Unique solvability of certain hybrid networks from their distances

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

April 29, 2005

Abstract: Phylogenetic relationships among taxa have usually been represented by rooted trees in which the leaves correspond to extant taxa and interior vertices correspond to extinct ancestral taxa. Recently, more general graphs than trees have been investigated in order to be able to represent hybridization, lateral gene transfer, and recombination events. A model is presented in which the genome at a vertex is represented by a binary string. In the presence of hybridization and the absence of convergent evolution and homoplasies, the evolution is modeled by an acyclic digraph. In general, it is shown how distances are computed in terms of the "originating weights" at vertices. An example shows that the distance between two vertices may not correspond to the sum of branch lengths on any path in the graph. If two vertices always have a most recent common ancestor, however, then distances can be measured along certain paths. Sufficient conditions are presented so that all the distances in a network are determined by the distances between leaves, including the root. In particular it is shown how to infer the originating weights at interior vertices from such information.

1 Introduction

Phylogenetic relationships among taxa have long been represented by graphs, especially rooted trees, in which the leaves correspond to extant taxa while the interior vertices correspond to ancestral, usually extinct, taxa. Typically, each edge has a branch length which summarizes the rate of substitutions in the genome between the taxa of its two end points. See, for example, Semple and Steel [14].

Recently there has been interest in utilizing graphs that are not necessarily trees. The additional edges permit a representation of such additional biological

events as hybridization, crossover or recombination, and gene transfer. Basic models of recombination were suggested by Hein [6], [7].

Given binary character information for the leaves, it is easy to find many networks that represent "perfect phylogenies", in which for each character the set of vertices with a particular value for that character is connected. Wang et al [15] consider the problem of finding a perfect phylogenetic network with recombination that has the smallest number of recombination events; the problem is natural since one expects recombination events to be very rare. Unfortunately, they show that the problem is NP-hard. They then consider 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. [2] 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 [3] they give a more detailed study of these node-disjoint cycles.

Interest in the use of graphs other than trees is also demonstrated by programs such as SplitsTree [9], T-REX [12], and Spectronet [8]. These programs have been used to visualize how the graphs describing phylogenetic relationships may deviate from trees.

In the analysis of trees, distance information rather than character information has often been found very useful. Perfect phylogenies usually do not exist for trees with real data, while distances can often be corrected plausibly and usefully (such as the corrections to raw distances given by the models of Jukes-Cantor [10], Kimura [11], or HKY [5]). Similarly, for more general graphs, distances have been found useful. Makarenkov and Legendre [13] describe an algorithm to build a connected, undirected reticulated network given distances between the taxa. The procedure starts by building a phylogenetic tree and then adds extra edges one at a time to minimize a least-squares optimization function.

Like [13], this paper directly concerns distance information. Thus, for extant taxa u and v we typically assume a distance $d(u, v)$ that measures the evolutionary change between u and v. Like Wang et al. [15] and Gusfield et al. [2] this paper models only the case of binary characters in rooted acyclic directed graphs. Unlike [15] and [2] we shall assume that the network itself is given. We then study the problem of inferring complete distance information (such as all the branch lengths) within the network.

Assume that the phylogenetic relationships are given by an acyclic directed graph in which the vertices correspond to taxa. Assume there is a distinguished vertex r, called the "root," corresponding to the most recent common ancestor of all taxa in the graph. Assume that evolution contains no "homoplasies"; this means that whenever the value of a character mutates away from the value at the root, then it never mutates back to the value at the root. Assume moreover that there is no convergent evolution; equivalently, assume that any mutation of a character occurs only once in the phylogenetic network. Then for any mutated character i there is a unique vertex u_i called the "originator" such that all taxa with the mutant form are descendents of u_i. A formalization of a

network in which this is true will be called here a "monotone marked network." The rootedness of the network will be extensively utilized, contrasting with [13]. Rootedness is an essential part of a system of biological meaning, since the root identifies the direction of time.

The principal concern in this paper will be the extent to which information about the leaves in a monotone marked network will uniquely determine information about the interior vertices and branch lengths. In particular, we shall typically assume distance information among the leaves and the root, and try to infer distance information about all the vertices in the network.

The theorems of Section 3 show that knowledge of the network and of the distance from the root r to each vertex v will determine all the distances between all vertices. The method is to relate the distances to the "originating weight" for each vertex. The Möbius Inversion Theorem provides a convenient framework for the formulas.

In Section 4, distances are computed in terms of the more familiar branch lengths. If there exists a most recent common ancestor (mrca) c of vertices u and v, then the distance $d(u, v)$ will be the sum of the branch lengths on the directed path from c to u plus the sum of the branch lengths on the directed path from c to v. Note in this case that the relevant path can be determined geometrically from the network. Unfortunately, an example shows a case in which no mrca exists and there is a pair of vertices whose distance is not the sum of branch lengths along any path. For such networks, the familiar analysis of distances in terms of branch lengths along some path will need to be abandoned. The author does not know whether such networks occur in nature.

Section 5 gives a sufficient condition such that knowledge of distances only between the leaves (including also the root) uniquely determines all the originating weights and hence all the branch lengths. In most circumstances, the true root r of the system is not known; instead, there is an outgroup r' consisting of an extant taxon which is assumed to be a child of the true root r. Under the model described in this paper, one may replace r by r' in the calculations and hence assume that distances between the leaves and the new root r' are known. Since such distance information is likely to be measurable, this result may be useful for practical applications.

2 Basic definitions

A *directed graph* or *digraph (V,E)* consists of a finite set V of vertices and an edge set E, which is a subset of $V \times V$. The edge $(u, v) \in E$ is directed from $u \in V$ to $v \in V$. There exists no edge $(v, v) \in E$ if $v \in V$. A *directed path* from v_0 to v_k is a sequence (v_0, v_1), (v_1, v_2), ..., (v_{k-1}, v_k) such that for $i = 1, 2, ..., k$, $(v_{i-1}, v_i) \in E$. The directed graph (V, E) is *acyclic* if it has no directed cycles—i.e., if there exists no directed path $(v_0, v_1), (v_1, v_2), (v_2, v_3), ..., (v_{k-1}, v_k)$ with $k \geq 1$ such that $v_0 = v_k$. The directed graph (V, E) is *rooted* if there exists a distinguished vertex $r \in V$ called the *root*, such that for each $v \in V$, $v \neq r$, there exists a directed path from r to v.

The edge $(u, v) \in E$ is *outgoing* from u and *incoming* to v. For each vertex v, the *indegree* of v is the number of edges $(u, v) \in E$ and the *outdegree* of v is the number of edges $(v, u) \in E$. The root has indegree 0. A vertex with outdegree 0 is called a *leaf*. A vertex with indegree 1 is called *regular*, while a vertex with indegree at least 2 is called *hybrid*. If $(u, v) \in E$, call u a *parent* of v and v a *child* of u.

If (V, E) is an acyclic digraph, then V has a partial order written $\leq$, defined as follows:
(1) $v \leq v$ for all $v \in V$.
(2) $u \leq v$ if there is a directed path from u to v.
Transitivity is immediate. Note that if $v \leq u$ and $u \leq v$ then $v = u$. If $u \in V$, $v \in V$ and it is false that $u \leq v$, then write $u \not\leq v$.

Let (V, E, r) be a rooted acyclic digraph. Let $A = \{0, 1\}$ be the 2-state alphabet, s be a positive integer, and let A^s denote the collection of s-tuples from A. We may regard A^s as the collection of strings of length s from the alphabet A. If $g \in A^s$, write $g = (g_1, g_2, ..., g_s)$.

We will regard A^s as an abstraction of the possible genomes for biological organisms. For the underlying model of genetic information, assume for each $v \in V$ there is a string $G(v) \in A^s$ called the *genome* of v. Each of the s positions is called a *character*, and $C = \{1, 2, ..., s\}$ denotes the set of characters. Let $M(v) = \{i : G(v)_i \neq G(r)_i\}$ be the *marker set* for v; it is the subset of C on which the genome of v differs from the genome at the root. Without loss of generality, we shall assume that, for each $i \in C$ there exists $v \in V$ such that $i \in M(v)$; otherwise, i could just be omitted from C.

Assume that each character $i \in C$ has a *weight* $w(i) > 0$ indicating some numerical property (such as the number of nucleotides in the corresponding region of a biological organism's physical genome). If M is a subset of C, then the weight of M is $w(M) = \sum[w(i) : i \in M]$.

In particular the weight $w(M(v))$ is a numerical measure of the amount of change in the genome from the root r to v. Note that $M(r) = \emptyset$ so $w(M(r)) = 0$. More generally, if $u \in V$ and $v \in V$ then the set difference
$\Delta(M(u), M(v)) = (M(u) - M(v)) \cup (M(v) - M(u))$
is the collection of characters on which $G(u)$ and $G(v)$ differ. Define
$d : V \times V \to \mathbb{R}$ by
$d(u, v) = w(\Delta(M(u), M(v)))$.
Then $d(u, v)$ measures the amount of difference in the genomes of u and v by summing the weights of positions at which $G(u)$ and $G(v)$ differ. We will call $d(u, v)$ the *(induced) distance* between u and v. Since $M(r) = \emptyset$ it follows that $d(r, v) = w(M(v))$ for every $v \in V$.

It is easy to see that d is a pseudometric on V; i.e.,
(1) $d(u, v) = 0$ if $u = v$;
(2) $d(u, v) \geq 0$ for all $u \in V, v \in V$;
(3) $d(u, v) = d(v, u)$ for all $u \in V, v \in V$;
(4) $d(t, v) \leq d(t, u) + d(u, v)$ for all $t \in V, u \in V, v \in V$.
Note that d will be a metric if the genomes $G(v)$ are distinct (i.e., if the map $G : V \to A^s$ is one-to-one). This is true because, if $u \neq v$, then because the

genomes are distinct and the alphabet is binary it follows that $M(u) \neq M(v)$, so $d(u, v) > 0$ since each $w(i) > 0$.

Let (V, E, r) be a rooted acyclic digraph. Suppose that there is a set C of characters and for each $u \in V$ there is a subset $M(u)$ of C, called the *marker set* of u, such that $M(r) = \emptyset$ and $\cup[M(u) : u \in V] = C$, together with a weight function w on C. We call (V, E, r, M, w) a *marked network*. We model the biological system by a marked network (V, E, r, M, w).

The marked network is *monotone* provided,
(1) whenever $u \leq v$ then $M(u) \subseteq M(v)$;
(2) for each $i \in C$, there exists $u_i \in V$ such that
(a) $i \in M(u_i)$, and
(b) whenever $i \in M(u)$, then $u_i \leq u$.
Call this vertex u_i the *originator* for i. Clearly, the originator u_i for i is uniquely determined.

Condition (1) implies that all the taxa in question carry as a sign of their inheritance all the changes from the root to any of their ancestors. When there is a hybridization event, (1) implies the inheritance by the hybrid of all the mutated positions of all parents. Condition (2) implies that every mutation occurred exactly once, so that there are no homoplasies and no convergent evolution. Whenever a character i changed its value, it did so only at the originator u_i, so that every vertex u with $i \in M(u)$ must be a descendent of u_i. In the context of trees, this condition is closely related to "perfect phylogeny." (See [14].) It is easy to construct examples where the converse of (1) fails, so that $M(u) \subseteq M(v)$ yet $u \not\leq v$.

Suppose that (V, E, r, M, w) is a monotone marked network containing a distinguished leaf r' called the "outgroup" such that $(r, r') \in E$. Assume that all the leaves correspond to extant taxa. There is a new monotone marked network (V', E', r', M', w) in which the root is the extant taxon r' and for which all leaves correspond to extant taxa. For example, suppose that r has exactly the two children r' and x. Define $V' = V - r$, $E' = (E - \{(r, x), (r, r')\}) \cup \{(r', x)\}$, $M'(u) = M(u) \cup M(r')$ when $u \in V'$ and $u \neq r'$, while $M'(r') = \emptyset$. It is easy to see that (V', E', r', M', w) is as claimed. For such a network, direct measurements of both the leaves and the root are possible.

3 Calculation of distances using originating weights

The most general approach to the calculation of distances between vertices relies on numbers called "originating weights," which are associated with each vertex, rather than branch lengths, which are associated with each edge.

Let (V, E, r, M, w) be a monotone marked network. Let u_i denote the originator for $i \in C$. If $v \in V$, let $H(v) = \{i \in C : v \text{ is the originator for } i\} = \{i \in C : v = u_i\}$, and let $h(v) = w(H(v))$. Call $H(v)$ the *originator set* of v and $h(v)$ the *originating weight* of v. From the definition, $i \in H(v)$ iff
(1) $i \in M(v)$; and
(2) whenever $u \in V$ and $i \in M(u)$, then $v \leq u$.

It follows that $i \in M(v)$ iff $u_i \leq v$. Note that $h(r) = 0$ since $M(r) = \emptyset$.

Theorem 3.1. *Let (V, E, r, M, w) be a monotone marked network.*
(1) For every $v \in V$, $d(r, v) = \Sigma[h(u) : u \in V, u \leq v]$.
(2) For every $v \in V$, $z \in V$,
$d(v, z) = \sum[h(u) : u \in V, u \leq v, u \not\leq z] + \sum[h(u) : u \in V, u \leq z, u \not\leq v]$.

Proof. For each $i \in C$, let u_i be the originator for i. Let $v \in V$. If $i \in M(v)$ then $u_i \leq v$ by the definition of originator; conversely, if $u_i \leq v$ then by monotonicity $i \in M(v)$. Hence
$d(r, v) = w(M(v)) = \sum[w(i) : i \in M(v)] = \sum[w(i) : u_i \leq v]$
$= \sum[\sum[w(i) : i \in C, u = u_i] : u \leq v] = \sum[w(H(u)) : u \leq v]$
$= \sum[h(u) : u \leq v]$. This proves (1).

For (2), let $v \in V, z \in V$. Then $d(v, z) = w(\Delta(M(v), M(z))$
$= w(M(v) - M(z)) + w(M(z) - M(v))$
$= \sum[w(i) : i \in M(v), i \notin M(z)] + \sum[w(i) : i \in M(z), i \notin M(v)]$
$= \sum[w(i) : u_i \leq v, u_i \not\leq z] + \sum[w(i) : u_i \leq z, u_i \not\leq v)]$
$= \sum[\sum[w(i) : i \in C, u = u_i] : u \leq v, u \not\leq z]$
$+ \sum[\sum[w(i) : i \in C, u = u_i] : u \leq z, u \not\leq v)]$
$= \sum[h(u) : u \in V, u \leq v, u \not\leq z] + \sum[h(u) : u \in V, u \leq z, u \not\leq v]$.
This proves (2). □

For $v \in V$, let $m(v) = w(M(v))$. Call $m(v)$ the *marker weight* of v. Note $m(r) = 0$ since $M(r) = \emptyset$.

Corollary 3.2. *For each $v \in V$, $m(v) = d(r, v) = \sum[h(u) : u \leq v]$.*

Proof. $m(v) = w(M(v)) = w(\Delta(M(r), M(v))) = d(r, v) = \sum[h(u) : u \leq v]$. □

Let μ be the Möbius function for $\leq$ on V (See, for example, [4] pp. 15-18, for the definition and existence theorem, which formulates a version of the Inclusion-Exclusion Principle.) For any $u \in V$, $v \in V$, $\mu(u, v)$ is an integer. The Möbius Inversion Theorem asserts that for any function f on the partially ordered set V, if $g(v) = \sum[f(u) : u \leq v]$, then $f(v) = \sum[g(u)\mu(u, v) : u \leq v]$. In particular, we obtain the following:

Theorem 3.3. *Let (V, E, r, M, w) be a monotone marked network. For every $v \in V$, $h(v) = \sum[m(u)\mu(u, v) : u \leq v]$*
where μ is the Möbius function for V.

In practice the use of the Möbius function can be avoided by an easy recursive procedure. For example, to compute h, given m, first note $h(r) = 0$ since $M(r) = \emptyset$. Assume v is a vertex such that $h(u)$ is known for each $u \in V$, $u \leq v$, $u \neq v$. Then, by Corollary 3.2, $h(v) = m(v) - \sum[h(u) : u \leq v, u \neq v]$ can be computed. This procedure can be utilized recursively until $h(v)$ is known for all $v \in V$.

Corollary 3.4. *Suppose $d(r, u)$ is known for each $u \in V$. Then for each $v \in V$, $z \in V$, $d(v, z)$ is determined.*

Proof. Note $m(u) = d(r, u)$ by Corollary 3.2. Then by Theorem 3.3, $h(v)$ is determined for each $v \in V$. Then by Theorem 3.1(2), $d(v, z)$ is determined. □

4 Calculation of distances using edges

Suppose (V, E, r, M, w) is a monotone marked network. We may visualize the digraph using edges with arrows. If (u, v) is an edge, give the edge the branch length $d(u, v) = w(\Delta(M(u), M(v)))$.

Of interest is the extent to which the distance between vertices is the sum of the branch lengths on paths (directed or undirected) between the vertices. Such an interpretation has been the usual interpretation of distances within trees. For our model, we will see that distances often can be found by adding certain branch lengths. Not all paths can be utilized, however. The determination of which branch lengths to add is entirely geometric.

The first result asserts that distances on directed paths add:

Theorem 4.1. *Let (V, E, r, M, w) be a monotone marked network. Let $u \in V$, $v \in V$ satisfy $u \leq v$. Let $(u_0, u_1), (u_1, u_2), ..., (u_{k-1}, u_k)$ be a directed path from $u = u_0$ to $v = u_k$. Then*

$$d(u, v) = \sum [d(u_{i-1}, u_i) : 1 \leq i \leq k].$$

Proof. Since the network is monotone, it follows that
$M(u_0) \subseteq M(u_1) \subseteq M(u_2) \subseteq ... \subseteq M(u_k)$.
Hence
$\Delta(M(u_0), M(u_k)) = M(u_k) - M(u_0)$
$= [M(u_1) - M(u_0)] + [M(u_2) - M(u_1)] + ... + [M(u_k) - M(u_{k-1})]$
where + indicates disjoint union of sets. It follows that
$d(u, v) = w(\Delta(M(u_0), M(u_k))) = d(u_0, u_1) + d(u_1, u_2) + ... + d(u_{k-1}, u_k)$. □

Corollary 4.2. *Let (V, E, r, M, w) be a monotone marked network. Suppose $t \in V, u \in V, v \in V$ satisfy $t \leq u \leq v$. Then $d(t, v) = d(t, u) + d(u, v)$.*

Proof. Let $(u_0, u_1), ..., (u_{k-1}, u_k)$ be a directed path from $t = u_0$ to $u = u_k$, so that $d(t, u) = \sum [d(u_{i-1}, u_i) : 1 \leq i \leq k]$. Similarly let $(u_k, u_{k+1}), ..., (u_{m-1}, u_m)$ be a directed path from $u = u_k$ to $v = u_m$, so that
$d(u, v) = \sum [d(u_{i-1}, u_i) : k + 1 \leq i \leq m]$.
Then $(u_0, u_1), ..., (u_{m-1}, u_m)$ is a directed path from t to v, and $d(t, v) = \sum [d(u_{i-1}, u_i) : 1 \leq i \leq m]$. The result follows. □

Let $u \in V, v \in V$. A *most recent common ancestor* for u and v is a vertex $c \in V$, such that
(1) $c \leq u, c \leq v$; and
(2) whenever $t \leq u$, and $t \leq$ v it follows that $t \leq c$.

Lemma 4.3. *Let $u \in V$, $v \in V$. If a most recent common ancestor for u and v exists, it is unique.*

Proof. Suppose c and c' both satisfy the condition for being most recent common ancestor for u and v. Then $c' \leq u$ and $c' \leq v$. Since c is a most recent common ancestor for u and v, it follows $c' \leq c$. Interchanging the roles of c and c', we also find $c \leq c'$. Hence $c = c'$. □

As a consequence, if there exists a most recent common ancestor for u and v, then we will unambiguously denote it by $\mathrm{mrca}(u, v)$.

Theorem 4.4. *Let (V, E, r, M, w) be a monotone marked network. Let $u \in V$, $v \in V$. If $c = mrca(u, v)$ exists, then $d(u, v) = d(c, u) + d(c, v)$. In particular, $d(u, v)$ is the length of any directed path from c to u summed with the length of any directed path from c to v.*

Proof. We show that $\Delta(M(u), M(v)) = \Delta(M(c), M(u)) + \Delta(M(c), M(v))$ where $+$ indicates disjoint union of sets. It will then follow that $d(u, v) = d(c, u) + d(c, v)$.

Since $c \leq u$ and $c \leq v$, by monotonicity it follows that $M(c) \subseteq M(u)$ and $M(c) \subseteq M(v)$. Hence $M(c) - M(u) = \emptyset$ and $M(c) - M(v) = \emptyset$. Thus we only need to show that

$$(M(u) - M(v)) + (M(v) - M(u)) = (M(u) - M(c)) + (M(v) - M(c)). \quad (1)$$

Suppose $i \in (M(u) - M(c)) \cap (M(v) - M(c))$. Let $u_i \in V$ be the originator for i. Hence $i \in M(u_i)$; moreover, since $i \in M(u)$ and $i \in M(v)$ it follows by monotonicity that $u_i \leq u$ and $u_i \leq v$. Since $c = \mathrm{mrca}(u, v)$, it follows that $u_i \leq c$. Since $i \in M(u_i)$ and by monotonicity $M(u_i) \subseteq M(c)$, it follows $i \in M(c)$. This contradicts the assumption on i, showing that no such i exists. It follows that the union on the right of (1) is disjoint. It is immediate that the union on the left of (1) is disjoint.

Suppose $i \in M(u) - M(v)$. By monotonicity, since $c \leq v$, it follows $M(c) \subseteq M(v)$. Hence $i \notin M(c)$. It follows that $i \in M(u) - M(c)$ and $i \notin M(v) - M(c)$. A similar argument shows that if $i \in M(v) - M(u)$, then $\in M(v) - M(c)$ but $i \notin M(u) - M(c)$. Hence the left side of (1) is included in the right side of (1).

Conversely, suppose $i \in M(u) - M(c)$. Let $u_i \in V$ be the originator for i. Since $i \in M(u)$, it follows $u_i \leq u$. If $i \in M(v)$, then we would also have $u_i \leq v$. Since $c = \mathrm{mrca}(u, v)$ it would follow $u_i \leq c$ whence $i \in M(c)$. This would contradict that $i \in M(u) - M(c)$. Hence we conclude that $i \notin M(v)$. Hence $i \in M(u) - M(v)$. A symmetric argument shows that if $i \in M(v) - M(c)$, then $i \in M(v) - M(u)$. This proves that the right side of (1) is included in the left side of (1), proving (1).

Since $d(c, u) = w(\Delta(M(u), M(c)) = w(M(u) - M(c))$
is the length of any directed path from c to u by Theorem 4.1, and similarly
$d(c, v) = w(\Delta(M(v), M(c)) = w(M(v) - M(c))$
is the length of any directed path from c to v, the proof is complete. □

Corollary 4.5. *Let (V, E, r, M, w) be a monotone marked network. Suppose $u \in V$, $v \in V$, $c \in V$ satisfy $c = mrca(u, v)$. Then*
(1) $d(r, c) = (d(r, u) + d(r, v) - d(u, v))/2$.

(2) $d(c,u) = (d(u,v) + d(u,r) - d(r,v))/2$.
(3) $d(c,v) = (d(v,u) + d(v,r) - d(r,u))/2$.

Proof. By Theorem 4.4 we have
(i) $d(u,v) = d(c,u) + d(c,v)$.
Since $r \leq c \leq u$, by Corollary 4.2 we have
(ii) $d(r,u) = d(r,c) + d(c,u)$; and similarly
(iii) $d(r,v) = d(r,c) + d(c,v)$.
Add (ii) and (iii), then subtract (i) to obtain
$d(r,u) + d(r,v) - d(u,v) = 2d(r,c)$,
proving (1). Similarly adding (i) and (ii) then subtracting (iii) yields
$d(u,v) + d(r,u) - d(r,v) = 2d(c,u)$, proving (2). Adding (i) and (iii) then subtracting (ii) yields (3). □

Note that if $u = v$, then $\mathrm{mrca}(u,v) = u$ and the results in Corollary 4.5 hold trivially.

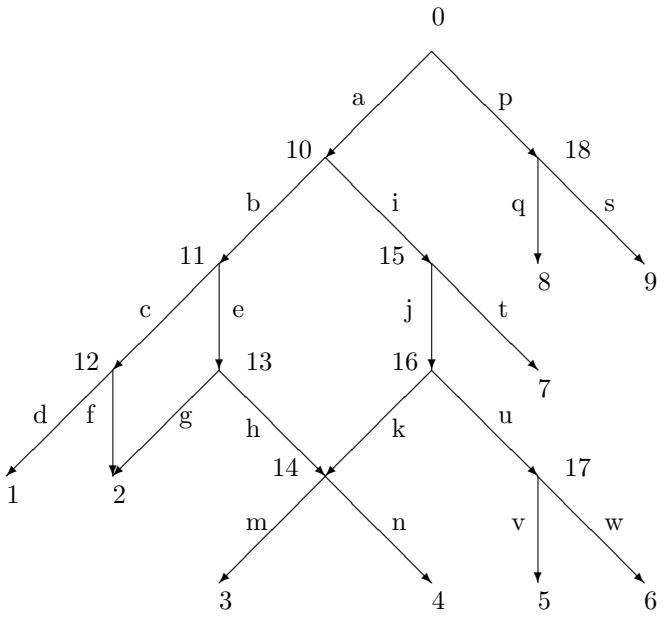

Figure 1: A network. The vertices have numbers and the branch lengths are letters. The root is 0.

Consider the directed graph shown in Figure 1. The root is 0. The leaves are 1, 2, 3, 4, 5, 6, 7, 8, and 9. There are two hybrid vertices 2 and 14. The branch length of each edge is indicated as a letter. Since there are two directed paths from 0 to 2, by Theorem 4.1 we find $d(0,2) = a + b + c + f =$

$a + b + e + g$. By inspection, $\text{mrca}(1, 13) = 11$, so by Theorem 4.4 it follows that $d(1, 13) = c + d + e$. In general, $d(1, 13) \neq d + f + g$ even though there is an (undirected) path following d, f, and g. Since $\text{mrca}(2, 5) = 10$, it follows that $d(2, 5) = b + e + g + i + j + u + v$. In this manner, for each vertex x and y, $d(x, y)$ can be written as a sum of the indicated branch lengths.

Let $X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Suppose one knew an estimate for $d(x, y)$ for all $x \in X, y \in X$ and one knew that the network in Figure 1 was correct. Then one could use least squares to estimate each indicated branch length, since each $d(x, y)$ would be expressed as a certain sum of the branch lengths.

The representation of distances as the sum of branch lengths along a path has been a useful tool in phylogenetic analysis of trees, and Theorems 4.1 and 4.4 show that often it can be used in networks as well. Figure 2, however, shows that there exist networks where no such paths can be utilized to identify distances, even in the idealized model discussed here.

In Figure 2, note that mrca(8,9) does not exist since $\{v \in V : v \leq 8, v \leq 9\} = \{0, 1, 6, 2\}$ and there is no $u \in V$ such that this set is $\{t \in V : t \leq u\}$. There is no (undirected) path connecting 8 and 9, the sum of whose branchlengths is $d(8, 9) = h(4) + h(5) + h(8) + h(9)$. To see this, note that the branch length $d(4, 8) = h(1) + h(6) + h(8)$ cannot be used since $h(6)$ does not occur in $d(8, 9)$. Similarly the branch length $d(6, 8) = h(2) + h(4) + h(8)$ cannot be used since $h(2)$ does not occur in $d(8, 9)$. Hence neither edge to vertex 8 can contribute to the desired path.

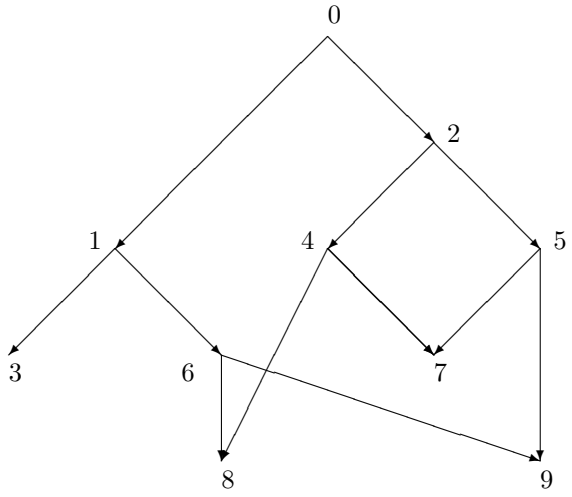

Figure 2: A network in which $d(8, 9)$ does not correspond to the length of any path from 8 to 9.

5 Unique solvability

Let (V, E, r, M, w) be a monotone marked network corresponding to a biological system. In this situation, the leaves typically correspond to extant taxa, and direct measurements are possible on these taxa, for example on their DNA. Moreover, typically the root r is determined by the choice of an appropriate outgroup, which is also an extant taxon, on which measurements can be performed. Other vertices, however, correspond to extinct ancestral taxa, and DNA measurements are usually not possible.

This situation is modelled by distinguishing a subset X of V about which extra information is available. An *X-network* is a network (V, E, r) with a subset X of V such that (1) $r \in X$; and (2) whenever $v \in V$ has outdegree less than 2, then $v \in X$. In particular, all leaves are in X. Vertices of outdegree 1 are usually indistinguishable from their children without additional information, so they are also required to lie in X. (Condition (2) is analogous to the usual assumption that vertices of trees with degree 2 lie in X; see [14].)

A *marked* X-network is a marked network (V, E, r, M, w) such that (V, E, r) is an X-network.

Let (V, E, r) be an X-network. For each $v \in V$, let $L(v) = \{x \in X : v \leq x\}$. Call $L(v)$ the *leaf-set* for v, since it tells primarily the leaves reachable from v by directed paths. The X-network is *X-injective* if the map $L : V \to 2^X$ is injective.

Theorem 5.1. *Let (V, E, r, M, w) be a monotone marked X-network. Assume (V, E, r) is X-injective. If $M(x)$ is known for each $x \in X$, then for each $v \in V$, $M(v)$ and $h(v)$ are determined.*

Proof. For $i \in C$ if there exists no $x \in X$ such that $i \in M(x)$, then there can exist no $v \in V$ with $i \in M(v)$ (since otherwise there must be a directed path from v to some leaf x, and by monotonicity it would follow that $i \in M(x)$). Since we assume that there exists $v \in V$ such that $i \in M(v)$, we may let $u_i \in V$ be the originator for i. Then for all $v \in V$, $i \in M(v)$ iff $u_i \leq v$. Let $X(i) = \{x \in X : i \in M(x)\}$. Then $X(i) = L(u_i)$. Since the network is X-injective, and $L(u_i)$ is assumed to be known, it follows that u_i is uniquely determined for each $i \in C$. In particular, for each $v \in V$, we conclude that $i \in M(v)$ iff $u_i \leq v$. Hence $M(v)$ is determined.

It follows that for each $v \in V$, $m(v) = w(M(v))$ is determined. By Theorem 3.3, $h(v)$ is determined. □

It follows that, if (V, E, r) is X-injective, then the distances $d(u, v)$ for all $u \in V$, $v \in V$ are determined once one knows $M(x)$ for all $x \in X$, as well as the weight function w. This type of information, however, is critically subject to the absence of any errors contradicting the assumption of monotonicity. For use with data, formulas involving distances would likely be more robust in the presence of errors. Hence it would be useful to have arguments depending on distances directly.

Let (V, E, r) be an X-network. Suppose M and w are given such that (V, E, r, M, w) is a monotone marked X-network, and let $d : V \times V \to \mathbb{R}$ be the induced distance function. Let $d_X : X \times X \to \mathbb{R}$ be the restriction of d. In the usual applications, M and d are not known, but d_X can be approximated. Here we assume that d_X is known exactly. The network (V, E, r) is *uniquely solvable* provided d can be uniquely determined from V, E, r, w, and d_X, without knowledge of M.

More explicitly, let (V, E, r) be an X-network. Form a matrix A, called the *weight-distance* matrix, as follows: The rows will be indexed by subsets $\{x, y\}$ of X with $x \neq y$; the columns will be indexed by $v \in V$, $v \neq r$. (Hence the number of rows of A will be $|X|(|X| - 1)/2$, and the number of columns will be $|V| - 1$.) The entry in row $\{x, y\}$ and column v, denoted $A_{\{x,y\},v}$, is defined by

$$A_{\{x,y\},v} = \begin{cases} 1 & \text{if } (v \leq x, v \not\leq y) \text{ or } (v \leq y, v \not\leq x) \\ 0 & \text{otherwise.} \end{cases}$$

Let h be the column vector whose entry with index v is $h(v)$; let d_X be the column vector whose entry with index $\{x, y\}$ is $d_X(x, y)$. Then since $h(r) = 0$, by Theorem 3.1, $Ah = d_X$. The X-network (V, E, r) is uniquely solvable iff A has nullspace $\{0\}$. Equivalently, the network is uniquely solvable iff the linear transformation taking h to Ah is one-to-one.

If (V, E, r) is uniquely solvable, then, by Theorem 3.1, for every $u \in V$, $v \in V$ a formula for $d(u, v)$ can be given in terms of the various $d_X(x, y)$ with $x \in X$ and $y \in X$. The formula will work for every monotone marked X-network (V, E, r, M, w). For such a network, all branch lengths and all originating weights are determined by the various distances $d_X(x, y)$ for $x \in X$ and $y \in X$.

If (V, E, r) is not uniquely solvable, then there is a nonzero vector h' such that $Ah' = 0$. All the values $d_X(x, y)$ for $x \in X$, $y \in X$ will not determine the values $h(v)$ for $v \in V$. I interpret this result to imply that the model is then inadequate to analyze the network (V, E, r) purely in terms of distances.

An *mrca-X-network* (V, E, r) is an X-network such that, for each $v \in V$, there exist $x \in X$ and $y \in X$ such that $v = \text{mrca}(x, y)$. A *monotone marked mrca-X-network* is a monotone marked network (V, E, r, M, w) such that (V, E, r) is an mrca-X-network.

Theorem 5.2. *Any monotone marked mrca-X-network is uniquely solvable.*

Proof. Let (V, E, r, M, w) be a monotone marked mrca-X-network. Suppose d_X is known. For each $v \in V$, let $x \in X$ and $y \in X$ be such that $v = \text{mrca}(x, y)$. By Corollary 4.5, it follows $d(r, v) = (d(r, x) + d(r, y) - d(x, y))/2$, whence, since $r \in X$, $x \in X$, $y \in X$, it follows $d(r, v) = (d_X(r, x) + d_X(r, y) - d_X(x, y))/2$. By Corollary 3.2, $m(v) = d(r, v)$ is determined for each $v \in V$ from d_X.

By Theorem 3.3 it follows that $h(v)$ is determined for each $v \in V$ from d_X. Indeed, $h(v) = \sum[\mu(u, v)d(r, u) : u \leq v]$.

From Theorem 3.1 it follows that for each $v \in V$, $z \in V$, $d(v, z)$ is determined.

For each $v \in V$ make one particular choice of $x \in X$ and $y \in X$ such that $v = \text{mrca}(x, y)$; let this choice be denoted by $x = x(v)$ and $y = y(v)$. With this notation we obtain more explicitly
$h(v) = \sum[\mu(u, v)(d_X(r, x(u)) + d_X(r, y(u)) - d_X(x(u), y(u)))/2 : u \leq v]$ and
$d(v, z) = \sum[h(u) : u \in V, u \leq v, u \not\leq z] + \sum[h(u) : u \in V, u \leq z, u \not\leq v]$. □

Example 1. The X-network in Figure 1 is uniquely solvable. Here $X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Each vertex can be written as the mrca of two members of X. For each member $x \in X$, note $x = \text{mrca}(x, x)$. For more complicated examples, note $11 = \text{mrca}(1, 3)$, $16 = \text{mrca}(3, 6)$, $14 = \text{mrca}(3, 4)$, $10 = \text{mrca}(1, 7)$. Hence if the network is a monotone marked mrca-X-network (V, E, r, M, w) with M unknown, then knowledge of d_X leads to the computation of $d(u, v)$ for all $u \in V$, $v \in V$. In particular, each branch length in Figure 1 is uniquely determined and explicitly computable.

Example 2. The X-network in Figure 2 is not uniquely solvable. The root is 0, and $X = \{0, 3, 7, 8, 9\}$. It is not an mrca-X-network since there are no $x \in X$, $y \in X$ such that $6 = \text{mrca}(x, y)$; the only possibility is $6 = \text{mrca}(8, 9)$, but $2 \leq 8$ and $2 \leq 9$ while $2 \not\leq 6$. To see that the network is not uniquely solvable, one lets $h_i = h(i)$ be an unknown for each vertex i, except that $h_0 = 0$. From Theorem 3.1, one finds the equations for $d_X(x, y)$ for all choices of $x \in X$ and $y \in X$. For example, $h_8 + h_9 + h_4 + h_5 = d_X(8, 9)$. It turns out that the relevant equations do not have a unique solution. If $h_2 - h_4 - h_5 - h_6 + h_7 + h_8 + h_9 = 0$ then $d(x, y) = 0$ for all $x \in X$, $y \in X$.

Acknowledgments

Thanks to the organizers of PCA04 in Uppsala, Sweden, and especially to Vincent Moulton, for an informative, friendly, and pleasant conference.

References

[1] D. Gusfield, Efficient algorithms for inferring evolutionary history, Networks 21 (1991)19-28.

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

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

[4] M. Hall, Jr., Combinatorial Theory, Second Edition, John Wiley & Sons, New York, 1986.

[5] M. Hasegawa, H. Kishino, and K. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, J. Mol. Evol. 22 (1985), 160-174.

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

[7] J. Hein, A heuristic method to reconstruct the history of sequences subject to recombination, J. Mol. Evol., 36 (1993) 396-405.

[8] K.T. Huber, M. Langton, D. Penny, V. Moulton, and M.Hendy, Spectronet: A package for computing spectra and median networks, Applied Bioinformatics 1(3) (2002) 159-161.

[9] D.H. Huson, SplitsTree: A program for analyzing and visualizing evolutionary data, Bioinformatics 141 (1998) 68-73.

[10] T.H. Jukes and C.R. Cantor, Evolution of protein molecules, in: S. Osawa and T. Honjo, eds., Evolution of Life: Fossils, Molecules, and Culture (Springer-Verlag, Tokyo, 1969) 79-95.

[11] M. Kimura, A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111-120.

[12] V. Makarenkov, T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks, Bioinformatics 17 (2001) 664-668.

[13] V. Makarenkov and P. Legendre, From a phylogenetic tree to a reticulated network, Journal of Computational Biology 11 (2004) 195-212.

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

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