Version dated: July 7, 2014

How much information is needed to infer reticulate evolutionary histories?

Katharina T. Huber[1], Leo van Iersel[2], Vincent Moulton[1] and Taoyang Wu[1]

[1] *School of Computing Sciences, University of East Anglia, Norwich, United Kingdom*

[2] *CWI, Amsterdam, Netherlands*

Corresponding author: Vincent Moulton, School of Computing Sciences, University of East Anglia, Norwich, United Kingdom; E-mail: vincent.moulton@cmp.uea.ac.uk.

Abstract.— Phylogenetic networks are a generalization of evolutionary trees and are an important tool for analyzing reticulate evolutionary histories. Recently, there has been great interest in developing new methods to construct rooted phylogenetic networks, that is, networks whose internal vertices correspond to hypothetical ancestors, leaves to sampled taxa, and in which vertices with more than one parent correspond to taxa formed by reticulate evolutionary events such as recombination or hybridization. Several methods for constructing evolutionary trees use the strategy of building up a tree from simpler building blocks (such as triplets or clusters), and so it is natural to look for ways to construct networks from smaller networks. In this paper we shall demonstrate a fundamental issue with this approach. Namely, we show that even if we are given *all* of the subnetworks induced on all proper subsets of the leaves of some rooted phylogenetic network, we still do not have all of the information required to completely determine that network. This implies that even if *all* of the building blocks for some reticulate evolutionary history were to be taken as the input to any given network building method, the method might still output an incorrect history. We also discuss some potential consequences of this result for constructing phylogenetic networks.

(Keywords: phylogenetic network, reticulate evolution, network reconstruction, evolutionary tree)

Introduction

Modern systematics assumes a tree as an integral component of the evolutionary model (Penny et al. 1992). However, genome science is also delivering a level of complexity previously under appreciated for many biological systems and organisms (e.g. Mallet 2007; Abbott et al. 2013; Liu et al. 2013; Muhlfeld et al. 2014). This growing appreciation has in turn motivated the development of phylogenetic networks (see e.g. Huson et al. 2010; Morrison 2011). These networks are a generalization of evolutionary trees and, in the broadest sense, can be any type of graph-theoretical network that is used to represent potentially complex patterns of evolutionary relationship.

Some networks, also referred to as data-display or split networks (Dress and Huson 2004; Morrison 2010), attempt only to represent bipartitions or splits in data, and the evidence these splits provide for contradictory relationships. In these networks the internal nodes usually have no explicit meaning. Such networks generalize unrooted evolutionary trees and have been used to visualize homoplasy and detect errors in human sequence data (e.g. Bandelt et al. 2000, 2001), for visualizing the support for particular bifurcating trees and hypotheses (e.g. Holland et al. 2005) and for exploring the genetic complexity of plant and animal datasets (e.g. Morrison 2005). There are numerous ways of computing these graphs (e.g. Huson and Bryant 2006). Methods essentially differ in the extent to which they visualize incompatibilities either because of the way they compute the splits and/or because of the dimensionality of the displayed network.

Other networks, also referred to as genealogical networks, are constructed to model evolutionary history wherein the evolution is suspected of being reticulate in nature. These networks, which are the focus of the present study, are typically rooted and contain internal vertices that represent hypothetical ancestors and leaves that represent taxa sampled from the data (extant or extinct). They are directed graphs with a single root vertex and leaves labelled with taxon names (see e.g. Fig. 1 and Mathematical Definitions section). They also contain no directed cycles, thus ensuring that no taxon can be a descendent of itself. In these networks vertices with more than one parent correspond to taxa that are formed by reticulate evolutionary events such as recombination or hybridization. In particular, a rooted evolutionary tree is a special type of rooted phylogenetic network which does not represent any reticulate evolutionary events. Genealogical networks are reviewed in e.g. (Huson et al. 2010), and have been used to study the evolution of organisms such as plants (Marcussen et al. 2011), viruses (Visser et al. 2012) and bacteria (Kunin et al. 2005).

Various methods have been proposed to construct genealogical phylogenetic networks, although it is generally agreed that there is still much more to be done in this direction (see e.g. Nakhleh 2011; Bapteste et al. 2013). Many of these methods follow a strategy that is also commonly used to build evolutionary trees (e.g. to construct supertrees), namely to infer networks from building blocks such as triplets (evolutionary trees with three leaves) (Huber et al. 2011), evolutionary trees (Kelk et al. 2012; D.Huson and Scornavacca 2012) or clusters/clades (van Iersel et al. 2010). However, a fundamental issue with this strategy is that the commonly used building blocks do not necessarily determine or *encode* networks, in contrast to evolutionary trees. In other words, there can be pairs of rooted phylogenetic networks that do not represent the same evolutionary histories, but still display exactly the same building blocks (see e.g. Gambette and Huber (2012) for triplets and clusters, and Willson (2011) for evolutionary trees). For example, considering the two networks in Fig. 1 (i) and (ii), the first one of which is adapted from the network pictured in (van Iersel et al. 2009, Fig. 10), which was constructed from a dataset of the yeast *Cryptococcus gattii* (see Hagen et al. 2013, for a following up study). Both networks display the same collection of evolutionary trees (pictured in the Supplementary Material), and therefore the same triplets, but they are not equivalent as networks. This is of importance since it implies that even if *all* of the building blocks for some reticulate evolutionary history were to be taken as the input to any given network building method, the method might still output an incorrect history.

To address this problem, it was recently proposed to try and construct networks using a network analogue of triplets called *trinets* (Huber and Moulton 2012). Trinets are rooted phylogenetic networks with three leaves (see e.g. Fig. 1 (iii)-(v)); they can be induced on any three leaves of a rooted network by taking the union of all paths from the root to one of the three leaves, and then removing all vertices that lie above the last vertex that is on all such paths, and suppressing parallel edges (see Supplementary Material for an illustration). For example, in Fig. 1 the trinet pictured in (iii) is induced on the three leaves c, e, f of the network pictured in (i). Note that in this example, even though the two networks in (i) and (ii) both induce the trinet pictured in (v), the trinets (iii) and (iv) that they induce on c, e, f are not equivalent. In particular, it follows that the networks in (i) and (ii) are also not equivalent. Thus, considering trinets could hold some promise for distinguishing between networks, especially since some special types of rooted phylogenetic networks (e.g. level-1, level-2 and tree-child) are in fact encoded by their trinets (see e.g. Huber and Moulton 2012; van Iersel and Moulton 2014).

Even so, when trying to extend these results on trinet encodings to more general networks we were somewhat surprised to discover that trinets do not necessarily encode networks. Indeed, more generally, in this paper we shall show that even if we are given the networks induced on *all* subsets of the leaves of a network except for the leaf-set itself, (which includes all possible trinets), we still do not necessarily have enough information to encode the network. More specifically, for any set X of taxa of size at least three, we shall present an example of two non-equivalent rooted, binary phylogenetic networks with leaf set X which both induce exactly the same network on any subset Y of X with $Y \neq X$ (see Theorem 3). As an illustration, we present these two networks in the case that X has four elements in Figure 2. In addition, in the Supplementary Material, we show that these networks also induce exactly the same set of evolutionary trees. Hence, even knowing all of the induced networks together with all of the induced trees for each of these two networks is still not enough information to distinguish between them.

Our examples were inspired by some results due to Thatte concerning the reconstructability of so-called pedigree graphs (Thatte 2008), which are used to represent ancestral relationships between individuals in a population. Thatte was able to show that a pedigree cannot in general be reconstructed from the collection of its proper subpedigrees. Although this result is similar in nature to ours, it is not a simple corollary, as pedigree graphs have quite a different structure to phylogenetic networks (e.g. a pedigree graph can have multiple roots or "founders" and all other vertices have two parents). Moreover, Thatte's concept of a subpedigree is different from our concept of a network induced on a subset of a network's leaves. Intriguingly, both Thatte's and our results are somewhat related to the Kelly-Ulam reconstruction conjecture that states that a graph is uniquely determined by all of its subgraphs. This conjecture is still open, although for directed graphs it is known to be false (see e.g. Stockmeyer 1977). Even so there are again important mathematical distinctions between graphs in general and phylogenetic networks and pedigrees (e.g. graphs are not labeled by a set of taxa and the concept of a subgraph is different from an induced network).

The contents of the rest of the paper are as follows. First we present some mathematical preliminaries on phylogenetic networks and also some terminology concerning binary sequences which will be key for constructing our examples. Then, given any leaf set of size at least three we present an example of two distinct *non-binary*, rooted phylogenetic networks having the same leaf set which both induce exactly the same network on any proper subset of their leaves. These were the first examples that we discovered, and at the time we were uncertain as to whether or not there could be examples of binary networks with this property, as there are various mathematical results in phylogenetics that hold for binary trees/networks but not for non-binary ones. However, by adapting our non-binary networks we are also able to construct two binary networks with the same property. Since the proof of this fact follows the same approach to that for the non-binary case but is considerably more technical, we shall present this in the Appendix. We conclude with a brief discussion of some ramifications and future directions as well as some potential consequences of our results for constructing reticulate evolutionary histories.

Mathematical Definitions

Digraphs

The basic graph-theoretical structure that underlies the phylogenetic networks in this paper is called a *digraph*. This is a connected, directed graph G consisting of a set of vertices $V(G)$ representing taxa (both hypothetical and sampled) and a set $E(G)$ of directed edges or *arcs* that join pairs of them. We denote an arc starting at vertex u and ending at vertex v by (u, v), and call u a *parent* of v and v a *child* of u. This represents the fact that u is a direct ancestor of v. The *in-* and *outdegree* of a vertex v in G is the number of arcs ending and starting at v, respectively. A vertex of G that has outdegree 0 is called a *leaf* of G (which corresponds to a sampled taxon, either extinct or extant), and the set of all leaves of G is denoted by $L(G)$. Note that vertices with indegree at most one and outdegree at least two represent speciations, whilst those with indegree at least two represent reticulations (e.g. evolutionary events such as hybridization and recombination). If a digraph G has a unique vertex with indegree zero, corresponding to a common ancestor of all of the taxa in question, then that vertex is called the *root* of G, denoted by $\rho(G)$, and we call G a *rooted digraph*. If G is rooted and G' is a further rooted digraph then we say that G and G' are *isomorphic (as digraphs)* if they are isomorphic in the usual graph-theoretical sense. If, in addition to being isomorphic, every leaf is mapped to itself by the underlying map, then G and G' are called *equivalent*.

A digraph with no directed cycles is called a *directed acyclic graph (DAG)*. For a rooted DAG G, a vertex in G that is neither a leaf nor the root is called an *interior vertex* of G. In addition, a vertex u in G is called an *ancestor* of a vertex v in G if u and v are equal[1] or there exists a directed path in G starting at u and ending at v. If u is an ancestor of v but $u \neq v$ then we say that v is *below* u. Thus, in a DAG a vertex v can never be below itself, which corresponds to the fact that v cannot be a biological descendent of itself. Furthermore, if G has at least three vertices and v is a vertex with outdegree one then we call v *degenerate* if indegree of v is not at least two. Finally, we call G *binary* if the outdegree of $\rho(G)$ is two and the sum of the indegree and outdegree of every interior vertex of G is three.

Phylogenetic Trees and Networks

Suppose for the remainder of the paper that X is some (non-empty) set of taxa. A *(phylogenetic) network* $\mathcal{N}$ *(on* X*)* is a rooted DAG without degenerate vertices whose set of leaves is X. Unless the phylogenetic network $\mathcal{N}$ in question has precisely two vertices, we always assume that the outdegree of the root of $\mathcal{N}$ is at least two. Note that a network $\mathcal{N}$ that does not contain vertices with indegree two or more is just an evolutionary or *phylogenetic tree (on* X*)*. As usual, we call a phylogenetic tree in which every leaf is the child of the root a *star tree.*

Now, suppose that Y is a non-empty subset of the set X of species. We now consider the subnet of $\mathcal{N}$ induced by restricting our attention to the leaves in Y. The *lowest stable ancestor* $\text{LSA}(Y)$ *of* Y *in* $\mathcal{N}$ is the vertex $w \in V(\mathcal{N}) - X$ that lies on *all* directed paths

[1] Although this means that in mathematical terms every vertex is considered to be an ancestor of itself, we adopt this mathematical convention as it simplifies the mathematics and is a common assumption in the theory of directed graphs.

from the root $\rho(\mathcal{N})$ of $\mathcal{N}$ to the elements in Y, so that no vertex of $\mathcal{N}$ below w enjoys this property. In case $\textsc{lsa}(X) = \rho(\mathcal{N})$, we call $\mathcal{N}$ *recoverable.* The *subnet* $\mathcal{N}|_Y$ of $\mathcal{N}$ induced by Y is defined as the phylogenetic network on Y obtained from $\mathcal{N}$ as follows: First, delete all vertices of $\mathcal{N}$ (and their incident arcs) that are not on a directed path from $\textsc{lsa}(Y)$ to some element in Y. Next, repeatedly *suppress* all resulting degenerate vertices (i.e. replace any such vertex v and the two arcs (u, v) and (v, w) containing it by a single arc (u, w)) and remove all parallel arcs until a phylogenetic network on Y is obtained. This definition for a subnet was introduced by Huber and Moulton (2012), and it aims to capture features that can be recovered from data (e.g. all degenerate vertices are suppressed as it would not be possible to decide how many degenerate vertices to include in a reconstructed network). Note that $\mathcal{N}|_X = \mathcal{N}$ if and only if $\mathcal{N}$ is recoverable. Also note that every subnet of $\mathcal{N}$ induced by restricting $\mathcal{N}$ to some non-empty subset of its leaves is necessarily recoverable.

We say that two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ on X are *network-equivalent* if for every non-empty, proper subset Y of X, the phylogenetic networks $\mathcal{N}|_Y$ and $\mathcal{N}'|_Y$ are equivalent. Thus two phylogenetic networks are equivalent if and only if they represent the same evolutionary histories. Note that the following useful observation concerning network-equivalence is an immediate consequence of our definitions.

Lemma 1. *Suppose that $\mathcal{N}_1$ and $\mathcal{N}_2$ are two recoverable phylogenetic networks on X. If $\mathcal{N}_1$ and $\mathcal{N}_2$ are equivalent, then $\mathcal{N}_1$ and $\mathcal{N}_2$ are network-equivalent.*

Binary sequences

All of our networks will be constructed using special types of binary sequences, that is, sequences over the alphabet $\{0, 1\}$. We use binary sequences since they provide a convenient way to encode the vertices of certain phylogenetic trees that will be relevant to our constructions.

As our examples rely on using some special types of binary sequences we now introduce some general terminology concerning such sequences. Suppose that n is a non-negative integer. We denote by $l(w)$ the *length* of a binary sequence w. We let $\emptyset$ denote the *empty sequence*, that is, the unique sequence with length 0, let $w_{k,n}$ be the binary sequence of length n with 0's in all but the k-th place, $1 \leq k \leq n$, and let $\mathbf{0_n}$, $\mathbf{1_n}$ be the binary sequences of length n consisting of all 0's and all 1's, respectively. We also let $\mathcal{B}_n$ denote the set of all binary sequences that have length n. Note that $\mathcal{B}_0 = \{\emptyset\}$.

Now, assume $n \geq 1$. For each sequence w in $\mathcal{B}_n$ and all $1 \leq i \leq n$, we denote by $[w]_i$ the i-th letter of w starting from the left. We define the *weight* of w as $\sum_{i=1}^{n} [w]_i$, that is, the number of 1's contained in w. Moreover, for each sequence $w \in \mathcal{B}_n$, we define the *support* $supp(w)$ of w to be the subset of $\{1, 2, \ldots, n\}$ consisting of all indices $i \in \{1, \ldots, n\}$ with $[w]_i = 1$. Finally, we denote by $\mathcal{B}_n^1$ and $\mathcal{B}_n^2$ the subsets of $\mathcal{B}_n$ consisting of sequences whose weights are odd and even, respectively. Note that we will assume that $\mathbf{0_n}$ is the only sequence in $\mathcal{B}_n$ that is contained in neither $\mathcal{B}_n^1$ nor $\mathcal{B}_n^2$. Thus, $|\mathcal{B}_n^1| = 2^{n-1}$ while $|\mathcal{B}_n^2| = 2^{n-1} - 1$. As an illustration of these definitions, $w_{2,3} = 010$, the weight of the sequence 011 is 2 and its support is $\{2, 3\}$, and $\mathcal{B}_3^1 = \{001, 010, 100, 111\}$, $\mathcal{B}_3^2 = \{110, 101, 011\}$.

Now, suppose that w and w' are two binary sequences. Then w' is called a *prefix of* w if $l(w') \leq l(w)$ and $[w']_i = [w]_i$ holds for all $1 \leq i \leq l(w')$. Note that the empty sequence is a prefix of every binary sequence. Also, if $\mathcal{B}$ is a set of binary sequences and $w \in \mathcal{B}$, then we call a sequence $w' \in \mathcal{B} - \{w\}$ a *precursor of* w *(in* $\mathcal{B}$*)* if w' is a prefix of w. And if, in addition, every precursor of w in $\mathcal{B}$ other than w' is also a precursor of w' in $\mathcal{B}$, then we say that w' is the *maximal precursor of* w *(in* $\mathcal{B}$*).* Note that if this exists, then it is unique. Finally, we call w a *common precursor of* $\mathcal{B}$ if, for every sequence $w' \in \mathcal{B} - \{w\}$, w is a precursor of w' in $\mathcal{B}$.

Main Results

Non-binary network examples

In this section we shall present two non-binary, phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ on an arbitrary set X with at least three elements that are not equivalent and prove that they are network-equivalent.

We begin by defining two rooted DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ from which we will obtain $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively. Let $n \geq 3$, let $X = \{x_1, \dots, x_n\}$, and let $Y = \{y_1, \dots, y_n\}$ be a set such that $X \cap Y = \emptyset$. For $i = 1, 2$, associate to X and Y the rooted DAG $\mathcal{D}_i$ with vertex set $X \cup Y \cup \mathcal{B}_n^i \cup \{\rho_i\}$, and arc set comprising of (i) for all $u \in \mathcal{B}_n^i$ the arcs (ρ_i, u), (ii) for all $1 \leq j \leq n$ the arcs (y_j, x_j), and (iii) for all $1 \leq j \leq n$ and $u \in \mathcal{B}_n^i$ the arcs (u, y_j) if and only if $[u]_j = 1$. Note that X is the set of leaves of $\mathcal{D}_i$ and ρ_i is the root. We illustrate these DAGs for $n = 4$ in Figure 3 and list the binary sequences that label the vertices in $\mathcal{B}_n^1$ and $\mathcal{B}_n^2$ in its caption. To obtain the phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ we just suppress all degenerate vertices of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Note that it is straight-forward to check that both $\mathcal{N}_1$ and $\mathcal{N}_2$ are recoverable.

We now prove the first of our main results.

Theorem 2. *For every $n \geq 3$, the networks $\mathcal{N}_1$ and $\mathcal{N}_2$ are not equivalent. However, $\mathcal{N}_1$ and $\mathcal{N}_2$ are network-equivalent.*

Proof. To see that $\mathcal{N}_1$ and $\mathcal{N}_2$ are not equivalent note that $\mathcal{B}_n^1 \cap \mathcal{B}_n^2 = \emptyset$ and that sequence $\mathbf{1_n}$ is contained in $\mathcal{B}_n^1 \cup \mathcal{B}_n^2$. Consequently, there exists a child of the root of $\mathcal{N}_1$ or $\mathcal{N}_2$ (but not both) that has outdegree n. Thus, $\mathcal{N}_1$ and $\mathcal{N}_2$ cannot be equivalent.

We next show that $\mathcal{N}_1$ and $\mathcal{N}_2$ are network-equivalent. Let $k \in \{1, \dots, n\}$ and put $X^k = X - \{x_k\}$, $Y^k = Y - \{y_k\}$. Note that $\mathcal{N}_1|_{X^k}$ and $\mathcal{N}_2|_{X^k}$ are recoverable as they are subnets of $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively. In view of Lemma 1, it therefore suffices to show that $\mathcal{N}_1|_{X^k}$ and $\mathcal{N}_2|_{X^k}$ are equivalent.

To this end, for any $k \in \{1, \dots, n\}$, we associate for $i = 1, 2$ a rooted DAG $\mathcal{D}_i^k$ to $\mathcal{D}_i$ with the leaf x_k removed. In particular, we define $\mathcal{D}_i^k$ to be the rooted DAG with leaf set X^k obtained from $\mathcal{D}_i$ by first deleting all arcs from $\mathcal{D}_i$ that do not lie on a path from the root ρ_i of $\mathcal{D}_i$ to a leaf in X^k and then removing all resulting isolated vertices. Note that $X^k \cup Y^k \cup \{\rho_i\} \subseteq V(\mathcal{D}_i^k) \subseteq X^k \cup Y^k \cup \{\rho_i\} \cup \mathcal{B}_n^1 - \{w_{k,n}\}$.

For brevity, for the rest of this proof, we let V_i^k and E_i^k denote the vertex set and edge set of D_i^k, respectively, and we put $w_k = w_{k,n}$. We define a map χ_k from V_1^k to V_2^k as follows. Let $\varphi_k = \varphi_{k,n}$ denote the map from $\mathcal{B}_n$ to $\mathcal{B}_n$ that "flips" precisely the k-th letter of a sequence in $\mathcal{B}_n$, i. e. the map given by, for all $w \in \mathcal{B}_n$, putting $[\varphi_k(w)]_j = [w]_j$ for $j \in \{1, \dots, n\} - \{k\}$, and $[\varphi_k(w)]_j = 1 - [w]_j$ for $j = k$. Note that $w_k \in \mathcal{B}_n^1$ and $supp(\varphi_k(w_k)) = \emptyset$. Moreover, the map φ_k induces a bijection $\overline{\varphi}_k$ from $\mathcal{B}_n^1 - \{w_k\}$ to $\mathcal{B}_n^2$. Using this bijection, we now define the map χ_k by putting, for $v \in V_1^k$,

$$\chi_k(v) = \begin{cases} v & \text{if } v \in X^k \cup Y^k, \\ \rho_2 & \text{if } v = \rho_1, \\ \overline{\varphi}_k(v) & \text{else.} \end{cases}$$

We shall show that this map is a bijection from V_1^k to V_2^k that extends to an isomorphism from $\mathcal{D}_1^k$ to $\mathcal{D}_2^k$ which maps every element in X^k to itself. This implies that $\mathcal{N}_1|_{X^k}$ and $\mathcal{N}_2|_{X^k}$ are equivalent.

Clearly, χ_k maps every element in X^k to itself. To see that χ_k is a bijection, note that since w_k is the only element in $\mathcal{B}_n^1$ that is contained in $V(\mathcal{D}_1)$ but not V_1^k, we have $\mathcal{B}_n^1 - \{w_k\} \subseteq V_1^k$. Combined with the fact that $\mathcal{B}_n^2 \subseteq V_2^k$ also holds and that $\overline{\varphi}_k$ is a bijection, it follows that χ_k is a bijection.

To see that χ_k induces an isomorphism between $\mathcal{D}_1^k$ and $\mathcal{D}_2^k$ it suffices to show for all

$v, w \in V_1^k$, that $(v, w) \in E_1^k$ if and only if $(\chi_k(v), \chi_k(w)) \in E_2^k$. In view of $\chi_k(u) = u$ holding for all $u \in V_1^k - \mathcal{B}_n^1$ and $(\rho_i, w) \in E(\mathcal{D}_i)$ holding for all $w \in \mathcal{B}_n^i$, it follows that we may restrict our attention to showing that for all $j \in \{1, \dots, n\} - \{k\}$ and all $w \in \mathcal{B}_n^1 - \{w_k\}$ we have that $(w, y_j) \in E_1^k$ if and only if $(\chi_k(w), \chi_k(y_j)) \in E_2^k$. So let $j \in \{1, \dots, n\} - \{k\}$ and $w \in \mathcal{B}_n^1 - \{w_k\}$. Assume first that $(w, y_j) \in E_1^k$. Then $[w]_j = 1$ and so $[\chi_k(w)]_j = [\overline{\varphi}_k(w)]_j = 1$ as $k \neq j$. Thus, $(\chi_k(w), \chi_k(y_j)) = (\chi_k(w), y_j) \in E_2^k$ as $\chi_k(y_j) = y_j$. Conversely, assume that $(\chi_k(w), \chi_k(y_j)) \in E_2^k$. Then since $\chi_k(y_j) = y_j$ we have $[\overline{\varphi}_k(w)]_j = [\chi_k(w)]_j = 1$, and, hence $[w]_j = 1$ in view of $j \neq k$. Thus, $(w, y_j) \in E_1^k$, as required. □

Binary network examples

We now extend the definitions of the networks $\mathcal{D}_1$ and $\mathcal{D}_2$ defined in the previous section so as to define two binary phylogenetic networks $\mathcal{H}_1$ and $\mathcal{H}_2$ that are not equivalent, but which are network-equivalent. We shall just present the definitions of $\mathcal{H}_1$ and $\mathcal{H}_2$; the proof of their network-equivalence is quite technical and can be found in the appendix.

Let $n \geq 3$ and $i \in \{1, 2\}$. Starting with the rooted DAGs $\mathcal{D}_i$ defined in the previous section, we shall define a sequence of three rooted DAGs all having leaf set X, the last one of which will yield $\mathcal{H}_i$. We illustrate this process in Figure 4, for the rooted DAG $\mathcal{D}_1$ depicted in Figure 3.

Step 1: We begin by replacing the star tree containing the root vertex of $\mathcal{D}_i$ by a tree with leaf set $\mathcal{B}_n^i$ that is a subtree of a certain tree $\mathcal{P}_n$ which is defined as follows. Let $\mathcal{A}_n$ be the set of all binary sequences with length at most n. The tree $\mathcal{P}_n$ is the rooted tree with vertex set $\mathcal{A}_n$ and arc set consisting of all pairs $(w, w') \in \mathcal{A}_n \times \mathcal{A}_n$ for which w is the maximal precursor of w' in $\mathcal{A}_n$. Note that the common precursor of $\mathcal{A}_n$ is clearly the root of $\mathcal{P}_n$. In addition, since each sequence $w \in \mathcal{A}_n$ is the maximal precursor of exactly two sequences in $\mathcal{A}_n$ if $w \notin \mathcal{B}_n$, and is not the maximal precursor of any sequence in $\mathcal{A}_n$ if $w \in \mathcal{B}_n$, it follows that $\mathcal{P}_n$ is a binary phylogenetic tree on $\mathcal{B}_n$. We depict the tree $\mathcal{P}_3$ in Figure 5(i).

Now, we replace the subgraph of $\mathcal{D}_i$ with vertex set consisting of the root ρ_i of $\mathcal{D}_i$ and the children of ρ_i (i .e. the star tree on $\mathcal{B}_n^i$ with root ρ_i) by the (necessarily binary) restriction $\mathcal{P}_n|_{\mathcal{B}_n^i}$ of $\mathcal{P}_n$ to $\mathcal{B}_n^i$. Let $\mathcal{D}_{i,1}$ denote the resulting rooted DAG (see e.g. Fig. 4(i)).

Step 2: We now replace each of the vertices y_j, $1 \leq j \leq n$, in the "bottom layer" of $\mathcal{D}_{i,1}$ by a tree $\mathcal{R}_j$. This tree is defined by reversing the direction of all arcs in the tree obtained by restricting the tree $\mathcal{P}_n$ to the set $\mathcal{B}_{n,j}^i$ of all binary sequences in $\mathcal{B}_n^i$ whose j-th letter is 1. Note that the unique leaf of $\mathcal{R}_j$ is y_j since for all $1 \leq j \leq n$ the source set of $\mathcal{R}_j$ is $\mathcal{B}_{n,j}^i$ and $\bigcup_{j=1}^n \mathcal{B}_{n,j}^i = \mathcal{B}_n^i$ which is the leaf set of $\mathcal{P}_n|_{\mathcal{B}_n^i}$.

Now, note that, for all $1 \leq j \leq n$, the indegree of y_j in $\mathcal{D}_i$ is $|\mathcal{B}_{n,j}^i| = 2^{n-2}$ and that therefore the indegree of y_j in $\mathcal{D}_{i,1}$ is also 2^{n-2}. We now replace for all $1 \leq j \leq n$ the subgraph of $\mathcal{D}_{i,1}$ induced on the set $\{y_j\} \cup \mathcal{B}_{n,j}^i$ by $\mathcal{R}_j$. Let $\mathcal{D}_{i,2}$ denote the resulting rooted DAG (see e.g. Fig. 4(ii)).

Step 3: The final stage of the construction involves replacing each of the vertices in the "middle layer" of $\mathcal{D}_{i,2}$ with another phylogenetic tree which is defined as follows.

Let $A = \{a_1, \dots, a_k\}$, $k \geq 1$, denote a set of positive integers with $a_1 < \cdots < a_k$. If $k = 1$, then we denote by $\mathcal{C}_A$ the phylogenetic tree whose unique leaf is labeled by the sole element in A. More generally, for $k \geq 2$ we denote by $\mathcal{C}_A$ the (up to equivalence) unique binary phylogenetic tree on A such that, over all non-leaf vertices v of $\mathcal{C}_A$, the collection of leaves below v is $\bigcup_{1 \leq j < k} \{\{a_j, a_{j+1}, \dots, a_k\}\}$. Note that $\mathcal{C}_A$ is an example of a rooted caterpillar tree (see e.g. Semple and Steel 2003). In Figure 5(ii) we present the tree $\mathcal{C}_A$ for $A = \{1, 2, 3, 5, 7\}$.

Now, any non-degenerate vertex w of $\mathcal{D}_{i,2}$ is not binary if and only if $w \in \mathcal{B}_n^i$ and $|\,supp(w)| > 2$. Therefore we shall consider vertices in $\mathcal{B}_n^i$ whose support has size at least three. We shall replace all such vertices w by a rooted tree that is derived from the tree $\mathcal{C}_{supp(w)}$ as follows (essentially, we replace w and its outgoing arcs by a rooted caterpillar whose leaves are the children of w). Put $Y_w := \{y_t \in Y \, : \, t \in supp(w)\}$. Then, since for all $1 \leq j < l \leq n$ the trees $\mathcal{R}_j$ and $\mathcal{R}_l$ defined in Step 2 do not share an interior vertex in $\mathcal{D}_{i,2}$, it follows that for every child w' of w there exists a unique vertex $y \in Y_w$ below w'. For all $t \in supp(w)$ let a_t denote the child of w in $\mathcal{D}_{i,2}$ that lies on the path from w to y_t so that, in particular, the set of children of w is $\{a_t \, : \, t \in supp(w)\}$. To obtain the final digraph $\mathcal{D}_{i,3}$ in our sequence, we replace, for each $w \in \mathcal{B}_n^i$ with $|\, supp(w)| > 2$, the subgraph of $\mathcal{D}_{i,2}$ induced on the set consisting of w and its children by the tree obtained from $\mathcal{C}_{supp(w)}$ by replacing each of its leaves $t \in supp(w)$ by the corresponding child a_t of w and replacing its root by w (see e.g. Fig. 4(iii)).

The phylogenetic network $\mathcal{H}_i$ is now defined to be the rooted DAG obtained from $\mathcal{D}_{i,3}$ by suppressing all degenerate vertices (see e.g. Fig. 4(iv)). Note that, by construction, the leaf set of $\mathcal{H}_i$ is X and $\mathcal{H}_i$ is binary. Also note that $\mathcal{H}_i$ is recoverable.

The proof of our second main result is quite technical and is given in the appendix.

Theorem 3. *For every $n \geq 3$, the binary phylogenetic networks $\mathcal{H}_1$ and $\mathcal{H}_2$ are not equivalent. However, $\mathcal{H}_1$ and $\mathcal{H}_2$ are network-equivalent.*

Discussion

Our examples illustrate a problem with generalising evolutionary models from rooted trees to rooted networks. We show that there are pairs of phylogenetic networks on an arbitrary set of taxa that are not equivalent, and yet display the same set of evolutionary trees (see Supplementary Material for additional details), as well as the same set of induced subnetworks. Although these examples are artificial in their construction, they still point to the possibility that this phenomenon could arise in nature, especially since phylogenetic networks can be extremely complex (see e.g. Dagan et al. 2008; Kunin et al. 2005).

The problem that we have presented has some potential ramifications to the development and use of new methods for constructing networks that explicitly represent evolution. First, as mentioned in the introduction, it implies that in practice we will have to be careful to ensure that the output from any network construction method is uniquely determined by its input. This in itself is not necessarily a great problem since even when we construct phylogenetic trees there can be multiple solutions (e.g. there can be several most parsimonious trees). Second, given that we know that there are cases where a network cannot be uniquely recovered from all of its induced subnetworks, it becomes important to characterize under which conditions these cases will be manifested, and we should try to understand how often biological data will actually meet these conditions. Finally, in the context of extending consensus and supertree methods to include phylogenetic networks, our result shows that, unlike trees, it will not be possible to develop supernetwork methods in general that are consistent, i.e. methods that are guaranteed to output a given network from all of its induced subnetworks. However, again it will be interesting to better understand how important this will actually be in practice.

Even though we have found that networks are not necessarily encoded by their induced subnets in general, some classes of networks are. For example, level-2 networks and thus also phylogenetic trees and level-1 networks are encoded by their trinets (note that the level of a binary phylogenetic network is the maximum number of indegree-2 vertices taken over all biconnected components of the network) (van Iersel and Moulton 2014). Hence it might be of interest to determine which types of networks are encoded by their induced subnets and also to possibly concentrate on developing methods to construct these special types of networks. Note that various methods have already been designed to construct special types of networks (see e.g. Willson 2012), but this obviously requires some care to ensure that the properties of the networks under consideration are realistic enough to represent real data. Note also that the level of the networks in our examples is exponential in $|X|$ (it is $(2^{n-2}-1)n$ with $n=|X|$). Hence it could be of interest to decide whether networks with reasonably low level relative to the size of their leaf set (e.g. linear level as function of $|X|$) are encoded by their subnets.

Even if we are not necessarily able to encode a network by its subnets, it could still be of interest to investigate whether at least some parameters (e.g. the number of reticulation vertices) can be determined or at least approximated by the knowledge of their induced subnets (or even trees). In addition, it could be useful to decide whether or not networks might be encoded if more information is available (e.g. if we are given branch lengths/dates for vertices or some model of evolution). Note that Thatte and Steel investigated reconstructability of pedigrees assuming a certain probabilistic model and were able to prove some encoding results for pedigrees in general (see e.g. Thatte and Steel 2008; Thatte 2013), so analogous results might also hold for phylogenetic networks.

There are some related mathematical problems that are also worth mentioning. It has been shown that a graph drawn uniformly at random is encoded by its subgraphs with probability 1, as the size of the vertex set goes to infinity (see e.g. Bollobás 1990). It would be interesting to work out the probability that a randomly selected phylogenetic network is encoded by its induced subnets. This might also provide some clues about whether or not networks arising in practice could be expected to be encoded by induced subnets or not. In particular, the aforementioned probabilistic result suggests that maybe networks on large sets of taxa that are not encoded by their induced subnetworks might be quite rare in practice. In addition, an interesting algorithmic question is the following: if we are given a phylogenetic network, can we decide efficiently if it is uniquely encoded by its induced subnets? And, if we are given a set of networks, can we efficiently decide if they are induced subnets of some network?

In conclusion, even if there may be more than one network that can induce the same set of trees and/or subnetworks, it is still useful to find ways to construct these networks so that alternative evolutionary scenarios can be explored. This has already proven a useful strategy in phylogenetics (for example, understanding the number of reconciliations of a gene tree with a species tree (Bansal et al. 2013)). In regards to this, it would be interesting to develop ways to determine how many networks can potentially display the same set of subnetworks. More generally, a better understanding of the structure of networks in terms of substructures could also give us a better understanding of the performance of current methods for network construction, and will hopefully also eventually help us to design new methods for confidently recovering reticulate evolutionary histories.

Appendix

In this appendix we prove Theorem 3. Assume that n is a positive integer and that $k \in \{1, \ldots, n\}$. We will use the following lemmas concerning the trees $\mathcal{P}_n$ and $\mathcal{C}_A$ used in the construction of $\mathcal{H}_1$ and $\mathcal{H}_2$. For brevity, for any sequence $w \in \mathcal{B}_n$ with non-empty support (with the natural order), we put $\mathcal{C}_w = \mathcal{C}_{supp(w)}$. Also and as in the proof of Theorem 2, we let $\varphi_k : \mathcal{B}_n \to \mathcal{B}_n$ be the map that "flips" precisely the k-th letter of a sequence in $\mathcal{B}_n$.

Lemma 4. *For each sequence $w \in \mathcal{B}_n - \{\mathbf{0}_n, w_{k,n}\}$, the trees $\mathcal{C}_w|_{supp(w)-\{k\}}$ and $\mathcal{C}_{\varphi_k(w)}|_{supp(\varphi_k(w))-\{k\}}$ are isomorphic.*

Proof. Put $w_k = w_{k,n}$. Let $w \in \mathcal{B}_n - \{\mathbf{0}_n, w_k\}$. Then neither $supp(w) = \emptyset$ nor $supp(w) = \{k\}$ holds. This implies that the trees $\mathcal{C}_w$ and $\mathcal{C}_{\varphi_k(w)}$ are both well-defined. Let w' denote the unique sequence in $\mathcal{B}_n$ whose support is $supp(w) - \{k\}$, which is clearly not the empty set. Then the tree $\mathcal{C}_{w'}$ is also well-defined and, as is straightforward to see, it is isomorphic to $\mathcal{C}_w|_{supp(w')}$. In view of $\emptyset \neq supp(w') \subseteq supp(\varphi_k(w))$, we also have that $\mathcal{C}_{w'}$ is isomorphic to $\mathcal{C}_{\varphi_k(w)}|_{supp(w')}$. Since $supp(w) - \{k\} = supp(w') = supp(\varphi_k(w)) - \{k\}$, it follows that $\mathcal{C}_w|_{supp(w)-\{k\}}$ must be isomorphic to $\mathcal{C}_{\varphi_k(w)}|_{supp(\varphi_k(w))-\{k\}}$. □

We now establish a similar result for the tree $\mathcal{P}_n$. For $\mathcal{A}_n$ as defined in Step 1 of the construction, a set $\mathcal{A} \subseteq \mathcal{A}_n$ of binary sequences is called *complete* if it contains a (necessarily unique) common precursor. Given such a set, generalizing the definition of $\mathcal{P}_n$, we let $\mathcal{P}[\mathcal{A}]$ denote the directed tree with vertex set $\mathcal{A}$ and arc set the set of all arcs $(w, w') \in \mathcal{A} \times \mathcal{A}$ for which w is the maximal precursor of w' in $\mathcal{A}$. Note that $\mathcal{P}_n = \mathcal{P}[\mathcal{A}_n]$.

Lemma 5. *(i) The trees $\mathcal{P}_n|_{\mathcal{B}^1_n - \{w_{n,k}\}}$ and $\mathcal{P}_n|_{\mathcal{B}^2_n}$ are isomorphic.*

(ii) For all $j \in \{1, \ldots, n\} - \{k\}$, the trees $\mathcal{P}_n|_{\mathcal{B}^1_{n,j}}$ and $\mathcal{P}_n|_{\mathcal{B}^2_{n,j}}$ are isomorphic.

Proof. Let $\mathcal{A}^*$ be the set of all binary sequences of finite length. For all $l \geq 1$, define the map $\psi_l : \mathcal{A}^* \to \mathcal{A}^*$ by

$$\psi_l : \mathcal{A}^* \to \mathcal{A}^* \; : \; w \mapsto \begin{cases} \varphi_{l,n}(w) & \text{if } w \in \mathcal{B}_n \text{ for some } l \leq n, \\ w & \text{else.} \end{cases}$$

Now, note that if a subset $\mathcal{A} \subseteq \mathcal{A}_n$ is complete, then the set $\psi_l(\mathcal{A})$ is complete since, for any two distinct sequences $w, w' \in \mathcal{A}$, we have that w is a precursor of w' in $\mathcal{A}$ if and only if $\psi_l(w)$ is a precursor of $\psi_l(w')$ in $\psi_l(\mathcal{A})$. Moreover, for $l = k$ the map ψ_k induces an isomorphism between the trees $\mathcal{P}[\mathcal{A}]$ and $\mathcal{P}[\psi_k(\mathcal{A})]$. In particular, ψ_k induces, for every subset $\mathcal{B} \subseteq \mathcal{B}_n$, an isomorphism between $\mathcal{P}_n|_{\mathcal{B}}$ and $\mathcal{P}_n|_{\psi_k(\mathcal{B})}$.

Both statements in the lemma now follow in view of the fact that $\psi_l(\mathcal{B}^1_n - \{w_{k,n}\}) = \varphi_l(\mathcal{B}^1_n - \{w_{k,n}\}) = \mathcal{B}^2_n$ holds for all $1 \leq l \leq n$ and $\psi_l(\mathcal{B}^1_{n,j}) = \mathcal{B}^2_{n,j}$ holds for all $1 \leq j, l \leq n$ with $j \neq l$. □

We now prove Theorem 3. We first show that $\mathcal{H}_1$ and $\mathcal{H}_2$ are not equivalent. Indeed, let $i \in \{1, 2\}$. If v is a vertex in $\mathcal{D}_i$ then the set of leaves below v equals X if and only if v is the root of $\mathcal{D}_i$ or $v = \mathbf{1_n}$. Since if m is even $\mathbf{1_m} \in \mathcal{B}^2_m$ and if m is odd $\mathbf{1_m} \in \mathcal{B}^1_m$, the number of vertices v in $\mathcal{D}_i$ for which the set of leaves below v equals X is different in $\mathcal{D}_1$ and $\mathcal{D}_2$. Thus, $\mathcal{H}_1$ and $\mathcal{H}_2$ cannot be equivalent.

We now prove that $\mathcal{H}_1$ and $\mathcal{H}_2$ are network-equivalent. Let $X^k = X - \{x_k\}$ and $Y^k = Y - \{y_k\}$. Note that the subnets $\mathcal{H}^k_1 = \mathcal{H}_1|_{X^k}$ and $\mathcal{H}^k_2 = \mathcal{H}_2|_{X^k}$ are recoverable as they are subnets of $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Hence, in view of Lemma 1, it suffices to show that $\mathcal{H}^k_1$ and $\mathcal{H}^k_2$ are equivalent.

Now, put $\mathcal{D}_{i,0} = \mathcal{D}_i$. For $0 \leq j \leq 3$, let $\mathcal{D}^k_{i,j} = (V^k_{i,j}, E^k_{i,j})$ denote the rooted DAG with leaf set X^k obtained from $\mathcal{D}_{i,j}$ by first deleting all arcs from $\mathcal{D}_{i,j}$ that do not lie on a path from the root of $\mathcal{D}_{i,j}$ to a leaf in X^k and then removing the resulting isolated vertices. Note that $\mathcal{D}^k_{i,0} = \mathcal{D}^k_i$ and that $\mathcal{H}^k_i$ is the phylogenetic network on X^k obtained from $\mathcal{D}^k_i$ by suppressing all degenerate vertices.

By Lemma 5(i), there exists a bijection $\chi^k_1 : V^k_{1,1} \to V^k_{2,1}$ that extends to an isomorphism from $\mathcal{D}^k_{1,1}$ to $\mathcal{D}^k_{2,1}$ such that, for all $v \in V^k_{1,1}$, we have $\chi^k_1(v) = \psi_k(v)$ if $v \in V^k_{1,1} - V^k_{1,0}$ and $\chi^k_1(v) = \chi_k(v)$ else where χ_k is the bijection from V^k_1 to V^k_2 given in the proof of Theorem 2. Using Lemma 5(ii), there also exists a bijection $\chi^k_2 : V^k_{1,2} \to V^k_{2,2}$ that extends to an isomorphism from $\mathcal{D}^k_{1,2}$ to $\mathcal{D}^k_{2,2}$ for which $\chi^k_2(v) = \chi^k_1(v)$ holds for all $v \in V^k_{1,1}$. Moreover, by Lemma 4, there also exists a bijection $\chi^k_3 : V^k_{1,3} \to V^k_{2,3}$ that extends to an isomorphism from $\mathcal{D}^k_{1,3}$ to $\mathcal{D}^k_{2,3}$ for which $\chi^k_3(v) = \chi^k_2(v)$ holds for all $v \in V^k_{1,2}$. Consequently, $\mathcal{H}^k_1$ and $\mathcal{H}^k_2$ must be equivalent, as required.

Supplementary material

An online-only appendix can be found in the Dryad data repository (**doi-???**).

Funding

LvI was supported by a Veni grant from The Netherlands Organisation for Scientific Research (NWO).

Acknowledgments

We would like to thank Bhalchandra Thatte for his helpful explanations regarding his pedigree reconstruction papers. We would also like to thank Peter Lockhart and two anonymous referees, as well as two editors, Frank E. Anderson and Tanja Stadler, for their helpful and constructive comments on the first version of this paper.

*

References

Abbott et al., 2013. Hybridization and speciation. Journal of Evolutionary Biology 26(2):229–246.

Bandelt, H.-J., P. Lahermo, M. Richards, and V. Macaulay. 2001. Detecting errors in mtdna data by phylogenetic analysis. International Journal of Legal Medicine 115:64–69.

Bandelt, H.-J., V. Macaulay, and M. Richards. 2000. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Molecular Phylogenetics and Evolution 16(1):8–28.

Bansal, M., E. Alm, and M. Kellis. 2013. Reconciliation revisited: handling multiple optima when reconciling with duplication, transfer, and loss. Pages 1–13 *in* Research in Computational Molecular Biology, Springer.

Bapteste, E., L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. McInerney, D. Morrison, L. Nakhleh, M. Steel, L. Stougie, and J. Whitfield. 2013. Networks: expanding evolutionary thinking. Trends in Genetics 29(8):439–441.

Bollobás, B. 1990. Almost every graph has reconstruction number three. Journal of Graph Theory 14:1–4.

Dagan, T., Y. Artzy-Randrup, and W. Martin. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proceedings of the National Academy of Sciences 105(29):10039–10044.

Dress, A. and D. Huson. 2004. Constructing splits graphs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(3):109–115.

Gambette, P. and K. Huber. 2012. On encodings of phylogenetic networks of bounded level. Journal of Mathematical Biology 65(1):157–180.

Hagen, F., et al. 2013. Ancient dispersal of the human fungal pathogen *Cryptococcus gattii* from the Amazon rainforest. PLoS One 8(8):e71148.

Holland, B., F. Delsuc, and V. Moulton. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. Systematic Biology 54(1):66 – 76.

Huber, K. and V. Moulton. 2012. Encoding and constructing 1-nested phylogenetic networks with trinets. Algorithmica 616:714–738.

Huber, K., L. van Iersel, S. Kelk, and R. Suchecki. 2011. A practical algorithm for reconstructing level-1 phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(3):635–649.

Huson, D. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23(2):254–267.

Huson, D. H., R. Rupp, and C. Scornavacca. 2010. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press.

Huson, D. and C. Scornavacca. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61(6):1061–1067.

Kelk, S., L. van Iersel, N. Lekić, S. Linz, C. Scornavacca, and L. Stougie. 2012. Cycle killer... qu'est-ce que c'est? on the comparative approximability of hybridization number and directed feedback vertex set. SIAM Journal on Discrete Mathematics 26(4):1635–1656.

Kunin, V., L. Goldovsky, N. Darzentas, and C. Ouzounis. 2005. The net of life: Reconstructing the microbial phylogenetic network. Genome Research 15:954–959.

Liu, Z., J. Müller, T. Li, R. M. Alvey, K. Vogl1, N.-. Frigaard, N. C. Rockwell, E. S. Boyd, L. P. Tomsho, S. C. Schuster, P. Henke, M. Rohde, J. Overmann, and D. A. Bryant. 2013. Genomic analysis reveals key aspects of prokaryotic symbiosis in the phototrophic consortium "chlorochromatium aggregatum". Genome Biology 14:R127.

Mallet, J. 2007. Hybrid speciation. Nature 446:279 – 283.

Marcussen, T., K. Blaxland, M. Windham, K. Haskins, and F. Armstrong. 2011. Establishing the phylogenetic origin, history, and age of the narrow endemic *viola guadalupensis* (violaceae). American Journal of Botany 98:1978–1988.

Morrison, D. 2005. Networks in phylogenetic analysis: new tools for population biology. International Journal for Parasitolog 35:567–582.

Morrison, D. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. Molecular Biology and Evolution 27(5):1044–1057.

Morrison, D. 2011. An introduction to phylogenetic networks. RJR Productions.

Muhlfeld, C. C., R. P. Kovach, L. A. Jones, R. Al-Chokhachy, and M. C. Boyer. 2014. Invasive hybridization in a threatened species is accelerated by climate change. Nature Climate Change doi:10.1038/nclimate2252.

Nakhleh, L. 2011. Evolutionary phylogenetic networks: models and issues. Pages 125–158 *in* Problem Solving Handbook in Computational Biology and Bioinformatics. Springer.

Penny, D., M. Hendy, and M. Steel. 1992. Progress with methods for constructing evolutionary trees. Trends in Ecology and Evolution 7:73 – 79.

Semple, C. and M. Steel. 2003. Phylogenetics. Oxford University Press.

Stockmeyer, P. 1977. The falsity of the reconstruction conjecture for tournaments. Journal of Graph Theory 1(1):19–25.

Thatte, B. 2008. Combinatorics of pedigrees i: counterexamples to a reconstruction problem. SIAM Journal of Discrete Mathematics 22(3):961–970.

Thatte, B. 2013. Reconstructing pedigrees: some identifiability questions for a recombination-mutation model. Journal of Mathematical Biology 66(1-2):37–74.

Thatte, B. and M. Steel. 2008. Reconstructing pedigrees: A stochastic perspective. Journal of Theoretical Biology 251(3):440–449.

van Iersel, L., J. Keijsper, S. Kelk, L. Stougie, F. Hagen, and T. Boekhout. 2009. Constructing level-2 phylogenetic networks from triplets. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6(4):667–681.

van Iersel, L., S. Kelk, R. Rupp, and D. Huson. 2010. Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. Bioinformatics 26(12):i124–i131.

van Iersel, L. and V. Moulton. 2014. Trinets encode tree-child and level-2 phylogenetic networks. Journal of Mathematical Biology 68:1707–1729.

Visser, J., D. Bellstedt, and M. Pirie. 2012. The recent recombinant evolution of a major crop pathogen, potato virus Y. PLoS One 7(11):e50631.

Willson, S. 2011. Regular networks can be uniquely constructed from their trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(3):785–796.

Willson, S. 2012. Reconstruction of certain phylogenetic networks from their tree-average distances. Bulletin of Mathematical Biology 75(10):1840–1878.

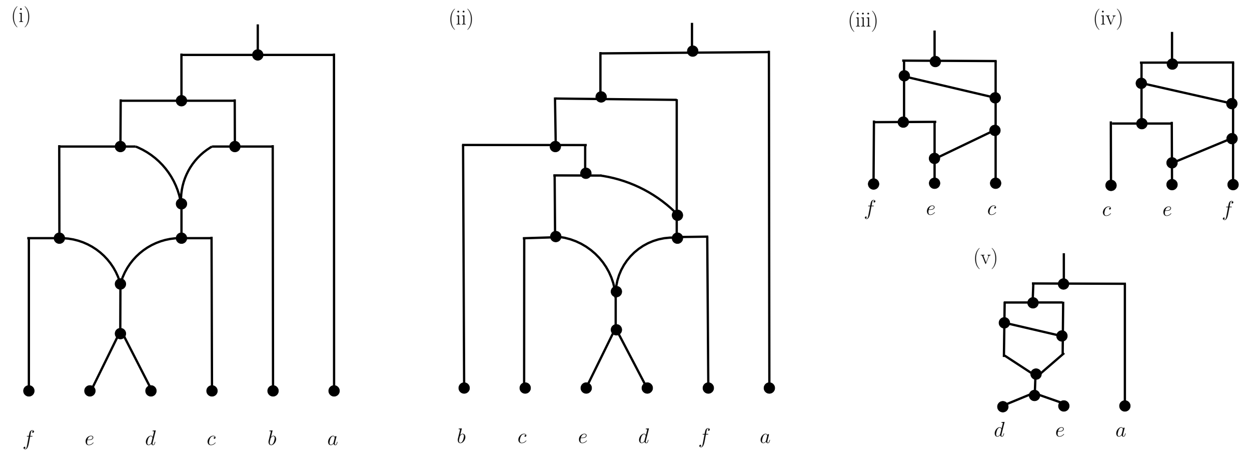

Figure 1: (i) and (ii): Two phylogenetic networks on the set of taxa $\{a, b, c, d, e, f\}$. (iii) and (iv): The trinets induced on the leaves c, e, f in networks in (i) and (ii), respectively. (v): The trinet induced on the leaves a, d, e by both of the networks in (i) and (ii). Here the network in (i) is the subnetwork of the network in (van Iersel et al. 2009, Fig. 10) computed from a dataset of the yeast *Cryptococcus gattii*, where taxa $a, b, \cdots, f$ correspond to taxa $1, 16, 8, 18, 7$ and 20, respectively, in the yeast network.

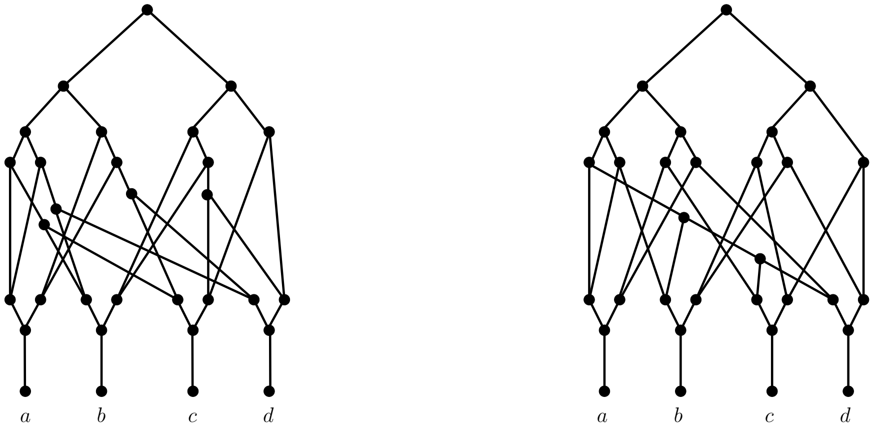

Figure 2: Two distinct rooted phylogenetic networks on the set of taxa $\{a, b, c, d\}$. The networks induce exactly the same set of trinets (pictured in the Supplementary Material) and also the same set of trees.

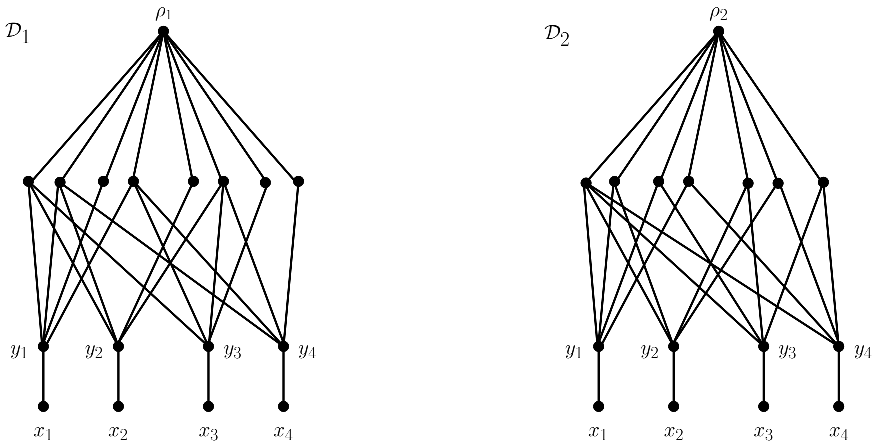

Figure 3: The rooted DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ for the case $n = 4$ and $X = \{x_1, x_2, x_3, x_4\}$. The labels of the vertices in $\mathcal{B}_4^i$, $i = 1, 2$, directly below the root in both DAGs are omitted; listed from left to right they are $1110, 1101, 1000, 1011, 0100, 0111, 0010, 0001$ for $\mathcal{D}_1$, and $1111, 1100, 1010, 1001, 0110, 0101, 0011$ for $\mathcal{D}_2$.

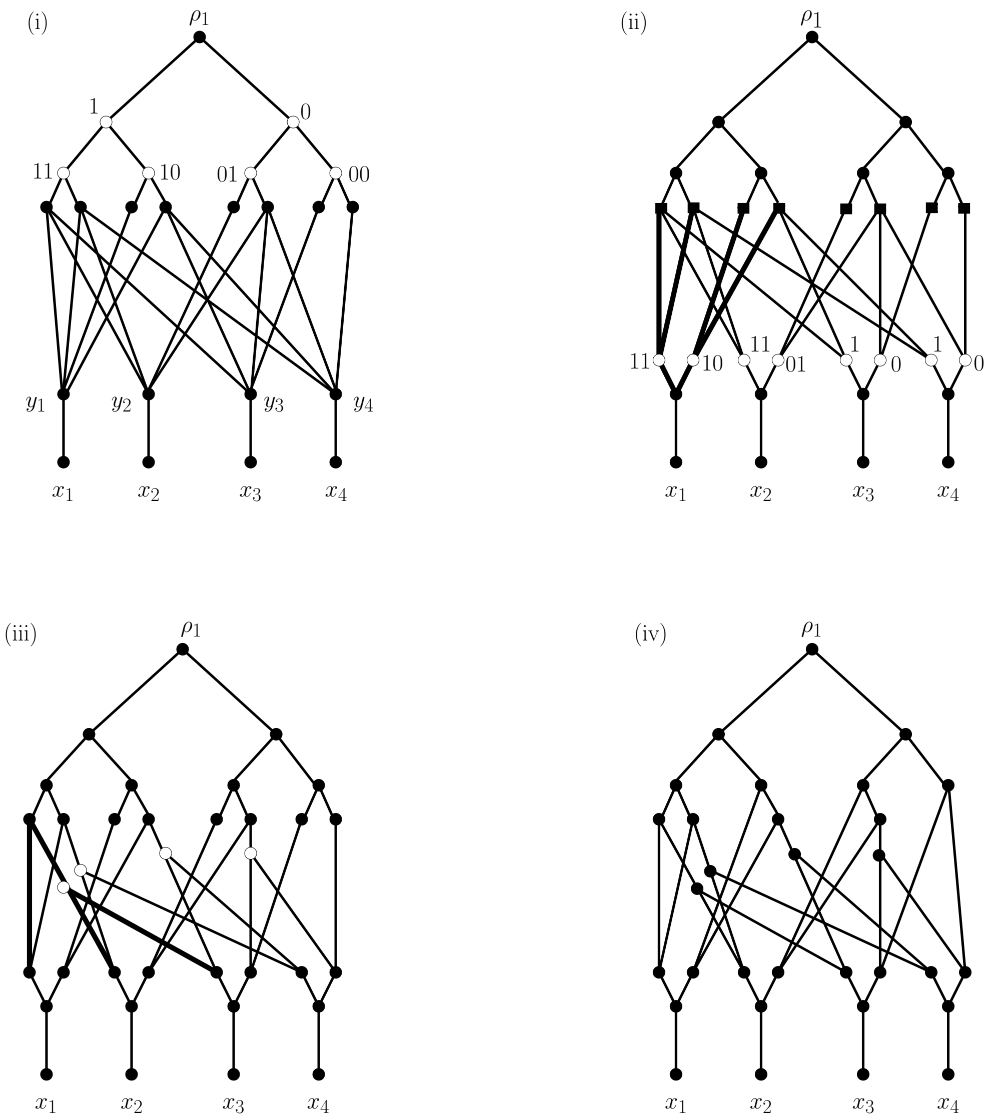

Figure 4: Constructing the network $\mathcal{H}_1$ from the network $\mathcal{D}_1$ in the case $|X| = 4$. At each stage we indicate those vertices that have been inserted by unfilled circles. (i) The network $\mathcal{D}_{1,1}$ in which the root has been replaced by the tree $\mathcal{P}_4$. (ii) The network $\mathcal{D}_{1,2}$ with the vertices in the middle layer indicated by squares. The tree $\mathcal{R}_1$ is indicated in bold. (iii) The network $\mathcal{D}_{1,3}$. The tree $\mathcal{C}_{supp(w)}$ associated to the binary sequence $w = 1110$ is indicated in bold. (iv) The network $\mathcal{H}_1$ obtained by suppressing all vertices in $\mathcal{D}_{1,3}$ having indegree and outdegree equal to one.

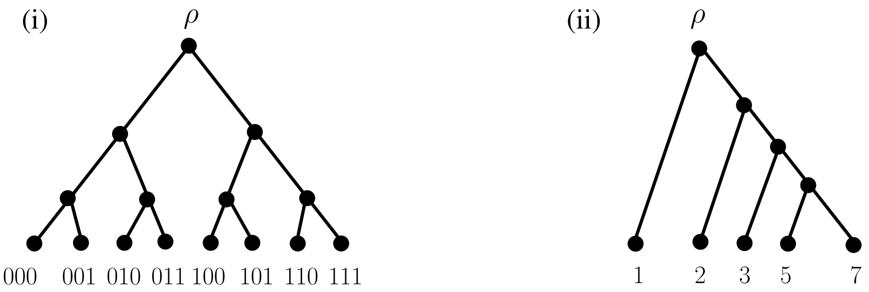

Figure 5: (i) The tree $\mathcal{P}_3$ on $\mathcal{B}_3$. (ii) The rooted caterpillar $\mathcal{C}_{\{1,2,3,5,7\}}$.