PHYLOGENETIC NETWORKS WITH EVERY EMBEDDED PHYLOGENETIC TREE A BASE TREE

CHARLES SEMPLE

Abstract. We show that the class of tree-child networks is precisely the class of tree-based networks with the property that every embedded phylogenetic tree is a base tree.

1. Introduction

Francis and Steel [4] recently introduced the class of phylogenetic networks called tree-based networks as a way of quantifying the concept of an 'underlying tree'. Intuitively, a phylogenetic network $\mathcal{N}$ is tree based if it can be obtained from a phylogenetic tree $\mathcal{T}$ by simply adding edges whose end-vertices subdivide edges of $\mathcal{T}$, in which case, $\mathcal{T}$ is called a base tree for $\mathcal{N}$. The notion of a tree-based network attempts to underlie the question of whether a phylogenetic network is simply a phylogenetic tree with some additional 'reticulation' edges or whether a phylogenetic network has little resemblance to a phylogenetic tree and so making the concept of an underlying tree meaningless. This notion is relevant to the continuing debate in the evolution of certain groups, such as prokaryotes, of whether evolution is tree-like with reticulation or whether it has no tree-like similarities at all [2, 3].

Horizontal gene transfer networks are tree based, but not all phylogenetic networks have this property. As well as a polynomial-time algorithm for deciding if an arbitrary phylogenetic network $\mathcal{N}$ is tree based, necessary and sufficient conditions for $\mathcal{N}$ to be a tree-based network are established in [4].

If $\mathcal{N}$ is a tree-based network and $\mathcal{T}$ is a base tree for $\mathcal{N}$, then $\mathcal{N}$ displays $\mathcal{T}$. But not every phylogenetic tree displayed by $\mathcal{N}$ is a base tree for $\mathcal{N}$. As a result, a number of natural questions arise, some of which are detailed in [4]. One particular question is the following. What is the class of phylogenetic networks $\mathcal{N}$ with the property that every embedding of a phylogenetic tree

Date: November 12, 2015.

Key words and phrases. Tree-child networks, tree-based networks.

The author was supported by the New Zealand Marsden Fund.

in $\mathcal{N}$ is a base tree for $\mathcal{N}$? In this paper, we prove that this class is the well-known class of tree-child networks.

The paper is organised as follows. In the rest of this section, we state some necessary definitions and the main result. The proof of this result is given in Section 2. Throughout the paper, X denotes a non-empty finite set. Notation and terminology follows Semple and Steel [5].

A *phylogenetic network* $\mathcal{N}$ *on* X is a rooted acyclic digraph with no edges in parallel and satisfying the following properties:

(i) the root has out-degree two;
(ii) a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is X;
(iii) all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one.

If $|X| = 1$, then $\mathcal{N}$ consists of the single vertex in X. In the literature, phylogenetic networks as described here are sometimes referred to as binary phylogenetic networks. Note that we could allow for parallel edges to be included in the definition of a phylogenetic network and, indeed, the results in this paper would still apply. However, in the literature, phylogenetic networks are typically defined without such edges.

Let $\mathcal{N}$ be a phylogenetic network. The vertices of $\mathcal{N}$ with out-degree zero are called *leaves*. Furthermore, the vertices with in-degree two and out-degree one are called *reticulations*, and the vertices with in-degree one and out-degree two are called *tree vertices*. The edges directed into a reticulation are *reticulation edges*, all other edges are *tree edges*. A *(binary) phylogenetic* X*-tree* is a phylogenetic network on X with no reticulations.

Let $\mathcal{T}$ be a phylogenetic X-tree and let $\mathcal{N}$ be a phylogenetic network on X. We say that $\mathcal{N}$ *displays* $\mathcal{T}$ if, up to contracting degree-two vertices, $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting edges and non-root vertices, in which case, the resulting acyclic digraph is an *embedding* of $\mathcal{T}$ in $\mathcal{N}$. Note that if $\mathcal{S}$ is an embedding of $\mathcal{T}$ in $\mathcal{N}$, then the root of $\mathcal{S}$ is the root of $\mathcal{N}$ and so it may have out-degree one.

A phylogenetic network $\mathcal{N}$ on X is a *tree-based network* if there is an embedding $\mathcal{S}$ of a phylogenetic X-tree $\mathcal{T}$ in $\mathcal{N}$ such that, for each edge e of $\mathcal{N}$ not in $\mathcal{S}$, both end-vertices of e are in $\mathcal{S}$. If this holds, then $\mathcal{S}$ (as well as $\mathcal{T}$) is a *base tree* for $\mathcal{N}$. Note that the definition of tree-based networks given here is different but equivalent to that in [4]. To illustrate, consider the phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, x_3, x_4, x_5\}$ shown in Figure 1. Here, $\mathcal{N}$ is a tree-based network. To see this, the embedding of the phylogenetic

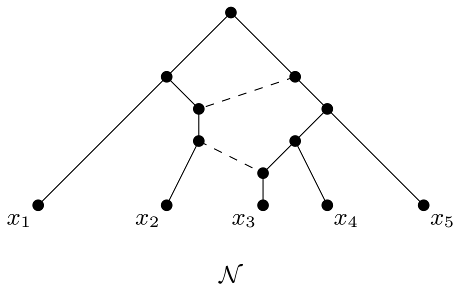

FIGURE 1. A tree-based network $\mathcal{N}$ and an embedding (solid edges) of a phylogenetic tree that is a base tree for $\mathcal{N}$.

X-tree shown with solid edges has the property that the end vertices of the remaining edges (dashed edges) of $\mathcal{N}$ are in the embedding.

A phylogenetic network $\mathcal{N}$ on X is a *tree-child network* if each non-leaf vertex v of $\mathcal{N}$ has a child that is either a tree vertex or a leaf. Introduced in [1], the class of tree-child networks is an increasingly prominent class of phylogenetic networks in the literature. It is easily checked that the phylogenetic network shown in Figure 1 is a tree-child network.

The main result of the paper is the following theorem. If u is a vertex of a phylogenetic network $\mathcal{N}$ and (u, v) is an edge in $\mathcal{N}$, we say v is a *child* of u and, conversely, u is a *parent* of v. More generally, u is an *ancestor* of a vertex w if there is a directed path from u to w in $\mathcal{N}$, in which case, w is a *descendant* of u.

Theorem 1.1. *The following statements are equivalent for a phylogenetic network $\mathcal{N}$ on X:*

(i) *$\mathcal{N}$ is a tree-child network.*
(ii) *No reticulation of $\mathcal{N}$ has a child reticulation and no tree vertex of $\mathcal{N}$ has two child reticulations.*
(iii) *Every embedded phylogenetic X-tree in $\mathcal{N}$ is a base tree for $\mathcal{N}$.*

The equivalence of (i) and (ii) in Theorem 1.1 is essentially no more than the definition of a tree-child network. However, it's included in Theorem 1.1 as this equivalence will be useful in its proof. An interesting problem that remains open is the following. Characterise the class of phylogenetic networks $\mathcal{N}$ with the property that every phylogenetic tree displayed by $\mathcal{N}$ is a base tree for $\mathcal{N}$. It is easily realised that this class strictly contains the class of tree-child networks. The subtleties of this problem is highlighted in the following example. Consider the phylogenetic network $\mathcal{N}_1$ on X_1 and the phylogenetic X_1-tree $\mathcal{T}_1$ shown in Figure 2, where $X_1 = \{x_1, x_2, x_3\}$. Now $\mathcal{N}_1$ displays $\mathcal{T}_1$, but the embedding (solid edges) of $\mathcal{T}_1$ in $\mathcal{N}_1$ is not a

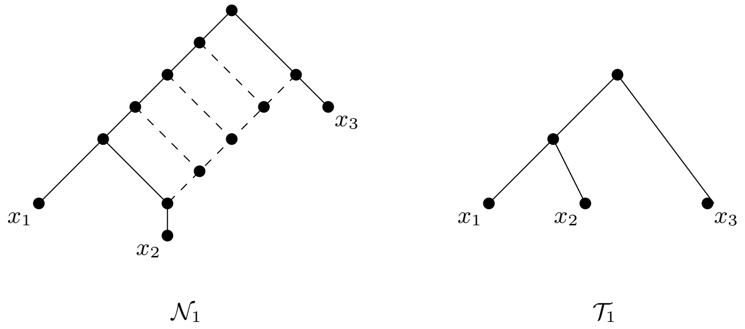

Figure 2. A phylogenetic network $\mathcal{N}_1$ and an embedding (solid edges) of a phylogenetic tree $\mathcal{T}_1$ that is not a base tree for $\mathcal{N}_1$.

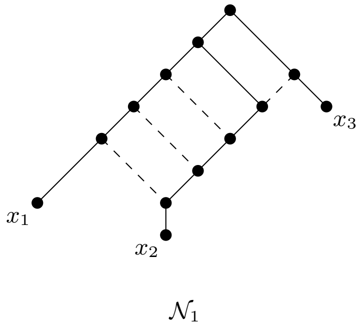

Figure 3. An embedding of $\mathcal{T}_1$ that is a base tree for $\mathcal{N}_1$.

base tree for $\mathcal{N}_1$. However, there does exist such an embedding as shown in Figure 3.

2. Proof of Theorem 1.1

In this section, we prove Theorem 1.1. We begin with a lemma. Let $\mathcal{N}$ be a phylogenetic network on X. Let u and e be a vertex and edge in $\mathcal{N}$, respectively. An embedding $\mathcal{S}$ of a phylogenetic X-tree in $\mathcal{N}$ *uses* u (respectively, e) if $\mathcal{S}$ contains u (respectively, e). Also, a directed path P in $\mathcal{N}$ ending at a leaf is a *tree path* if every intermediate vertex in P is a tree vertex.

Lemma 2.1. *Let $\mathcal{N}$ be a phylogenetic network on X and let $\mathcal{S}$ be an embedding of a phylogenetic X-tree displayed by $\mathcal{N}$. If P is a tree path in $\mathcal{N}$, then $\mathcal{S}$ uses each of the edges and vertices in P.*

Proof. Let P be a tree path in $\mathcal{N}$ starting at a vertex u and ending at a leaf ℓ. Suppose that there is a vertex or edge of P not used by $\mathcal{S}$. Then, as $\mathcal{S}$ uses ℓ, there is a vertex $v \neq u$ in P used by $\mathcal{S}$ such that all of the vertices and edges of the subpath of P from v to ℓ are used by $\mathcal{S}$ but not the edge of P directed into v. But then v is a reticulation and so, as $v \neq u$, the path P is not a tree path; a contradiction. This completes the proof of the lemma. □

It is easily seen (and well known) that if $\mathcal{N}$ is a tree-child network, then, for all vertices u in $\mathcal{N}$, there is a tree path in $\mathcal{N}$ starting at u. This fact is freely used in the proof of Theorem 1.1.

Proof of Theorem 1.1. We first prove that (i) implies (iii). Suppose that $\mathcal{N}$ is a tree-child network, and let $\mathcal{S}$ be an embedding in $\mathcal{N}$ of a phylogenetic X-tree $\mathcal{T}$. We need to show that $\mathcal{S}$ is a base tree for $\mathcal{N}$. The proof is by induction on the number r of reticulations in $\mathcal{N}$. If $r = 0$, then $\mathcal{N}$ is a phylogenetic X-tree and (iii) trivially holds. Now assume that (iii) holds whenever a tree-child network has at most $r - 1$ reticulations, where $r \geq 1$.

Let u be a reticulation in $\mathcal{N}$, and let p and q denote the parents of u. Since $\mathcal{N}$ is a tree-child network, both p and q are tree vertices. Furthermore, there is a tree path P_u in $\mathcal{N}$ starting at u. By Lemma 2.1, $\mathcal{S}$ uses P_u and, in turn, this implies that $\mathcal{S}$ uses exactly one of (p, u) and (q, u). Without loss of generality, we may assume that $\mathcal{S}$ uses (p, u). Again, as $\mathcal{N}$ is a tree-child network, $\mathcal{N}$ has tree paths P_p and P_q starting at p and q, respectively. Since u is a reticulation, neither P_p nor P_q contains u. By Lemma 2.1, $\mathcal{S}$ uses the vertices and edges in $P_p \cup P_q$ as well as the unique edges directed into p and q. Observe that the end vertices q and u of the edge (q, u) are used by $\mathcal{S}$.

Now let $\mathcal{N}'$ be the phylogenetic network on X obtained from $\mathcal{N}$ by deleting the edge (q, u) and contracting the resulting degree-two vertices q and u. As the child of q that is not u is a tree vertex and the unique child of u is a tree vertex, it is easily checked that $\mathcal{N}'$ is a tree-child network. Note that this holds whether or not either p or q is ancestor of the other. Let $\mathcal{S}'$ be the embedding of $\mathcal{T}$ in $\mathcal{N}'$ that is obtained from $\mathcal{S}$ by contracting q and u. Since $\mathcal{N}'$ has one less reticulation than $\mathcal{N}$, it follows by the induction assumption that $\mathcal{S}'$ is a base tree for $\mathcal{N}'$. In turn, this implies that $\mathcal{S}$ is a base tree for $\mathcal{N}$ and so (iii) holds.

We next prove that (iii) implies (ii). We establish the contrapositive of this implication. In particular, we show that if there exists reticulations u and v of $\mathcal{N}$ such that v is the unique child of u, or there exists a tree vertex of N with two child reticulations, then there exists an embedding of a phylogenetic X-tree in $\mathcal{N}$ that is not a base tree for $\mathcal{N}$. Let ρ denote the root of $\mathcal{N}$. First suppose that $\mathcal{N}$ contains reticulations u and v such that v

is the unique child of u. Let p be the parent of v that is not u. Let P be a directed path in $\mathcal{N}$ starting at ρ, traversing (p, v), and ending at a leaf. Since $\mathcal{N}$ is rooted and acyclic, such a path exists. Moreover, it is easily seen that there is an embedding $\mathcal{S}$ of a phylogenetic X-tree in $\mathcal{N}$ that uses all of the vertices and edges in P. But $\mathcal{S}$ cannot use (u, v) nor the two edges directed into u. Hence $\mathcal{S}$ is not a base tree for $\mathcal{N}$.

Now suppose $\mathcal{N}$ has distinct reticulations u and v sharing a parent p. By the last paragraph, we may assume that $\mathcal{N}$ has no parent-child reticulations. Let q_u and q_v be the parent of u and v, respectively, that is not p. Note that q_u and q_v need not be distinct. We next show that $\mathcal{N}$ has an embedding of a phylogenetic X-tree using (q_u, u) and (q_v, v). If there is a directed path in $\mathcal{N}$ starting at ρ, containing (q_u, u) and (q_v, v), and ending at a leaf, then such an embedding exists. So assume that there is no such path in $\mathcal{N}$. Then, as $\mathcal{N}$ is rooted and acyclic, there is a directed path P_{q_u} starting at ρ, ending at q_u, and avoiding (q_v, v) and any descendant of u and v. Similarly, there is a path P_{q_v} in $\mathcal{N}$ starting at ρ, ending at q_v, and avoiding (q_u, u) and any descendant of u and v.

Let P_u and P_v be directed paths in $\mathcal{N}$ starting at u and v, respectively, and ending at a leaf. Note that neither P_u nor P_v contains an ancestor of q_u or q_v. If P_u and P_v are vertex disjoint, then it is easily seen that there is an embedding of a phylogenetic X-tree in $\mathcal{N}$ using all the vertices and edges in

$$P_{q_u} \cup P_{q_v} \cup P_u \cup P_q \cup \{(q_u, u), (q_v, v)\}.$$

Assume P_u and P_v are not vertex disjoint. Let w be the first vertex in which P_u and P_v meet. Clearly, w is a reticulation. Let p_w be the parent of w, where $p_w \in P_u$. By assumption, $\mathcal{N}$ has no parent-child reticulations, so p_w is a tree vertex. Let t denote the child of p_w not equal to w. Let P'_u be a directed path in $\mathcal{N}$ obtained by taking the subpath of P_u from u to p_w, traversing (p_w, t), and adjoining a directed path from t to a leaf. As above, P'_u has the property of P_u that it does not contain an ancestor of q_u or q_v. If P'_u and P_v are vertex disjoint, then $\mathcal{N}$ has an embedding of a phylogenetic X-tree using all the vertices and edges in

$$P_{q_u} \cup P_{q_v} \cup P'_u \cup P_q \cup \{(q_u, u), (q_v, v)\}.$$

If P'_u and P_v are not vertex disjoint, repeat this process. Since $\mathcal{N}$ is acyclic, this process eventually constructs vertex disjoint paths P^*_u and P_q starting at u and v, respectively, ending at leaves, and containing no ancestors of q_u or q_v. In turn, this implies that there is an embedding $\mathcal{S}$ of a phylogenetic X-tree in $\mathcal{N}$ using all the vertices and edges in

$$P_{q_u} \cup P_{q_v} \cup P^*_u \cup P_q \cup \{(q_u, u), (q_v, v)\}.$$

In particular, $\mathcal{S}$ uses (q_u, u) and (q_v, v). But then $\mathcal{S}$ cannot use (p, u), (p, v), and the unique edge directed into p. Hence $\mathcal{S}$ is not a base tree for $\mathcal{N}$. It now follows that (iii) implies (ii).

Lastly, suppose that (ii) holds. Let u be a vertex of $\mathcal{N}$. If u is a tree vertex, then, as both of its children cannot be reticulations, it has at least one child that is a tree vertex. If u is a reticulation, then, as its unique child cannot be a reticulation, its child is a tree vertex. By definition, $\mathcal{N}$ is a tree-child network and so (i) holds. This completes the proof of the theorem. □

References

[1] G. Cardona, F. Rossello, and G. Valiente, Comparison of tree-child phylogenetic networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics 6 (2009) 552–569.

[2] T. Dagan and W. F. Martin, The tree of one percent, Genome Biology 7 (2006) 118.

[3] W. F. Doolittle and E. Bapteste, Pattern pluralism and the Tree of Life hypothesis, Proceedings of the National Academy of Sciences USA 104 (2007) 2043–2049.

[4] A.R. Francis and M. Steel, Which phylogenetic networks are merely trees with additional arcs?, Systematic Biology, 64 (2015) 768–777.

[5] C. Semple and M. Steel, Phylogenetics, Oxford University Press, New York, 2003.

School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: charles.semple@cantrebury.ac.nz