Unique reconstruction of tree-like phylogenetic networks from distances between leaves

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

June 14, 2005

Abstract: In this paper, a class of rooted acyclic directed graphs (called TOM-networks) is defined that generalizes rooted trees and allows for models including hybridization events. It is argued that the defining properties are biologically plausible. Each TOM-network has a distance defined between each pair of vertices. For a TOM-network N, suppose that the set X consisting of the leaves and the root is known, together with the distances between members of X. It is proved that N is uniquely determined from this information and can be reconstructed in polynomial time. Thus, given exact distance information on the leaves and root, the phylogenetic network can be uniquely recovered, provided that it is a TOM-network. An outgroup can be used instead of a true root.

1 Introduction

Phylogenetic relationships have been represented by graphs ever since Darwin. Most commonly the graphs have been rooted trees in which the leaves correspond to extant taxa and internal nodes correspond to ancestral usually extinct taxa. The edges correspond to genetic transformation by mutation. As such, the edges have a branch length that quantifies the amount of genetic change.

Recently there has been considerable interest in graphs that are not necessarily trees. Such graphs may result from modeling biological events in addition to substitutions, insertions, and deletions in DNA. The extra events include hybridization, crossover or recombination, and gene transfer. Basic models of recombination were suggested by Hein [11], [12]. Some general frameworks are found in [2], [3], [18], and [19].

A common approach is to seek networks in which, for every character, the set of vertices with a particular value of that character is a connected subset of the graph. Such a graph represents a "perfect phylogeny" [5]. Networks with

this property are easy to construct when it is not required that they be trees. Since recombination events are expected to be quite rare, Wang et al. [22] consider the problem of finding a perfect phylogenetic network with recombination that has the smallest number of recombination events. They show that the problem is NP-hard, and they then consider a restricted problem in which the recombination events may be considered independent. Gusfield et al. [8], [9] make a further study of these networks, which they call "galled-trees." Baroni et al. [3] consider the problem, given a collection of rooted trees, of simultaneously displaying these trees in a single reticulated network using the minimum number of hybridization events. This current paper differs from these in that the networks need not be galled-trees, and there is no explicit minimization of the number of recombination events.

The input to the procedures in this paper will be distance information. In the analysis of trees, such information has often been found very useful. Perfect phylogenies usually do not exist for trees with real data. In contrast, raw distances from DNA can often be corrected plausibly and usefully to lead to good estimates of the amount of genomic change. Examples of such corrections are given by the models of Jukes-Cantor [14], Kimura [15], or HKY [10]. More generally they arise naturally in the computation of the likelihoods of trees for the common method of maximum likelihood [4]. Similarly, for more general graphs, distances have been found useful. The method of "split decomposition" [2], [13] utilizes distance data to find graphs indicating phylogenetic relationships. Makarenkov and Legendre [17] 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.

This paper has two main parts. The first part presents a model for a certain class of rooted directed graphs with distances (called TOM-networks) which are not necessarily trees but nevertheless have mathematical restrictions that are biologically plausible. The second part exhibits the mathematical tractability of these networks by showing that, given exact distance information among the extant taxa and the root, one may precisely reconstruct the graph, including the internal nodes and the distances between all the nodes. This latter reconstruction idealizes the basic biological problem of reconstructing evolutionary history given data available in the present, since DNA permits estimation of the evolutionary distances between extant taxa. The method is constructive and can be done in polynomial time.

Here is a more detailed overview of the principal properties of TOM-networks. An exact definition is given in Section 5.

(1) Assume that the network is a directed graph, there is a root r, and there are no directed cycles.

(2) In rooted trees for any two vertices x and y there is a uniquely determined most recent common ancestor $\mathrm{mrca}(x, y)$ such that all common ancestors of both x and y are also ancestors of $\mathrm{mrca}(x, y)$. The vertex $\mathrm{mrca}(x, y)$ is of considerable use in interpreting the tree. If, for example, $\mathrm{mrca}(x, y)$ is dated to 100 million years ago, one infers that the taxonomic groups of x and y diverged 100 million

years ago. In a TOM-network we likewise assume that for any two vertices x and y there is a uniquely determined vertex $\mathrm{mrca}(x, y)$, so that the same interpretation is possible.

(3) If ancestral species are to be inferred from extant species, they must leave a historical mark of some sort identifying some features of their genome. Since mutations involve rare random events, it is considered a suspicious coincidence if the same character derived independently more than once. Hence we assume that a character that changes once never changes again and such changes in characters leave a permanent record.

(4) If x, y, and z are distinct vertices, then by (2) they have a most recent common ancestor $\mathrm{mrca}(x, y, z)$. In trees, $\mathrm{mrca}(x, y, z)$ must be identical with either $\mathrm{mrca}(x, y)$, $\mathrm{mrca}(x, z)$, or $\mathrm{mrca}(y, z)$. We call networks with this property "tree-like" and we assume that TOM-networks are tree-like. Figure 1 shows a tree-like network that is not a tree. Any failure to be tree-like forces there to exist a vertex with outdegree 3 and highly coordinated hybridization events. (The current notion of "tree-like" differs from a notion with the same name in [1].)

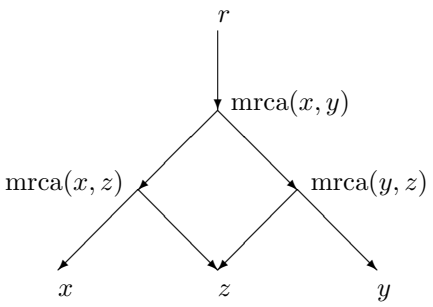

Figure 1: A tree-like network that is not a tree.

(5) Hybridization or recombination events in which two or more species merge to yield a new species are the essence of a directed graph that is not a tree. Here we assume that a hybridization event involves only two parental species, never more than two.

(6) An arc (x, y) from taxon x to taxon y should represent a direct ancestry. Thus if there is an arc (x, y) and an arc (y, z), it follows that x has already exhibited its ancestry to z and an arc (x, z) would be redundant. We assume that a TOM-network has no such redundant arcs.

(7) To each vertex x there is associated a number $h(x) \geq 0$ which quantifies the amount of genetic innovation at x. The distance $d(x, y)$ between any two taxa x and y can be expressed in terms of these quantities and forms a metric on the set of vertices.

Many more details about these networks are given in the corresponding sections.

The main theorem of the paper (Theorem 5.10) asserts roughly the following: Suppose that the evolutionary history of a system is given by an (unknown) TOM-network N. Let the set X include both the root r and all leaves of N. Suppose that the correct distances $d(x, y)$ are given for all x and y in X. (The set X corresponds to the set of extant taxa, and the distances $d(x, y)$ correspond to measurements on the extant taxa. An outgroup can be substituted for the true root r.) Then the remainder of the network can be uniquely reconstructed. Thus the internal vertices of N, their arcs, and the numbers $h(v)$ for all the vertices v are uniquely determined. Theorem 5.12 shows that the reconstruction can be done in polynomial time.

Theorem 5.10 is analogous to well-known and frequently used results under the assumption that evolutionary history is determined by trees. For example, the method of Neighbor-Joining in Saitou and Nei [20] has the analogous property: If the exact distances are given between the leaves of a tree, then the internal nodes are determined as are the distances between them. Other distance methods such as Sattath and Tversky [21], Fitch [6], and Li [16] also have analogous theorems. The innovation in the current method is that it applies to networks that are not trees.

The proof of Theorem 5.10 may be described roughly as follows: The input consists of a collection X of taxa including the root r and all leaves, and also distances $d(x, y)$ between every pair of points $x \in X$ and $y \in X$. The procedure modifies the set X as it reconstructs the network. While initially it is likely that all members of X except r are leaves of the network, under the reconstruction a modified set X may also include inferred internal vertices some of which are ancestral to other members of the same set X. We show how to distinguish leaves from such internal vertices using the distances. When a leaf x is identified such that $h(x) > 0$, we show how to compute $h(x)$ exactly and identify a single parent p of x. If $p \in X$ then this identifies the arc from p to x as well as its branch length. If $p \notin X$, then we insert a new point p into X. Formulas based on (2), (3), (4), and (7) above let us infer all distances from members of X to p. The role of x has then been completely identified, and x can be removed from X. This reduces the problem to a different set X. When a leaf is identified such that $h(x) = 0$, it must be a hybrid vertex which by (5) has exactly two parents p and q. Again, formulas permit the computation of all distances from members of X to p and to q. One now can insert p and q into X and remove x, reducing the problem once again. Since the input is assumed to arise from a network, each of these reductions leads to a smaller network, so the process ultimately terminates.

Sections 2 and 3 describe the model-building process by which the assumptions (1) through (6) above are made plausible and specific. Section 4 discusses (7) and derives formulas for the distances between taxa. These formulas are essential in the proof of Theorem 5.10 since they permit the inductive step. Finally Section 5 presents the proof of Theorem 5.10 and an analysis of the complexity of the method.

In real biological systems, the idealized system described by TOM-networks is unlikely to arise. True distances between extant taxa are usually not known exactly. One may expect, however, that good approximations may occur, especially when certain ill-behaved characters are ignored. (The identification of which characters are ill-behaved then becomes a major problem.) In such a situation, Theorem 5.10 gives hope that a historically plausible TOM-network can be found by methods similar to those used in Theorem 5.10.

The proof of Theorem 5.10 is constructive and therefore yields a procedure for reconstructing a TOM-network. The author has written computer code implementing this procedure. The correctness of the procedure relies on the assumption that the exact distances are known. The procedure works well when there are small inaccuracies in the input distances, but more research is needed to yield a procedure that is still more robust under perturbations.

2 Marked networks

A *directed graph* or *digraph* (V, A) consists of a finite set V of vertices and an arc set A, which is a subset of $V \times V$. The *arc* $(v, w) \in A$ is directed from $v \in V$ to $w \in V$. 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 $i = 1, 2, ..., k$, $(v_{i-1}, v_i) \in A$. The directed graph (V, A) 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 $v_0 = v_k$. In particular, if (V, A) is acyclic it contains no loop of form (v, v). The directed graph (V, A) 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 arc $(u, v) \in A$ is *outgoing* from u and *incoming* to v. For each vertex v, the *indegree* of v is the number of arcs $(u, v) \in A$ and the *outdegree* of v is the number of arcs $(v, u) \in A$. 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 $(u, v) \in A$, call u a *parent* of v and v a *child* of u.

If (V, A) 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) $u \leq v$ for $u \neq v$ iff there is a directed path from u to v. If $u \in V$, $v \in V$ and it is false that $u \leq v$, then write $u \not\leq v$. If $u \leq v$ and $u \neq v$, then v is a *descendent* of u.

The purpose of a graphical representation for phylogeny is to indicate the essential pattern of inheritance of genomes. The essential information of genetic inheritance is given by the partial order $\leq$, and the digraph provides a visual summary of this inheritance. The inclusion of an arc (u, v) whenever $u \leq v$, $u \neq v$, would make the digraph a confusing and redundant representation. In the interest of simplicity we therefore assume that any arc (u, v) for which there is a path from u to v other than this arc may be removed. This is formalized as follows:

An acylic digraph (V, A) is *proper* provided there is an arc from u to v if and only if (1) $u \leq v$, (2) $u \neq v$, and (3) there exists no vertex $t \in V$, $t \neq u$,

$t \neq v$, such that $u \leq t \leq v$.

Let (V, A, r) be a rooted acyclic digraph. Let $\mathcal{A} = \{0, 1\}$ be the 2-state *alphabet*, s be a positive integer, and let $\mathcal{A}^s$ denote the collection of s-tuples from $\mathcal{A}$. Thus $\mathcal{A}^s$ is the collection of strings of length s from the alphabet $\mathcal{A}$. If $g \in \mathcal{A}^s$, write $g = (g_1, g_2, ..., g_s)$. We will regard $\mathcal{A}^s$ 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 $G(v) \in \mathcal{A}^s$ called the *genome* of v, whose ith entry is $G(v)_i$. Each of the s positions is called a *character*, and $C = \{1, 2, ..., s\}$ denotes the set of characters. The collection of subsets of C will be denoted 2^C. 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. We assume that two taxa with the same genome are identical.

A *marked network* (V, A, r, C, M) consists of a proper rooted acyclic digraph (V, A, r), a set C of characters, and a function $M : V \to 2^C$ satisfying that (1) $M(r) = \emptyset$, and (2) $M(u) = M(v)$ iff $u = v$. We model the biological system by a marked network (V, A, r, C, M).

Not every subset U of C will typically occur together in some vertex $v \in V$. A nonempty subset U of C is *realized* if there exists $v \in V$ such that $U \subseteq M(v)$. The characters in a realized subset occur together at some vertex.

Evolution involves very rare events governed by randomness and natural selection that transform biological populations. In forming an idealized model of evolution, we assume that events such as changes in a particular character i will occur only once. Thus we assume in the idealized model that homoplasies (in which the same character changes more than once on a path from the root) do not occur, and that convergence (in which the same combinations of genes arise independently more than once) does not occur. We will regard such events (which of course do occur in reality) as imperfections to be dealt with at a later time after the basic framework has been determined. The assumption that characters (and combinations of characters) change only once is interpreted as assuming that for every realized combination U of characters, there is a first taxon u_U in which the combination occurred, and every taxon with the combination is a descendent to u_U. More formally we will assume the network is "monotone," as defined below:

A marked network (V, A, r, C, M) is *monotone* provided that, whenever U is a realized subset of C, then there exists $u_U \in V$ such that for all $v \in V$,

$$u_U \leq v \text{ if and only if } U \subseteq M(v).$$

Call u_U the *originator* for U. Clearly u_U is uniquely determined.

Theorem 2.1. *Let (V, A, r, C, M) be a monotone marked network. For $u \in V$ and $v \in V$, $u \leq v$ if and only if $M(u) \subseteq M(v)$.*

Proof. Suppose $u \leq v$. For each $i \in M(u)$ we know that the originator u_i satisfies $u_i \leq u$. Since $u \leq v$ it follows that $u_i \leq v$, so $i \in M(v)$. Since this is true for all $i \in M(u)$, we obtain that $M(u) \subseteq M(v)$.

Conversely, suppose $M(u) \subseteq M(v)$. Then $u_{M(u)}$ exists. Since $M(u) \subseteq M(v)$, it follows that $u_{M(u)} \leq v$ from monotonicity. I claim that $u_{M(u)} = u$. To see

this, note from monotonicity that $u_{M(u)} \leq u$. Hence by the first part of 2.1, proved above, $M(u_{M(u)}) \subseteq M(u)$. From monotonicity $M(u) \subseteq M(u_{M(u)})$. Hence $M(u) = M(u_{M(u)})$. Since both u and $u_{M(u)}$ have the same marker set, it follows that $u = u_{M(u)}$. Finally, since $u_{M(u)} \leq v$, it follows that $u \leq v$. □

Suppose U is a nonempty set of vertices. A *most recent common ancestor* of U is a vertex $c \in V$ such that (1) for all $u \in U$, $c \leq u$; and (2) whenever $t \in V$ and for all $u \in U$, $t \leq u$, it follows that $t \leq c$.

It is easy to construct marked networks in which there exist distinct $u \in V$, $v \in V$ for which no most recent common ancestor for $\{u, v\}$ exists.

Lemma 2.2. *Suppose U is a nonempty set of vertices. If a most recent common ancestor for U exists, it is unique.*

Proof. Suppose c and c' both satisfy the condition for being a most recent common ancestor for U. Then for each $u \in U$, $c' \leq u$. Since c is a most recent common ancestor for U 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, then we may unambiguously denote it by $\mathrm{mrca}(U)$. If U is a finite set such as $\{u, v, w\}$ we may shorten the notation to $\mathrm{mrca}(u, v, w)$.

Theorem 2.3. *Let (V, A, r, C, M) be a monotone marked network. Suppose U is a nonempty set of vertices. Then mrca(U) exists. Moreover, mrca$(U) = u_{\cap\{M(u):u\in U\}}$ and $M(\mathrm{mrca}(U)) = \cap\{M(u) : u \in U\}$.*

Proof. Let $V = \cap\{M(u) : u \in U\}$. By monotonicity, the originator u_V must exist. I claim that $u_V = \mathrm{mrca}(U)$. To see this, first note that for each $u \in U$, $V \subseteq M(u)$, whence by monotonicity, $u_V \leq u$. Next suppose $t \in V$ satisfies that for all $u \in U$, $t \leq u$. Then by 2.1, for all $u \in U$, $M(t) \subseteq M(u)$. It follows that $M(t) \subseteq \cap\{M(u) : u \in U\} = V \subseteq M(u_V)$. By 2.1, $t \leq u_V$. This proves that u_V satisfies the conditions for being $\mathrm{mrca}(U)$, so by 2.2, $\mathrm{mrca}(U) = u_V$.

Note that $V \subseteq M(u_V)$ by monotonicity. Conversely, suppose $i \in M(u_V)$ and $u \in U$. Then $u_i \leq u_V \leq u$, whence $i \in M(u)$. It follows that $i \in V$, whence $M(u_V) \subseteq V$. We conclude that $M(\mathrm{mrca}(U)) = M(u_V) = V$. □

In the absence of all errors, Theorem 2.3 can be used to reconstruct networks from their leaves. Let (V, A, r, C, M) be a monotone marked network. Suppose that we know the collection L of leaves (extant taxa), together with the root r and we know $M(x)$ for each $x \in L$. For each nonempty subset U of L let $v_U = \cap\{M(x) : x \in U\}$. Let $V' = \{v_U : U \text{ is a nonempty subset of } L\}$. Define the partial order $\leq'$ on V' by $v_U \leq' v_W$ iff $v_U \subseteq v_W$.

Let $V'' = \{v \in V :$ there exists a nonempty subset U of L for which $v = \mathrm{mrca}(U)\}$. Define $f : V'' \to V'$ by $f(\mathrm{mrca}(U)) = M(\mathrm{mrca}(U))$.

The set V' and the partial order $\leq'$ are easily constructed from L and $M|L$. Moreover, they carry the essential information from (V, A, r, C, M), as shown in the following result:

Theorem 2.4. *Let (V, A, r, C, M) be a monotone marked network. The map $f : V'' \to V'$ is one-to-one. Moreover, if u and v are in V'', then $u \leq v$ iff $f(u) \leq' f(v)$.*

Proof. If U and W are nonempty subsets of L such that $\mathrm{mrca}(U) \leq \mathrm{mrca}(W)$, then by 2.1, $M(\mathrm{mrca}(U)) \subseteq M(\mathrm{mrca}(W))$ so $f(\mathrm{mrca}(U)) \leq' f(\mathrm{mrca}(W))$. Conversely, if $f(\mathrm{mrca}(U)) \leq' f(\mathrm{mrca}(W))$, then $M(\mathrm{mrca}(U)) \subseteq M(\mathrm{mrca}(W))$ so, by 2.1, $\mathrm{mrca}(U) \leq \mathrm{mrca}(W)$.

To see that f is one-to-one, suppose $f(\mathrm{mrca}(U)) = f(\mathrm{mrca}(W))$. Then $\mathrm{mrca}(U) \leq \mathrm{mrca}(W)$ and $\mathrm{mrca}(W) \leq \mathrm{mrca}(U)$. It follows that $\mathrm{mrca}(U) = \mathrm{mrca}(W)$. □

In principle, Theorem 2.4 may be used to reconstruct much of the information in (V, A, r, C, M) from L and $M|L$. The procedure, however, is sensitive to even a single homoplasy. In practice, only some characters C satisfy the requirements assumed above, and there will probably be some homoplasies. It is difficult to identify these characters in advance given only the character sets at the leaves.

Consequently, the emphasis of this paper will be to try to reconstruct the networks instead by using distance information. Distances typically include corrections to estimate the number of homoplasies. If there are many characters for which the assumptions above are satisfied, it may be hoped that their signal will outweigh that of characters involved in homoplasies.

It is essential to our results that $M(r) = \emptyset$. In a typical biological setting, the true root r is an unknown ancestral taxon with unknown genome and its location is usually determined by an extant outgroup taxon r'. The following result shows that effectively one may replace $M(v)$ by $M'(v) = \{i : G(v)_i \neq G(r')_i\}$ and still obtain the essential results. Thus, r may be replaced by r' without loss of generality. In particular, we may assume that the genome of the root is known.

Lemma 2.5. *Let (V, A, r, C, M) be a monotone marked network, and for $v \in V$ let $G(v)$ denote the genome of v. Let the outgroup $r' \in V$ be a distinguished leaf with parent r. Let $M'(v) = \{i : G(v)_i \neq G(r')_i\}$. Then*
(1) For $v \in V - r'$, $M(v) \cap M(r') = \emptyset$.
(2) For $v \in V - r'$, $M'(v) = M(v) \cup M(r')$.
(3) For $u \in V - r'$ and $v \in V - r'$, $u \leq v$ if and only if $M'(u) \subseteq M'(v)$.

Proof. Let $i \in M(r')$. Then $u_i = r'$ because the parent of r' is r and $M(r) = \emptyset$. For $v \in V - r'$, $r' \not\leq v$, whence $u_i \not\leq v$, whence $i \notin M(v)$. Hence $M(r')$ is disjoint from $M(v)$ for all $v \in V - r'$, proving (1).

Let $v \in V - r'$. If $i \in M(r')$, then $G(r')_i \neq G(r)_i$ and by (1) $G(v)_i = G(r)_i$, whence $G(v)_i \neq G(r')_i$ so $i \in M'(v)$, whence $M(r') \subseteq M'(v)$. If $i \in M(v)$, then $G(v)_i \neq G(r)_i$ but by (1) $i \notin M(r')$ so $G(r')_i = G(r)_i$, whence $G(v)_i \neq G(r')_i$ so $i \in M'(v)$ and $M(v) \subseteq M'(v)$. Hence $M(v) \cup M(r') \subseteq M'(v)$. Conversely, if $i \in M'(v)$ and $i \notin M(r')$ then $G(v)_i \neq G(r')_i$ and $G(r')_i = G(r)_i$, so $i \in M(v)$. Hence $M(v) \cup M(r') = M'(v)$. This proves (2).

For (3), suppose $u \in V - r'$ and $v \in V - r'$. If $u \leq v$ then by 2.1, $M(u) \subseteq M(v)$, whence by (2) $M'(u) \subseteq M'(v)$. Conversely, suppose $M'(u) \subseteq M'(v)$. Then by (1) and (2) it follows that $M(u) \subseteq M(v)$ so $u \leq v$ by 2.1. □

3 Tree-like networks

This section concerns properties that depend only on the proper acyclic digraph (V, A, r) and not on all the structure of a marked network.

An *mrca-network* is a proper rooted acyclic digraph (V, A, r) such that whenever U is a nonempty subset of V, then mrca(U) exists. If (V, A, r, C, M) is a monotone marked network, then from 2.3 it follows that (V, A, r) is an mrca-network.

Recall that a vertex v is *hybrid* if it has more than one parent. If a network has no hybrids then it is a tree. It is common to assume that hybridization events are very rare. For example, [22] studies the problem of finding a network that has as few hybrid vertices as possible. A hybridization event in which a vertex has three parents would require an exact coordination of very rare events. The reconstruction in Section 5 will assume that such rare events don't occur. More precisely, we will assume that each hybrid vertex has exactly two parents.

An mrca-network is called *tree-like* provided, for all distinct vertices x, y, z, either mrca$(x, y, z) =$ mrca(x, y) or mrca$(x, y, z) =$ mrca(x, z) or mrca$(x, y, z) =$ mrca(y, z).

It has been common since Darwin to assume that phylogenetic relationships are described well by a directed rooted tree. A natural weakening of the notion of a tree is given by the notion of a tree-like network. We shall prove in Corollary 3.5 that directed trees are tree-like, justifying the name. Figure 1 shows a tree-like network with 3 leaves that is not a tree, while Figure 2 shows an mrca-network with 3 leaves that is not tree-like.

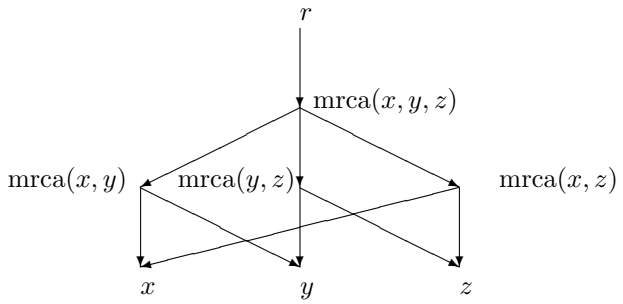

Figure 2: An mrca-network with 3 leaves x, y, and z that is not tree-like.

The network in Figure 2 contains a vertex $\mathrm{mrca}(x, y, z)$ with three children. It also has three hybrid vertices x, y, and z. Their hybridization events are delicately coordinated in that the three vertices have together only a total of three parents. If hybridization events are rare then three such delicately coordinated hybridization events must be extremely rare. It is easy to see that any mrca-network that is not tree-like must contain a vertex of outdegree 3 as in Figure 2. In Section 5 we shall ignore the possibility of such unlikely networks by assuming that the mrca-network is tree-like.

A network is *binary* if every vertex that is not a leaf has exactly two children. Since a non-tree-like network requires a vertex of outdegree at least 3, it follows that a binary mrca-network is always tree-like. The assumption that a network is tree-like is weaker than the assumption that a network is binary. Tree-like networks need not be galled trees in the sense of [8]. For example, the network in Figure 3 is tree-like but not galled:

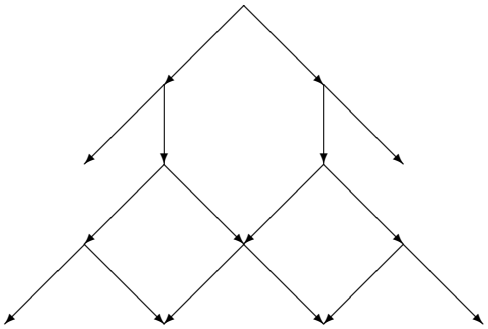

Figure 3: A tree-like network that is not a galled tree.

The next three lemmas will be used in Section 5. They are used to identify $\mathrm{mrca}(x, y)$ for certain vertices x and y.

Lemma 3.1. *Let (V, A, r) be an mrca-network. If p has distinct children x and y, then $p = \mathrm{mrca}(x, y)$.*

Proof. The proof will be by contradiction. Suppose $p \neq \mathrm{mrca}(x, y)$. Since x and y are children of p, it follows that $p \leq x$ and $p \leq y$, whence $p \leq \mathrm{mrca}(x, y)$. Since $p \neq \mathrm{mrca}(x, y)$ there is a nontrivial directed path from p to $\mathrm{mrca}(x, y)$, so the child c of p along that path satisfies $c \leq \mathrm{mrca}(x, y)$. In particular, $c \leq x$ and $c \leq y$. Hence $p \leq c \leq x$ and $p \leq c \leq y$. Since the graph is proper and $p \leq c \leq x$ either $c = p$ or $c = x$, and since c is a child of p it follows $c = x$. Analogously it follows $c = y$. Hence $x = y$, a contradiction. □

Lemma 3.2. *Let (V, A, r) be an mrca-network. Suppose p has distinct children x and y, and $y \leq z$. Then either $mrca(x, z) = p$ or $mrca(x, z) = x$.*

Proof. Note $p \leq x$ and $p \leq z$ so $p \leq \text{mrca}(x, z)$. If $p \neq \text{mrca}(x, z)$, then there is a child q of p (possibly coinciding with x or y) such that $q \leq x$ and $q \leq z$. If $q \neq x$ then $p \leq q \leq x$ with p, q, and x distinct, and since the network is proper there cannot be an arc from p to x, contradicting that x is a child of p. Hence $q = x$ and $x \leq z$. It follows $x \leq \text{mrca}(x, z) \leq x$ whence $\text{mrca}(x, z) = x$. □

Lemma 3.3. *Let (V, A, r) be an mrca-network. Let x have distinct parents p and q. Suppose y is a child of p, $y \neq x$. Then $q \not\leq y$.*

Proof. Suppose instead $q \leq y$. Since $q \leq x$ it follows $q \leq \text{mrca}(x, y)$. But $\text{mrca}(x, y) = p$ by 3.1. Hence $q \leq p \leq x$. Yet p, q, and x are distinct, so, since the digraph is proper, there is no arc from q to x. This contradicts that q is a parent of x. □

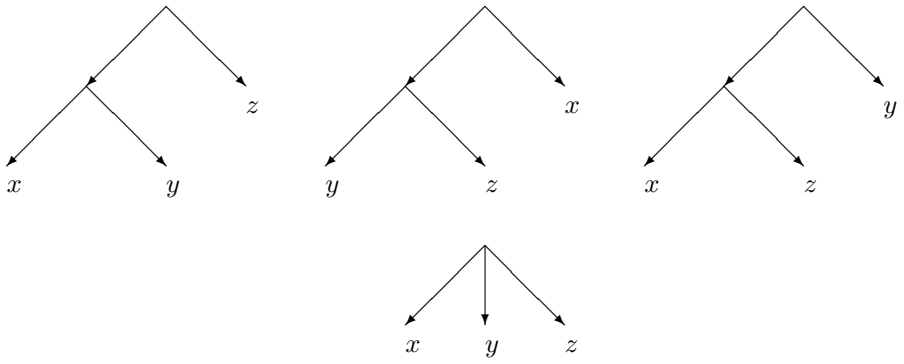

Figure 4: Possible rooted trees with leaves x, y, z.

The remainder of this section will be devoted to proving that trees are tree-like, justifying the name.

Theorem 3.4. *Let $N = (V, A, r)$ be an mrca-network. Assume N has no vertices with outdegree 1. Then N is a tree iff for all vertices x, y, and z either*
(1) $mrca(x, y, z) = mrca(x, y) = mrca(x, z)$; or
(2) $mrca(x, y, z) = mrca(x, y) = mrca(y, z)$; or
(3) $mrca(x, y, z) = mrca(x, z) = mrca(y, z)$.

Proof. Suppose N is a tree. We prove the stated condition. We may assume x, y, and z are distinct since otherwise the conclusion is trivial. Then N restricted to $\{x, y, z\}$ is either $xy|z$ or $yz|x$ or $xz|y$ or (xyz) as in Figure 4. If $yz|x$ then

$\text{mrca}(x,y,z) = \text{mrca}(x,y) = \text{mrca}(x,z)$ and (1) holds. Similarly if $xz|y$ then (2) holds and if $xy|z$ then (3) holds. Finally if (xyz) then $\text{mrca}(x,y,z) = \text{mrca}(x,y) = \text{mrca}(x,z) = \text{mrca}(y,z)$ so all of (1), (2), and (3) hold.

Conversely assume that for all x, y, z we have either (1), (2), or (3). To show that N is a tree, it suffices to show that no vertex has indegree greater than one. Suppose to the contrary that x has distinct parents p and q. Then p has outdegree at least 2, so p has a child y, $y \neq x$. Similarly, q has outdegree at least 2, so q has a child z, $z \neq x$. By 3.3, $p \not\leq z$ and $q \not\leq y$. Hence $y \neq z$. By 3.1, $\text{mrca}(x,y) = p$ and $\text{mrca}(x,z) = q$. We show that none of (1), (2), or (3) can hold. If (1) held, then $\text{mrca}(x,y,z) = p$ and $p \leq z$, contradicting 3.3. If (2) held then $\text{mrca}(x,y,z) = p$ and $p \leq z$, contradicting 3.3. If (3) held, then $\text{mrca}(x,y,z) = q$ and $q \leq y$, contradicting 3.3. Hence no such vertex x can occur. □

Corollary 3.5. *Let (V,A,r) be an mrca-network that is a tree. Then it is tree-like.*

4 Distances in weighted networks

A *weighted network* (V,A,r,C,M,w) is a monotone marked network (V,A,r,C,M) such that each character $i \in C$ has a *weight* $w(i) \geq 0$ indicating some numerical property (such as the number of nucleotides in the corresponding region of a biological organism's physical genome). If T is a subset of C, then the *weight* of T is $w(T) = \Sigma[w(i) : i \in T]$.

In particular if $x \in V$, the weight $w(M(x))$ is a numerical measure of the amount of change in the genome from the root r to x. Note that $M(r) = \emptyset$ so $w(M(r)) = 0$. More generally, if $x \in V$ and $y \in V$ then the set difference

$$\Delta(M(x), M(y)) = (M(x) - M(y)) \cup (M(y) - M(x))$$

is the collection of characters on which $M(x)$ and $M(y)$ differ. Since all characters are binary, it is also the collection of characters on which the genomes $G(x)$ and $G(y)$ differ. Define $d : V \times V \to \mathbb{R}$ by

$$d(x,y) = w(\Delta(M(x), M(y))).$$

Then $d(x,y)$ measures the amount of difference in the genomes of x and y. We will call $d(x,y)$ the *(induced) distance* between x and y. Since $M(r) = \emptyset$ it follows that $d(r,x) = w(M(x))$ for every $x \in V$.

Let (V,A,r,C,M,w) be a weighted network. If $x \in V$, let $H(x) = \{i \in C :$ x is the originator u_i for $i\}$, and let $h(x) = w(H(x))$. Call $H(x)$ the *originating character set* of x and $h(x)$ the *originating weight* of x. From the definition, $i \in H(x)$ iff (1) $i \in M(x)$, and (2) whenever $y \in V$ and $i \in M(y)$, then $x \leq y$.

Theorem 4.1 below shows that the distance can be expressed entirely in terms of the partial order and the originating weights. The proofs of 4.1 and some other preparatory results may be found in [23].

Theorem 4.1. *Let (V, A, r, C, M, w) be a weighted network with induced distance d. For $x \in V$ let $h(x)$ be the originating weight of x. Then for all $x \in V$, $y \in V$,*

$$d(x, y) = \Sigma[h(u) : u \in V, u \leq x, u \not\leq y] + \Sigma[h(u) : u \in V, u \leq y, u \not\leq x].$$

Proof. See Theorem 3.1 of [23]. □

We abstract the essential features of a weighted network by the idea of an originating mrca-network. An *originating* mrca-network (V, A, r, h, d) is an mrca-network (V, A, r) together with a function $h : V \to \mathbb{R}$ such that (1) for every $x \in V$, $h(x) \geq 0$, (2) $h(r) = 0$, and (3) for every $x \in V$, $y \in V$,

$$d(x, y) = \Sigma[h(u) : u \in V, u \leq x, u \not\leq y] + \Sigma[h(u) : u \in V, u \leq y, u \not\leq x].$$

If (V, A, r, C, M, w) is a weighted network with induced distance d, it follows from 4.1 that (V, A, r, h, d) is an originating mrca-network.

Theorem 4.2. *Let (V, A, r, h, d) be an originating mrca-network.*
(1) For every $x \in V$, $y \in V$,
$d(x, y) = \Sigma[h(u) : u \in V, u \leq x, u \not\leq y] + \Sigma[h(u) : u \in V, u \leq y, u \not\leq x]$.
(2) If $x \leq y$ then $d(x, y) = \Sigma[h(u) : u \in V, u \leq y, u \not\leq x]$.
(3) For every $x \in V$, $d(r, x) = \Sigma[h(u) : u \in V, u \leq x]$.

Proof. See 3.1 and 3.2 of [23]. □

The next two results give conditions such that d will be a metric on V.

A *pseudometric* on V is a function $d : V \times V \to \mathbb{R}$ such that (1) $d(x, y) = 0$ if $x = y$; (2) $d(x, y) \geq 0$ for all $x \in V$, $y \in V$; (3) $d(x, y) = d(y, x)$ for all $x \in V$, $y \in V$; (4) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x \in V$, $y \in V$, $z \in V$.

Lemma 4.3. *If (V, A, r, h, d) is an originating mrca-network, then d is a pseudometric.*

Proof. All is obvious except for the triangle inequality (4). To prove this, note
$d(x, z) = \Sigma[h(u) : u \leq x, u \not\leq z] + \Sigma[h(u) : u \leq z, u \not\leq x]$
$= \Sigma[h(u) : u \leq x, u \not\leq z, u \leq y] + \Sigma[h(u) : u \leq x, u \not\leq z, u \not\leq y]$
$+\Sigma[h(u) : u \leq z, u \not\leq x, u \leq y] + \Sigma[h(u) : u \leq z, u \not\leq x, u \not\leq y]$
$\leq \Sigma[h(u) : u \not\leq z, u \leq y] + \Sigma[h(u) : u \leq x, u \not\leq y]$
$+\Sigma[h(u) : u \not\leq x, u \leq y] + \Sigma[h(u) : u \leq z, u \not\leq y]$
$= d(x, y) + d(y, z)$. □

If $h(v) > 0$ for all vertices v, then it is clear that d is a metric. We will find it convenient instead to allow the possibility that a hybrid vertex v satisfies $h(v) = 0$, and the following Theorem 4.4 will be useful:

An originating mrca-network (V, A, r, h, d) is *strict* provided
(1) $h(v) > 0$ whenever v is regular (with indegree 1);
(2) $h(r) = 0$; and
(3) if v is hybrid, then each parent p of v satisfies $h(p) > 0$.

Theorem 4.4. *Let (V, A, r, h, d) be a strict originating mrca-network. Then d is a metric on V.*

Proof. By 4.3, d is a pseudometric. There remains to show only that, if $d(x, y) = 0$, then $x = y$. Assume $d(x, y) = 0$. Then by 4.2(1),
$0 = d(x, y) = \Sigma[h(u) : u \leq x, u \not\leq y] + \Sigma[h(u) : u \leq y, u \not\leq x]$.

I claim that $x \leq y$. If not then $x \leq x$ but $x \not\leq y$. Hence x would contribute to the first term above, whence $h(x) = 0$. But if x were regular, this contradicts that $h(x) > 0$. Hence x is hybrid. Suppose p is a parent of x, so $h(p) > 0$. If $p \not\leq y$ then since $p \leq x$, p would contribute to the first term above, whence $h(p) = 0$, a contradiction. Hence $p \leq y$. Now $p \leq \text{mrca}(x, y) \leq x$, whence by properness either $\text{mrca}(x, y) = p$ or $\text{mrca}(x, y) = x$ since there is an arc from p to x. But since x is hybrid there is a second parent q of x, and by the same argument either $\text{mrca}(x, y) = q$ or $\text{mrca}(x, y) = x$. Since p, q, and x are distinct, it follows that $\text{mrca}(x, y) = x$, whence $x \leq y$.

A symmetric argument shows $y \leq x$. Hence $x = y$. □

The next result shows that the distance function behaves naturally. If $x \leq y \leq z$ then the distances add, so that $d(x, z) = d(x, y) + d(y, z)$. More generally, if $c = \text{mrca}(x, y)$, then $d(x, y)$ is the length of any directed path from c to x summed with the length of any directed path from c to y.

Theorem 4.5. *Let (V, A, r, h, d) be an originating mrca-network.*
(1) Suppose $x \in V$, $y \in V$, $z \in V$ satisfy $x \leq y \leq z$. Then
$d(x, z) = d(x, y) + d(y, z)$.
(2) Let $x \in V$, $y \in V$. If $c = mrca(x, y)$, then
(2a) $d(x, y) = d(c, x) + d(c, y)$;
(2b) $d(r, c) = (d(r, x) + d(r, y) - d(x, y))/2$;
(2c) $d(c, x) = (d(x, y) + d(x, r) - d(r, y))/2$.

Proof. See [23]. □

The proof of the main result (Theorem 5.10) will require that, given vertices x, y, and z we can infer $d(x, \text{mrca}(y, z))$ from $d(r, x)$, $d(r, y)$, $d(r, z)$, $d(x, z)$, $d(x, y)$, and $d(y, z)$. These messy formulas are developed in Lemmas 4.6, 4.7, and 4.8.

Lemma 4.6. *Let (V, A, r, h, d) be an originating mrca-network. Let x, y, z be distinct vertices.*
(1) If $mrca(x, y, z) = mrca(x, y)$ then
$d(x, mrca(y, z)) = (2d(x, y) + d(r, z) - d(y, z) - d(r, y))/2$.
(2) If $mrca(x, y, z) = mrca(y, z)$ then
$d(x, mrca(y, z)) = (2d(r, x) + d(y, z) - d(r, y) - d(r, z))/2$.

Proof. Let $c = \text{mrca}(x, y, z) = \text{mrca}(x, \text{mrca}(y, z))$. For (1) note
$d(x, \text{mrca}(y, z)) = d(x, c) + d(\text{mrca}(y, z), c)$ [by 4.5(2a)]

$= d(x, \mathrm{mrca}(x,y)) + d(\mathrm{mrca}(y,z), c)$ [since $\mathrm{mrca}(x,y,z) = \mathrm{mrca}(x,y)$]
$= d(x, \mathrm{mrca}(x,y)) + d(r,y) - d(\mathrm{mrca}(y,z), y) - d(r,c)$ [since $r \le c \le \mathrm{mrca}(y,z) \le y$, whence, using 4.5(1), $d(r,y) = d(r,c) + d(c, \mathrm{mrca}(y,z)) + d(\mathrm{mrca}(y,z), y)$]
$= d(x, \mathrm{mrca}(x,y)) + d(r,y) - d(\mathrm{mrca}(y,z), y) - d(r, \mathrm{mrca}(x,y))$
$= (d(x,y) + d(x,r) - d(r,y))/2 + d(r,y) - (d(y,z) + d(y,r) - d(r,z))/2$
$-(d(r,x) + d(r,y) - d(x,y))/2$ [by 4.5(2b)]
$= (2d(x,y) + d(r,z) - d(y,z) - d(r,y))/2$. This proves (1).

For (2), note $r \le \mathrm{mrca}(y,z) = \mathrm{mrca}(x,y,z) \le x$. Hence
$d(x, \mathrm{mrca}(y,z)) = d(r,x) - d(r, \mathrm{mrca}(y,z))$ [by 4.5(1)]
$= d(r,x) - (d(r,y) + d(r,z) - d(y,z))/2$ [by 4.5(2)]
$= (2d(r,x) + d(y,z) - d(r,y) - d(r,z))/2$.
This proves (2).

□

Lemma 4.7. *Let (V, A, r, h, d) be a strict originating mrca-network. Let x, y, z be distinct vertices. If $mrca(x,y,z) = mrca(x,y)$ then*
(i) $d(r,x) + d(y,z) \le d(r,z) + d(x,y)$
with strict inequality unless $mrca(x,y) = mrca(y,z)$, and
(ii) $d(r,y) + d(x,z) \le d(r,z) + d(x,y)$
with strict inequality unless $mrca(x,y) = mrca(x,z)$.

Proof. For (i), let $c = \mathrm{mrca}(x,y,z)$. Then
$d(r,z) + d(x,y) - d(r,x) - d(y,z) = d(r,z) + d(\mathrm{mrca}(x,y), x)$
$+d(\mathrm{mrca}(x,y), y) - d(r,x) - d(\mathrm{mrca}(y,z), y) - d(\mathrm{mrca}(y,z), z)$ [by 4.5(2a)]
$= [d(r,c) + d(c, \mathrm{mrca}(y,z)) + d(\mathrm{mrca}(y,z), z)] + d(c,x)$
$+[d(c, \mathrm{mrca}(y,z)) + d(\mathrm{mrca}(y,z), y)] - [d(r,c) + d(c,x)] - d(\mathrm{mrca}(y,z), y)$
$-d(\mathrm{mrca}(y,z), z)$ [by 4.5(1) since $\mathrm{mrca}(x,y) = c \le \mathrm{mrca}(y,z) \le z$]
$= d(r,c) + d(c, \mathrm{mrca}(y,z)) + d(\mathrm{mrca}(y,z), z) + d(c,x) + d(c, \mathrm{mrca}(y,z))$
$+d(\mathrm{mrca}(y,z), y) - d(r,c) - d(c,x) - d(\mathrm{mrca}(y,z), y) - d(\mathrm{mrca}(y,z), z)$
$= 2d(c, \mathrm{mrca}(y,z)) \ge 0$ with strict inequality unless $\mathrm{mrca}(x,y,z) = \mathrm{mrca}(y,z)$ since d is a metric by 4.4. This proves (i).

The case (ii) is symmetric to case (i). □

Lemma 4.8. *Let (V, A, r, h, d) be a strict originating mrca-network that is tree-like, and let x, y, z be distinct vertices. Then the following cases are exhaustive:*
(1) If $d(r,x) + d(y,z) \le d(r,z) + d(x,y)$ and $d(r,y) + d(x,z) \le d(r,z) + d(x,y)$ then $d(x, mrca(y,z)) = (2d(x,y) + d(r,z) - d(y,z) - d(r,y))/2$.
(2) If $d(r,z) + d(x,y) \le d(r,x) + d(y,z)$ and $d(r,y) + d(x,z) \le d(r,x) + d(y,z)$ then $d(x, mrca(y,z)) = (2d(r,x) + d(y,z) - d(r,y) - d(r,z))/2$.
(3) If $d(r,x) + d(y,z) \le d(r,y) + d(x,z)$ and $d(r,z) + d(x,y) \le d(r,y) + d(x,z)$ then $d(x, mrca(y,z)) = (2d(x,z) + d(r,y) - d(y,z) - d(r,z))/2$.

Proof. To prove (1), we break into cases.
Case a. First assume $d(r,x) + d(y,z) < d(r,z) + d(x,y)$ and $d(r,y) + d(x,z) < d(r,z) + d(x,y)$. Since N is tree-like, either $\mathrm{mrca}(x,y,z) = \mathrm{mrca}(x,y)$ or $\mathrm{mrca}(x,y,z) = \mathrm{mrca}(x,z)$ or $\mathrm{mrca}(x,y,z) = \mathrm{mrca}(y,z)$. If $\mathrm{mrca}(x,y,z) =$

mrca(x,z), then by 4.7 it would follow that $d(r,z)+d(x,y) \leq d(r,y)+d(x,z)$, a contradiction. If mrca(x,y,z) = mrca(y,z), then by 4.7 it would follow that $d(r,z)+d(x,y) \leq d(r,x)+d(y,z)$, also a contradiction. Hence mrca(x,y,z) = mrca(x,y). Now 4.6(1) applies, proving the claim.

Case b. Next assume $d(r,x)+d(y,z) = d(r,z)+d(x,y)$ and $d(r,y)+d(x,z) < d(r,z)+d(x,y)$. Since N is tree-like, either mrca(x,y,z) = mrca(x,y) or mrca(x,y,z) = mrca(x,z) or mrca(x,y,z) = mrca(y,z). By 4.7 we cannot have mrca(x,y,z) = mrca(x,z) since $d(r,y)+d(x,z) < d(r,z)+d(x,y)$. If mrca$(x,y,z)$ = mrca(x,y) the conclusion of 4.6(1) holds and we are done. If mrca(x,y,z) = mrca(y,z) then mrca(x,y,z) = mrca(y,z) = mrca(x,y) because the relevant strict inequality of 4.7 does not apply. In this situation, let c = mrca(x,y,z) = mrca(y,z) = mrca(x,y), and let e = mrca(x,z). Then
$d(r,z) = d(r,c)+d(c,e)+d(e,z)$ [by 4.5(1), since $r \leq c \leq e \leq z$],
$d(r,y) = d(r,c)+d(c,y)$ [by 4.5(1)],
$d(x,y) = d(c,y)+d(c,x)$ [by 4.5(2a)],
$d(y,z) = d(c,y)+d(c,z) = d(c,y)+d(c,e)+d(e,z)$ [by 4.5(1) and 4.5(2a)].
Hence
$(2d(x,y)+d(r,z)-d(y,z)-d(r,y))/2$
$= (2(d(c,y)+d(c,x))+(d(r,c)+d(c,e)+d(e,z))-(d(c,y)+d(c,e)+d(e,z))$
$-(d(r,c)+d(c,y)))/2 = d(c,x)$ proving the claim for Case b.

Case c. Next assume $d(r,x)+d(y,z) < d(r,z)+d(x,y)$ and $d(r,y)+d(x,z) = d(r,z)+d(x,y)$. This case is symmetric to Case b by interchanging x and y, so the proof of b applies.

Case d. Next assume $d(r,x)+d(y,z) = d(r,z)+d(x,y)$ and $d(r,y)+d(x,z) = d(r,z)+d(x,y)$. Since N is tree-like, either mrca(x,y,z) = mrca(x,y) or mrca(x,y,z) = mrca(x,z) or mrca(x,y,z) = mrca(y,z). If mrca(x,y,z) = mrca(x,y), then the conclusion follows. In either other case, because of the absence of strict inequalities 4.7 shows that mrca(x,y,z) = mrca(y,z) = mrca(y,x) = mrca(x,z). The result now follows as for Case b with $c = e$.

This completes the proof of (1). The proofs of (2) and (3) are symmetric, proving 4.8. □

5 Reconstruction of a tree-like originating mrca-network from distances.

A *tree-like originating mrca-network* (or *TOM-network*) (V,A,r,h,d) is a strict originating mrca-network that is tree-like and such that each hybrid vertex has exactly two parents. A TOM-network will be our model for the underlying phylogenetic network. It typically arises from a weighted network.

For convenience, here is a summary of the assumed properties of a TOM-network:
(1) (V,A) is an acyclic directed graph with vertex set V and arc set A.
(2) The root $r \in V$ satisfies that, for all $v \in V$, $r \leq v$.
(3) If p is a parent of x, then there exists no $v \in V$ distinct from p and x, such

that $p \leq v \leq x$.
(4) For any nonempty subset U of V, mrca(U) exists.
(5) (V, A, r) is tree-like in that, for all $x \in V$, $y \in V$, $z \in V$ either $\text{mrca}(x, y, z) = \text{mrca}(x, y)$ or $\text{mrca}(x, y, z) = \text{mrca}(x, z)$ or $\text{mrca}(x, y, z) = \text{mrca}(y, z)$.
(6) To each $v \in V$ there exists $h(v) \geq 0$.
(7) $h(r) = 0$.
(8) If v is a regular vertex (a vertex with indegree 1) then $h(v) > 0$.
(9) If x is a hybrid vertex (a vertex with indegree greater than 1) then x has exactly two parents p and q (hence has indegree exactly 2). Moreover, $h(p) > 0$ and $h(q) > 0$.
(10) For every $x \in V$, $y \in V$,
$d(x, y) = \Sigma[h(u) : u \leq x, u \not\leq y] + \Sigma[h(u) : u \leq y, u \not\leq x]$.

If (V, A, r, h, d) is a TOM-network, a *base set* X is a subset of V that contains r, each leaf, and every vertex of outdegree 1. (Often there will be no vertices of outdegree 1.) The set X will correspond to the taxa for which measurements are possible. In particular we will assume that $d(x, y)$ is known exactly for $x \in X$, $y \in X$. Theorem 5.10 will show how, given only X (including the identity of r) and the distance $d(x, y)$ between members x and y of X, we shall be able to reconstruct V, A, h, and therefore d, effectively reconstructing the TOM-network. The reconstruction will be shown to require only polynomial time (Theorem 5.12). These results suggest that TOM-networks may be useful in biological settings.

In practice it will be possible to select r to be an outgroup, as discussed in Section 2.

If $N = (V, A, r, h, d)$ is a TOM-network, suppose x is a hybrid vertex with parents p and q. Call x a *positive hybrid* if $h(x) > 0$. If x is a positive hybrid, perform an operation to produce a new network N^x as follows: Insert a new vertex x' called the *separated vertex* at x. Delete the arcs (p, x) and (q, x). Insert new arcs (p, x'), (q, x'), and (x', x), and let $h(x') = 0$. This procedure will be called *separating* x, and N^x is the *separated network*. See Figure 5. Clearly N^x will be a TOM-network.

The separated vertex x' has biological meaning. Suppose that the true system is a monotone marked network. In the actual act of hybridization of taxa p and q, a new taxon x' was produced such that $M(x') = M(p) \cup M(q)$. Further mutation from x' led to the taxon x such that $H(x) = M(x) - M(x')$. Thus x' denotes the presumed first hybrid offspring and x denotes the descendent of x', slightly mutated, which first left sufficient record to be detectable in the TOM-network.

Here is an overview of how we shall try to reconstruct the TOM-network $N = (V, A, r, h, d)$ with base set X, given distances on X. The problem will recursively be reduced to an easier problem. We will define a nonnegative function $\delta : X - r \to \mathbb{R}$ called the *stem* function, which will turn out to have the following properties:
(1) If $\delta(x) > 0$, then x is a leaf (Lemma 5.3).
(2) If x is a leaf then $h(x) = \delta(x)$ (Lemmas 5.4 and 5.5).

Let $|X|$ denote the cardinality of X. If $|X| \leq 2$ then it is easy to see that N

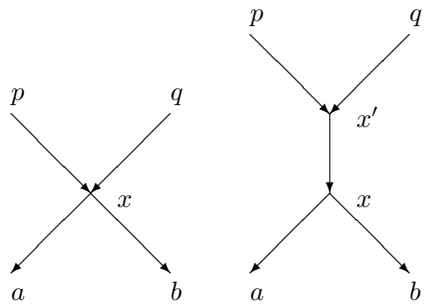

Figure 5: In separation of the hybrid vertex x, the left graph N is replaced by the right graph N^x in which $h(x') = 0$.

is determined. Hence we assume $|X| \geq 3$.

If there exists x with $\delta(x) > 0$, then by (1) it is a leaf and by (2) $h(x) = \delta(x)$ is known. If x was a regular leaf then it has a parent $p \in V$. Form a new TOM-network (V', A', r, h, d) with base set X' such that $V' = V - x$ and $X' = (X - x) \cup \{p\}$. For every $u \in X$, we will be able to infer $d(u, p)$ using the results of Section 4. We will thus have reduced to an easier problem since in this new problem V' has one fewer member.

If x was a positive hybrid leaf, then let x' be the separated vertex of x (so that $h(x') = 0$). Form a new TOM-network (V', A', r, h, d) with base set X' by letting $V' = (V - x) \cup \{x'\}$, and $X' = (X - x) \cup \{x'\}$. We will again have reduced to an easier problem since, while there are the same number of vertices in V' as in V, there will be one fewer positive hybrid vertex.

If every $x \in X$ satisfies $\delta(x) = 0$, then we will be able to pick $x \in X$ which is a hybrid leaf with $h(x) = 0$. Using the fact that $\delta(x) = 0$ by (2), we will be able to identify $a \in X$ and $b \in X$ such that a, b, and x are distinct and the parents of x are $p = \text{mrca}(x, a)$ and $q = \text{mrca}(x, b)$. For every $u \in X$, the distances $d(u, p)$ and $d(u, q)$ may be calculated by Theorem 4.8 since the network is tree-like. Hence we may remove x from X and V, store $h(x) = 0$, and add the vertices p and q to X. We will have reduced to an easier problem since the new V will have one fewer vertex.

Ultimately the procedure performed recursively will come to a situation where V has only two vertices, X has only two points and must agree with V, and the network is then obvious.

We now present the details of the reconstruction process. Throughout, (V, A, r, h, d) will be a TOM-network with base set X. The series of lemmas will culminate in Theorem 5.10, showing that the TOM-network has been uniquely determined.

The *stem function* $\delta : X - r \to \mathbb{R}$ is defined, for $x \in X$, $x \neq r$, by

$$\delta(x) = \min\{(d(x,y) + d(x,z) - d(y,z))/2 : y \in X, z \in X; x, y, z \text{ are distinct}\}.$$

By the triangle inequality $d(y,z) \leq d(y,x) + d(x,z)$, so $\delta(x) \geq 0$ for all $x \in X$.

Note that $\delta(x)$ can be computed entirely using the distances defined on members of X. We now show properties of the stem function. Lemma 5.3 will show that the stem function will let us identify a leaf.

Lemma 5.1. *For any $x \in X$, $y \in X$, $z \in X$, x, y, z distinct,*

$$(d(x,y)+d(x,z)-d(y,z))/2 = \Sigma[h(u) : u \leq x, u \not\leq y, u \not\leq z]+\Sigma[h(u) : u \leq y, u \not\leq x, u \leq z].$$

Proof. $(d(x,y) + d(x,z) - d(y,z))/2$
$= (\Sigma[h(u) : u \leq x, u \not\leq y] + \Sigma[h(u) : u \leq y, u \not\leq x] + \Sigma[h(u) : u \leq x, u \not\leq z]$
$+\Sigma[h(u) : u \leq z, u \not\leq x] - \Sigma[h(u) : u \leq y, u \not\leq z] - \Sigma[h(u) : u \leq z, u \not\leq y])/2$
$= (\Sigma[h(u) : u \leq x, u \not\leq y, u \leq z] + \Sigma[h(u) : u \leq x, u \not\leq y, u \not\leq z]$
$+\Sigma[h(u) : u \leq y, u \not\leq x, u \leq z] + \Sigma[h(u) : u \leq y, u \not\leq x, u \not\leq z]$
$+\Sigma[h(u) : u \leq x, u \leq y, u \not\leq z] + \Sigma[h(u) : u \leq x, u \leq\not\leq y, u \not\leq z]$
$+\Sigma[h(u) : u \leq z, u \not\leq x, u \leq y] + \Sigma[h(u) : u \leq z, u \not\leq x, u \not\leq y]$
$-\Sigma[h(u) : u \leq x, u \leq y, u \not\leq z] - \Sigma[h(u) : u \not\leq x, u \leq y, u \not\leq z]$
$-\Sigma[h(u) : u \leq x, u \leq z, u \not\leq y] - \Sigma[h(u) : u \not\leq x, u \leq z, u \not\leq y])/2$
$= \Sigma[h(u) : u \leq x, u \not\leq y, u \not\leq z] + \Sigma[h(u) : u \leq y, u \not\leq x, u \leq z].$ □

Corollary 5.2. *If $x \in X$, $x \neq r$, and x is a leaf, then $\delta(x) \geq h(x)$.*

Proof. For any y and z in X, with x, y, z distinct, $(d(x,y)+d(x,z)-d(y,z))/2 = \Sigma[h(u) : u \leq x, u \not\leq y, u \not\leq z] + \Sigma[h(u) : u \leq y, u \not\leq x, u \leq z]$ by 5.1. But since x is a leaf, it follows $x \not\leq y$ and $x \not\leq z$. Hence x contributes to the first sum. Since $h(u) \geq 0$ for all u, we obtain $(d(x,y) + d(x,z) - d(y,z))/2 \geq h(x)$. Since this is true for all such y and z, we obtain 5.2. □

Lemma 5.3. *If $x \in X$, $x \neq r$, and $\delta(x) > 0$ then x is a leaf.*

Proof. Suppose $x \in X$, $x \neq r$, and x is not a leaf. Then there is an arc (x,y). If y is not a leaf, then starting from y one can find a maximal directed path, which necessarily ends at a leaf z, which necessarily lies in the base set X. In either case, there exists $z \in X$, $z \neq x$, such that there is a directed path from x to z. Since $x \neq r$, there is a directed path from r to x. Note $r \neq z$ since otherwise there would be a directed cycle. Hence x, r, and z are distinct members of X. But then $0 \leq \delta(x) \leq (d(x,r) + d(x,z) - d(r,z))/2 = 0$, the last by 4.5 since $r \leq x \leq z$. Hence $\delta(x) = 0$, a contradiction. □

The next lemmas 5.4 and 5.5 will show that for any leaf x, whether regular or hybrid, the originating weight $h(x)$ equals the (known) value of $\delta(x)$. This result is key to reconstructing all the originating weights of the TOM-network.

Lemma 5.4. *Assume $|X| \geq 3$. If $x \in X$ is a regular leaf, then $\delta(x) = h(x)$.*

Proof. Let p denote the parent of x. We reduce to three cases, in each of which we show that $\delta(x) \leq h(x)$. It will then follow by 5.2 that $\delta(x) = h(x)$, proving the lemma.

Case 1. Assume $p \in X$, $p \neq r$. We can utilize $y = p$, $z = r$. Then
$\delta(x) \leq (d(x,p) + d(x,r) - d(p,r))/2$
$= \Sigma[h(u) : u \leq x, u \not\leq p, u \not\leq r] + \Sigma[h(u) : u \leq p, u \not\leq x, u \leq r]$ [by 5.1]
$= \Sigma[h(u) : u \leq x, u \not\leq p, u \not\leq r]$ [since if $u \leq r$ then $u = r$ and $h(r) = 0$]
$= h(x) + \Sigma[h(u) : u \leq x, u \neq x, u \not\leq p, u \not\leq r]$
$= h(x)$ [since if $u \leq x$ and $u \neq x$ then $u \leq p$].

Case 2. Assume $p = r$. Since $|X| \geq 3$, choose $z \in X$ so r, x, z are distinct. Then by 5.1,
$\delta(X) \leq (d(x,r) + d(x,z) - d(r,z))/2$
$= \Sigma[h(u) : u \leq x, u \not\leq r, u \not\leq z] + \Sigma[h(u) : u \leq r, u \not\leq x, u \leq z]$
$= \Sigma[h(u) : u \leq x, u \not\leq r, u \not\leq z]$ [since if $u \leq r$ then $u = r$ and $h(r) = 0$]
$= h(x) + \Sigma[h(u) : u \leq x, u \not\leq r, u \not\leq z, u \neq x] = h(x)$ [since if $u \leq x$ and $u \neq x$, then $u \leq p = r$, so $u = r$ and $h(r) = 0$]

Case 3. Assume $p \notin X$. Since X contains all leaves and all vertices of outdegree 1 and $p \notin X$, the outdegree of p is at least 2. Hence there exists a child q of p, $q \neq x$, and then a directed path from q to some leaf z. Since z is a leaf, $z \in X$ and $p \leq z$. If $z = x$ then $p \leq q \leq x$ with p, q, and x distinct implies that there is no arc from p to x, a contradiction. Hence $z \neq x$. Now by 5.1,
$\delta(X) \leq (d(x,r) + d(x,z) - d(r,z))/2$
$= \Sigma[h(u) : u \leq x, u \not\leq r, u \not\leq z] + \Sigma[h(u) : u \leq r, u \not\leq x, u \leq z]$
$= \Sigma[h(u) : u \leq x, u \not\leq r, u \not\leq z]$
$= h(x) + \Sigma[h(u) : u \leq x, u \not\leq r, u \not\leq z, u \neq x]$
$= h(x)$ [since if $u \leq x$, $u \neq x$, then $u \leq p$, whence $u \leq z$]. □

Lemma 5.5. *Let $x \in X$ be a hybrid leaf. Then $\delta(x) = h(x)$.*

Proof. Since x is hybrid, it has exactly two parents p and q. Choose $a \in X$ such that $a \neq x$ and $p \leq a$. (If $p \in X$, choose $a = p$. Otherwise p has a child z different from x. A maximal directed path from z must terminate at a leaf, which we will call a. Since a is a leaf, $a \in X$. If $a = x$ then $p \leq z \leq x$ implies that there was no arc from p to x, a contradiction. Hence $a \neq x$, and $p \leq z \leq a$.) Similarly choose $b \in X$ such that $b \neq x$ and $q \leq b$.

By 3.2, either $\mathrm{mrca}(x,a) = p$ or $\mathrm{mrca}(x,a) = x$. But if $\mathrm{mrca}(x,a) = x$ then $x \leq a$ contradicting that x is a leaf. Hence $\mathrm{mrca}(x,a) = p$. Similarly $\mathrm{mrca}(x,b) = q$. It follows that $a \neq b$, since otherwise $p = \mathrm{mrca}(x,a) = \mathrm{mrca}(x,b) = q$.

Suppose $u \leq x$, $u \neq x$. Since p and q are the only parents of x, it follows that either $u \leq p$ or $u \leq q$. If $u \leq p$ then $u \leq p \leq a$ while if $u \leq q$, then $u \leq q \leq b$. It follows that either $u \leq a$ or $u \leq b$. Hence $\Sigma[h(u) : u \leq x, u \not\leq a, u \not\leq b]$
$= h(x) + \Sigma[h(u) : u \leq x, u \neq x, u \not\leq a, u \not\leq b]$
$= h(x)$. (*)

Note a, b, and x are distinct vertices. Let $c = \mathrm{mrca}(a,b,x)$. Since the network is tree-like, either $c = \mathrm{mrca}(a,b)$ or $c = \mathrm{mrca}(a,x)$ or $c = \mathrm{mrca}(b,x)$. If

$c = \text{mrca}(a, x)$, then $c = p$ so since $c \leq b$ it follows $p \leq b$. Since $p \leq b$ and $p \leq x$ it follows $p \leq \text{mrca}(b, x) = q$. Hence $p \leq q \leq x$ which by properness contradicts that there is an arc from p to x. It follows that $c \neq \text{mrca}(a, x)$. A symmetric argument shows that $c \neq \text{mrca}(b, x)$. We conclude that $c = \text{mrca}(a, b)$. It follows that $\text{mrca}(a, b) \leq x$.

Now x, a, and b are distinct members of X. Hence by 5.1,
$\delta(x) \leq (d(x, a) + d(x, b) - d(a, b))/2$
$= \Sigma[h(u) : u \leq x, u \not\leq a, u \not\leq b] + \Sigma[h(u) : u \leq a, u \not\leq x, u \leq b]$
$= h(x) + \Sigma[h(u) : u \leq a, u \not\leq x, u \leq b]$ [by (*)]
$= h(x)$ [since if $u \leq a$ and $u \leq b$, then $u \leq \text{mrca}(a, b) \leq x$]

By 5.2, it follows that $\delta(x) = h(x)$. □

Suppose the vertex x is a leaf and v is an arbitrary vertex. The inductive proof of Theorem 5.10 requires that we can infer $d(v, p)$ where p is any parent of x. The following Lemma 5.6 gives the required formula in case x is a regular leaf. Lemma 5.7 gives the formula in case x is a hybrid leaf with parent a separated vertex x', while Lemma 5.8 gives the formula in case x is a hybrid leaf that has already been separated, so that $\delta(x) = 0$.

Lemma 5.6. *Assume $|X| \geq 3$. Suppose x is a regular leaf. Let $p \in V$ be the parent of x. Then for all $v \in V$ such that $v \neq x$, $d(v, p) = d(v, x) - \delta(x)$.*

Proof. By 4.2(1),
$d(v, x) = \Sigma[h(u) : u \leq v, u \not\leq x] + \Sigma[h(u) : u \leq x, u \not\leq v]$
$= \Sigma[h(u) : u \leq v, u \not\leq x] + \Sigma[h(u) : u \leq x, u \not\leq v, u \neq x] + h(x)$
$= \Sigma[h(u) : u \leq v, u \not\leq p] + \Sigma[h(u) : u \leq p, u \not\leq v] + h(x)$ [since $u \leq p$ iff $u \leq x$ and $u \neq x$]
$= d(v, p) + h(x)$ [by 4.2(1)]
$= d(v, p) + \delta(x)$ [by 5.4]. □

Lemma 5.7. *Suppose x is a hybrid leaf with parents p and q. Suppose $\delta(x) > 0$. Form N^x by inserting a separated vertex x'. Then for all $v \in V$, $v \neq x$, $d(v, x') = d(v, x) - \delta(x)$.*

Proof. Note that $x' \notin V$ since x' has become a single parent of x, where originally x had two parents. Since N^x is a TOM-network,
$d(v, x) = \Sigma[h(u) : u \leq v, u \not\leq x] + \Sigma[h(u) : u \leq x, u \not\leq v]$
$= \Sigma[h(u) : u \leq v, u \not\leq x] + \Sigma[h(u) : u \leq x, u \not\leq v, u \neq x] + h(x)$ [since $x \not\leq v$ because x is a leaf]
$= \Sigma[h(u) : u \leq v, u \not\leq x'] + \Sigma[h(u) : u \leq x', u \not\leq v] + h(x)$ [since $h(x') = 0$]
$= d(v, x') + h(x) = d(v, x') + \delta(x)$ [by 5.5]. □

Lemma 5.8. *Suppose x is a hybrid leaf with parents p and q such that $\delta(x) = 0$. Let $a \in X$ and $b \in X$ satisfy that a, b, and x are distinct and $(d(x, a) + d(x, b) - d(a, b))/2 = 0$. Then*
(1) Either ($p \leq a$ and $q \leq b$) or ($p \leq b$ and $q \leq a$).
Assume (by interchanging p and q if necessary) that $p \leq a$ and $q \leq b$. Then
(2) $p = mrca(x, a)$ and $q = mrca(x, b)$,

(3) $d(p,x) = (d(x,r) + d(x,a) - d(r,a))/2$
(4) $d(q,x) = (d(x,r) + d(x,b) - d(r,b))/2$,
(5) $d(r,p) = (d(r,a) + d(r,x) - d(a,x))/2$,
(6) $d(r,q) = (d(r,x) + d(r,b) - d(b,x))/2$,
(7) $d(p,q) = d(p,x) + d(q,x) = (2d(x,r) + d(x,a) + d(x,b) - d(r,a) - d(r,b))/2$.

Proof. By 5.1, $0 = (d(x,a) + d(x,b) - d(a,b))/2$
$= \Sigma[h(u) : u \leq x, u \not\leq a, u \not\leq b] + \Sigma[h(u) : u \leq a, u \not\leq x, u \leq b]$.

Observe $p \leq x$. If $p \not\leq a$ and $p \not\leq b$, then p would contribute to the left term . Since the sum is zero, it follows that $h(p) = 0$, which contradicts that $h(p) > 0$. We conclude that either $p \leq a$ or $p \leq b$. Similarly it must be true that either $q \leq a$ or $q \leq b$.

Suppose $p \leq a$ and $q \leq a$. By 3.2 since $p \leq x$, $\mathrm{mrca}(x,a)$ is either p or x, and since x is a leaf, it follows $\mathrm{mrca}(x,a) = p$. But similarly since $q \leq x$ it follows that $\mathrm{mrca}(x,a) = q$. Since p and q are distinct, this situation cannot occur. Hence it cannot be true that $p \leq a$ and $q \leq a$. Similarly it cannot be true that $p \leq b$ and $q \leq b$. This proves (1).

Now, possibly after interchanging p and q, we may assume $p \leq a$ and $q \leq b$. Since $p \leq a$, $p \leq x$, $a \neq x$, and x is a leaf, by 3.2 it follows $\mathrm{mrca}(x,a) = p$. Now (3) and (5) follow from 4.5

For (7), by 5.1
$(d(x,p) + d(x,q) - d(p,q))/2$
$= \Sigma[h(u) : u \leq x, u \not\leq p, u \not\leq q] + \Sigma[h(u) : u \leq p, u \not\leq x, u \leq q]$
$= \Sigma[h(u) : u \leq x, u \not\leq p, u \not\leq q]$ [since if $u \leq p$ then it follows $u \leq x$]
$= \Sigma[h(u) : u \leq x, u \neq x, u \not\leq p, u \not\leq q] + h(x) = h(x)$ [since if $u \neq x$, $u \leq x$ then there must be a path from u to a parent p or q of x]
$= 0$ [by 5.5 since $\delta(x) = 0$].
Hence $d(p,q) = d(p,x) + d(x,q)$
$= (d(x,r) + d(x,a) - d(r,a))/2 + (d(x,r) + d(x,b) - d(r,b))/2$ [by (3) and (4)]
$= (2d(x,r) + d(x,a) + d(x,b) - d(r,a) - d(r,b))/2$. This proves (7). □

Let X be a base subset of V. A vertex $x \in X$ is *interior* provided $x \neq r$ and x is not a leaf.

Lemma 5.9. *x is interior iff there exists $y \in X$, $y \neq x$, $y \neq r$, such that $d(r,y) = d(r,x) + d(x,y)$.*

Proof. If x is interior, then x is not a leaf. Hence a maximal directed path starting at x must terminate at a leaf y, which will lie in X since X is a base set. Then $r \leq x \leq y$, whence $d(r,y) = d(r,x) + d(x,y)$ by 4.5. Conversely suppose $y \neq x$, $y \neq r$, $y \in X$, and $d(r,y) = d(r,x) + d(x,y)$. By 4.5(2c), $d(x, \mathrm{mrca}(x,y)) = (d(x,r) + d(x,y) - d(r,y))/2 = 0$. Since d is a metric by 4.4, $x = \mathrm{mrca}(x,y)$, so $x \leq y$ and x is not interior. □

We may now prove the main theorem of this paper. The proof is by induction, in which the notion of "smaller" is as follows: If (V, A, r, h, d) is a TOM-network, a vertex $v \in V$ is *positive* if $h(v) > 0$. The set of positive vertices will be denoted $P(V, A, r, h, d)$. A TOM-network (V, A, r, h, d) will be

called *t-smaller* than a TOM-network (V', A', r', h', d') if (i) $|V| < |V'|$; or (ii) $|V| = |V'|$ but $|P(V, A, r, h, d)| < |P(V', A', r', h', d')|$.

Theorem 5.10. *Let $N = (V, A, r, h, d)$ be a TOM-network. Let X be a base subset of V containing r, all leaves, and all vertices of outdegree 1. Assume that $d : X \times X \to \mathbb{R}$ is known. Then N is uniquely determined.*

Proof. The proof will be by induction. We shall assume the theorem is true for all TOM-networks t-smaller than N and prove it for N.

For the base of the induction note that if $|X| = 1$ then N consists of the vertex r alone. If $|X| = 2$, then X consists of r and another vertex $x \neq r$. Clearly then A consists of the single arc (r, x) and $h(r) = 0$ while $h(x) = d(r, x)$.

Assume the result for all TOM-networks t-smaller than $N = (V, A, r, h, d)$. We may assume $|X| \geq 3$. Since d is determined from V, A, r, and h, we will omit it from the notation (V, A, r, h). There will be two cases for the induction.

Case 1. Assume that there exists $x \in X$ such that $\delta(x) > 0$. Select one such x. By 5.3, x is a leaf of N, and $h(x) = \delta(x)$ by 5.4 or 5.5. Let $X' = X - \{x\} \cup \{u\}$ where u is a new point. For $v \in V$, $v \neq x$, let $d'(v, u) = d(v, x) - \delta(x)$, while if $v \in V$, $w \in V$, $v \neq x$, $w \neq x$ let $d'(v, w) = d(v, w)$. If there exists $v \in V$, $v \neq x$, such that $d'(v, u) = 0$, then identify v with u (so now $X' = X - \{x\}$ in that case). To see that this construction reduces to a t-smaller TOM-network, we break into two subcases:

Subcase 1a. If x is a regular leaf, then x has a unique parent p. By 5.6 for every $v \in V$, $v \neq x$, $d(v, p) = d(v, x) - \delta(x) = d'(v, u)$, so we may identify p with u. Let $N' = (V', A', r, h')$, $V' = V - \{x\}$, $A' = A - \{(p, x)\}$, $h' = h|V'$. Note that N' is t-smaller than N since $|V'| < |V|$. Moreover, X' is a base set for N' where u is identified with p and d' is the metric on N' by 5.6. By the inductive hypothesis, N' is uniquely identified.

We now reconstruct N from N' and X. First note that we can identify which member of X' is p: If $X' \neq X$, then p is the member of X' not in X; if $X' = X$, then p is the member of X such that $d(p, x) = \delta(x)$. Now N is determined since $V = V' \cup \{x\}$, $A = A' \cup \{(p, x)\}$, h agrees with h' except at x and also satisfies $h(x) = \delta(x)$ by 5.4.

Subcase 1b. If x is a hybrid leaf with parents p and q, let x' be the separated vertex of x. Identify x' with u. By 5.7 for every $v \in V$, $v \neq x$, $d(v, x') = d(v, x) - \delta(x) = d'(v, u)$. Let N'' denote N^x in which x has been separated. Thus $N'' = (V'', A'', r, h'')$ with base set X where $V'' = V \cup \{x'\}$, $A'' = A - \{(p, x), (q, x)\} \cup \{(p, x'), (q, x'), (x', x)\}$, h'' agrees with h but $h''(x') = 0$. Let $N' = (V', A', r, h')$ with base set X' where $V' = V'' - \{x\}$, $A' = A'' - \{(x', x)\}$, $X' = X \cup \{x'\} - \{x\}$, $h' = h''|V'$. Then N' is t-smaller than N since $|V'| = |V|$ but the positive vertex x in N is replaced by the nonpositive vertex x' in N'. Moreover, X' is a base set for N' where u is identifed with x' and d' is the metric on N' by 5.7. By the inductive hypothesis N' is determined from the information.

Now we reconstruct N from N', X', and X in the obvious manner: First note that x' will be the member of X' not in X. We construct N'' by $V'' = V' \cup \{x\}$, $A'' = A' \cup \{(x', x)\}$, h'' agrees with h' but $h''(x) = \delta(x)$. Second, note that p

and q are identified as the parents of x'. We construct N by $V = V'' - \{x'\}$, $A = A'' - \{(p, x'), (q, x'), (x', x)\} \cup \{(p, x), (q, x)\}$, $h = h''|V$.

Case 2. Assume that $|X| \geq 3$ and there exists no $x \in X$, $x \neq r$, such that $\delta(x) > 0$. Then every member of X is either r or an interior vertex or a hybrid leaf such that $h(x) = 0$. Select $x \in X$, $x \neq r$, such that $d(r, x)$ is maximal. Then x is not interior since if x were interior by 5.9 there would be $y \in X$ such that $d(r, y) = d(r, x) + d(x, y) > d(r, x)$. Hence x is a hybrid leaf with $\delta(x) = h(x) = 0$.

Let p and q be the parents of x. Since $\delta(x) = 0$, there exist $a \in X$ and $b \in X$ such that a, b, and x are distinct and $(d(x, a) + d(x, b) - d(a, b))/2 = 0$. By 5.8 we may assume that $p \leq a$, $q \leq b$, $p = \text{mrca}(a, x)$, and $q = \text{mrca}(b, x)$.

Let $N' = (V', A', r, h')$ with base set X' where $V' = V - \{x\}$, $A' = A - \{(p, x), (q, x)\}$, $h' = h|V'$, $X' = (X - \{x\}) \cup \{p, q\}$. Then N' is t-smaller than N since $|V'| < |V|$. Moreover $d|X' \times X'$ is known: If y and z are in X, then $d(y, z)$ was known by hypothesis. If $y \neq p$, $y \neq q$, then by 4.8 since $p = \text{mrca}(a, x)$, we may compute $d(y, p)$ as follows:
(1) If $d(r, y) + d(a, x) \leq d(r, x) + d(y, a)$ and $d(r, a) + d(y, x) \leq d(r, x) + d(y, a)$ then $d(y, p) = d(y, \text{mrca}(a, x)) = (2d(y, a) + d(r, x) - d(a, x) - d(r, a))/2$.
(2) If $d(r, x) + d(y, a) \leq d(r, y) + d(a, x)$ and $d(r, a) + d(y, x) \leq d(r, y) + d(a, x)$ then $d(y, p) = d(y, \text{mrca}(a, x)) = (2d(r, y) + d(a, x) - d(r, a) - d(r, x))/2$.
(3) If $d(r, y) + d(a, x) \leq d(r, a) + d(y, x)$ and $d(r, x) + d(y, a) \leq d(r, a) + d(y, x)$ then $d(y, p) = d(y, \text{mrca}(a, x)) = (2d(y, x) + d(r, a) - d(a, x) - d(r, x))/2$.
There is an analogous formula for $d(y, q)$ when $y \neq p$, $y \neq q$. Finally, $d(p, q) = (2d(x, r) + d(x, a) + d(x, b) - d(r, a) - d(r, b))/2$ by 5.8(7).

Hence by the inductive hypothesis, N' is uniquely determined.

In reconstructing N from N' we must identify which members of X' are p and q. Note that p will be the point such that $d(p, p) = 0$ where $d(y, p)$ is computed by 4.8. More explicitly it will be the point p such that either
(1) $d(r, p) + d(a, x) \leq d(r, x) + d(p, a)$, $d(r, a) + d(p, x) \leq d(r, x) + d(p, a)$ and $0 = (2d(p, a) + d(r, x) - d(a, x) - d(r, a))/2$; or
(2) $d(r, x) + d(p, a) \leq d(r, p) + d(a, x)$, $d(r, a) + d(p, x) \leq d(r, p) + d(a, x)$ and $0 = (2d(r, p) + d(a, x) - d(r, a) - d(r, x))/2$; or
(3) $d(r, p) + d(a, x) \leq d(r, a) + d(p, x)$, $d(r, x) + d(p, a) \leq d(r, a) + d(p, x)$ and $0 = (2d(p, x) + d(r, a) - d(a, x) - d(r, x))/2$. It will be possible to check which point is p since all distances not involving x were reconstructed in N'; $d(r, x)$ and $d(a, x)$ were assumed known; and $d(p, x) = (d(x, r) + d(x, a) - d(r, a))/2$ by 5.8.

Similarly it will be possible to identify the parent q.

We now reconstruct N by $V = V' \cup \{x\}$, $A = A' \cup \{(p, x), (q, x)\}$, $h = h'$ except that $h(x) = 0$. □

The remainder of this section is devoted to showing that the reconstruction of the TOM-network can be accomplished in polynomial time.

Lemma 5.11. *Let (V, A, r, h, d) be a TOM-network with base set X. Then $|V| \leq 3|X| - 2$.*

Proof. The proof is by induction on $|V|$. If $|V| = 1$, then $V = \{r\} = X$, $|X| = 1$, and $|V| \leq 3|X| - 2$. Assume the result when $|V| \leq n$, and consider a TOM-network $N = (V, A, r, h, d)$ with $|V| = n + 1$. There must be a leaf w. Consider two cases:

Case (1). Assume w is a regular vertex with indegree 1. Let (p, w) be the unique arc to w. There are two subcases depending on the outdegree of p.

Subcase (1a). Assume that the outdegree of p is at least 3. Form a new network N' with base set X' by deleting w and the arc (p, w) from N and letting $X' = X - w$. Then $|V'| = |V| - 1$ and $|X'| = |X| - 1$. By the inductive hypothesis, $|V'| \leq 3|X'| - 2$, from which it follows $|V| \leq 3|X| - 4 \leq 3|X| - 2$.

Subcase (1b). Assume that the outdegree of p is 2, with outgoing arcs (p, w) and (p, v). Form a new network N' with vertex set $V' = V - \{w, p\}$. The arcs of N' are found by removing the arcs (p, w) and (p, v), and replacing each arc (u, p) by (u, v). The base set is $X' = X - w$. Now $|V'| = |V| - 2$ and $|X'| = |X| - 1$. By the inductive hypothesis, $|V'| \leq 3|X'| - 2$, from which it follows $|V| \leq 3|X| - 3 \leq 3|X| - 2$.

Case (2). Assume that w is a hybrid vertex. Then it has exactly two parents p and q. There are several subcases:

Subcase (2a). Assume that both p and q have outdegree 2. Let the arcs from p be (p, w) and (p, v), while the arcs from q are (q, w) and (q, u). Then $u \neq v$ by 3.3. Form a new network N' with vertex set $V' = V - \{w, p, q\}$. The arcs of N' are found by deleting (p, w) and (p, v) from A; moreover, remove each arc (s, p) and replace it by (s, v); and similarly remove each arc (s, q) and replace it by (s, u). The base set $X' = X - w$. Now $|V'| = |V| - 3$ and $|X'| = |X| - 1$. By the inductive hypothesis, $|V'| \leq 3|X'| - 2$, from which it follows $|V| \leq 3|X| - 2$.

It is easy to deal, as in Subcase (1a), with the three remaining subcases where p or q or both have outdegree at least 3. □

Theorem 5.12. *Let $N = (V, A, r, h, d)$ be a TOM-network with base set X. The reconstruction of N from $d : X \times X \to \mathbb{R}$ takes time at worst $O(|V|^4)$ or $O(|X|^4)$.*

Proof. The proof of 5.10 shows that N is reconstructed by recursively reconstructing a number of different networks. Each step either (i) removes a vertex v from V with $h(v) > 0$; or (ii) replaces a hybrid vertex x with $h(x) > 0$ by its separated vertex x' with $h(x') = 0$; or (iii) removes a hybrid vertex. Hence the number of steps is at most $|V| + |H|$ where H is the set of hybrid vertices of N. Since $|H| \leq |V|$, it follows that the number of steps is at most $2|V|$.

Each step requires the recomputation of $\delta(x)$ for each member x of the current X set (which in the worst case is as large as V), and such a computation requires in the worst case computing $(d(x, a) + d(x, b) - d(a, b))/2$ for all possible a and b in X distinct from x. This requires time $O(|V|^3)$. The remaining operations take time which is less than $O(|V|^3)$.

Hence the total time in the worst case is at most $O(|V|)$ cases, each taking time $O(|V|^3)$ hence total time $O(|V|^4)$. By 5.11, $|V| = O(|X|)$, so the worst time is $O(|X|^4)$. □

6 Acknowledgments

The author thanks the anonymous referees for many helpful suggestions on the presentation. He also thanks Raul Piaggio for useful questions about the complexity.

References

[1] Bandelt, H.-J. and A. Dress (1986). Reconstructing the shape of a tree from observed dissimilarity data. Advances in Applied Mathematics 7, 309-343.

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

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

[4] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17, 368-376.

[5] Felsenstein, J. (2004). Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

[6] Fitch, W.M. (1981). A non-sequential method for constructing trees and hierarchical classifications. J. Mol. Evol. 18, 30-37.

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

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

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

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

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

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

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

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

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

[16] Li, W.-H. (1981). A simple method for construcing phylogenetic trees from distance matrices. Proc. Natl. Acad. Sci. USA 78, 1085-1089.

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

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

[20] Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 , 406-425.

[21] Sattath, S. and A. Tversky (1977). Additive similarity trees. Psychometrika 42, 319-345.

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

[23] Willson, S.J (2005). Unique solvability of certain hybrid networks from their distances. To appear in Annals of Combinatorics.