Tree-average distances on certain phylogenetic networks have their weights uniquely determined

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

May 11, 2012

Abstract: A phylogenetic network N has vertices corresponding to species and arcs corresponding to direct genetic inheritance from the species at the tail to the species at the head. Measurements of DNA are often made on species in the leaf set, and one seeks to infer properties of the network, possibly including the graph itself. In the case of phylogenetic trees, distances between extant species are frequently used to infer the phylogenetic trees by methods such as neighbor-joining.

This paper proposes a *tree-average* distance for networks more general than trees. The notion requires a *weight* on each arc measuring the genetic change along the arc. For each displayed tree the distance between two leaves is the sum of the weights along the path joining them. At a hybrid vertex, each character is inherited from one of its parents. We will assume that for each hybrid there is a probability that the inheritance of a character is from a specified parent. Assume that the inheritance events at different hybrids are independent. Then for each displayed tree there will be a probability that the inheritance of a given character follows the tree; this probability may be interpreted as the probability of the tree. The *tree-average* distance between the leaves is defined to be the expected value of their distance in the displayed trees.

For a class of rooted networks that includes rooted trees, it is shown that the weights and the probabilities at each hybrid vertex can be calculated given the network and the tree-average distances between the leaves. Hence these weights and probabilities are uniquely determined. The hypotheses on the networks include that hybrid vertices have indegree exactly 2 and that vertices that are not leaves have a tree-child.

Key words: digraph, distance, metric, hybrid, network, tree-child, normal network, phylogeny,

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 leaves typically 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. General frameworks for phylogenetic networks are discussed in [1], [2], [3], and [4]. See also the recent book [5].

There are many methods to reconstruct phylogenetic trees from information such as the DNA of extant species. The most generally accepted methods include maximum parsimony, maximum likelihood, and Bayesian. See [6] for an overview. These methods, however, are only heuristic, do not guarantee an optimal solution, and can be very time-consuming for a moderate number of species.

Suppose X denotes the set of extant species for some analysis, 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 [7], Kimura [8], HKY [9], and log determinant [10], [11] distances. The log determinant distance is especially interesting in that it can be proved that typically the distances add along the paths, so that the distance along a path is the sum of the distances for 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 [12]. It is computationally fast. 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 [13], [14] 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 have been rarely used to construct phylogenetic networks that are not necessarily trees. It is true that distances occur in common exploratory methods to display the diversity of trees for the same species such as the split decomposition (see [15] or an overview in [5]). These distances, however, are not derived from any biologically based model of evolution.

This paper studies a distance on rooted directed networks that is based upon a model of evolution. Consider, for example, the network N in Figure 1. The root is 1 and there is a hybridization event at 7 with parents 6 and 8. Vertex 7 is called a *hybrid* vertex or a *reticulation* vertex. For some characters, the character state at 7 is inherited from the parental species 6, while for other

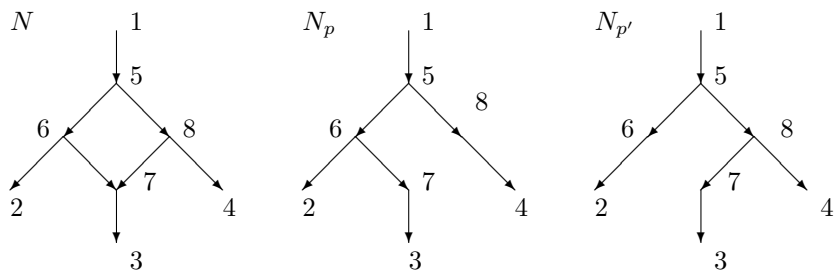

Figure 1: A network N with root 1, and the two trees N_p and $N_{p'}$ that it displays. If N is equiprobable, then 7 inherits approximately half its characters from 6 and the other characters from 8.

characters the character state at 7 is inherited from species 8. For character states inherited from 6 the evolutionary history is best described by the displayed tree N_p, while for character states inherited from 8 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 $p(7) = 6$ while $p'(7) = 8$. 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. This paper explores the possibility that an appropriate distance between the vertices in the network N is a weighted average of the distances in N_p and $N_{p'}$.

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 will be defined to be the expected value of the distances in the various trees N_p:

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

If a hybrid vertex h satisfies that each parent q of h has the same probability, we will call the inheritance *equiprobable at h*. This special case assumes that

the contribution from each parent to h is the same; if there are two parents, each contributes approximately 50%.

In Figure 1 note that, for each species in the leafset $X = \{1, 2, 3, 4\}$, it is plausible that the DNA is available since 2, 3, 4 correspond to extant species and 1 to an extant outgroup species. Hence it is plausible that we know $d(x, y; N)$ for distinct x and y in X, hence $\binom{4}{2} = 6$ nonzero distances. Nevertheless, N has 8 arcs and hence it is not likely that from the 6 known distances we could compute 8 independent weights for these arcs. Indeed, the equations obtained in this paper for this network have infinitely many solutions. There is a possibility of simultaneous identical mutations between 6 and 7 and between 8 and 7 which might be confused with mutations between 7 and 3.

In this paper we will assume that the weight of an arc into a hybrid vertex is 0. Thus in Figure 1, the weights of arcs $(6, 7)$ and $(8, 7)$ will be zero. Under this assumption vertex 7 corresponds roughly to the immediate offspring of a hybridization event, in which some characters came intact from 6 and the remainder intact from 8. Further mutation occurred before species 3 evolved from 7.

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 $x, y \in \{1, 2, 3, 4\}$, we might be able to recover the weights for each of the 6 arcs in N that are not directed into the hybrid vertex 7. These same weights would be utilized in distances for both N_p and $N_{p'}$. On the other hand, we should like to determine an additional parameter $\alpha(6, 7)$ telling the probability of inheritance by 7 of a character from 6. It is unlikely that six equations, one for each $d(x, y; N)$, will uniquely and generically determine seven real parameters. Indeed, the methods of this paper for this example lead to six equations in seven unknowns such that for certain values of the distances the weights and probabilities are not uniquely determined. Consequently for the situation in Figure 1 we will assume that $\alpha(6, 7) = \alpha(8, 4) = 1/2$; we call the inheritance *equiprobable at 7*.

By contrast, Figure 2 shows another network with $X = \{r, x_1, x_2, x_3, y\}$ containing a single hybrid vertex h_0. In this case there are $\binom{5}{2} = 10$ distances 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 $\alpha_1 = \alpha(q_1, h_0)$ satisfying $0 < \alpha_1 < 1$. In fact, this paper will show how to determine all 9 parameters. Then $\alpha(q_2, h_0) = 1 - \alpha_1$ is also determined. In Figure 2 we will not need to assume equiprobability at h_0.

In order to obtain interesting results, assumptions must be made about the network N. As an extreme case it would be easy to add many more internal vertices and edges to the network N of Figure 1 without adding any additional leaves yet increasing arbitrarily the number of arcs. For example, Figure 3 shows a network in which the network N of Figure 1 has been modified by the addition of other arcs. The 6 distances do not determine the weights for all 7 arcs that do not lead to a hybrid vertex in Figure 3.

Particular kinds of acyclic networks have been studied in various papers. Wang *et al.* [16] and Gusfield *et al.* [17] study "galled trees" in which all

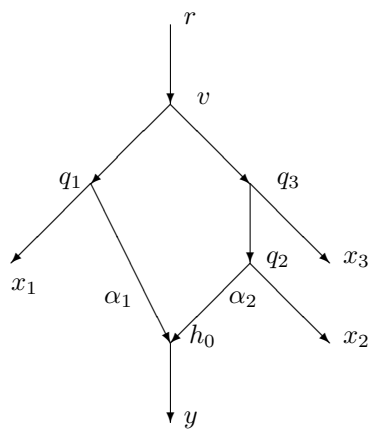

Figure 2: A minimal configuration needed to be able to find the probability $\alpha_1 = \alpha(q_1, h_0) = 1 - \alpha_2$ that a character state in h_0 is inherited from q_1.

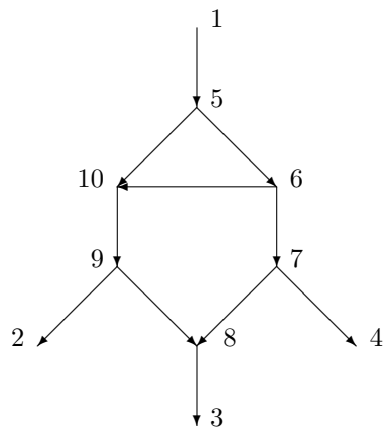

Figure 3: This tree-child network has $X = \{1, 2, 3, 4\}$. There are 7 arcs not leading to a hybrid vertex but only 6 distances, and the weights are not uniquely determined. The network is not normal because arc $(5, 10)$ is redundant.

recombination events are associated with node-disjoint recombination cycles; the idea occurs also earlier in [18]. Choy *et al.* [19] and Van Iersel *et al.* [20] 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.* [21] discussed "tree-child" networks, in which every vertex 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 [22] which are both tree-child and contain no redundant arc.

Most results in this paper assume that the network is *normal*. This means, briefly, that every vertex not in X and not a leaf has a tree-child (a child with indegree one); and moreover, there is no redundant arc. For example, if $X = \{1, 2, 3, 4\}$ then the network in Figure 1 is normal while the network in Figure 3 is not normal since arc (5,10) is redundant. With the assumption that there are no redundant arcs we show in Section 3 that for a given network N, the tree-average distance d is a metric on X. With the assumption of normality we also show that different parent maps p yield different displayed trees N_p. Hence the average over the parent maps p is the same as the average over displayed trees. This result eliminates the logical possibility that different parent maps p_1 and p_2 might yield displayed trees that are topologically the same, yielding an uncertainty about which is the correct average to use in the definition.

The main result, Theorem 4.1, assumes that the network N is normal and also that for all hybrid vertices the indegree is exactly 2 and the outdegree is exactly 1. At each hybrid vertex h we assume either equiprobability or else that h has a grandparent on at least one side of the reticulation cycle, as in Figure 2 but not Figure 1. Then 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. This calculation is, of course, trivial if the network is equiprobable at h.

A model for a distance function containing certain parameters is called *identifiable* if the parameters can be reconstructed from the (exact) values of the distance function. Theorem 4.1 thus asserts 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.

A major problem, of course, is the reconstruction of N itself from a distance function d. I have obtained partial results (not included in this paper) which give a reconstruction of N itself when the distance d is the tree-average distance and when the network N satisfies the hypotheses of Theorem 4.1 and some additional hypotheses. The reconstruction of N is possible because of the simple forms of the formulas obtained in this paper. Essentially, the formulas are simple enough that they can be used recursively when only part of the network is yet known. I plan a subsequent paper which will utilize the results in the current paper to reconstruct N from the tree-average distances.

The assumption that all hybrid vertices have indegree 2, assumed in Theorem 4.1, is plausible biologically since in sexually reproducing species an offspring arises from one egg and one sperm.

The assumption that there be no redundant arcs is essential for Theorem 4.1. Figure 3 displays a tree-child network N with $X = \{1, 2, 3, 4\}$. There are 6 independent nonzero distances between the members of X, yet there are 7 arcs not directed into hybrid vertices. It is easy to choose positive values for the

tree-average distances such that there are infinitely many positive choices of the weights given the network. Note that each vertex not a leaf has a tree-child, so the network is a tree-child network [21]. Hence Theorem 4.1 cannot be extended to general tree-child networks.

Some other extensions of the current results and problems are discussed in the concluding section 6.

2 Fundamental Concepts

A *directed graph* or *digraph* (V, A) consists of a finite set V of *vertices* and a finite set A of *arcs*, each consisting of an ordered pair (u, v) where $u \in V$, $v \in V$, $u \neq v$. 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. 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.

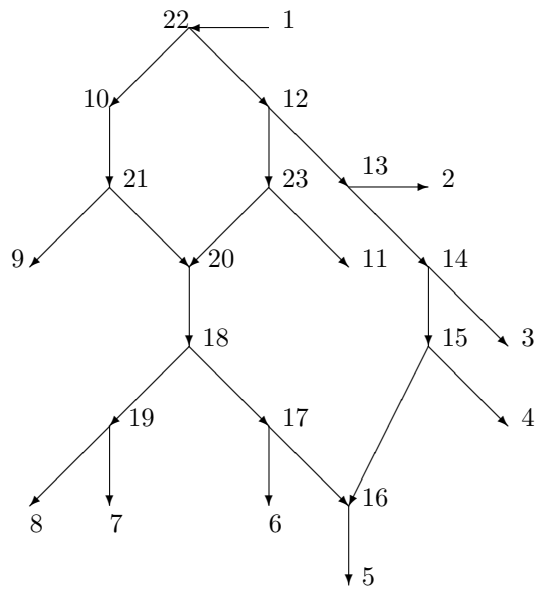

Figure 4: A normal phylogenetic X-network with $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ and root 1. The root corresponds to an outgroup species. Measurements on DNA are assumed possible on members of X.

Figure 4 shows a phylogenetic X-network N with base-set $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$. The root is $r = 1$. Note that the leaves are in X by (i), $1 \in X$ by (ii), and $10 \in X$ by (iii). Measurements such as DNA are assumed possible on members of X. Since the root 1 is actually an outgroup and the leaves are all extant, this is plausible for all members of X except 10. We are perhaps here assuming that, by some fortuitous chance, some historical DNA of 10 is also available.

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 networks in Figure 2 and 4 are normal, while the network of Figure 3 is not normal. The usage here of "normal" differs slightly from that in [22] in that here hybrid vertices that are not leaves may have outdegree 1, whereas in [22] hybrid vertices that were not leaves had outdegree 2 or higher. There is an obvious one-to-one relationship between normal networks

in the current sense and normal networks in the previous sense.

A normal network N is *semibinary* if each hybrid node has indegree 2 and outdegree 1. It follows from normality that the child of the hybrid node is necessarily 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 4, the path 20, 18, 19, 8 is normal and is a normal path from 20 to X. The path 18, 17, 16, 5 is not normal since 16 is hybrid. The trivial path 3 is normal.

Suppose N is normal and $v \in V$. Then there is a normal path from v to X. To see this, if $v \in X$, then the trivial path is a normal path from v to X. If $v \notin X$, then v has a tree child v_1. If $v_1 \in X$, then the path v_0, v_1 is a normal path to v_1 in X. Otherwise v_1 has a tree-child v_2. If $v_2 \in X$ then the path v_0, v_1, v_2 is a normal path from v to v_2 in X. Proceeding in this manner, we obtain the result.

Suppose two normal paths shared a common vertex x, say the normal paths $v = v_0, \cdots, v_k = x$ and $w = w_0, \cdots, w_j = x$. If $k > 0$ and $j > 0$ then since x is normal with a unique parent, it follows that $v_{k-1} = w_{j-1}$. Repeating the argument we find that either there is an i such that $v = w_i$ or else there is an i such that $w = v_i$. This argument, of frequent use, is called *following the normal paths backwards.*

A *graph* (or, for emphasis, an *undirected graph*) (V, E) consists of a finite set V of *vertices* and a finite set E of *edges*, each a subset $\{v_1, v_2\}$ of V consisting of two distinct vertices. Thus an edge has no direction, while an arc has a direction. If $N = (V, A, r, X)$ is a phylogenetic X-network, there is an associated undirected graph $Und(N) = (V, E)$ in which every arc in A has its direction ignored; thus $E = \{\{a, b\} : (a, b) \in A \text{ or } (b, a) \in A\}$.

3 The Tree-Average Distance

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 4, an example of a parent map p satisfies $p(20) = 23$, $p(16) = 17$, and for all other vertices v besides 1, $p(v)$ is the unique parent of v.

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. 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, if h is

hybrid in N, then in N_p the vertex h has indegree 1 from the arc $(p(h), h)$ and outdegree 1, yet need not lie in 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.

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 $\omega(h, c)$.

In any rooted tree $T = (V, A, r)$, two vertices u and v have a unique *most recent common ancestor* $\text{mrca}(u, v) = \text{mrca}(u, v; T) \in V$ that satisfies
(1) $\text{mrca}(u, v) \leq u$ and $\text{mrca}(u, v) \leq v$;
(2) whenever $z \leq u$ and $z \leq v$, then $z \leq \text{mrca}(u, v)$.
In a network that is not a tree, two vertices u and v need not have a $\text{mrca}(u, v)$.

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 $m = \text{mrca}(u, v; N_p)$, 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 fraction of the genome that h inherits from u. We may interpret $\alpha(u, h)$ as the probability that a character is inherited by h from u, so 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 $Pr(p) = \prod[\alpha(p(h), h) : h \in H]$.

The *tree-average distance* $d(u, v; N)$ between u and v in N is defined 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*. If the network N is equiprobable at h for all $h \in H$, then we call the network *equiprobable*, and for each u and v in X, $d(u, v; N)$ is the average of the values $d(u, v; N_p)$ for $p \in Par(N)$.

For example, for the network N in Figure 1 suppose that the arcs have weights given by $\omega(1, 5) = 1 = \omega(5, 6) = \omega(7, 3)$, while $\omega(5, 8) = \omega(8, 4) = 2$ and $\omega(6, 2) = 4$. Since 7 is hybrid, $\omega(6, 7) = \omega(8, 7) = 0$. Suppose, as in Figure 1, the parent map p satisfies $p(7) = 6$ while the parent map p' satisfies $p'(7) = 8$. Then N_p shown in Figure 1 is obtained from N by deleting the arc $(8, 7)$ while $N_{p'}$ is obtained from N by deleting the arc $(6, 7)$. Assume $\alpha(6, 7) = 1/3$ and $\alpha(8, 7) = 2/3$, so $Pr(p) = 1/3$, $Pr(p') = 2/3$. To compute $d(1, 3; N)$ we find $d(1, 3; N_p) = \omega(1, 5) + \omega(5, 6) + \omega(6, 7) + \omega(7, 3) = 1 + 1 + 0 + 1 = 3$,
$d(1, 3; N_{p'}) = \omega(1, 5) + \omega(5, 8) + \omega(8, 7) + \omega(7, 3) = 1 + 2 + 0 + 1 = 4$. Hence $d(1, 3; N) = (1/3)d(1, 3; N_p) + (2/3)d(1, 3; N_{p'}) = (1/3)(3) + (2/3)(4) = 11/3$. For another example $d(1, 2; N_p) = d(1, 2; N_{p'}) = 6$ so $d(1, 2; N) = (1/3)(6) + (2/3)(6) = 6$.

Given u and v, the vertices $\mathrm{mrca}(u, v; N_p)$ may differ for different p. This is seen in Figure 1 where $\mathrm{mrca}(2, 3; N_p) = 6$ while $\mathrm{mrca}(2, 3; N_{p'}) = 5$.

Theorem 3.1. *Assume $N = (V, A, r, X)$ is a phylogenetic X-network that has no redundant arcs. Assume N has a weight function ω satisfying that $\omega(a, b) > 0$ if b is normal. Then the tree-average distance on X from N is a metric on X.*

Proof. A metric d on X must satisfy
(1) For all x and y in X, $d(x, y) \geq 0$ and $d(x, y) = 0$ iff $x = y$.
(2) For all x and y in X, $d(x, y) = d(y, x)$.
(3) For all $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.
For (2), suppose $x, y \in X$. For all p, $d(x, y; N_p) = d(y, x; N_p)$, whence $d(x, y; N) = d(y, x; N)$.

For (3) suppose $x, y, z \in X$. For each N_p, $d(x, z; N_p) \leq d(x, y; N_p) + d(y, z; N_p)$ from the truth of the four-point condition, see [23], p 147. Hence the result follows for distances in N as well.

For (1) it is clear that for each p, $d(x, y; N_p) \geq 0$, whence $d(x, y; N) \geq 0$. Moreover, for each p, $d(x, x; N_p) = 0$, whence $d(x, x; N) = 0$.

To finish the proof of (1), suppose $d(x, y; N) = 0$; we show $x = y$. Assume instead $x \neq y$. Since the weights are nonnegative, for every p we have

$d(x, y; N_p) = 0$. Hence for every $p \in Par(N)$, in N_p the unique path between x and y contains only arcs (a, b) with b hybrid in N.

If x and y are both normal, then for every p the unique path between x and y in N_p must consist of a directed path from $v = \text{mrca}(x, y; N_p)$ to x and a path from v to y; hence it contains a normal arc whence $d(x, y; N_p) > 0$. Thus we may assume that one vertex, say y, is hybrid.

In N choose a directed path $P = y_0, y_1, y_2, \cdots, y_k = y$ such that y_1 is not hybrid but $y_2, \cdots, y_k$ are hybrid. This is always possible because there is a directed path from r to y, say $u_0 = r, u_1, u_2, \cdots, u_k = y$. The child u_1 of r cannot be hybrid, because if it were, then its other parent q besides r must also have a path to q from r, and this path combined with the arc (q, u_1) would make the arc (r, u_1) redundant. Moreover, we may choose this path so that x does not lie in $\{y_1, \cdots, y_k\}$ since whenever y_i is hybrid there are at least two choices of the parent y_{i-1}, and we may select y_{i-1} to be distinct from x.

If x is normal in N, let Q be the trivial path $z_0 = x$. Otherwise we may choose a directed path $Q = z_0, z_1, \cdots, z_s = x$ such that z_0 is not hybrid but all other vertices are hybrid. Moreover, we may assume that the vertices of Q are all distinct from the vertices of P. This is because, if z_i is hybrid, it cannot have two parents q_1 and q_2 which are on P since then there must be a directed path from say q_1 to q_2, whence the arc (q_1, z_i) is redundant.

Since the vertices on P and Q are distinct, there exists a parent map p that agrees with all the choices made in constructing both P and Q. Hence in N_p, P is a path from y_0 to y, Q is a path form z_0 to x, and the paths are disjoint. In N_p let $v = \text{mrca}(y_0, z_0; N_p)$. Then in N_p the unique path between x and y consists of P, Q, a path from v to y_0, and a path from v to z_0. Since y_1 and z_0 are normal, this path includes a normal arc, so $d(x, y; N_p) > 0$. It follows that $d(x, y; N) > 0$, a contradiction.

□

Corollary 3.2. *Assume N is a normal network with weight function ω such that $\omega(a, b) > 0$ if b is normal. Then the tree-average distance on X from N is a metric on X.*

The tree-average distance is defined as a weighted average in terms of parent maps. Any tree that arises as N_p for some parent map p is said to be *displayed* in N. There is a logical possibility that several different parent maps p could yield essentially the same displayed tree. The next theorem gives sufficient conditions so that in fact the displayed trees are all distinct. Hence the tree-average distance becomes a weighted average over all the distinct displayed trees.

The proof requires the notion of a *split*. A *split of* X is a partition of X into exactly two nonempty subsets; if these are A and B, we write the split $A|B$. Two splits $A_1|B_1$ and $A_2|B_2$ are *compatible* if at least one of the sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, and $B_1 \cap B_2$ is the empty set. Removal of any edge e (but not its endpoints) from a tree T produces a split $\Sigma(e)$ consisting of vertices in the connected components of T with e removed. The set of splits of a tree T will be

denoted $\Sigma(T)$. If T is directed, then the splits of T are obtained by reference only to the undirected tree so $\Sigma(T) = \Sigma(Und(T))$. By the Splits-Equivalence Theorem (see [23], p. 44) any two splits of a tree are compatible.

Theorem 3.3. *Assume $N = (V, A, r, X)$ is a normal phylogenetic X-network. Suppose that every hybrid vertex that is not a leaf satisfies that it has outdegree 1 and that its unique child is normal. Suppose p and q are distinct parent maps for N. Then N_p and N_q are topologically distinct trees.*

Proof. We show that $\Sigma(N_p)$ and $\Sigma(N_q)$ are distinct. Since $p \neq q$ there exists a hybrid vertex h such that $p(h) \neq q(h)$. Let $q_1 = p(h)$ and $q_2 = q(h)$. Choose a normal path in N from q_1 to $x_1 \in X$, a normal path from q_2 to $x_2 \in X$, and a normal path from h to $y \in X$. Note that each normal path is a path in both N_p and N_q. Moreover, q_1 is normal in N because otherwise its unique child would not be a tree-child. Similarly q_2 is normal in N.

If $\Sigma(N_p) = \Sigma(N_q)$, then each pair of splits would be compatible. In N_p consider the split $\Sigma(a, q_1)$ where a is the unique parent of q_1 and we remove the arc (a, q_1) from N_p. We may write $\Sigma(a, q_1)$ as $A_1|B_1$ where A_1 contains r. The directed path in N_p from r to y includes the arc (a, q_1), so B_1 contains y. Then $x_1 \in B_1$ because there is a path from q_1 to x_1 and from h to y, neither of which includes (a, q_1). Moreover, $x_2 \in A_1$. To see this, since N_p is rooted, there is a directed path from r to q_2. If it included the arc (a, q_1), then there would be a directed path in N_p from q_1 to q_2; this is not possible since in that case the arc (q_1, h) would be redundant in N, contradicting normality of N. Since N_p contains the directed path from q_2 to x_2 missing the arc (a, q_1), it follows that $x_2 \in A_1$. Hence $\{r, x_2\} \subseteq A_1$ and $\{y, x_1\} \subseteq B_1$.

In N_q consider the split $\Sigma(b, q_2)$ where b is the parent of q_2 and we remove the arc (b, q_2) from N_q. Similarly to the case of N_p we may write $\Sigma(b, q_2) = A_2|B_2$ where $\{r, x_1\} \subseteq A_2$ and $\{y, x_2\} \subseteq B_2$. If N_p were topologically the same as N_q, then these splits would need to be compatible. Yet $r \in A_1 \cap A_2$, $x_2 \in A_1 \cap B_2$, $x_1 \in B_1 \cap A_2$, and $y \in B_1 \cap B_2$, contradicting compatibility. □

Corollary 3.4. *Suppose $N = (V, A, r, X)$ is a phylogenetic X-network that is normal. Suppose every hybrid vertex that is not a leaf has outdegree 1 and its unique child is normal. Suppose that there are exactly k hybrid vertices h_1, h_2, $\cdots, h_k$ and that for $i = 1, \cdots, k$, hybrid vertex h_i has indegree d_i. Then the total number of distinct trees displayed by N and the total number of parent maps are both $\prod[d_i : i = 1, \cdots, k]$.*

4 Finding the weight function from d and N

In this section we prove the main theorem, that the weights are determined by knowledge of N and the tree-average distances between members of X. For each hybrid vertex h we will assume either equiprobability at h or else a more complicated situation resembling Figure 2. The assumptions can be different at different hybrid vertices.

Theorem 4.1. *Suppose $N = (V, A, r, X)$ is a phylogenetic X-network which is normal and semibinary. Let ω be a weight function on A satisfying $\omega(a, b) = 0$ if b is hybrid and $\omega(a, b) \geq 0$ if b is normal. Assume that N is known and that the tree-average distance $d(x, y; N)$ is known for each x and y in X.*

For each hybrid vertex h with parents q_1 and q_2, assume either
(1) the inheritance is equiprobable at h; or
(2) at least one parent (say q_2) satisfies that there exists q_3 such that
(a) there is a normal path from q_3 to q_2;
(b) there is a normal path from q_3 to some x_3 in X which is disjoint from the normal path from q_3 to q_2 except for the vertex q_3;
(c) there is no directed path from q_3 to q_1.
Then the weight function ω is uniquely determined and can be computed explicitly. Moreover, for each hybrid h, the probabilities $\alpha(q_i, h)$ for each parent q_i of h are uniquely determined and can be computed explicitly.

See Figure 2 to understand the assumptions about h in (2). Throughout this section we will assume the hypotheses of Theorem 4.1.

The proof primarily consists of a number of cases to handle different situations. We will present several of these special situations as lemmas and then later relate these together. Each lemma tells how certain distances or weights relate to distances between members of X.

Lemma 4.2. *Assume the hypotheses of Theorem 4.1. Suppose there is a normal path from a to b. Suppose there is a normal path from a to $x \in X$ which meets the normal path from a to b only in a. Suppose b has normal paths to y and z in X which are disjoint except at b. Then*
$d(a, b; N) = [d(r, y; N) + d(x, z; N) - d(r, x; N) - d(y, z; N)]/2.$

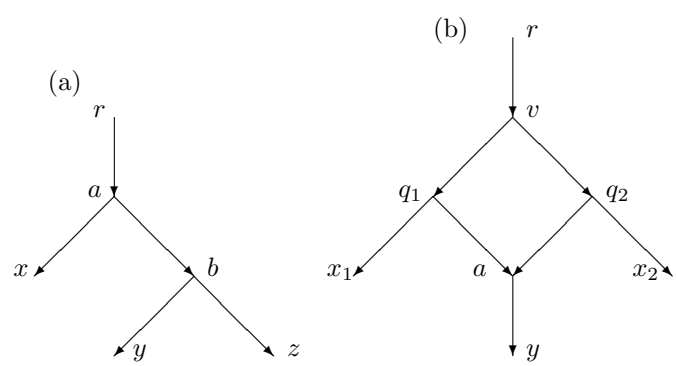

Figure 5: (a) The situation of Lemma 4.2. (b) The situation of Lemmas 4.4 and 4.7. If p is a parent map with $p(a) = q_1$, the figure shows part of N_p together with the arc (q_2, a).

Proof. For each $p \in Par(N)$, the path from a to b, the path from a to x, the path from b to y, and the path from b to z must lie in N_p since none of the arcs enters a hybrid vertex. Moreover, there must be a path from r to a which includes none of the arcs on the other paths mentioned above. See Figure 5a. Hence for each $p \in Par(N)$ one can verify

$d(r, y; N_p) = d(r, a; N_p) + d(a, b; N_p) + d(b, y; N_p)$
$d(x, z; N_p) = d(a, x; N_p) + d(a, b; N_p) + d(b, z; N_p)$
$d(r, x; N_p) = d(r, a; N_p) + d(a, x; N_p)$
$d(y, z; N_p) = d(b, y; N_p) + d(b, z; N_p)$.

It follows that

$$[d(r, y; N_p) + d(x, z; N_p) - d(r, x; N_p) - d(y, z; N_p)]/2 = d(a, b; N_p).$$

Taking expected values we see $d(a, b; N) = \sum[Pr(p)d(a, b; N_p) : p \in Par(N)]$
$= \sum[Pr(p)[d(r, y; N_p) + d(x, z; N_p) - d(r, x; N_p) - d(y, z; N_p)]/2 : p \in Par(N)]$
$= [d(r, y; N) + d(x, z; N) - d(r, x; N) - d(y, z; N)]/2.$ □

Lemma 4.3. *Assume the hypotheses of Theorem 4.1.*

(1) Suppose (a, b) is an arc where $a \in X$ and b is normal. Suppose b has normal paths to y and z in X which are disjoint except at b. Then $\omega(a, b) = [d(a, y; N) + d(a, z; N) - d(y, z; N)]/2$.

(2) Suppose there is a normal path from a to $b \in X$. Suppose there is a normal path from a to $x \in X$ which intersects the path from a to b only in a. Then $d(a, b; N) = [d(b, r; N) + d(b, x; N) - d(r, x; N)]/2$.
In particular, suppose (a, b) is an arc, $b \in X$ is normal, and there is normal path from a to $x \in X$ which does not include b. Then
$\omega(a, b) = [d(b, r; N) + d(b, x; N) - d(r, x; N)]/2$.

(3) Suppose (a, b) is an arc and b is normal. Suppose there is a normal path from a to $x \in X$ which does not include the vertex b. Suppose b has normal paths to y and z in X which are disjoint except at b. Then
$\omega(a, b) = [d(r, y; N) + d(x, z; N) - d(r, x; N) - d(y, z; N)]/2$.

Proof. For (1) we take $a = x$ in Lemma 4.2 and note that $d(r, y; N) - d(r, a; N) = d(a, y; N)$. For (2) we take $b = y = z$ in Lemma 4.2 and note that $d(y, z; N) = 0$. For (3), we use the normal path a, b as the path from a to b. □

Lemma 4.4. *Assume the hypotheses of Theorem 4.1. Suppose there is a normal path from a to $y \in X$ where a is hybrid with indegree 2 and parents q_1 and q_2. Assume q_1 and q_2 have normal paths to x_1 and x_2 respectively in X. Then $d(a, y; N) = [d(y, x_1; N) + d(y, x_2; N) - d(x_1, x_2; N)]/2$.*

Proof. See Figure 5b. We first show that the portion of the figure including the paths from q_1 to x_1, from q_2 to x_2, from a to y and the arcs (q_1, a) and (q_2, a) accurately represents the hypotheses of the lemma. (The network in Figure 3, which is not normal, has this situation with $a = 10$, $q_1 = 5$, $q_2 = 6$, $x_1 = x_2 = 4$, $y = 2$. Hence Figure 5b is wrong for the network in Figure 3, primarily because the normal paths from q_1 to x_1 and from q_2 to x_2 intersect.) I claim that for normal networks the normal paths from q_1 to x_1 and from q_2 to x_2 have no

vertex in common. To see this, suppose there were such a common vertex w. In that case by following the normal paths backwards from w we infer that either q_1 lies on the path from q_2 to x_2 or else q_2 lies on the path from q_1 to x_1. In the former case there is a directed path from q_2 to q_1, whence the arc (q_2, a) is redundant, contradicting the normality of the network. In the latter case (q_1, a) is redundant. It follows that the paths are disjoint. In particular, $x_1 \neq x_2$.

Similarly, neither path can intersect the normal path from a to y. If, for example, the path from q_1 to x_1 intersected the path from a to y, then by following the normal paths backwards we would have that either q_1 lies on the path from a to y or else a lies on the path from q_1 to x_1. In the former case there would be a directed cycle from q_1 to a to q_1, contradicting that the network is acyclic. In the latter case the hybrid vertex a would lie on the normal path from q_1 to x_1, contradicting that it is a normal path.

Suppose $p \in Par(N)$ is a parent map that satisfies $p(a) = q_1$, and let p' denote the complementary parent map that agrees with p except that $p'(a) = q_2$. Thus N_p and $N_{p'}$ agree except that N_p contains the arc (q_1, a) while $N_{p'}$ contains instead the arc (q_2, a). In particular they both contain the same paths from q_1 to x_1, from q_2 to x_2, and from a to y. Let $v = \text{mrca}(q_1, q_2; N_p)$. There is a directed path from r to v since r is the root (possibly $r = v$). There are directed paths from v to q_1 and v to q_2 in N_p which are disjoint except for v. Figure 5b thus shows a portion of N_p relevant to the lemma, together with the arc (q_2, a).

In N_p we see from Figure 5b that
$d(y, x_1; N_p) = d(a, y; N_p) + \omega(q_1, a) + d(q_1, x_1; N_p)$,
$d(y, x_2; N_p) = d(a, y; N_p) + \omega(q_1, a) + d(q_1, q_2; N_p) + d(q_2, x_2; N_p)$,
$d(x_1, x_2; N_p) = d(q_1, x_1; N_p) + d(q_1, q_2; N_p) + d(q_2, x_2; N_p)$.

By substituting these formulas we see that $[d(y, x_1; N_p) + d(y, x_2; N_p) - d(x_1, x_2; N_p)]/2 = d(a, y; N_p) + \omega(q_1, a)$. Since $\omega(q_1, a) = 0$ because a is hybrid, it follows
$[d(y, x_1; N_p) + d(y, x_2; N_p) - d(x_1, x_2; N_p)]/2 = d(a, y; N_p)$

The network $N_{p'}$ is the same except that (q_1, a) is replaced by (q_2, a). A symmetric argument then shows
$[d(y, x_1; N_{p'}) + d(y, x_2; N_{p'}) - d(x_1, x_2; N_{p'})]/2$
$= d(a, y; N_{p'}) + \omega(q_2, a) = d(a, y; N_{p'})$.

Since the indegree of a is 2, every parent map p satisfies either $p(a) = q_1$ or $p(a) = q_2$. It follows that for every $p \in Par(N)$,
$[d(y, x_1; N_p) + d(y, x_2; N_p) - d(x_1, x_2; N_p)]/2 = d(a, y; N_p)$.
When we take the expected value over all $p \in Par(N)$ we obtain by linearity
$[d(y, x_1; N) + d(y, x_2; N) - d(x_1, x_2; N)]/2 = d(a, y; N)$. □

Lemma 4.5. *Assume the hypotheses of Theorem 4.1. Suppose (a, b) is an arc such that b is normal, and a is hybrid with indegree 2 and parents q_1 and q_2. Assume q_1 and q_2 have normal paths to x_1 and x_2 respectively in X. Suppose b has normal paths to w and z in X where the paths are disjoint except for b. Then*
$\omega(a, b) = [d(x_1, w; N) + d(x_2, z; N) - d(x_1, x_2; N) - d(w, z; N)]/2$.

Proof. Since b is normal and the paths from b to w and from b to z are normal and disjoint except for b, we have
$d(w,z;N_p) = d(b,w;N_p) + d(b,z;N_p)$ for every parent map p, whence
$d(w,z;N) = d(b,w;N) + d(b,z;N)$. Similarly $d(a,w;N) = \omega(a,b) + d(b,w;N)$ and $d(a,z;N) = \omega(a,b) + d(b,z;N)$.

Hence $[d(a,w;N) + d(a,z;N) - d(w,z;N)]/2$
$= [\omega(a,b) + d(b,w;N) + \omega(a,b) + d(b,z;N) - d(b,w;N) - d(b,z;N)]/2$
$= \omega(a,b)$.

In addition, Lemma 4.4 applies with y replaced by w since the path from a to b to w is normal. Hence
$d(a,w;N) = [d(w,x_1;N) + d(w,x_2;N) - d(x_1,x_2;N)]/2$.
Lemma 4.4 also applies with y replaced by z. Hence
$d(a,z;N) = [d(z,x_1;N) + d(z,x_2;N) - d(x_1,x_2;N)]/2$.
By substitution it follows $\omega(a,b) = [d(a,w;N) + d(a,z;N) - d(w,z;N)]/2$
$= [d(w,x_1;N)+d(w,x_2;N)-2d(x_1,x_2;N)+d(z,x_1;N)+d(z,x_2;N)-2d(w,z;N)]/4$.

But symmetry shows that for each parent map p,
$d(w,x_2;N_p) + d(z,x_1;N_p) = d(w,x_1;N_p) + d(z,x_2;N_p)$.
Hence by taking the expected value over $p \in Par(N)$, we have
$d(w,x_2;N) + d(z,x_1;N) = d(w,x_1;N) + d(z,x_2;N)$.

Thus $\omega(a,b) = [2d(w,x_1;N) - 2d(x_1,x_2;N) + 2d(z,x_2;N) - 2d(w,z;N)]/4$
$= [d(w,x_1;N) - d(x_1,x_2;N) + d(z,x_2;N) - d(w,z;N)]/2$. □

For the next calculations we require a preliminary result. Suppose h_0 is hybrid with indegree 2 and parents q_1 and q_2. For a given parent map p with $p(h_0) = q_1$, 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_0). Let H be the set of hybrid vertices of N. For each $p \in Par(N)$ satisfying $p(h_0) = q_1$, let $W(p) = \prod[\alpha(p(h),h) : h \in H, h \neq h_0]$. Hence $Pr(p) = \alpha(q_1,h_0)W(p)$ and $Pr(p') = \alpha(q_2,h_0)W(p)$.

Lemma 4.6. *For any X-network M which is a subnetwork of N, suppose $C(M)$ is a linear combination of expressions of form $d(a,b;M)$. Then*
(1) $C(G_p) = \alpha(q_1,h_0)C(N_p) + \alpha(q_2,h_0)C(N_{p'})$.
(2) $C(N) = \sum[W(p)C(G_p) : p \in Par(N), p(h_0) = q_1]$.

Proof. For (1), $d(x,y;G_p) = \alpha(q_1,h_0)d(x,y;N_p) + \alpha(q_2,h_0)d(x,y;N_{p'})$.

For (2) each term $d(a,b;N) = \sum Pr(p)d(a,b;N_p)$. Hence
$C(N) = \sum Pr(p)C(N_p)$ by linearity
$= \sum[Pr(p)C(N_p) + Pr(p')C(N_{p'}) : p(h_0) = q_1]$
$= \sum[\alpha(q_1,h_0)W(p)C(N_p) + \alpha(q_2,h_0)W(p)C(N_{p'}) : p(h_0) = q_1]$
$= \sum[W(p)[\alpha(q_1,h_0)C(N_p) + \alpha(q_2,h_0)C(N_{p'})] : p(h_0) = q_1]$
$= \sum[W(p)C(G_p) : p \in Par(N), p(h_0) = q_1]$. □

Lemma 4.7. *Assume the hypotheses of Theorem 4.1. Suppose a is hybrid with indegree 2 and parents q_1 and q_2. Assume the inheritance is equiprobable at a. Suppose there is a normal path from q_1 to $x_1 \in X$, from q_2 to $x_2 \in X$, and from a to $y \in X$. Then*
$d(q_1,x_1;N) = d(x_1,y;N) - d(r,y;N) + [d(r,x_1;N) + d(r,x_2;N) - d(x_1,x_2;N)]/2$.

Proof. See Figure 5b. As in the proof of Lemma 4.4, the portion of the figure including the paths from q_1 to x_1, from q_2 to x_2, from a to y and the arcs (q_1, a) and (q_2, a) accurately represents the hypotheses of the lemma since N is normal. Suppose $p \in Par(N)$ satisfies $p(a) = q_1$. Let p' denote the complementary parent map such that $p'(a) = q_2$. Then all three normal paths in the statement lie in both N_p and $N_{p'}$ since they contain no hybrid arcs. Note that N_p contains (q_1, a) and not (q_2, a), while $N_{p'}$ contains (q_2, a) but not (q_1, a). Moreover, the path in N_p between q_1 and q_2 must be the same as the path in $N_{p'}$ between q_1 and q_2. Let $v = \text{mrca}(q_1, q_2; N_p)$; then v is also $\text{mrca}(q_1, q_2; N_{p'})$.

For any phylogenetic X-network M with the same base-set X write
$L(M) = d(x_1, y; M) - d(r, y; M) + [d(r, x_1; M) + d(r, x_2; M) - d(x_1, x_2; M)]/2$.
Note that L is a linear expression.

In both N_p and $N_{p'}$,
$d(r, x_1) = d(r, v) + d(v, q_1) + d(q_1, x_1)$
$d(r, x_2) = d(r, v) + d(v, q_2) + d(q_2, x_2)$
$d(x_1, x_2) = d(q_1, x_1) + d(v, q_1) + d(v, q_2) + d(q_2, x_2)$.
Hence $[d(r, x_1) + d(r, x_2) - d(x_1, x_2)]/2 = d(r, v)$.

In N_p we find
$d(x_1, y; N_p) = d(x_1, q_1; N_p) + \omega(q_1, a) + d(a, y; N_p)$, and
$d(r, y; N_p) = d(r, v; N_p) + d(v, q_1; N_p) + \omega(q_1, a) + d(a, y; N_p)$.
Hence $L(N_p)$
$= d(x_1, y; N_p) - d(r, y; N_p) + [d(r, x_1; N_p) + d(r, x_2; N_p) - d(x_1, x_2; N_p)]/2$
$= d(x_1, y; N_p) - d(r, y; N_p) + d(r, v; N_p)$
$= d(x_1, q_1; N_p) + \omega(q_1, a) + d(a, y; N_p) - d(r, v; N_p) - d(v, q_1; N_p) - \omega(q_1, a) - d(a, y; N_p) + d(r, v; N_p)$
$= d(x_1, q_1; N_p) - d(v, q_1; N_p)$.

In $N_{p'}$ we find
$d(x_1, y; N_{p'}) = d(q_1, x_1; N_{p'}) + d(v, q_1; N_{p'}) + d(v, q_2; N_{p'}) + \omega(q_2, a) + d(a, y; N_{p'})$
$d(r, y; N_{p'}) = d(r, v; N_{p'}) + d(v, q_2; N_{p'}) + \omega(q_2, a) + d(a, y; N_{p'})$.
Hence $L(N_{p'})$
$= d(x_1, y; N_{p'}) - d(r, y; N_{p'}) + [d(r, x_1; N_{p'}) + d(r, x_2; N_{p'}) - d(x_1, x_2; N_{p'})]/2$
$= d(x_1, y; N_{p'}) - d(r, y; N_{p'}) + d(r, v; N_{p'}) = d(q_1, x_1; N_{p'}) + d(v, q_1; N_{p'})$.

Thus $L(N_p) + L(N_{p'})$
$= d(q_1, x_1; N_p) - d(v, q_1; N_p) + d(q_1, x_1; N_{p'}) + d(v, q_1; N_{p'})$
$= d(q_1, x_1; N_p) + d(q_1, x_1; N_{p'})$
since $d(v, q_1; N_p) = d(v, q_1; N_{p'})$.

Using Lemma 4.6(1) with $h_0 = a$, we see that
$L(G_p) = \alpha(q_1, a)L(N_p) + \alpha(q_2, a)L(N_{p'})$ so
$L(G_p) = (1/2)[L(N_p) + L(N_{p'})]$ by equiprobability at a.
From above it follows
$L(G_p) = (1/2)d(q_1, x_1; N_p) + (1/2)d(q_1, x_1; N_{p'})$.

By Lemma 4.6(2)
$L(N) = \sum[W(p)L(G_p) : p \in Par(N), p(a) = q_1]$
$= \sum[W(p)(1/2)d(q_1, x_1; N_p) + W(p)(1/2)d(q_1, x_1; N_{p'}) : p(a) = q_1]$
$= \sum[Pr(p)d(q_1, x_1; N_p) + Pr(p')d(q_1, x_1; N_{p'}) : p \in Par(N), p(a) = q_1]$
$= \sum[Pr(p)d(q_1, x_1; N_p) : p \in Par(N)]$

$= d(q_1, x_1; N)$. $\square$

It is interesting in the proof that different choices of the parent map p may yield different vertices v; nevertheless all these choices cancel out.

Lemma 4.8. *Assume the hypotheses of Theorem 4.1. Suppose h is hybrid with indegree 2 and parents q_1 and q_2. Assume equiprobable inheritance at h. Suppose there is a normal path from q_2 to $x_2 \in X$ and from h to $y \in X$. Suppose q_1 has normal child b and there are normal paths from b to $z_1 \in X$ and from b to $z_2 \in X$ such that these paths intersect only at b. Then*
$\omega(q_1, b) = [2d(z_1, y; N) - 4d(r, y; N) + d(r, z_1; N) + 2d(r, x_2; N) - d(z_1, x_2; N) + 2d(z_2, y; N) + d(r, z_2; N) - d(z_2, x_2; N) - 2d(z_1, z_2; N)]/4$.

In particular, if b is a leaf, then $\omega(q_1, b) = [2d(b, y; N) - 2d(r, y; N) + d(r, b; N) + d(r, x_2; N) - d(b, x_2; N)]/2$.

Proof. By an argument like that for Lemma 4.2, for each $p \in Par(N)$ we have
$\omega(q_1, b) = [d(q_1, z_1; N_p) + d(q_1, z_2; N_p) - d(z_1, z_2; N_p)]/2$
whence by averaging over $p \in Par(N)$ we find
$\omega(q_1, b) = [d(q_1, z_1; N) + d(q_1, z_2; N) - d(z_1, z_2; N)]/2$.

But the paths from q_1 to z_1 and from q_2 to z_2 are normal, so by Lemma 4.7
$d(q_1, z_1; N) = d(z_1, y; N) - d(r, y; N) + [d(r, z_1; N) + d(r, x_2; N) - d(z_1, x_2; N)]/2$
and
$d(q_1, z_2; N) = d(z_2, y; N) - d(r, y; N) + [d(r, z_2; N) + d(r, x_2; N) - d(z_2, x_2; N)]/2$.
Hence $\omega(q_1, b) = [d(z_1, y; N) - d(r, y; N) + [d(r, z_1; N) + d(r, x_2; N) - d(z_1, x_2; N)]/2$
$+ d(z_2, y; N) - d(r, y; N) + [d(r, z_2; N) + d(r, x_2; N) - d(z_2, x_2; N)]/2 - d(z_1, z_2; N)]/2$
$= [2d(z_1, y; N) - 2d(r, y; N) + d(r, z_1; N) + d(r, x_2; N) - d(z_1, x_2; N) + 2d(z_2, y; N) - 2d(r, y; N) + d(r, z_2; N) + d(r, x_2; N) - d(z_2, x_2; N) - 2d(z_1, z_2; N)]/4$
$= [2d(z_1, y; N) - 4d(r, y; N) + d(r, z_1; N) + 2d(r, x_2; N) - d(z_1, x_2; N) + 2d(z_2, y; N) + d(r, z_2; N) - d(z_2, x_2; N) - 2d(z_1, z_2; N)]/4$.

If b is a leaf we may take $b = z_1 = z_2$ to obtain
$\omega(q_1, b) = [2d(b, y; N) - 4d(r, y; N) + d(r, b; N) + 2d(r, x_2; N) - d(b, x_2; N) + 2d(b, y; N) + d(r, b; N) - d(b, x_2; N) - 2d(b, b; N)]/4$
$= [4d(b, y; N) - 4d(r, y; N) + 2d(r, b; N) + 2d(r, x_2; N) - 2d(b, x_2; N) - 2d(b, b; N)]/4$
$= [2d(b, y; N) - 2d(r, y; N) + d(r, b; N) + d(r, x_2; N) - d(b, x_2; N)]/2$. $\square$

We next prove analogues of Lemma 4.7 and Lemma 4.8 for the case where the hybrid is not equiprobable and we are dealing with the situation in Figure 2 rather than Figure 5b.

Lemma 4.9. *Assume the hypotheses of Theorem 4.1. Suppose h_0 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) $w_{q_2x_2}(M) = d(r, x_2; M) - w_{rv}(M) - w_{vq_3}(M) - w_{q_3q_2}(M)$
(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) $d(q_2, x_2; N) = w_{q_2x_2}(N)$.

Proof. Suppose $p \in Par(N)$ is a parent map satisfying $p(h_0) = q_1$ and p' is the complementary parent map agreeing with p except that $p'(h_0) = q_2$. Let $G_p = N_p$ with the additional arc (q_2, h_0), so $G_p = N_p \cup N_{p'}$. A portion of G_p is shown in Figure 2. Note that Figure 2 is accurate for every p (although the vertex v may differ for different p) because of the hypotheses on q_1, q_2, q_3, h_0, x_1, x_2, x_3, and y.

Write $u_{rv} = d(r, v; G_p)$, $u_{vq_1} = d(v, q_1; G_p)$, $u_{vq_3} = d(v, q_3; G_p)$, $u_{q_3x_3} = d(q_3, x_3; G_p)$, $u_{q_3q_2} = d(q_3, q_2; G_p)$, $u_{q_2x_2} = d(q_2, x_2; G_p)$, $u_{hy} = d(h, y; G_p)$, $u_{q_1x_1} = d(q_1, x_1; G_p)$.

The definition of the tree-average distance yields the following ten equations for G_p, where $\alpha = \alpha(q_1, h_0)$:
$d(r, x_1; G_p) = u_{rv} + u_{vq_1} + u_{q_1x_1}$
$d(r, x_3; G_p) = u_{rv} + u_{vq_3} + u_{q_3x_3}$
$d(r, x_2; G_p) = u_{rv} + u_{vq_3} + u_{q_3q_2} + u_{q_2x_2}$
$d(r, y; G_p) = \alpha[u_{rv} + u_{vq_1} + u_{hy}] + (1 - \alpha)[u_{rv} + u_{vq_3} + u_{q_3q_2} + u_{hy}]$
$= u_{rv} + u_{hy} + \alpha u_{vq_1} + (1 - \alpha)(u_{vq_3} + u_{q_3q_2})$
$d(x_1, x_3; G_p) = u_{q_1x_1} + u_{vq_1} + u_{vq_3} + u_{q_3x_3}$
$d(x_1, x_2; G_p) = u_{q_1x_1} + u_{vq_1} + u_{vq_3} + u_{q_3q_2} + u_{q_2x_2}$
$d(x_1, y; G_p) = \alpha[u_{q_1x_1} + u_{hy}] + (1 - \alpha)[u_{q_1x_1} + u_{vq_1} + u_{vq_3} + u_{q_3q_2} + u_{hy}]$
$= u_{q_1x_1} + u_{hy} + (1 - \alpha)[u_{vq_1} + u_{vq_3} + u_{q_3q_2}]$
$d(x_3, x_2; G_p) = u_{q_3x_3} + u_{q_3q_2} + u_{q_2x_2}$
$d(x_3, y; G_p) = \alpha[u_{q_3x_3} + u_{vq_3} + u_{vq_1} + u_{hy}] + (1 - \alpha)[u_{q_3x_3} + u_{q_3q_2} + u_{hy}]$
$= u_{q_3x_3} + u_{hy} + \alpha(u_{vq_3} + u_{vq_1}) + (1 - \alpha)(u_{q_3q_2})$
$d(x_2, y; G_p) = \alpha[u_{q_2x_2} + u_{q_3q_2} + u_{vq_3} + u_{vq_1} + u_{hy}] + (1 - \alpha)[u_{q_2x_2} + u_{hy}]$
$= u_{q_2x_2} + u_{hy} + \alpha(u_{q_3q_2} + u_{vq_3} + u_{vq_1})$

We now solve this system of ten equations.

It is straightforward by simplifying the expressions that
$[d(r,x_1;G_p)+d(r,x_3;G_p)-d(x_1,x_3;G_p)]/2=u_{rv}$
so a comparison with (a) shows that $w_{rv}(G_p)=u_{rv}$. Similarly $[d(r,x_1;G_p)+d(r,x_2;G_p)-d(x_1,x_2;G_p)]/2=u_{rv}$ so the two expressions in (a) for $w_{rv}(G_p)$ are the same.

Likewise from the ten equations,
$[d(r,x_3;G_p)+d(x_1,x_2;G_p)-d(r,x_1;G_p)-d(x_3,x_2;G_p)]/2=u_{vq_3}$
so $w_{vq_3}(G_p)=u_{vq_3}$;
$[d(r,x_3;G_p)+d(x_3,x_2;G_p)-d(r,x_2;G_p)]/2=u_{q_3x_3}$
so $w_{q_3x_3}(G_p)=u_{q_3x_3}$;
$[d(y,x_2;G_p)+d(y,x_1;G_p)-d(x_1,x_2;G_p)]/2=u_{hy}$ so $w_{hy}(G_p)=u_{hy}$.

From the system of ten equations we see
$E_2(G_p)=u_{q_1x_1}+(1-\alpha)u_{vq_1}-\alpha u_{vq_1}$
$=u_{q_1x_1}+(1-2\alpha)u_{vq_1}$.

Since $d(r,x_1;G_p)=u_{rv}+u_{vq_1}+u_{q_1x_1}$ it follows $d(r,x_1;G_p)=u_{rv}+u_{vq_1}+E_2(G_p)-(1-2\alpha)u_{vq_1}$ whence

$$2\alpha u_{vq_1}=d(r,x_1;G_p)-E_2(G_p)-u_{rv} \tag{1}$$

Similarly $E_4(G_p)=u_{q_2x_2}+u_{hy}+\alpha(u_{q_3q_2}+u_{vq_3}+u_{vq_1})-u_{rv}-u_{hy}-\alpha u_{vq_1}-(1-\alpha)(u_{vq_3}+u_{q_3q_2})+u_{rv}$
$=u_{q_2x_2}+\alpha(u_{q_3q_2}+u_{vq_3})-(1-\alpha)(u_{vq_3}+u_{q_3q_2})=u_{q_2x_2}+(2\alpha-1)(u_{vq_3}+u_{q_3q_2})$.

But from $d(r,x_2;G_p)=u_{rv}+u_{vq_3}+u_{q_3q_2}+u_{q_2x_2}$ it follows $u_{q_2x_2}=d(r,x_2;G_p)-u_{rv}-u_{vq_3}-u_{q_3q_2}$ so $E_4(G_p)=d(r,x_2;G_p)-u_{rv}-u_{vq_3}-u_{q_3q_2}+(2\alpha-1)(u_{vq_3}+u_{q_3q_2})$. This can be solved to show

$$(2-2\alpha)(u_{vq_3}+u_{q_3q_2})=d(r,x_2;G_p)-u_{rv}-E_4(G_p) \tag{2}$$

Since $d(x_3,y;G_p)=u_{q_3x_3}+u_{hy}+\alpha(u_{vq_3}+u_{vq_1})+(1-\alpha)(u_{q_3q_2})$ we obtain

$$\alpha(u_{vq_3}+u_{vq_1})+(1-\alpha)(u_{q_3q_2})=d(x_3,y)-u_{q_3x_3}-u_{hy} \tag{3}$$

Note (1), (2), and (3) are equations in the unknowns α, w_{vq_1}, $w_{q_3q_2}$ in terms of known quantities such as w_{rv}, $w_{q_3x_3}$, w_{hy}, w_{vq_3}, $E_4(G_p)$. These three equations in three unknowns can be solved to yield for G_p (for any $p\in Par(N)$ with $p(h)=q_1$) the following:
$\alpha(G_p)=[2d(x_3,y;G_p)-2w_{q_3x_3}-2w_{hy}-d(r,x_1;G_p)+E_2(G_p)+2w_{rv}+E_4(G_p)-d(r,x_2;G_p)+2w_{vq_3}]/[4w_{vq_3}]$
$w_{vq_1}(G_p)=[d(r,x_1;G_p)-E_2(G_p)-w_{rv}]/[2\alpha(G_p)]$
$w_{q_3q_2}(G_p)=[d(x_3,y;G_p)-w_{q_3x_3}-w_{hy}-\alpha(w_{vq_3}+w_{vq_1})]/(1-\alpha(G_p))$.

Moreover, the value of α is independent of the choice of p.

We thus have $C(G_p)=\alpha D(G_p)$ for each p satisfying $p(h_0)=q_1$.

By Lemma 4.6, $C(N)=\sum[W(p)C(G_p):p(h_0)=q_1]$
and $D(N)=\sum[W(p)D(G_p):p(h_0)=q_1]$.

Hence $C(N)=\sum[W(p)\alpha D(G_p):p(h_0)=q_1]$
$=\alpha\sum[W(p)D(G_p):p(h_0)=q_1]=\alpha D(N)$.

It follows that $\alpha = C(N)/D(N)$. This proves (i).

Similarly, for any $p \in Par(N)$ satisfying $p(h_0) = q_1$, since the path from q_1 to x_1 is normal, $d(q_1, x_1; N) = d(q_1, x_1; G_p) = w_{q_1 x_1}(G_p)$. By Lemma 4.10 $d(q_1, x_1; N) = \sum[W(p)d(q_1, x_1; G_p) : p \in Par(N), p(h_0) = q_1]$
$= \sum[W(p)w_{q_1 x_1}(G_p) : p \in Par(N), p(h_0) = q_1]$
$= w_{q_1 x_1}(N)$, proving (ii). Similarly $d(q_2, x_2; N) = w_{q_2 x_2}(N)$, proving (iii). □

Lemma 4.10. *Assume the hypotheses of Theorem 4.1. Suppose h_0 is hybrid with indegree 2 and parents q_1 and q_2. Suppose there is a normal path from q_3 to q_2, from q_2 to $x_2 \in X$, from q_1 to $x_1 \in X$, from h_0 to $y \in X$, and from q_3 to $x_3 \in X$ but no directed path from q_3 to q_1.*
(a) Suppose q_1 has normal child b and there are normal paths from b to $x_1 \in X$ and from b to $z_1 \in X$ such that these paths intersect only at b. Then $\omega(q_1, b) = [d(q_1, x_1; N) + d(q_1, z_1; N) - d(x_1, z_1; N)]/2$, where $d(q_1, x_1; N)$ and $d(q_1, z_1; N)$ are determined by Lemma 4.9.
(b) Suppose q_2 has normal child c and there are normal paths form c to $x_2 \in X$ and from c to $z_2 \in X$ such that these paths intersect only at c. Then $\omega(q_2, c) = [d(q_2, x_2; N) + d(q_2, z_2; N) - d(x_2, z_2; N)]/2$, where $d(q_2, x_2; N)$ and $d(q_2, z_2; N)$ are determined by Lemma 4.9.

Proof. For (a), Lemma 4.9 applies to yield $d(q_1, x_1; N)$. By a parallel computation with z_1 replacing x_1, Lemma 4.9 also yields $d(q_1, z_1; N)$. Since the paths from q_1 to x_1 and z_1 are normal, it follows that $\omega(q_1, b) = d(q_1, b; N)$
$= [d(q_1, x_1; N) + d(q_1, z_1; N) - d(x_1, z_1; N)]/2$ by an argument like that of Lemma 4.2. A similar argument shows (b). □

We now turn to the proof of the main theorem 4.1:

Proof. We seek to reconstruct each weight $\omega(a, b)$ and each probability. If b is hybrid, then by assumption $\omega(a, b) = 0$. Hence we may assume b is normal.

At the tail a we have the following exhaustive list of possibilities:

Case A_1. There is a normal path from a to some $w \in X$ such that the path does not go through b. This includes the possibility where $a \in X$ (in which case the trivial path at a satisfies the condition). Since $r \in X$, this includes the case $a = r$.

Case A_2. a is hybrid and b is its unique child. Since a is hybrid it has two parents q_1 and q_2. Choose a normal path from q_1 to $w_1 \in X$ and from q_2 to $w_2 \in X$.

Case A_3. a has a hybrid child h' with other parent q'. Choose a normal path from q' to $w_1 \in X$ and from h' to $w_2 \in X$.

At the head b, either $b \in X$ or else b is not a leaf and b has at least two children, at least one of which must be normal. Hence we have the following exhaustive list of possibilities:

Case B_1. $b \in X$.

Case B_2. b has two normal children c_1 and c_2. For $i = 1, 2$ there is a normal path from c_i to $x_i \in X$.

Case B_3. b has one normal child c and a hybrid child h for which there is exactly one other parent q. There is a normal path from c to $x \in X$, from h to $y \in X$, and from q to $z \in X$.

Since there are 3 cases for a and three cases for b, we must consider 9 cases. The case where A_i is combined with B_j will be denoted Case A_iB_j. We will compute $\omega(a, b)$. To compute the probabilities, it suffices to compute $\alpha(a, h')$ in situation A_3.

Case A_1B_1. Assume there is a normal path from a to some $w \in X$ such that the path does not go through b, and $b \in X$. Then Lemma 4.3(2) shows that
$\omega(a, b) = [d(r, b; N) + d(w, b; N) - d(r, w; N)]/2$.

Case A_1B_2. Assume there is a normal path from a to some $w \in X$ such that the path does not go through b. Assume b has two normal children c_1 and c_2. For $i = 1, 2$ there is a normal path from c_i to $x_i \in X$. In this case, Lemma 4.3(3) shows that
$\omega(a, b) = [d(r, x_1; N) + d(w, x_2; N) - d(r, w; N) - d(x_1, x_2; N)]/2$.

Case A_2B_1. Assume a is hybrid and b is its unique child. Assume $b \in X$. Since a is hybrid it has two parents q_1 and q_2. Choose a normal path from q_1 to $w_1 \in X$ and from q_2 to $w_2 \in X$. In this case, Lemma 4.4 shows that
$\omega(a, b) = [d(b, w_1; N) + d(b, w_2; N) - d(w_1, w_2; N)]/2$.

Case A_2B_2. Assume a is hybrid and b is its unique child. Since a is hybrid it has two parents q_1 and q_2. Choose a normal path from q_1 to $w_1 \in X$ and from q_2 to $w_2 \in X$. Assume b has two normal children c_1 and c_2. For $i = 1, 2$ there is a normal path from c_i to $x_i \in X$. In this case by Lemma 4.5 we have
$\omega(a, b) = [d(w_1, x_1; N) + d(w_2, x_2; N) - d(w_1, w_2; N) - d(x_1, x_2; N)]/2$.

Case A_3B_1. Assume a has a hybrid child h' with other parent q'. Choose a normal path from q' to $w_1 \in X$ and from h' to $w_2 \in X$. Assume $b \in X$. In the equiprobable case, Lemma 4.7 with $q_1 = a$, $x_1 = b$, $x_2 = w_2$ shows
$\omega(a, b) = d(a, b; N) = d(b, w_2; N) - d(r, w_2; N) + [d(r, b; N) + d(r, w_1; N) - d(b, w_1; N)]/2$.
In the other case, Lemma 4.9(ii) with $q_1 = a$ and $x_1 = b$ yields $\omega(a, b)$ while Lemma 4.9(i) yields $\alpha(a, h')$.

Case A_3B_2. Assume a has a hybrid child h' with other parent q'. Choose a normal path from q' to $w_1 \in X$ and from h' to $w_2 \in X$. Assume b has two normal children c_1 and c_2. For $i = 1, 2$ there is a normal path from c_i to $x_i \in X$. In the equiprobable case, Lemma 4.8 with $q_1 = a$, $y = w_2$, $z_1 = x_1$, $z_2 = x_2$, $h = h'$, $q_2 = q'$, $x_2 = w_1$ shows
$\omega(a, b) = [2d(x_1, w_2; N) - 4d(r, w_2; N) + d(r, x_1; N) + 2d(r, w_1; N) - d(x_1, w_1; N) + 2d(x_2, w_2; N) + d(r, x_2; N) - d(x_2, w_1) - 2d(x_1, x_2)]/4$.
In the non-equiprobable case Lemma 4.10a applies to determine $\omega(a, b)$, while Lemma 4.9(i) determines $\alpha(a, h')$.

Case A_1B_3. Assume that there is a normal path from a to some $w \in X$ such that the path does not go through b. Assume b has one normal child c and a hybrid child h for which there is exactly one other parent q. There is a normal path from c to $x \in X$, from h to $y \in X$, and from q to $z \in X$. See Figure 6. Since N is normal, an argument like that for Lemma 4.4 shows that Figure 6 is accurate for the situation.

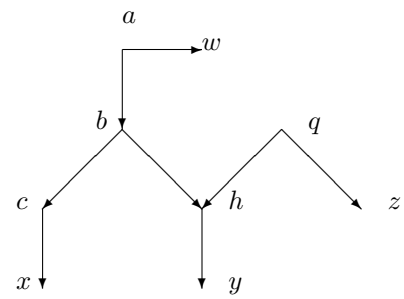

Figure 6: Case A_1B_3.

In this situation, by Lemma 4.4(2), $d(a,x;N) = [d(x,r;N) + d(x,w;N) - d(r,w;N)]/2$. In the equiprobable case, by Lemma 4.7, with $b = q_1$, $x_1 = x$, $z = x_2$,
$d(b,x;N) = d(x,y;N) - d(r,y;N) + [d(r,x;N) + d(r,z;N) - d(x,z;N)]/2$.
Finally $\omega(a,b) = d(a,x;N) - d(b,x;N)$. In the non-equiprobable case, Lemma 4.9 with $a = q_3$ and $b = q_2$ yields the computation of $\omega(a,b) = w_{q_3,q_2}(N)$ and Lemma 4.9(i) shows $\alpha(b,h) = \alpha(q_2,h;N) = 1 - \alpha(q_1,h;N)$.

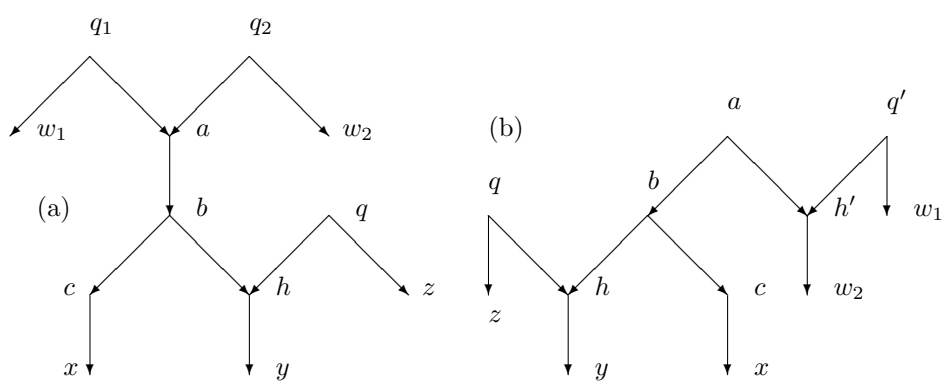

Figure 7: (a) Case A_2B_3. (b) Case A_3B_3.

Case A_2B_3. Assume a is hybrid and b is its unique child. Since a is hybrid it has two parents q_1 and q_2. Choose a normal path from q_1 to $w_1 \in X$ and from q_2 to $w_2 \in X$. Assume b has one normal child c and a hybrid child h for which there is exactly one other parent q. Choose a normal path from c to $x \in X$, from h to $y \in X$, and from q to $z \in X$.

See Figure 7a. An argument like that for Lemma 4.4 shows that the figure

accurately represents what is needed in the argument. In particular, the normal paths from q_1 to w_1, from q_2 to w_2, and from q to x have no vertex in common. Similarly the paths from q to z, from b to x, and from h to y have no vertex in common.

By Lemma 4.4, $d(a,x;N) = [d(x,w_1;N) + d(x,w_2;N) - d(w_1,w_2;N)]/2$.
In the equiprobable case, by Lemma 4.7, $d(b,x;N) = d(x,y;N) - d(r,y;N) + [d(r,x;N) + d(r,z;N) - d(x,z;N)]/2$.
In the non-equiprobable case, Lemma 4.9(ii) or 4.9(iii) similarly yields $d(b,x;N)$. But $\omega(a,b) = d(a,x;N) - d(b,x;N)$ since the path from a to x is normal, so subtracting these formulas leads to a formula for $\omega(a,b)$.

Case A_3B_3. Assume that a has a hybrid child h' with other parent q'. Choose a normal path from q' to $w_1 \in X$ and from h' to $w_2 \in X$. Assume b has one normal child c and a hybrid child h for which there is exactly one other parent q. Choose is a normal path from c to $x \in X$, from h to $y \in X$, and from q to $z \in X$.

See Figure 7b. The argument will make two uses of Lemma 4.7 or 4.9, and Figure 7b accurately represents the situation by arguments like those in Lemma 4.4.

In the equiprobable case, by Lemma 4.7,
$d(a,x;N) = d(x,w_2;N) - d(r,w_2;N) + [d(r,x;N) + d(r,w_1;N) - d(x,w_1;N)]/2$,
$d(b,x;N) = d(x,y;N) - d(r,y;N) + [d(r,x;N) + d(r,z;N) - d(x,z;N)]/2$.
But then $\omega(a,b) = d(a,x;N) - d(b,x;N)$ since the path from a to x is normal. In the other case, Lemma 4.9(ii) or 4.9(iii) yields $d(a,x;N)$ and $d(b,x;N)$ and again $\omega(a,b)$ is determined. Moreover, Lemma 4.9(i) yields $\alpha(a,h')$ and $\alpha(q,h)$.

Since all 9 cases yield a formula for $\omega(a,b)$ and also any relevant probability when a is parent to a hybrid and b is a normal child of a, the proof of the theorem is complete. □

Corollary 4.11. *Suppose $N = (V,A,r,X)$ is a normal phylogenetic X-network such that each hybrid vertex has indegree 2 and, if it is not a leaf, outdegree 1. Let $n = |X|$ and a be the total number of arcs directed into any normal vertex. Then $a \leq \binom{n}{2}$.*

Proof. We may assume that the arcs have weights and that each hybrid is equiprobable. Each of the weights $\omega(u,v)$ if (u,v) is an arc directed into a normal vertex v is uniquely determined from the $\binom{n}{2}$ linear equations obtained from the $\binom{n}{2}$ distances given by the tree-average distance function. Hence there are at most $\binom{n}{2}$ variables. □

Figure 1 gives an example in which $n = 4$ and there are exactly $\binom{4}{2} = 6$ arcs directed into a normal vertex. Hence the bound in Corollary 4.11 is tight.

5 An example

We illustrate the calculations of Section 4 to find the values of the weight function given the network and the tree-average distance. Figure 4 exhibits a phylo-

genetic X-network $N = (V, A, r, X)$ with $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ and root 1 which satisfies the hypotheses of Theorem 4.1. Observe that by Corollary 3.4, N displays exactly 4 trees, and there are exactly four parent maps. Let ω be a weight function on A such that $\omega(a, b) = 0$ when b is hybrid but $\omega(a, b) \geq 0$ when b is normal. Let $d(x, y) = d(x, y; N)$ denote the resulting tree-average distance between x and y in X. Suppose first that we assume equiprobability about the network, so each $\alpha(a, b) = 1/2$ when b is hybrid. There are 24 arcs for which we compute the weights as follows:

First, since 16 and 20 are hybrid, we have $\omega(17, 16) = \omega(15, 16) = \omega(21, 20) = \omega(23, 20) = 0$.

By Lemma 4.3(2),
$\omega(19, 8) = [d(8, 1) + d(7, 8) - d(1, 7)]/2$,
$\omega(19, 7) = [d(7, 1) + d(7, 8) - d(1, 8)]/2$,
and we similarly find $\omega(14, 3)$, $\omega(13, 2)$, and $\omega(22, 10)$.

By Lemma 4.3(1),
$\omega(1, 22) = [d(1, 9) + d(1, 11) - d(9, 11)]/2$.

By Lemma 4.3(3),
$\omega(18, 19) = [d(1, 8) + d(6, 7) - d(1, 6) - d(7, 8)]/2$,
$\omega(12, 13) = [d(1, 2) + d(11, 3) - d(1, 11) - d(2, 3)]/2$,
and we similarly find $\omega(13, 14)$ and $\omega(22, 12)$.

By Lemma 4.5,
$\omega(20, 18) = [d(9, 8) + d(11, 6) - d(9, 11) - d(8, 6)]/2$.

By Lemma 4.4,
$\omega(16, 5) = [d(5, 4) + d(5, 6) - d(4, 6)]/2$.

By Lemma 4.7 in the equiprobable case,
$\omega(21, 9) = d(9, 7) - d(1, 7) + [d(1, 9) + d(1, 11) - d(9, 11)]/2$,
$\omega(23, 11) = d(11, 7) - d(1, 7) + [d(1, 9) + d(1, 11) - d(9, 11)]/2$,
and we similarly find $\omega(17, 6)$ and $\omega(15, 4)$.

By Lemma 4.3(2),
$d(18, 6) = [d(6, 1) + d(6, 7) - d(1, 7)]/2$.
But then $\omega(18, 17) = d(18, 6) - \omega(17, 6)$.
Similarly by Lemma 4.2(2) $d(14, 4) = [d(4, 1) + d(4, 3) - d(1, 3)]/2$
and then $\omega(14, 15) = d(14, 4) - \omega(15, 4)$.
Similarly by Lemma 4.3(2) $d(12, 11) = [d(11, 1) + d(11, 2) - d(1, 2)]/2$
and then $\omega(12, 23) = d(12, 11) - \omega(23, 11)$.
Finally, $d(10, 9)$ is known since $10 \in X$, so $\omega(10, 21) = d(10, 9) - \omega(21, 9)$.

This concludes the calculation of all the weights for N in the equiprobable case. Note that in several of these calculations, there were alternative choices possible. For example, we also have
$\omega(22, 12) = [d(1, 4) + d(9, 11) - d(1, 9) - d(4, 11)]/2$.

The general case where we do not assume equiprobability proceeds in a similar manner, different from the above only in the use of Lemma 4.9 in place of Lemma 4.7. We compute $\omega(21, 9)$, $\omega(23, 11)$, $\alpha(21, 20)$, and $\alpha(23, 20)$ using Lemma 4.9 with $x_1 = 9$, $x_2 = 11$, $x_3 = 2$, and $y = 7$. We compute $\omega(17, 6)$, $\omega(15, 4)$, $\alpha(17, 16)$, and $\alpha(15, 16)$ using Lemma 4.9 with $x_1 = 6$, $x_2 = 4$, $x_3 = 3$, $y = 5$.

6 Extensions

Theorem 4.1 applies only to normal phylogenetic networks for which the indegree of each hybrid vertex is 2. It would be interesting to see whether the same results are true without the restriction on the indegree of a hybrid vertex. Whereas I have verified this for several individual networks with vertices of indegree 3 or 4, I do not have a general proof.

In the event of a true hybridization between two sexual species, it is plausible to assume that the indegree is 2 and that each parent contributes approximately equally. Hence in this case it is plausible that we would obtain the tree-average distance utilized in Theorem 4.1. Nevertheless, backcrossing of the hybrid h with one of the parental species q_1 could easily increase the fraction of the genome of q_1 in h, changing it from 50%. Similarly, if the reticulation is actually a horizontal gene transfer, common between bacteria, then there is no guarantee that the sources contribute approximately equally. Hence the occurrence of probabilities different from 1/2 seems likely.

Competing interests. The author has no competing interests.

Acknowledgments. I am indebted to Jesper Jansson for references about the first use of certain kinds of networks. I also thank Mukund Thattai and Mike Steel for useful discussions about the probabilities at hybrid vertices. Finally I am endebted to the anonymous referees for many helpful suggestions.

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] 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.

[4] 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.

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

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

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

[8] 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.

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

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

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

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

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