Properties of normal phylogenetic networks

Stephen J. Willson
Department of Mathematics
Iowa State University
Ames, IA 50011 USA
swillson@iastate.edu

August 13, 2009

Abstract. A phylogenetic network is a rooted acyclic digraph with vertices corresponding to taxa. Let X denote a set of vertices containing the root, the leaves, and all vertices of outdegree 1. Regard X as the set of vertices on which measurements such as DNA can be made. A vertex is called normal if it has one parent, and hybrid if it has more than one parent. The network is called normal if it has no redundant arcs and also from every vertex there is a directed path to a member of X such that all vertices after the first are normal. This paper studies properties of normal networks.

Under a simple model of inheritance that allows homoplasies only at hybrid vertices, there is essentially unique determination of the genomes at all vertices by the genomes at members of X if and only if the network is normal. This model is a limiting case of more standard models of inheritance when the substitution rate is sufficiently low.

Various mathematical properties of normal networks are described. These properties include that the number of vertices grows at most quadratically with the number of leaves and that the number of hybrid vertices grows at most linearly with the number of leaves.

Key words: normal network; hybrid; recombination; speciation; genome; ancestral reconstruction.

1 Introduction

Phylogenetic relationships are most commonly represented by rooted trees. The extant taxa correspond to leaves of the trees, while internal vertices correspond to ancestral species. The arcs correspond to genetic change, and bifurcations correspond to speciation events.

There has been increased interest recently in phylogenetic networks that are not necessarily trees. Such networks could include such additional events as

hybridization, recombination, or lateral gene transfer. Basic models of recombination were suggested by Hein [13]. General frameworks are discussed in [1], [2], [18], and [19]. Typically these frameworks model phylogenies by acyclic rooted directed graphs.

Recent evidence suggests that such reticulation is not rare. Nevertheless a common approach has been to seek networks with as few recombination events as possible, in part to give a lower bound. Wang *et al.* [21] considered the problem of finding a perfect phylogenetic network (see [20] p. 69) with recombination that has the smallest number of recombination events. They suggested that the problem is NP-hard, and a full proof was given by Bordewich and Semple in [5].

Various authors have considered restrictions on networks in order to obtain tractable problems. Wang *et al.* considered a restricted problem in which all recombination events are associated with node-disjoint recombination cycles, and they present a sufficient condition to identify such networks. Gusfield *et al.* [11] gave necessary and sufficient conditions to identify these networks, which they call "galled-trees," and they added a much more specific and realistic model of recombination events. The notion of galled trees has been generalized to the notion of "level-k" networks, i.e., networks in which every biconnected component contains at most k reticulation nodes [14]. A tree is level-0 while a galled tree is level-1. Baroni, Semple, and Steel [2] introduced the idea of a "regular" network, which coincides with its cover digraph. Cardona *et al.* [9] discussed "tree-child" networks, in which every vertex not a leaf has a child that is not a reticulation vertex.

This paper concerns a restricted family of rooted directed acyclic networks here called "normal" with respect to a "base-set X". The base-set X corresponds to the set of taxa on which direct information (such as DNA sequences) are known. We assume that X includes all the leaves, the root, all vertices of outdegree 1, and perhaps some other vertices as well. (An outgroup can be used in place of the true root.) A vertex is called "normal" if it has only one parent and "hybrid" if it has more than one parent. In normal networks, every vertex not in X has a child which is not a hybrid vertex. Related restrictions on networks include the "tree-sibling" networks [18], [10], and [6].

Moreover, an important assumption for normal networks is that there are no "redundant" arcs (u, v) such that there is already a directed path from vertex u to vertex v without use of the arc (u, v). In the literature some authors allow redundant arcs, and others do not. "Regular networks" [2] do not allow redundant arcs since each such network must be isomorphic with its "cover digraph" in which redundant arcs are excluded. The important class of "tree-child" networks [9] permits redundant arcs. They are, however, eliminated in the related class of "tree-child time-consistent networks" [7], [8] as a consequence of a notion of time-consistency, just as they are for similar reasons in [3] where networks are assumed to have a "temporal labeling."

Previous work by the author [22], [23], [24] includes theorems about polynomial-time methods for reconstructing normal networks from information on their leaves (with additional assumptions as well). These papers show that some basic network-reconstruction problems are tractable on certain families of normal

networks. (In some of those papers the term "regular" is used where in this paper we utilize the term "normal". The reason for the change is to distinguish the notion from that of regularity as defined by [2] since both notions are important in the current paper.)

Networks are an idealized tool for understanding biological relationships. In this paper, we consider some idealized uses of phylogenetic trees and then ask which networks can be used for the same idealized purposes. The basic theme is that only normal networks can be utilized for many of these purposes and still be determinable from data on extant species.

Indeed, suppose that T is a rooted tree with the set X of leaves and no vertices of total degree 2. Suppose every character is binary with state 0 at the root and there is "perfect phylogeny" (see [20], p. 69) in the sense that for each state of each character the set of vertices whose genomes have that state is connected. Then the following are true:

(A) For each character i, the vertex u^i at which the character mutates to state 1 from the allele 0 at the root is uniquely determined by the characters at members of X.

(B) If $K = \{x \in X :$ the character i at x has allele 1$\}$, then u^i is the most recent common ancestor of the members of K.

(C) The genome of each vertex is uniquely determined.

It follows that, for rooted trees, since taxa in X correspond to extant species, measurements can in principle reconstruct the genomes at internal vertices in this idealized situation. Under a more realistic model of evolution, homoplasies occur and the genomes will not be completely determinable at the internal vertices. Nevertheless, many individual characters may still satisfy the assumptions of the idealized situation; for these characters a plausible reconstruction of the character at the internal vertices can be made.

Now suppose that instead there is a phylogenetic network N (not necessarily a tree) with the base-set X. Again suppose that every character is binary. Under what assumptions on N and on the model for inheritance are the genomes at each vertex uniquely determined from data on X in a similar manner?

This paper treats two simple models of inheritance, both closely related to perfect phylogeny. The first is called Accumulation Phylogeny and was introduced by Baroni and Steel [4]. The second model is much less strict than Accumulation Phylogeny at hybrid vertices and is called Relaxed Accumulation.

Both models assume that there are never any "parallel mutations" in which the same mutation occurs at the same site at different parts of the phylogenetic network and also that at normal vertices there are never any "back-mutations". The two models differ on the nature of inheritance at hybrid vertices.

Both models can be regarded as limiting cases of inheritance when substitutions are so rare that no two mutations ever occur at the same site. For example, a back-mutation would require two mutations at the same site, hence could never occur in the limiting case. Accumulation Phylogeny is the limiting case where inheritance at hybrid vertices is similar to that in polyploidy, where the full genome of both parents is inherited. Relaxed Phylogeny is the limiting case when the hybrid child inherits only some genes from each parent.

The main results, roughly stated, are as follows:
(1) (Theorem 3.2) Suppose (A) and (C) hold under the model of Accumulation Phylogeny. Then N must be regular in the sense of [2].
(2) (Theorem 3.4) If N is normal, then N is regular.
(3) (Theorem 3.3) If N is normal, then (A), (B), and (C) hold under Accumulation Phylogeny.
(4) (Theorem 4.1) N is normal iff, under Relaxed Accumulation, (A) is true up to a certain geometrically defined ambiguity.
(5) (Theorems 4.2 and 4.3) If N is normal, then under Relaxed Accumulation, (B) and (C) are true up to a certain ambiguity.

It follows that, if (A), (B), and (C) are to be true for a network N under either Accumulation Phylogeny or Relaxed Accumulation (up to certain specified ambiguities), then N must be normal.

While neither Accumulation nor Relaxed Accumulation is realistic, one would expect that any method of reconstructing ancestral genomes under a more realistic model of evolution (allowing, for example, homoplasies at normal vertices) should be able to deal with characters arising from these simple models. This should be true since Relaxed Accumulation is a limiting case when the mutation rate is sufficiently low. Even when the mutation rate is somewhat higher, some characters should still satisfy the assumptions of Relaxed Accumulation, and for these characters reconstruction at the ancestral vertices should be possible. See further discussion of this point in Section 6.

Figure 1 shows a phylogenetic network that is regular but not normal. It is not normal because every child of 9 is hybrid. Observations can be made only at the members of $X = \{1, 2, 3, 4, 5\}$. Suppose a binary character a is observed at vertices 2 and 3 but nowhere else in X. Under Accumulation Phylogeny, one could infer that character a originated at 8 , the most recent common ancestor of 2 and 3. Under Relaxed Accumulation, however, it might originate at 8, but it might also originate at 7, then be inherited by 8, 9, 2, and 3 but not by 4 (which would inherit the allele from 10 rather than 9). Hence it is not determined whether vertex 7 has character a; and (A), (B), and (C) fail for this network. The theorems in this paper show that such failure is not possible for a normal network.

Section 5 shows that normality is a very serious restriction on a network. If the base-set consists only of the root and the leaves, then (Theorem 5.1) the number of vertices in a normal network grows at most quadratically with the number of leaves, and (Theorem 5.2) the number of hybrid vertices grows at most linearly with the number of leaves. By contrast, in regular networks the number of vertices and and the number of hybrid vertices can grow exponentially with the number of leaves. In general rooted acyclic directed graphs, the numbers of vertices and hybrid vertices are unbounded with respect to the number of leaves.

Normality is a property largely independent of "level" [14]. A normal network can have arbitrarily high level, and even a level-1 network need not be normal.

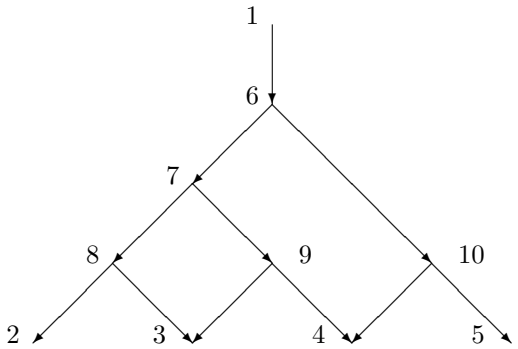

Figure 1: A phylogenetic network $N = (V, A, r, X)$. Here $X = \{1, 2, 3, 4, 5\}$. The network is regular but not normal. Under Relaxed Accumulation it is not possible to decide whether a character observed only in 2 and 3 is present at 7.

2 Fundamentals

In this section basic concepts for phylogenetic networks are defined, including the notion of regularity. Features common to both models of evolution are also described.

A *directed graph* or *digraph* (V, A) consists of a finite set V of *vertices* and a finite set A of *arcs*, each consisting of an ordered pair (u, v) where $u \in V$, $v \in V$, $u \neq v$, interpreted as an arrow from u to v. There are no multiple arcs and no loops. If $(u, v) \in A$, say that u is a *parent* of v and v is a *child* of u. A *directed path* is a sequence $u_0, u_1, \cdots, u_k$ of vertices such that for $i = 1, \cdots, k$, $(u_{i-1}, u_i) \in A$. The path is *trivial* if $k = 0$. Write $u \leq v$ if there is a directed path starting at u and ending at v. The digraph is *acyclic* if there is no nontrivial directed path starting and ending at the same point. If the digraph is acyclic, it is easy to see that $\leq$ is a partial order on V.

The digraph (V, A) has *root* r if there exists $r \in V$ such that for all $v \in V$, $r \leq v$. The graph is *rooted* if it has a root.

The *indegree* of vertex u is the number of $v \in V$ such that $(v, u) \in A$. The *outdegree* of u is the number of $v \in V$ such that $(u, v) \in A$. If (V, A) is rooted at r then r is the only vertex of indegree 0. A *leaf* is a vertex of outdegree 0. A *normal* (or *tree-child*) vertex is a vertex of indegree 1. A *hybrid* vertex (or *recombination vertex* or *reticulation node*) is a vertex of indegree at least 2.

A *base-set* X is a subset of V that contains the root, all leaves, and all vertices of outdegree 1. The interpretation of X is that its members correspond to taxa on which direct measurements may be made. The leaves correspond typically to extant taxa on which such measurements as DNA are possible. While the root is usually thought of as being a remote ancestor, it may be replaced in practice by an outgroup on which measurements can be made. In consideration of trees, it is common to suppress any vertices of outdegree 1 other than possibly the

root (hence total degree 2) because nothing in the tree uniquely identifies such a taxon. Hence if such exists, it is because special information is known about such a taxon, whence we assume measurements on it can be made and it is in X.

An arc (u, v) is *redundant* if there exists $w \in V$ such that u, v, and w are distinct and $u \leq w \leq v$. The inclusion of a redundant arc is problematic since it duplicates much genetic information while adding to both indegrees and outdegrees.

In this paper a *phylogenetic network* $N = (V, A, r, X)$ is an acyclic rooted digraph (V, A) with root r and base-set X such that there are no redundant arcs.

The fundamental problem is to learn about N from information on X only.

Let W be a nonempty subset of V. The *most recent common ancestor* of W, denoted mrca(W), is the vertex $u \in V$, if it exists, such that the following hold:

(1) For all $w \in W$, $u \leq w$.

(2) Suppose v satisfies that for all $w \in W$, $v \leq w$. Then $v \leq u$.

If mrca(W) exists, it is unique. This is because, if u_1 and u_2 both satisfy the definition then by (1) for all $w \in W$, $u_2 \leq w$, whence by (2) $u_2 \leq u_1$. By a symmetric argument, $u_1 \leq u_2$. Hence $u_1 = u_2$. It is easy to construct examples in networks, however, where mrca(W) does not exist.

Let $\mathcal{P}(X)$ denote the collection of subsets of X. Following [2] define the *cluster map* $c : V \to \mathcal{P}(X)$ by $c(v) = \{x \in X : v \leq x\}$.

Baroni, Semple, Steel [2] defined the class of regular networks. The network N (with no redundant arcs) is *regular* iff

(1) c is one-to-one; and

(2) $u \leq v$ iff $c(v) \subseteq c(u)$.

Since there are no redundant arcs, by (2) there is an arc (u, v) iff $c(v) \subset c(u)$ and there is no vertex w such that $c(v) \subset c(w) \subset c(u)$. To see this, if there is an arc (u, v) then $u \leq v$, whence, by (2), $c(v) \subseteq c(u)$ and by (1), $c(v) \subset c(u)$. If there existed w such that $c(v) \subset c(w) \subset c(u)$, then the three vertices would be distinct, and by (2), $u \leq w \leq v$, so the arc (u, v) would be redundant.

A directed path $u = u_0, u_1, u_2, \cdots, u_k = v$ (where (u_{i-1}, u_i) is an arc for $i = 1, \cdots, k$) is a *normal path from u to v* provided for $i > 0$ u_i is normal. Note u may or may not be hybrid. There is a *normal path from u to X* if there is a normal path from u to some x such that $x \in X$. If $u \in X$, then the path $u = u_0$ with $k = 0$ is a *trivial normal path* to X. The network N is *normal* if from every vertex $u \in V$ there is a normal path to X.

This paper will consider a number of models of evolution of the genome. All these models will have the following in common: Let $N = (V, A, r, X)$ be a phylogenetic network. There is a finite set C of *characters*. Each character is binary, having exactly two states 0 and 1. The genome at the root r is known (since $r \in X$) and the state of each character at the root will be assumed to be 0. (Otherwise, for that character, exchange the names of the two states.) Hence the genome at each vertex v may be completely described by a set

$$M(v) = \{i \in C : \text{ the state of character } i \text{ at } v \text{ is } 1\}$$

called the *mutated genome* at v. Thus $M(v)$ identifies those characters whose state at v differs from the state at the root. Trivially, $M(r) = \emptyset$. For each $i \in C$ assume that there exists a unique vertex $u^i \in V$ such that $i \in M(u^i)$ and such that there is no parent p of u^i such that $i \in M(p)$. Call u^i the *originating vertex* for i, since it is the vertex at which there is a mutation for the first time of character i from the state 0 at the root. Since u^i is assumed to be unique, we are assuming that there is no parallel evolution and the same mutation never arises spontaneously two different times. This assumption is plausible if the mutation rate is sufficiently low. It follows that, if $i \in M(v)$, then $u^i \leq v$. Assume furthermore that for each $v \in V$, $v \neq r$, there exists $i \in C$ such that $u^i = v$. This second assumption asserts that each vertex exhibits some genetic innovation.

For $u \in V$, define $O(v) = \{i \in C : u^i = v\}$ and call $O(v)$ the *originating set* at v. Note $O(v) \subseteq M(v)$.

3 Regular networks

In this section the model of evolution called Accumulation Phylogeny, described by Baroni and Steel [4], is reviewed. It was shown in [4] that if a phylogenetic network is regular, then the genomes of all vertices under Accumulation Phylogeny are uniquely determined by the genomes at members of X. Here a partial converse Theorem 3.2 is proved. Theorem 3.4 shows that every normal network is regular. Theorems 3.3 and Corollary 3.6 give simple ways to identify the genomes at vertices in a normal network under Accumulation Phylogeny.

Let $N = (V, A, r, X)$ be a phylogenetic network. There is inheritance under *Accumulation Phylogeny* provided that for all $i \in C$,

(A1) there exists a unique vertex $u^i \in V$ such that $i \in M(u^i)$ and no parent p of u^i satisfies $i \in M(p)$; and

(A2) $i \in M(v)$ iff $u^i \leq v$.

Thus every character i originates exactly once (at u^i), and every descendent of u^i exhibits the modified character i. Since hybrid children inherit all the modifications of all parents, this model is plausible primarily when hybridization events are polyploidies.

Clearly, under Accumulation Phylogeny, for each $v \in V$, $M(v) = \{i \in C : u^i \leq v\}$. Moreover, if $u \leq v$ then $M(u) \subseteq M(v)$.

The following uniqueness theorem is a consequence of Theorem 3.2 of [4].

Theorem 3.1. *Let $N = (V, A, r, X)$ be a regular phylogenetic network. Assume X consists of the root r and all the leaves. Assume that evolution by Accumulation Phylogeny results in the mutated genomes $M(x)$ for $x \in X$. Then N is the only regular phylogenetic network with base set consisting of the root and all leaves such that Accumulation Phylogeny results in $M(x)$ for all $x \in X$. Moreover, for all $v \in V$, $M(v)$ is uniquely determined.*

Under the assumption that N is regular, it follows that the graph (V, A) is uniquely determined by $M(x)$ for all $x \in X$. Moreover, $M(v)$ will then be uniquely determined for all $v \in V$, not just $v \in X$, whence the genomes at all vertices are uniquely determined by $M(x)$ for all $x \in X$.

The next theorem shows that networks that are easily interpretable under the Accumulation Phylogeny model will be regular:

Theorem 3.2. *Let $N = (V, A, r, X)$ be a phylogenetic network. Assume that the mutated genomes M obey Accumulation Phylogeny. Suppose*
(1) N has the genomes of all vertices determined by the genomes at members of X; and
(2) N exhibits all the genetic influences that could determine the genomes at members of X.
Then N is regular.

Note that (1) asserts that knowledge of $M(x)$ for all $x \in X$ will uniquely determine $M(v)$ for all $v \in V$. Similarly (2) asserts that there is no additional nonredundant arc (u, v) between vertices u and v in V which is compatible under Accumulation Phylogeny with the sets $M(x)$ for all $x \in X$.

To illustrate Theorem 3.2, Figure 2 shows a network N that is not regular and in which (1) fails. N is not regular because $c(6) = c(7) = \{2, 3, 4, 5\}$. Suppose character a is observed in $M(2)$, $M(3)$, $M(4)$, $M(5)$ but not in $M(1)$. Under Accumulation Phylogeny there is no way to decide whether a originated at 6 or at 7; hence the genome of 6 is not determined.

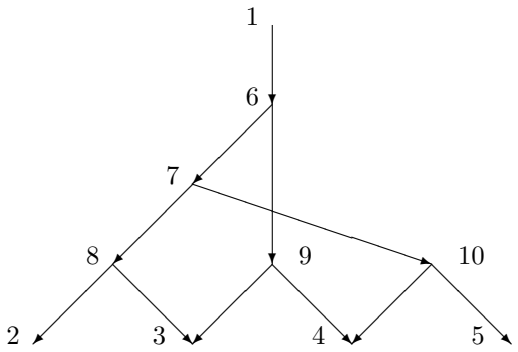

Figure 2: A phylogenetic network $N = (V, A, r, X)$. Here $X = \{1, 2, 3, 4, 5\}$. The network is not regular. Under Accumulation Phylogeny there is no way to tell whether a character observed only in 2, 3, 4, and 5 is present at 6.

Proof. We first show that $c : V \to \mathcal{P}(X)$ is one-to-one. Suppose $c(u) = c(v)$; we show that $u = v$. Choose $i \in C$ such that $u^i = u$. By Accumulation Phylogeny it follows that for $x \in X$ we have $i \in M(x)$ iff $u^i \leq x$ iff $x \in c(u^i)$ iff $x \in c(u)$.

Now consider a second Accumulation Phylogeny M' on N such that the only difference from M is that $u'^i = v$. Then $i \in M'(x)$ iff $x \in c(v)$. Since by assumption $c(u) = c(v)$, it follows that $i \in M(x)$ iff $i \in M'(x)$. Hence for all $x \in X$, $M(x) = M'(x)$. By (1), it follows that for all $w \in V$, $M(w) = M'(w)$. In particular, $M(u) = M'(u)$ and $M(v) = M'(v)$. But $i \in M(u)$ since $u^i = u$, whence $i \in M'(u)$, whence $u'^i \leq u$, whence $v \leq u$. Similarly $i \in M'(v)$ since $u'^i = v$, whence $i \in M(v)$ and $u^i \leq v$, whence $u \leq v$. It follows that $u = v$.

We next show the second condition of regularity. Suppose $u \leq v$. If $x \in c(v)$, then $v \leq x$, whence $u \leq v \leq x$ so $x \in c(u)$. Hence $c(v) \subseteq c(u)$.

Conversely, suppose $c(v) \subseteq c(u)$. I claim $u \leq v$. Let each vertex have mutated genome M under Accumulation Phylogeny. Suppose it is false that $u \leq v$. Adjoin the arc (u, v) yielding $A' = A \cup \{(u, v)\}$ to form the network $N' = (V, A', r, X)$ in which, assuming Accumulation Phylogeny, vertices have mutated genomes denoted M', the cluster map is denoted c', and write $a \leq' b$ if there is a directed path in N' from a to b.

For all $a \in V$, I claim that $c(a) = c'(a)$. To see this, observe that if $x \in c(a)$, then $a \leq x$ whence $a \leq' x$ since each arc of N is an arc of N', whence $x \in c'(a)$. Conversely, if $x \in c'(a)$ then $a \leq' x$. There are two cases. In the event that a directed path from a to x in N' does not involve the arc (u, v), then $a \leq x$ in N, whence $x \in c(a)$. On the other hand if a directed path from a to x in N' involves the arc (u, v), then $a \leq u$ in N and $v \leq x$ in N. Hence $x \in c(v)$, whence $x \in c(u)$ since $c(v) \subseteq c(u)$, whence $x \in c(a)$ since $a \leq u$ and $u \leq x$.

It follows that for each character i and for each $x \in X$, we have $i \in M(x)$ iff $x \in c(u^i)$ iff $x \in c'(u^i)$ iff $i \in M'(x)$.

Thus the inclusion of the additional genetic influence given by (u, v) does not change the genome at any member of X. Hence N did not exhibit all the genetic influences required by (2), a contradiction. □

The next theorem shows that when N is normal, under Accumulation Phylogeny, the identity of the originator for any character i is easily computed.

Theorem 3.3. *Assume that $N = (V, A, r, X)$ is a normal phylogenetic network. Then for all $v \in V$, mrca$(c(v))$ exists and $v =$ mrca$(c(v))$. In particular, if character i originates at u^i under Accumulation Phylogeny, then*

$$u^i = \text{mrca}(\{x \in X : i \in M(x)\}).$$

Proof. The proof will be by induction. Assume that the result is true for all children of v.

If v is a leaf, then $v \in X$ and $c(v) = \{v\}$. Trivially then $v = \text{mrca}(\{v\})$.

Assume that v is not a leaf. Let the children of v be denoted $c_1, \cdots, c_k$ in V.

Case 1. Assume $v \in X$. Then $c(v) = \{v\} \cup \bigcup [c(c_i) : i = 1, \cdots, k]$. I claim $v = \text{mrca}(c(v))$. To see this, note that $v \leq x$ for $x \in c(v)$ since if $x \in c(c_i)$ then $v \leq c_i \leq x$. Conversely, suppose $w \leq x$ for all $x \in c(v)$; then in particular, $w \leq v$.

Case 2. Assume $v \notin X$. Then $c(v) = \cup[c(c_i) : i = 1, \cdots, k]$. I claim $v = \text{mrca}(c(v))$.

Note that if $x \in c(v)$, then $x \in c(c_i)$ for some i, whence $v \leq c_i \leq x$. There remains to show that if $w \leq x$ for all $x \in c(v)$, then $w \leq v$.

Since $v \notin X$ and there is a normal path from v to X, it follows that some child c_i is normal. Without loss of generality, assume c_1 is normal. Since $w \leq x$ for all $x \in c(v)$, it follows $w \leq x$ for all $x \in c(c_1)$. By induction $c_1 = \text{mrca}(c(c_1))$, whence $w \leq c_1$. Since c_1 is normal with unique parent v, either $w = c_1$ or $w \leq v$. Since $v \notin X$, v has outdegree at least 2, so $k \geq 2$. Since $w \leq x$ for all $x \in c(c_2)$, it follows $w \leq \text{mrca}(c(c_2)) = c_2$. If $w = c_1$, then $v < c_1 < c_2$, which makes the arc (v, c_2) redundant, contradicting that N is a phylogenetic network. Hence $w \neq c_1$, whence $w \leq v$. This proves that $v = \text{mrca}(c(v))$.

If i originates at u^i under Accumulation Phylogeny, then $c(u^i) = \{x \in X : i \in M(x)\}$ is the observable set of extant taxa at which the mutated allele of i is observed, and $u^i = \text{mrca}(c(u^i))$. □

The next theorem shows that if N is normal, then N is regular, whence it has properties studied by [2].

Theorem 3.4. *Suppose $N = (V, A, r, X)$ is normal. Then N is regular.*

Proof. For each $v \in V$, by Theorem 3.3 we have $v = \text{mrca}(c(v))$. It follows that c is one-to-one, since if $c(v) = c(w)$, then $v = \text{mrca}(c(v)) = \text{mrca}(c(w)) = w$.

For property (2) of regularity, suppose $u \leq v$. Then for $x \in X$, if $v \leq x$ then $u \leq x$, whence $c(v) \subseteq c(u)$. Conversely, suppose $c(v) \subseteq c(u)$. Then for each $x \in c(v)$ we have $x \in c(u)$, whence $u \leq x$ for each $x \in c(v)$. Since $\text{mrca}(c(v)) = v$ by Theorem 3.3, it follows $u \leq \text{mrca}(c(v)) = v$. This proves regularity. □

Note that by regularity of N and Theorem 3.1, under the assumption of Accumulation Phylogeny, the arcs of a normal phylogenetic network N are determined uniquely by $M(x)$ for $x \in X$.

The next two results show that under Accumulation Phylogeny all the mutated genomes $M(v)$ are uniquely determined by the mutated genomes $M(x)$ for $x \in X$ when the network is normal.

Lemma 3.5. *Suppose K is a subset of X and $v = mrca(K)$ exists. Then under Accumulation Phylogeny, $M(v) = \cap[M(x) : x \in K]$.*

Proof. If $x \in K$, then $v \leq x$. Hence $M(v) \subseteq M(x)$, whence $M(v) \subseteq \cap[M(x) : x \in K]$. Conversely, suppose $i \in \cap[M(x) : x \in K]$; we show $i \in M(v)$. For $x \in K$, $i \in M(x)$, whence $u^i \leq x$ according to Accumulation Phylogeny. It follows that $u^i \leq \text{mrca}(K) = v$, whence $i \in M(v)$. □

Corollary 3.6. *Suppose N is normal. Then under Accumulation Phylogeny for all $v \in V$, $M(v) = \cap[M(x) : x \in c(v)]$.*

Proof. By Theorem 3.3, $\text{mrca}(c(v)) = v$, whence the result follows from Lemma 3.5. □

The next few results give properties of normal networks that will be needed in Section 5 for an investigation of how restrictive is the assumption of normality.

Lemma 3.7. *Let the vertex x have distinct children c and d. Let $c = c_1, c_2, \cdots, c_k = x_1$ be a normal path from c to x_1 and $d = d_1, d_2, \cdots, d_j = x_2$ be a normal path from d to x_2. Then $x_1 \neq x_2$.*

Proof. Otherwise, $x_1 = x_2$. We prove the result by induction on j+k.

Case (1) If $k = 1$ then $x_1 = c = x_2$. Hence $x < d \leq x_2 = c$. By nonredundancy of the arc (x, c), we obtain $d = c$, a contradiction.

Case (2) If $j = 1$ then $x_1 = x_2 = d$. Hence $x < c \leq x_1 = d$. By nonredundancy of the arc (x, d) we obtain $d = c$, a contradiction.

Case (3) Otherwise $k > 1$ and $j > 1$. By normality c_k has unique parent c_{k-1} and d_j has unique parent d_{j-1}, whence $c_{k-1} = d_{j-1}$. But $(j-1)+(k-1) < j+k$. By the inductive hypothesis it follows $c_{k-1} \neq d_{j-1}$, a contradiction. □

Lemma 3.8. *Suppose v has distinct children a and b. Suppose there is a normal path from a to x_1 and a normal path from b to x_2. Assume a is normal. Then $v = mrca(x_1, x_2)$.*

Proof. Let the normal paths be $a = a_0, a_1, a_2, \cdots, a_j = x_1$ and $b = b_0, b_1, b_2, \cdots, b_k = x_2$. Suppose $u \leq x_1$ and $u \leq x_2$. To show $v = \text{mrca}(x_1, x_2)$, we prove that $u \leq v$.

By normality we have either $u = x_1 = a_j$, or $u \leq a_{j-1}$. If $u \leq a_{j-1}$ then either $u = a_{j-1}$ or $u \leq a_{j-2}$. In like manner we see that either $u = a_j$, or $u = a_{j-1}, \cdots$, or $u = a_1$, or $u \leq a_0 = a$. Similarly, either $u = x_2$, or $u = b_{k-1}$, .., or $u = b_1$, or $u \leq b_0 = b$. By Lemma 3.7 we cannot have simultaneously $u = a_m$ and $u = b_n$ for any m and n. Hence the possibilities are
Case (i) $u \leq a$, $u \leq b$.
Case (ii) $u \leq a$, $u = b_n$ for some n.
Case (iii) $u = a_m$ for some m, $u \leq b$.

In case (ii), $v < b \leq b_n = u \leq a$, contradicting the nonredundancy of the arc (v, a). In case (iii), $v < a \leq a_m = u \leq b$, contradicting the nonredundancy of the arc (v, b). Hence (i) must apply. But then since a is normal and $u \leq a$, we have either $u = a$ or $u \leq v$. If $u = a$, then $v < a = u \leq b$ contradicts the nonredundancy of arc (v, b). Hence $u \leq v$. This shows $v = \text{mrca}(x_1, x_2)$. □

Theorem 3.9. *Assume that $N = (V, A, r, X)$ is normal. For each vertex v that is not in X, there exist x_1 and x_2 in X such that $v = mrca(x_1, x_2)$.*

Proof. Since $v \notin X$, v has outdegree at least 2 hence has distinct children a and b. By normality we may assume a is normal. Choose x_1 and x_2 as in Lemma 3.8. Then $v = \text{mrca}(x_1, x_2)$. □

4 Inheritance under Relaxed Accumulation

In this section the model of Accumulation Phylogeny is modified to allow more inheritance possibilities at a hybrid vertex, yielding a model called Relaxed

Accumulation. The cluster map c used to define regularity is generalized to a "generalized cluster map" gc. The main result Theorem 4.1 is that normality of a network is equivalent to gc being "generalized monotone." For a normal network, Theorem 4.2 gives a simple way to identify where a character originates under Relaxed Accumulation, up to a clearly described ambiguity. Similarly Theorem 4.3 tells how to determine the genome at each vertex from the genomes at X up to a clearly described ambiguity.

Genetic inheritance via Accumulation Phylogeny is very restrictive, especially at hybrid vertices. Suppose that v has parents p and q. Except in the case of polyploidy (where the child inherits all chromosomes of each parent) it is unlikely that v inherits all the mutated genomes of both p and q.

We therefore seek to extend the results of section 3 to a model of evolution that is less restrictive at hybrid vertices. Model 2, called Relaxed Accumulation, assumes that normal children inherit the mutated genomes of their parents, just as in Accumulation Phylogeny. Hybrid children, however, may or may not inherit a mutated character from a parent. The assumption is plausible when the substitution rate is sufficiently low. In this view, multiple substitutions at a site are regarded only as noise to be ignored.

Following are the assumptions of Relaxed Accumulation:
Let $N = (V, A, r, X)$ be a phylogenetic network. Assume
(A1) Each hybrid vertex has indegree exactly two, hence has exactly two parents.
(A2) For each character $i \in C$, there exists a unique vertex $u^i \in V$ such that $i \in M(u^i)$ and no parent p of u^i satisfies $i \in M(p)$.
(A3) For each character $i \in C$, the set $FE(i) = \{v \in V : i \in M(v)\}$ satisfies
(A3a) $u^i \in FE(i)$.
(A3b) If $u \in FE(i)$ and w is a normal child of u, then $w \in FE(i)$.
(A3c) If $u \in FE(i)$ and w is a hybrid child of u, then w may or may not be in $FE(i)$.
(A3d) All members of $FE(i)$ are obtained via (A3a) through (A3c).
Call $FE(i)$ the *full expression set for* i since it is the set of all vertices at which the mutated allele of character i is expressed.

Note that, under (A3), if v is hybrid with parents p and q, there are the following eight possibilities for each character i:
(i) $p \in FE(i)$, $q \in FE(i)$, $v \in FE(i)$
(ii) $p \in FE(i)$, $q \in FE(i)$, $v \notin FE(i)$
(iii) $p \in FE(i)$, $q \notin FE(i)$, $v \in FE(i)$
(iv) $p \in FE(i)$, $q \notin FE(i)$, $v \notin FE(i)$
(v) $p \notin FE(i)$, $q \in FE(i)$, $v \in FE(i)$
(vi) $p \notin FE(i)$, $q \in FE(i)$, $v \notin FE(i)$
(vii) $p \notin FE(i)$, $q \notin FE(i)$, $v \in FE(i)$
(viii) $p \notin FE(i)$, $q \notin FE(i)$, $v \notin FE(i)$.

If (vii) occurs, then $i \in O(v)$ and $v = u^i$. By (A2) there is exactly one vertex u^i, and here that vertex is v. If (ii) occurs then there was a homoplasy at v, since the character i reverts to the state at the root even though neither parent exhibited the mutated form. If we restrict the inheritance by disallowing (ii)

then it is easy to see that the result is a perfect phylogeny in the sense that whenever two vertices u and v satisfy that $i \in M(u) \cap M(v)$, then there is a path in N (not necessarily directed) such that for all w on the path, $i \in M(w)$.

If $i \in C$, define

$$E(i) = \{x \in X : i \in M(x)\}.$$

Call $E(i)$ the *observed expression set* for i since all observations are on members of X. Note $E(i) = FE(i) \cap X$.

Suppose $v = u^i$. Under Accumulation Phylogeny, $E(i) = c(v)$. Under Relaxed Accumulation there may be many more possibilities for $E(i)$.

Figure 3 exhibits a network N which is a motivating example.

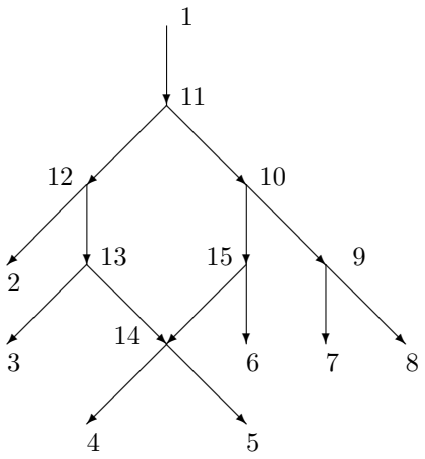

Figure 3: A normal phylogenetic network $N = (V, A, r, X)$. Here $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

If $u^i = v$ is a leaf, then $E(i) = \{v\}$. If $u^i = 9$ then $E(i) = \{7, 8\}$. If $u^i = 10$ then the two possibilities are $E(i) = \{4, 5, 6, 7, 8\}$ and $E(i) = \{6, 7, 8\}$. If $u^i = 11$ then the possibilities are $E(i) = \{2, 3, 4, 5, 6, 7, 8\}$ and $E(i) = \{2, 3, 6, 7, 8\}$. If $u^i = 12$ then the possibilities are $E(i) = \{2, 3, 4, 5\}$ and $E(i) = \{2, 3\}$. If $u^i = 13$ then the possibilities are $E(i) = \{3, 4, 5\}$ and $E(i) = \{3\}$. If $u^i = 14$ then $E(i) = \{4, 5\}$. If $u^i = 15$ then the possibilities are $E(i) = \{4, 5, 6\}$ and $E(i) = \{6\}$. Since 1 is the root, we cannot have $u^i = 1$.

Each of the indicated sets except $\{3\}$ and $\{6\}$ occurs exactly once. The set $\{3\}$ occurred as a possible $E(i)$ both if $u^i = 3$ and if $u^i = 13$. The set $\{6\}$ occurred as a possible $E(i)$ both if $u^i = 6$ and if $u^i = 15$. If, for example, i occurs in the genome of exactly $E(i) = \{2, 3, 6, 7, 8\}$ then we know i originated at 11, or $u^i = 11$.

Define the *generalized cluster map*

$$gc : V \to \mathcal{P}(\mathcal{P}(X))$$

by $gc(v) = \{U \subseteq X$: there is an inheritance under Relaxed Accumulation such that $v = u^i$ and $U = E(i)\}$. More precisely, let $v \in V$. Suppose $i \in C$ and $v = u^i$. Under Relaxed Accumulation, the full expression set $FE(i) \subseteq V$ satisfies that $E(i) = FE(i) \cap X$ and $E(i) \in gc(v)$. All members of $gc(v)$ are obtained in this manner.

In Figure 3 we have $gc(10) = \{\{4, 5, 6, 7, 8\}, \{6, 7, 8\}\}$.

A desirable condition is that, if $U \in gc(v)$ for some $v \in V$, then v is uniquely determined. If this condition were true for all U and v, then from knowledge of N and $U = E(i)$, we could always pinpoint the unique vertex v where the character i originated, since $U \in gc(v)$. Even in the simple example of Figure 3, however, this strong condition does not hold since $U = \{3\}$ is in both $gc(13)$ and $gc(3)$; and similarly $U = \{6\}$ is in both $gc(15)$ and $gc(6)$.

Note, nevertheless, that the failures occur in a narrowly defined situation. For example, the only normal child of vertex 13 is vertex 3; and there is then ambiguity only about whether $U = \{3\}$ arises from $u^i = 3$ or $u^i = 13$. This kind of ambiguity is instantly recognizable from the geometry of the network N.

With this motivation we make the following definition:

Let u and v be vertices. Say that there is a *choice-free normal path* from u to v if there is a normal path $u = u_0, u_1, \cdots, u_k = v$ such that for $i = 1, \cdots, k$, u_i is the only normal child of u_{i-1}.

Figure 4 shows a choice-free normal path from u to v. In Figure 4 the same $E(i)$ can lie in both $gc(u)$ and $gc(v)$. In Figure 3 there is a choice-free normal path from 13 to 3 and also from 15 to 6.

The generalized cluster map gc is *generalized monotone* provided that, whenever $U \subseteq X$ appears in both $gc(u)$ and $gc(v)$ for distinct u and v then either there is a choice-free normal path from u to v, or there is a choice-free normal path from v to u.

By inspection we verify that the generalized cluster map for Figure 3 is generalized monotone. Generalized monotonicity is desirable since then, given a character that appears exactly in the members $gc(u)$ of X, the originating vertex u is uniquely determined up to the easily recognizable ambiguity associated with a choice-free normal path.

Theorem 4.1 shows that normality is equivalent to the desirable property that gc be generalized monotone.

Theorem 4.1. *Let $N = (V, A, r, X)$ be a phylogenetic network. N is normal if and only if the generalized cluster map gc is generalized monotone.*

Proof. Suppose that N is not normal. Then there exists $u \in V$ such that there is no normal path from u to X. By induction there exists $u \in V$ such that there is no normal path from u to X but from each child v of u there is a normal path from v to X. If $u \in X$, then the trivial normal path would suffice, whence $u \notin X$. If u has outdegree 0 or 1, then $u \in X$, whence u has outdegree at least two. Since there is a normal path from each child v of u to X, it follows that

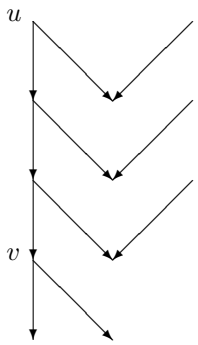

Figure 4: A choice-free normal path from u to v.

each child of u is hybrid. Let v be a child of u. Under Relaxed Accumulation, choose a character i with $u^i = v$ and with full expression set $FE(i)$. Now choose a character j with $u^j = u$ such that $u \in FE(j)$ and the only child of u in $FE(j)$ is v, and such that $FE(j) = \{u\} \cup FE(i)$. Since each child of u is hybrid, $FE(j)$ is a valid full expression set. Since $u \notin X$, we see that $X \cap FE(i) = X \cap FE(j)$, whence $U = X \cap FE(i)$ is in both $gc(u)$ and $gc(v)$, showing that gc is not generalized monotone.

Conversely, suppose that N is normal. Suppose u and v in V, and there exists $U \subseteq X$ such that $U \in gc(u)$ and $U \in gc(v)$. Say $U = FE(i) \cap X$ where $FE(i)$ is obtained as above starting at $u^i = u$, and also $U = FE(j) \cap X$ where $FE(j)$ is obtained as above starting at $u^j = v$. By normality choose a normal path from u to $x \in X$, say $u = u_0, u_1, \cdots, u_n = x$. Then each vertex on the path is in $FE(i)$, so $x \in U$ whence $x \in FE(j)$. Note that $v \leq x = u_n$. By normality, either $v = u_n$ or $v \leq u_{n-1}$; if $v \leq u_{n-1}$ then either $v = u_{n-1}$ or $v \leq u_{n-2}$. In this manner we see that either $v = u_k$ for some $k > 0$ or else $v \leq u_0 = u$. In particular, either $v \leq u$ or $u < u_k = v$.

Similarly by exchanging the roles of u and v, we have either $u \leq v$ or $v < u$ with a normal path from v to u. If $u = v$, we are done. Otherwise, by exchanging possibly the roles of u and v, we may assume there is a normal path $u = u_0, u_1, \cdots, u_m = v$ with $m > 0$.

I claim that u_1 is the only normal child of u. Otherwise, suppose b is a normal child of u, $b \neq u_1$. Choose a normal path $b = b_0, b_1, \cdots, b_k = y$ from b to $y \in X$ by normality. Then $y \in FE(i)$, whence $y \in FE(j)$ and $v \leq y$. If $v = y = b_k$, then $u_{m-1} = b_{k-1}$ by normality. Iterating the argument, we have either $u = b_p$ for some p (whence there is a directed cycle u to b to $b_p = u$, a contradiction), or for some p, $u_p = b$ (whence the arc (u, b) is redundant, a contradiction). Hence $v \neq b_k$, whence $v \leq b_{k-1}$. Repeating the argument we find that $v \leq b$, whence the arc (u, b) is redundant, a contradiction. Hence u_1

is the only normal child of u.

In similar manner we see that for $i = 1, \cdots, m$, u_i is the only normal child of u_{i-1}. Hence the normal path $u_0, u_1, \cdots, u_m$ is choice-free. □

The next theorem generalizes Theorem 3.3 to the Relaxed Accumulation model to show that mrca(U) often exists in a normal network.

Theorem 4.2. *Assume that $N = (V, A, r, X)$ is a normal network. For all $v \in V$, if $U \in gc(v)$, then mrca(U) exists. Moreover, if $u = mrca(U)$ then there exists a choice-free normal path from v to u (possibly $u = v$).*

Proof. The proof will be by induction; we will assume the result for all children of v and prove the result for v.

If v is a leaf, then $v \in X$ and $gc(v) = \{\{v\}\}$. Trivially then $v = \text{mrca}(\{v\})$. Otherwise, v is not a leaf. If $v \in X$, then $v \in U$. Then for all $u \in U$, $v \leq u$; and moreover, if $w \in V$ and $w \leq u$ for all $u \in U$, then $w \leq v$. Thus $v = \text{mrca}(U)$. Hence we may assume $v \notin X$ whence v has outdegree at least 2.

Let $FE(i)$ be the full expression vertex set for some character i such that $v = u^i$ and such that $U = FE(i) \cap X$. By normality there is a normal path $v = v_0, v_1, \cdots, v_k = x$ from v to some $x \in X$. By normality note that for all j, $0 \leq j \leq k$, $v_j \in FE(i)$. Without loss of generality we may assume that $v_j \notin X$ for $j < k$. Since $v \notin X$, $k > 0$ and v_1 is a child of v. Choose maximal p such that $0 \leq p \leq k$ and for $1 \leq j \leq p$, v_j is the only child of v_{j-1}that lies in $FE(i)$. Hence if $p = 0$ then $v = v_0$ has a child b distinct from v_1 such that $b \in FE(i)$. If $p > 0$ then $v_0, v_1, \cdots, v_p$ is a choice-free normal path (since another normal child of v_{j-1} besides v_j would have to lie in $FE(i)$); and moreover, either ($p = k$, $v = v_p \in X$) or else ($p < k$, $v_p \notin X$, and v_p has another child b distinct from v_{p+1} such that $b \in FE(i)$). By the Relaxed Accumulation model, for all $u \in U$, $v_p \leq u$.

We show that $v_p = \text{mrca}(U)$.

Suppose $v_p \in X$. Then for all $u \in U$, $v_p \leq u$; moreover, $v_p \in U = FE(i) \cap X$, whence if $w \leq x$ for all $x \in U$, then $w \leq v_p$. Hence $v_p = \text{mrca}(U)$.

Thus we may assume $v_p \notin X$, whence $p < k$ and v_p has another child b distinct from v_{p+1} such that $b \in FE(i)$. By normality there is a normal path $b = b_0, b_1, \cdots, b_m = y$ from b to some $y \in X$. By normality of the paths, note that $v_j \in FE(i)$ for all $j \leq k$; and since $b \in FE(i)$, it follows $b_j \in FE(i)$ for all $j \leq m$. In particular $x \in U$ and $y \in U$.

Note that v_{p+1} and b are both children of v_p, v_{p+1} is normal, and there are normal paths from v_{p+1} to x and from b to y. By Lemma 3.8 $v_p = \text{mrca}(x, y)$. Suppose that $w \leq u$ for all $u \in U$. Then $w \leq x$ and $w \leq y$, whence $w \leq \text{mrca}(x, y) = v_p$. But for all $u \in U$, $v_p \leq u$. Hence $\text{mrca}(U) = v_p$. □

For each u and for each character i, clearly $i \in M(u)$ iff $u \in FE(i)$. Hence the genome of u is determined by data on X to the same extent that each $FE(i)$ is determined from data on X. The following theorem shows that, when N is normal, $FE(i)$ is determined by $E(i)$ and the identity of u^i. Hence the genomes are determined up to the ambiguity associated with choice-free normal paths.

Theorem 4.3. *Let $N = (V, A, r, X)$ be a normal phylogenetic network and assume the Relaxed Accumulation model. Suppose $u^i = u$ and $E(i) = U$. Then $FE(i)$ is uniquely determined. Indeed,*

$FE(i) = \{v \in V : u \leq v,$ *and there exists $x \in U$ and a normal path from v to x*$\}$.

Proof. Suppose $v \in FE(i)$. Then $i \in M(v)$, so $u^i = u \leq v$. Moreover, by normality there is a normal path from v to some $x \in X$. For each vertex y on this normal path, we have $i \in M(y)$, whence $y \in FE(i)$. In particular, $i \in M(x)$, whence $i \in FE(i) \cap X = E(i) = U$. Thus $FE(i) \subseteq \{v \in V : u \leq v,$ and there exists $x \in U$ and a normal path from v to $x\}$.

Conversely, suppose $v \in V$, $u \leq v$, and there exists $x \in U$ and a normal path $v = v_0, v_1, \cdots, v_n = x$ from v to x. I claim $v \in FE(i)$, or equivalently $i \in M(v)$. Since $x \in U = E(i)$, $i \in M(x)$. Note $u \leq x$. By Relaxed Accumulation it follows that either $u = x = v_n$ or $i \in M(v_{n-1})$. If $u = v_n$, then since $u \leq v = v_0 \leq v_n = u$, we have $n = 0$, whence $v = u$ and $v \in FE(i)$. Hence we may assume $i \in M(v_{n-1})$. If $n = 1$ then $i \in M(v)$, whence $v \in FE(i)$. Otherwise, by Relaxed Accumulation it follows that either $u = v_{n-1}$ or else $i \in M(v_{n-2})$. If $u = v_{n-1}$, then $u \leq v \leq v_{n-1} = u$ shows $v = v_{n-1}$ whence $v \in FE(i)$. Hence we may assume $i \in M(v_{n-2})$. If $n = 2$ then $i \in M(v)$ so $v \in FE(i)$. Otherwise since $u \leq v_{n-2}$ we have either $u = v_{n-2}$ or $i \in M(v_{n-3})$. In this manner we see that $v \in FE(i)$. □

5 Comparison of regular and normal networks

This section shows that the class of normal networks is very much more restricted than the class of regular networks. For a normal network, Theorem 5.1 shows that the total number of vertices is at most a quadratic function of $|X|$, and Theorem 5.2 shows that the total number of hybrid vertices is at most $|X| - 1$. By contrast, for a regular network the number of vertices and the number of hybrid vertices can be exponential in $|X|$.

Theorem 5.1. *Suppose that $N = (V, A, r, X)$ is an acyclic rooted digraph with base-set X. Assume X consists only of the leaves and the root (so there are no vertices of outdegree 1). Let $v = |V|$ and $n = |X|$.*
(1) If N is a tree, then $v \leq 2n - 2$.
(2) If N is normal, then $v \leq (n^2 - n + 2)/2$.
(3) If N is regular, then $v \leq 2^{n-1}$.
(4) If N is not regular, then v is unbounded.

Proof. (1). Let T be the unrooted tree corresponding to N. If the root r of N is a leaf of T, then T has n leaves and it is well-known that $v \leq 2n - 2$. (See, for example, [20], p. 8.) If the root r is not a leaf of T (since r has degree 2 in T), then T has $n-1$ leaves, whence it has at most $2(n-1)-2 = 2n-4$ vertices other than r, whence $v \leq 2n - 3$.

(2) For every vertex w of N not in X, by Theorem 3.9 there exist x_1 and x_2 in X such that $w = \mathrm{mrca}(x_1, x_2)$. Note that neither x_1 nor x_2 can be the root r. Hence x_1 and x_2 must be chosen from the $(n-1)$ members of X that are not the root. The number of such vertices w must then be at most $\binom{n-1}{2} = (n-1)(n-2)/2$. Hence $v \leq \binom{n-1}{2} + n = (n^2 - n + 2)/2$.

(3) Since N is regular, the cluster map $c : V \to \mathcal{P}(X)$ is one-to-one. Note $c(r) = X$, while if $u \neq r$, then $c(u)$ is a nonempty subset of $X - \{r\}$, of which there are $2^{n-1} - 1$. Hence $v \leq 2^{n-1} - 1 + 1 = 2^{n-1}$.

(4) Let $N_{h,k}$ denote the following network with k generations each of size h, together with a root r. Let the jth generation be denoted $w_1^j, w_2^j, \cdots, w_h^j$. There is an arc from r to each member w_i^1 of the first generation. If $j < k$, then there is an arc from w_i^j to w_i^{j+1} and to $w_{i+1 \mod h}^{j+1}$. There are no other arcs. Thus in each generation other than the last, two adjacent taxa form the parents of a hybrid child. Figure 5 shows $N_{3,4}$. Let X consist of the h leaves and the root, so $n = h + 1$. Thus $v = hk + 1 = (n-1)k + 1$. Since k is arbitrary, v is unbounded. □

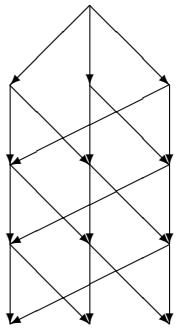

Figure 5: The network $N_{3,4}$

The upper bound in (3) is attained. Given a finite set X with a distinguished member r, let $X' = X - \{r\}$. Form $N = (V, A, R, Z)$ with vertex set V the nonempty subsets U of X' together with the vertex X. Define an arc (U, W) iff $W \subset U$ and there is no Y with $W \subset Y \subset U$. Let $R = X$. Let Z consist of all singleton subsets of X' together with R. Then N is a phylogenetic network. Note that Z may be identified with the members of X. Moreover, N is regular, with $c(U) = U$. Then $v = 2^{n-1}$. Indeed, N is essentially the cover graph for $\mathcal{P}(X')$, see [2].

It is well-known that the upper bound in (1) is attained. In fact, the upper bound in (2) is also attained. Let $X = \{1, 2, 3, 4\}$. Let $N = (V, A, r, X)$ with $r = 1$, $V = \{1, 2, 3, 4, \{2, 3, 4\}, \{2, 3\}, \{3, 4\}\}$, $A = \{(1, \{2, 3, 4\}), (\{2, 3, 4\}, \{2, 3\}),$

$(\{2,3,4\},\{3,4\})$, $(\{2,3\},2)$, $(\{2,3\},3)$, $(\{3,4\},3)$, $(\{3,4\},4)\}$. It is easily verified that N is normal, $n = 4$, $v = 7$, and the inequality in (2) is an equality. More generally, it is not hard to find an example for arbitrary n where the upper bound in (2) is attained.

Theorem 5.2. *Let $N = (V, A, r, X)$ be a normal phylogenetic network. Let H be the set of hybrid vertices. Then $|H| \leq |X| - 1$.*

Proof. (Following the proof of [9] Prop 1). Define $\phi : V - X \to V - \{r\}$ such that for all $v \in V - X$, $\phi(v)$ is a normal child of v. Since N is normal, if $v \notin X$, then v has a normal child, so ϕ can be defined. Moreover, ϕ is one-to-one since if $\phi(v) = \phi(w) = u$, then u is normal, so u has exactly one parent, which must be both v and w. Hence $|\phi(V - X)| = |V| - |X|$. No hybrid vertex lies in the image of ϕ. The root is not hybrid, nor is $r \in \phi(V - X)$. Hence $|H| \leq |V| - 1 - |\phi(V - X)| = |V| - 1 - |V| + |X| = |X| - 1$. □

By contrast, if $N = (V, A, R, Z)$ is the regular network described above to show that the upper bound in (3) is attained, then every vertex U with $|U| < n - 2$ is hybrid since there are at least two choices of $x \in X'$, $x \notin U$, such that $U \cup \{x\} \neq X'$, and each such $U \cup \{x\}$ is a parent of U. Thus the only vertices which are not hybrid are X, X', and each subset of X' with $|X'| - 1$ members. Hence $|H| = 2^{n-1} - 2 - \binom{n-1}{n-2} = 2^{n-1} - n - 1$. Thus a regular network can have exponentially many hybrid vertices.

6 Discussion

This paper concerns the class of normal phylogenetic networks. Section 3 indicated that networks should be regular if ancestral reconstruction is to be possible under the evolutionary model of Accumulation Phylogeny. Section 4 showed that the networks should be normal if ancestral reconstruction is to be possible under the Relaxed Accumulation model. Since Theorem 3.4 shows that each normal network is regular, normal networks allow reconstruction under both models.

The author expects that most more realistic models of evolution should allow both models as special cases in which the rate of substitution is sufficiently low; hence normal networks will have special importance under more realistic models. For example, suppose that one studies evolution on networks under a model similar to the Kimura 2-parameter model [17] along arcs to a normal child, but with inheritance at a hybrid vertex h with parents p and q under which some alleles are inherited from p and others from q. In the limiting case where the mutation rates are low and there are a large number of characters, it is likely that the data will closely resemble data arising from Relaxed Accumulation. (Homoplasies at normal vertices will be negligible since the probability of two substitution events at the same site will be negligible.) If the network is normal, the results of this paper suggest that one might draw inferences about the genomes at non-leaf vertices. Without normality of the network, this will not be possible, even up to the ambiguity of choice-free normal paths.

Hence, when trying to reconstruct a phylogenetic network from data, a user might first try to reconstruct a normal network if possible. Previous results by the author [24] indicate one such method with some strong assumptions in addition to normality. This current paper suggests that the assumption of normality cannot be eliminated.

Important current methods for the reconstruction of trees include parsimony and likelihood. The paper [15] generalizes likelihood methods to networks, and similarly [16] generalizes parsimony methods to networks. The problems are shown to be NP-hard, but promising heuristic methods are suggested. Such extensions are important for biologists if they are to interpret networks in a manner similarly to that in which they currently interpret trees. Since Relaxed Accumulation is a limiting case of evolution models that are utilized by likelihood methods, the current results should apply to likelihood reconstructions when mutation rates are sufficiently low. Similarly, when mutation rates are sufficiently low, there would be negligible back-mutations and parallel evolution, so the current results should apply to parsimony reconstructions as well.

There is a fast algorithm to check whether a given network $N = (V, A, r, X)$ with base-set X is normal. Recursively we define a set $S \subseteq V$ of vertices. Initially let $S = X$. Recursively, if $s \in S$ and s is normal (ie., s has indegree 1 in N), adjoin the unique parent p of s to S and check off s (so it won't be tested again). Repeat until there are no further additions, yielding a final set S. If $S = V$ then N is normal. Otherwise N is not normal, and vertices in $V - S$ have no normal path to X. Let $|X| = n$ and $|V| = v$. Then each vertex is tested at most once. If the arcs to a vertex are given as a list, then the time to check the indegree of a single vertex is constant, so the time complexity is $O(v)$. If the arcs are given by an incidence matrix then the time to check the indegree of a single vertex is $O(v)$, whence the time complexity to check normality is $O(v^2)$.

Acknowledgments

I wish to thank the Isaac Newton Institute in Cambridge UK for its hospitality in a wonderful setting while I developed this paper. I thank Mike Steel, Katharina Huber, Tandy Warnow, and Devin Bickner for helpful conversations. Finally, I thank the anonymous referees for improvements on the original version and for further references.

References

[1] H.-J. Bandelt and A. Dress, (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data, Molecular Phylogenetics and Evolution 1, 242-252.

[2] M. Baroni, C. Semple, and M. Steel, (2004), A framework for representing reticulate evolution, Annals of Combinatorics 8, 391-408.

[3] M. Baroni, C. Semple, and M. Steel, (2006), Hybrids in real time, Syst. Biol. 55, 46-56.

[4] M. Baroni and M. Steel, (2006), Accumulation phylogenies, Annals of Combinatorics 10, 19-30.

[5] M. Bordewich and C. Semple, (2007), Computing the minimum number of hybridization events for a consistent evolutionary history, Discrete Applied Mathematics 155, 914-928.

[6] G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, (2008), A distance metric for a class of tree-sibling phylogenetic networks, Bioinformatics 24, 1481-1488.

[7] G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, (2009), Metrics for phylogenetic networks I: Generalizations of the Robinson-Foulds metric, to appear in IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[8] G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, (2009), Metrics for phylogenetic networks II: Nodal and triplets metrics, to appear in IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[9] G. Cardona, F. Rosselló, and G. Valiente, (2009), Comparison of tree-child phylogenetic networks, to appear in IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[10] G. Cardona, F. Rosselló, and G. Valiente, (2008), Tripartitions do not always discriminate phylogenetic networks, Math. Biosci. 211, 356-370.

[11] D. Gusfield, S. Eddhu, and C. Langley, (2004), Optimal, efficient reconstruction of phylogenetic networks with constrained recombination, Journal of Bioinformatics and Computational Biology 2, 173-213.

[12] D. Gusfield, S. Eddhu, and C. Langley, (2004), The fine structure of galls in phylogenetic networks, INFORMS J. of Computing 16(4), 459-469.

[13] J. Hein, (1990), Reconstructing evolution of sequences subject to recombination using parsimony, Math. Biosci. 98, 185-200.

[14] L. van Iersel, J. Keijsper, S. Kelk, and L. Stougie, (2007), Constructing level-2 phylogenetic networks from triplets, arXiv:0707.2890v1 [q-bio.PE]

[15] G. Jin, L. Nakhleh, S. Snir, and T. Tuller, (2006), Maximum likelihood of phylogenetic networks, Bioinformatics 22, 2604-2611.

[16] G. Jin, L. Nakhleh, S. Snir, and T. Tuller, (2007), Efficient parsimony-based methods for phylogenetic network reconstruction, Bioinformatics 23, e123-e128.

[17] M. Kimura, (1980), A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, Journal of Molecular Evolution 16, 111-120.

[18] B.M.E. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme, (2004), Phylogenetic networks: modeling, reconstructibility, and accuracy, IEEE/ACM Transactions on Computational Biology and Bioinformatics 1, 13-23.

[19] L. Nakhleh, T. Warnow, and C.R. Linder, (2004), Reconstructing reticulate evolution in species–theory and practice, in P.E. Bourne and D. Gusfield, eds., Proceedings of the Eighth Annual International Conference on Computational Molecular Biology (RECOMB '04, March 27-31, 2004, San Diego, California), ACM, New York, 337-346.

[20] C. Semple and M. Steel, (2003), Phylogenetics, Oxford University Press, Oxford.

[21] L. Wang, K. Zhang, and L. Zhang, (2001), Perfect phylogenetic networks with recombination, Journal of Computational Biology 8, 69-78.

[22] S.J. Willson, (2007), Unique determination of some homoplasies at hybridization events, Bulletin of Mathematical Biology 69, 1709-1725.

[23] S. J. Willson, (2007), Reconstruction of some hybrid pylogenetic networks with homoplasies from distances, Bulletin of Mathematical Biology 62, 2561-2590.

[24] S.J. Willson, (2008), Reconstruction of certain phylogenetic networks from the genomes at their leaves, Journal of Theoretical Biology 252, 338-349.