Annals of Combinatorics 8 (2004) 391-408
0218-0006/04/040391-18
DOI 10.1007/s00026-004-0228-0

Annals of Combinatorics

A Framework for Representing Reticulate Evolution*

Mihaela Baroni, Charles Semple, and Mike Steel

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
mbaroni@ugal.ro, {c.semple, m.steel}@math.canterbury.ac.nz

Received November 12, 2003

AMS Subject Classification: 05C05, 92D15

Abstract. Acyclic directed graphs (ADGs) are increasingly being viewed as more appropriate for representing certain evolutionary relationships, particularly in biology, than rooted trees. In this paper, we develop a framework for the analysis of these graphs which we call *hybrid phylogenies*. We are particularly interested in the problem whereby one is given a set of phylogenetic trees and wishes to determine a hybrid phylogeny that 'embeds' each of these trees and which requires the smallest number of hybridisation events. We show that this quantity can be greatly reduced if additional species are involved, and investigate other combinatorial aspects of this and related questions.

Keywords: directed acyclic graph, reticulate evolution, hybrid species, subtree prune and regraft

1. Introduction

Creating a 'tree of life' has been a primary goal of systematic biology since Charles Darwin's first sketch of an evolutionary tree in 1837. It has become an accepted dogma that such a tree would describe how all present-day species had evolved from a common ancestor. However, accumulating data suggest that evolution is more complex than this, because many species are mosaics of genes derived from different ancestors. This pattern may be the result of processes such as hybridisation (the formation of a species that contains genetic contributions from more than one ancestral species), a process that is widely recognised in certain plant and fish species. Nearly 20 years ago, Funk [8] cautioned "it is difficult to overemphasise the importance of hybridization and polyploidy in evolution." Other mechanisms, such as the horizontal transfer of genes between species may also be important sources of reticulate (non-tree like) evolution particularly for deep divergences in the tree of life. The situation here has been recently summarised by Doolittle [6] who wrote that "molecular phylogeneticists will have failed to find the 'true tree', not because their methods are inadequate or because

* The authors thank the New Zealand Marsden Fund (UOC-MIS-005) for supporting this research.

they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree."

To model reticulate evolution it seems increasingly appropriate to represent the evolution of the species under study with a directed graph, where the vertices correspond to extant and ancestral species, while each arc represents the transfer of genetic material from one species to another—for example, by hybridisation or horizontal gene transfer. This gives rise to several interesting mathematical and computational problems. One question is how best to represent and reconstruct these digraphs. To date much of the analysis in the biological literature has been somewhat ad-hoc. For example, starting from a tree, one can introduce additional arcs in a heuristic fashion to see if there is an improvement in the 'fit' to data. Such an approach was described by Legendre and Makarenkov [11] for inferring a reticulation network ('reticulogram') from a given distance matrix, and applied to examples from biogeography, population microevolution, and hybridisation. Other aspects of the problem of representing hybridisation in biology are discussed in [4, 9, 10, 13, 15, 17–19].

Another strategy for describing reticulate evolution has been to apply existing mathematical procedures that generate graphs (rather than trees) to biological data. Lapointe [9], reviewed four such approaches. These methods—pyramids [5], weak hierarchies [2], splitsgraphs [7] and reticulograms—were applied to the same data set and the results are compared. However, it is not clear that such general techniques for constructing graphs, often developed for quite different processes, are precisely the right tool for representing hybrid evolution.

In this paper, we take an alternative approach, developing a digraph representation that reflects directly the biological questions we consider. We call these digraphs, subject to simple constraints, 'hybrid phylogenies'. In Section 2, we formally describe these phylogenies and identify an important subclass—the 'regular' hybrid phylogenies (these are naturally isomorphic to the cover digraph of their associated cluster system). By restricting our attention to regular hybrid phylogenies, we avoid many pathologies that can arise in the infinite set of possible hybrid phylogenies on a given set of extant species. Indeed, Section 4 shows for application purposes no generality is lost in confining ourselves to regular hybrid phylogenies.

One of the themes throughout this paper is to use this formalism to study a fundamental question of interest for biologists: given a collection of trees on sets of species that faithfully represent the (tree-like) evolution of different parts of various species genomes, we would like to know how these trees can be 'displayed' by a single hybrid phylogeny. In particular it is of interest to determine the smallest number of hybrid events that are required for the trees to be simultaneously displayed by a single hybrid phylogeny. This number then sets a lower bound on the degree of hybridisation that has occurred in the evolution of the species under consideration. Proposition 4.2 shows that the restriction to regular hybrid phylogenies does not change this minimum number.

In order to study these concepts it is first necessary to formalise the notion of what it means for a hybrid phylogeny to 'display' a rooted phylogenetic tree. We do this in Section 3 and show that, for any given collection $\mathcal{P}$ of rooted phylogenetic trees, there is a canonical (and regular) hybrid phylogeny that displays each of the trees in this collection. This particular hybrid for when $\mathcal{P}$ consists of two trees is considered further in Section 6. In general, this canonical hybrid exhibits more hybrid events than

are required. Section 5 shows that the number of hybrid events can be greatly reduced if other species (not mentioned by any of the input trees) are permitted—a phenomenon that is biologically relevant, since systematists generally sample only a subset of species from a group, and other species (including ones that may now be extinct) may have been involved in hybridisation in the past. Finally, Section 7 investigates the question of whether the canonical hybrid phylogeny associated with two rooted phylogenetic trees uniquely determines these two trees. We show that provided the two trees are sufficiently similar (one can be transformed into the other by a single subtree transfer operation) this is the case, but, in general, it is not so.

We hope that the results in this paper will provide a basis for further investigations into the representation and analysis of hybrid evolution. Unless otherwise stated, the notation and terminology in this paper coincides with [16].

2. Hybrid Phylogenies

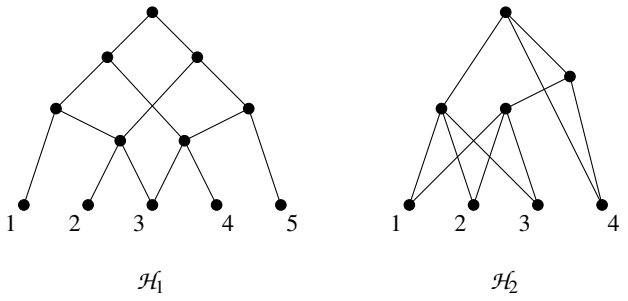
We first recall some basic terminology concerning digraphs. For additional background, see [3]. A *directed graph* (or *digraph*) D is an ordered pair (V, A) consisting of a non-empty set V of *vertices* and a subset A of $V \times V$ of *arcs*. We sometimes denote the vertex set of D by $V(D)$ and the arc set of D by $A(D)$. The *out-degree* (respectively, *in-degree*) of a vertex v of D, denoted $d^+(v)$ (respectively, $d^-(v)$) is the number of arcs in A whose first (respectively, second) component is v. A *directed cycle* of a digraph $D = (V, A)$ is a sequence $v_0, a_1, v_1, a_2, v_2, \ldots, v_{k-1}, a_k, v_k$ of vertices and arcs in which the first and last vertices are equal, $a_i = (v_{i-1}, v_i)$ for all i, and, apart from v_0 and v_k, no vertex or arc appears more than once. A digraph is *acyclic* if it has no directed cycles. An acyclic digraph D with no underlying parallel edges is *rooted* if there is a distinguished vertex ρ, called the *root*, with the properties that $d^-(\rho) = 0$ and there is a directed path from ρ to every vertex of D. Observe that, because of the restrictions placed on D, except for ρ, no other vertex has in-degree zero.

A *hybrid phylogeny* $\mathcal{H}$ (*on* X) is an ordered pair $(D;\phi)$ consisting of a rooted acyclic digraph $D = (V, A)$ and a bijective map ϕ from X into the set of vertices of V with out-degree zero such that the root has out-degree at least two and, for all $v \in V - \phi(X)$ with $d^-(v) = 1$, we have $d^+(v) \geq 2$. The vertices of in-degree at least two are called *hybridisation* vertices. The set X is called the *label* set of $\mathcal{H}$ and is denoted by $\mathcal{L}(\mathcal{H})$. Furthermore, for a collection $\mathcal{P}$ of hybrid phylogenies we use $\mathcal{L}(\mathcal{P})$ to denote $\bigcup_{\mathcal{H}\in\mathcal{P}} \mathcal{L}(\mathcal{H})$. Throughout this paper, we will often refer to hybrid phylogenies as *hybrids* and always draw them with their arcs directed downwards, and so omit the arrow heads. Two hybrid phylogenies are shown in Figure 1.

Two hybrids $\mathcal{H}_1 = (D_1;\phi_1)$ and $\mathcal{H}_2 = (D_2;\phi_2)$ on X, where $D_1 = (V_1, A_1)$ and $D_2 = (V_2, A_2)$, are *isomorphic* if there exists a bijection $\psi\colon V_1 \to V_2$ that induces a bijection between A_1 and A_2 that preserves orientation, and $\phi_2 = \psi \circ \phi_1$. We write $\mathcal{H}_1 \cong \mathcal{H}_2$ if $\mathcal{H}_1$ is isomorphic to $\mathcal{H}_2$.

For a hybrid $\mathcal{H}$ on X, let

$$h(\mathcal{H}) = \sum_{v\neq\rho} (d^-(v) - 1).$$



$\mathcal{H}_1$ $\mathcal{H}_2$

Figure 1: Two hybrid phylogenies.

We call $h(\mathcal{H})$ the *hybridisation number* of $\mathcal{H}$. Observe that $h(\mathcal{H}) \geq 0$ and $h(\mathcal{H}) = |A| - |V| + 1$.

Rooted phylogenetic trees are special types of hybrid phylogenies. In particular, a hybrid phylogeny $\mathcal{H}$ is a rooted phylogenetic tree if and only if $h(\mathcal{H}) = 0$. As a consequence of this, we call the vertices of $\mathcal{H}$ with out-degree zero the *leaves* of $\mathcal{H}$. Note here that there is no restriction on the in-degree of a leaf. Viewing a hybrid phylogeny as representing the evolutionary history of a collection of present-day species, the hybridisation number quantifies the number of associated hybridisation events.

A class of hybrid phylogenies that will play an important role in this paper are the 'regular hybrids'. Although a certain type of hybrid, we will see that little generality is lost in working with regular hybrids. We describe regular hybrids next.

A collection $\mathcal{C}$ of non-empty subsets of X is a *cluster system* (*on* X) if $X \in \mathcal{C}$ and, for all $x \in X$, we have $\{x\} \in \mathcal{C}$. Let $\mathcal{C}$ be a cluster system on X and consider the *cover digraph* of $\mathcal{C}$; that is, the digraph that has vertex set $\mathcal{C}$ and an arc from C_1 to C_2 whenever $C_2 \subset C_1$ and there is no $C_3 \in \mathcal{C}$ with $C_2 \subset C_3 \subset C_1$. A natural way to obtain a hybrid on X is to begin with a cluster system $\mathcal{C}$ on X, take the cover digraph of $\mathcal{C}$ to be our rooted directed graph with root X, and define a map $\phi\colon X \to \mathcal{C}$ by setting $\phi(x) = \{x\}$ for all $x \in X$. We call this hybrid the *cover hybrid of* $\mathcal{C}$ and denote it by $\mathcal{H}(\mathcal{C})$.

Conversely, given a hybrid $\mathcal{H} = (D;\phi)$ on X with vertex set V and root ρ, there is a canonical way to obtain a cluster system on X. For all $v \in V$, let $c(v)$ denote the subset of X consisting of the elements x for which there is a directed path in D from v to $\phi(x)$. We call $c(v)$ the *cluster* corresponding to v. Observe that $c(\rho) = X$, $c(v) \neq \emptyset$ for all $v \in V$, and $c(\phi(x)) = \{x\}$ for all $x \in X$. The collection $\mathcal{C}(\mathcal{H}) = \{c(v)\colon v \in V\}$ is the *set of clusters of* $\mathcal{H}$ and is a cluster system on X. If $\mathcal{H}$ is a rooted phylogenetic tree, this definition differs slightly with that given in [16]. In particular, here we associate a cluster with the root, whereas, in [16], no cluster is associated with the root.

Let $\mathcal{H}$ be a hybrid with cluster set $\mathcal{C}$. We say that $\mathcal{H}$ is *regular* if the map from the vertex set of $\mathcal{H}$ into the vertex set of $\mathcal{H}(\mathcal{C})$ defined by $v \mapsto c(v)$ induces an isomorphism between $\mathcal{H}$ and $\mathcal{H}(\mathcal{C})$. In Figure 1, $\mathcal{H}_1$ is a regular hybrid. However, $\mathcal{H}_2$ is not regular as it contains two distinct vertices whose corresponding clusters are the same. Alternatively, $\mathcal{H}_2$ contains an edge joining the leaf labelled 4 and the root and yet there is another vertex whose corresponding cluster properly contains 4 and is a proper subset of $\{1,2,3,4\}$. Unlike hybrids in general, regular hybrids are determined by their sets

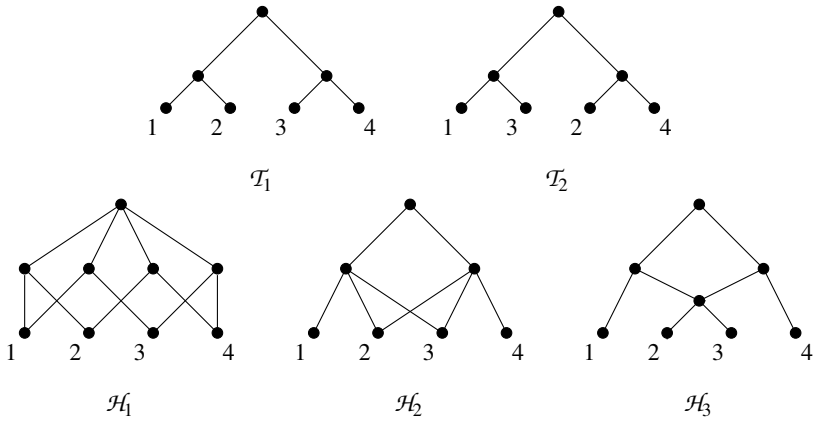


Figure 2: Five regular hybrids.

of clusters. Rooted phylogenetic trees are special types of regular hybrids.

It can be easily shown that a hybrid $\mathcal{H}$ on X is regular if and only if it is isomorphic to $\mathcal{H}(\mathcal{C})$ for some cluster system $\mathcal{C}$ on X (and in that case $\mathcal{C} = \mathcal{C}(\mathcal{H})$). We will give a graph-theoretic characterisation for regularity in Proposition 4.1.

Note that if $\mathcal{H}$ is a regular hybrid, then the clusters associated to the vertices of $\mathcal{H}$ are all distinct and are strictly nested along any directed path. Consequently, the longest directed path in any regular hybrid on X has at most $|X|$ vertices. Also, a regular hybrid has no vertices of out-degree 1. Finally, two regular hybrids on X are isomorphic if and only if they have the same set of clusters.

3. Displaying Hybrids

For rooted phylogenetic trees, the mathematical notion of 'display' captures the concept of preserving ancestral relationships when viewed as evolutionary trees and it has become a fundamental notion in phylogenetics. In this section, we extend the definition of display for rooted phylogenetic trees to hybrids.

A rooted digraph D' is a *rooted subdigraph* of a rooted digraph D if $V(D')$ and $A(D')$ are subsets of $V(D)$ and $A(D)$, respectively. Furthermore, for two hybrid phylogenies $\mathcal{H}$ and $\mathcal{H}'$ on the same leaf set, we say that $\mathcal{H}'$ is a *refinement* of $\mathcal{H}$ if $\mathcal{H}$ can be obtained from $\mathcal{H}'$ by contracting internal edges. These two definitions enable us to say what it means for one hybrid to be displayed by another. Let $\mathcal{H}$ be a hybrid on X and let $\mathcal{H}'$ be a hybrid on X'. We say that $\mathcal{H}'$ *displays* $\mathcal{H}$ if $X \subseteq X'$ and a rooted subdigraph of $\mathcal{H}'$ is a refinement of $\mathcal{H}$. For example, in Figure 2, $\mathcal{H}_1$ and $\mathcal{H}_2$ display both $\mathcal{T}_1$ and $\mathcal{T}_2$, but $\mathcal{H}_3$ displays neither $\mathcal{T}_1$ nor $\mathcal{T}_2$. If $\mathcal{H}$ and $\mathcal{H}'$ are rooted phylogenetic trees, then this definition of display coincides with the usual notion of display for rooted phylogenetic trees (see [16]). However, as we now show, some of the results that hold for rooted phylogenetic trees do not hold in the hybrid setting.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic trees on X. Then $\mathcal{T}'$ displays $\mathcal{T}$ if and only if $\mathcal{C}(\mathcal{T}) \subseteq \mathcal{C}(\mathcal{T}')$. However, for two regular hybrids on X, the analogous result

is not true. For example, in Figure 2, $\mathcal{H}_2$ displays $\mathcal{T}_1$ but $\mathcal{C}(\mathcal{T}_1)$ is not a subset of $\mathcal{C}(\mathcal{H}_2)$. Furthermore, $\mathcal{C}(\mathcal{H}_2) \subseteq \mathcal{C}(\mathcal{H}_3)$, but $\mathcal{H}_3$ does not display $\mathcal{H}_2$. Although this last example shows that the converse does not hold, it is easily checked that if $\mathcal{H}'$ is a refinement of $\mathcal{H}$, then $\mathcal{C}(\mathcal{H}) \subseteq \mathcal{C}(\mathcal{H}')$. Despite these examples, we do have the following proposition.

Proposition 3.1. *Let $\mathcal{H}$ be a regular hybrid phylogeny on X and let $\mathcal{T}$ be a rooted phylogenetic tree on X. If $\mathcal{C}(\mathcal{T}) \subseteq \mathcal{C}(\mathcal{H})$, then $\mathcal{H}$ displays $\mathcal{T}$.*

Proof. The proof is by induction on the height g of $\mathcal{T}$. Clearly, the proposition holds if $g = 1$. Now assume that $g \geq 2$ and the proposition holds for all regular hybrid phylogenies and rooted phylogenetic trees on X where the height of the latter is less than g.

Let ρ and ρ' denote the roots of $\mathcal{T}$ and $\mathcal{H}$, respectively. Evidently, $c(\rho) = c(\rho')$. Let $v_1, v_2, \ldots, v_k$ be the vertices of $\mathcal{T}$ that are immediate descendants of ρ and, for each i, let A_i denote the cluster of $\mathcal{T}$ corresponding to v_i. Observe that $\{A_1, A_2, \ldots, A_k\}$ is a partition of X. Since $\mathcal{C}(\mathcal{T}) \subseteq \mathcal{C}(\mathcal{H})$, for each i, there is a vertex u_i of $\mathcal{H}$ with $c(u_i) = A_i$. For all i, let $\mathcal{H}_i$ be the regular hybrid whose set of clusters consists of the subsets of A_i in $\mathcal{C}(\mathcal{H})$. Since $\{A_1, A_2, \ldots, A_k\}$ partitions X, every vertex w of $\mathcal{H}$ that is a descendant of one of the vertices $u_1, u_2, \ldots, u_k$ has the property that $c(w)$ is a subset of exactly one of the sets $A_1, A_2, \ldots, A_k$. This means that $\mathcal{H}_i$ is the regular hybrid obtained from $\mathcal{H}$ by restricting $\mathcal{H}$ to u_i and its descendants together with their incident edges. Furthermore, since $\mathcal{C}(\mathcal{T}) \subseteq \mathcal{C}(\mathcal{H})$, it follows that the rooted phylogenetic tree $\mathcal{T}_i$ obtained from $\mathcal{T}$ by restricting $\mathcal{T}$ to A_i has the property that $\mathcal{C}(\mathcal{T}_i) \subseteq \mathcal{C}(\mathcal{H}_i)$ for all i. As the height of $\mathcal{T}_i$ is less than g for all i, we deduce by the induction assumption that $\mathcal{H}_i$ displays $\mathcal{T}_i$.

To see that $\mathcal{H}$ displays $\mathcal{T}$, it suffices to observe that, for each i, there exists a (directed) path in $\mathcal{H}$ from ρ' to u_i that avoids the vertices $u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_k$. Noting that no directed path exists between two of the vertices $u_1, u_2, \ldots, u_k$, this is indeed the case. Hence $\mathcal{H}$ displays $\mathcal{T}$. ∎

An immediate consequence of Proposition 3.1 is that, given a collection $\mathcal{P}$ of rooted phylogenetic X-trees, the regular hybrid whose set of clusters is $\bigcup_{\mathcal{T} \in \mathcal{P}} \mathcal{C}(\mathcal{T})$ displays $\mathcal{P}$.

The converse of Proposition 3.1 does not hold. To see this, observe that, in Figure 2, $\mathcal{H}_2$ displays $\mathcal{T}_1$, but $\mathcal{C}(\mathcal{T}_1)$ is not a subset of $\mathcal{C}(\mathcal{H}_2)$.

4. Regular Hybrids

In this section, we consider more closely the concept of a regular hybrid. We begin with a useful graph-theoretic characterisation of regular hybrids.

Proposition 4.1. *A hybrid $\mathcal{H}$ is regular if and only if, for all distinct vertices $v_1, v_2 \in V(\mathcal{H})$, the following conditions hold*:

(i) $c(v_1) \neq c(v_2)$.
(ii) *If $c(v_2) \subset c(v_1)$, then there exists a directed path from v_1 to v_2.*
(iii) *If there are two distinct directed paths connecting v_1 and v_2, then neither path consists of a single arc.*

Proof. Clearly, if $\mathcal{H}$ is regular, then conditions (i)–(iii) hold. Now assume that the vertices of $\mathcal{H}$ satisfy (i)–(iii) in the statement of the proposition. Let $\mathcal{C}$ be the set of clusters of $\mathcal{H}$. To prove the converse, we show that the map $v \mapsto c(v)$ induces an isomorphism between $\mathcal{H}$ and the cover hybrid $\mathcal{H}(\mathcal{C})$ of $\mathcal{C}$. Since $\mathcal{H}$ satisfies (i), this map is a bijection.

Let (v_1, v_2) be an arc of $\mathcal{H}$. Then $c(v_2) \subset c(v_1)$, and so, by the definition of $\mathcal{H}(\mathcal{C})$, there is a directed path in $\mathcal{H}(\mathcal{C})$ from $c(v_1)$ to $c(v_2)$. If this directed path does not consist of a single arc, then there is a vertex $c(u)$ of $\mathcal{H}(\mathcal{C})$ that lies in this path such that $c(v_2) \subset c(u) \subset c(v_1)$. Since the vertices of $\mathcal{H}$ satisfy (ii), this implies that there is a directed path in $\mathcal{H}$ from v_1 to v_2 that passes through u. But then there are two directed paths from v_1 to v_2, one of which consists of a single arc. This contradicts the fact that $\mathcal{H}$ satisfies (iii). Hence $(c(v_1), c(v_2))$ is an arc of $\mathcal{H}(\mathcal{C})$.

Now suppose that $(c(v_1), c(v_2))$ is an arc of $\mathcal{H}(\mathcal{C})$. Then $c(v_2) \subset c(v_1)$. Therefore, as $\mathcal{H}$ satisfies (ii), there is a directed path in $\mathcal{H}$ from v_1 to v_2. This path must consist of a single arc. To see this, assume this is not the case. Then there is a vertex, u say, on this path distinct from v_1 and v_2 such that $c(v_2) \subset c(u) \subset c(v_1)$. This implies that $(c(v_1), c(v_2))$ is not an arc of $\mathcal{H}(\mathcal{C})$; a contradiction. Thus (v_1, v_2) is an arc of $\mathcal{H}$. This completes the proof of the proposition. ∎

Although regular hybrids are a special type of hybrid phylogeny, Proposition 4.2 shows that, for any hybrid $\mathcal{H}$, there is always a regular hybrid that displays $\mathcal{H}$ and has the same hybridisation number as $\mathcal{H}$. Such a regular hybrid can be obtained from $\mathcal{H}$ by a sequence of operations. We describe these operations first, before presenting Proposition 4.2.

Let $\mathcal{H}$ be a hybrid phylogeny, and consider the hybrid phylogeny that is constructed from $\mathcal{H}$ by performing the following sequence of operations for all distinct $v_1, v_2 \in V(\mathcal{H})$:

(I) If $c(v_1) = c(v_2)$, then, for each i, adjoin a new leaf vertex to v_i via a new arc and assign the new leaf vertex a new label.
(II) If $c(v_2) \subset c(v_1)$, but there is no directed path from v_1 to v_2, then adjoin a new leaf vertex to v_2 via a new arc and assign the new leaf vertex a new label.
(III) If there are two distinct directed paths from v_1 to v_2 one of which is an arc e, then subdivide e with a single vertex and adjoin a new leaf vertex to this single vertex via a new arc and assign the new leaf vertex a new label.

Note that throughout the above process, each of the newly created labels are distinct.

Proposition 4.2. *Let $\mathcal{H}$ be a hybrid and let $\mathcal{H}'$ be a hybrid obtained from $\mathcal{H}$ by applying operations* (I)–(III) *above. Then $\mathcal{H}'$ is regular. Furthermore,*

(i) *the hybridisation number of $\mathcal{H}$ is equal to the hybridisation number of $\mathcal{H}'$ and*
(ii) *any hybrid displayed by $\mathcal{H}$ is also displayed by $\mathcal{H}'$.*

Proof. Clearly, (i) and (ii) hold as the in-degree of the vertices of $\mathcal{H}$ do not change under the construction of $\mathcal{H}'$ and each new vertex has in-degree one. We now show that $\mathcal{H}'$ is regular. By Proposition 4.1, it suffices to show that, for all distinct $v_1, v_2 \in V(\mathcal{H}')$, (i)–(iii) of Proposition 4.1 is satisfied.

As each new leaf is assigned a new label in the construction, it follows that, for all distinct $v_1, v_2 \in V(\mathcal{H}')$, we have $c(v_1) \neq c(v_2)$. Furthermore, because of (II) in this construction, there is always a directed path from v_1 to v_2 if $c(v_2) \subset c(v_1)$. Lastly, because of (III) in the construction, there are no two distinct directed paths connecting two vertices one of which is an arc. ∎

We end this section by describing a way of comparing regular hybrids with the same label set. For a fixed set X, there is a natural metric on the collection of regular hybrids on X. Recall that the *symmetric difference* of two sets A and B, denoted $A \triangle B$, is the set

$$A \triangle B = (A \cup B) - (A \cap B) = (A - B) \cup (B - A).$$

For two hybrids $\mathcal{H}_1$ and $\mathcal{H}_2$ on X, define $d(\mathcal{H}_1, \mathcal{H}_2)$ to be

$$d(\mathcal{H}_1, \mathcal{H}_2) = |C(\mathcal{H}_1) \triangle C(\mathcal{H}_2)|.$$

The proof of the following result is straightforward and omitted.

Proposition 4.3. *The function d defined above is a metric on the collection of regular hybrids on X.*

This metric, when restricted to rooted phylogenetic X-trees, is the well-known 'Robinson-Foulds' metric [14].

5. The Hybridisation Number of a Collection of Trees

Although hybridisation events do occur in biology, they are still relatively rare. Consequently, for a collection $\mathcal{P}$ of rooted phylogenetic trees, a fundamental problem is to determine the smallest number of hybridisation events that are required for the existence of a hybrid phylogeny $\mathcal{H}$ to simultaneously display each of the trees in $\mathcal{P}$. In this section, we consider this problem and, in particular, show that there can be a vast disparity if we restrict $\mathcal{H}$ to have the same label set as $\mathcal{P}$ or allow $\mathcal{H}$ to have additional labels. The reason for allowing the latter is that it is usually the case that not all species of a group are sampled.

Let $\mathcal{P}$ be a collection of regular hybrids. A hybrid $\mathcal{H}$ is said to *display* $\mathcal{P}$ if each hybrid in $\mathcal{P}$ is displayed by $\mathcal{H}$. Let

$$h(\mathcal{P}) = \min\{h(\mathcal{H}) \colon \mathcal{H} \text{ regular hybrid that displays } \mathcal{P} \text{ and } \mathcal{L}(\mathcal{P}) = \mathcal{L}(\mathcal{H})\}$$

and

$$h^+(\mathcal{P}) = \min\{h(\mathcal{H}) \colon \mathcal{H} \text{ regular hybrid that displays } \mathcal{P} \text{ and } \mathcal{L}(\mathcal{P}) \subseteq \mathcal{L}(\mathcal{H})\}.$$

If $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$, we will denote the number $h(\mathcal{P})$ by $h(\mathcal{T}_1, \mathcal{T}_2)$. Clearly, $h^+(\mathcal{P}) \leq h(\mathcal{P})$. The next two lemmas are needed for the proof of Proposition 5.3.

Lemma 5.1. *Let $\mathcal{H}$ be a regular hybrid on X. Then*

$$h(\mathcal{H}) \geq |V| - 2|X| + 1.$$

Proof. Let V and ϕ denote the vertex set and labelling map of $\mathcal{H}$, respectively. Since $\mathcal{H}$ is regular, $d^+(v) \geq 2$ for all $v \in V - \phi(X)$. Therefore, as the out-degree of each vertex in $\phi(X)$ is zero, $|A| = \sum_{v \in V} d^+(v) \geq 2(|V| - |X|)$. This implies that

$$h(\mathcal{H}) = |A| - |V| + 1 \geq 2(|V| - |X|) - |V| + 1 = |V| - 2|X| + 1. \qquad \blacksquare$$

Lemma 5.2. *Let $\mathcal{T}$ be a rooted phylogenetic tree on X and let $\mathcal{H}$ be a regular hybrid. Suppose that $\mathcal{H}$ displays $\mathcal{T}$. Then there is a map $\psi\colon \mathcal{C}(\mathcal{T}) \to \mathcal{C}(\mathcal{H})$ such that*

(i) *for all $x \in X$, $\psi(\{x\}) = \{x\}$ and*
(ii) *for all $A, B \in \mathcal{C}(\mathcal{T})$ with $A \subset B$, we have $\psi(A) \subset \psi(B)$.*

Proof. Since $\mathcal{H}$ displays $\mathcal{T}$, there is a rooted subdigraph $\mathcal{T}'$ of $\mathcal{H}$ that is a refinement of $\mathcal{T}$. Let $\psi_1\colon V(\mathcal{T}) \to V(\mathcal{T}')$ be the one-to-one map defined by setting $\psi_1(v)$ to be the vertex v' of $\mathcal{T}'$ such that $c(v) = c(v')$ where $d^+(v') \geq 2$ if v is a non-leaf vertex and $d(v') = 1$ if v is a leaf vertex. It is easily seen that ψ_1 is well-defined. Since $\mathcal{T}'$ is a subdigraph of $\mathcal{H}$, each vertex of $\mathcal{T}'$ is a vertex of $\mathcal{H}$. With this in mind, let $\psi\colon \mathcal{C}(\mathcal{T}) \to \mathcal{C}(\mathcal{H})$ be the map defined by setting $\psi(A)$ to be the cluster of $\mathcal{H}$ whose associated vertex is the vertex of $\mathcal{T}'$ that is assigned the vertex of $\mathcal{T}$ corresponding to A under ψ_1. We now show that ψ satisfies (i) and (ii) in the statement of the lemma.

Clearly, ψ satisfies (i). Furthermore, if A and B are clusters of $\mathcal{T}$, and $A \subset B$, then, as $\mathcal{T}'$ is a rooted subdigraph of $\mathcal{H}$, the vertex corresponding to $\psi(A)$ is a descendant of the vertex corresponding to $\psi(B)$. As $\mathcal{H}$ is regular, this implies that $\psi(A) \subset \psi(B)$. It follows that ψ satisfies (ii). $\blacksquare$

The next proposition shows that for a pair of rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ the difference between $h(\mathcal{T}_1, \mathcal{T}_2)$ and $h^+(\mathcal{T}_1, \mathcal{T}_2)$ can be arbitrarily large.

Proposition 5.3. *For all $n \geq 3$, there exist two binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $h(\mathcal{T}_1, \mathcal{T}_2) = n - 2$, but $h^+(\mathcal{T}_1, \mathcal{T}_2) = 1$.*

Proof. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the two rooted binary phylogenetic trees shown in Figure 3. Since $\mathcal{T}_1$ and $\mathcal{T}_2$ are binary, and neither $\mathcal{T}_1$ nor $\mathcal{T}_2$ displays the other, any regular hybrid that displays both $\mathcal{T}_1$ and $\mathcal{T}_2$ requires at least one hybridisation. Let $\mathcal{H}^+$ be the regular hybrid as shown in Figure 3, where $x \notin \{1, 2, \ldots, n\}$. As $\mathcal{H}^+$ displays both $\mathcal{T}_1$ and $\mathcal{T}_2$, it follows that $h^+(\mathcal{T}_1, \mathcal{T}_2) = 1$. We next show that $h(\mathcal{T}_1, \mathcal{T}_2) = n - 2$.

Let $\mathcal{H}$ be the regular hybrid as shown in Figure 3. Now $\mathcal{H}$ displays both $\mathcal{T}_1$ and $\mathcal{T}_2$, and $h(\mathcal{H}) = n - 2$. To see that $h(\mathcal{T}_1, \mathcal{T}_2)$ is not less than $n - 2$, let $\mathcal{H}'$ be a regular hybrid on $\{1, 2, \ldots, n\}$ that displays both $\mathcal{T}_1$ and $\mathcal{T}_2$. Since $\mathcal{H}'$ displays $\mathcal{T}_1$, it follows by Lemma 5.2 that there exists a map $\psi_1\colon \mathcal{C}(\mathcal{T}_1) \to \mathcal{C}(\mathcal{H}')$ such that

$$\{1\} = \psi_1(\{1\}) \subset \psi_1(\{1,2\}) \subset \cdots \subset \psi_1(\{1, 2, \ldots, n-1\}) \subset \psi_1(\{1, 2, \ldots, n\}).$$

Since $\mathcal{T}_1$ and $\mathcal{H}'$ have the same label set, $\psi_1(\{1, 2, \ldots, n\}) = \{1, 2, \ldots, n\}$. This now implies that, for all i, $\psi_1(\{1, 2, \ldots, i\}) = \{1, 2, \ldots, i\}$. Hence $\mathcal{C}(\mathcal{T}_1) \subseteq \mathcal{C}(\mathcal{H}')$. Similarly, $\mathcal{C}(\mathcal{T}_2) \subseteq \mathcal{C}(\mathcal{H}')$. Therefore $|V(\mathcal{H}')| \geq n + 1 + 2(n-2) = 3n - 3$. By Lemma 5.1, this implies that

$$h(\mathcal{H}') \geq |V(\mathcal{H}')| - 2n + 1 \geq (3n - 3) - 2n + 1 = n - 2.$$

Hence $h(\mathcal{T}_1, \mathcal{T}_2) = n - 2$. $\blacksquare$

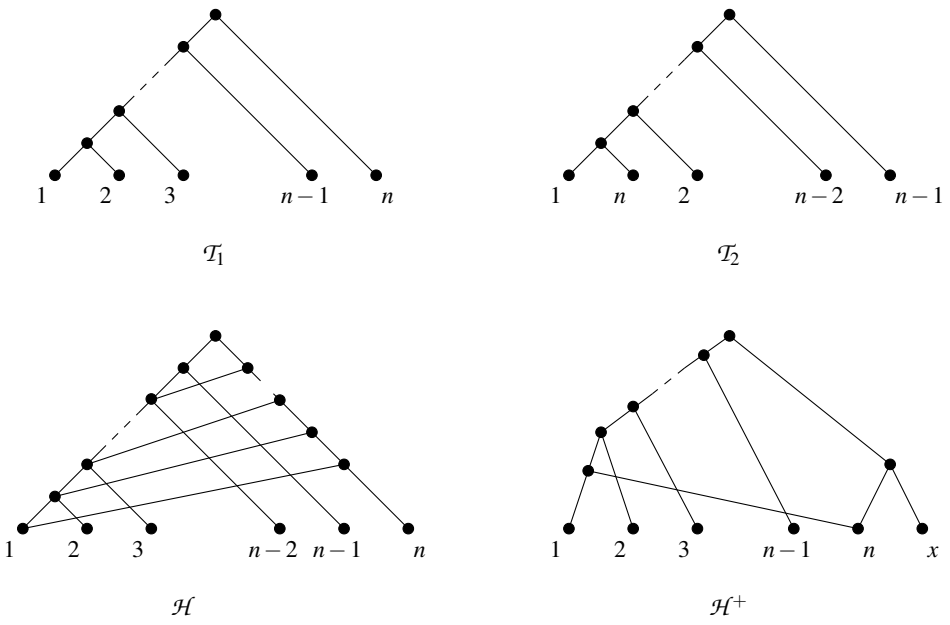


Figure 3: Two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, and two regular hybrids that display both $\mathcal{T}_1$ and $\mathcal{T}_2$.

For a collection $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ of rooted phylogenetic trees and a subset U of $\mathcal{L}(\mathcal{P})$, let

$$\mathcal{P}|U = \{\mathcal{T}_1|U, \mathcal{T}_2|U, \ldots, \mathcal{T}_k|U\}.$$

Lemma 5.4. *Let $\mathcal{T}$ be a rooted phylogenetic tree and let $\mathcal{H}$ be a hybrid that displays $\mathcal{T}$. Then $\mathcal{H}$ displays $\mathcal{T}|U$ for all subsets U of $\mathcal{L}(\mathcal{T})$.*

Proof. Since $\mathcal{H}$ displays $\mathcal{T}$ there is a rooted subdigraph $\mathcal{T}'$ of $\mathcal{H}$ that is a refinement of $\mathcal{T}$. Consider the minimal rooted subtree of $\mathcal{T}'$ that connects the elements in U. This minimal rooted subtree is a rooted subdigraph of $\mathcal{H}$ and, moreover, it is a refinement of $\mathcal{T}|U$. It follows that $\mathcal{H}$ displays $\mathcal{T}|U$. ∎

Proposition 5.5. *Let $\mathcal{P}$ be a collection of rooted phylogenetic trees. Then, for all subsets U of $\mathcal{L}(\mathcal{P})$, we have $h^+(\mathcal{P}|U) \leq h^+(\mathcal{P})$.*

Proof. Let $\mathcal{H}$ be a regular hybrid that displays $\mathcal{P}$ with the property that $h^+(\mathcal{P}) = h(\mathcal{H})$. By Lemma 5.4, $\mathcal{H}$ displays $\mathcal{P}|U$. It follows that $h^+(\mathcal{P}|U) \leq h^+(\mathcal{P})$ as required. ∎

The inequality in Proposition 5.5 does not necessarily hold if h^+ is replaced with h. In particular, we have the following proposition.

Proposition 5.6. *For all non-negative integers k, there is a collection $\mathcal{P}$ of rooted binary phylogenetic trees on X and a subset U of X such that*

$$h(\mathcal{P}|U) \geq h(\mathcal{P}) + k.$$

Proof. Let k be a non-negative integer and let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$, where $\mathcal{T}_1$ and $\mathcal{T}_2$ are the two rooted binary phylogenetic trees shown in Figure 4. Let $U = \{1, 2, \ldots, k+3, k+4\}$. Then, up to isomorphism, $\mathcal{P}|U$ consists of the two rooted binary phylogenetic trees shown in Figure 3. By Proposition 5.3, $h(P|U) = k+2$. Furthermore, the hybrid $\mathcal{H}$ shown in Figure 4 is regular and displays both $\mathcal{T}_1$ and $\mathcal{T}_2$. Since $h(\mathcal{H}) = 2$, this implies that $h(\mathcal{P}) \leq 2$. The proposition now follows. ∎

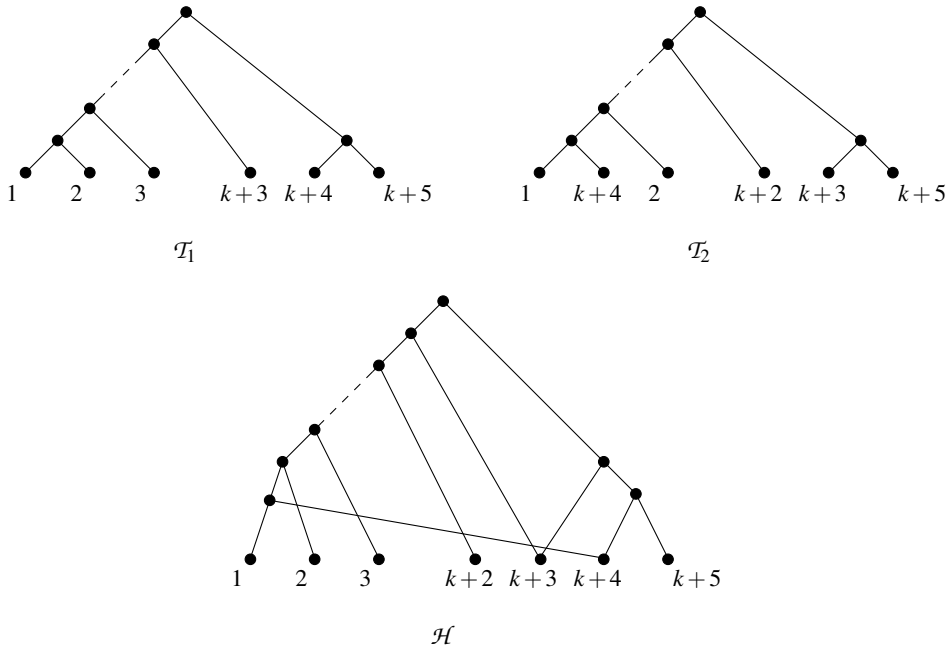


Figure 4: Two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, and a regular hybrid phylogeny that displays $\mathcal{T}_1$ and $\mathcal{T}_2$.

6. The Cluster Union Hybrid

As mentioned at the end of Section 3, a natural way to obtain a regular hybrid $\mathcal{H}$ that displays a collection $\mathcal{P}$ of rooted phylogenetic trees on the same leaf set is by taking $\mathcal{H}$ to be the regular hybrid whose set of clusters is the union of the sets of clusters of the trees in $\mathcal{P}$. In this section, we consider the special case when $\mathcal{P}$ consists of two rooted phylogenetic trees.

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic trees on X. We denote the cover hybrid phylogeny $\mathcal{H}(\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2))$ of $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$ by $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$.

Lemma 6.1. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic trees on X. Then*

(i) *The in-degree of any vertex of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ is at most two.*

(ii) *The hybridisation number of* $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ *is equal to*

$$|\{v \in V(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2])\colon d^-(v) = 2\}|.$$

Proof. To prove (i), assume that there is a vertex v of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ of in-degree at least three. Then (at least) two of the immediate ancestors of v, u_1 and u_2 say, have the property that $c(u_1)$ and $c(u_2)$ are either both clusters of $\mathcal{T}_1$ or both clusters of $\mathcal{T}_2$. Furthermore, $c(v) \subseteq c(u_1) \cap c(u_2)$ and so $c(u_1) \cap c(u_2) \neq \emptyset$. But since $c(u_1)$ and $c(u_2)$ are distinct, and clusters of the same rooted phylogenetic tree this implies that either $c(u_1) \subset c(u_2)$ or $c(u_2) \subset c(u_1)$. In the former (respectively, the latter), (u_2, v) (respectively, (u_1, v)) is not an arc of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$, a contradiction.

The second part of Lemma 6.1 now follows from part (i) and the definition of the hybridisation number of a hybrid phylogeny. ∎

For two rooted phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$, let $f\colon \mathcal{C}(\mathcal{T}_1) \times \mathcal{C}(\mathcal{T}_2) \to 2^{2^X}$ and $g\colon \mathcal{C}(\mathcal{T}_1) \times \mathcal{C}(\mathcal{T}_2) \to 2^{2^X}$ be the maps defined by

$$g(A, B) = \{S \subseteq A \cap B\colon S \in \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)\},$$

and

$$f(A, B) = \begin{cases} \emptyset, & \text{if } A \cap B \in \{\emptyset, A, B\}, \\ \max g(A, B), & \text{otherwise,} \end{cases}$$

where $\max g(A, B)$ denotes the set of maximal elements of $g(A, B)$ under set inclusion. The next proposition describes exactly the hybridisation vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$.

Proposition 6.2. *Let* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *be two rooted phylogenetic trees on* X. *Then*

$$\bigcup_{(A,B) \in \mathcal{C}(\mathcal{T}_1) \times \mathcal{C}(\mathcal{T}_2)} f(A, B)$$

is the collection of clusters corresponding to the hybridisation vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$.

Proof. Throughout the proof, we denote $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ by $\mathcal{H}$. By Lemma 6.1(ii), it suffices to prove that

$$\{c(v)\colon v \in V(\mathcal{H}), d^-(v) = 2\} = \bigcup \{f(A, B)\colon A \cap B \notin \{\emptyset, A, B\}\}.$$

We will establish equality of these two sets by showing set inclusion in both directions.

Let $v \in V(\mathcal{H})$ with $d^-(v) = 2$. Then v has two immediate ancestors, v_1 and v_2 say, such that $c(v) \subseteq c(v_1) \cap c(v_2)$ and, without loss of generality, $c(v_i) \in \mathcal{C}(\mathcal{T}_i)$ for each $i \in \{1, 2\}$. Set $S = c(v)$, $A = c(v_1)$, and $B = c(v_2)$. Clearly, $A \cap B$ is non-empty. Also, A is not a subset of B, for otherwise $c(v) \subset c(v_1) \subset c(v_2)$ contradicting the definition of $\mathcal{H}$. Similarly, B is not a subset of A, and so $A \cap B \notin \{\emptyset, A, B\}$. Now S is a maximal element of $g(A, B)$. To see this, suppose there exists $S' \in \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$ with $S \subset S' \subseteq A \cap B$. Then there exists $v' \in V(\mathcal{H})$ such that $S' = c(v')$. But then $c(v) \subset c(v') \subset c(v_1)$ again contradicting the definition of $\mathcal{H}$. Thus if $d^-(v) = 2$, then $c(v) \in \bigcup \{f(A, B)\colon A \cap B \notin \{\emptyset, A, B\}\}$.

Now suppose that $S \in f(A, B)$, where $A \in \mathcal{C}(\mathcal{T}_1)$, $B \in \mathcal{C}(\mathcal{T}_2)$, and $A \cap B \notin \{\emptyset, A, B\}$. Then there exist vertices v, v_1, and v_2 such that $c(v) = S$, $c(v_1) = A$, and $c(v_2) = B$. As S is a subset of $A \cap B$, there exist paths in $\mathcal{H}$ from v_1 to v and from v_2 to v. By Lemma 6.1(i), it suffices to show that $d^-(v) \neq 1$. Assume that $d^-(v) = 1$, and let v' be the immediate ancestor of v in $\mathcal{H}$. Let $S' = c(v')$. Each of the above paths must pass through v'. Therefore $S' \subseteq A$ and $S' \subseteq B$, contradicting the fact that S is a maximal element of $g(A, B)$. Hence $S \in \{c(v) : v \in V(\mathcal{H}), d^-(v) = 2\}$. This completes the proof of the proposition. ∎

Since the hybridisation vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ are precisely the vertices that have in-degree two, it follows from Proposition 6.2 that the hybridisation number of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ is equal to $|\bigcup\{f(A, B) : A \cap B \notin \{\emptyset, A, B\}\}|$. Furthermore, a bound on the number of vertices that have in-degree two can be easily obtained as follows. As a rooted phylogenetic tree with n leaves has at most $2n-1$ vertices, such a tree has at most $n-2$ interior vertices. Therefore the cover hybrid $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ of two rooted phylogenetic trees with the same label set of size n has at most $2(n-2)+n$ vertices excluding the root. At least two of these vertices are adjacent to the root, in which case they have in-degree equal to 1. Hence the number of vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ with in-degree two is at most $2(n-2)+n-2 = 3n-6$. Thus $h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) \leq 3n-6$.

7. The Incompatibility Graph

A classical result in phylogenetics is the following lemma (see [16]).

Lemma 7.1. *Let $\mathcal{C}$ be a collection of subsets of X that contains X and each singleton subset. Then there exists a rooted phylogenetic tree $\mathcal{T}$ on X whose set of clusters is $\mathcal{C}$ if and only if, for all $A, B \in \mathcal{C}$,*

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

Moreover, if $\mathcal{C}$ is such a collection, then $\mathcal{T}$ is the unique rooted phylogenetic tree on X that has $\mathcal{C}$ as its set of clusters.

One consequence of Lemma 7.1 is that a rooted phylogenetic tree can be recovered from its set of clusters. In the last section of this paper, we investigate an extension of this, in particular, for two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, what information can be inferred about $\mathcal{T}_1$ and $\mathcal{T}_2$ from $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$.

For a set X, we denote $\{X\} \cup \{\{x\} : x \in X\}$ by X_{triv}. Now let $\mathcal{C}$ be a collection of subsets of X. The *incompatibility graph of $\mathcal{C}$* is the graph that has vertex set $\mathcal{C}$ and an edge joining two vertices A and B precisely if $A \cap B \notin \{\emptyset, A, B\}$. A graph G is said to be 2-*colourable* if each vertex of G can be assigned one of two colours so that adjacent vertices are assigned different colours.

Proposition 7.2. *Let G be the incompatibility graph of a collection $\mathcal{C}$ of subsets of X. Then G is 2-colourable if and only if there exists a pair of rooted phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that*

$$\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2) = \mathcal{C} \cup X_{\text{triv}}.$$

Proof. Assume that G is 2-colourable. Under a 2-colouring of G, let $\mathcal{C}_1$ be the set of vertices of G assigned one colour and let $\mathcal{C}_2$ be the set of vertices of G assigned the other colour. Now, for all $A, B \in \mathcal{C}_1$, we have $A \cap B \in \{\emptyset, A, B\}$. By Lemma 7.1, this implies that $\mathcal{C}_1 \cup X_{\text{triv}}$ is the collection of clusters of a rooted phylogenetic tree on X. Similarly, $\mathcal{C}_2 \cup X_{\text{triv}}$ is the collection of clusters of a rooted phylogenetic tree on X.

For the converse, assume that there exists a pair of rooted phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2) = \mathcal{C} \cup X_{\text{triv}}$. Consider the incompatibility graph G of $\mathcal{C}$. Colour the vertices of G in $\mathcal{C}(\mathcal{T}_1)$ one colour and the vertices of G in $\mathcal{C}(\mathcal{T}_2)$ another colour. The choice of colour for a vertex in both sets is arbitrary as all such vertices are isolated in G. To show that this is a 2-colouring of G, let $\{A, B\}$ be an edge of G. Then $A \cap B \notin \{\emptyset, A, B\}$. Therefore, by Lemma 7.1, A and B are not clusters of the same tree, and so A and B are assigned different colours. Thus this assignment of colours is indeed a 2-colouring of G. ∎

It is a straightforward exercise to show that if G is a connected graph that is 2-colourable, then the partition of the vertex set of G induced by such a colouring is unique. We use this fact and Proposition 7.2 freely in the proof of the next proposition. For a fixed set X, let $\mathcal{P}(X)$ denote the collection of rooted phylogenetic trees on X.

Proposition 7.3. *Let $\mathcal{C}$ be a non-empty collection of subsets of X, and suppose that the incompatibility graph G of $\mathcal{C}$ is 2-colourable. Let k be the number of components of G with at least two vertices and let m be the number of isolated vertices of G that are not in X_{triv}. Then the number of 2-element subsets $\{\mathcal{T}_1, \mathcal{T}_2\}$ of $\mathcal{P}(X)$ for which $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2) = \mathcal{C} \cup X_{\text{triv}}$ is equal to $2^{k-1}3^m$ if $k \geq 1$ and $\frac{1}{2}(1+3^m)$ if $k = 0$.*

Proof. Clearly, it suffices to show that the result holds when $\mathcal{C}$ does not contain any element of X_{triv}. Suppose an appropriate assignment of colours for the non-isolated vertices of G is made. Then, by Lemma 7.1, any subset of the isolated vertices can be added to either of the two subsets of vertices of G assigned a particular colour and the resulting set together with X and the singleton subsets of X are exactly the clusters of a unique rooted phylogenetic tree on X. Indeed, all desirable 2-element subsets of $\mathcal{P}(X)$ can be obtained in this way.

We partition the proof into two cases depending upon whether (i) $k \geq 1$ or (ii) $k = 0$. Consider (i). Up to choosing colours, the number of ways of colouring the components of G that have at least two vertices is 2^{k-1}. Now each such colouring partitions the vertex set of G restricted to these components into two non-empty parts V_1 and V_2. By the last paragraph, each isolated vertex is either added to V_1, V_2, or both. Since there are m such vertices, there are 3^m ways of doing this. It now follows that, for $k \geq 1$, there are $2^{k-1}3^m$ 2-element subsets $\{\mathcal{T}_1, \mathcal{T}_2\}$ of $\mathcal{P}(X)$ for which $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2) = \mathcal{C} \cup X_{\text{triv}}$.

Now consider (ii). The only significant difference with this case is that $\mathcal{T}_1$ and $\mathcal{T}_2$ are not equal, and that we do not double count the 2-element subsets $\{\mathcal{T}_1, \mathcal{T}_2\}$ of $\mathcal{P}(X)$ for which $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2) = \mathcal{C} \cup X_{\text{triv}}$. It follows that there are $\frac{3^m-1}{2} + 1 = \frac{1+3^m}{2}$ distinct such sets. ∎

Corollary 7.4 is an immediate consequence of Proposition 7.3.

Corollary 7.4. *Let $\mathcal{C}$ be a non-empty collection of subsets of X that does not contain any element of X_{triv}, and let G be the incompatibility graph of $\mathcal{C}$. Then there is exactly*

one 2-element subset $\{\mathcal{T}_1, \mathcal{T}_2\}$ *of* $\mathcal{P}(X)$ *for which* $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2) = \mathcal{C} \cup X_{\text{triv}}$ *if and only if* G *is a 2-colourable connected graph with at least two vertices.*

The statement of Corollary 7.4 can be strengthened if we restrict $\mathcal{T}_1$ and $\mathcal{T}_2$ to be rooted binary phylogenetic trees. To obtain this strengthening (Corollary 7.6), we use the next lemma.

Lemma 7.5. *Let* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *be two rooted binary phylogenetic trees. Then* A *is an isolated vertex of the incompatibility graph* G *of* $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$ *if and only if* A *is a cluster of both* $\mathcal{T}_1$ *and* $\mathcal{T}_2$.

Proof. It follows from Lemma 7.1 that if A is a cluster of both $\mathcal{T}_1$ and $\mathcal{T}_2$, then A is an isolated vertex of the incompatibility graph G.

To prove the converse, suppose that A is an isolated vertex of G. Without loss of generality, we may assume that A is a cluster of $\mathcal{T}_1$. Assume that A is not a cluster of $\mathcal{T}_2$, and let A' be the minimal cluster of $\mathcal{T}_2$ that contains A. By assumptions, $A \neq A'$ and $|A'| \geq 3$. Since $\mathcal{T}_2$ is binary and $|A'| \geq 3$, it follows that $\mathcal{T}_2$ has a cluster B that is properly contained in A' and contains elements of A and $X - A$. Furthermore, as A' is the minimal cluster of $\mathcal{T}_2$ that contains A, we deduce that $A \cap B \notin \{\emptyset, A, B\}$. Hence, in G, there is an edge joining A and B. This contradicts the isolation of A in G. Thus A is a cluster of $\mathcal{T}_2$. ∎

Corollary 7.6 is a straightforward consequence of Proposition 7.3 and Lemma 7.5.

Corollary 7.6. *Let* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *be two rooted binary phylogenetic trees, and let* $\mathcal{C} = \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$. *If the incompatibility graph* G *of* $\mathcal{C}$ *has at most one component with at least two vertices, then* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *is the only pair of trees of* $\mathcal{P}(X)$ *whose union of clusters is* $\mathcal{C}$. *Furthermore, if* G *consists of isolated vertices, then* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *are isomorphic.*

For two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ for which the incompatibility graph G of $\mathcal{C}$ has at most one component with at least two vertices, $\mathcal{T}_1$ and $\mathcal{T}_2$ can be reconstructed from G. This is done by constructing the clusters of $\mathcal{T}_1$ and $\mathcal{T}_2$ as follows. If G consists of isolated vertices, then $\mathcal{T}_1$ and $\mathcal{T}_2$ are isomorphic, and the vertex set of G is the set of clusters of $\mathcal{T}_1$. On the other hand, if G has exactly one component C with at least two vertices, 2-colour this component and, for each $i \in \{1, 2\}$, set $\mathcal{C}_i$ to be the union of the set of isolated vertices of G and the subset of vertices of C assigned one particular colour. Then, for each i, $\mathcal{C}_i$ is the set of clusters of $\mathcal{T}_i$.

Tree rearrangement operations have been extensively studied for rooted and unrooted binary phylogenetic trees (for example, see [1, 12]). Such operations include 'nearest neighbour interchange' and 'subtree prune and regraft'. Here we are interested in the latter operation for rooted binary phylogenetic trees. Let $\mathcal{T}$ be a rooted binary phylogenetic tree and let $e = \{u, v\}$ be an edge of $\mathcal{T}$ where $u \leq_{\mathcal{T}} v$. Here, $u \leq_{\mathcal{T}} v$ means that u is in the path from the root of $\mathcal{T}$ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic tree obtained from $\mathcal{T}$ by deleting e, adding an edge between v and a vertex that subdivides an edge in the component containing u, and then suppressing any resulting degree-two vertex. In the case u is the root of $\mathcal{T}$, we must also delete the other edge incident with u. We also allow the reverse of this operation. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a single *subtree prune and regraft* (SPR) operation.

Theorem 7.7. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two distinct rooted binary phylogenetic X-trees. Suppose that $\mathcal{T}_2$ can be obtained from $\mathcal{T}_1$ by a single* SPR *operation. Then*

(i) *The incompatibility graph of $\mathcal{C} = \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$ has exactly one component with at least two vertices.*

(ii) *Both $\mathcal{T}_1$ and $\mathcal{T}_2$ can be reconstructed from the incompatibility graph of $\mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$.*

Proof. Consider a single SPR operation that takes $\mathcal{T}_1$ to $\mathcal{T}_2$. Let $e = \{u, v\}$ be the edge of $\mathcal{T}_1$ that is pruned in performing this operation and let $e' = \{u', v'\}$ be the edge of $\mathcal{T}_1$ that is subdivided, where $u \leq_{\mathcal{T}_1} v$ and $u' \leq_{\mathcal{T}_1} v'$. We may assume that e and e' are not adjacent, for otherwise, $\mathcal{T}_1$ and $\mathcal{T}_2$ are equal. Let A be the cluster of $\mathcal{T}_1$ associated with v. Furthermore, let w be the vertex of $\mathcal{T}_2$ that is the result of subdividing e'. There are two cases to consider depending upon whether (a) the paths from u and u' to the root contain neither u' nor u, respectively, or (b) the path from u to the root or u' to the root includes u' or u, respectively.

First consider case (a). It is easily checked that $\mathcal{C}(\mathcal{T}_1) \cap \mathcal{C}(\mathcal{T}_2)$ consists of all of the clusters of $\mathcal{T}_1$ whose associated vertices do not lie on the path from u to u' except for the vertex t of $\mathcal{T}_1$ that is the most recent common ancestor of u and u'. Let $\mathcal{D}$ be the set of clusters of $\mathcal{T}_1$ and $\mathcal{T}_2$ that are not in this intersection. To prove (a), we will show that the incompatibility graph G of $\mathcal{D}$ consists of a single component. Using the fact that the paths from u to t and t to u' are vertex disjoint apart from t, it is easily seen that every element of $\mathcal{D}$ that is in $\mathcal{C}(\mathcal{T}_1)$ and contains A is joined to every element of $\mathcal{D}$ that is in $\mathcal{C}(\mathcal{T}_2)$ and contains A. To complete the proof of (a), it suffices to show that, for each remaining element of $\mathcal{D}$, there is an edge joining it to an element of $\mathcal{D}$ that contains A.

Let B be an element of $\mathcal{D}$ that does not contain A. First assume that B is an element of $\mathcal{C}(\mathcal{T}_2)$. Now $B \cup A$ is a cluster of $\mathcal{T}_1$. Furthermore, as B is not a cluster of $\mathcal{T}_1$, there is a cluster $B' \cup A$ in $\mathcal{T}_1$ such that $B' \cup A$ is a proper subset of $B \cup A$ and $B' \cup A \in \mathcal{D}$. It follows that B is joined to $B' \cup A$ by an edge in G. By symmetry, if B is a cluster of $\mathcal{T}_1$, then B is joined by an edge to a vertex of G that contains A. This completes the proof of (a).

The proof of (b) is similar and is omitted. This completes the proof of (i). Moreover, combining the construction described after Corollary 7.6 with (i), we obtain (ii). ∎

Theorem 7.7(ii) cannot be extended to two rooted binary phylogenetic trees that require at least two SPR operations for one to be obtained from the other. To see this, consider the rooted binary phylogenetic trees shown in Figure 5. It is routinely checked that two SPR operations are required to transform $\mathcal{T}_1$ to $\mathcal{T}_2$ and $\mathcal{T}_1'$ to $\mathcal{T}_2'$. The union of the clusters of $\mathcal{T}_1$ and $\mathcal{T}_2$, and the union of the clusters of $\mathcal{T}_1'$ and $\mathcal{T}_2'$ are equal. In particular, the associated incompatibility graphs are identical.

We end this section by noting a connection between the subtree prune and regraft operation and the hybridisation number of two rooted binary phylogenetic trees. Observe that, in Figure 3, the minimum number of SPR operations required to obtain $\mathcal{T}_2$ from $\mathcal{T}_1$ is 1. Furthermore, in the proof of Proposition 5.3, we showed that $h^+(\mathcal{T}_1, \mathcal{T}_2) = 1$. The fact that this hybridisation number is equal to 1 is no coincidence. The proof of Proposition 7.8 is straightforward and omitted.

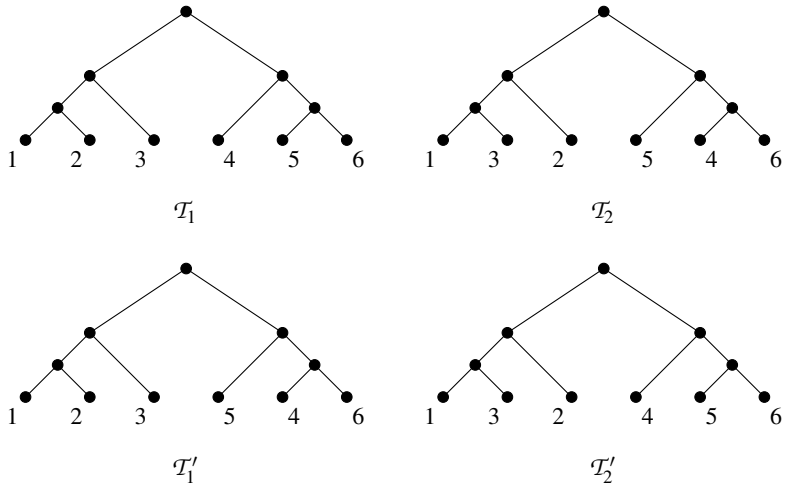


Figure 5: Two pairs of trees with an identical union of clusters.

Proposition 7.8. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic trees. Suppose that $d_{SPR}(\mathcal{T}_1, \mathcal{T}_2) = 1$. Then $h^+(\mathcal{T}_1, \mathcal{T}_2) = 1$.*

References

1. B.L. Allen and M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Comb. **5** (2001) 1–13.
2. H.-J. Bandelt and A.W.M. Dress, Weak hierarchies associated with similarity measures - an additive clustering technique, Bull. Math. Biol. **51** (1989) 133–166.
3. J.B. -Jensen and G. Gutin, Digraphs: Theory, Algorithms and Applications, Springer-Verlag, London, 2001.
4. K. Bremer and H.-E. Wanntorp, Hierarchy and reticulation in systematics, Syst. Zool. **28** (4) (1979) 624–627.
5. E. Diday and P. Bertrand, An extension of hierarchical clustering: the pyramidal presentation, In: Pattern Recognition in Practice II, E.S. Gelsema and L.N. Kanal, Eds., North-Holland, Amsterdam, 1986, pp. 411–423.
6. W.F. Doolittle, Phylogenetic classification and the universal tree, Science **284** (1999) 2124–2128.
7. 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.
8. V.A. Funk, Phylogenetic patterns and hybridization, Ann. Missouri Bot. Gand. **72** (1985) 681–715.
9. F.-J. Lapointe, How to account for reticulation events in phylogenetic analysis: a comparison of distance-based methods, J. Classification **17** (2000) 175–184.
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, Syst. Biol. **51** (2) (2002), 199–216.

12. M. Li, J. Tromp, and L. Zhang, On the nearest neighbour interchange distance between evolutionary trees, Journal of Theoretical Biology **182** (1996) 463–467.
13. G. Nelson, Reticulation in cladograms, In: Advances in Cladistics, vol. 2, N.I. Platnick and V.A. Funk, Eds., Columbia University Press, 1983, pp. 105–111.
14. D.F. Robinson and L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. **53** (1981) 131–147.
15. F.J. Rohlf, Phylogenetic models and reticulations, J. Classification **17** (2000) 185–189.
16. C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.
17. P.E. Smouse, Reticulation inside the species boundary, J. Classification **17** (2000) 165–173.
18. P.H.A. Sneath, Reticulate evolution in bacteria and other organisms: how can we study it?, J. Classification **17** (2000) 159-163.
19. H.-E. Wanntorp, Reticulated cladograms and the identification of hybrid taxa, In: Advances in Cladistics, vol. 2, N.I. Platnick and V.A. Funk, Eds., Columbia University Press, 1983, pp. 81–88.