Annals of Combinatorics 10 (2006) 19-30
0218-0006/06/010019-12
DOI 10.1007/s00026-006-0271-0

Annals of Combinatorics

Accumulation Phylogenies

Mihaela Baroni and Mike Steel

Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
mihaela.baroni@ugal.ro, m.steel@math.canterbury.ac.nz

Received September 13, 2004

AMS Subject Classification: 05C05, 05C20, 92D15

Abstract. Directed acyclic graphs provide a convenient representation of reticulate evolution in systematic biology. In this paper we formalize and analyse a simple model in which evolved characteristics are passed on to all descendant species. We show that the resulting observed sets of characteristics for the species at the leaves uniquely determine the digraph that described the evolution of the species, under certain restrictions. We also provide a characterisation for when this digraph is actually a tree.

Keywords: digraph, tree, reticulate evolution

1. Introduction

Trees and graphs provide useful representations of evolutionary relationships in biology [6–8, 10, 11, 14]. For such graphs, a subset of the vertices correspond to present-day species, and characteristics of these species are used to estimate the structure of the remainder of the graph. One type of data which is becoming increasingly applied is the presence or absence of various genetic signals (the most simple of which is simply the presence or absence of a gene). In this paper we investigate settings in which such signals propagate down the directed graph that represents the evolutionary history of the species. In this model genetic signals arise just once, and are never (completely) lost. We are particularly interested in determining when the underlying network can be uniquely recovered from the accumulated signals at the terminal vertices, and in characterising when this graph is a tree.

The structure of this paper is as follows. In the remainder of this section we introduce hybrid phylogenetic digraphs, and the concept and biological motivation for accumulation phylogenies. In Section 2 we study properties of accumulation maps and in Section 3 we deal with a restricted class of hybrid phylogenies ('regular hybrid phylogenies') for which we can state our main result (Theorem 3.2). In Section 4 we describe a necessary and sufficient condition for the hybrid phylogeny described in Section 3 to be a tree, along with some concluding comments.

1.1. Preliminaries

In this paper we deal with digraphs (directed graphs), which can be represented by a pair (V, A) where V is a set (of *vertices*), and A is a subset of $V \times V$ of *arcs*. For each $v \in V$, let $d^-(v)$ and $d^+(v)$ denote, respectively the in-degree and out-degree (number of arcs for which v is the second, respectively first co-ordinate). In this paper we will be concerned only with *acyclic* digraphs (that is, digraphs that have no directed cycles). More details regarding digraphs can be found in [1]. A *rooted digraph* is an acylic digraph $\mathcal{D} = (V, A)$ with a distinguished vertex ρ (called the *root*) that has the property that $d^-(\rho) = 0$ and for which there exists a directed path from ρ to every vertex in $V - \{\rho\}$.

Let X be a non-empty set, and suppose that $\mathcal{D} = (V, A)$ is a rooted digraph, for which X is the set of vertices of out-degree 0 (the *leaves* of $\mathcal{D}$). We say that $\mathcal{D}$ is a *hybrid phylogeny* (on X) provided that for all $v \in V - X$ with $d^-(v) \leq 1$ we have $d^+(v) > 1$. Two hybrid phylogenies on X, $\mathcal{D}$ and $\mathcal{D}'$ say, are regarded as equivalent if there is a digraph isomorphism from the vertices of $\mathcal{D}$ to the vertices of $\mathcal{D}'$ that restricts to the identity function on X.

Two examples of hybrid phylogenies are shown in Figure 1 (for $X = \{1, 2, 3, 4\}$ and $X = \{1, 2, 3\}$, and with the arcs oriented downwards).

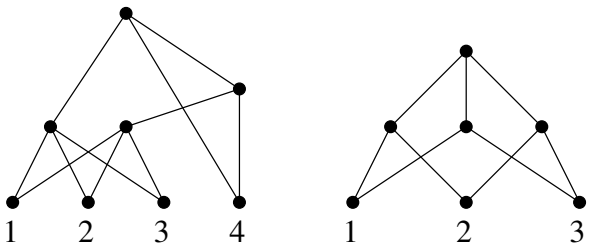

Figure 1: Two examples of hybrid phylogenies. The example on the right is *regular*.

A particular subclass of hybrid phylogenies are those for which (V, A) is a rooted tree - these are usually called *rooted phylogenetic trees* (on X), which we will denote using the symbol $\mathcal{T}$. If in addition $d^+(v) \in \{0, 2\}$ for every vertex v of $\mathcal{T}$ we say that $\mathcal{T}$ is a *rooted binary phylogenetic tree* (on X).

Rooted phylogenetic trees are widely used in evolutionary biology, (and other areas of classification such as linguistics) to represent evolutionary relationships between present-day species, which are represented by the out-degree 0 vertices (the 'leaves'). More recently hybrid phylogenies have been suggested as suitable for modelling evolution in the presence of reticulation (whereby new species contain genetic contributions from two or more ancestral species) [8–14]. Further mathematical background on hybrid phylogenies can be found in [2], while for rooted phylogenetic trees the reader is referred to [15].

1.2. Accumulation Phylogenies

Let S be any (finite or infinite) set. An *accumulation phylogeny* on X is a triple (V, A, f) where (V, A) is a hybrid phylogeny on X with the root ρ, and f is a function from $A \cup \{\rho\}$ to 2^S (the power set of S) that satisfies the following two properties:

(P1) the set $\{f(a)\colon a \in A \cup \{\rho\}\}$ is a collection of pairwise disjoint sets; and

(P2) for each vertex $v \in V - \{\rho\}$ we have $f(a) \neq \emptyset$ for at least one arc a that ends at v.

Note that $f(\rho)$, and $f(a)$ (for certain arcs a) can be the empty set.

1.3. Relevance of Accumulation Phylogenies to Molecular Systematics

In molecular evolutionary biology we may view (V, A) as modelling the evolution of a collection X of extant species from a common ancestor ρ, allowing for hybrid or reticulate evolution. In this setting we may regard S as a set of genes, $f(\rho)$ as the genes present in the most recent common ancestor of the species in X, and for each arc $a = (u, v)$, $f(a)$ denotes the genes that arose in the lineage leading from (ancestral species) u to v. An accumulation phylogeny models the situation where (i) genes arise at most once (condition (P1)), a situation that is commonly assumed and goes under the name 'gene genesis'; and (ii) S is sufficiently extensive that each descendant species acquires at least one new gene (condition (P2)).

Recently, the gene content of species has been used for reconstructing phylogenies (see [16]) by constructing measures of (dis)similarity based on the amount of genes shared by two species (we consider this further in the next section). The approach we consider here may provide an alternative technique for reconstructing the phylogenetic history of species (not necessarily based on a phylogenetic tree) using gene content or other types of genomic markers. Indeed our results will apply to gene content data whenever genes evolve according to the two properties of an accumulation phylogeny (as described in the previous paragraph) and in such a way that whenever a gene arises then either that gene, or some 'trace' of it is identifyable in all descendant species in X.

1.4. Accumulation Maps

Suppose that (V, A, f) is an accumulation phylogeny on X. Consider the map $\Gamma\colon V \to 2^S$, defined as follows. For a vertex $v \in V$, let $\pi(v)$ be the set of arcs in A that lie in at least one path from ρ to v. Then set

$$\Gamma(v) := f(\rho) \cup \left(\bigcup_{a \in \pi(v)} f(a) \right).$$

Let $\gamma := \Gamma | X$, the restriction of the map Γ to X. We refer to γ (respectively Γ) as the *accumulation map* on X (respectively on V) induced by (V, A, f) (or, more briefly, by (V, A)).

Informally, the map Γ describes how the elements of S 'accumulate' as one moves down the digraph, and $\gamma(x)$ describes the resulting subset of S at the leaf x. For example, for either of the hybrid phylogenies shown in Figure 2 together with the values $f(\rho)$ and $f(a)$ as indicated (from the set $\{s_0, s_1, \ldots s_{10}\}$), one has $\gamma(2) = \{s_0, s_1, s_3, s_4, s_5, s_6, s_7\}$. Our interest in this paper is in studying what the $\gamma(x)$ values tell us about the underlying accumulation phylogeny (V, A, f). As Figure 2 shows, different accumulation phylogenies can give rise to the same accumulation map on X.

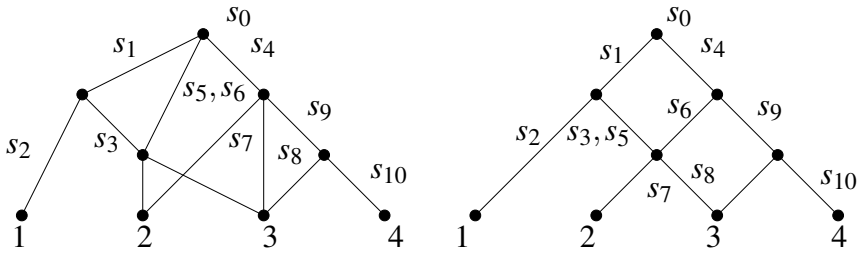

Figure 2: Two accumulation phylogenies on $X = \{1, 2, 3, 4\}$ with the same induced accumulation map γ.

2. Properties of γ and Γ

We now provide three lemmas that describe the basic properties of accumulation maps. These results will be useful for us later in the paper. First we define some terminology.

Consider a map $\gamma\colon X \to 2^S$. For $s \in S$, let

$$A(s) := \{x \in X\colon s \in \gamma(x)\},$$

and consider the associated family of subsets of X:

$$\mathcal{A}(\gamma) := \{A(s) : s \in S\} \cup \{X\}.$$

Suppose we have a hybrid phylogeny $\mathcal{D} = (V, A)$ on X. For $v, v' \in V$ write $v <_{\mathcal{D}} v'$ if there is a directed path in $\mathcal{D}$ from v to v', and $v \leq_{\mathcal{D}} v'$ if $v <_{\mathcal{D}} v'$ or $v = v'$. For $v \in V$, let $c(v) = \{x \in X\colon v \leq_{\mathcal{D}} x\}$, which we refer to as a *cluster* of $\mathcal{D}$.

Lemma 2.1. *Suppose that (V, A, f) is an accumulation phylogeny on X, with induced accumulation map γ, and with $S = \cup_{x\in X}\gamma(x)$.*

(i) *For each arc $a = (u, v) \in A$, and each $s \in f(a)$ we have $A(s) = c(v)$.*
(ii) *For each $s \in S$, $A(s) = c(v)$ for at least one vertex $v \in V$.*
(iii) *For each vertex $v \in V - \{\rho\}$ there exists at least one element $s \in S$ such that $c(v) = A(s)$.*
(iv) *$\mathcal{A}(\gamma) = \{c(v)\colon v \in V\}$. In particular, $\mathcal{A}(\gamma)$ does not depend on the choice of f.*

Proof. Part (i). First suppose that there exists $a \in A$ with $s \in f(a)$, and $a = (u, v)$ for some $u \in V$. If $x \in c(v)$ then $v \leq_{\mathcal{D}} x$ and so $s \in \gamma(x)$. Consequently, $c(v) \subseteq A(s)$. To establish the reverse inclusion, suppose that $x \in A(s)$. Then by (P1) there exists a path from v to x and so $x \in c(v)$. Hence $A(s) \subseteq c(v)$, and thus $A(s) = c(v)$ as claimed.

Part (ii). Suppose that $s \in S$. Since $S = \cup_{x\in X}\gamma(x)$, there exists an element $w \in A \cup \{\rho\}$ with $s \in f(w)$. By (P1) this element w is unique. If $w = \rho$ then $A(s) = X = c(\rho)$. Otherwise, if w is an arc, let v denote the end vertex of w. By *Part* (i) we have that $A(s) = c(v)$.

Part (iii). For any vertex $v \in V - \{\rho\}$, (P2) guarantees that there is at least one arc a that ends at v for which $f(a) \neq \emptyset$. Select any $s \in f(a)$. Then $c(v) = A(s)$ by *Part* (i).

Part (iv). Let $A \in \mathcal{A}(\gamma)$. Then $A = X = c(\rho)$ or there exists $s \in S$ with $A = A(s)$. In the latter case, from (ii), it follows that $A(s) = c(v)$ for some $v \in V$. Conversely, let $v \in V$. If $v = \rho$, then $c(v) = X \in \mathcal{A}(\gamma)$. If $v \neq \rho$, from (iii), it follows that there exists $s \in S$ such that $c(v) = A(s)$, therefore $c(v) \in \mathcal{A}(\gamma)$. ∎

Lemma 2.2. *Let (V, A, f) be an accumulation phylogeny on X, and let γ and Γ be the induced accumulation maps on X and V respectively. Let $\mathcal{D} = (V, A)$.*

(1) *For any $u, v \in V$, the following conditions are satisfied*:

(i) $u \neq v \Leftrightarrow \Gamma(u) \neq \Gamma(v)$;
(ii) $u <_{\mathcal{D}} v \Leftrightarrow \Gamma(u) \subset \Gamma(v)$;
(iii) $u \leq_{\mathcal{D}} v \Leftrightarrow \Gamma(u) \subseteq \Gamma(v)$.

(2) *For any $v \in V$,*

$$\Gamma(v) \subseteq \bigcap_{x \in c(v)} \gamma(x).$$

(3) *Given $v \in V$,*

$$\bigcap_{x \in c(v)} \gamma(x) \subseteq \Gamma(v) \Leftrightarrow (c(v) \subseteq c(u) \Rightarrow u \leq_{\mathcal{D}} v).$$

Proof. Part (1)(i). Suppose $u \neq v$. The conditions $u <_{\mathcal{D}} v$ and $v <_{\mathcal{D}} u$ cannot be satisfied simultaneously. Assuming that $u <_{\mathcal{D}} v$ does not hold, it follows that $u \neq \rho$ and any arc ending in u does not lie on a path from ρ to v. Let a be such an arc with $f(a) \neq \emptyset$. Then $f(a) \subseteq \Gamma(u) - \Gamma(v)$, hence $\Gamma(u) \neq \Gamma(v)$.

Part (1)(ii). Assume that $u <_{\mathcal{D}} v$ and let $s \in \Gamma(u)$. Then either $s \in f(\rho)$ or $s \in f(a)$ for some arc a in a path from ρ to u. In the former case $s \in \Gamma(v)$ and in the latter a is an arc of a path from ρ to v. It follows from Part 1(i) that the inclusion is strict.

Conversely, suppose that $\Gamma(u) \subset \Gamma(v)$. If $u = \rho$, then $u <_{\mathcal{D}} v$. Assume now that $u \in V - \{\rho\}$. According to (P2) there exists s in S such that $s \in f(a)$ for some arc a ending in u. Therefore $s \in \Gamma(u)$, hence $s \in \Gamma(v)$. From (P1) it follows that $s \in f(b)$ entails $b = a$. Since $s \in \Gamma(v)$, there is a path from ρ to v containing a. Consequently, $u <_{\mathcal{D}} v$.

Part 1(iii). This follows from Parts 1(i) and 1(ii).

Part (2). Let $x \in c(v)$. Since $v \leq_{\mathcal{D}} x$, we have $\Gamma(v) \subseteq \gamma(x)$. Consequently, $\Gamma(v) \subseteq \cap_{x \in c(v)} \gamma(x)$.

Part (3). Let $v \in V$ and suppose that $\cap_{x \in c(v)} \gamma(x) \subseteq \Gamma(v)$. Let u be a vertex of V with $c(v) \subseteq c(u)$. It follows that

$$\Gamma(u) \subseteq \bigcap_{x \in c(u)} \gamma(x) \subseteq \bigcap_{x \in c(v)} \gamma(x) \subseteq \Gamma(v),$$

and so by Part 1(iii) of this lemma, $u \leq_{\mathcal{D}} v$.

Conversely, let $v \in V$ and suppose that $c(v) \subseteq c(u)$ entails $u \leq_{\mathcal{D}} v$. Let $s \in \cap_{x \in c(v)} \gamma(x)$. If $s \in f(\rho)$ then $s \in \Gamma(v)$. Otherwise, $s \in f(a)$ for some arc $a = (w, u)$. For any $x \in c(v)$ there is at least one path from ρ to x, containing a. It follows that $c(v) \subseteq c(u)$, thus $u \leq_{\mathcal{D}} v$. Therefore, $s \in \Gamma(v)$. ∎

Given a rooted digraph (V, A) and any vertex v of V let $\mathrm{end}(v)$ be the set

$$\mathrm{end}(v) := \begin{cases} \{\rho\}, & \text{if } v = \rho; \\ \{(u, v) \colon (u, v) \in A\}, & \text{otherwise;} \end{cases}$$

and for any function $f \colon A \cup \{\rho\} \to 2^S$ let

$$f(\mathrm{end}(v)) := \{s \in S \colon s \in f(a) \text{ for some } a \in \mathrm{end}(v)\}.$$

Lemma 2.3. *Let (V, A, f) be an accumulation phylogeny on X, and let Γ be the associated accumulation map on V. For all vertices v of V,*

$$f(\mathrm{end}(v)) = \Gamma(v) - \bigcup_{(u,v)\in A} \Gamma(u).$$

Proof. Let $v \in V$. If $v = \rho$ then $f(\mathrm{end}(v)) = f(\rho) = \Gamma(\rho)$ and there is no arc (u, v) in A. If $v \neq \rho$, then

$$\Gamma(v) = \left(\bigcup_{(u,v)\in A} \Gamma(u) \right) \cup f(\mathrm{end}(v))$$

and the two sets in the union are disjoint. The lemma now follows. ∎

3. Regular Hybrid Phylogenies

Given a collection $\mathcal{C}$ of non-empty subsets of X, the *cover digraph* for $\mathcal{C}$ is the digraph with vertex set $\mathcal{C}$ and an arc (A, B) whenever $B \subset A$ and there is no $C \in \mathcal{C}$ with $B \subset C \subset A$. We say that a hybrid phylogeny $\mathcal{D}$ is *regular* if the map $v \mapsto c(v)$ is an isomorphism from $\mathcal{D}$ to the cover digraph of its set of clusters. For example, the hybrid phylogeny on the right of Figure 1 is regular, while that on the left is not. Clearly, every rooted phylogenetic tree is a regular hybrid phylogeny. Note that if $\mathcal{D}$ is a regular hybrid, then the clusters associated to the vertices of $\mathcal{D}$ are all distinct.

For accumulation phylogenies for which the underlying digraph $\mathcal{D}$ is regular, the induced map γ suffices (along with $\mathcal{D}$) to reconstruct the sets $\Gamma(v)$ and $f(\mathrm{end}(v))$ for every vertex v of V, as we now show.

Proposition 3.1. *Suppose $\mathcal{D} = (V, A)$ is a regular hybrid phylogeny, and that γ is an accumulation map induced by (V, A, f). Then for any $v \in V$,*

(i) $\Gamma(v) = \bigcap_{x \in c(v)} \gamma(x)$,
(ii) $f(\mathrm{end}(v)) = \bigcap_{x \in c(v)} \gamma(x) - \bigcup_{(u,v)\in A} \bigcap_{x \in c(u)} \gamma(x)$.

Proof. Part (i). We have $\Gamma(v) \subseteq \bigcap_{x \in c(v)} \gamma(x)$, by Lemma 2.2(2). The reverse inclusion follows from Lemma 2.2(3), since if (V, A) is regular then for $u, v \in V$ we have $c(v) \subseteq c(u)$ entails $u \leq_{\mathcal{D}} v$.

Part (ii). By Lemma 2.3 we may identify $f(\mathrm{end}(v))$ with $\Gamma(v) - \bigcup_{(u,v)\in A} \Gamma(u)$. By Part (i) of Proposition 3.1 we have $\Gamma(v) = \bigcap_{x \in c(v)} \gamma(x)$ and $\Gamma(u) = \bigcap_{x \in c(u)} \gamma(x)$ from which Part (ii) now follows. ∎

3.1. Representations of Accumulation Maps on X

Our first main theorem considers the existence and uniqueness of representations of an arbitrary map γ by an accumulation phylogeny. The existence question has a fairly straightforward solution, however the uniqueness question is more interesting – as Figure 2 showed, γ does not, in general, uniquely determine the underlying hybrid phylogeny (V, A). However the following result shows that if we restrict our attention to regular hybrid phylogenies then one can uniquely recover (V, A) (along with the sets $f(\text{end}(v))$) from γ.

Theorem 3.2. *Let S be an arbitrary set, X a finite non-empty set, and γ a map from X to 2^S with $S = \bigcup_{x\in X}\gamma(x)$. Then, γ is the accumulation map induced by at least one accumulation phylogeny if and only if*

$$\gamma(x) - \bigcup_{y\in X:\, y\neq x} \gamma(y) \neq \emptyset, \text{ for all } x \in X. \tag{3.1}$$

Moreover, when this holds, there is a unique regular hybrid phylogeny (V, A) on X such that γ is the accumulation map induced by (V, A, f) for at least one choice of a map f. Although f is not necessarily uniquely determined by γ the sets $f(\text{end}(v))$, $v \in V$ are uniquely determined by γ.

Proof. If γ is the accumulation map on X for an accumulation phylogeny (V, A, f) then inequality (3.1) follows from property (P2).

Assume now that (3.1) holds. For each $s \in S$, let $A(s) = \{x \in X : s \in \gamma(x)\}$ and let $x \in X$. It follows that $A(s) = \{x\}$ for each $s \in \gamma(x) - \cup_{y\neq x}\gamma(y)$. Let $\mathcal{C} = \{A(s) : s \in S\} \cup \{X\}$ and let $\mathcal{D} = (V, A)$ be the cover digraph of $\mathcal{C}$. Note that $\mathcal{D}$ is a regular hybrid phylogeny. Then define $\Gamma : V \to 2^S$, by setting

$$\Gamma(v) = \bigcap_{x\in c(v)} \gamma(x),$$

and for each $v \in V$, let

$$S_v = \Gamma(v) - \bigcup_{(u,v)\in A} \Gamma(u).$$

It suffices to construct a map f from $A \cup \{\rho\}$ to 2^S such that $f(\text{end}(v)) = S_v$ for each $v \in V$. To this end, for each $v \in V - \{\rho\}$, let

$$g_v : S_v \to \text{end}(v)$$

be an arbitrary function. Now define $f : A \cup \{\rho\} \to 2^S$, by setting $f(\rho) = \Gamma(\rho)$, and $f((u, v)) = \{s : g_v(s) = (u, v)\}$. Clearly, (V, A, f) is an accumulation phylogeny on X and Γ is its induced accumulation map; furthermore $\gamma = \Gamma|X$. This establishes the first part of Theorem 3.2. The uniqueness of (V, A) follows from Lemma 2.1(iv) since we are assuming (V, A) is regular, and so it is determined by $\{c(v) : v \in V\}$. The uniqueness of $f(\text{end}(v))$ follows from Proposition 3.1(ii) . ∎

Remark 3.3. (1) Referring to Theorem 3.2, since (V, A) is uniquely determined by γ, so are the numbers $d^-(v)$. There are then exactly

$$\prod_{v \neq \rho} (d^-(v))^{|S_v|}$$

ways to define the map $f: A \cup \{\rho\} \to 2^S$. Consequently, f (and thereby the accumulation phylogeny (V, A, f)) is uniquely determined by γ if and only if (V, A) is a tree.

(2) Note that the unique regular hybrid phylogeny (V, A) that furnishes a representation of γ can be reconstructed from γ by an algorithm that runs in polynomial time (in $|X|$). Similarly, constructing a function f such that γ is the accumulation map for (V, A, f) is computationally easy.

4. Recognising Trees

Suppose an accumulation map $\gamma: X \to 2^S$ is induced by some accumulation phylogeny. We would like to be able to characterise when the unique regular hybrid phylogeny $\mathcal{D}$ that induces γ is a rooted phylogenetic tree.

It is a well-known, classical result (see, for example, [5]) that any collection $\mathcal{C}$ of sets (in particular $\mathcal{A}(\gamma)$) forms the clusters of a rooted tree if and only if $\mathcal{C}$ is a *hierarchy* — that is, any two sets in $\mathcal{C}$ are either disjoint, or one is contained in the other. This provides a simple characterisation for when the regular digraph $\mathcal{D}$ that induces an accumulation map is a tree. In this section we provide an alternative characterisation based on properties of the pairwise intersections $\gamma(x) \cap \gamma(y)$ (for all distinct pairs $x, y \in X$), since pairwise comparisons are often used for tree-building in molecular biology. We show that the cardinality of these pairwise comparisons does not suffice to characterise when the regular digraph $\mathcal{D}$ that induces γ is a tree (suggesting that simple 'distance-based' approaches, such as in [16] could be problematic in failing to detect horizontal gene transfer). However there is a simple characterisation involving the full structure of the sets. We begin with some terminology.

Recall that an *ultrametric* on X is any symmetric function $d: X \times X \to \mathbb{R}$ that satisfies the following two properties:

- $d(x, x) = 0$ for all $x \in X$, and
- for any three distinct elements $x, y, z \in X$, $d(x, y) \leq \max\{d(x, z), d(y, z)\}$.

An example of an ultrametric on X is provided as follows: Suppose we have a rooted phylogenetic tree $\mathcal{T}$ on X together with any function h from the vertices of $\mathcal{T}$ into the reals that is increasing along any path from a leaf to the root. Then for each distinct pair $x, y \in X$ set $d(x, y)$ equal to the h–value of the vertex of $\mathcal{T}$ that is the most recent common ancestor of the pair x, y, and set $d(x, x) = 0$ for all $x \in X$. In this case we say that $\mathcal{T}$ provides a *representation* of d. It is a classic result that for any ultrametric d there is a unique rooted phylogenetic tree on X that provides a representation of d (see, for example, [15]).

If $\mathcal{D}$ is a phylogenetic tree it is easy to recover $\mathcal{D}$ from any induced accumulation map on X, and to show how, we first make a further definition. For any map $\gamma\colon X \to 2^S$ define $d_\gamma\colon X \times X \to \mathbb{R}$ by setting

$$d_\gamma(x,y) = \begin{cases} 0, & \text{if } x = y; \\ -|\gamma(x) \cap \gamma(y)|, & \text{otherwise.} \end{cases}$$

The proof of the following result is straightforward and omitted.

Proposition 4.1. *Suppose that γ is an accumulation map on X induced by a rooted phylogenetic tree $\mathcal{T}$. Then d_γ is an ultrametric on X and $\mathcal{T}$ is the* (*unique*) *rooted phylogenetic tree on X that provides a representation of d_γ.*

A natural question at this point is whether the converse of Proposition 4.1 holds - that is, if d_γ is an ultrametric for an accumulation map γ induced by a hybrid phylogeny $\mathcal{D}$ does this imply that $\mathcal{D}$ is a rooted phylogenetic tree? The following example shows that the answer is no.

Example 4.2. Let $X = \{1,2,3\}$ and consider the cover digraph $\mathcal{D}$ of the following subsets: $\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}$; this digraph is shown on the right of Figure 1. For the associated function f take $S = A$ and assign the singleton set $\{a\}$ to each arc a of $\mathcal{D}$ and set $f(\rho) = \emptyset$. Then d_γ is an ultrametric yet $\mathcal{D}$ is not a rooted phylogenetic tree.

This example shows that it is impossible to distinguish between a tree and a non-tree hybrid phylogeny solely on the basis of the number of elements of S that each pair of elements x and y differ on. This raises a further question of whether it is possible to recognise when the underlying digraph is a tree if one uses more of the structure of the set system $\{\gamma(x) \cap \gamma(y)\colon x, y \in X, x \neq y\}$ than just the cardinality of these sets.

To address this question it is helpful to consider any map $\gamma\colon X \to 2^S$ and define

$$D_\gamma\colon X \times X \to 2^S$$

by $D_\gamma(x,y) = \gamma(x) \cap \gamma(y)$. It turns out that when γ is induced by a rooted phylogenetic tree then D_γ is an example of what Böcker and Dress [3] have called a *symbolic ultrametric*, and which is defined as follows.

Let M be an arbitrary set. A map δ from $X \times X$ into M is said to be an *symbolic ultrametric* (on X) if each of the following conditions is satisfied:

(U1) $\delta(x,y) = \delta(y,x)$, for all $x, y \in X$;
(U2) $|\{\delta(x,y), \delta(x,z), \delta(y,z)\}| \leq 2$, for all $x, y, z \in X$;
(U3) there are no pairwise distinct elements x, y, w, and z of X with

$$\delta(x,y) = \delta(y,w) = \delta(w,z) \neq \delta(y,z) = \delta(z,x) = \delta(x,w).$$

Returning to our arbitrary map $\gamma\colon X \to 2^S$, it is trivial that the associated function $D_\gamma\colon X \times X \to 2^S$ satisfies condition (U1). Moreover, D_γ also satisfies condition (U3) by the following lemma.

Lemma 4.3. *There are no mutually distinct sets A_1, A_2, A_3, and A_4 such that*

$$A_1 \cap A_2 = A_2 \cap A_3 = A_3 \cap A_4 \neq A_2 \cap A_4 = A_4 \cap A_1 = A_1 \cap A_3. \tag{4.1}$$

Proof. Suppose there exist mutually distinct sets $A_i \in \{1, \ldots, 4\}$ that satisfy (4.1). Then we obtain that:

$$A_1 \cap A_2 \cap A_3 \cap A_4 = A_3 \cap A_4 = A_2 \cap A_4,$$

a contradiction. ∎

Consequently, D_γ is a symbolic ultrametric on X if and only if it satisfies condition (U2). With this in mind we say that D_γ satisfies the *set-ultrametric property* if for any three distinct elements $x, y, z \in X$, two of the sets $D_\gamma(x, y)$, $D_\gamma(y, z)$ and $D_\gamma(x, z)$ are equal (this is equivalent to condition (U2)). If in addition, the two equal sets are always contained (respectively strictly contained) in the third we say that D_γ has the *nested* (*respectively strictly-nested*) *set-ultrametric property*.

We are now ready to state the pairwise characterisation result. Although part of this result may be proved using the main result [3] (on the representation of symbolic ultrametrics by discriminating maps on rooted phylogenetic trees), we have chosen instead to provide a self-contained argument.

Theorem 4.4. *Let γ be the accumulation map on X induced by an accumulation phylogeny, and let $\mathcal{D} = (V, A)$ be the unique regular hybrid phylogeny on X that induces γ (as provided by Theorem 3.2).*

(1) *The following are equivalent*:

  (i) *$\mathcal{D}$ is a rooted phylogenetic tree,*
  (ii) *D_γ has the set-ultrametric property,*
  (iii) *D_γ has the nested set-ultrametric property.*

(2) *$\mathcal{D}$ is a rooted binary phylogenetic tree if and only if D_γ has the strictly-nested set-ultrametric property.*

Proof. Part (1). We first show that (i) implies (iii). Assume that $\mathcal{D}$ is a rooted phylogenetic tree on X and let $x, y, z \in X$. On the one hand, for any vertices u and v of V, $\Gamma(u) \cap \Gamma(v) = \Gamma(\mathrm{lca}(u, v))$, where $\mathrm{lca}(u, v)$ denotes the most recent common ancestor of u and v in $\mathcal{D}$. On the other hand, two of the vertices $\mathrm{lca}(x, y)$, $\mathrm{lca}(y, z)$, and $\mathrm{lca}(z, x)$ are equal. We may assume that $\mathrm{lca}(x, y) = \mathrm{lca}(y, z)$. In this case, $\mathrm{lca}(x, y) \leq_{\mathcal{D}} \mathrm{lca}(x, z)$, hence $\Gamma(\mathrm{lca}(x, y)) = \Gamma(\mathrm{lca}(y, z)) \subseteq \Gamma(\mathrm{lca}(x, z))$. Therefore, since $D_\gamma(x_1, x_2) = \Gamma(\mathrm{lca}(x_1, x_2))$, for any pair $x_1, x_2 \in X$ we have $\mathcal{D}_\gamma(x, y) = \mathcal{D}_\gamma(y, z) \subseteq \mathcal{D}_\gamma(x, z)$.

Condition (iii) clearly implies condition (ii). We now show that (ii) implies (i) - or equivalently, that the negation of (i) implies the negation of (ii). Thus let us assume that $\mathcal{D}$ is not a tree, then there exists a vertex v with $d^-(v) \geq 2$. Let u_1 and u_2 be two vertices such that $(u_1, v) \in A$, $(u_2, v) \in A$ and $u_1 \neq u_2$ (Figure 3). Since $\mathcal{D}$ is regular there is no path from u_1 to u_2 or from u_2 to u_1. Consequently, neither of the inclusions $c(u_1) \subseteq c(u_2)$ and $c(u_2) \subseteq c(u_1)$ hold, so there exist two leaves x and y with $x \in c(u_1) - c(u_2)$ and $y \in c(u_2) - c(u_1)$. Let $z \in c(v)$.

The hybrid $\mathcal{D}$ is regular, so $u_1 \neq \rho \neq u_2$, hence, $d^-(u_1) \neq 0 \neq d^-(u_2)$. For each $i \in \{1, 2\}$ let a_i be an arc ending in u_i with $f(a_i) \neq \emptyset$. Then, $f(a_1) \subseteq \gamma(x) \cap \gamma(z)$, $f(a_2) \subseteq$

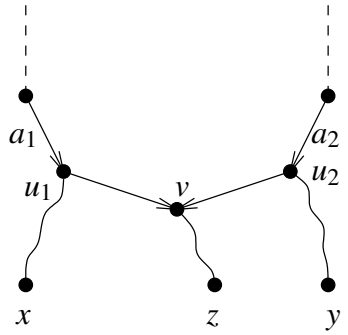

Figure 3.

$\gamma(y) \cap \gamma(z)$, $f(a_1) \cap \gamma(y) = \emptyset = f(a_2) \cap \gamma(x)$. Therefore, the sets $\mathcal{D}_\gamma(x, y)$, $\mathcal{D}_\gamma(y, z)$, and $\mathcal{D}_\gamma(x, z)$ are mutually distinct, which contradicts the set-ultrametric property.

Part (2). If, in addition, $\mathcal{D}$ is binary and $x, y, z \in X$ are mutually distinct, then $\text{lca}(x, y) \neq \text{lca}(x, z)$, hence the containment is strict. Conversely, assume that the strictly-nested set-ultrametric property is satisfied. According to (i), $\mathcal{D} = (V, A)$ is a tree. If $\mathcal{D}$ is not binary one can find the vertices u, v_1, v_2, v_3 such that $(u, v_i) \in A$, $1 \leq i \leq 3$. If $x_i \in c(v_i)$, then $D_\gamma(x_1, x_2) = D_\gamma(x_2, x_3) = D_\gamma(x_3, x_1) = \Gamma(u)$, which is contradictory to the strictly-nested set-ultrametric property. ∎

4.1. Concluding Comments

It would be interesting to investigate variations on the model described - either by weakening one (or both) of the properties (P1) or (P2), or by allowing elements of S to be 'lost'. We describe briefly this last variation. Rather than requiring, for each arc $a = (u, v) \in A$ that

$$\Gamma(v) = \Gamma(u) \cup f(a)$$

we may weaken this to require merely that

$$\Gamma(v) = B \cup f(a),$$

where $B \subseteq \Gamma(u)$ to thereby model the situation where elements of S become lost over time. In this model we maintain (P1) and (P2) but allow some flexibility in the definition of Γ. As might be expected, one can say much less about the underlying hybrid phylogeny (V, A) from the induced accumulation map $\gamma = \Gamma|X$. However γ does still convey some phylogenetic information - for example, suppose we know that $\mathcal{D} = (V, A)$ is a tree. Then if, for some subset Y of $X - \{x\}$, we have

$$\gamma(x) \subseteq \cup_{y \in Y} \gamma(y),$$

then this places a constraint on $\mathcal{D}$ - namely, the most recent common ancestor of the species in Y must also be an ancestor of x.

Acknowledgments. The authors thank the New Zealand Marsden Fund for supporting this research. We also thank Stefan Grünwald, and an anonymous referee for some helpful comments on an earlier version of this manuscript.

References

1. J. Bang-Jensen and G. Gutin, Digraphs: Theory, Algorithms and Applications, Springer-Verlag, London, 2001.
2. M. Baroni, C. Semple, and M. Steel, A framework for representing reticulate evolution, Ann. Comb. **8** (4) (2004) 391–408.
3. S. Böcker and A.W.M. Dress, Recovering symbolically dated, rooted trees from symbolic ultrametrics, Adv. Math. **138** (1998) 105–125.
4. D. Bryant and V. Moulton, NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks, Mol. Biol. Evol. **21** (2) (2004) 255–265.
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, D. Huson, and V. Moulton, Analysing and visualizing sequence and distance data using SPLITSTREE, Discrete Appl. Math. **71** (1996) 95–109.
7. J. Felsenstein, Inferring Phylogenies, Sinauer Press, 2004.
8. D. Gusfield, S. Eddhu, and C. Langley, The fine structure of galls in phylogenetic networks, INFORMS J. Comput. **16** (2004) 459–469.
9. J. Jansson and W.–K. Sung, Inferring a level–1 phylogenetic network from a dense set of rooted triplets, In: Proceedings of the Tenth International Computing and Combinatorics Conference, Springer–Verlag, 2004.
10. P. Legendre, Biological applications of reticulate analysis, J. Classification **17** (2000) 191–195.
11. P. Legendre and V. Makarenkov, Reconstruction of biogeographic and evolutionary networks using reticulograms, Systematic Biol. **51** (2) (2002) 199–216.
12. B.M.E. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R.E. Timme, Phylogenetic networks: modeling, reconstructibility, and accuracy, IEEE/ACM Trans. Comput. Biology Bioinform. **1** (1) (2004) 13–23.
13. L. Nakhleh, J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, and A. Tholse, Towards the development of computational tools for evaluating phylogenetic network reconstruction methods, In: Proceedings of the Eighth Pacific Symposium on Biocomputing, (2003) pp. 315–326.
14. L. Nakhleh, T. Warnow, and C.R. Linder, Reconstructing reticulate evolution in species – theory and practice, In: Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology, (2004) pp. 337–346.
15. C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.
16. B. Snel, P. Bork, and M.A. Huynen, Genome phylogeny based on gene content, Nat. Genet. **21** (1999) 108–110.