sum

$$D(x_1 \cdots x_r\colon k) := \sum_{J \in \binom{\{1,\ldots,r\}}{k}} D(\{x_j\colon j \in J\}).$$

Finally, given any map $\mu\colon \mathcal{S}^*(X) \to \mathcal{G}$ in $\mathcal{S}^*(X|\mathcal{G})$ and any split system $\mathcal{S} \subseteq \mathcal{S}^*(X)$ for X, put

$$\mu(\mathcal{S}) := \sum_{S \in \mathcal{S}} \mu(S),$$

and let D_μ denote the map

$$D_\mu\colon \mathcal{P}(X) \to \mathcal{G}\colon Y \mapsto D_\mu(Y) := \mu(\mathcal{S}^*(X\colon Y)).$$

We refer to maps $\mu\colon \mathcal{S}^*(X) \to \mathcal{G}$ as above as (*proper*) *$\mathcal{G}$-valued split assignments* (for X), and we will seek for specific conditions on μ that will allow us to reconstruct μ from the induced maps $D_\mu^{(s)}$ for the particular values $s = 2$ and $s = 3$ (what happens for larger values of s, apparently also worth to be studied, will be considered in subsequent papers).

Here is the context in which this task has been found to be of particular interest in previous work, and which will be the focus of this paper: An *X-tree* is a tree $T =$

(V_T, E_T) for which X is a subset of V_T and X includes all vertices of degree at most 2. In case X is precisely the set of leaves of T, we say that T is a *phylogenetic X-tree*. Given any finite tree $T = (V_T, E_T)$, and any edge $e \in E_T$, we let $S_X(e) \in \mathcal{S}^*(X)$ denote the (necessarily nontrivial) X-split associated with edge e (i.e., the split $\{A, B\}$ consisting of the two subsets of vertices from X in the corresponding two connected components of the graph $(V_T, E_T - \{e\})$; further background on X-trees can be found in [17]).

Let $T = (V_T, E_T)$ be an X-tree and

$$\lambda\colon E_T \to \mathcal{G}$$

be any assignment of elements of $\mathcal{G}$ to the edges of T — we refer to this as a *$\mathcal{G}$-valued edge assignment for T* — and associate, to any such assignment λ, the induced map

$$\lambda_X\colon \mathcal{S}^*(X) \to \mathcal{G}\colon S \mapsto \lambda_X(S) := \sum_{\substack{e\in E_T\\ S_X(e)=S}} \lambda(e)$$

and the corresponding map $D_\lambda := D_{\lambda_X}\colon \mathcal{P}(X) \to \mathcal{G}\colon Y \mapsto \lambda_X(\mathcal{S}^*(X\colon Y))$.

Observe that if we denote, for any subset Y of X, the (unique!) smallest subtree of T containing all vertices in Y by $T(Y)$ and its edge set by $E_T(Y)$, we have $e \in E_T(Y)$ for some edge e in E_T if and only if the X-split $S_X(e)$ associated to e is contained in $\mathcal{S}^*(X\colon Y)$ which, in turn, implies that $D_\lambda(Y)$ coincides with the *total λ-length*

$$\left\|\lambda_{T(Y)}\right\| := \sum_{e\in E_T(Y)} \lambda(e)$$

of $T(Y)$.

1.2. Examples

In the context of X-trees, there are two choices of $\mathcal{G}$ of particular interest.

(i) $\mathcal{G}$ is the additive group $\mathbb{R} = (\mathbb{R}, +)$ of real numbers.

When λ is required to take nonnegative values, only, on the edges of the tree, the restriction $D_\lambda^{(2)}$ of D_λ to the subsets of size at most 2 $\big($or, rather, the associated map $X \times X \to \mathbb{R}\colon (x, y) \mapsto D_\lambda(xy)\big)$, is called a *tree metric* (for X) and has been widely studied for more than 30 years. If the subsets Y of X taken into consideration are allowed to range over larger subsets of X, the map D_λ has been applied to quantify biological diversity of subsets of species and, in this context, $D_\lambda(Y)$ is referred to as the *phylogenetic diversity* of Y. This measure, introduced by Faith in 1992 [11], provides some indication of how much evolutionary 'heritage' each possible subset Y contains in relation to the entire tree (by comparing $D_\lambda(Y)$ to $D_\lambda(X)$) and has been proposed as a tool for 'managing' the conservation of endangered species (or, depending on one's interest and perspective, for prioritizing their extinction). For further details, we refer the reader to [3] and the references therein.

An example of this concept is illustrated in Figure 1; if we take $Y = \{a, b, g, e\}$ then $D_\lambda(Y) = \|\lambda_{T(Y)}\| = 18$.

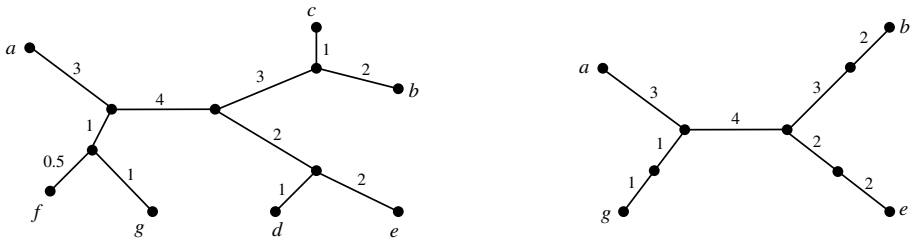

Figure 1: *Left*: A phylogenetic X-tree with edges weighted by real numbers. *Right*: The induced edge weighted tree on the subset $Y = \{a, b, g, e\}$ with $D_\lambda(Y) = 18$.

There have been two recent mathematical investigations of D_λ in the case where $\mathcal{G}$ is the additive group of real numbers. Pachter and Speyer [15] addressed the following question: For which value of m do the subsets Y of X of size m suffice to recover T? Steel [18] showed how the subsets of size m that maximize $D_\lambda(Y)$ can be constructed using a greedy algorithm.

In addition, maps in $\mathcal{P}(X, 3 \mid \mathcal{G})$ have been studied in the context of classification (such as psychology) under the general term 'three-way distances', as discussed further in [13, 14].

(ii) $\mathcal{G}$ is an elementary abelian 2-group.

This arises in the following context which is closely connected with the study of aligned genetic sequences in evolutionary biology. Consider the set of sequences of length k over an alphabet $\mathcal{A}$ of either 2 (e.g., the 'purines, pyrmidines' in molecular biology) or 4 letters (e.g., the DNA bases A, C, G, T). Note that there is canonical transitive and faithful action of an elementary abelian 2-group $\mathcal{G} = \mathcal{G}_\mathcal{A}$ of order $\#\mathcal{A}$ on $\mathcal{A}$, namely the unique nontrivial action of the cyclic of order 2 on $\mathcal{A}$ in case $\#\mathcal{A} = 2$, and the action of the normal elementary abelian 4-subgroup of the full permutation group $S_\mathcal{A}$ of $\mathcal{A}$, i.e., the well-known *Klein* 4-*subgroup* of $S_\mathcal{A}$, in case $\#\mathcal{A} = 4$. This allows us to view $\mathcal{A}$ as an *affine space* over $\mathcal{G}_\mathcal{A}$. In other words, as was first noted and exploited by Evans and Speed [10], there exists in case $\#\mathcal{A} = 2$ or 4 a unique map

$$\tau\colon \mathcal{A} \times \mathcal{A} \to \mathcal{G}_\mathcal{A}\colon (x, y) \mapsto \tau_{x\to y}$$

from $\mathcal{A} \times \mathcal{A}$ into that group that associates, to any pair (x, y) of letters in $\mathcal{A}$, the unique *translation* $\tau_{xy} \in \mathcal{G}_\mathcal{A}$ for which $\tau_{x\to y}(x) = y$ holds. In consequence, there exists also, for every $k \in \mathbb{N}$ and for $\mathcal{G} := {\mathcal{G}_\mathcal{A}}^k$, a unique map

$$\tau^{(k)}\colon \mathcal{A}^k \times \mathcal{A}^k \to \mathcal{G}\colon (x, y) \mapsto \tau^{(k)}_{x\to y}$$

that associates, to any pair (x, y) of sequences of length k in $\mathcal{A}^k$, the unique translation $\tau^{(k)}_{x\to y} \in \mathcal{G}$ for which $\tau^{(k)}_{x\to y}(x) = y$ holds. Note that

(i) $\tau^{(k)}_{x\to y} = 0 \iff x = y$,

(ii) $\tau^{(k)}_{y\to z} + \tau^{(k)}_{x\to y} = \tau^{(k)}_{x\to z}$,

(iii) and $\tau^{(k)}_{x\to y} = \tau^{(k)}_{y\to x}$

holds for all $x, y, z \in \mathcal{A}^k$.

In case $\mathcal{A}$ is the set $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ of DNA bases, one of the three non-zero elements in $\mathcal{G}_\mathcal{A}$ corresponds, in molecular biology, to what is called 'transition', the other two to 'transversions'.

Now, suppose we have an X-tree $T = (V_T, E_T)$ and a vertex assignment $\psi\colon V_T \to \mathcal{A}^k$ that assigns a sequence $\psi(v)$ in $\mathcal{A}^k$ to each vertex v in T, and consider, with $\mathcal{G} := (\mathcal{G}_\mathcal{A})^k$ as above, the map

$$\lambda = \lambda_\psi\colon E_T \to \mathcal{G}\colon \{u, v\} \mapsto \tau^{(k)}_{\psi(u)\to\psi(v)} \tag{1.1}$$

and the associated map $D_\lambda = D_{\lambda_\psi} \in \mathcal{P}(X, 3 \mid \mathcal{G})$.

Note that $D_\lambda(xy) = \tau^{(k)}_{\psi(x)\to\psi(y)}$ holds for all $x, y \in X$. Thus, for any subset Y of X in $\binom{X}{\leq 2}$, the value $D_\lambda(Y)$ is determined solely by the restriction $f := \psi|_X$ of the map ψ to X — a property that does not extend to the 3-element subsets of X.

Nevertheless, one can design various procedures for selecting a value $D(xyz) \in \mathcal{G} = (\mathcal{G}_\mathcal{A})^k$ for 3-element sets $\{x, y, z\} \in \binom{X}{\leq 3}$ that D_λ could take purely as a function of f. In case $\#\mathcal{A} = 2$, a rather natural way to define $D(xyz) \in \mathcal{G} = (\mathcal{G}_\mathcal{A})^k$ is, after identifying $\mathcal{G}_\mathcal{A}$ with the group $\mathbb{Z}_2 := \mathbb{Z}/2\mathbb{Z}$, to define its i-th component $D_i(xyz)$, for any three elements x, y, z in X, by $D_i(xyz) := 0 \mod 2$ in case the i-th components $f_i(x), f_i(y), f_i(z)$ of the three sequences $f(x), f(y)$, and $f(z)$ in $\mathcal{A}^k$ coincide, and $D_i(xyz) := 1 \mod 2$ else. In other words, one may define

$$D_i(xyz) := \#\{f_i(x), f_i(y), f_i(z)\} - 1 \mod 2$$

for each $i = 1, \ldots, k$. Note that the resulting map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$, at least, extends the restriction $D^{(2)}_\lambda$ of $D_\lambda = D_{\lambda_\psi}$ to $\binom{X}{\leq 2}$, i.e., one has $D(xxy) = D_\lambda(xy)$ for all $x, y \in X$. Later (in Proposition 3.7), we will characterize those maps $f\colon X \to \mathcal{A}^k$ for which this particular choice of D is of the form $D = D_\lambda$ for some $\mathcal{G}$-valued edge assignment of an X-tree T.

1.3. Outline of Results to Come

Given an X-tree $T = (V_T, E_T)$ and a $\mathcal{G}$-valued edge assignment λ for T, the support $\mathrm{supp}(\lambda_X)$ $\big(:= \{S \in \mathcal{S}(X)\colon \lambda_X(S) \neq 0_\mathcal{G}\}\big)$ of λ_X is necessarily a compatible split system* for X as it is contained in the set

$$\mathcal{S}(X|T) = \{S_X(e)\colon e \in E_T\} \subseteq \mathcal{S}^*(X)$$

* A split system $\mathcal{S} \subseteq \mathcal{S}(X)$ is called *compatible* if any two splits are compatible, i.e., if for all $\{A, B\}, \{A', B'\}$ in $\mathcal{S}$, one of the four intersections $A \cap A'$, $A \cap B'$, $B \cap A'$, and $B \cap B'$ is empty.

of all nontrivial X-splits associated with the edges of T. More specifically, defining

$$\Lambda(X|\mathcal{G}) := \{\mu \in \mathcal{S}^*(X|\mathcal{G}) \colon \text{supp}(\mu) \text{ is a compatible split system for } X\},$$

it follows from the standard results regarding X-trees and compatible split systems quoted above that $\Lambda(X|\mathcal{G})$ consists exactly of all those maps μ from $\mathcal{S}^*(X)$ into $\mathcal{G}$ that are of the form λ_X for some X-tree T and some $\mathcal{G}$-valued edge assignment λ for T in which case a unique such pair (T, λ) exists — unique up to canonical isomorphism, of course — for which $\lambda(e) \neq 0_{\mathcal{G}}$ holds for every edge e in T.

Given any integer $s \in \mathbb{N}$, we will denote by $\Theta_s(X|\mathcal{G})$ the map

$$\Theta_s(X|\mathcal{G}) \colon \Lambda(X|\mathcal{G}) \to \mathcal{P}(X, s\,|\,\mathcal{G}) \colon \mu \mapsto D_\mu^{(s)},$$

and we will say that a map $D \in \mathcal{P}(X, s\,|\,\mathcal{G})$ has an *arboreal representation* (of order s) if D is contained in the image of $\Theta_s(X|\mathcal{G})$. We will mostly be concerned in this paper with the cases $s = 2$ and $s = 3$ as these will turn out to be the crucial values to be considered. In the remainder of this paper we will:

- recall results from [2] which state that, in case $\mathcal{G}$ contains no element of order 2, the map $\Theta_2(X|\mathcal{G})$ is an injective map from $\Lambda(X|\mathcal{G})$ onto the set of all maps $D\colon \binom{X}{\leq 2} \to \mathcal{G}$ for which $D(x) = 0$,
$$D(xy) + D(yz) + D(zx) \in 2\mathcal{G} \quad (:= \{2\gamma \colon \gamma \in \mathcal{G}\}),$$
and
$$\#\{D(xy) + D(uv), D(xu) + D(yv), D(xv) + D(yu)\} \leq 2$$
holds for all $x, y, z, u, v \in X$,
- show that the map $\Theta_3(X|\mathcal{G})$ is always injective (i.e., it is injective for every abelian group $\mathcal{G}$ whether it contains elements of order 2 or not),
- characterize the image of $\Theta_3(X|\mathcal{G})$ in a similar (yet slightly more complicated) fashion,
- and describe some further results that are particular to elementary abelian 2-groups.

To establish these results, the following definition will be crucial:

Definition 1.1. *For any map $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ in $\mathcal{P}(X, 3\,|\,\mathcal{G})$, any $x \in X$, and any two further elements $a, b \in X$, set*

$$D_x(ab) := D(abx) - D(ab). \tag{1.2}$$

Notice that D_x is symmetric, i.e., we have $D_x(ab) = D_x(ba)$ for all $a, b \in X$, and that $D_x(ab) = 0_{\mathcal{G}}$ holds for all a, b in X with $x \in \{a, b\}$. Note also that, when $\mathcal{G}$ is the additive group of real numbers and

$$2D(abc) = D(ab) + D(bc) + D(ca)$$

holds for all $a, b, c \in X$, then D_x is precisely the *Farris transform* of D defined by

$$D_x^F(a, b) := \frac{1}{2}(D(ax) + D(bx) - D(ab)),$$

(see [9] for a recent survey of uses of this transformation, and [6] for further recent applications). For a general abelian group, D_x will play the same role as this transform (when $\mathcal{G}$ has elements of order 2, the direct group-theoretic analogue of the Farris transform is not well defined).

2. Injectivity Results

The following result was shown in [2]:

Proposition 2.1. *The map $\Theta_2(X|\mathcal{G})$ from $\Lambda(X|\mathcal{G})$ to $\mathcal{P}(X,2\,|\,\mathcal{G})$ is injective whenever $\mathcal{G}$ contains no elements of order 2. Moreover, any map $\mu \in \Lambda(X|\mathcal{G})$ can be reconstructed from its image $D_\mu^{(2)} = \Theta_2(X|\mathcal{G})(\mu)$ by a polynomial-time algorithm (in #X).*

This proposition can be used to provide a weak result for general abelian groups as follows: Let $\mathcal{G}_2$ denote the 2-torsion subgroup of $\mathcal{G}$ consisting of all elements having order a power of 2. Then, given an X-tree T and a $\mathcal{G}$-valued edge assignment λ that takes values in $\mathcal{G} - \mathcal{G}_2$, only, on all edges of T, one can reconstruct T (but not λ!) from $D_\lambda^{(2)}$. To see this, we simply observe that the quotient group $\mathcal{G}/\mathcal{G}_2$ has no elements of order 2 and if $\psi\colon\ \mathcal{G} \to \mathcal{G}/\mathcal{G}_2$ denotes the quotient map, then $\psi \circ \lambda$ provides a $\mathcal{G}/\mathcal{G}_2$-valued edge assignment for T that is never equal to the identity element of $\mathcal{G}/\mathcal{G}_2$ by the assumption that λ takes values in $\mathcal{G} - \mathcal{G}_2$. Since $\psi \circ D = D_{\psi\circ\lambda}$, the claim now follows from Proposition 2.1.

The exclusion of elements of order 2 from $\mathcal{G}$ is necessary in order for $D_\mu^{(2)}$ to determine T as the following example shows:

Example 2.2. For $X = \{1, 2, \ldots, 6\}$, consider the X-trees with leaf set X that have the shape shown in Figure 2. Up to isomorphism, there are precisely 15 such X-trees. Take $\mathcal{G} := \mathbb{Z}_2$ and, for each of these trees, consider the constant $\mathcal{G}$-valued edge assignment λ that assigns the nonzero element of $\mathcal{G}$ to each of its edges. Then, $D_\lambda(xx') = 0_\mathcal{G}$ holds for all $x, x' \in X$ and, so, each of these 15 trees induce the same map $D_\lambda^{(2)}$ even though their respective edge assignments λ are never equal to $0_\mathcal{G}$.

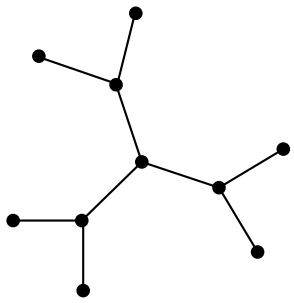

Figure 2.

Returning to the case where $\mathcal{G}$ is a group without elements of order 2, D has an arboreal representation precisely if, for any (or for all) $x \in X$, its Farris transform D_x satisfies the properties of a 'symbolic ultrametric' on $X - \{x\}$ and, for this situation, we can apply the *representation theory of symbolic ultrametrics* as developed in [4]. We will show that the same holds for D_x as defined in Definition 1.1, and we first recall the theory of symbolic ultrametrics.

For an X-tree $T = (V, E)$, any map t from V into an arbitrary set M is called a *symbolic dating map for* T. Further, t is said to be *discriminating* if $t(u) \neq t(v)$ holds for all edges $\{u, v\}$ of T. Observe that no restrictions are placed on M. Now, define a quadruple (X, T, t, z) with

- X a finite set, $T = (V, E)$ an X-tree, t symbolic dating map for T, and z an element of X

to be a *symbolic representation* of a pair (Y, δ) where Y is a finite set and δ is a map from $\binom{Y}{\leq 2}$ into M if $Y = X - \{z\}$ holds, and we have

$$\delta(xy) = t\left(\mathrm{Med}_T(x, y, z)\right)$$

for all $x, y \in Y$ — here Med_T is the median function that assigns, to any three vertices u, v, w of T the unique vertex $\mathrm{Med}_T(u, v, w)$ in T that is on each of the three paths in T connecting the three vertices u, v, w. Similarly, (X, T, t, z) is called a *discriminating* symbolic representation of (Y, δ) if, in addition, t is discriminating and z is not a leaf in T.

Theorem 2.3 (below) characterizes those maps $\delta\colon \binom{Y}{\leq 2} \to M$ that have a (discriminating) symbolic representation. Indeed, define δ to be a *symbolic ultrametric* (*on* Y) if the following two conditions are satisfied:

(U1) $\#\{\delta(ab), \delta(bc), \delta(ca)\} \leq 2$ holds for all $a, b, c \in Y$,
(U2) and there are no elements $a, b, c, d \in Y$ with

$$\delta(ab) = \delta(bc) = \delta(cd) \neq \delta(bd) = \delta(da) = \delta(ac).$$

Then, the following holds (cf. [4]):

Theorem 2.3. *Let Y and M be finite sets, and let δ be a map from $Y \times Y$ into M. Then, there exists a symbolic representation of δ if and only if δ is a symbolic ultrametric in which case there exists* (*up to isomorphism*) *a unique discriminating symbolic representation of δ.*

We pause to note an interesting consequence of this result for molecular systematics that had been the motivation for developing the theory presented in [4]:

Corollary 2.4. *Suppose one has a collection X of species whose evolution is described by some unknown X-tree $T = (V, E)$. Suppose further that a 'phylogenetic character' $\chi\colon V \to M$ with values in some set M can be defined such that, for any three species $x, y, z \in X$, it is possible to accurately reconstruct the value $\chi(x, y, z) \in M$ that χ attains for the ancestral species $a(x, y, z)$ that is represented by the vertex $\mathrm{Med}_T(x, y, z)$ of T. Let T' denote the X-tree whose splits correspond to those edges of T for which the character values at its endpoints are distinct* (*obtained from T by contracting all other edges of T*). *Then, T' can be accurately reconstructed from these data. In particular, T itself can be reconstructed from these data provided $\chi(u) \neq \chi(v)$ holds for all $u, v \in V$ with $\{u, v\} \in E$.*

Proof. Select any element $z \in X$ and define $D\colon \binom{X-\{z\}}{\leq 2} \to M$ by $D(xy) := \chi(z, x, y)$. Then, D is a symbolic ultrametric and has a unique discriminating symbolic representation on T'. This completes the proof. ∎

Returning to arbitrary abelian groups, we have the following result concerning the injectivity of the map $\Theta_3(X|\mathcal{G})$ from $\Lambda(X|\mathcal{G})$ to $\mathcal{P}(X, 3\,|\,\mathcal{G})$.

Theorem 2.5. *Given an X-tree $T = (V_T, E_T)$ and a $\mathcal{G}$-valued edge assignment λ of T with $\lambda(e) \neq 0_{\mathcal{G}}$ for all $e \in E_T$, one can, by a polynomial-time algorithm, reconstruct T and λ (up to canonical isomorphism) from the map $D := D_\lambda^{(3)}$.*

Proof. First, suppose that $D = D_{\lambda'}^{(3)}$ holds for some pair (T', λ') where $T' = (V', E')$ is an X-tree and λ' is a $\mathcal{G}$-valued edge assignment for T' that does not vanish on the edges of T'. We first show that T' is isomorphic to T: Consider a fixed element $z \in X$ and define a symbolic dating map $t' = t'_{\lambda'}$ on the vertices v' of T' as follows: Set $t'(v')$ equal to the sum of the group elements assigned by λ' to the edges of T' on the path from z to v'. As $\lambda'(e') \neq 0_{\mathcal{G}}$ holds, by assumption, for any edge e' of T', t' is a discriminating map (i.e., $t'(u') \neq t'(v')$ holds for every edge $\{u', v'\}$ of T'). Furthermore, if $v' = \mathrm{Med}_{T'}(x, y, z)$ holds for some vertex v' in V' and two elements x, y in X, the definition of $D_z = (D_{\lambda'}^{(3)})_z$ implies immediately that $t'(v') = D_z(xy)$ holds. Thus, (T', t') provides a discriminating symbolic representation of D_z considered as a map from $(X - \{z\}) \times (X - \{z\})$ into $\mathcal{G}$.

However, in view of Theorem 2.3, there is at most one such pair (T', t') that does so. Thus, T and T' must be canonically isomorphic and — modulo this isomorphism — $t'_{\lambda'}$ must coincide with the symbolic dating map $t = t_\lambda$ that is induced by λ. Moreover, we have $\lambda(e) = t(v) - t(u)$ for any edge $e = \{u, v\}$ in E_T for which u is contained in the path connecting z and v in T. Thus, λ is also determined by t.

Finally, one can reconstruct T and λ from D in polynomial time by applying, e.g., the algorithm described in [17] for constructing a discriminating symbolic representation of a symbolic ultrametric. ∎

Remark 2.6. Note, by the way, that Theorem 2.5 easily implies Proposition 2.1: This follows directly from the fact that, in view of the identity $2D(x_1x_2x_3) = D(x_1x_2x_3 : 2)$ that is easily seen to hold[†] for any map D in $\mathcal{P}(X, 3\,|\,\mathcal{G})$ that is of the form $D = D_\mu^{(3)}$ for some map $\mu \in \Lambda(X|\mathcal{G})$, the map $D^{(2)}$ determines the map D uniquely in case $\mathcal{G}$ contains no elements of order 2.

3. The Image of $\Theta_3(X|\mathcal{G})$

We now consider the following question: How can we characterize those maps $D \in \mathcal{P}(X, s\,|\,\mathcal{G})$ that do have an arboreal representation? In the light of Theorem 2.5, we will be concerned only with the cases $s = 2, 3$.

To begin with, let us suppose that, for an arbitrary abelian group $\mathcal{G}$, we have a map $D \in \mathcal{P}(X, 2\,|\,\mathcal{G})$. If there exists an X-tree $T = (V_T, E_T)$ and a $\mathcal{G}$-valued edge assignment $\lambda\colon E_T \to \mathcal{G}$ with $D = D_\lambda^{(2)}$, then D satisfies the so-called *generalized four-point condition* (with respect to $\mathcal{G}$): This condition states that $D(a) = 0$ holds for every

[†] Indeed, it will be shown in [1] that a map D in $\mathcal{P}(X, 3\,|\,\mathcal{G})$ is of the form $D = D_\mu^{(3)}$ for some map $\mu \in \mathcal{S}^*(X|\mathcal{G})$ if and only if $D(x_1x_2x_3x_4 : 3) = D(x_1x_2x_3x_4 : 2)$ and $D(x) = 0_{\mathcal{G}}$ holds for all $x_1, x_2, x_3, x_4, x \in X$ which readily (putting $x_3 = x_4$) also implies that $2D(x_1x_2x_3) = D(x_1x_2x_3 : 2)$ must hold for any such map D and all $x_1, x_2, x_3 \in X$.

$a \in X$ and that, for any four (not necessarily distinct) elements a, b, c, d in X, one has

$$\#\{D(ab)+D(cd), D(ac)+D(bd), D(ad)+D(bc)\} \leq 2$$

as well

$$D(ab)+D(bc)+D(ca) \in 2\mathcal{G}.$$

In [2], it was shown that, provided $\mathcal{G}$ has no element of order 4, this generalized four-point condition is not only a necessary, but also a sufficient condition for a map D as above to have an arboreal representation (of order 2). Moreover, when $\mathcal{G}$ is an elementary abelian 2-group, the following holds:

Proposition 3.1. *For any map D from $\binom{X}{\leq 2}$ into an elementary abelian 2-group $\mathcal{G}$, the following assertions are equivalent*:

(i) *D has an arboreal representation on some phylogenetic X-tree*;
(ii) *D has an arboreal representation on* every *phylogenetic X-tree*;
(iii) *D satisfies the generalized four-point condition with respect to $\mathcal{G}$*;
(iv) *there exists a map $f\colon X \to \mathcal{G}$ such that $D(xy) = f(x)+f(y)$ holds for all $x, y \in X$.*

Furthermore, if all of this holds, f is clearly uniquely determined by its value on one single point $z \in X$ while, conversely, one may specify any value γ for $f(z)$ and put $f(x) := D(xz)+\gamma$ for every $x \in X$, to construct a map $f\colon X \to \mathcal{G}$ with the required properties.

Proof. The equivalence of (i)–(iii) was established in [2, Proposition 2]. To establish the equivalence of (i) and (iv), first suppose that there exists a map f having the property described in part (iv), and consider the phylogenetic X-tree T with leaf set X that has no interior edges (the 'star' tree). For any edge $e \in E$, let x_e denote the leaf $x \in X$ contained in e and put $\lambda(e) := f(x_e)$. Then, it is easily seen that $D = D_\lambda$ holds.

Conversely, suppose that $D = D_\lambda$ holds for some X-tree $T = (V_T, E_T)$ and some $\mathcal{G}$-valued edge assignment $\lambda\colon E_T \to \mathcal{G}$, and select an arbitrary element $z \in X$ and an arbitrary element $\gamma \in \mathcal{G}$. For each $x \in X$, let $E(x)$ denote the set of edges occurring in the unique shortest path from z to x in T, i.e., put $E(x) := \{e \in E(T), S_X(e) \in \mathcal{S}^*(X|\{z,x\})\}$, and put

$$f(x) := \gamma + \sum_{e \in E(x)} \lambda(e).$$

Then,

$$f(z)+f(x) = 2\gamma + \sum_{e \in E(x)} \lambda(e) = D_\lambda(z,x) = D(z,x)$$

holds since 2γ vanishes in $\mathcal{G}$. Furthermore, because D takes values in the elementary abelian 2-group $\mathcal{G}$, we have

$$D(xy) = D(xz)+D(zy) = (f(z)+f(x))+(f(z)+f(y)) = f(x)+f(y),$$

as required. The uniqueness of f, once z and $f(z)$ are specified, is clear from the identity $f(x) = D(xz)+f(z)$. ∎

We now turn to the question of arboreal representations of order 3. As any map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$ that has an arboreal representation is necessarily of the form $D = D_\mu^{(3)}$ for some map $\mu \in \mathcal{S}^*(X|\mathcal{G})$, a necessary condition for a map D in $\mathcal{P}(X, 3 \mid \mathcal{G})$ to have an arboreal representation (see [1] for more details) is that

$$D(x) = 0_{\mathcal{G}} \tag{3.1}$$

and

$$D(x_1x_2x_3x_4 : 2) = D(x_1x_2x_3x_4 : 3) \tag{3.2}$$

holds for all $x, x_1, x_2, x_3, x_4 \in X$. Regarding these identities, the following observations are of some interest and easily established: Given any map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$,

(i) the identity

$$2D(x_1x_2x_3) = D(x_1x_2x_3 : 2) \tag{3.3}$$

holds for all $x_1, x_2, x_3 \in X$ provided that D satisfies the conditions (3.1) and (3.2) for all $x, x_1, x_2, x_3, x_4 \in X$,

(ii) the identity (3.3) holds for all $x_1, x_2, x_3 \in X$ if and only if $D(x) = 0$ holds for all $x \in X$ and (3.3) holds for any three distinct elements x_1, x_2, x_3 in X,

(iii) if (3.3) holds for all $x_1, x_2, x_3 \in X$, then (3.2) holds for all $x_1, x_2, x_3, x_4 \in X$ with $\#\{x_1, x_2, x_3, x_4\} \leq 3$, $2D(x_1x_2x_3x_4 : 2) = 2D(x_1x_2x_3x_4 : 3)$ holds for all x_1, x_2, x_3, $x_4 \in X$, and one has

$$2D_x(ab) = 2D(xab) - 2D(ab) = D(xa) + D(xb) - D(ab),$$

for all $x, a, b \in X$,

(iv) if (3.3) holds for all $x_1, x_2, x_3 \in X$ and x is a fixed element in X, then (3.2) holds for all x_1, x_2, x_3, x_4 in X if and only if it holds for all x_1, x_2, x_3, x_4 in X with $x \in \{x_1, x_2, x_3, x_4\}$.

The proofs are left, as a simple exercise, to the reader.

Thus, if (3.3) holds, for all x_1, x_2, x_3 in X, for a map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$ and if $\mathcal{G}$ has no elements of order 2, then $D^{(2)}$ satisfies the generalized four-point condition if and only if

$$\#\{D_x(ab), D_x(bc), D_x(ca)\} \leq 2$$

holds for all $a, b, c, x \in X$: Indeed, putting

$$\gamma := D(ax) + D(bx) + D(cx),$$

we have

$$\gamma - 2D_x(ab) = D(cx) + D(ab), \quad \gamma - 2D_x(bc) = D(ax) + D(bc),$$

and

$$\gamma - 2D_x(ca) = D(bx) + D(ca)$$

and, therefore,

$$\begin{aligned}\#\{D_x(ab), D_x(bc), D_x(ca)\} &= \#\{\gamma - 2D_x(ab), \gamma - 2D_x(bc), \gamma - 2D_x(ca)\} \\ &= \#\{D(ab) + D(cx), D(ac) + D(bx), D(ax) + D(bc)\}.\end{aligned}$$

Thus, it remains to note that (3.3) implies also that

$$D(ab) + D(bc) + D(ca) = 2D(abc) \in 2\mathcal{G}$$

holds for all $a, b, c \in X$.

Consequently, if $\mathcal{G}$ has no elements of order 2, we have:

Proposition 3.2. *A map $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ into an abelian group with no elements of order 2 has an arboreal representation if and only if the following two conditions are satisfied*:

(i) *one has $\#\{D_x(ab), D_x(bc), D_x(ca)\} \leq 2$ for all $a, b, c, x \in X$,*
(ii) *D satisfies the three-point condition* (3.3).

It follows that, if $\mathcal{G}$ has no elements of order 2, then $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ has an arboreal representation if and only if, for all subsets U of X of size at most four, its restriction

$$D_U\colon \binom{U}{\leq 3} \to \mathcal{G}\colon A \mapsto D(A)$$

to $\binom{U}{\leq 3}$ has an arboreal representation. Note also that, for all $a, b, c, x \in X$, we have

$$\#\{D_x(ab), D_x(bc), D_x(ca)\} \leq 2, \tag{3.4}$$

for every abelian group $\mathcal{G}$ and any map $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ that has an arboreal representation.

However, this four-point condition (even if combined with other three- or four-point conditions) is not sufficient for D to have an arboreal representation in case $\#X \geq 5$:

Example 3.3. Let X denote the set of vertices of a regular pentagon P in the plane, and define a map $D\colon \binom{X}{\leq 3} \to \mathbb{Z}_2$ as follows: For each $A \in \binom{X}{2}$, set $D(A) := 0$; and for $A \in \binom{X}{3}$, set $D(A) := 1$ if the vertices in A appear consecutively in P, otherwise set $D(A) = 0$. Then, the restriction

$$D_U\colon \binom{U}{\leq 3} \to \mathbb{Z}_2\colon A \mapsto D(A)$$

of D to each 4-subset U of X has an arboreal representation; yet, D itself does not.

In contrast, we will see below (Theorem 3.5) that, for any abelian group $\mathcal{G}$, a map $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ has an arboreal representation if (and only if) the restriction

$$D_U\colon \binom{U}{\leq 3} \to \mathbb{Z}_2\colon A \mapsto D(A)$$

of D relative to each subset U of X of cardinality at most 5 has such a representation, and we will provide a simple *five-point condition* that, combined with the three- and four-point conditions discussed above, characterizes those maps D that have an arboreal representation. First note that characterizing arboreal representations is particularly easy in case $\#X = 3$:

Lemma 3.4. *Suppose that $\mathcal{G}$ is an arbitrary abelian group and X has cardinality* 3. *Then, a map $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ has an arboreal representation if and only if D satisfies the three-point condition* (3.3)*.*

Proof. The necessity of (3.3) is clear. Conversely, suppose that (3.3) holds and consider the tree T that has X as its set of leaves, all attached to an unlabelled vertex of degree 3. For $x \in X$, assign the $\mathcal{G}$-value $\lambda(e_x) := D(X) - D(X - \{x\})$ to the edge e_x incident with the leave x. Then, $D(xy) = \lambda(e_x) + \lambda(e_y) = D_\lambda(xy)$ holds for any two distinct elements $x, y \in X$ as $x, y \in X$ and $x \neq y$ (together with our assumptions $\#X = 3$ and (3.3)) implies that

$$2D(X) = D(X - \{x\}) + D(X - \{y\}) + D(xy)$$

and, hence,

$$\begin{aligned} D_\lambda(xy) &= \lambda(e_x) + \lambda(e_y) \\ &= D(X) - D(X - \{x\}) + D(X) - D(X - \{y\}) \\ &= 2D(X) - D(X - \{x\}) - D(X - \{y\}) \\ &= D(xy) \end{aligned}$$

must hold while

$$\begin{aligned} D_\lambda(X) &= \sum_{x \in X} \lambda(e_x) \\ &= \sum_{x \in X} \big(D(X) - D(X - \{x\})\big) \\ &= 3D(X) - \sum_{x \in X} D(X - \{x\}) \\ &= 3D(X) - 2D(X) \\ &= D(X) \end{aligned}$$

follows again from applying (3.3) to the 3-element set X as this implies also that $2D(X) = \sum_{x \in X} D(X - \{x\})$ must hold. ∎

We now state the main result of this section:

Theorem 3.5. *Given a set X of cardinality at least* 4*, an element $x_0 \in X$, an arbitrary abelian group $\mathcal{G}$, and a map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$, the following assertions are equivalent*:

(i) *D has an arboreal representation*;
(ii) *The following four conditions all hold*:

(3PC) *For all $a, b, c \in X$, D satisfies the three-point condition* (3.3)*,*
(4PC) *for all $x, a, b, c \in X$, D satisfies the four-point condition* (3.4)*,*
(4PC*) *the identity* (3.2) *holds for all $x, a, b, c \in X$,*

(5PC) *for all* $x, a, b, c, d \in X$, D *satisfies the following* 5*-point condition*:

$$D_x(ab) = D_x(bc) = D_x(cd)$$

together with

$$D_x(ca) = D_x(ad) = D_x(db)$$

implies

$$D_x(ab) = D_x(ca);$$

(iii) D *satisfies the conditions* (4PC), (4PC*), *and* (5PC) *described in part* (ii) *for* $x := x_0$ *and all* $a, b, c, d \in Y := X - \{x_0\}$.

In particular, a map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$ *has an arboreal representation if and only if the restriction* D_U *relative to every subset* U *of* X *of size at most* 5 *has an arboreal representation.*

Proof. Given a map $D \in \mathcal{P}(X, 3 \mid \mathcal{G})$ and an element $x \in X$, let D_x in $\mathcal{P}(X - \{x\}, 2 \mid \mathcal{G})$ be defined as in (1.2). To establish the implication (i) $\Rightarrow$ (ii), suppose that D has an arboreal representation and choose some phylogenetic X-tree $T = (V, E)$ and some map $\lambda\colon E \to \mathcal{G}$ that is not equal to $0_{\mathcal{G}}$ for any interior edge e with $D = D_\lambda$. It follows that the map D_x is a symbolic ultrametric on $X - \{x\}$ since it has a (unique) discriminating symbolic representation as described in the proof of Theorem 2.5. By Theorem 2.3, it follows that D_x satisfies properties (U1) and (U2) which are precisely the conditions (4PC) and (5PC). Condition (4PC*) can also be verified directly.

The implication (ii) $\Rightarrow$ (iii) is trivial, so it remains to establish the implication (iii) $\Rightarrow$ (i).

So, suppose that D satisfies the properties (3PC), (4PC), (4PC*), and (5PC) for $x := x_0$ and all $a, b, c, d \in Y = X - \{x_0\}$. Conditions (4PC) and (5PC) imply that D_{x_0} satisfies the properties (U1) and (U2) required for a symbolic ultrametric on Y, and so, again by Theorem 2.3, D_{x_0} has a discriminating symbolic representation (X', T', t', z) where z is an arbitrary additional element not contained in X or in any other set considered so far, X' is the set $Y \cup \{z\}$, $T' = (V', E')$ is an X'-tree, and $t'\colon V' \to \mathcal{G}$ is an appropriate symbolic dating map, i.e., a map such that

$$t'(\mathrm{Med}(a, b, z)) = D_{x_0}(ab) = D(abx_0) - D(ab)$$

holds for all $a, b \in Y$.

Now, consider the X-tree $T := (V, E)$ with vertex set $V := V' \cup \{x_0\}$ and edge set $E := E' \cup \{\{x_0, z\}\}$, define a dating map t for T that just extends t' by putting $t(x_0) := 0_{\mathcal{G}}$, and note that

$$t(\mathrm{Med}(a, b, x_0)) = D_{x_0}(ab) = D(abx_0) - D(ab),$$

and hence, in particular, also

$$t(a) = t(\mathrm{Med}(a, a, x_0)) = D_{x_0}(a) = D(x_0 a)$$

holds for all $a, b \in X$.

Define a $\mathcal{G}$-valued edge assignment $\lambda = \lambda_t$ for T by putting $\lambda(\{u, v\}) := t(v) - t(u)$ for every edge $\{u, v\}$ of T for which u is on the path between v and x_0. For any two vertices u, v in V, let $\lambda(u, v)$ denote the sum of the λ-values of the edges on the path from u to v so that $\lambda(v, u) = t(v) - t(u)$ holds for any two vertices u, v in V for which u is on the path between v and x_0 and, therefore, in particular, $\lambda(v, x_0) = t(v) - t(x_0) = t(v)$ for all v in V.

Next, put $D' := D_\lambda^{(3)}$ and note that $D'(ab) = \lambda(a, b)$ holds for all a, b in X. We claim that $D' = D$ holds. Indeed, given any two elements a, b in X, we have

$$D'(ax_0) = \lambda(a, x_0) = t(a) = D(x_0a),$$

as well as, with $v := \mathrm{Med}_T(a, b, x_0)$ and, hence,

$$t(v) = D_{x_0}(ab) = D(abx_0) - D(ab),$$

also

$$\begin{aligned} D'(ab) &= \lambda(a, v) + \lambda(b, v) \\ &= (t(a) - t(v)) + (t(b) - t(v)) \\ &= D(x_0b) + D(x_0a) - 2(D(x_0ab) - D(ab)) \\ &= D(x_0b) + D(x_0a) - (D(x_0b) + D(x_0a) + D(ab)) + 2D(ab) \\ &= D(ab) \end{aligned}$$

and

$$\begin{aligned} D'(abx_0) &= \lambda(a, v) + \lambda(b, v) + \lambda(x_0, v) \\ &= D(ab) + t(v) \\ &= D(ab) + (D(abx_0) - D(ab)) \\ &= D(abx_0). \end{aligned}$$

Thus, noting that $D(x_1x_2x_3x_4 : 2) = D(x_1x_2x_3x_4 : 3)$ holds for all x_1, x_2, x_3, x_4 in X by assumption and $D'(x_1x_2x_3x_4 : 2) = D'(x_1x_2x_3x_4 : 3)$ by construction and that, therefore $D(x_1x_2x_3x_4 : 3) = D'(x_1x_2x_3x_4 : 3)$ must also hold for all x_1, x_2, x_3, x_4 in X, we see finally that

$$\begin{aligned} D(abc) &= D(abcx_0 : 3) - D(abx_0) - D(acx_0) - D(bcx_0) \\ &= D'(abcx_0 : 3) - D'(abx_0) - D'(acx_0) - D'(bcx_0) \\ &= D'(abc) \end{aligned}$$

also holds for all a, b, c in X. ∎

Remarkably, one can also derive Proposition 3.2 as a simple corollary of Theorem 3.5. To this end we first establish the following result:

Lemma 3.6. *If a map* $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ *satisfies the three-point and the four-point condition* (3PC) *and* (4PC)*, then one has* $4g_1 = 4g_2$ *for any two elements* $g_1, g_2 \in \mathcal{G}$ *for which there exists five elements* $a, b, c, d, x \in X$ *with* $g_1 = D_x(ab) = D_x(bc) = D_x(cd)$ *and* $g_2 = D_x(ca) = D_x(ad) = D_x(db)$.

Proof. Note that, putting

$$g := D(xa) + D(xb) + D(xc) + D(xd),$$

our assumptions imply that

$$D(ab) + D(cd) = g - 4g_1,$$

$$D(ac) + D(bd) = g - 4g_2,$$

and

$$D(ad) + D(bc) = g - 2g_1 - 2g_2$$

holds. Thus, (4PC) implies that either $4g_1 = 4g_2$, $4g_1 = 2g_1 + 2g_2$, or $4g_2 = 2g_1 + 2g_2$ holds which in turn implies that $4g_1 = 4g_2$ must, in any case, hold, as claimed. ∎

To derive Proposition 3.2 from Theorem 3.5 suppose that $D\colon \binom{X}{\leq 3} \to \mathcal{G}$ satisfies the three-point and the two four-point conditions (3PC) and (4PC). Then, if $\mathcal{G}$ has no elements of order 2, it is easily checked that D also satisfies condition (4PC*), and Lemma 3.6 implies that D also satisfies condition (5PC). Thus, by Theorem 3.5, D has an arboreal representation. ∎

Finally, we return to the specific setting where sequences over a 2-element set $\mathcal{A}$ are associated to the elements in X by a map

$$f = (f_1, \ldots, f_k)\colon X \to \mathcal{A}^k,$$

and, identifying $\mathcal{G}_{\mathcal{A}}$ with the group $\mathbb{Z}_2$ as in Section 1.2, we consider further aspects regarding arboreal representations of the associated map

$$D = D_f\colon \binom{X}{\leq 3} \to {\mathbb{Z}_2}^k\colon \{x, y, z\} \to (D_i(xyz))_{i=1,\ldots,k} \tag{3.5}$$

defined, as above, by

$$D_i(xyz) := \#\{f_i(x), f_i(y), f_i(z)\} - 1 \mod 2.$$

An obvious question arises in this situation: For which functions $f\colon X \to \mathcal{A}^k$ does D_f have an arboreal representation? Our final proposition shows that, informally speaking, D_f has an arboreal representation if and only if the X-splits induced by f are *compatible* (*cf.* the footnote in Section 1.3).

To make this more precise, we introduce the following notation: Given a map $f = (f_1, \ldots, f_k)\colon X \to \mathcal{A}^k$, set $S_i(f) := \{f_i^{-1}(\alpha_1), f_i^{-1}(\alpha_2)\}$ where α_1, α_2 denote the two elements in $\mathcal{A}$, and set

$$\mathcal{S}(f) = \{S_i(f)\colon i = 1, 2, \ldots, k\} \cap \mathcal{S}^*(X).$$

It is a well-known and classical result that the 'Buneman complex' $B(\mathcal{S}(f))$ obtained from $\mathcal{S}(f)$ is a tree — and can, thus, be viewed in a natural way as an X-tree — if and only if $\mathcal{S}(f)$ is a compatible split system (*cf.* [5], see also [7, 8] for the definition of the Buneman complex $B(\mathcal{S})$ associated with an arbitrary split system $\mathcal{S}$). This translates into

Proposition 3.7. *Given a map $f\colon X \to \mathcal{A}^k$, let $D_f\colon \binom{X}{\leq 3} \to (\mathcal{G}_{\mathcal{A}})^k$ be defined as in* (3.5)*. Then D_f has an arboreal representation if and only if $\mathcal{S}(f)$ is a compatible split system in which case*

- *$B(\mathcal{S}(f))$ provides an underlying X-tree for such a representation,*
- *f extends to a map ψ defined on the vertex set of the X-tree $B(\mathcal{S}(f))$ that is defined by identifying, for each vertex v, some elements x, y, z in X so that v is the median of x, y, z in $B(\mathcal{S}(f))$ as defined in Section 2 and then associating to v the median of $f(x), f(y)$, and $f(z)$, i.e., the sequence $f(v) = (f_1(v), \ldots, f_k(v))$ whose i-th component $f_i(v)$ is that element in $\mathcal{A}$ that occurs at least twice among the elements $f_i(x), f_i(y)$, and $f_i(z)$ in the 2-element set $\mathcal{A}$,*
- *and D coincides with the map D_λ associated with the resulting edge weighting λ $(= \lambda_\psi$, as defined in* (1.1)*) of $B(\mathcal{S}(f))$.*

To see (most of) this, all one has to observe is that D_f coincides with the map D_μ for the map $\mu \in \mathcal{S}^*(X|\mathcal{G})$ that maps any split $S \in \mathcal{S}^*(X)$ onto the sequence $(\delta(S, S_1(f)), \ldots, \delta(S, S_k(f)))$ $\big($with $\delta(S, S') := 0 \mod 2$ in case $S \neq S'$ and $\delta(S, S') := 1 \mod 2$ in case $S = S'$, as usual$\big)$.

Remark 3.8. It may be interesting to consider extensions of the results of this paper to *non*-abelian groups, in the setting where each edge has two elements of the group assigned, one for each of the two directions by which the edge can be traversed. This setting was explored in [16] for functions $D\colon X \times X \to \mathcal{G}$ ($\mathcal{G}$ non-abelian), thereby providing non-abelian analogues (and generalizations) of the results from [2]. As in our current paper, elements of order 2 in $\mathcal{G}$ played an important role in [16], so the consideration of functions D from X^3 $\big($rather than from $\binom{X}{3}\big)$ into $\mathcal{G}$ might also be useful in the non-abelian setting.

References

1. A.W.M. Dress, Split decomposition over an abelian group, Part I: Generalities, Ann. Comb., to appear.
2. H.-J. Bandelt and M.A. Steel, Symmetric matrices representable by weighted trees over a cancellative abelian monoid, SIAM J. Discrete Math. **8** (1995) 517–525.
3. G.M. Barker, Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation, Biol. J. Linnean Soc. **76** (2002) 165–194.
4. S. Böcker and A.W.M. Dress, Recovering symbolically dated, rooted trees from symbolic ultrametrics, Adv. Math. **138** (1998) 105–125.
5. P. Buneman, The recovery of trees from measures of dissimilarity, In: Mathematics in the Archaeological and Historical Sciences, F.R. Hodson, D.G. Kendall, and P. Tautu, Eds., Edinburgh University Press (1971) pp. 387–395.

6. A.W.M. Dress, B. Holland, K. Huber, J. Koolen, V. Moulton, and J. Weyer-Menkhoff, Δ additive and Δ ultra-additive maps Gromov's trees, and the Farris transform, Discrete Appl. Math. **146** (2005) 51–73.
7. A.W.M. Dress, M. Hendy, K. Huber, and V. Moulton, On the number of vertices and edges of the Buneman graph, Ann. Comb. **1** (1997) 329–337.
8. A.W.M. Dress, K. Huber, and V. Moulton, Some variations on a theme by Buneman, Ann. Comb. **1** (1997) 339–352.
9. A.W.M. Dress, K. Huber, and V. Moulton, Some uses of the Farris transform in Mathematics and Phylogenetics — A Review, Ann. Comb. **11** (2007) 1–37.
10. S.N. Evans and T.P. Speed, Invariants of some probability models used in phylogenetic inference, Ann. Statist. **21** (1993) 355–377.
11. D.P. Faith, Conservation evaluation and phylogenetic diversity, Biol. Conservation **61** (1992) 1–10.
12. J. Felsenstein, Inferring Phylogenies, Sinauer Associates Sunderland, Mass., 2004.
13. W.J. Heiser and M. Bennani, Triadic distance models: axiomatization and least squares representation, J. Math. Psych. **41** (1997) 189–206.
14. S. Joly and G. Le Calvé, Three-way distances, J. Classification **12** (1995) 191–205.
15. L. Pachter and D. Speyer, Reconstructing trees from subtree weights, Appl. Math. Lett. **17** (6) (2004) 615–621.
16. C. Semple and M.A. Steel, Tree representations of non-symmetric, group-valued proximities, Adv. Appl. Math. **23** (1999) 300–321.
17. C. Semple and M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.
18. M. Steel, Phylogenetic diversity and the greedy algorithm, Systematic Biol. **54** (2005) 527–529.