Reconstruction of certain phylogenetic networks from their tree-average distances

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

October 10, 2012

Abstract: Trees are commonly utilized to describe the evolutionary history of a collection of biological species, in which case the trees are called *phylogenetic trees*. Often these are reconstructed from data by making use of distances between extant species corresponding to the leaves of the tree. Because of increased recognition of the possibility of hybridization events, more attention is being given to the use of phylogenetic networks that are not necessarily trees. This paper describes the reconstruction of certain such networks from the *tree-average distances* between the leaves. For a certain class of phylogenetic networks, a polynomial-time method is presented to reconstruct the network from the tree-average distances. The method is proved to work if there is a single reticulation cycle.

Keywords: phylogeny, network, metric, phylogenetic network, tree, tree-average distance

1 Introduction

In phylogeny, the evolution of a collection of species is modelled via a directed graph in which the vertices are species and the arcs indicate direct descent, usually with modification as mutations accumulate. The tips correspond to extant species, while internal vertices typically correspond to presumed ancestors. It has been common to assume that the directed graphs are trees, but more recently more general networks have also been studied so as to include the possibility of hybridization of species or lateral gene transfer [9], [4]. General frameworks for phylogenetic networks are discussed in [1], [2], [19], and [20]. See also the recent book [13]. An example of published non-tree network published by a biologist is [18].

For a particular analysis, suppose X denotes the set of extant species, including an outgroup which is used to locate the root. The DNA information may be summarized via the computation of distances between members of X. If $x, y \in X$, then $d(x, y)$ summarizes the amount of genetic difference between the DNA strings of x and y. In order to compensate at least partially for the possibility of repeated mutation at the same site, a number of different distances are in use, based on different models of mutation. Notable examples include the Jukes-Cantor [15], Kimura [16], HKY [11], and log determinant [17], [23] distances. The log determinant distance is especially interesting in that if the evolution is along a phylogenetic tree, then it can be proved that the distances add along the paths, so that the distance along a path is the sum of the distances on each edge along the path.

Some fast methods to reconstruct phylogenetic trees make use of distances between members of X. Probably the most common distance-based method is Neighbor-joining [21]. It is computationally fast, requiring $O(n^3)$ time for n species in X. It often gives a good initial tree with which heuristic methods begin in order to find an improved tree by other methods. Another more recent method FastME [7], [8] is based on the principle of balanced minimum evolution, in which one assumes that the correct tree is the one that exhibits the minimal total amount of evolution, suitably measured.

Distance-based methods that are founded on biological models of evolution have been rarely used to construct phylogenetic networks that are not necessarily trees. Sometimes distances occur in exploratory methods to display the diversity of trees for the same species such as the split decomposition (see [12] or an overview in [13]); but these distances, however, are not based on a biological model of evolution.

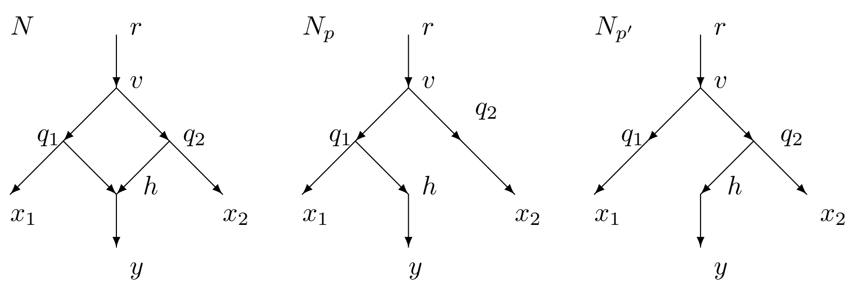

Figure 1: A network N with root r, and the two trees N_p and $N_{p'}$ that it displays. Here h inherits some of its characters from q_1 and the other characters from q_2.

This paper assumes that the appropriate model of evolution is a directed network that is not necessarily a tree. Consider, for example, the network N in Figure 1. The root is r and there is a hybridization event at h with parents q_1 and q_2. Vertex h is called a *hybrid* vertex or a *reticulation* vertex. For some characters, the character state at h is inherited from the parental species q_1, while for other characters the character state at h is inherited from species q_2.

For character states inherited from q_1 the evolutionary history is best described by the displayed tree N_p, while for character states inherited from q_2 the history is best described by the tree $N_{p'}$. Here p and p' are *parent maps* telling the parent of every non-root vertex. In the example $p(h) = q_1$ while $p'(h) = q_2$. Each parent map p leads to a displayed tree N_p.

In Figure 1, each arc might have a numerical *weight* measuring the amount of genetic change on the arc. In either tree N_p or $N_{p'}$ the distance between two vertices might be plausibly defined as the sum of the weights of the edges on the unique path between the vertices. In this paper we will utilize the *tree-average distance* so that for each x and y in X, their distance $d(x, y; N)$ in the network N is a weighted average of their distances in N_p and $N_{p'}$. This distance was studied in [27].

More generally, the trees displayed by a network N will be conveniently indexed as N_p where p ranges over all the parent maps. Let $Par(N)$ denote the set of all parent maps for N. For each hybrid vertex h, the probability that a character of h is inherited from a particular parent vertex q_i will be denoted $\alpha(q_i, h)$. Assume that these inheritances at different hybrid vertices are independent events. Then for each $p \in Par(N)$ we obtain that the probability $Pr(p)$ that the tree N_p models the inheritance of a particular character is given by

$$Pr(p) = \prod [\alpha(p(h), h) : h \text{ is hybrid}].$$

If x and y are vertices, then the distance between x and y in N_p, written $d(x, y; N_p)$, is the sum of the weights of arcs on the unique path joining x and y in N_p. The *tree-average distance* $d(x, y; N)$ between x and y in N was defined in [27] to be the expected value of the distances in the various trees N_p:

$$d(x, y; N) = \sum [Pr(p)d(x, y; N_p) : p \in Par(N)].$$

In this paper as in [27] we will assume that the weight of an arc into a hybrid vertex is 0. Thus in Figure 1, the weights of arcs (q_1, h) and (q_2, h) will be zero. Under this assumption vertex h corresponds roughly to the immediate offspring of a hybridization event, in which some characters came intact from q_1 and the remainder intact from q_2. Further mutation occurred before species y evolved from h, as measured by the weight of arc (h, y).

Note that the number of arcs of N in Figure 1 that are not directed into a hybrid vertex is 6. It is therefore plausible that given the 6 numbers $d(x, y; N)$ for $x, y \in \{r, x_1, x_2, y\}$, we might be able to recover the weights for each of the 6 arcs in N that are not directed into the hybrid vertex h. These same weights would be utilized in distances for both N_p and $N_{p'}$. On the other hand, the additional parameter $\alpha(q_1, h)$ telling the probability of inheritance by h of a character from q_1 would also be of interest. It is unlikely that six equations, one for each $d(x, y; N)$, will uniquely and generically determine seven real parameters. Consequently for the situation in Figure 1 we will assume that $\alpha(q_1, h) = \alpha(q_2, h) = 1/2$; in this case we call the inheritance *equiprobable at h.*

By contrast, Figure 2 shows another network with $X = \{r, x_1, x_2, x_3, y\}$ containing a single hybrid vertex h. In this case there are $\binom{5}{2} = 10$ distances

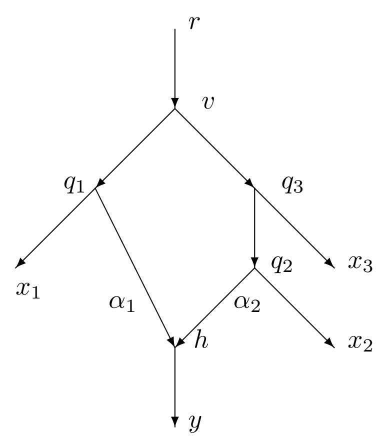

Figure 2: A minimal configuration needed to be able to find the probability $\alpha_1 = \alpha(q_1, h) = 1 - \alpha_2$ that a character state in h is inherited from q_1.

and 8 arcs not into a hybrid vertex, so it is plausible that the 10 equations would allow us to uniquely determine a ninth parameter $\alpha_1 = \alpha(q_1, h)$ satisfying $0 < \alpha_1 < 1$. In fact, [27] shows how to determine all 9 parameters. Then $\alpha(q_2, h) = 1 - \alpha_1$ is also determined.

Particular kinds of acyclic networks have been studied in various papers. Wang *et al.* [24] and Gusfield *et al.* [10] study "galled trees" in which all recombination events are associated with node-disjoint recombination cycles; the idea occurs also earlier in [25]. Choy *et al.* [6] and Van Iersel *et al.* [14] generalized galled trees to "level-k" networks. Baroni, Semple, and Steel [2] introduced the idea of a "regular" network, which coincides with its cover digraph. Cardona *et al.* [5] discussed "tree-child" networks, in which every vertex that is not a leaf has a child that is not a reticulation vertex. An arc (a, b) is *redundant* if there is a directed path from a to b that that does not utilize this arc. The current author has utilized "normal" networks [26], [27] which are both tree-child and contain no redundant arc.

Most results in this paper assume that the network is *normal.*

This paper builds on the results in [27]. These assume that the network N is normal and also that for all hybrid vertices the indegree is exactly 2 and the outdegree is exactly 1. Then [27] shows that from knowledge both of N and of the tree-average distance function d, the weights for all arcs are uniquely determined and indeed can be computed by explicit formulas. Moreover, the probabilities of inheritance at each hybrid vertex are uniquely determined and can be computed by explicit formulas.

The current paper concerns the reconstruction of N itself from the tree-average distance function d. What makes the reconstruction possible is that the formulas in [27] have simple forms which can be used recursively when only part of the network is yet known.

A model is called *identifiable* if the parameters can in principle be reconstructed from the (exact) values of the data that are modeled. The results in [27] assert that, if the tree-average distance function d on X and the network N are known, then the real parameters of the model (i.e., the weights and the probabilities) are identifiable in various cases. The results in the current paper imply that, under the given assumptions, the network N itself is also determined from d, so the network is itself identifiable.

Here is an outline of the current paper. Theorem 3.2 shows that a recursive reconstruction of N will be possible provided that one can correctly recognize two situations using the tree-average distance d. These two situations are a cherry $\{x, y\}$ and a hybrid h with parents q_1 and q_2 such that h has a leaf-child y, q_1 has a leaf-child x_1, and q_2 has a leaf-child x_2 as in Figure 2. For each of these situations, one can use the formulas of [27] to find the weights and probabilities needed to simplify the problem to a smaller problem. In the case of a cherry, one can reduce to a simpler problem in which both x and y have been removed; in the case of a hybrid one can reduce to a simpler problem in which h, y, x_1, and x_2 have been removed.

Section 4 deals with the problem of recognizing a cherry $\{x, y\}$. The main result is Theorem 4.5, which gives a necessary and sufficient condition that $\{x, y\}$ be a cherry, given in terms of the tree-average distance. As a result of Theorem 4.5, there remains only the problem of correctly recognizing the hybrid situation.

The recognition of a hybrid is studied in Section 5. Several necessary conditions in terms of the tree-average distance are described. The main Theorem 6.1 asserts that these conditions are sufficient when there is a single reticulation cycle. A proof is given in Section 6. Section 7 gives a more complicated example with two reticulation cycles in which the conditions are also sufficient.

Some extensions of the current results and problems are discussed in the concluding Section 8.

2 Fundamental Concepts

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

The *indegree* of vertex u is the number of $v \in V$ such that $(v, u) \in A$. The *outdegree* of u is the number of $v \in V$ such that $(u,v) \in A$. A *leaf* is a vertex of outdegree 0. A *normal vertex* (or *tree vertex*) is a vertex of indegree 1. A *hybrid* vertex (or *reticulation vertex*) is a vertex of indegree at least 2. An arc (u,v) is a *normal arc* if v is a normal vertex.

A digraph (V,A) is *rooted* if it has a unique vertex $r \in V$ with indegree 0 such that, for all $v \in V$, $r \leq v$. This vertex r is called the *root*.

Let X denote a finite set. Typically in phylogeny, X is a collection of species. Measurements are assumed to be possible among members of X, so that we may assume that, for example, their DNA is known for each $x \in X$.

A *phylogenetic* X*-network* $N = (V,A,r,X)$ is a rooted acyclic digraph $G = (V,A)$ with root r such that there is a one-to-one map $\phi : X \to V$ whose image contains all vertices v such that either
(i) v is a leaf; or
(ii) $v = r$; or
(iii) v has indegree 1 and outdegree 1.
There may be additional vertices in X. We will identify each $x \in X$ with its image $\phi(x)$. The set X will be called the *base-set* for N.

In biology the network gives a hypothesized relationship among the members of X. It is quite common also that a certain extant *outgroup* species r' is assumed to have evolved separately from the rest of the species in question. When this happens, we identify the species r' with the root r. Thus extant species (the leaves) are in X by (i) since measurements can be made on them. The outgroup r', which is identified with the root, is in X by (ii). If a vertex has indegree 1 and outdegree 1 then nothing uniquely determines it unless, for fortuitous reasons, it is possible to make measurements on its DNA, in which case it lies in the base-set X.

An X*-tree* is a phylogenetic X-network such that the underlying digraph is a tree.

Figure 2 shows a phylogenetic X-network N with base-set $X = \{r, x_1, x_2, x_3, y\}$. The root is r. Measurements such as DNA are assumed possible on members of X. Since the root r is actually an outgroup and the leaves are all extant, this is plausible for all members of X.

An arc $(u,v) \in A$ is *redundant* if there exists $w \in V$ such that u, v, and w are distinct and $u \leq w \leq v$. The removal of a redundant arc (u,v) still leaves $u \leq v$ in the network.

A phylogenetic X-network $N = (V,A,r,X)$ with base-set X is *normal* provided (1) whenever $v \in V$ and $v \notin X$, then v has a tree-child c; and (2) there are no redundant arcs. The network in Figure 2 is normal.

A *normal path* in N from v to x is a directed path $v = v_0, v_1, \cdots, v_k = x$ such that for $i = 1, \cdots, k$, v_i is normal. A *normal path from v to X* is a normal path starting at v and ending at some $x \in X$. For example, in Figure 2, the path v, q_3, q_2, x_2 is normal and is a normal path from v to X. The path q_1, h, y is not normal since h is hybrid. The trivial path x_1 is normal.

If $N = (V,A,r,X)$ is a phylogenetic X-network, then a *parent map* p for N consists of a map $p : V - \{r\} \to V$ such that, for all $v \in V - \{r\}$, $p(v)$ is a parent of v. Note that r has no parent. If v is normal, then there is only one possibility for $p(v)$, while if v is hybrid, there are at least two possibilities for $p(v)$. In Figure 2, an example of a parent map p satisfies $p(h) = q_2$, and for all other vertices w besides r, $p(w)$ is the unique parent of w.

Write $Par(N)$ for the set of all parent maps for N. In general if there are k distinct hybrid vertices and they have indegrees respectively $i_1, i_2, \cdots, i_k$, then the number of distinct parent maps p is $|Par(N)| = \prod[i_j : j = 1, \cdots, k]$. If N is a network with k distinct hybrid vertices, each of indegree 2, then $|Par(N)| = 2^k$.

Given $p \in Par(N)$ the set A_p of *p-arcs* is $A_p = \{(p(v), v) : v \in V - \{r\}\}$. The *induced tree* N_p is the directed graph (V, A_p) with root r. See N_p in Figure 1 when $p(h) = q_1$. Note that each vertex in $V - \{r\}$ has a unique parent in N_p. Thus N_p is a tree with vertex set V. The set X, however, need not be a base-set of N_p. For example, in Figure 1, N_p contains the vertex h with indegree 1 and outdegree 1, yet $h \notin X$.

Several of the proofs will require the notion of "complementary parents". Suppose $p \in Par(N)$ and h is a particular hybrid vertex with exactly two parents q_1 and q_2. Assume $p(h) = q_1$. The *complementary parent map* p' of p *with respect to* h is defined by

$$p'(v) = \begin{cases} p(v) & \text{if } v \neq h \\ q_2 & \text{if } v = h. \end{cases}$$

Thus p' agrees with p except at h, where p' chooses the other parent from that chosen by p. Of occasional use will be the network $G_p = N_p \cup N_{p'}$.

A phylogenetic X-network is *weighted* provided that for each arc $(a, b) \in A$ there is a non-negative number $\omega(a, b)$ called the *weight of* (a, b) such that
(1) if b is hybrid, then $\omega(a, b) = 0$;
(2) if b is normal, then $\omega(a, b) \geq 0$.
We call the function ω from the set of arcs to the reals the *weight function* of N. We interpret $\omega(a, b)$ as a measure of the amount of genetic change from species a to species b. If h is hybrid with parents q_1 and q_2 and unique child c, then the hybridization event is essentially assumed to be instantaneous between q_1 and q_2 with no genetic change in those character states inherited by h from q_1 or q_2 respectively. Further mutation then occurs from h to c, as measured by $\omega(h, c)$.

If $N = (V, A, r, X)$ is a phylogenetic X-network and $S \subset X$, then the restriction of N to S, denoted $N|S$, consists of that part of N which includes all possible ancestors of members of S. More formally $N|S = (V', A', r, S)$ where $V' = \{v \in V : \text{there exists } s \in S \text{ such that } v \leq s\}$, $A' = \{(u, v) \in A : u \in V', v \in V'\}$. It is easy to see that $N|S$ is a phylogenetic S-network. If N is weighted, then $N|S$ is also weighted, using the same weight function but restricted to A'.

In any rooted network $N = (V, A, r, X)$, *a most recent common ancestor* of two vertices u and v is a vertex w such that (1) $w \leq u$ and $w \leq v$, and (2) there is no vertex w' such that $w' \leq u$, $w' \leq v$, $w < w'$. In general a most recent common ancestor of u and v exists, but it need not be uniquely determined. In any rooted tree, however, there is a unique most recent common ancestor of u and v.

Suppose that $N = (V, A, r, X)$ is a weighted phylogenetic X-network with weight function ω. For each $p \in Par(N)$ and for each $u, v \in V$, define the distance $d(u, v; N_p)$ as follows: in N_p there is a unique undirected path $P(u, v)$ between u and v; define $d(u, v; N_p)$ to be the sum of the weights of arcs along $P(u, v)$. More precisely, since N_p is a tree, there exists a most recent common ancestor $m = \mathrm{mrca}(u, v; N_p)$, a directed path P_1 given by $m = u_0, u_1, \cdots, u_k = u$ from m to u, and a directed path P_2 given by $m = v_0, v_1, \cdots, v_j = v$ from m to v. Define

$$d(u, v; N_p) = \sum[\omega(u_i, u_{i+1}) : i = 0, \cdots, k-1] + \sum[\omega(v_i, v_{i+1}) : i = 0, \cdots, j-1].$$

We shall refer to $d(u, v; N_p)$ as the *distance between u and v in N_p*.

Let H denote the set of hybrid vertices of N. For each $h \in H$, let $P(h)$ denote the set of parents of h, i.e. the set of vertices u such that $(u, h) \in A$. Since $h \in H$, $|P(h)| \geq 2$. For each $u \in P(h)$, let $\alpha(u, h)$ denote the probability that a character is inherited by h from u. As an approximation, $\alpha(u, h)$ measures the fraction of the genome that h inherits from u. Note for all $h \in H$, $\sum[\alpha(u, h) : u \in P(h)] = 1$.

If h and h' are distinct members of H, we will assume that the inheritances at h and h' are independent. More generally, suppose for every $h \in H$ that q_h is a parent of h. Then we assume that the events that a character at h is inherited from q_h are independent. It is then easy to see that for each $p \in Par(N)$ the probability that inheritance follows the parent map p is $Pr(p) = \prod[\alpha(p(h), h) : h \in H]$.

Following [27] we define the *tree-average distance* $d(u, v; N)$ between u and v in N by

$$d(u, v; N) = \sum[Pr(p) d(u, v; N_p) : p \in Par(N)].$$

It is thus the expected value of the distances between u and v in the various N_p.

The simplest situation has each parent of h equally likely, so $\alpha(p(h), h) = 1/|P(h)|$ for each $p \in Par(N)$. If this situation occurs, we call the network *equiprobable at h.* Further details and examples are in [27].

If $T = (V, A, , X)$ is a undirected phylogenetic tree with leaf set X, a 4-set $\{x_1, x_2, x_3, x_4\}$ from X is a *quartet*. When T is restricted to a quartet, the result is called a *quartet tree*. The only possible quartet trees are denoted $x_1x_2|x_3x_4$, $x_1x_3|x_2x_4$, $x_1x_4|x_2x_3$, and $x_1x_2x_3x_4$. In $x_1x_2|x_3x_4$ removal of the internal edge disconnects T so that one component contains x_1 and x_2 while the other component contains x_3 and x_4. The *star* is denoted $x_1x_2x_3x_4$. For additive distances on trees, it is well-known [22] that $x_1x_2|x_3x_4$ if and only if $d(x_1, x_2) + d(x_3, x_4) < d(x_1, x_3) + d(x_2, x_4) = d(x_1, x_4) + d(x_2, x_3)$.

Let $N = (V, A)$ be an acyclic digraph. A *pseudocycle* in N is a sequence of vertices $x_0, x_1, x_2, \cdots, x_n$ from V with $n > 0$ such that $x_n = x_0$ and for each i (taken mod n) either

(1) (x_i, x_{i+1}) is an arc; or

(2) x_i is hybrid with distinct parents x_{i-1} and x_{i+1} and (x_{i+1}, x_i) is an arc.

A pseudocycle is not a cycle since it is not a directed path. Nevertheless it is very similar to a cycle since time is moving forward on most parts of the sequence. The existence of a pseudocycle indicates a lack of "time consistency". For example, if there is a temporal representation on the network [3] then each vertex v has a "time" $f(v)$ such that when v has parents p and q, then $f(p) = f(q)$; and when c is a child of u, then $f(u) < f(c)$. Following a pseudocycle we see that the successive hybrid parents must exist later in time and yet loop back to the original hybrid node, an impossibility. Hence the network can have no pseudocycle.

3 Overview of the reconstruction

Throughout the reconstruction we will make the following assumptions:

Assumptions 3.1. *Let $N = (V, A, r, X)$ be a rooted directed network. Assume*
(0) N is normal.
(1) All hybrids have indegree 2 and outdegree 1, and the child is a tree-child.
(2) Every weight of an arc to a hybrid vertex is 0.
(3) The weight of every arc to a normal vertex is positive.
(4) All normal vertices have outdegree 0 or 2.
(5) N has no pseudocycles.
(6) X consists of the set of leaves of N together with r.

A *cherry* $\{x, y\}$ is a set of two vertices x and y in X such that
(1) both x and y are leaves which are normal vertices;
(2) both x and y have the same parent q;
(3) q is normal;
(4) q has outdegree 2.

Suppose that the tree-average distance d is known between all members of X. We wish to see how to reconstruct N. The first key result, Theorem 3.2, asserts that either there is a cherry or else there is a hybrid vertex of a particularly simple kind.

Theorem 3.2. *Assume Assumptions 3.1 . Suppose N has no cherry and at least 4 vertices in X. Then there exists a hybrid vertex h with parents q_1 and q_2 such that each of these has a tree-child which is a leaf.*

The conclusion of Theorem 3.2 is illustrated in Figure 2. In the figure, there is no cherry but h is hybrid with parents q_1 and q_2. Then h has tree-child y which is a leaf, q_1 has tree-child x_1 which is a leaf, and q_2 has tree-child x_2 which is a leaf. If these conditions occur, we will say that $(y; x_1, x_2)$ is a *hybrid triple.*

Proof. Choose a maximal path (with the most arcs) from r ending at v_1 with parent w_1. Then v_1 is a leaf, hence normal. If w_1 has another child c, then c cannot be hybrid, since then c would have a child and a longer path from r could have been obtained. Hence c is normal. Moreover, if c had a child then a longer path could have been obtained; hence c is a normal leaf. Since v_1 is normal then $\{c, v_1\}$ is a cherry, a contradiction. Hence v_1 has no sibling whence w_1 is hybrid with two parents q_{11} and q_{12}. We choose the labeling so that q_{11} is on the given maximal path from r to w_1. Note there are two arcs on the path after q_{11}. By normality, q_{11} has a normal child z. If z is not a leaf it has at least two children c_1 and c_2 . By maximality each child c_i is a leaf. But no leaf is hybrid, whence both are normal, so $\{c_1, c_2\}$ is a cherry, a contradiction. Hence z is a normal leaf.

By normality, q_{12} has a normal child u_1. If u_1 is a leaf then we are done with $h = w_1$, $q_1 = q_{11}$, $q_2 = q_{12}$, $y = v_1$, $x_1 = z$, $x_2 = u_1$. Otherwise choose a maximal directed path starting at u_1. Repeat the argument. Since the vertex set is finite there is ultimately a repetition leading to a pseudocycle. This is impossible, so the procedure terminates. □

We may now present the general idea of the reconstruction. Suppose that $N = (V, A, r, X)$ is a phylogenetic X-network satisfying Assumptions 3.1. Suppose we are given all the tree-average distances $d(x, y; N)$ for x and y in X. Initially a network $R = (W, B)$ has vertex set $W = X$ and arc set $B = \emptyset$. We recursively simplify the network to a new network N' as follows:

(1) For each distinct x and y in X we check, using Corollary 4.6 (in the next section) whether $\{x, y\}$ is a cherry.

(2) If a cherry $\{x, y\}$ is recognized, then we proceed as follows:
By [27] Lemma 4.3(2), the parent q of both x and y satisfies that $\omega(q, x) = [d(x, y; N) + d(x, r; N) - d(r, y; N)]/2$ and $\omega(q, y) = [d(y, x; N) + d(y, r; N) - d(r, x; N)]/2$. Moreover, by additivity of the distances, for every $z \in X$, z other than x or y, $d(z, q; N) = d(q, x; N) - \omega(q, x) = d(q, y; N) - \omega(q, y)$.

We construct a new phylogenetic network $N' = (V', A', r, X')$ with distance d' and a network $R' = (W', B')$ as follows: Since $\{x, y\}$ is recognized as a cherry, there exists in N a vertex q which is the parent of x and y. Let $V' = V - \{x, y\}$, $X' = X - \{x, y\} \cup \{q\}$, $A' = A - \{(q, x), (q, y)\}$. Moreover, for $z \in X'$, $d(z, q; N') = d(z, x; N) - \omega(q, x)$ is known. Finally $d'(u, v; N') = d(u, v; N)$ for $\{u, v\} \subset X'$ if neither u nor v is q; and $d'(z, q; N') = d(q, x; N) - \omega(q, x)$.

There is a new vertex q (identified with the q in N) such that $W' = W \cup \{q\}$ and $B' = B \cup \{(q, x), (q, y)\}$ where $\omega(q, x)$ and $\omega(q, y)$ are given as above.

(3) Suppose no cherry $\{x, y\}$ is recognized. Then by Theorem 4.5 no cherry exists in N, and by Theorem 3.2 there exists a hybrid triple $(y_0; x_{10}, x_{20})$. For each possible choice of $(y; x_1, x_2)$ we use Section 5 to determine whether $(y; x_1, x_2)$ is a hybrid triple. By Theorem 3.2, this will succeed for some choice. By Theorem 6.1, no triple that is not a hybrid triple will be falsely identified, under certain additional assumptions.

(3a) Suppose $(y; x_1, x_2)$ is identified as an equiprobable hybrid triple. By [27] we know $\omega(h, y)$, $\omega(q_1, x_1)$, and $\omega(q_2, x_2)$. We modify N to $N' = (V', A', r, X')$ where $V' = V - \{h, y, x_1, x_2\}$, $A' = A - \{(h, y), (q_1, x_1), (q_2, x_2)\}$, $X' = X - \{y, x_1, x_2\} \cup \{q_1, q_2\}$. We know for $v \in V'$, v other than q_1, q_2 by [27] $d(v, q_1; N) = d(v, x_1; N) - \omega(q_1, x_1)$ $d(v, q_2; N) = d(v, x_2; N) - \omega(q_2, x_2)$. Moreover, $d(q_1, q_2; N) = d(x_1, x_2; N) - \omega(q_1, x_1) - \omega(q_2, x_2)$.

We modify $R = (W, B)$ to $R' = (W', B')$ where $W' = W \cup \{q_1, q_2, h\}$, $B' = B \cup \{(h, y), (q_1, x_1), (q_2, x_2), (q_1, h), (q_2, h)\}$, $\alpha(q_1, h) = 1/2$, $\alpha(q_2, h) = 1/2$, $\omega(q_1, h) = \omega(q_2, h) = 0$. Moreover, by [27], $\omega(h, y)$, $\omega(q_1, x_1)$, and $\omega(q_2, x_2)$ are given by the formulas $arv = (d(r, x_1) + d(r, x_2) - d(x_1, x_2))/2$, $\omega(q_1, x_1) = d(x_1, y) - d(r, y) + arv$, $\omega(q_2, x_2) = d(x_2, y) - d(r, y; M) + arv$, $\omega(h, y) = (d(y, x_1) + d(y, x_2; M) - d(x_1, x_2))/2$.

(3b) Suppose $(y; x_1, x_2)$ is identified as a hybrid triple such that x_3 is identified as a normal descendant of an appropriate ancestor of x_2. Then [27] Lemma 4.9 gives formulas in this situation for $\omega(h, y)$, $\omega(q_1, x_1)$, $\omega(q_2, x_2)$ as well as $\alpha(q_1, h)$ and $\alpha(q_2, h)$. Then we proceed as in (3a) except that we use these alternative formulas for these quantities given by [27].

4 Recognition of a cherry

In this section we see necessary and sufficient conditions to recognize whether $\{x, y\}$ is a cherry.

Suppose w and z in X satisfy that w, x, y, z are distinct in X. For any network M with base-set X, let $W_x(M) = d(w, x; M) + d(z, y; M)$, $W_y(M) = d(w, y; M) + d(x, z; M)$, $W_z(M) = d(w, z; M) + d(x, y; M)$, using the tree-average distance d on M.

Lemma 4.1. *Suppose x and y are leaves that form a cherry in the network N. Suppose w and z in X satisfy that w, x, y, z are distinct in X. Then $W_z(N) < W_x(N) = W_y(N)$.*

Proof. For every parent map p we have $wz|xy$ in N_p, so

$$d(w, z; N_p) + d(x, y; N_p) < d(w, x; N_p) + d(z, y; N_p) = d(w, y; N_p) + d(z, x; N_p)$$

with the strict inequality since the common parent q of x and y is normal so the arc into q has positive weight. Hence $W_z(N_p) < W_x(N_p) = W_y(N_p)$. Taking averages over p weighted by $Pr(p)$, we see that $W_z(N) < W_x(N) = W_y(N)$. $\square$

Theorem 4.2 is the converse of Lemma 4.1. Together these two results give a necessary and sufficient condition for $\{x, y\}$ to be a cherry.

Theorem 4.2. *Assume Assumptions 3.1. Suppose x and y are in X. Suppose that for all choices of w and z in X such that w, x, y, z are distinct, we have that $W_z(N) < W_x(N) = W_y(N)$. Then $\{x, y\}$ is a cherry.*

The proof of Theorem 4.2 will require a lemma. The lemma shows that if for various parent maps of N we have exactly two of the possibilities among $W_z < W_x = W_y$, $W_x < W_z = W_y$, $W_y < W_z = W_x$ then for the tree-average distance we cannot have any of those conditions.

Lemma 4.3. *Suppose x and y are in X. Pick w and z in X so that w, x, y, z are distinct.*

(1) Assume for a nonempty collection of parent maps p we have $W_z(N_p) < W_x(N_p) = W_y(N_p)$ and for a nonempty collection of parent maps p we have $W_x(N_p) < W_z(N_p) = W_y(N_p)$ but we never have a parent map p for which $W_y(N_p) < W_z(N_p) = W_x(N_p)$. Then we can't have $W_z(N) < W_x(N) = W_y(N)$.

(2) Assume for a nonempty collection of parent maps p we have $W_y(N_p) < W_z(N_p) = W_x(N_p)$. and for a nonempty collection of parent maps p we have $W_x(N_p) < W_z(N_p) = W_y(N_p)$ but we never have a parent map p for which $W_z(N_p) < W_x(N_p) = W_y(N_p)$. Then we can't have $W_z(N) < W_x(N) = W_y(N)$.

(3) Assume for a nonempty collection of parent maps p we have $W_z(N_p) < W_x(N_p) = W_y(N_p)$ and for a nonempty collection of parent maps p we have $W_y(N_p) < W_z(N_p) = W_x(N_p)$ but we never have a parent map p for which $W_x(N_p) < W_z(N_p) = W_y(N_p)$. Then we can't have $W_z(N) < W_x(N) = W_y(N)$.

Here is a geometric interpretation of Lemma 4.3: Suppose w, x, y, z are distinct members of X. Suppose there exist parent maps p such that N_p displays the quartet $wz|xy$ and parent maps p such that N_p displays the quartet $wx|yz$ but no N_p displays the quartet $wy|xz$. Then the tree average distance cannot appear to have quartet $wz|xy$ via $d(w,z) + d(x,y) < d(w,x) + d(y,z) = d(w,y) + d(x,z)$. Nor can it appear to have quartet $wy|xz$ via $d(w,y) + d(x,z) < d(w,x) + d(y,z) = d(w,z) + d(x,y)$.

Proof. (1) Write A_z for the sum of the $W_z(N_p)$ for p such that $W_z(N_p) < W_x(N_p) = W_y(N_p)$ weighted by the probability of p; thus
$A_z = \sum[Pr(p)W_z(N_p) : W_z(N_p) < W_x(N_p) = W_y(N_p)]$. Similarly let $B_z = \sum[Pr(p)W_z(N_p) : W_x(N_p) < W_z(N_p) = W_y(N_p)]$. Then $W_z(N) = A_z + B_z$ since these exhaust all the parent maps p under the assumptions of (1). Similarly define $A_x = \sum[Pr(p)W_x(N_p) : W_z(N_p) < W_x(N_p) = W_y(N_p)]$
$B_x = \sum[Pr(p)W_x(N_p) : W_x(N_p) < W_z(N_p) = W_y(N_p)]$,
$A_y = \sum[Pr(p)W_y(N_p) : W_z(N_p) < W_x(N_p) = W_y(N_p)]$
$B_y = \sum[Pr(p)W_x(N_p) : W_x(N_p) < W_z(N_p) = W_y(N_p)]$. Thus $W_x(N) = A_x + B_x$ and $W_y(N) = A_y + B_y$.

Suppose (1) is false, so $A_z + B_z < A_x + B_x = A_y + B_y$.

By linearity, $A_z < A_x = A_y$ and $B_x < B_z = B_y$.

Since $A_x + B_x = A_y + B_y$ and $A_x = A_y$, it follows that $B_x = B_y$. This contradicts $B_x < B_y$, proving (1).

(2) Let $A_z = \sum[Pr(p)W_z(N_p) : W_y(N_p) < W_z(N_p) = W_x(N_p)]$
$B_z = \sum[Pr(p)W_z(N_p) : W_x(N_p) < W_z(N_p) = W_y(N_p)]$
and similarly define A_x, A_y, B_x, B_y. Then $A_y < A_z = A_x$ and $B_x < B_z = B_y$. Moreover, $W_z(N) = A_z + B_z$ since these exhaust all the parent maps p under the assumptions of (2). If (2) is false and $W_z(N) < W_x(N) = W_y(N)$ then $A_z + B_z < A_x + B_x = A_y + B_y$.

But $A_z = A_x$ so $A_x + B_z < A_x + B_x$ whence $B_z < B_x$, a contradiction, proviing (2).

(3) follows symmetrically with the proof of (1). □

Corollary 4.4. *Suppose N is a phylogenetic network satisfying Assumptions 3.1. Suppose w, x, y, z are distinct leaves. Assume*

(i) there exists a quartet $wx|yz$ or $wy|xz$ or $wz|xy$ such that there is no parent map p for which this quartet occurs in N_p.
(ii) there exists a parent map p for which $wx|yz$ or $wy|xz$ occurs in N_p.

Then it cannot be the case that $W_z(N) < W_x(N) = W_y(N)$.

Proof. If there exist two different quartets that arise in some N_p but not three, then one of (1), (2), or (3) in Lemma 4.3 occurs and the conclusion follows. By (i) we cannot have all three quartets occuring in N_p for various parent maps p. Hence the only case that remains is that only one quartet occurs in N_p for various p. By (ii) it is either $wx|yz$ or $wy|xz$. In the former case we have $W_x < W_y = W_z$ and in the latter we have $W_y < W_x = W_z$. □

We can now prove Theorem 4.2.

Proof. Both x and y are normal leaves. Let q_y be the parent of y and q_x be the parent of x. I claim $q_x = q_y$.

If $q_x \neq q_y$, then there exists a most recent common ancestor v of x and y and it must satisfy that either $v < q_x$ or $v < q_y$ or both. Without loss of generality assume $v < q_y$. Hence there is a directed path $P(v, q_y)$ from v to q_y of positive length such that no vertex of $P(v, q_y)$ except v is ancestral to x. In particular we do not have $q_y \leq x$. There are 5 cases to consider.

Assume first that q_y is normal, hence of indegree 1. Since it has a child and its outdegree cannot be 1, its outdegree is 2. Since the outdegree of q_y is 2, it has another child c, which is either normal or hybrid.

Case 1. Suppose c is normal. By normality of N, we may choose a normal path from c to $z \in X$. See Figure 3a.

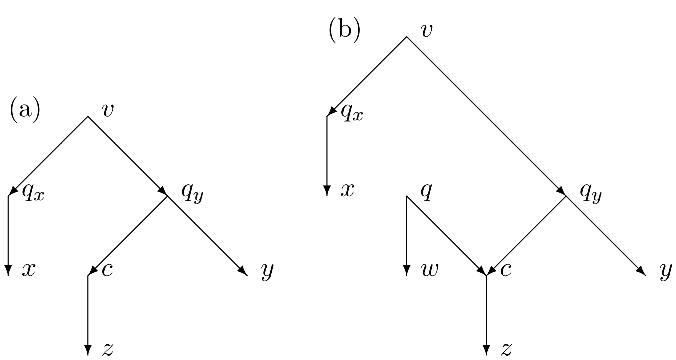

Figure 3: (a) Case 1 of Theorem 4.2. (b) Case 2.

I claim that for all p, in N_p we have $yz|rx$. To see this, note that each N_p contains the arcs (q_y, y) and (q_y, c) and the normal path $P(c, z)$. Their union forms the undirected path $P(y, z)$ in N_p between y and z. But then $P(y, z)$ is disjoint from the undirected path $P(r, x)$ in N_p. (Otherwise, they would meet in a vertex on $P(q_y, z)$ and it would follow that $q_y \leq x$, a contradiction.)

Since N_p is a tree, it follows that for all p, with $w = r$, $W_x(N_p) < W_y(N_p) = W_z(N_p)$. If we take the averages weighted by the probabilities, we see $W_x(N) < W_y(N) = W_z(N)$, contradicting the hypotheses. Thus Case 1 cannot occur.

Case 2. Suppose c is hybrid with other parent q. Since there are no hybrid leaves, choose a nontrivial normal path $P(c, z)$ from c to $z \in X$. Since q has outdegree 2 we may choose a normal path $P(q, w)$ from q to $w \in X$. Assume q is not $\leq x$. See Figure 3b.

Claim 2a. There is no parent map p such that N_p has $wy|xz$.

To see Claim 2a, we show that for any p, $P(w, y; N_p)$ intersects $P(x, z; N_p)$. First, there exists t such that $P(w, y; N_p)$ is the union of the directed paths $P(t, w; N_p)$ and $P(t, y; N_p)$. And there exists s such that $P(x, z; N_p) = P(s, x; N_p) \cup P(s, z; N_p)$.

First, observe that $P(t, y; N_p)$ must include q_y since q_y is the unique parent of y. Moreover, since $t \leq w$ in N_p and $P(q, w)$ is normal in N, either t lies on $P(q, w)$ or else $t \leq q$ and $P(t, y; N_p)$ contains $P(q, w)$. But if t lies on $P(q, w)$ then $q \leq t \leq q_y$ so (q, c) is redundant. Thus $P(t, y; N_p)$ contains $P(q, w)$ and in particular q. Hence $P(t, y; N_p)$ must contain both q and q_y.

Second, observe that $P(x, z; N_p)$ must contain either q or q_y. To see this, since $s \leq z$ and $P(c, z)$ is normal, either $s \leq c$ or else s lies on $P(c, z)$. But if s lies on $P(c, z)$ then $q_y \leq c \leq s \leq x$ contradicting that q_y is not $\leq x$. Hence $s \leq c$ in N_p. Since c has only the two parents q and q_y, it follows that $P(s, z)$ contains either q or q_y.

Thus $P(w, y; N_p)$ intersects $P(x, z; N_p)$ as claimed.

Claim 2b. There exists a parent map p such that N_p displays $yz|wx$.

To see Claim 2b, suppose p satisfies $p(c) = q_y$. Since y is normal and $P(c, z; N)$ is normal, it follows that $P(y, z; N_p)$ is the union of (q_y, y), (q_y, c), and $P(c, z)$. There exists t such that $P(w, x; N_p) = P(t, w; N_p) \cup P(t, x; N_p)$. I claim that $P(w, x)$ is disjoint from $P(y, z)$. To see this note

(i) $P(t, w; N_p)$ cannot contain the leaf y.

(ii) $P(t, w; N_p)$ cannot contain q_y. If $P(t, w; N_p)$ contains q_y then $q_y \leq w$. Since $P(q, w)$ is normal, either $q_y \leq q$ or else q_y lies on $P(q, w)$. If $q_y \leq q$ then (q_y, c) is redundant, a contradiction. If q_y lies on $P(q, w)$ then $q \leq q_y$ so (q, c) is redundant, a contradiction.

(iii) $P(t, w; N_p)$ cannot intersect $P(c, z)$. If they met in u, then $c \leq u \leq w$. Since $P(q, w)$ is normal either $c \leq q$ or else c lies on $P(q, w)$. But if $c \leq q$ there is a directed cycle in N. If c lies on $P(q, w)$ then since $q \neq c$ the arc (q, c) is redundant.

(iv) $P(t, x; N_p)$ cannot contain the leaf y.

(v) $P(t, x; N_p)$ cannot contain q_y. Otherwise, $q_y \leq x$, a contradiction.

(vi) $P(t, x; N_p)$ cannot intersect $P(c, z)$. If they met in u, then $c \leq u \leq x$ whence $q_y \leq x$, a contradiction.

This proves Claim 2b.

By Corollary 4.4, we cannot have $W_z(N) < W_x(N) = W_y(N)$. Hence Case 2 cannot occur.

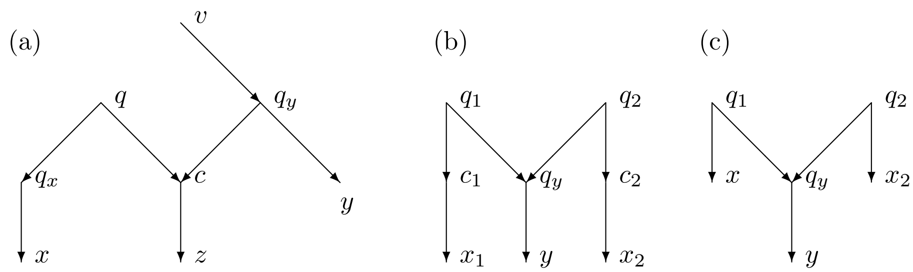

Figure 4: (a) Case 3 of Theorem 4.2. (b) Case 4. (c) Case 5.

Case 3. Suppose c is hybrid with other parent q. Since there are no hybrid leaves, we may choose a normal path from c to $z \in X$. Suppose we may also choose a normal path from q to $x \in X$. Let $w = r$. See Figure 4a.

Claim 3a. There is no parent map p such that $wz|xy$ in N_p.

To see this, suppose p is a parent map. Recall $w = r$. We show that the undirected path $P(w, z; N_p)$ from w to z in N_p always intersects $P(x, y; N_p)$. In the rooted tree N_p, note that q and q_y have a most recent common ancestor u. N_p must contain either (q, c) or (q_y, c). Thus the path $P(w, z; N_p)$ must contain $P(r, u; N_p)$, $P(c, z; N_p)$, and either $P(u, q; N_p)$ or $P(u, q_y; N_p)$. In particular $P(w, z; N_p)$ must contain either q or q_y.

On the other hand, $P(x, y; N_p)$ must contain q_y since it is the unique parent of the leaf y. Moreover, $P(x, y; N_p)$ must contain q. This is because $P(x, y; N_p) = P(s, x) \cup P(s, y)$ for some s. Since $P(q, x)$ is normal and $s \leq x$, either $s \leq q$ or s lies on $P(q, x)$ and $s \neq q$. In the latter case, $q < s \leq y$, whence $q \leq q_y$ and (q, c) is redundant, a contradiction.

Claim 3b. There exists a parent map p such that $wx|yz$ in N_p.

To see this, choose p such that $p(c) = q_y$. I claim $P(w, x; N_p)$ does not intersect $P(y, z; N_p)$. Note that $P(r, x; N_p)$ must consist of directed paths $P(r, u; N_p)$ together with $P(r, q; N_p)$ and $P(q, x; N_p)$ since $P(q, x)$ is normal in N. On the other hand $P(y, z; N_p)$ consists of (q_y, y), (q_y, c) and $P(c, z)$ since $P(c, z)$ is normal in N. It is now clear that $P(w, x; N_p)$ is disjoint from $P(y, z; N_p)$.

By Corollary 4.4 we cannot have $W_z(N) < W_x(N) = W_y(N)$, showing that Case 3 cannot occur.

Now instead of assuming that q_y is normal, we assume q_y is hybrid and the leaf y is the unique child of q_y. Let q_1 and q_2 be the parents of q_y. Choose normal children c_i of q_i respectively and normal paths from c_i to $x_i \in X$ respectively.

Case 4. Assume that we may choose x_1 and x_2 so that x, y, x_1, and x_2 are all distinct. See Figure 4b.

Let $w = x_1$, $z = x_2$.

Claim 4a. There is no parent map p such that in N_p we have $wz|xy$. To see this, suppose p is a parent map. Suppose $wz|xy$, so that $P(x_1, x_2; N_p)$ is disjoint from $P(x, y; N_p)$.

Note that $P(x_1, x_2; N_p)$ is the union of $P(s, x_1; N_p)$ and P$(s, x_2; N_p)$ for some s. Since $s \leq x_1$ in N_p, either $s \leq q_1$ or else s lies on $P(q_1, x_1; N_p)$ and is distinct from q_1. But in the latter case $q_1 < s \leq x_2$ in N_p, whence $q_1 < s \leq q_2$ in N_p and (q_1, q_y) is redundant in N, a contradiction. Hence $s \leq q_1$ in N_p. A similar argument shows $s \leq q_2$ in N_p. Hence $P(x_1, x_2; N_p)$ includes both q_1 and q_2.

On the other hand, $P(x, y; N_p)$ must contain the leaf y, hence its unique child q_y hence the parent of q_y in N_p hence either q_1 or q_2. This shows that $P(x, y; N_p)$ cannot be disjoint from $P(x_1, x_2; N_p)$, proving the claim.

For every p, N_p is binary, so in any N_p it follows we must have either $wx|yz$ or $wy|xz$.

By Corollary 4.4 it follows that we cannot have $W_z(N) < W_x(N) = W_y(N)$, so Case 4 cannot occur.

Since Case 4 cannot occur, it follows that we cannot select x_1 and x_2 so that x_1, x_2, x, and y are distinct. Hence (possibly by interchanging x_1 and x_2) we may assume that the only leaf descendant of q_1 by a normal path is x. Thus the only remaining case is the following case 5. See Figure 4c.

Case 5. The vertex q_y is hybrid with parents q_1 and q_2. There is a normal path from q_1 to x and a normal path from q_2 to x_2. Let $w = r$, $z = x_2$.

Claim 5a. There is no parent map p such that $ry|xx_2$ in N_p.

To see claim 5a, it suffices to show $P(r, y; N_p)$ and $P(x, x_2; N_p)$ must intersect. Let u denote the most recent common ancestor of x and x_2 in N_p, so $P(x, x_2; N_p)$ is the union of $P(u, x; N_p)$ and $P(u, x_2; N_p)$. But $u \leq x$ in N_p, so either u lies on $P(q_1, x)$ or $u \leq q_1$ in N_p. If u lies on $P(q_1, x)$ then in N_p we have $q_1 \leq u \leq x_2$, whence in N_p we have $q_1 \leq q_2$, whence $q_1 \leq q_2$ in N, which implies that (q_1, q_y) is redundant, a contradiction. Hence $u \leq q_1$ in N_p. A similar argument shows $u \leq q_2$ in N_p. Hence $P(x, x_2; N_p)$ contains both q_1 and q_2. On the other hand $P(r, y; N_p)$ must include y, hence its unique parent q_y, hence either q_1 or q_2, so $P(r, y; N_p)$ intersects $P(x, x_2; N_p)$.

Claim 5b. There exists a parent map p such that $rx|yx_2$ in N_p.

To see claim 5b, choose p such that $p(q_y) = q_2$. Then in N_p we have that $P(y, x_2; N_p)$ is the union of $P(q_2, x_2; N_p)$ and $P(q_2, q_y, y; N_p)$. If $u = \mathrm{mrca}(q_1, q_2; N_p)$, then $P(r, x)$ is the union of $P(r, u)$, $P(u, q_1)$, and $P(q_1, x)$. Suppose $P(r, x)$ meets $P(y, x_2; N_p)$ in s.

(i) If $s \in P(q_1, x) \cap P(q_2, x_2)$ then $q_1 \leq s \leq x_2$, forcing $q_1 \leq s \leq q_2$, making (q_1, q_y) redundant.

(ii) If $s = y$, there is a contradiction since y has no children, and if $s = q_y$ there is a contradiction since the only proper descendant is y.

(iii) If $s \in P(u, q_1) \cap P(q_2, x_2)$ then $q_2 \leq s \leq q_1$, so (q_2, q_y) is redundant.

(iv) If $s \in P(r, u) \cap P(q_2, x_2)$ then $q_2 \leq s \leq u \leq q_1$ so (q_2, q_y) is redundant.

Hence there can be no intersection of $P(r, x)$ and $P(y, x_2)$, so $rx|yx_2$ in N_p.

By Corollary 4.4, it follows that we cannot have $W_z(N) < W_x(N) = W_y(N)$, so Case 5 cannot occur.

Cases 1 through 5 show that the assumption that $q_x \neq q_y$ is impossible, so $q_x = q_y$. Since q_x has outdegree at least two, it must have outdegree exactly 2 by the hypotheses, and it must be normal (since a hybrid vertex has outdegree one). Hence $\{x,y\}$ form a cherry, as asserted. This completes the proof of the theorem. □

We may combine Lemma 4.1 and Theorem 4.2 into the following summary:

Theorem 4.5. *(a) If $|X| \geq 4$, then $\{x,y\}$ is a cherry if and only if for all w and z in X such that $\{w,x,y,z\}$ are distinct, we have $W_z(N) < W_x(N) = W_y(N)$. (b) If $|X| = 3$, say $X = \{r,x,y\}$, then $\{x,y\}$ is a cherry.*

Proof. (a) is immediate from Lemma 4.1 and Theorem 4.2. If $|X| = 3$ then there can be no hybrid vertex and (b) is immediate. □

5 Recognition of a hybrid

Suppose that we seek to reconstruct $N = (V,A,r,X)$ from the tree-average distances on X. From Section 4 we know how to recognize a cherry $\{x,y\}$. Hence we may assume there is no cherry, and by Theorem 3.2 there exists a hybrid vertex h with parents q_1 and q_2 such that h has a child y which is a leaf, q_1 has a child x_1 which is a leaf, and q_2 has a child x_2 which is a leaf. The essential step is to identify such y, x_1, and x_2. To do this, we consider all possibilities for y, x_1, and x_2 and find a choice which satisfies certain necessary criteria.

We present five necessary criteria, labeled B through F.

5.1 Criterion B: Clustering conditions

This criterion is the most useful for quickly eliminating false candidates for hybrids.

Lemma 5.1. *Assume that h is hybrid with parents q_1 and q_2 and both $\alpha(q_1,h) > 0$ and $\alpha(q_2,h) > 0$. Suppose h has a normal leaf child y, q_1 has a normal leaf child x_1, and q_2 has a normal leaf child x_2. Suppose $w \in X$ is distinct from y, x_1, and x_2. For each network M with the same X let $W_y(M) = d(w,y;M) + d(x_1,x_2;M)$, $W_{x_1}(M) = d(w,x_1;M) + d(y,x_2;M)$, $W_{x_2}(M) = d(w,x_2;M) + d(y,x_1;M)$.*
Then for all such w, $W_{x_1}(N) < W_y(N)$ and $W_{x_2}(N) < W_y(N)$.

The geometric content of Lemma 5.1 is seen in Figure 2. Suppose $w \in X$ is distinct from y, x_1, and x_2 somewhere in the network (unspecified in Figure 2). Note that for a parent map p with $p(h) = q_1$, for the 4-set $\{w,y,x_1,x_2\}$ we necessarily have the quartet tree $yx_1|wx_2$. For the complementary parent map p' with $p'(h) = q_2$ we necessarily have the quartet tree $yx_2|wx_1$. Lemma 5.1 essentially says that there is no parent map p such that $wy|x_1x_2$ in N_p.

The proof will require the following definition. For a given parent map p with $p(h) = q_1$, let p' denote the complementary parent map and $G_p = N_p \cup N_{p'}$ be the network N_p with the additional arc (q_2,h). Let H be the set of hybrid vertices of N. For each $p \in Par(N)$ satisfying $p(h) = q_1$, let $W(p) = \prod[\alpha(p(h'),h') : h' \in H, h' \neq h]$. Hence $Pr(p) = \alpha(q_1,h)W(p)$ and $Pr(p') = \alpha(q_2,h)W(p)$.

Proof. Each parent map satisfies either $p(h) = q_1$ or $p(h) = q_2$. If $p(h) = q_1$ then for every $w \in X$ distinct from y, x_1, x_2 we have that $\{y, x_1\}$ is a cherry in N_p, so $W_{x_2}(N_p) < W_{x_1}(N_p) = W_y(N_p)$. If $p(h) = q_2$ then for all such w we have that $\{y, x_2\}$ is a cherry, so $W_{x_1}(N_p) < W_{x_2}(N_p) = W_y(N_p)$. In particular, if $p(h) = q_1$ we have $W_{x_2}(N_p) < W_{x_1}(N_p) = W_y(N_p)$ and if p' is the complementary parent map (so $p'(h) = q_2$) then $W_{x_1}(N_{p'}) < W_{x_2}(N_{p'}) = W_y(N_{p'})$. It follows that

$$\alpha(q_1, h)W_{x_2}(N_p) < \alpha(q_1, h)W_{x_1}(N_p) = \alpha(q_1, h)W_y(N_p)$$

and

$$\alpha(q_2, h)W_{x_1}(N_{p'}) < \alpha(q_2, h)W_{x_2}(N_{p'}) = \alpha(q_2, h)W_y(N_{p'})$$

We combine N_p and $N_{p'}$ into the network $G_p = N_p \cup N_{p'}$. When we take into account the probabilities at h, we see $W_y(G_p) = \alpha(q_1, h)W_y(N_p) + \alpha(q_2, h)W_y(N_{p'})$.

Take the sum over all parent maps. Since each p satisfying $p(h) = q_1$ has its complementary p', we see that
$W_y(N) = \sum[W(p)W_y(G_p); p(h) = q_1]$. Similarly,
$W_{x_1}(G_p) = \alpha(q_1, h)W_{x_1}(N_p) + \alpha(q_2, h)W_{x_1}(N_{p'})$,
$W_{x_1}(N) = \sum[W(p)W_{x_1}(G_p) : p(h) = q_1]$,
$W_{x_2}(G_p) = \alpha(q_1, h)W_{x_2}(N_p) + \alpha(q_2, h)W_{x_2}(N_{p'})$,
$W_{x_2}(N) = \sum[W(p)W_{x_2}(G_p) : p(h) = q_1]$.

Hence $W_{x_1}(G_p) = \alpha(q_1, h)W_{x_1}(N_p) + \alpha(q_2, h)W_{x_1}(N_{p'}) < \alpha(q_1, h)W_y(N_p) + \alpha(q_2, h)W_y(N_{p'}) = W_y(G_p)$ and the inequality is strict since the case $p'(h) = q_2$ occurs and $\alpha(q_2, h) > 0$ so $W_{x_1}(G_p) < W_y(G_p)$. Similarly $W_{x_2}(G_p) = \alpha(q_1, h)W_{x_2}(N_p) + \alpha(q_2, h)W_{x_2}(N_{p'}) < \alpha(q_1, h)W_y(N_p) + \alpha(q_2, h)W_y(N_{p'}) = W_y(G_p)$ but the inequality is strict since $p(h) = q_1$ occurs and $\alpha(q_1, h) > 0$ so $W_{x_2}(G_p) < W_y(G_p)$.

Now $W_{x_1}(N) = \sum[W(p)W_{x_1}(G_p) : p(h) = q_1] < \sum[W(p)W_y(G_p) : p(h) = q_1] = W_y(N)$ so $W_{x_1}(N) < W_y(N)$. Similarly $W_{x_2}(N) < W_y(N)$. □

We say that $(y; x_1, x_2)$ passes Criterion B provided that the conclusion of Lemma 5.1 holds. Alternatively, Lemma 5.1 says that if y, x_1, and x_2 have the hypothesized relationship with a hybrid vertex, then $(y; x_1, x_2)$ passes Criterion B.

5.2 Criterion C: Exact relationships among distances relating y, x_1, and x_2

The following Lemma 5.2 is useful since it gives an exact relationship that must hold for any z between $d(z, y)$, $d(z, x_1)$, $d(z, x_2)$ and $\alpha(q_1, h)$. Assume that $N = (V, A, r, X)$ has hybrid h with parents q_1, q_2, such that q_1 has a child x_1 which is a normal leaf, q_2 has a child x_2 which is a normal leaf, and h has a child y which is a normal leaf.

Lemma 5.2. *For every $z \in X$ other than y, x_1, x_2*
(1) $d(z, h) = \alpha(q_1, h)d(z, q_1) + \alpha(q_2, h)d(z, q_2)$

(2) $d(z,y)-\omega(h,y) = \alpha(q_1,h)[d(z,x_1)-\omega(q_1,x_1)]+\alpha(q_2,h)[d(z,x_2)-\omega(q_2,x_2)]$.
In particular, in the equiprobable case,
(3) $d(z,y)-\omega(h,y) = (1/2)[d(z,x_1)-\omega(q_1,x_1)] + (1/2)[d(z,x_2)-\omega(q_2,x_2)]$.

Proof. For each parent map p such that $p(h) = q_1$, let p' denote the complementary parent map whch agrees with p except that $p'(h) = q_2$. Then every parent map has the form either p or p'. For each $z \in X$, z other than y, x_1, x_2, note
$d(z,h;N_p) = d(z,q_1;N_p)$ since $\omega(q_1,h) = 0$ and
$d(z,h;N_{p'}) = d(z,q_2;N_{p'})$ since $\omega(q_2,h) = 0$.
Hence if $p(h) = q_1$, then
$d(z,h;G_p) = \alpha(q_1,h)d(z,h;N_p) + \alpha(q_2,h)d(z,h;N_{p'})$.
By Lemma 4.6 of [27], for each parent map p with $p(h) = q_1$, $d(z,h;N) = \sum[W(p)d(z,h:G_p);p(h)=q_1]$ where $W(p) = \prod(\alpha(p(h'),h') : h' \neq h)$ whence
$d(z,h;N)$
$= \sum[\alpha(q_1,h)W(p)d(z,h;N_p) + \sum\alpha(q_2,h)W(p)d(z,h;N_{p'}) : p(h) = q_1]$
$= \sum\alpha(q_1,h)W(p)d(z,h;N_p) + \sum\alpha(q_2,h)W(p)d(z,h;N_{p'})$
$= \sum\alpha(q_1,h)W(p)d(z,q_1;N_p) + \sum\alpha(q_2,h)W(p)d(z,q_2;N_{p'})$
(since $\omega(q_1,h) = \omega(q_2,h) = 0$)
$= \alpha(q_1,h)\sum[W(p)d(z,q_1;N_p)] + \alpha(q_2,h)\sum[W(p)d(z,q_2;N_{p'})]$.

But $d(z,q_1;N_p) = d(z,q_1;N_{p'})$ and $d(z,q_2;N_p) = d(z,q_2;N_{p'})$ since the path connecting z to q_1 in either case does not pass through h. Hence by Lemma 4.6 of [27] $d(z,q_1;N) = \sum[W(p)d(z,q_1;G_p) : p(h) = q_1]$ and similarly $d(z,q_2;N) = \sum[W(p)d(z,q_2;G_p) : p(h) = q_1]$. It follows that $d(z,h;N) = \alpha(q_1,h)d(z,q_1;N) + \alpha(q_2,h)d(z,q_2;N)$ proving (1).

Since $d(z,q_1;N)+\omega(q_1,x_1) = d(z,x_1;N)$ we have $d(z,q_1;N) = d(z,x_1;N) - \omega(q_1,x_1)$. Similarly, $d(z,q_2;N) = d(z,x_2;N) - \omega(q_2,x_2)$ and $d(z,h;N) = d(z,y;N) - \omega(h,y)$. If we substitute these into (1), we obtain (2). Finally we obtain (3) from (2) in the equiprobable case since then $\alpha(q_1,h) = 1/2$. □

Corollary 5.3. *The value* $d(z,y)-\omega(h,y)$ *should lie between the values* $[d(z,x_1)-\omega(q_1,x_1)]$ *and* $[d(z,x_2)-\omega(q_2,x_2)]$.

We say that the hybrid passes Criterion C2 if (2) from Lemma 5.2 holds, and it passes Criterion C3 if (3) from Lemma 5.2 holds.

5.3 Criterion D. Relationship among r, x_1, x_2, x_3

In the event of a non-equiprobable hybrid, Lemma 5.4 gives a relationship that most hold among r, x_1, x_2, and x_3.

Lemma 5.4. *In the case of a non-equiprobable hybrid* $(y;x_1,x_2,x_3)$ *we must have that*

$$d(r,x_1;N)+d(x_2,x_3;N) < d(r,x_2;N)+d(x_1,x_3;N) = d(r,x_3;N)+d(x_1,x_2;N)$$

Proof. From Figure 2, we see that for every parent map p we have that $rx_1|x_2x_3$ is a quartet in N_p. Hence

$$d(r,x_1;N_p)+d(x_2,x_3;N_p) < d(r,x_2;N_p)+d(x_1,x_3;N_p) = d(r,x_3;N_p)+d(x_1,x_2;N_p)$$

for each p. Taking the weighted sum where N_p is weighted by $Pr(p)$, we obtain the result. □

5.4 Criterion E. Conditions on signs in the equiprobable case.

This subsection gives some inequalities that must hold in the equiprobable case.

Let y, x_1, x_2 be distinct leaves (and distinct from r). In the equiprobable case for the network M define
$w_{rv}(M) := (d(r,x_1;M) + d(r,x_2;M) - d(x_1,x_2;M))/2$
$w_{q_1x_1}(M) := d(x_1,y;M) - d(r,y;M) + w_{rv}(M)$
$w_{q_2x_2}(M) := d(x_2,y;M) - d(r,y;M) + w_{rv}(M)$
$w_{vq_1}(M) := d(r,x_1;M) - w_{rv}(M) - w_{q_1x_1}(M)$
$w_{vq_2}(M) := d(r,x_2;M) - w_{rv}(M) - w_{q_2x_2}(M)$
$w_{hy}(M) := (d(y,x_1;M) + d(y,x_2;M) - d(x_1,x_2;M))/2$

These definitions are made plausible from the diagram of N in Figure 1. In the diagram from [27] $w_{rv}(M)$ is the estimate for the distance between r and v; $w_{q_1x_1}(N)$ is the estimate for the distance between q_1 and x_1; aq_2x_2 is the estimate for $d(q_2,x_2;N)$; $w_{vq_1}(N)$ estimates $d(v,q_1;N)$; $w_{vq_2}(N)$ estimates $d(v,q_2;N)$; and $w_{hy}(N)$ estimates $d(h,y;N)$. We now show that, if the distances are exactly known, then these estimates tell the exact values.

Lemma 5.5. *Assume that h is an equiprobable hybrid with parents q_1 and q_2. Suppose h has a normal child y which is a leaf. Suppose q_1 and q_2 have normal children x_1 and x_2 respectively which are leaves. Then the quantities $w_{rv}(N)$, $w_{q_1x_1}(N)$, $w_{q_2x_2}(N)$, $w_{vq_1}(N)$, $w_{vq_2}(N)$, and $w_{hy}(N)$ are all positive. Moreover, $d(q_1,x_1) = w_{q_1x_1}(N)$, $d(q_2,x_2) = w_{q_2x_2}(N)$, and $d(h,y) = w_{hy}(N)$.*

Proof. Note that for every complementary pair p and p', if $G_p = N_p \cup N_{p'}$ then the network in Figure 1a depicts part of G_p, where v is the most recent common ancestor of q_1 and q_2 in both N_p and $N_{p'}$. If w is another vertex distinct from r, y, x_1, and x_2 there are three possibilities for the placement of w: it could be attached on the path from r to v, on the path from v to q_1, or on the path from v to q_2.

Then in G_p we have the distances determined as follows:
$w_{rv}(G_p) := d(r,v;G_p) = (d(r,x_1;G_p) + d(r,x_2;G_p) - d(x_1,x_2;G_p))/2$
$w_{q_1x_1}(G_p) := d(q_1,x_1;G_p) = d(x_1,y;G_p) - d(r,y;G_p) + d(r,v;G_p)$
$w_{q_2x_2}(G_p) := d(q_2,x_2;G_p) = d(x_2,y;G_p) - d(r,y;G_p) + d(r,v;G_p)$
$w_{vq_1}(G_p) := d(v,q_1;G_p) = d(r,x_1;G_p) - w_{rv}(G_p) - w_{q_1x_1}(G_p)$
$w_{vq_2}(G_p) := d(v,q_2;G_p) = d(r,x_2;G_p) - w_{rv}(G_p) - w_{q_2x_2}(G_p)$
$w_{hy}(G_p) := d(h,y;G_p) = (d(y,x_1;G_p) + d(y,x_2;G_p) - d(x_1,x_2;G_p))/2$
and all are positive.

By Lemma 4.6 of [27] $w_{rv}(N) = \sum[W(p)w_{rv}(G_p) : p \in Par(N), p(h) = q_1]$ which is positive since $W(p) > 0$ and $w_{rv}(G_p) > 0$.

A similar argument proves the other conclusions. The identification of the values for $d(q_1, x_1)$, $d(q_2, x_2)$, and $d(h, y)$ follows from [27]. $\square$

A choice of $(y; x_1, x_2)$ will be said to satisfy Criterion E in the equiprobable case provided that the conclusion of Lemma 5.5 holds.

5.5 Criterion F. Conditions on signs in the general case

The material in this subsection is like that for Criterion E but applies to the general case which is not equiprobable.

For the general case in which the hybrid need not be equiprobable, we assume the existence of x_3 as in Figure 2. From [27], Lemma 4.9, we obtain the following explicit formulas:

Lemma 5.6. *Suppose h is hybrid with indegree 2 and parents q_1 and q_2. Suppose there is a normal path from q_1 to $x_1 \in X$, from q_2 to $x_2 \in X$, and from h to $y \in X$. Assume q_3 is such that there is a normal path from q_3 to q_2, a normal path from q_3 to $x_3 \in X$, but no directed path from q_3 to q_1. Suppose M is a phylogenetic X-network that is a subnetwork of N. Let*
(a) $w_{rv}(M) = [d(r, x_1; M) + d(r, x_3; M) - d(x_1, x_3; M)]/2$
$= [d(r, x_1; M) + d(r, x_2; M) - d(x_1, x_2; M)]/2$
(b) $w_{vq_3}(M) = [d(r, x_3; M) + d(x_1, x_2; M) - d(r, x_1; M) - d(x_3, x_2; M)]/2$
(c) $w_{q_3x_3}(M) = [d(r, x_3; M) + d(x_3, x_2; M) - d(r, x_2; M)]/2$
(d) $w_{hy}(M) = [d(y, x_2; M) + d(y, x_1; M) - d(x_1, x_2; M)]/2$
(e) $E_2(M) = d(x_1, y; M) - d(r, y; M) + w_{rv}(M)$
(f) $E_4(M) = d(x_2, y; M) - d(r, y; M) + w_{rv}(M)$
(g) $\alpha(M) = [2d(x_3, y; M) - 2w_{q_3x_3}(M) - 2w_{hy}(M) - d(r, x_1; M) + E_2(M) + 2w_{rv}(M) + E_4(M) - d(r, x_2; M) + 2w_{vq_3}(M)]/[4w_{vq_3}(M)]$
(h) $w_{vq_1}(M) = [d(r, x_1; M) - E_2(M) - w_{rv}(M)]/[2\alpha(M)]$
(i) $w_{q_3q_2}(M) = [d(x_3, y; M) - w_{q_3x_3}(M) - w_{hy}(M) - \alpha(M)(w_{vq_3}(M) + w_{vq_1}(M))]/(1 - \alpha(M))$
(j) $w_{q_1x_1}(M) = d(r, x_1; M) - w_{rv}(M) - w_{vq_1}(M)$
(k) $w_{q_2x_2}(M) = d(r, x_2; M) - w_{rv}(M) - w_{vq_3}(M) - w_{q_3,q_2}(M)$
(l) $C(M) = 2d(x_3, y; M) - 2w_{q_3x_3}(M) - 2w_{hy}(M) - d(r, x_1; M) + E_2(M) + 2w_{rv}(M) + E_4(M) - d(r, x_2; M) + 2w_{vq_3}(M)$
(m) $D(M) = 4w_{vq_3}(M)$.

Then
(i) $\alpha(q_1, h; N) = \alpha(N) = C(N)/D(N)$.
(ii) $d(q_1, x_1; N) = w_{q_1x_1}(N)$.
(iii) $d(q_2, x_2; N) = w_{q_2x_2}(N)$.

Indeed, $w_{rv}(N)$, $w_{vq_1}(N)$, $w_{q_1x_1}(N)$, $w_{hy}(N)$, $w_{vq_3}(N)$, $w_{q_3x_3}(N)$, $w_{q_3q_2}(N)$, $w_{q_2x_2}(N)$, and $\alpha(N)$ estimate the respective quantities $d(r, v; N)$, $d(v, q_1; N)$, $d(q_1, x_1; N)$, $d(h, y; N)$, $d(v, q_3; N)$, $d(q_3, x_3; N)$, $d(q_3, x_2; M)$, $\alpha(q_1, h; N)$ and give the exact values when the hypotheses are satisfied.

Lemma 5.7. *In the general case, the quantities $w_{rv}(N)$, $w_{vq_1}(N)$, $w_{q_1x_1}(N)$, $w_{hy}(N)$, $avq_3(N)$, $aq_3x_3(N)$, $aq_3q_2(N)$, $w_{q_2x_2}(N)$ of Lemma 5.6 are all positive. Moreover* $0 < \alpha(q_1, h; N) < 1$.

The proof is similar to that of Lemma 5.5.

A choice of $(y; x_1, x_2)$ with x_3 will be said to satisfy Criterion F provided that the conclusion of Lemma 5.7 holds.

5.6 Summary of the test for a hybrid.

Suppose we are given the tree-average distances for N. Using Theorem 4.5 we may eliminate all cherries. Hence we may assume that there are no cherries. By Theorem 3.2 there exists a hybrid vertex h with parents q_1 and q_2 such that h has a child y which is a leaf, q_1 has a child x_1 which is a leaf, and q_2 has a child x_2 which is a leaf. We consider all possibilities for y, x_1 and x_2. For each choice of (y, x_1, x_2) we perform the following checks:
(i) equiprobable(y, x_1, x_2): The choice passes the test provided it passes Criteria B, C3, and E.

(ii) general(y, x_1, x_2): Consider all possible $x_3 \in X$ distinct from y, x_1, x_2. The choice of (y, x_1, x_2) with x_3 passes the test provided that it passes Criteria B, C2, D, and F. In checking Criterion C2, we utilize the formulas of Lemma 5.6 to estimate $\omega(h, y)$, $\omega(q_1, x_1)$, $\omega(q_2, x_2)$, and $\alpha(q_1, h)$; note $\alpha(q_2, h) = 1 - \alpha(q_1, h)$.

We accept any (y, x_1, x_2) that passes either (i) or (ii).

6 Proof that the algorithm works if there is only one reticulation cycle

The main theorem of this paper is the following result:

Theorem 6.1. *Assume Assumptions 3.1 . Suppose $N = (V, A, r, X)$ has a single reticulation cycle and has its exact tree-average distances known. Then N is reconstructed by the algorithm from its tree-average distances.*

By Theorem 4.5 we may recognize any cherry that occurs in N and remove it by following the method described in Section 3. Hence we may assume that N has no cherries. Then N appears as in Figure 5 (possibly with some vertices deleted).

Suppose that the hybrid h has normal child $y \in X$ and parents q_1 and q_2 with respective normal children x_1 and x_2 in X. We say $A(v; v_1, v_2)$ is true if $v = y$, $v_1 = x_1$, $v_2 = x_2$ passes Criterion B. In other words, for $w \in X$ let $W_v(N) = d(w, v; N) + d(v_1, v_2; N)$, $W_{v_1}(N) = d(w, v_1; N) + d(v, v_2; N)$, $W_{v_2}(N) = d(w, v_2; N) + d(v, v_1; N)$. Then $A(v; v_1, v_2)$ is true iff for all $w \in X$ other than v, v_1, v_2 we have both $W_{v_1}(N) < W_v(N)$ and $W_{v_2}(N) < W_v(N)$. By lemma 5.1, $A(y; x_1, x_2)$ is true. Note that by symmetry, $A(a; b, c)$ is true if and only if $A(a; c, b)$ is true.

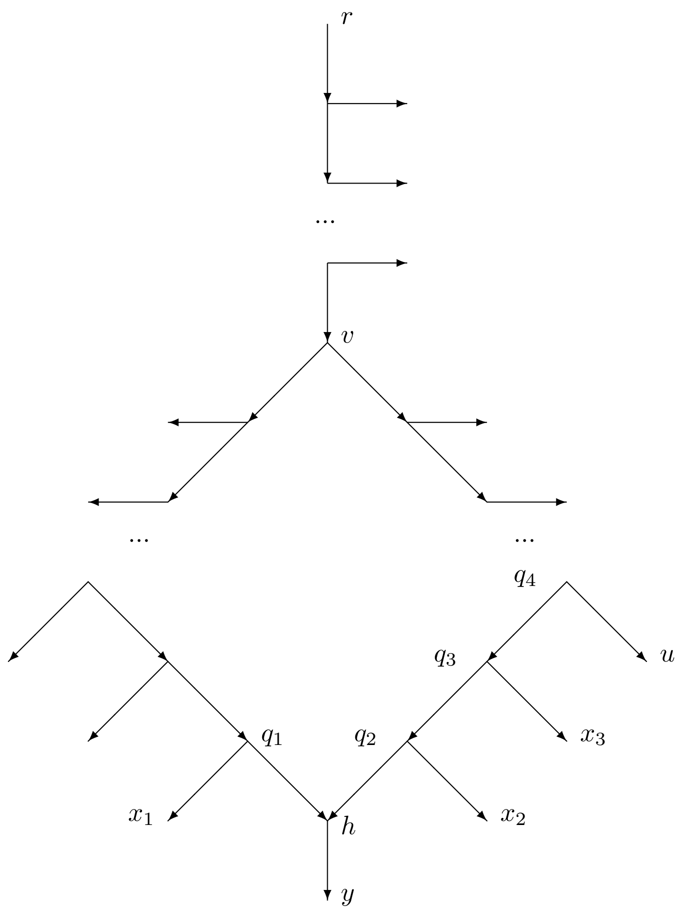

Figure 5: The reality when there is a single reticulation cycle to hybrid h and there are no cherries. Here the parents of h are q_1 and q_2. Moreover, h, q_1, and q_2 have respectively normal children y, x_1, x_2 in X.

Observe that we are interested only in possibilities for a, b, c in X such that $A(a; b, c)$ is true and also such that a, b, c are all possibilities for being children of a hybrid and the parents of a hybrid. Consequently none of a, b, c can be the root r. On the other hand, w in the test could possibly equal r.

Lemma 6.2. *Suppose $A(a; b, c)$ is true. Then for all $e \in X$, $e \notin \{a, b, c\}$ both $A(a; b, e)$ and $A(a; e, c)$ are false.*

Proof. If $A(a; b, e)$ is true then choosing $w = c$ we have $d(c, b) + d(a, e) < d(a, c) + d(b, e)$. But since $A(a; b, c)$ is true, then choosing $w = e$ yields $d(e, b) + d(a, c) < d(e, a) + d(b, c)$, a contradiction. Hence $A(a; b, e)$ is false. A symmetric argument shows $A(a;e,c)$ is false. $\square$

Lemma 6.3. *Suppose there is a 4-set $\{a,b,c,e\}$ such that for all parent maps p, the same quartet tree is in N_p. Then $A(a;b,c)$ is false.*

Proof. Suppose that the common quartet tree is $uv|xy$ for a permutation u, v, x, y of a, b, c, e. Then $d(u,v)+d(x,y) < d(u,x)+d(v,y) = d(u,y)+d(v,x)$. But if $A(a;b,c)$ is true then there is a unique strict maximum among the three quantities $d(a,b)+d(c,e)$, $d(a,c)+d(b,e)$, $d(a,e)+d(b,c)$, a contradiction. $\square$

Lemma 6.4. *Suppose there is a subset S of X such that $|S| \geq 4$ and for all parent maps p, the restriction $N_p|S$ is the same tree. Then for $\{a,b,c\} \subseteq S$, $A(a;b,c)$ is false.*

Proof. Since $|X| \geq 4$ we may suppose $w \in S - \{a,b,c\}$. Then for all p, the trees N_p induce the same quartet on the 4-set $\{a,b,c,w\}$. By Lemma 6.3, $A(a;b,c)$ is false. $\square$

Lemma 6.5. *Assume that for a nonempty collection of parent maps p, we have $wx|yz$ in N_p and for a nonempty collection of parent maps p we have $wy|xz$ in N_p but there is no parent map p such that in N_p we have $wz|xy$. Then $A(y;x,z)$, $A(y;z,x)$, $A(x;y,z)$, and $A(x;z,y)$ are false.*

Restatement. Suppose for the 4-set $\{w,x,y,z\}$ exactly two quartet trees appear in the various N_p, including $wy|xz$. Then $A(y;x,z)$, $A(y;z,x)$, $A(x;y,z)$, and $A(x;z,y)$ are false. This leaves open the possibility that $A(z;x,y)$ is true.

Proof. Let $P_1 \subset Par(N)$ be the set of p such that $wx|yz$ in N_p, and let P_2 be the set of p such that $wy|xz$ in N_p. Then $Par(N) = P_1 \cup P_2$ since the other quartet never occurs.

For $p \in P_1$,

$$d(w,x;N_p)+d(y,z;N_p) < d(w,y;N_p)+d(x,z;N_p) = d(w,z;N_p)+d(x,y;N_p).$$

For $p \in P_2$,

$$d(w,y;N_p)+d(x,z;N_p) < d(w,x;N_p)+d(y,z;N_p) = d(w,z;N_p)+d(x,y;N_p).$$

Taking the weighted sums it follows
$\sum[Pr(p)d(w,y;N_p) : p \in P_2] + \sum Pr(p)d(x,z;N_p)p \in P_2]$
$< \sum[Pr(p)d(w,z;N_p) : p \in P_2] + \sum[Pr(p)d(x,y;N_p) : p \in P_2]$.
Add to each side $\sum[Pr(p)d(w,y;N_p) : p \in P_1] + \sum[Pr(p)d(x,z;N_p) : p \in P_1]$.
We obtain
$\sum[Pr(p)d(w,y;N_p) : p \in P_2] + \sum Pr(p)d(x,z;N_p)p \in P_2]$
$+ \sum[Pr(p)d(w,y;N_p) : p \in P_1] + \sum[Pr(p)d(x,z;N_p) : p \in P_1]$
$< \sum[Pr(p)d(w,z;N_p) : p \in P_2] + \sum[Pr(p)d(x,y;N_p) : p \in P_2]$
$+ \sum[Pr(p)d(w,y;N_p) : p \in P_1] + \sum[Pr(p)d(x,z;N_p) : p \in P_1]$.

But the left side
$= [\sum[Pr(p)d(w,y;N_p) : p \in P_1] + \sum[Pr(p)d(w,y;N_p) : p \in P_2]]$

$+[\sum[Pr(p)d(x,z;N_p) : p \in P_1] + \sum Pr(p)d(x,z;N_p) : p \in P_2]]$
$= d(w,y;N) + d(x,z;N)$
and the right side $= [\sum[Pr(p)d(w,z;N_p) : p \in P_2] + \sum[Pr(p)d(x,y;N_p) : p \in P_2]]$
$+[\sum[Pr(p)d(w,z;N_p) : p \in P_1] + \sum[Pr(p)d(x,y;N_p) : p \in P_1]]$
$= d(w,z;N) + d(x,y;N)$.

Hence $d(w,y;N) + d(x,z;N) < d(w,z;N) + d(x,y;N)$. Yet if $A(y;x,z)$ is true then $d(w,z;N)+d(x,y;N) < d(w,y;N)+d(x,z;N)$, a contradiction. Thus $A(y;x,z)$ is false, and equivalently $A(y;z,x)$ is false.

If we interchange x and y we obtain by symmetry that $A(x;y,z)$ and $A(x;z,y)$ are false. □

Note that Lemma 6.5 may be contrasted with Lemma 5.1, which says that if for all w, $wx|yz$ in some N_p and $wz|xy$ in some N_p but never $wy|xz$ then $A(y;x,z)$ is true.

Lemma 6.6. *Suppose that $N = (V, A, r, X)$ satisfies that $A(y_1;x_{11},x_{12})$, $A(y_1;x_{12},x_{11})$, $A(y_2;x_{21},x_{22})$, $A(y_2;x_{21},x_{21})$, $\cdots$, $A(y_k;x_{k1},x_{k2})$, $A(y_k;x_{k2},x_{k1})$ are the only acceptances that are true. Form $N' = (V', A', r, X')$ by picking a leaf $z \in X$ and replacing it by a cherry $\{z_1, z_2\}$. (Thus we add z_1, z_2, and arcs (z, z_1), (z, z_2), with positive weights, removing z from X but adding z_1 and z_2 to make X'.) Then in N'*
(i) If there is no i such that z is y_i or x_{i1} or x_{i2}, then $A(y_i;x_{i1},x_{i2})$ and $A(y_i;x_{i2},x_{i1})$ remain true in N'.
(ii) Suppose $A(a;b,c)$ is true in N'. Then $\{a,b,c\} \subset X - \{z\}$, so none of a, b, c is z_1 or z_2.

Proof. Note that for each $u \in X$, $d(u,z_1) = d(u,z) + \omega(z,z_1)$ and $d(u,z_2) = d(u,z) + \omega(z,z_2)$ by the construction.

(i) Suppose z is not y_i or x_{i1} or x_{i2}. In N we have $A(y_i;x_{i1},x_{i2})$ is true. Hence for every $w \in X$ other than y_i, x_{i1}, x_{i2} we have
$d(w,x_{i1};N) + d(y_i,x_{i2};N) < d(w,y_i;N) + d(x_{i1},x_{i2};N)$
$d(w,x_{i2};N) + d(y_i,x_{i1};N) < d(w,y_i;N) + d(x_{i1},x_{i2};N)$.

In N', for each $w \neq z_1, z_2$ the same inequalities will still hold. I claim the same inequalities also hold for $w = z_1$ or z_2. To see this note that choosing $w = z$ yields
$d(z,x_{i1};N) + d(y_i,x_{i2};N) < d(z,y_i;N) + d(x_{i1},x_{i2};N)$ and
$d(z,x_{i2};N) + d(y_i,x_{i1};N) < d(z,y_i;N) + d(x_{i1},x_{i2};N)$.

Hence when we add $\omega(z,z_1)$ to each side we obtain
$\omega(z,z_1) + d(z,x_{i1};N) + d(y_i,x_{i2};N) < \omega(z,z_1) + d(z,y_i;N) + d(x_{i1},x_{i2};N)$ and
$\omega(z,z_1) + d(z,x_{i2};N) + d(y_i,x_{i1};N) < \omega(z,z_1) + d(z,y_i;N) + d(x_{i1},x_{i2};N)$.

Hence $d(z_1,x_{i1};N') + d(y_i,x_{i2};N') < d(z_1,y_i;N') + d(x_{i1},x_{i2};N')$ and
$d(z_1,x_{i2};N') + d(y_i,x_{i1};N') < d(z_1,y_i;N') + d(x_{i1},x_{i2};N')$
so $A(y_i;x_{i1},x_{i2})$ is true in N'. The same argument shows $A(y_i;x_{i2},x_{i1})$ is true in N'.

(ii) Suppose $A(a;b,c)$ is true in N'. I claim that $\{a,b,c\} \subset X - \{z\}$, so none of a, b, c is z_1 or z_2.

Case 1. Suppose $a = z_1$ so $A(z_1; b, c)$ in N' and neither b nor c equals z_2. Then for all w, $d(w, b) + d(z_1, c) < d(w, z_1) + d(b, c)$ and $d(w, c) + d(z_1, b) < d(w, z_1) + d(b, c)$. In particular if $w = z_2$ then $d(z_2, b) + d(z_1, c) < d(z_2, z_1) + d(b, c)$ and $d(z_2, c) + d(z_1, b) < d(z_2, z_1) + d(b, c)$. But relating distances to z_2 and z_1 we have
$\omega(z, z_2) + d(z, b) + \omega(z, z_1) + d(z, c) < \omega(z, z_2) + \omega(z, z_1) + d(z, z) + d(b, c)$ and
$\omega(z, z_2) + d(z, c) + \omega(z, z_1) + d(z, b) < \omega(z, z_2) + \omega(z, z_1) + d(z, z) + d(b, c)$.

Hence $d(z, b) + d(z, c) < d(b, c)$ which contradicts the triangle inequality, shown to be true in [27].

Case 2. Suppose $a = z_1$ so $A(z_1; b, c)$ in N' and $b = z_2$, so $A(z_1; z_2, c)$ is true in N'. Then for all $w \in X - \{z\}$, $d(w, z_2) + d(z_1, c) < d(w, z_1) + d(z_2, c)$ and $d(w, c) + d(z_1, z_2) < d(w, z_1) + d(z_2, c)$. Substituting $d(w, z_2) = d(w, z) + \omega(z, z_2)$ etc. we obtain $\omega(z, z_2) + d(w, z) + d(z, c) + \omega(z, z_1) < d(w, z) + \omega(z, z_1) + d(z, c) + \omega(z, z_2)$ and $d(w, c) + d(z, z) + \omega(z, z_1) + \omega(z, z_2) < d(w, z) + \omega(z, z_1) + d(z, c) + \omega(z, z_2)$. Hence $d(w, z) < d(w, z)$ contradicting that d is a metric.

It follows that a cannot be z_1, and a symmetric argument shows $a \neq z_2$.

Case 3. Suppose $a \notin \{z_1, z_2\}$ but $b = z_1$, $c \neq z_2$ yet $A(a; z_1, c)$ is true. Then for all w, $d(w, z_1) + d(a, c) < d(w, a) + d(z_1, c)$ and $d(w, c) + d(z_1, a) < d(w, a) + d(z_1, c)$. In particular, for $w = z_2$, $d(z_2, , z_1) + d(a, c) < d(z_2, a) + d(z_1, c)$ and $d(z_2, c) + d(z_1, a) < d(z_2, a) + d(z_1, c)$.

Thus $\omega(z, z_2) + \omega(z, z_1) + d(a, c) < d(z, a) + d(z, c) + \omega(z, z_2) + \omega(z, z_1)$ and $\omega(z, z_2) + \omega(z, z_1) + d(z, c) + d(z, a) < d(z, a) + d(z, c) + \omega(z, z_2) + \omega(z, z_1)$.

This shows that $d(a, c) < d(z, a) + d(z, c)$ and $d(z, c) + d(z, a) < d(z, a) + d(z, c)$ so $0 < 0$, a contradiction.

Case 4. Suppose $a \notin \{z_1, z_2\}$ but $b = z_1$, $c = z_2$ and $A(a; z_1, z_2)$ is true. Then for w distinct from z_1, z_2, and a, we have $d(w, z_1) + d(z_2, a) < d(w, a) + d(z_1, z_2)$ and $d(w, z_2) + d(z_1, a) < d(w, a) + d(z_1, z_2)$.

Since $d(w, z_1) = d(w, z) + \omega(q, z_1)$, it follows that
$\omega(z, z_1) + \omega(z, z_2) + d(w, z) + d(z, a) < \omega(z, z_1) + \omega(z, z_2) + d(w, a)$
and $\omega(z, z_1) + \omega(z, z_2) + d(w, z) + d(z, a) < \omega(z, z_1) + \omega(z, z_2) + d(w, a)$. Hence $d(w, z) + d(z, a) < d(w, a)$ and $d(w, z) + d(z, a) < d(w, a)$ contradicting the triangle inequality. This completes the proof of (ii). $\square$

Lemma 6.7. *Assume the hypotheses of Lemma 6.6 and the construction of N'. If $z = y_i$, then $A(z; x_{i1}, x_{i2})$ is meaningless for N' since $z \notin X'$. Moreover, $A(z_1; x_{i1}, x_{i2})$ and $A(z_2; x_{i1}, x_{i2})$ are false for N'. Similarly if $z = x_{i1}$ so $A(y_i; z, x_{i2})$ is true in N, then $A(y_i; z, x_{i2})$ is meaningless in N', while $A(y; z_1, x_{i2})$ is false in N'.*

Proof. Suppose $z = y_i$. Since $A(z; x_{i1}, x_{i2})$ is true for every $w \in X$ other than $y_i = z$, x_{i1}, x_{i2} we have
$d(w, x_{i1}; N) + d(z, x_{i2}; N) < d(w, z; N) + d(x_{i1}, x_{i2}; N)$ and
$d(w, x_{i2}; N) + d(z, x_{i1}; N) < d(w, z; N) + d(x_{i1}, x_{i2}; N)$.

I claim that $A(z_1; x_{i1}, x_{i2})$ is false. Adding $\omega(z, z_1)$ to the above yields that for $w \in X$ other than z, x_{i1}, x_{i2} we have
$d(w, x_{i1}; N) + \omega(z, z_1) + d(z, x_{i2}; N) < d(w, z; N) + \omega(z, z_1) + d(x_{i1}, x_{i2}; N)$
$d(w, x_{i2}; N) + d(z, x_{i1}; N) + \omega(z, z_1) < d(w, z; N) + \omega(z, z_1) + d(x_{i1}, x_{i2}; N)$.

Hence $d(w, x_{i1}; N') + d(z_1, x_{i2}; N') < d(w, z_1; N') + d(x_{i1}, x_{i2}; N')$ and $d(w, x_{i2}; N') + d(z_1, x_{i1}; N') < d(w, z_1; N') + d(x_{i1}, x_{i2}; N')$.

The only remaining issue is whether the same is true for $w = z_2$. We ask whether
$d(z_2, x_{i1}; N') + d(z_1, x_{i2}; N') < d(z_2, z_1; N') + d(x_{i1}, x_{i2}; N')$ and
$d(z_2, x_{i2}; N') + d(z_1, x_{i1}; N') < d(z_2, z_1; N') + d(x_{i1}, x_{i2}; N')$.

But $d(z_2, z_1) = \omega(z, z_1) + \omega(z, z_2)$ and $d(z_2, x_{i1}; N') = d(z, x_{i1}; N') + \omega(z, z_2)$. Hence these inequalities would imply $d(z, x_{i1}; N) + \omega(z, z_2) + d(z, x_{i2}; N) + \omega(z, z_1) < \omega(z, z_1) + \omega(z, z_2) + d(x_{i1}, x_{i2}; N)$ and
$\omega(z, z_2) + d(z, x_{i2}; N) + \omega(z, z_1) + d(z, x_{i1}; N) < \omega(z, z_1) + \omega(z, z_2) + d(x_{i1}, x_{i2}; N)$.

But then we obtain $d(z, x_{i1}; N) + d(z, x_{i2}; N) < d(x_{i1}, x_{i2}; N)$ and $d(z, x_{i2}; N) + d(z, x_{i1}; N) < d(x_{i1}, x_{i2}; N)$, which violates the triangle inequality. Hence $A(z_1; x_{i1}, x_{i2})$ is false. A symmetric argument shows $A(z_2; x_{i1}, x_{i2})$ is false.

A similar argument applies if $z = x_{i1}$ or x_{i2}. The lemma follows. □

We now give the proof of Theorem 6.1. We are assuming there are no cherries. See Figure 5.

Since there is a single reticulation cycle, if h and y are removed, then we obtain a tree T. Consider the path P_1 in T from the root r to x_1 and the path P_2 from r to x_2. These paths diverge at a point v (possibly $v = r$). Suppose x is a leaf other than y, x_1, or x_2. Then the path from r to x in T must depart from $P_1 \cup P_2$ at a point w, whence there is a path P_x from w to x disjoint from $P_1 \cup P_2$ except for w. But a path from w that starts along P_x toward x and has a maximal number of arcs must end at a leaf with parent q; and if this maximal length is greater than one, then the other child of q must also be a leaf by maximality. Since neither can be hybrid (since h is the only hybrid), it follows that if the maximal path from w in that direction has length greater than one, then N has a cherry, a contradiction. It follows that x is a leaf with parent w.

It is possible that h is equiprobable. It is also possible that q_2 has an ancestor q_3 such that there is a normal path from q_3 to q_2 and there is no path from q_3 to q_1, and there is a normal path from q_3 to $x_3 \in X$ that is disjoint from the normal path to q_2 except for q_3. In either situation if we can correctly identify y, x_1, x_2 and $\alpha(q_1, h)$, then we may remove the correct hybrid. Once this has occurred, there are no more hybrid vertices and successive identification of cherries may take place. Hence the theorem is true provided we can correctly identify y, x_1, x_2, and $\alpha(q_1, h)$. Of course, in either case the correct choice satisfies the criteria of Section 5. Hence we must show that there is no false signal to accept a different choice.

Write $A(v; v_1, v_2, \alpha)$ if the algorithm accepts the hybrid h' as having child v and parents q_1 and q_2 with respective normal children v_1 and v_2 in X, with $\alpha(q_1, h') = \alpha$. Clearly, if $A(v; v_1, v_2)$ is false, then for all α, $A(v; v_1, v_2, \alpha)$ is false.

The proof of Theorem 6.1 will involve a sequence of claims.

Claim 1. *Suppose none of v, v_1, v_2 is y. Then $A(v; v_1, v_2)$ is false.*

Proof. Let $S = X - \{y\}$. Then Lemma 6.4 applies since $\{v, v_1, v_2\} \subseteq S$. $\square$

Hence if $A(v; v_1, v_2)$ is true then one of v, v_1, v_2 is y.

Claim 2. *Suppose $A(y; x_1, v)$ is true. Then $v = x_2$.*

Proof. Since $A(y; x_1, x_2)$ is true, then by Lemma 6.2 for $v \neq x_2$ we must have that $A(y; x_1, v)$ is false. $\square$

Similarly we have

Claim 3. *Suppose $A(y; x_2, v)$ is true. Then $v = x_1$.*

Claim 4. *If $\{u_1, u_2\}$ is disjoint from $\{x_1, x_2\}$, then $A(y; u_1, u_2)$ is false.*

Proof. Let $T = N$ with h removed as well as all arcs involving h, so T is a directed tree. See Figure 5. Note that if $p(h) = q_1$ then N_p consists of T with x_1 replaced by a cherry $\{x_1, y\}$, while if $p'(h) = q_2$ the $N_{p'}$ consists of T with x_2 replaced by a cherry $\{x_2, y\}$. All leaves other than x_1 and x_2 have their parent on the path P_1 or P_2. In particular, u_1 and u_2 are situated in this manner. The vertex where P_1 and P_2 diverge is denoted v.

We analyze 7 cases. Case 1. Suppose u_1 and u_2 both branch off the path from r to v. Then the quartet for $\{u_1, u_2, y, x_1\}$ in N_p is always $u_1u_2|yx_1$. Hence $A(y; u_1, u_2)$ is false by Lemma 6.3.

Case 2. Suppose u_1 branches off the path from r to v but u_2 branches off the path from v to x_2. If $p(h) = q_1$ then the quartet for $\{u_1, u_2, y, x_1\}$ in N_p is $u_1u_2|yx_1$, while if $p'(h) = q_2$ then the quartet in $N_{p'}$ is $yu_2|u_1x_1$. Hence by Lemma 6.5 with $w = x_1$, $A(y; u_1, u_2)$ is false.

Case 3. Suppose u_1 branches off the path from r to v but u_2 branches off the path from v to x_1. An argument symmetric to that of Case 2 with $w = x_2$ yields a contradiction.

Case 4. Suppose u_2 branches off the path from r to v but u_1 branches off the path from v to x_2. An argument symmetric to that of Case 2 yields a contradiction.

Case 5. Suppose u_2 branches off the path from r to v but u_1 branches off the path from v to x_1. An argument symmetric to that of Case 3 yields a contradiction.

Case 6. Suppose u_1 and u_2 both branch off the path from v to x_1. By symmetry we may assume that the branch for u_1 is closer to r than the branch for u_2. Let $w = x_1$ and consider the quartets for $\{x_1, y, u_1, u_2\}$. If $p(h) = q_1$ then N_p has the quartet $x_1y|u_1u_2$ while if $p'(h) = q_2$ then $N_{p'}$ has the quartet $x_1u_2|u_1y$. Hence by Lemma 6.5, $A(y; u_1, u_2)$ is false.

Case 7. Suppose u_1 and u_2 both branch off the path from v to x_2. An argument symmetric to Case 6 yields a contradiction.

This completes the proof of Claim 4. $\square$

It follows from Claims 2, 3, and 4 that if $A(v; u_1, u_2)$ is true, then either the correct hybrid is found via $v = y$; $\{u_1, u_2\} = \{x_1, x_2\}$ or else we may assume $u_1 = y$. We must eliminate the possibility that $u_1 = y$.

Claim 5. *Suppose $A(k; y, u)$ is true. Then $k = x_1$ or x_2.*

Proof. As with Claim 4, the proof considers 7 cases.

Case 1. Suppose k and u both branch off the path from r to v. Then for all parent maps p, N_p has the quartet $x_1y|uk$. Hence $A(k; y, u)$ is false by Lemma 6.3.

Case 2. Suppose k and u both branch off the path from v to x_1, $k \neq x_1$, $u \neq x_1$, k further from r than is u. Let $w = x_1$. If $p(h) = q_1$ then N_p has the quartet $yx_1|ku$ and if $p'(h) = q_2$ then $N_{p'}$ has the quartet $x_1k|yu$. Hence by Lemma 6.5, $A(k; y, u)$ is false.

Case 3. Suppose k and u both branch off the path from v to x_1, $k \neq x_1$, $u \neq x_1$, k closer to r than is u. Let $w = x_2$. If $p(h) = q_1$ then N_p has the quartet $yu|kx_2$ and if $p'(h) = q_2$ then $N_{p'}$ has the quartet $ku|yx_2$. By Lemma 6.5, $A(k; y, u)$ is false.

Case 4. Suppose k and u both branch off the path from v to x_2, $k \neq x_2$, $u \neq x_2$, k further from r than is u. Then $A(k; y, u)$ is false by an argument symmetric to Case 2.

Case 5. Suppose k and u both branch off the path from v to x_2, $k \neq x_2$, $u \neq x_2$, k closer to r than is u. Then $A(k; y, u)$ is false by an argument symmetric to the proof of Case 3.

Case 6. Suppose k branches off the path from v to x_1, $k \neq x_1$, while u branches off the path from v to x_2, $u \neq x_2$. Then $A(k; y, u)$ is false. To see this, let $w = x_1$. If $p(h) = q_1$ then N_p has the quartet $yx_1|ku$ and if $p'(h) = q_2$ then $N_{p'}$ has the quartet $x_1k|yu$. By Lemma 6.5, $A(k; y, u)$ is false.

Case 7. Suppose k branches off the path from v to x_2, $k \neq x_2$, while u branches off the path from v to x_1, $u \neq x_1$. Then $A(k; y, u)$ is false by an argument symmetric to Case 6.

This completes the proof of Claim 5. □

As a result of Claim 5, the only possible accepted false hybrids are $A(x_1; y, u)$ or $A(x_2; y, u)$. Since the cases are symmetric, we will study only the possibility of $A(x_2; y, u)$.

Claim 6. *Suppose the attachment point of u lies on the path from v to x_1. Then $A(x_2; y, u)$ is false.*

Proof. Let $w = r$. We consider the 4-set $\{x_2, y, u, r\}$. If $p(h) = q_1$ then we obtain $rx_2|yu$. If $p'(h) = q_2$ then we obtain $ru|yx_2$. By Lemma 6.5, $A(x_2; y, u)$ is false. □

Claim 7. *Suppose the attachment point of u lies on the path from r to v. Then $A(x_2; y, u)$ is false.*

Proof. Note that $u \neq r$. Consider the quartet $\{r, x_2, y, u\}$. For every parent map p, N_p displays $ru|x_2y$. Hence $A(x_2; y, u)$ is false by Lemma 6.3. □

If $A(x_2; y, u)$ is true, it follows that the attachment point of u lies on the path from v to x_2. The next claim shows that this attachment point must be the parent of q_2.

Claim 8. *Suppose $A(x_2; y, u)$ is true. If the parent of x_2 is q_2, then the parent q_3 of q_2 satisfies that q_3 is not $\leq q_1$, and $u = x_3$ is the other child of q_3 besides q_2.*

Proof. Suppose first that the parent q_3 of q_2 satisfies that q_3 is not $\leq q_1$ and that $x_3 \neq u$ is the other child of q_3. Let $w = x_3$. We consider the quartet for $\{x_2, y, x_3, u\}$. If $p(h) = q_1$ then the quartet tree is $x_3x_2|yu$; if $p'(h) = q_2$ then the quartet is $x_3u|yx_2$. Hence $A(x_2; y, u)$ is false by Lemma 6.5.

Now suppose that the parent of q_2 is $v \leq q_1$ (so no x_3 exists). Then u must satisfy the hypotheses of either Claim 6 or Claim 7, and then that claim would lead to a contradiction. Thus this case cannot occur. □

It follows that the only other possibility besides the correct $A(y; x_1, x_2)$ is that $A(x_2; y, x_3)$ holds (or a symmetric case $A(x_1; y, x_0)$ where the parent of x_0 is the grandparent of x_1). The proofs above show that Lemma 5.1 (Criterion B) eliminates all other possibilities. In fact, $A(x_2; y, x_3)$ is consistent with Lemma 5.1. To see this, if $w \neq y$, x_2, x_3, then if $p(h) = q_1$ we have $wy|x_2x_3$ in N_p, while if $p'(h) = q_2$ then $N_{p'}$ displays $wx_3|yx_2$, the same predictions as $A(x_2; y, x_3)$. Hence other criteria besides Criterion B are now needed.

Claim 9. *$A(x_2; y, x_3, 1/2)$ is false. More specifically, the equiprobable case is eliminated by Criterion C.*

Proof. Suppose in a network M , x_2 is the child of a hybrid, while y and x_3 are the normal children of the parents of the hybrid, yet the tree-average distances are those given for N. See Figure 6 for two possibilities M_1 and M_2 for M.

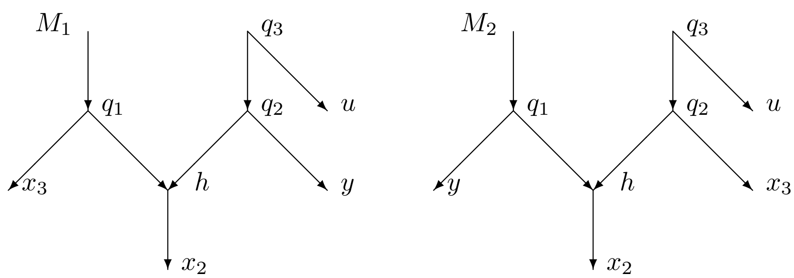

Figure 6: Networks M_1 or M_2 might suggest that $A(x_2; y, x_3, \alpha)$ is true.

We test whether M could pass the tests as an equiprobable hybrid. This would require by Criterion C, that for any z,
$d(z, x_2) - \omega(h, y) = (1/2)[d(z, y) - \omega(q_1, x_1)] + (1/2)[d(z, x_3) - \omega(q_2, x_2)]$. Moreover, by Lemma 5.5 we have
$w_{rv}(M) = (d(r, y) + d(r, x_3) - d(y, x_3))/2$
$\omega(q_1, x_1) = w_{q_1x_1}(M) := d(y, x_2) - d(r, x_2) + w_{rv}(M)$
$\omega(q_2, x_2) = w_{q_2x_2}(M) := d(x_3, x_2) - d(r, x_2) + w_{rv}(M)$
$\omega(h, y) = w_{hy}(M) := (d(x_2, y) + d(x_2, x_3) - d(y, x_3))/2.$

This simplifies to
$2d(z,x_2) + d(r,y) + d(r,x_3) = d(z,y) + 2d(r,x_2) + d(z,x_3)$

To check whether this equation holds, we substitute the true values obtained from N.
$d(r,y) = d(r,v) + \omega(h,y) + (1/2)d(v,q_1) + (1/2)d(v,q_3) + (1/2)\omega(q_3,q_2)$
$d(r,x_3) = d(r,v) + d(v,q_3) + \omega(q_3,x_3)$
$d(r,x_2) = d(r,v) + d(v,q_3) + \omega(q_3,q_2) + \omega(q_2,x_2)$

This leads to $2d(z,x_2) + d(r,v) + \omega(h,y) + (1/2)d(v,q_1) + (1/2)d(v,q_3) + (1/2)\omega(q_3,q_2) + d(r,v) + d(v,q_3) + \omega(q_3,x_3) = d(z,y) + 2d(r,v) + 2d(v,q_3) + 2\omega(q_3,q_2) + 2\omega(q_2,x_2) + d(z,x_3)$

This must hold for all $z \in X$ distinct from x_2, y, x_3, in particular for $z = x_1$. The true values are $d(x_1,x_2) = \omega(q_1,x_1)+d(v,q_1)+d(v,q_3)+\omega(q_3,q_2)+\omega(q_2,x_2)$
$d(x_1,y) = \omega(q_1,x_1) + \omega(h,y) + (1/2)d(v,q_1) + (1/2)d(v,q_3) + (1/2)\omega(q_3,q_2)$
$d(x_1,x_3) = \omega(q_1,x_1) + d(v,q_1) + d(v,q_3) + \omega(q_3,x_3)$.

When these values are substituted, the equation simplifies to $d(v,q_1) = 0$, which contradicts Assumption A(3).

Hence $A(x_2;y,x_3,1/2)$ is false since it fails Criterion C under the assumption that the hybrid is equiprobable. $\square$

There remains only the possibility that $A(x_2;y,x_3,\alpha)$ is true, where $\alpha \neq 1/2$, and in reality the attachment point of x_3 is the parent of x_2.

Claim 10. *If $\alpha \neq 1/2$, then $A(x_2;y,x_3,\alpha)$ is false.*

Proof. Suppose that $A(x_2;y,x_3,\alpha)$ is true. Since $\alpha \neq 1/2$, there must be an ancestor of the hybrid on one side of the reticulation cycle with a normal child u in X. Thus we assume that one of M_1 and M_2 in Figure 6 passes all the tests for being a hybrid with child x_2.

If M_1 is true then for every p, N_p must display $r,x_3|u,y$ by Criterion D. Hence $d(r,x_3) + d(u,y) < d(r,u) + d(x_3,y) = d(r,y) + d(u,x_3)$. If u attaches between v and x_1 then when $p(h) = q_1$ it follows that N_p displays $uy|rx_3$ while when $p'(h) = q_2$ then $N_{p'}$ displays $ur|yx_3$ whence by Lemma 3.5 we cannot have $d(r,x_3) + d(u,y) < d(r,u) + d(x_3,y) = d(r,y) + d(u,x_3)$. If u attaches between r and v then for all p, N_p displays $ru|yx_3$, whence $d(r,u) + d(y,x_3) < d(r,y) + d(u,x_3) = d(r,x_3) + d(u,y)$, a contradiction. Finally, if u attaches between v and x_2 then when $p(h) = q_1$, N_p displays $ux_3|ry$ while when $p'(h) = q_2$, $N_{p'}$ displays $yx_3|ru$; hence by Lemma 3.5 we cannot have $d(r,x_3) + d(u,y) < d(r,u) + d(x_3,y) = d(r,y) + d(u,x_3)$. Hence in every case, Criterion D fails, so in the event of M_1, $A(x_2;y,x_3,\alpha)$ is false.

Hence we must assume that M_2 depicts the assumed situation. In this situation, there is no path from q_3 to q_1, but there is a normal path from q_3 to q_2. It need not be the case that q_3 is actually a parent of q_2.

Note that if M_2 is true, then for every p, N_p must display $ry|ux_3$. Hence $d(r,y) + d(u,x_3) < d(r,u) + d(y,x_3) = d(r,x_3) + d(u,y)$. But the reality is Figure 5. If u attaches between r and v then when $p(h) = q_1$ N_p must display $ru|yx_3$ while when $p'(h) = q_2$ then $N_{p'}$ must display $ru|yx_3$. Hence we must have $d(r,u)+d(y,x_3) < d(r,y)+d(u,x_3) = d(r,x_3)+d(u,y)$, a contradiction. If

u attaches between v and x_1 then when $p(h) = q_1$ N_p must display $uy|rx_3$ while when $p'(h) = q_2$, $N_{p'}$ must display $yx_3|ru$. Since these are different by Lemma 3.5 we cannot have $d(r, y) + d(u, x_3) < d(r, u) + d(y, x_3) = d(r, x_3) + d(u, y)$.

It follows that in reality u must attach between v and x_2. Since x_3 attaches to an ancestor of x_2, it follows that the attachment point q_4 of u must lie between v and q_3. If $p(h) = q_1$ then for the 4-set $\{u, y, x_3, r\}$, N_p must display $ux_3|ry$, so Criterion B does not prevent $A(x_2; yx_3)$.

We now show that the value of α computed from Lemma 5.6 cannot satisfy $0 < \alpha < 1$, contradicting Criterion F.

Since we are assuming M_2, the formulas in Lemma 5.6 must be utilized with x_1 replaced by y, y replaced by x_2, x_2 replaced by x_3, and x_3 replaced by u. These become:
(a) $w_{rv} = [d(r, y) + d(r, u) - d(y, u)]/2$
$= [d(r, y) + d(r, x_3) - d(y, x_3)]/2$
(b) $w_{vq_3} = [d(r, u) + d(y, x_3) - d(r, y) - d(u, x_3)]/2$
(c) $w_{q_3u} = [d(r, u) + d(u, x_3) - d(r, x_3)]/2$
(d) $w_{hx_2} = [d(x_2, x_3) + d(x_2, y) - d(y, x_3)]/2$
(e) $E_2 = d(y, x_2) - d(r, x_2) + w_{rv}$
(f) $E_4 = d(x_3, x_2) - d(r, x_2) + w_{rv}$
(g) $\alpha = [2d(u, x_2) - 2w_{q_3u} - 2w_{hx_2} - d(r, y) + E_2 + 2w_{rv} + E_4 - d(r, x_3) + 2w_{vq_3}]/[4w_{vq_3}]$
(h) $w_{vq_1} = [d(r, y) - E_2 - w_{rv}]/[2\alpha]$
(i) $w_{q_3q_2} = [d(u, x_2) - w_{q_3u} - w_{hx_2} - \alpha(w_{vq_3} + w_{vq_1})]/(1 - \alpha)$
(j) $w_{q_1y} = d(r, y) - w_{rv} - w_{vq_1}$
(k) $w_{q_2x_3} = d(r, x_3) - w_{rv} - w_{vq_3} - w_{q_3q_2}$
(l) $C = 2d(u, x_2) - 2w_{q_3u} - 2w_{hx_2} - d(r, y) + E_2 + 2w_{rv} + E_4 - d(r, x_3) + 2w_{vq_3}$
(m) $D = 4w_{vq_3}$.

Then
(i) $\alpha(q_1, h) = \alpha = C/D$.
(ii) $d(q_1, y) = w_{q_1y}$.
(iii) $d(q_2, x_3) = w_{q_2x_3}$.

We must substitute the true quantities from the true network given in Figure 5.
$d(r, y) = d(r, v) + d(h, y) + \alpha d(v, q_1) + d(v, q_4) + d(q_4, q_3) + d(q_3, q_2) - \alpha d(v, q_4) - \alpha d(q_4, q_3) - \alpha d(q_3, q_2)$
$d(u, x_3) = d(q_4, u) + d(q_4, q_3) + d(q_3, x_3)$
$d(r, u) = d(r, v) + d(v, q_4) + d(q_4, u)$
$d(y, x_3) = d(h, y) + d(q_3, x_3) + \alpha d(v, q_1) + \alpha d(v, q_4) + \alpha d(q_4, q_3) + d(q_3, q_2) - \alpha d(q_3, q_2)$
$d(r, x_3) = d(r, v) + d(v, q_4) + d(q_4, q_3) + d(q_3, x_3)$
$d(y, x_2) = d(h, y) + d(q_2, x_2) + \alpha d(v, q_1) + \alpha d(v, q_4) + \alpha d(q_4, q_3) + \alpha d(q_3, q_2)$
$d(r, x_2) = d(r, v) + d(v, q_4) + d(q_4, q_3) + d(q_3, q_2) + d(q_2, x_2)$
$d(u, x_2) = d(q_4, u) + d(q_4, q_3) + d(q_3, q_2) + d(q_2, x_2)$
$d(x_3, x_2) = d(q_3, x_3) + d(q_3, q_2) + d(q_2, x_2)$
$d(y, u) = d(h, y) + d(q_4, u) + \alpha d(v, q_1) + \alpha d(v, q_4) + d(q_4, q_3) - \alpha d(q_4, q_3) + d(q_3, q_2) - \alpha d(q_3, q_2)$.

When we make these substitutions and simplify, we find after considerable algebra:
$C = -4d(q_4, q_3) + 4\alpha d(q_4, q_3)$
$D = 4\alpha d(v, q_4) + 4\alpha d(q_4, q_3) - 4d(q_4, q_3)$
Hence $\alpha = C/D = [-4d(q_4, q_3)+4\alpha d(q_4, q_3)]/[4\alpha d(v, q_4)+4\alpha d(q_4, q_3)-4d(q_4, q_3)]$

We require $0 < \alpha < 1$ by Criterion F.

Case 1. Suppose $D > 0$. Then $0 < C/D < 1$ requires $0 < C < D$ so $0 < -4d(q_4, q_3)+4\alpha d(q_4, q_3) < 4\alpha d(v, q_4)+4\alpha d(q_4, q_3)-4d(q_4, q_3)$. In particular $0 < 4(\alpha - 1)d(q_4, q_3)$ which is impossible since $d(q_4, q_3) > 0$ and $1 - \alpha > 0$.

Case 2. Suppose $D < 0$. Then $0 < C/D < 1$ requires $0 > C > D$, $0 > -4d(q_4, q_3) + 4\alpha d(q_4, q_3) > 4\alpha d(v, q_4) + 4\alpha d(q_4, q_3) - 4d(q_4, q_3)$.

In particular, $-4d(q_4, q_3)+4\alpha d(q_4, q_3) > 4\alpha d(v, q_4)+4\alpha d(q_4, q_3)-4d(q_4, q_3)$ so $0 > 4\alpha d(v, q_4)$ which is impossible since $\alpha > 0$ and $d(v, q_4) > 0$.

Case 3. Suppose $D = 0$ so $w_{vq_3} = 0$ But then from [27] it follows that $0 = w_{vq_3} = d(v, q_3)$ whence $v = q_3$. This is a contradiction since in Figure 5 we saw that u must attach strictly between v and q_3.

This completes the proof of Claim 10, which completes the proof of Theorem 6.1. □

In summary suppose the truth was that there was a hybrid h with child y and parents q_1 and q_2 with respective children x_1 and x_2, where y, x_1, and x_2 are in X. Then Criterion B eliminates all false possibilities except $A(x_2; y, x_3)$ and the symmetric possibility $A(x_1; y, x_0)$ (where x_3 attaches to the parent of q_2 or x_0 attaches to the parent of q_1). The elimination of these two false possibilities makes use of the other criteria.

It is clear that the reconstruction is polynomial. Indeed

Theorem 6.8. *Suppose $N = (V, A, r, X)$ is normal and satisfies the assumptions 3.1. Let $|X| = n$. Given the tree-average distances $d(x, y; N)$ for all x, y, in X, then the procedure reconstructs N in time $O(n^7)$.*

Proof. By [26] the number of vertices is $O(n^2)$. For each vertex the analysis of the possibilities of a hybrid includes for a given y, x_1, x_2, x_3, a check for all w distinct from these, hence time $O(n^5)$. The analysis of the possibilities of a cherry involves for a given $\{x, y\}$ a check for all $\{w, z\}$ distinct from these, hence time $O(n^4) \leq O(n^5)$. Hence we need $O(n^2)$ steps, each using at most time $O(n^5)$, for a total time of $O(n^7)$. □

7 A more complicated example

The methods of the paper also work for some normal networks that do not contain a single reticulation cycle. It is easy to see, for example, that they work for galled trees [10] that satisfy assumptions 3.1. They also work for the normal network N shown in Figure 7.

Suppose we were given the tree-average distances between members of $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, where 1 is the root. We would check all pairs $\{x, y\}$

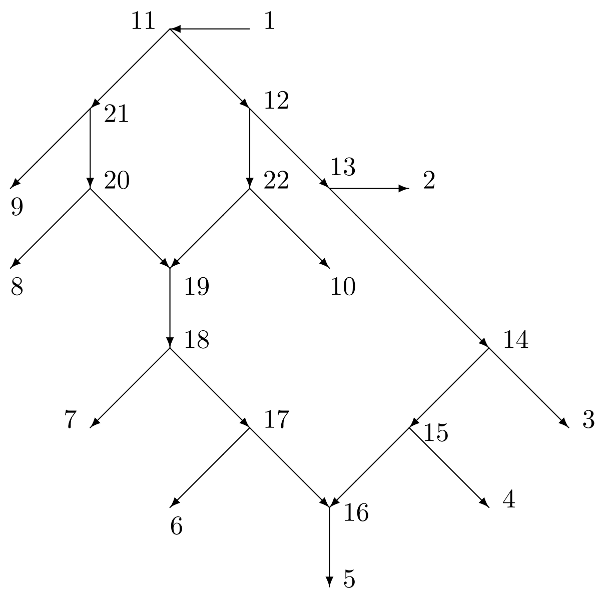

Figure 7: A network N with two reticulation cycles.

for possible cherries by seeing whether the conclusion of Theorem 4.5 holds. We would conclude that the network has no cherry. Checking for hybrids we would locate the hybrid with
(a) $y = 5$, $x_1 = 6$, $x_2 = 4$, $x_3 = 3$, satisfying all the criteria.
This same hybrid would also be recognized as
(b) $y = 5$, $x_1 = 6$, $x_2 = 4$, $x_3 = 2$
(c) $y = 5$, $x_1 = 4$, $x_2 = 6$, $x_3 = 7$.
All of these descriptions will lead to the correct identification, with the correct weights $\omega(17, 6)$, $\omega(16, 5)$, $\omega(15, 4)$ and probability $\alpha(17, 16)$. If $\alpha(17, 16) = 1/2$ then an equiprobable hybrid
(d) $y = 5$, $x_1 = 6$, $x_2 = 4$
would also be accepted and would result in the same weights.

The only issue so far is whether an incorrect hybrid might have been identified instead of (a), (b), (c), or (d). If the correct hybrid is identified, then it is removed. The resulting network has only a single reticulation cycle, so Theorem 6.1 applies. Thus the network N is correctly reconstructed.

In fact arguments like those for the proof of theorem 6.1 show that no incorrect hybrid would be identified in that first step. The arguments are made more complicated by the fact that N has four different trees of form N_p rather than just two, as in Theorem 6.1. The hypotheses of Lemma 6.5 require a bit more checking, since we must avoid the possibility of having all three quartets arise for a certain choice of w. (This is immediate in Theorem 6.1 since there are only two trees N_p.) In addition, there are many more cases. We omit further analysis.

8 Conclusions and extensions

In this section we remark on some related issues.

(a) Dealing with errors in the distances

The methods in this paper assume that the correct tree-average distances are known exactly. A major difficulty is that some of the criteria for a cherry or a hybrid as stated require some exact equalities. For example, the criterion that $\{x, y\}$ is a cherry requires that for all w and z such that x,y,w,z are distinct, we must have $d(x, y) + d(w, z) < d(x, w) + d(y, z) = d(x, z) + d(w, y)$. Such a condition is unlikely to hold if the true distances are subject to small errors, since the equality will almost certainly fail. Similarly, the formulas used in Theorem 6.1 for recognition of a hybrid involve an equality.

More generally, if we are to be able to use the results on real data, it would be useful to have a more robust calculation that will work when the data have sufficiently small errors. While the author has a computer program that works well when the exact tree-average distances are input or are input with only very small perturbations, the program does not appear to be reliable yet with real data.

(b) Taxon sampling

The results raise the issue of taxon sampling. Suppose that the true network is as in Figure 2, with the probability $\alpha(q_1, h) \neq 1/2$, say $\alpha(q_1, h) = 1/3$. Suppose that the taxon x_3 were not present, so $X = \{r, x_1, y, x_2\}$ and the true tree-average distances were still known. Our method could correctly find that there is no cherry and use Criterion B to find that y is child to the hybrid and the parents have children x_1 and x_2. But we would be forced to accept the hybrid as equiprobable and we would not reconstruct the correct $\alpha(q_1, h)$. The tree-average distances in the reconstructed network would not match the input distances.

With real data, of course, we would not expect an exact match in any event. A more serious problem, however, is that unless x_3 is present, there is no possibility of finding the correct $\alpha(q_1, h)$. Thus the method needs to be applied only when x_3 is present. On the other hand, in advance it is not possible to guarantee that all hybrids have a taxon in a position analogous to x_3. The collection of taxa must satisfy properties that are not knowable in advance.

(c) Hybrids of indegree 3 or higher

It is quite possible that a network could have hybrids with indegree 3 or higher. The results of [27] do not apply to give explicit formulas for the weights even in the equiprobable case when each parent of the lower hybrid has probability 1/3 and each parent of the upper hybrid has probability 1/2. With inadequate taxon sampling, such a normal network might well be the best description of a system in which several species in sequence experience gene transfer and/or hybridization to produce ultimately species 3.

References

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

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

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

[4] A. Boc and V. Makarenkov (2003) New efficient algorithm for detection of horizontal gene transfer events. In Benson, G. and R.D., M. (Eds.). Proceedings of the WABI03, pp. 190-201, Lecture Notes in Computer Science # 2812.

[5] G. Cardona, F. Rosselló, and G. Valiente, (2009), Comparison of tree-child phylogenetic networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6(4): 552-569.

[6] C. Choy, J. Jansson, K. Sadakane, and W.-K. Sung, (2005) Computing the maximum agreement of phylogenetic networks, Theoretical Computer Science 335(1):93-107.

[7] R. Desper and O. Gascuel (2002) Fast and accurate phylogeny reconstrution algorihms based on the minimum-evolution principle, Journal of Computational Biology, 9(5): 687-705.

[8] R. Desper and O. Gascuel (2004) Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting, Molecultar Biology and Evolution, 21(3): 587-598.

[9] W. F. Doolittle et al. (2003) How big is the iceberg of which organella genes in nuclear genomes are but the tip? Philos. Trans. R. Soc. Lond. B. Biol. Sci., 358, 39-47

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

[11] M. Hasegawa, 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.

[12] D. Huson, (1998), SplitsTree: analyzing and visualizing evolutionary data, Bioinformatics, 14(10):68-73.

[13] D. Huson, R. Rupp, and C. Scornavacca, (2010), Phylogenetic Networks: Concepts, Algoriithms and Applications, Cambridge, Cambridge University Press.

[14] L. J. J. van Iersel, J. C. M. Keijsper, S. M. Kelk, L. Stougie, F. Hagen, and T. Boekhout, (2009), Constructing level-2 phylogenetic networks from triplets, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6(43): 667-681.

[15] T.H. Jukes 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.

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

[17] J.A. Lake (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances, Proc. Natl. Acad. Sci. USA 91, 1455-1459.

[18] T. Marcussen et al, (2012). Inferring species networks from gene trees in high-polyploid North American and Hawaiian Violets (Viola, Violacae), Systematic Biology 61, 107-126.

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

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

[21] N. Saitou and M. Nei, (1987), The neighbor-joining method: A new method for reconstructing phylogenetic trees, Molecular Biology and Evolution 4: 406-425.

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

[23] M.A. Steel (1994) Recovering a tree from the leaf colorations it generates under a Markov model, Appl. Math. Lett. 7(2) 19-23.

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

[25] L. Wang, B. Ma, M. Li, (2000), Fixed topology alignment with recombination, Discrete Applied Mathematics 104(1-3): 281-300.

[26] S.J. Willson, (2010), Properties of normal phylogenetic networks, Bulletin of Mathematical Biology 72, 340-358.

[27] S.J. Willson, (2012), Tree-average distances on certain phylogenetic networks have their weights uniquely determined, Algorithms for Molecular Biology (2012) 7:13, doi:10.1186/1748-7188-7-13.