Bulletin of Mathematical Biology (2007) 69: 1709–1725
DOI 10.1007/s11538-006-9187-4

ORIGINAL ARTICLE

Unique Determination of Some Homoplasies at Hybridization Events

Stephen J. Willson

Department of Mathematics, Iowa, State University, Ames, IA 50011, USA

Received: 12 May 2006 / Accepted: 7 December 2006 / Published online: 23 February 2007

Abstract Phylogenetic relationships may be represented by rooted acyclic directed graphs in which each vertex, corresponding to a taxon, possesses a genome. Assume the characters are all binary. A homoplasy occurs if a particular character changes its state more than once in the graph. A vertex is "regular" if it has only one parent and "hybrid" if it has more than one parent. A "regular path" is a directed path such that all vertices after the first are regular. Assume that the network is given and that the genomes are known for all leaves and for the root. Assume that all homoplasies occur only at hybrid vertices and each character has at most one homoplasy. Assume that from each vertex there is a regular path leading to a leaf. In this idealized setting, with other mild assumptions, it is proved that the genome at each vertex is uniquely determined. Hence, for each character the vertex at which a homoplasy occurs in the character is uniquely determined. Without the assumption on regular paths, an example shows that the genomes and homoplasies need not be uniquely determined.

Keywords Phylogeny · Network · Phylogenetic network · Hybridization · Homoplasy

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 direct genetic inheritance, typically involving genetic change such as substitutions, insertions, and deletions in the DNA. Each site of the DNA is called a "character," but there are more general notions of character that include morphological data for a taxon or even entire genes. To be of phylogenetic interest, a character should have more than one "state" or "allele" present in different taxa.

*Corresponding author.
E-mail address: swillson@iastate.edu (S. J. Willson).

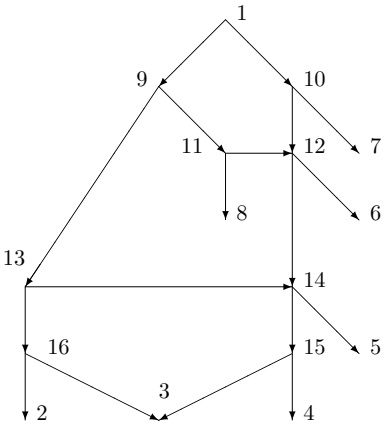

Fig. 1 A phylogenetic network with $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

For a given character, if the set of vertices with a particular state of the character is connected, the character is called "compatible" with the tree. In a "perfect phylogeny" on a tree, every character is compatible. Character compatibility is treated in Semple and Steel (2003) (pp. 65–83) and Felsenstein (2006) (pp. 87–96). A "homoplasy" in a tree is an event giving rise to incompatibility. In a tree with a perfect phylogeny, there are no homoplasies. One kind of homoplasy could arise in a rooted tree if a character has state a in taxon 1, which then changes to state b in a speciation event at taxon 2, and then some descendent taxon 3 of taxon 2 again has the character in state a. Another type of homoplasy occurs if there are two different speciation events in which the same new state of the character arises.

There has been increased interest recently in phylogenetic networks that are directed graphs but 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 (1990, 1993). Some general frameworks are found in Bandelt and Dress (1986, 1992), Baroni et al. (2004), Moret et al. (2004), and Nakhleh et al. (2004). In such a network, a "hybrid" vertex can arise which has more than one parental taxon.

For us, a phylogenetic network will be an acyclic directed graph with a set V of vertices, a set A of arcs, and a root. We assume that there are no "redundant" arcs. (See Section 2.) Let X denote the set of vertices on which the characters are known. We assume that X contains all the leaves together with the root. An example is shown in Fig. 1. The problem is then to infer the characters at all vertices of the network, given the network and the characters at each member of X.

The simplest model for reconstruction of ancestral genomes in a phylogenetic tree was given by Camin and Sokal (1965) (see also Felsenstein, 2006, pp. 73–74). This model assumes that characters are binary and that there are no homoplasies. In this case, reconstruction of the genome at each vertex in a phylogenetic tree is easy.

Wang et al. (2001) consider rooted acyclic directed graphs that need not be trees. They study a 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. (2004a) give necessary and sufficient conditions to identify these networks, which they call "galled-trees," and they add a much more specific and realistic model of recombination events. In Gusfield et al. (2004b), they give a more detailed study of these node-disjoint cycles.

Like Wang et al. (2001), this paper approaches phylogeny allowing hybridization or recombination events. Like Camin and Sokal (1965) and Wang et al. (2001), we assume that the characters are binary and that the states of the characters at all leaves of the network (as well as the root) are known. In particular, we may as well assume that all characters at the root are 0. (An outgroup can be used in place of the true root, so we may realistically assume that the genome of the root is known.) Our model does not require that the networks be galled-trees. Nor do we need the special assumptions on the hybridization process given in Gusfield et al. (2004b).

Previous work by the author in Willson (2006a,b) also assumed binary characters and allowed the possibility of hybridization. It, however, assumed that there were no homoplasies. Since homoplasies appear to be likely at hybridization events, a complete model should include that possibility. The models in Willson (2006a,b) would apply most directly to the case of polyploidy, in which the genome of both parents appears in the hybrid, so that homoplasies need not occur at the hybridization. An alternative approach to polyploidy may be found in Huber and Moulton (2006) and in Huber et al. (2006), utilizing multi-labelled trees.

The innovation in the current paper is a model allowing homoplasies at the hybrid vertices. We assume that a state can change from 0 to 1 only once in the network. We also assume that, if a state changes from 1 to 0 (thus introducing a homoplasy), such a change occurs at a unique hybridization event such that all further descendents have state 0.

A list of assumptions A1 through A4 is given at the beginning of Section 4. We make two principal assumptions in this model. Assumption A1 is that all homoplasies occur at hybrid vertices. This assumption is justified in Section 2.2. A vertex is called "regular" if it is not hybrid, and a directed path is "regular" if all vertices on it, except possibly the initial vertex, are regular. The second principal assumption (A4) is that for every vertex v there is a regular path starting at v and ending at some member of X. The path may be trivial if v is already in X. This path provides a direct connection between v and the known genome at some member of X.

We also make mild assumptions (justified in Section 3) that there are no trivial homoplasies. Assumption A2 is that any vertex of outdegree 1 lies in X. Typically this means in terms of the initial data that the only such vertex, if any exist, is the root. Without this assumption, there is no way to distinguish events at such a vertex from events at its child.

Second, consider the phylogenetic network in Fig. 1. Note that vertex 16 has children 2 and 3. An "immediate homoplasy" would occur if a character had state 0 at vertex 13, state 1 at 16, 0 at 2, and 1 at 3. The character would have mutated to state 1 in vertex 16, then immediately had a homoplasy to revert to state 0 at vertex 2. This is not distinguishable from the character having state 0 at vertices 13,

16, and 2, but state 1 at vertex 3 since the characters are known only at vertices in X. The possibility of an immediate homoplasy is instantly recognizable and hence trivial. Assumption A3 is that there are no immediate homoplasies.

The main result (Theorem 4.9) asserts that, under these assumptions, when the network is known and the genomes at the members of X are known, then the vertices at which the various mutation events occur are uniquely determined and the genomes at each vertex are uniquely determined. An example in Section 5 shows that, when even one vertex has no regular path to a member of X, the locations of the changes of state may not be determined, even when the network is itself known. The fact that the genomes are not determined even in this idealized situation when some vertex has no regular path to a member of X shows that an assumption analogous to the regular path assumption will need to be made in future more realistic analyses.

The results in this paper assume knowledge of the network. A very interesting question is the reconstruction of the network itself from knowledge of the genomes at all members of X. For the monotonic case where there are no homoplasies but with additional assumptions, a procedure was given in Willson (2006b) to perform the reconstruction even in the presence of hybridizations. The analogous reconstruction under the current hypotheses is the subject of further research.

The analysis in this paper is very sensitive to any small deviations from the idealized model and hence inappropriate for direct calculations with real data. Ongoing research is directed toward modifying the assumptions to make the calculations more relevant to real problems.

2. Fundamentals

2.1. Networks

A *directed graph* $D = (V, A)$ consists of a set V of *vertices* and a set A of *arcs*. Each arc is an ordered pair (x, y), where $x \in V$ and $y \in V$, $x \neq y$, which we visualize as directed from x to y. No multiple arcs or loops are allowed. A *directed path* σ given by $x_0, x_1, \ldots, x_n$ is a sequence of vertices $x_i \in V$ such that for $i = 1, \ldots, n$, $(x_{i-1}, x_i) \in A$. Such a directed path would be called a *directed path from* x_0 *to* x_n of *length* n. In particular, for every x_0 there is a directed path from x_0 to x_0 of length 0, called the *trivial path* at x_0. D is *acyclic* if there is no directed path $x_0, x_1, \ldots, x_n$ with $n \geq 1$ such that $x_0 = x_n$. We will assume D is acyclic; this assumption is natural since each arc (x, y) should correspond to an event requiring a positive quantity of time from x to y. D is *rooted* if there exists a *root* $r \in V$ such that for all $x \in V$ there is a directed path from r to x. Define a partial order $\leq$ on V by $a \leq b$ iff there is a directed path from a to b. In particular, for all $a \in V$, $a \leq a$ by the trivial path. We write $a < b$ if $a \leq b$ and $a \neq b$.

The arcs of D are intended to describe the most fundamental transfer of genomic information between the taxa in V. Directed paths describe transfers that are a consequence of these arcs. Accordingly, we assume that the arcs in D are *nonredundant* in the sense that no arc (x, y) exists when there is a directed path from x to y of length greater than 1. Equivalently, (x, y) is an arc iff $x \leq y$, $x \neq y$, and whenever $x \leq z \leq y$ then either $z = x$ or $z = y$.

For any vertex $v \in V$, a *parent* of v is $u \in V$ such that $(u, v) \in A$, a *child* of v is $u \in V$ such that $(v, u) \in A$, a *grandparent* of v is a parent of a parent of v, a *descendent* of v is any $u \in V$ such that $v \leq u$, and an *ancestor* of v is any $u \in V$ such that $u \leq v$. The *indegree* of v is the number of parents of v; the *outdegree* of v is the number of children of v. A vertex of outdegree 0 is called a *leaf*. The root is the only vertex of indegree 0. A vertex is *regular* if its indegree is 1 and *hybrid* if its indegree is greater than 1. We shall frequently use the following obvious fact: suppose p is the parent of c where c is regular; if $x \leq c$, then either $x = c$ or $x \leq p$. If v is a vertex, a *regular path* from v to z is a directed path $v = x_0, x_1, \ldots, x_n = z$, where for $i = 1, \ldots, n$, (x_{i-1}, x_i) is an arc and x_i is regular. Note that x_0 need not be regular. The trivial path at x_0 is also considered a regular path.

A *base set* for a rooted acyclic directed graph D is a subset X of V that includes the root r and each leaf. In typical applications, X corresponds to the set of taxa on which measurements can be made. The leaves are typically extant taxa, so that their DNA is available. While the true root is usually only inferred, in typical applications an extant outgroup is utilized to locate the root, so its DNA is also typically known.

If K is a nonempty subset of V, then a *most recent common ancestor* for K is a vertex v such that

(1) for all $k \in K$, $v \leq k$; and
(2) whenever $u \in V$ satisfies that for all $k \in K$, $u \leq k$, then it follows $u \leq v$.

If a most recent common ancestor for K exists, it is easily seen to be unique; it will be denoted $\mathrm{mrca}(K)$. It often happens that $\mathrm{mrca}(K)$ will not exist. The notation may sometimes be simplified, so we might write $\mathrm{mrca}(a, b)$ for $\mathrm{mrca}(\{a, b\})$.

A *phylogenetic network* $N = (V, A, r, X)$ is an acyclic directed graph (V, A) with root r that has base set X, such that there are no redundant arcs.

2.2. *Genomic assignments*

Let $N = (V, A, r, X)$ be a phylogenetic network. Let $C = \{1, \ldots, k\}$ denotes a finite set, called the set of *characters*. Each character $i \in C$ is assumed to be binary, possessing two *states* (0 or 1). Associated with each vertex $v \in V$ there is a binary string $G(v) = G(v)_1, G(v)_2, \ldots, G(v)_k$ called the *genome*, where $G(v)_i$ is the state (0 or 1) of character i. Label each state so that for each character i, 0 is the state in the root r. Then for each $v \in V$ define $M(v) = \{i \in C : G(v)_i = 1\}$. Call $M(v)$ the *mutated genome* at taxon v since it is the set of characters at which v has mutated from the root. Note $M(r) = \emptyset$. We assume that each $i \in C$ is *relevant* in that there exists $u \in V$ such that $i \in M(u)$. For each $i \in C$, let $K(i) = \{v \in V : i \in M(v)\}$. Assume that for each $i \in C$, $K(i)$ is nonempty; otherwise, i is irrelevant to our analysis.

A simplifying biological assumption is that changes of any character from one state to the other state are sufficiently rare that the same change would never happen twice. For each character $i \in C$, there must be a unique taxon $u_i \in V$ where i first appeared with state 1. It follows that i has state 1 in u_i and for every $v \in V$ such that $i \in M(v)$ we must have $u_i \leq v$. Hence $\mathrm{mrca}(K(i))$ exists, $u_i = \mathrm{mrca}(K(i))$, and $i \in M(u_i)$. Call u_i the *originator* for i.

For each $u \in V$, define the *originating set* to be

$$O(u) = \{i \in C : u_i = u\}.$$

Equivalently, $O(u)$ is the set of characters for which u is the originator. It is clear that the sets $O(u)$ are pairwise disjoint. Moreover, since $M(r) = \emptyset$ it follows $O(r) = \emptyset$.

Once a character i has first mutated in u_i from the state at the root, we assume that all descendents of u_i carry the mutated version of the character unless there is another taxon $x_i \in V$ where a reversion occurs so that the character changes back to the unmutated form in the root. Necessarily, $u_i < x_i$. Under the assumption that no mutation ever occurs twice, it follows that every descendent of x_i carries the version of the character i at the root since the mutated version never reoccurs in a descendent of x_i.

Thus a complete description of where the mutated character i occurs is as follows: either

(1) there exists a unique vertex u_i such that $i \in M(v)$ iff $u_i \leq v$; or
(2) there exist unique vertices u_i and x_i such that $u_i < x_i$ and

$$i \in M(v) \quad \text{iff} \quad (u_i \leq v \quad \text{and} \quad x_i \not\leq v).$$

Case (1) can be regarded as Case (2) where x_i does not exist.

Referring to Case (2), define

$$H(u, x) = \{i \in C : u = u_i \text{ and } x = x_i\}.$$

Call $H(u, x)$ the *homoplasy set* from u at x. It follows that $H(u, x) = \emptyset$ unless $u < x$. Note $H(u, x) \subseteq O(u)$. In particular, $H(r, x) = \emptyset$ since $O(r) = \emptyset$. It is now immediate that for all v, the mutated genome $M(v)$ is

$$M(v) = \cup[O(u) : u \leq v] - \cup[H(u, x) : u \leq v, x \not\leq v].$$

If x is hybrid with parents p and q, then (except in the case of polyploidy) we expect part of the genome to be inherited from p and part from q. If $i \in O(u) \cap M(p)$ but $i \notin M(q)$, then it is possible that $i \notin M(x)$ because the relevant portion of the genome was inherited from q and not p and the mutated state of the character is lost for the descendents of x. Thus we expect that frequently $H(u, x) \neq \emptyset$ when x is hybrid.

By contrast, when x is regular with unique parent p, the entire genome of x is inherited from p with some additional mutations. Suppose $u < x$ and $i \in O(u) \cap M(p)$. A reversion of character i from state 1 to state 0 in $M(x)$ would correspond to two independent mutations at the same site corresponding to i in u and in x. Under the assumption that mutations are rare, we expect such a coincidence of mutations to be unlikely. We therefore will treat such mutations as not occurring in the general case (and treat their occurrence in reality as a second-order effect or perturbation). More specifically, we shall assume (Assumption A1) that regular

vertices have no homoplasies: if $u < x$ and x is regular, then $H(u, x) = \emptyset$. This assumption resembles the assumption of monotonicity in Willson (2006a), but only applied to regular vertices.

A shortcoming of the model appears in the situation where x is hybrid with parents p and q, $u \leq p, u \leq q$ and there exists v with $i \in H(u, v), u < v \leq p, v \not\leq q$. Then $i \in M(q)$ but $i \notin M(p)$. By our assumptions on homoplasy sets, $i \notin M(x)$ since $u < v \leq x$. In the hybridization event at x, how does the genome "know" that i is to be excluded from $M(x)$ because it is not in $M(p)$, even though it is in $M(q)$? For example, consider Fig. 1. If $i \in H(9, 14)$, then it follows $i \in M(16)$, but $i \notin M(15)$. Our assumption is that $i \notin M(3)$ since $14 \leq 3$, even though the situation between parents 16 and 15 appears symmetric. Our assumption must be based on the expectation that such situations are to be treated as small corrections on the model. An interesting problem would be to extend the current results to a more general model, where in Fig. 1 we might or might not have such i in $M(3)$.

We summarize the discussion above in the notion of a "genomic assignment." A *genomic assignment* (O, H, M) on a phylogenetic network $N = (V, A, r, X)$ is a collection of sets $O(u)$ for each $u \in V$, $H(u, v)$ for each $u \in V$ and $v \in V$, and $M(u)$ for each $u \in V$ such that

(1) the sets $O(u)$ are pairwise disjoint;
(2) if $u \not< v$ then $H(u, v) = \emptyset$;
(3) for each $u \in V$ and for each v such that $u < v$, $H(u, v) \subseteq O(u)$;
(4) for each fixed $u \in V$ the sets $H(u, v)$ are pairwise disjoint;
(5) $O(r) = \emptyset$;
(6) for each $v \in V$, $M(v) = \cup[O(u) : u \leq v] - \cup[H(u, x) : u \leq v, x \not\leq v]$.

We assume that for each x in the base set X, $M(x)$ is known. This is natural since X corresponds to the (usually extant) taxa on which measurements can be made. A fundamental problem is to infer as much as possible about the phylogenetic network and its genomic assignment given only information on X.

3. Special situations

In this section, we investigate some special situations where information in the network N may not be uniquely determined by information on the base set. Each situation will lead to a simplifying assumption for our analysis.

3.1. Vertices of outdegree 1

Let $N = (V, A, r, X)$ be a phylogenetic network. Suppose that the network N is given, as well as $M(x)$ for $x \in X$.

Lemma 3.1. *Suppose v has outdegree 1 with unique child w and suppose $v \notin X$. Then the situations (1) and (2) are indistinguishable, where*

(1) $i \in O(v) - \cup[H(v, z) : v < z]$, or
(2) $i \in O(w) - \cup[H(w, z) : w < z]$.

Likewise situations (3) and (4) are indistinguishable, where
(3) $i \in H(v, z)$ *for some* z *such that* $w < z$, *or*
(4) $i \in H(w, z)$.
Finally, (5) and (6) are indistinguishable, where
(5) $i \in H(v, w)$, *or*
(6) i *is omitted from* C.

Proof: For each pair of situations $M(x)$ will be unchanged between the two situations for every $x \in X$. We see this by several cases. Note that for $x \in X$ we have $w \leq x$ iff $v \leq x$ because $v \notin X$. Moreover, in general, we have

(a) if $w \leq z$ then $v \leq z$;
(b) if $v \leq z$ and $z \neq v$, then $w \leq z$.

If $i \in M(x)$ assuming (1), then $v \leq x$, whence $w \leq x$, so $i \in M(x)$ assuming (2). The converse is symmetric. This shows that situations (1) and (2) are indistinguishable.

If $i \in M(x)$ assuming (3) then $v \leq x$ and $z \not\leq x$. Since $x \in X$ we have $w \leq x$ and $z \not\leq x$, whence $i \in M(x)$ assuming (4). The converse is similar. This shows that (3) and (4) are indistinguishable.

If (5) is assumed, then the only vertex x such that $i \in M(x)$ is $x = v$. Hence the character i with state 1 is never observed in any member of X and may be omitted from the analysis. □

It follows that if $v \notin X$ and v has outdegree 1, then all the characters in $O(v)$ can be transferred to its child w (or omitted in case (5)). Hence the situation is indistinguishable from that in which $O(v) = \emptyset$. In this situation, v can be omitted entirely from V by removing v from V, removing the arc (v, w) from A, and replacing each arc $(u, v) \in A$ by a new arc (u, w). For simplicity, in Section 4, we shall assume (Assumption A2) that no such v occurs.

3.2. Immediate homoplasies

Suppose p has exactly two children a and b. If $i \in H(p, a)$, then there is an *immediate homoplasy* in that $i \in O(p)$ but $i \notin M(a)$, so character i first mutates to state 1 in p but immediately reverts to state 0 in p's child a. In general, it will be impossible to distinguish this situation from that in which $i \in O(b)$, so i first appears in b. This is shown in the following result.

Lemma 3.2. *Assume that vertex* $p \notin X$ *and* p *has exactly two children* a *and* b. *Suppose* $i \in H(p, a)$ *in one genomic assignment* (O, H, M).

(1) If a *and* b *have no common descendent, let another genomic assignment* (O', H', M') *be the same as* (O, H, M) *except that* $i \in O'(b)$. *Then for all* $x \in X$, $M(x) = M'(x)$.
(2) If a *and* b *have a common descendent, assume*

$$c = mrca\{v \in V : a \leq v, b \leq v\} \quad \text{exists.}$$

Suppose another genomic assignment (O', H', M') *is the same as* (O, H, M) *except that* $i \in H'(b, c)$. *Then for all* $x \in X$, $M(x) = M'(x)$.

Proof: We show that $M(x) = M'(x)$ for all $x \in X$. Assume first that a and b have no common descendent. Suppose $i \in M(x)$ for some $x \in X$. Then $p \leq x$ and $a \not\leq x$. Since $p \neq x$, it follows that $b \leq x$, so $i \in M'(x)$. Conversely, suppose $i \in M'(x)$. Then $b \leq x$ whence $p \leq x$. But if $a \leq x$, then x is a common descendent of a and b, a contradiction. Hence $p \leq x$ and $a \not\leq x$ so $i \in M(x)$.

Now assume that a and b have a common descendent and c is as described. It follows $a \leq c$ and $b \leq c$ by the definition of mrca. Suppose $x \in X$ and $i \in M(x)$. Then $p \leq x$ and $a \not\leq x$. It follows $b \leq x$. Moreover, if $c \leq x$ then $a \leq c \leq x$ contradicting $a \not\leq x$. Hence $c \not\leq x$, and $i \in M'(x)$. Conversely, if $i \in M'(x)$, then $b \leq x$ and $c \not\leq x$. Hence $p \leq b \leq x$. If in addition $a \leq x$, then since $a \leq x$ and $b \leq x$ it follows $c \leq x$, a contradiction. Hence $a \not\leq x$, whence $i \in M(x)$. □

These two situations are indistinguishable in terms of the input data but in a trivial way since the ambiguity is always present and always easily identified. In Section 4, we will resolve this ambiguity by assuming (Assumption A3) that there are no immediate homoplasies, even if the mrca(a, b) does not exist.

4. Results

The goal of this section is to show that, given a phylogenetic network $N = (V, A, r, X)$ with genomic assignment (O, H, M), and given $M(x)$ for all $x \in X$, we can uniquely determine all the sets $M(u)$, $O(u)$, and $H(u, v)$.

Let $N = (V, A, r, X)$ be a phylogenetic network and(O, H, M) be a genomic assignment on N. The full results require the following assumptions, collected here for convenience:

A1. Regular vertices have no homoplasies. More specifically, if $x \in V$ is regular and $u < x$ then $H(u, x) = \emptyset$.
A2. If v has outdegree 1, then $v \in X$.
A3. Nonimmediacy of homoplasies. Suppose that the vertex $p \notin X$ has exactly two children a and b. Then $H(p, a)$ and $H(p, b)$ are both empty.
A4. For every vertex $v \in V$, there is a regular path from v to a member of X.

Assumption A1 was justified in Section 2, A2 was justified by 3.1, and A3 was justified by 3.2. Section 5 will contain an example of failure without A4.

Our first results culminate in Theorem 4.4, showing that if all $M(u)$ are known, then so are all $O(u)$ and all $H(u, v)$. For this purpose, we express $O(u)$ and $H(u, v)$ in terms of various $M(x)$.

Lemma 4.1. *Assume A1. Suppose* $a \in V$ *is regular with parent* p. *Then*

(1) $M(a) = M(p) \cup O(a)$, *where* $M(p)$ *and* $O(a)$ *are disjoint; and*
(2) $O(a) = M(a) - M(p)$.

Proof: Recall $M(a) = \cup[O(u) : u \leq a] - \cup[H(u, v) : u < v \leq a]$. Since a is regular, $H(u, a) = \emptyset$ for $u < a$. Hence

$$M(a) = O(a) \cup \bigcup[O(u) : u < a] - \cup[H(u, v) : u < v < a].$$

Since p is the unique parent of a, $u < a$ iff $u \leq p$. Hence

$$M(a) = O(a) \cup \bigcup[O(u) : u \leq p] - \cup[H(u, v) : u < v \leq p].$$

But

$$M(p) = \cup[O(u) : u \leq p] - \cup[H(u, v) : u < v \leq p],$$

whence $M(a) = O(a) \cup M(p)$. Since the $O(u)$ are pairwise disjoint, $O(a) \cap M(p) = \emptyset$. Hence $O(a) = M(a) - M(p)$. □

Lemma 4.2. *Assume A1. If z is hybrid with parents $p_1, p_2, \ldots, p_m$, then*

$$O(z) = M(z) - (M(p_1) \cup \cdots \cup M(p_m)).$$

Proof: Recall $M(z) = \cup[O(u) : u \leq z] - \cup[H(u, v) : u < v \leq z]$.

Suppose $i \in O(z)$. Then $i \notin H(u, v)$ for $u < z$, whence $i \in M(z)$. But, for $j = 1, \ldots, m$, $i \notin M(p_j)$ since otherwise $z \leq p_j$. Hence

$$O(z) \subseteq M(z) - (M(p_1) \cup ... \cup M(p_m)).$$

Conversely, suppose $i \in M(z) - (M(p_1) \cup \cdots \cup M(p_m))$. If $i \notin O(z)$ then there exists $u < z$ so $i \in O(u)$. Since $u < z$ there exists j with $u \leq p_j$. Since $i \notin M(p_j)$, there exists v with $u < v \leq p_j$ and $i \in H(u, v)$. But then $i \notin M(z)$. Hence $M(z) - (M(p_1) \cup \cdots \cup M(p_m)) \subseteq O(z)$.

The lemma follows. □

Lemma 4.3. *Assume A1. Suppose z is hybrid with parents $p_1, p_2, \ldots, p_m$. Let $u \in V$ with $u < z$. Let $K = \{j : u \leq p_j\}$. Then*

$$H(u, z) = O(u) \cap \bigcap[M(p_j) : j \in K] - M(z).$$

Proof: Since $u < z$, note K is nonempty. Suppose

$$i \in O(u) \cap \bigcap[M(p_j) : j \in K] - M(z).$$

Since $u \leq z$ but $i \notin M(z)$, there exists v with $u < v \leq z$ and $i \in H(u, v)$. For $j \in K$, since $i \in M(p_j)$, we cannot have $u < v \leq p_j$. If $j \notin K$, then $u \not\leq p_j$ so we cannot have $u < v \leq p_j$. Hence there is no parent p_j with $v \leq p_j$, whence $v = z$ and $i \in H(u, z)$.

Conversely, suppose $i \in H(u, z)$. Then $i \in O(u)$. For $j \in K$,

$$M(p_j) = \cup[O(u) : u \le p_j] - \cup[H(u, v) : u < v \le p_j].$$

Since the sets $H(u, v)$ are pairwise disjoint, and $i \in H(u, z)$, it follows that $i \notin H(u, v)$ for $u < v \le p_j$. Hence $i \in M(p_j)$. But $i \notin M(z)$. Hence

$$i \in \left(O(u) \cap \bigcap [M(p_j) : j \in K]\right) - M(z). \qquad \square$$

Theorem 4.4. *Assume A1. If $M(v)$ is known for all $v \in V$, then $O(u)$ and $H(u, v)$ are determined for all u and v.*

Proof: If v is regular with parent p, then $O(v)$ is determined by 4.1. If v is hybrid with parents $p_1, p_2, \ldots, p_m$, then $O(v)$ is determined by 4.2. If $v = r$, then $O(v) = \emptyset$. Hence $O(v)$ is determined for all $v \in V$.

Suppose u and v are given. If $u \not< v$ then $H(u, v) = \emptyset$. If $u < v$ and v is regular, then $H(u, v) = \emptyset$ since regular vertices have no homoplasies. If $u < v$ and v is hybrid with parents $p_1, p_2, \ldots, p_m$, let $K = \{j : u \le p_j\}$. Since $u < v$, K is nonempty. But then $H(u, v)$ is determined by 4.3.

The determination of $M(u)$ for all vertices u depends crucially on the use of the regular paths. Lemma 4.5 gives the simple relationship between the mutated genomes at the ends of a regular path. $\square$

Lemma 4.5. *Assume A1. Let $x_0, x_1, \ldots, x_n$ be a regular path. Then*

$$M(x_n) = M(x_0) \cup \bigcup_{i=1}^{n} O(x_i).$$

Proof: Recall $M(x_n) = \cup[O(u) : u \le x_n] - \cup[H(u, v) : u < v \le x_n]$. Since $x_1, \ldots, x_n$ are regular, if $u \le x_n$ it follows either $u \le x_0$ or $u = x_i$ for some i such that $1 \le i \le n$. Hence

$$\begin{aligned} M(x_n) &= \cup[O(u) : u \le x_0] \cup O(x_1) \cup \cdots \cup O(x_n) \\ &\quad - \cup[H(u, v) : u < v \le x_0] - \cup[H(u, x_1) : u < x_1] \\ &\quad - \cup[H(u, x_2) : u < x_2] - \cdots - \cup[H(u, x_n) : u < x_n] \\ &= \left[\bigcup [O(u) : u \le x_0] \cup O(x_1) \cup \cdots \cup O(x_n)\right] - \bigcup [H(u, v) : u < v \le x_0] \end{aligned}$$

since the regular vertices x_i for $i > 0$ have no homoplasies

$$\begin{aligned} &= \left[\bigcup [O(u) : u \le x_0] - \cup[H(u, v) : u < v \le x_0]\right] \cup O(x_1) \cup \cdots \cup O(x_n) \\ &= M(x_0) \cup O(x_1) \cup \cdots \cup O(x_n). \end{aligned}$$

The next three results give the calculation of $M(u)$ in different situations. $\square$

Lemma 4.6. *Assume A1. Suppose there are regular paths $a = x_0, x_1, \ldots, x_n$ and $a = y_0, y_1, \ldots, y_m$ where $x_1 \ne y_1$. Then $M(a) = M(x_n) \cap M(y_m)$.*

Proof: We first see that, if $x_i = y_j$, then $i = j = 0$. This is proved via several cases.

Case 1. Suppose $i = 0$. If $j > 0$ then $a = y_0 < y_j = x_0 = a$, a contradiction.
Case 2. Suppose $j = 0$. If $i > 0$ then $a = x_0 < x_i = y_0 = a$, a contradiction.
Case 3. Suppose $i > 0$ and $j > 0$. Then since x_i and y_j are regular, their parents are unique, whence $x_{i-1} = y_{j-1}$. Repeating this argument it follows that we may assume that either $i = 1$ or $j = 1$ and both $i > 0$ and $j > 0$. Without loss of generality, assume $i = 1$, $j > 0$, and $x_1 = y_j$. If $j = 1$, this contradicts $x_1 \neq y_1$. If $j > 1$ then $x_0 = a = y_0 < y_1 < y_j = x_1$ contradicts the nonredundancy of the arc (x_0, x_1). Hence Case 3 cannot occur. Thus the points $a, x_1, \ldots, x_n, y_1, \ldots, y_m$ are all distinct.

By Lemma 4.5,

$$M(x_n) = M(a) \cup \bigcup_{i=1}^{n} O(x_i)$$

and

$$M(y_m) = M(a) \cup \bigcup_{i=1}^{m} O(y_i).$$

Since the points $x_1, \ldots, x_n, y_1, \ldots, y_m$ are all distinct, the sets $O(x_1), \ldots, O(x_n)$, $O(y_1)$, $O(y_2), \ldots, O(y_m)$ are all pairwise disjoint. Hence $M(x_n) \cap M(y_m) = M(a)$. □

Lemma 4.7. *Assume A1 and A3. Assume p is regular with parent p'. Suppose p has at least two children. Suppose that the child c of p is regular and there is a regular path from c to a. Suppose that all the other children $z_1, z_2, \ldots, z_k$ of p are hybrid, and there is a regular path from z_j to w_j. Then*

$$M(p) = M(p') \cup \big(M(a) \cap (\cup_{i=1}^{k} M(w_i))\big).$$

Proof: $M(p) = \cup[O(u) : u \leq p] - \cup[H(u, v) : u < v \leq p]$.
If $i \in M(p')$ then there exists $u \leq p'$ such that $i \in O(u)$ and there is no v, $u < v \leq p'$, with $i \in H(u, v)$. Hence $u \leq p$. If $i \in H(u, v)$ for some v satisfying $u < v \leq p$, then $v = p$ since otherwise by regularity of p, $v \leq p'$; but $H(u, p) = \emptyset$ since regular vertices have no homoplasies. Hence $i \in M(p)$. Thus $M(p') \subseteq M(p)$.

Let $c = x_1, \ldots, x_n = a$ be a regular path from c to a. Then $p = x_0, x_1, \ldots, x_n$ is a regular path from p to a. By 4.5

$$M(a) = M(p) \cup O(x_1) \cup \cdots \cup O(x_n).$$

For j such that $1 \leq j \leq k$, let $z_j = y_0, y_1, \ldots, y_m = w_j$ be a regular path from z_j to w_j. Then by 4.5,

$$M(w_j) = M(z_j) \cup O(y_1) \cup \cdots \cup O(y_m).$$

As in the proof of 4.6, the vertices $x_1, x_2, \ldots, x_n, y_1, \ldots, y_m$ are distinct, so their originating sets are pairwise disjoint. Hence $M(a) \cap M(w_j) = M(p) \cap M(z_j)$. It follows that $M(a) \cap M(w_j) \subseteq M(p)$. This proves

$$M(p') \cup \bigcup_{i=1}^{k} (M(a) \cap M(w_i)) \subseteq M(p).$$

Conversely, suppose $i \in M(p)$. Then there exists $u \leq p$ with $i \in O(u)$ such that there is no v with $u < v \leq p$ and $i \in H(u, v)$.

If $u = p$, then $i \in M(a)$ since $M(p) \subseteq M(a)$ by 4.5. Moreover, if $u = p$ and there are exactly two children of p of which z_1 is one, then $i \in M(z_1)$ since there are no immediate homoplasies by Assumption A3, whence $i \in M(a) \cap M(w_1)$ by 4.5. If $u = p$ and p has more than two children so $k > 1$ then since the sets $H(p, z_j)$ are pairwise disjoint, there exists j such that $i \notin H(p, z_j)$. Hence $i \in M(z_j)$, whence by 4.5 $i \in M(w_j)$, so $i \in M(a) \cap M(w_j)$. Thus in any event if $u = p$ then

$$i \in \cup_{i=1}^{k} (M(a) \cap M(w_i)).$$

If $u \neq p$, then since p is regular with parent p', it follows $u \leq p'$. If there were v with $u < v \leq p'$ and $i \in H(u, v)$, then $u < v \leq p$ and $i \notin M(p)$, a contradiction. Hence no such v exists. It follows $i \in M(p')$. Hence

$$M(p) \subseteq M(p') \cup \bigcup_{i=1}^{k} (M(a) \cap M(z_i)).$$

The lemma follows. □

Lemma 4.8. *Assume A1 and A3. Assume p is hybrid with parents $p'_1, p'_2, \ldots, p'_m$. Suppose p has at least two children. Suppose that the child c is regular and there is a regular path from c to a. Suppose that all the other children $z_1, z_2, \ldots, z_k$ are hybrid. Suppose for $i = 1, \ldots, k$ there is a regular path from z_i to w_i. Then*

$$M(p) = \bigcup_{i=1}^{m} (M(a) \cap M(p'_i)) \cup \bigcup_{i=1}^{k} (M(a) \cap M(w_i)).$$

Proof: $M(p) = \cup[O(u) : u \leq p] - \cup[H(u, v) : u < v \leq p]$.
Let $c = c_1, c_2, \ldots, c_n = a$ be a regular path from c to a, so $p = c_0, c_1, c_2, \ldots, c_n = a$ is a regular path from p to a since c is regular. For $i = 1, \ldots, k$, let $z_i = z_{i,0}$, $z_{i,1}, \ldots, z_{i,m_i} = w_i$ be a regular path from z_i to w_i.

By 4.5,

$$M(a) = M(p) \cup O(c_1) \cup O(c_2) \cup \cdots \cup O(c_n),$$

and

$$M(w_i) = M(z_i) \cup O(z_{i,1}) \cup \cdots \cup O(z_{i,m_i}).$$

For $j = 1, \ldots, m$,

$$M(a) \cap M(p'_j) = M(p) \cap M(p'_j) \subseteq M(p)$$

since $O(c_i) \cap M(p'_j) = \emptyset$ because otherwise $c_i \leq p'_j$ and there would be a directed cycle $p'_j < p \leq c_i \leq p'_j$.

The points $z_{i,0}, z_{i,1}, \ldots, z_{i,m_i}, c_1, \ldots, c_n$ are distinct as in the proof of 4.6, so their originating sets are pairwise disjoint. Hence

$$M(a) \cap M(w_i) = M(p) \cap M(z_i) \subseteq M(p).$$

It follows that

$$\bigcup_{i=1}^{m} (M(a) \cap M(p'_i)) \cup \bigcup_{i=1}^{k} (M(a) \cap M(w_i)) \subseteq M(p).$$

Conversely, suppose $i \in M(p)$, whence $i \in M(a)$. Then there exists $u \leq p$ with $i \in O(u)$ and there is no v with $u < v \leq p$ and $i \in H(u, v)$.

If $u = p$, then $i \in M(a)$ since $M(p) \subseteq M(a)$ by 4.5. Moreover, if $u = p$ and there are exactly two children of p (so $k = 1$), then $i \in M(z_1)$ since there are no immediate homoplasies by Assumption A3, whence $i \in M(a) \cap M(z_1) = M(a) \cap M(w_1)$. If $u = p$ and p has more than two children so $k > 1$ then since the sets $H(p, z_j)$ are pairwise disjoint, there exists j such that $i \notin H(p, z_j)$. Hence $i \in M(z_j)$, whence $i \in M(a) \cap M(w_j)$. Thus in any event, if $u = p$ then

$$i \in \cup_{j=1}^{k} (M(a) \cap M(w_j)).$$

If $u \neq p$, then $u \leq p'_j$ for some j. If there were v with $u < v \leq p'_j$ and $i \in H(u, v)$, then $u < v \leq p$ and $i \notin M(p)$, a contradiction. Hence no such v exists. It follows $i \in M(p'_j)$, whence $i \in M(a) \cap M(p'_j)$. Hence

$$M(p) \subseteq \bigcup_{i=1}^{m} (M(a) \cap M(p'_i)) \cup \bigcup_{i=1}^{k} (M(a) \cap M(w_i)).$$

The result follows.

We now have the tools available for our main result. □

Theorem 4.9. *Let $N = (V, A, r, X)$ be a phylogenetic network with a genomic assignment (O, H, M). Assume A1, A2, A3, and A4. Assume $M(x)$ is known for each $x \in X$. Then for all $v \in V$, $M(v)$ is determined and $O(v)$ is determined. For all $u < v$, $H(u, v)$ is determined.*

Proof: We prove the result by induction. Assume that $M(u)$ is determined for all $u < v$. We show that $M(v)$ is determined. The base of the induction is that $M(r) = \emptyset$.

Case 1. Suppose that v has outdegree 0 or 1. Then $v \in X$, so $M(v)$ is given.

Case 2. Suppose that v has two regular children c_1 and c_2. There is a regular path from c_1 to a member a of X, and there is a regular path from c_2 to a member b of X. Hence there are two regular paths, one from v to a and one from v to b, passing through different children of v. By 4.6, $M(v) = M(a) \cap M(b)$. Since a and b are in X, $M(a)$ and $M(b)$ are known, so $M(v)$ is determined.

Case 3. Suppose $v \notin X$ and v does not have two regular children. Since there is a regular path from v to a member of X, v must have one regular child c. All the other children $z_1, \ldots, z_k$ must be hybrid. There is a regular path from c to some vertex $a \in X$, and for $j = 1, \ldots, k$ there is a regular path from z_j to a vertex $w_j \in X$.

If v is regular with parent p', then by 4.7

$$M(v) = M(p') \cup \bigcup_{i=1}^{k} (M(a) \cap M(w_i)).$$

Since $a, w_1, \ldots, w_k$ are in X, it follows $M(a), M(w_1), \ldots, M(w_k)$ are known. Since $p' < v$, $M(p')$ is known by the inductive hypothesis. Hence $M(v)$ is determined.

If v is hybrid with parents $p'_1, \ldots, p'_m$, then by 4.8

$$M(v) = \bigcup_{i=1}^{m} (M(a) \cap M(p'_i)) \cup \bigcup_{i=1}^{k} (M(a) \cap M(w_i)).$$

Since $a, w_1, \ldots, w_k$ are in X, it follows $M(a), M(w_1), \ldots, M(w_k)$ are known. Since $p'_j < v$ for all j, $M(p'_j)$ is known by the inductive hypothesis. Hence $M(v)$ is determined.

Hence $M(v)$ is determined in all cases. By induction, it follows that for all $u \in V$, $M(u)$ is determined. From 4.4 it follows that for all v, $O(v)$ is determined, and for all $u < v$, $H(u, v)$ is determined. □

5. Examples

Example 1 is shown in Fig. 1. Here $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The hybrid vertices are 12, 14, and 3. The network satisfies the hypotheses of Theorem 4.9, so all $M(u)$, $O(u)$, $H(u, v)$ are determined uniquely by $M(x)$ for $x \in X$. Note that the network is not a galled tree as described in Gusfield et al. (2004b). On the other hand, suppose (V, A) is a galled tree with root r, and suppose X consists of its root, its leaves, and all vertices of outdegree 1. Since the cycles in (V, A) are node-disjoint, it is easy to see that (V, A, r, X) is a phylogenetic network satisfying A2 and A4. Hence the

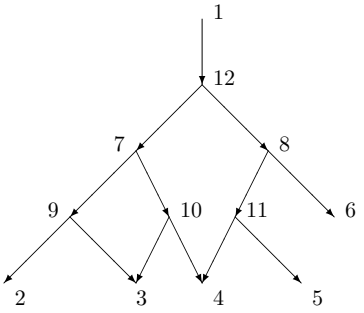

Fig. 2 The network in Example 2. $O(9)$ and $H(7, 4)$ are not determined by $M(x)$ for $x \in X$.

current results apply to any galled tree with a genomic assignment satisfying A1 and A3.

Example 2 is shown in Fig. 2. Here $X = \{1, 2, 3, 4, 5, 6\}$. This network fails the hypotheses of Theorem 4.9 because there is no regular path from 10 to X since both children 3 and 4 of 10 are hybrid. One cannot distinguish between $i \in O(9)$ and $i \in H(7, 4)$; either way the only members of X whose genomes contain i are 2 and 3. Hence $O(9)$ and $H(7, 4)$ are not determined.

Acknowledgements

I wish to thank Mike Steel for clarifying conversations and the anonymous referees for suggestions to improve the manuscript.

References

Bandelt, H.-J., Dress, A., 1986. Reconstructing the shape of a tree from observed dissimilarity data. Adv. Appl. Math. 7, 309–343.

Bandelt, H.-J., Dress, A., 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1, 242–252.

Baroni, M., Semple, C., Steel, M., 2004. A framework for representing reticulate evolution. Ann. Comb. 8, 391–408.

Camin, J.H., Sokal, R.R., 1965. A method for deducing branching sequences in phylogeny. Evolution 19, 311–326.

Felsenstein, J., 2006. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Gusfield, D., 1991. Efficient algorithms for inferring evolutionary history. Networks 21, 19–28.

Gusfield, D., Eddhu, S., Langley, C., 2004a. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. J. Bioinform. Comput. Biol. 2, 173–213.

Gusfield, D., Eddhu, S., Langley, C., 2004b. The fine structure of galls in phylogenetic networks. INFORMS J. Comput. 16, 459–469.

Hein, J., 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. 98, 185–200.

Hein, J., 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. 36, 396–405.

Huber, K.T., Moulton, V., 2006. Phylogenetic networks from multi-labelled trees. J. Math. Biol. 52, 613–632.

Huber, K.T., Oxelman, B., Lott, M., Moulton, V., 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. Mol. Biol. Evol. 23, 1784–1791.
Moret, B.M.E., Nakhleh, L., Warnow, T., Linder, C.R., Tholse, A., Padolina, A., Sun, J., Timme, R., 2004. Phylogenetic networks: Modeling, reconstructibility, and accuracy. IEEE/ACM Trans. Comput. Biol. Bioinform. 1, 13–23.
Nakhleh, L., Warnow, T., Linder, C.R., 2004. Reconstructing reticulate evolution in species–theory and practice. In: Bourne, P.E., Gusfield, D. (Eds.), Proceedings of the Eighth Annual International Conference on Computational Molecular Biology (RECOMB '04, March 27–31, 2004, San Diego, California), ACM, New York, pp. 337–346.
Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford.
Wang, L., Zhang, K., Zhang, L., 2001. Perfect phylogenetic networks with recombination. J. Comput. Biol. 8, 69–78.
Willson, S.J., 2006a. Unique solvability of certain hybrid networks from their distances. Ann. Combin. 10, 165–178.
Willson, S.J., 2006b. Unique reconstruction of tree-like phylogenetic networks from distances between leaves. Bull. Math. Biol. 68, 919–944.