Phylogenetic Networks: Modeling, Reconstructibility, and Accuracy

Bernard M.E. Moret, Luay Nakhleh, Tandy Warnow, C. Randal Linder, Anna Tholse, Anneke Padolina, Jerry Sun, and Ruth Timme

Abstract—Phylogenetic networks model the evolutionary history of sets of organisms when events such as hybrid speciation and horizontal gene transfer occur. In spite of their widely acknowledged importance in evolutionary biology, phylogenetic networks have so far been studied mostly for specific data sets. We present a general definition of phylogenetic networks in terms of directed acyclic graphs (DAGs) and a set of conditions. Further, we distinguish between model networks and reconstructible ones and characterize the effect of extinction and taxon sampling on the reconstructibility of the network. Simulation studies are a standard technique for assessing the performance of phylogenetic methods. A main step in such studies entails quantifying the topological error between the model and inferred phylogenies. While many measures of tree topological accuracy have been proposed, none exist for phylogenetic networks. Previously, we proposed the first such measure, which applied only to a restricted class of networks. In this paper, we extend that measure to apply to *all* networks, and prove that it is a metric on the space of phylogenetic networks. Our results allow for the systematic study of existing network methods, and for the design of new accurate ones.

Index Terms—Phylogenetic networks, reticulate evolution, error metric, Robinson-Foulds, bipartitions, tripartitions.

1 Introduction

Phylogenies are the main tool for representing evolutionary relationships among biological entities. Their pervasiveness has led biologists, mathematicians, and computer scientists to design a variety of methods for their reconstruction (see, e.g., [28]). Almost all such methods, however, construct trees; yet, biologists have long recognized that trees oversimplify our view of evolution, since they cannot take into account such events as hybrid speciation and horizontal gene transfer. These nontree events, usually called *reticulations*, give rise to edges that connect nodes from different branches of a tree, creating a directed acyclic graph structure that is usually called a *phylogenetic network*. To date, no accepted methodology for network reconstruction has been proposed. Many researchers have studied closely related problems, such as the compatibility of tree splits [2], [3] and other indications that a tree structure is inadequate for the data at hand [15], detection and identification of horizontal gene transfer [7], [8], and, more generally, detection and identification of recombination events [19], [20]. A number of biological studies of reticulation have also appeared [21], [22], [23], [25]. Our group has proposed a first, very simple method [17] based on an observation of Maddison [13], but it remains limited to just a few reticulations.

- *B.M.E. Moret and A. Tholse are with the Department of Computer Science, University of New Mexico, Albuquerque, NM 87131. E-mail: {moret, tholse}@cs.umn.edu.*
- *L. Nakhleh, T. Warnow, and J. Sun are with the Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712. E-mail: {nakhleh, tandy, jsun}@cs.utexas.edu.*
- *C.R. Linder, A. Padolina, and R. Timme are with the School of Biological Sciences, University of Texas at Austin, Austin, TX 78712. E-mail: {rlinder, annekes, retimme}@mail.utexas.edu.*

Manuscript received 23 Jan. 2004; revised 21 May 2004; accepted 25 May 2004.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0005-0104.

In phylogenetic reconstruction, a standard methodology for assessing the quality of reconstruction methods is simulation [9], [10]. Simulation allows a direct comparison between the "true" phylogeny and its reconstruction, something that is generally not possible with real data, where the true phylogeny is at best only partially known. In a simulation study, a model phylogeny is generated, typically in two stages: A topology is created, and then the evolution of a set of molecular sequences is simulated on that topology. The set of sequences obtained at the leaves is then fed to the reconstruction methods under study and their output compared with the model phylogeny. While computationally expensive (the large parameter space and the need to obtain statistically significant results necessitate a very large number of tests), simulation studies provide an unbiased assessment of the quality of reconstruction, as well as a first step in the very difficult process of deriving formal bounds on the behavior of reconstruction methods.

A crucial part of a simulation study is the comparison between the reconstructed and the true phylogenies. Many methods have been devised to compute the error rate between two phylogenetic trees but, with the exception of our work [16], no such method has been proposed for phylogenetic networks. We proposed a measure of the relative error between two phylogenetic networks based on the tripartitions induced by the edges of the network [16], a measure that naturally extends a standard error measure used for trees (the Robinson-Foulds measure), and provided experimental results showing that our measure exhibited desirable properties.

In this paper, we formalize our results about this measure. We establish a framework for phylogenetic networks in terms of directed acyclic graphs (DAGs) and define a set of properties that a DAG must have in order to reflect a realistic phylogenetic network. A crucial aspect of our model is our distinction between *model networks* and *reconstructible networks*: The former represent what really happened (at the level of simplification of the model, of

1545-5963/04/$20.00 © 2004 IEEE Published by the IEEE CS, NN, and EMB Societies & the ACM

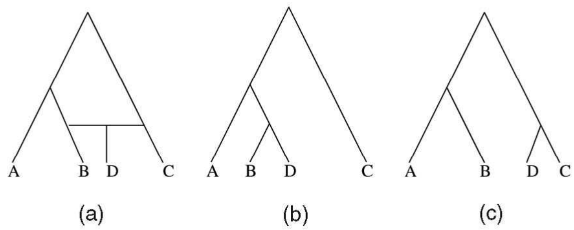

Fig. 1. Hybrid speciation: the network and its two induced trees.

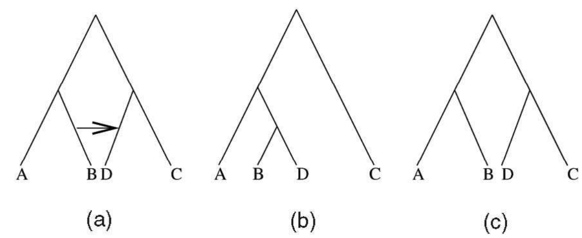

Fig. 2. Horizontal transfer: the network and its two induced trees.

course), while the latter represent what can be inferred from data on current organisms. Using the properties of model and reconstructible networks, we extend our original measure [16] to obtain a true metric, thereby showing that the combination of these DAG properties and our distance measure provides a sound theoretical as well as practical basis for the analysis of phylogenetic networks and for the assessment of network reconstruction methods.

The rest of the paper is organized as follows: In Section 2, we briefly review phylogenetic trees, bipartitions, and the Robinson-Foulds error measure. In Section 3, we define model phylogenetic networks as DAGs that obey certain properties. In Section 4, we discuss the ramifications of missing taxa and other problems arising with biological datasets on the identifiability of reticulation events, and distinguish between model networks and reconstructible ones. In Section 5, we briefly review the measure we introduced in [16] and prove that our measure is a metric on the space of phylogenetic networks. In Section 6, we discuss future work needed to bring network reconstruction on a par with tree reconstruction.

2 Phylogenetic Trees and Bipartitions

A *phylogenetic tree* is a leaf-labeled tree that models the evolution of a set of a taxa (species, genes, languages, placed at the leaves) from their most recent common ancestor (placed at the root). The internal nodes of the tree correspond to the speciation events. Many algorithms have been designed for the inference of phylogenetic trees, mainly from biomolecular (DNA, RNA, or amino-acid) sequences [28]. Similarly, several measures have been used to assess the accuracy of tree reconstruction; the most commonly used measure is the *Robinson-Foulds (RF)* metric [24], which we now define.

Every edge e in an unrooted leaf-labeled tree T defines a bipartition π_e on the leaves (induced by the deletion of e), so that we can define the set $C(T) = \{\pi_e : e \in E(T)\}$, where $E(T)$ is the set of all internal edges of T. If T is a model tree and T' is the tree inferred by a phylogenetic reconstruction method, we define the *false positives* to be the edges of the set $C(T') - C(T)$ and the *false negatives* to be those of the set $C(T) - C(T')$.

- The *false positive rate* (FP) is $|C(T') - C(T)|/(n-3)$.
- The *false negative rate* (FN) is $|C(T) - C(T')|/(n-3)$.

(When both trees are binary, we have $FP = FN$.) Since an unrooted binary tree on n leaves has $n-3$ internal edges, the false positive and false negative rates are values in the range $[0,1]$. The RF distance between T and T' is simply the average of these two rates, $(FN + FP)/2$. Measures, such as the RF distance, that quantify the distance between two trees in terms of their edge structure, are often called measures of *topological*[1] *accuracy*.

Theorem 1 [24]. *The pair $(\mathcal{T}, m)$, where $\mathcal{T}$ is the space of phylogenetic trees on n leaves and m is the RF distance, is a metric space.*

A *metric space* is a set of objects with an equivalence relation, $\equiv$, and a binary distance function, d, which together satisfy the following three conditions, for every three objects x, y, and z:

- positivity: $d(x,y) \geq 0 \wedge [d(x,y) = 0 \Leftrightarrow x \equiv y]$,
- symmetry: $d(x,y) = d(y,x)$, and
- triangle inequality: $d(x,y) + d(y,z) \geq d(x,z)$.

3 Model Phylogenetic Networks

3.1 Nontree Evolutionary Events

We consider two types of evolutionary events that give rise to network (as opposed to tree) topologies: *hybrid speciation* and *horizontal gene transfer* (also called *lateral* gene transfer).

In hybrid speciation, two lineages recombine to create a new species. We can distinguish *diploid hybridization*, in which the new species inherits one of the two homologs for each chromosome from each of its two parents—so that the new species has the same number of chromosomes as its parents, and *polyploid hybridization*, in which the new species inherits the two homologs of each chromosome from both parents—so that the new species has the sum of the numbers of chromosomes of its parents. Under this last heading, we can further distinguish *allopolyploidization*, in which two lineages hybridize to create a new species whose ploidy level is the sum of the ploidy levels of its two parents (the expected result), and *autopolyploidization*, a regular speciation event that does not involve hybridization, but which doubles the ploidy level of the newly created lineage. Prior to hybridization, each site on each homolog has evolved in a tree-like fashion, although, due to meiotic recombination, different strings of sites may have different histories. Thus, each site in the homologs of the parents of the hybrid evolved in a tree-like fashion on one of the trees induced by (contained inside) the network representing the hybridization event, as illustrated in Fig. 1.

In horizontal gene transfer, genetic material is transferred from one lineage to another without resulting in the production of a new lineage. In an evolutionary scenario involving horizontal transfer, certain sites are inherited through horizontal transfer from another species, while all others are inherited from the parent, as symbolized in Fig. 2.

When the evolutionary history of a set of taxa involves processes such as hybrid speciation or horizontal gene

1. The word "topological" is not used here in a mathematical sense, but only to signify that the measure of accuracy relates only to the structure of the tree, not to any other parameters.

transfer, trees can no longer represent the evolutionary relationship; instead, we turn to rooted directed acyclic graphs (rooted DAGs).

3.2 Notation

Given a (directed) graph G, let $E(G)$ denote the set of (directed) edges of G and $V(G)$ denote the set of nodes of G. Let (u, v) denote a directed edge from node u to node v; u is the *tail* and v the *head* of the edge and u is a *parent* of v.[2] The *indegree* of a node v is the number of edges whose head is v, while the *outdegree* of v is the number of edges whose tail is v. A node of indegree 0 is a *leaf* (often called a *tip* by systematists). A directed path of length k from u to v in G is a sequence $u_0u_1 \cdots u_k$ of nodes with $u = u_0$, $v = u_k$, and $\forall i, 1 \leq i \leq k, (u_{i-1}, u_i) \in E(G)$; we say that u is the tail of p and v is the head of p. Node v is *reachable* from u in G, denoted $u \leadsto v$, if there is a directed path in G from u to v; we then also say that u is an *ancestor* of v. A *cycle* in a graph is a directed path from a vertex back to itself; trees never contain cycles: in a tree, there is always a unique path between two distinct vertices. Directed acyclic graphs (or DAGs) play an important role on our model; note that every DAG contains at least one vertex of indegree 0. A *rooted directed acyclic graph*, in the context of this paper, is then a DAG with a single node of indegree 0, the *root*; note that all all other nodes are reachable from the root by a (directed) path of edges.

3.3 Phylogenetic Networks

Strimmer et al. [26] proposed DAGs as a model for describing the evolutionary history of a set of sequences under recombination events. They also described a set of properties that a DAG must possess in order to provide a realistic model of recombination. Later, Strimmer et al. [27] proposed adopting *ancestral recombination graphs* (ARGs), due to Hudson [11] and Griffiths and Marjoram [5], as a more appropriate model of phylogenetic networks. ARGs are rooted graphs that provide a way to represent linked collections of trees (assuming that the trees are ultrametric or nearly so) by a single network. Hallett and Lagergren [6] described a similar set of conditions on rooted DAGs to use them as models for evolution under horizontal transfer events. (Another network-like model is pedigrees, designed to represent the parentage of individual organisms that propagate through sexual reproduction—so that the indegree of each internal node of a pedigree is either 0 or 2 [18].)

Like Strimmer et al. [26], we use DAGs to describe the topology of our phylogenetic networks and, like Hallett and Lagergren [6], we add a set of (mostly simpler) conditions to ensure that the resulting DAGs reflect the properties of hybrid speciation. Unlike these authors, however, we carefully distinguish between *model networks* (the representation of the actual evolutionary scenario) and *reconstructible networks*: The latter are more loosely structured than the former, reflecting the lack of data (due to extinction or to sampling difficulties) that often characterizes biological data.

3.4 Model Networks

A phylogenetic network $N = (V, E)$ is a rooted DAG obeying certain constraints. We begin with a few definitions.

Definition 1. *A node $v \in V$ is a* tree node *if and only if one of these three conditions holds:*

- *$indegree(v) = 0$ and $outdegree(v) = 2$: root,*
- *$indegree(v) = 1$ and $outdegree(v) = 0$: leaf, or*
- *$indegree(v) = 1$ and $outdegree(v) = 2$: internal tree node.*

A node v is a network node *if and only if we have $indegree(v) = 2$ and $outdegree(v) = 1$.*

2. We shall use "parent" in this graph-theoretical sense, unless otherwise noted.

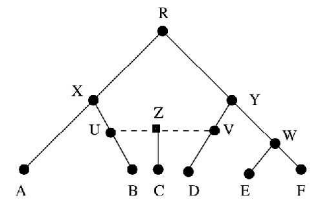

Fig. 3. A phylogenetic network N on six species. Disks denote the tree nodes and squares denote the network nodes, while solid lines denote the tree edges and dashed lines denote the network edges.

Tree nodes correspond to regular speciation or extinction events, whereas network nodes correspond to reticulation events (such as hybrid speciation and horizontal gene transfer). We clearly have $V_T \cap V_N = \emptyset$ and can easily verify that we have $V_T \cup V_N = V$. We extend the node categories to corresponding edge categories as follows:

Definition 2. *An edge $e = (u, v) \in E$ is a* tree edge *if and only if v is a tree node; it is a* network edge *if and only if v is a network node.*

The tree edges are directed from the root of the network toward the leaves and the network edges are directed from their tree-node endpoint towards their network-node endpoint. Fig. 3 shows an example of a network in which the species at node Z is the product of (homoploid or allopolyploid) hybrid speciation. In such a network, a species appears as a directed path p that does not contain any network edge (since a network edge connects two existing species or an existing species and a newly created one). If p_1 and p_2 are two directed paths that define two distinct species, then p_1 and p_2 must be edge-disjoint, that is, the two paths cannot share edges. For example, the directed path p from node X to node B in Fig. 3 could define species B (as could the path from R to B), whereas the directed path from node X to node C does not define a species, since it contains a network edge.

A phylogenetic network $N = (V, E)$ defines a partial order on the set V of nodes. We can also assign times to the nodes of N, associating time $t(u)$ with node u; such an assignment, however, must be consistent with the partial order. Call a directed path p from node u to node v that contains at least one tree edge a *positive-time directed path*. If there exists a positive-time directed path from u to v, then we must have $t(u) < t(v)$. Moreover, if $e = (u, v)$ is a network edge, then we must have $t(u) = t(v)$, because a reticulation event is effectively instantaneous at the scale of evolution; thus reticulation events act as synchronization points between lineages.

The combination of a time flow along tree edges and synchronization along network edges enables us to compare times across branches (something that cannot be done with trees). Positive-time directed paths, however, do not capture all temporal constraints imposed by reticulation events: Fig. 4 illustrates this point. In that figure, we have $t(Y) = t_1$ and $t(X) = t_4$; reticulation events H_1 and H_2 occur at times t_2 and t_3, respectively. The two reticulation events

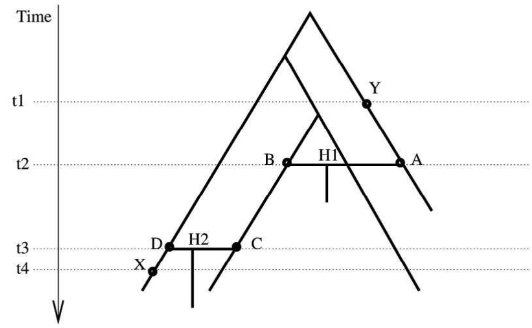

Fig. 4. A scenario illustrating two nodes X and Y that cannot coexist in time.

clearly imply $t_1 < t_2 < t_3 < t_4$ which, in turn, implies that X and Y cannot coexist in time and, hence, cannot be the "parents" of a reticulation event, yet there does not exist a positive-time directed path from Y to X. However, there does exist a *sequence* $P = \langle p_1, p_2, p_3 \rangle$ of positive-time directed paths, where p_1 is the directed path from Y to A, p_2 is the directed path from B to C, and p_3 is the directed path from D to X. We generalize this observation as follows:

Definition 3. *Given a network N, two nodes u and v cannot* coexist *(in time) if there exists a sequence $P = \langle p_1, p_2, \ldots, p_k \rangle$ of paths such that:*

- *p_i is a positive-time directed path, for every $1 \leq i \leq k$,*
- *u is the tail of p_1, and v is the head of p_k, and*
- *for every $1 \leq i \leq k-1$, there exists a network node whose two parents are the head of p_i and the tail of p_{i+1}.*

Obviously, if two nodes x and y cannot coexist in time, then a reticulation event between them cannot occur.

We are finally ready to define a model phylogenetic network.

Definition 4. *A* model phylogenetic network *is a rooted DAG obeying the following constraints:*

1. *Every node has indegree and outdegree defined by one of the four combinations $(0, 2)$, $(1, 0)$, $(1, 2)$, or $(2, 1)$—corresponding to, respectively, root, leaves, internal tree nodes, and network nodes.*
2. *If two nodes u and v cannot coexist in time, then there does not exist a network node w with edges (u, w) and (v, w).*
3. *Given any edge of the network, at least one of its endpoints must be a tree node.*

4 Reconstructible Phylogenetic Networks

Our definition of model phylogenetic networks assumes that complete information about every step in the evolutionary history is available. Such is the case in simulations and in artificial phylogenies evolved in a laboratory setting—hence, our use of the term *model*. When attempting to reconstruct a phylogenetic network from sample data, however, a researcher will normally have only incomplete information, due to a combination of extinctions, incomplete sampling, and abnormal model conditions. Extinctions and incomplete sampling have the same consequences: The data do not reflect all of the various lineages that contributed to the current situation. Abnormal conditions include insufficient differentiation along edges, in which case some of the edges may not be reconstructible, leading to polytomies and, thus, to nodes of outdegree larger than 2. All three types of problems may lead to the reconstruction of networks that violate the constraints of Definition 4. (The distinction between a model phylogeny and a reconstructible phylogeny is common with trees as well: For instance, model trees are always rooted, whereas reconstructed trees are usually unrooted. In networks, both the model network and the reconstructed network must be rooted: Reticulations only make sense with directed edges.)

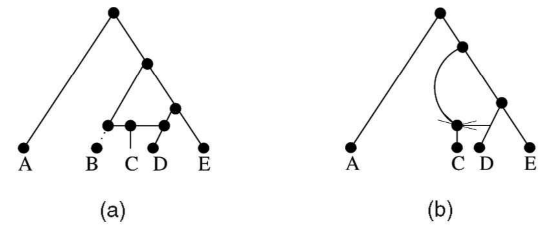

Fig. 5. A small phylogenetic network illustrating the distinction between model networks and reconstructible networks. (a) The model network and (b) its best reconstruction.

We illustrate the problem in Fig. 5. Fig. 5a shows a network on a set of five species, where species B became extinct (or was not sampled). Fig. 5b shows the reconstruction of the same network; note that this network violates time coexistence and also contains a pair of edges, one of which is a tree edge and the other a network edge, both incident into a network node.

Clearly, then, a reconstructible network will require changes from the definition of a model network. We must relax the degree constraints to allow arbitrary outdegrees for both network nodes and internal tree nodes. In addition, we need to reconsider the time coexistence property in some detail.

4.1 Identifiability of Reticulation Events

Fig. 5 illustrates a scenario in which identifying a reticulation event in a network is not possible based solely on topological considerations. Based on *separate analyses* of the data, the two trees in Fig. 6 are the gene trees consistent with the data. Those two gene trees may in turn be *reconciled* into two different species networks as shown in Fig. 7, each of which indicates a different reticulation event: In Fig. 7a, the reticulation produces lineage C, but in Fig. 7b, it produces lineage D.

Nevertheless, identifiability can be preserved if sufficient data are available to obtain an accurate estimate of the actual

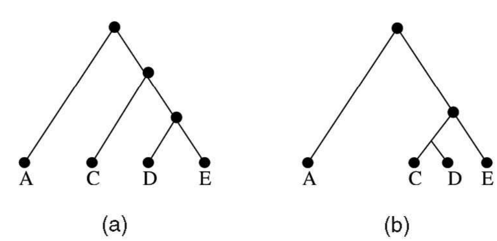

Fig. 6. The two gene trees consistent with the network of Fig. 5a. (a) First gene tree and (b) second gene tree.

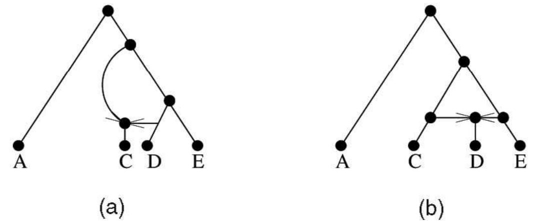

Fig. 7. The two possible reconciliations of the gene trees of Fig. 6. (a) First reconciliation and (b) second reconciliation.

number of evolutionary changes on the edges of the gene trees. Using again the example of Fig. 5a, we illustrate the process in Fig. 8. In Fig. 8a, we reproduce the network of Fig. 5a, but with edge lengths indicated; in Figs. 8b and 8c, we show the two gene trees of Fig. 6, again with edge lengths. We can then verify that it is possible to associate edge lengths to the network of Fig. 7a so as to be consistent with the two gene trees, as shown in Fig. 8d, but that the same cannot be achieved with the network of Fig. 7b. (Recall that true network edges have a length of zero: It is then a simple matter to write a system of linear equations to describe the edge lengths of the network in terms of those of the gene trees, and verify that the system obtained from the network in Fig. 7b does not admit a solution.) The network of Fig. 8d captures the same evolutionary history as that depicted in the original network of Fig. 5a, after eliminating species B and removing the resulting internal node of indegree and outdegree 1. In contrast, the network of Fig. 7b, eliminated because we cannot consistently assign it edge lengths, captures an evolutionary history different from that of the true phylogeny. Another piece of information that may enable us to identify events in such situations is the level of ploidy: If our modern species show different levels of ploidy, then we know that lineages with high levels of ploidy have gone through additional polyploidization events.

4.2 The Effect of Missing Taxa

As stated in Section 3.4, a reticulation event takes place between two lineages that coexist in time, requiring a phylogenetic network to satisfy the time coexistence property. However, as described above, missing taxa from

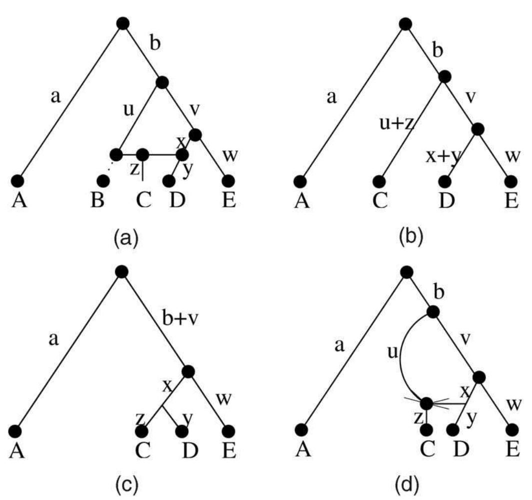

Fig. 8. Identifying reticulations for the network of Fig. 5a by using edge lengths. (a) Model network, (b) first gene tree, (c) second gene tree, and (d) reconciliation using lengths.

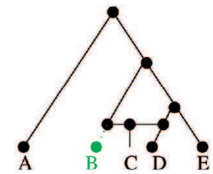

Fig. 9. Augmenting the network of Fig. 5b with nodes of indegree and outdegree 1.

a phylogenetic analysis may lead to a phylogenetic network that violates time coexistence. Fig. 9 illustrates (using the network of Fig. 5b) how to augment such a phylogenetic network to remedy that violation, by introducing nodes of indegree and outdegree 1. Such introduced nodes imply in most cases that a taxon (or a clade) was missing.

If another taxon from Fig. 5a goes missing, say taxon D, we obtain the situation depicted in Fig. 10. Fig. 10a shows the model network, with the two extinct taxa. Now, the two gene trees have identical topologies, but different edge lengths, as shown in Figs. 10b and 10c. Thus, in absence of edge lengths, we would simply return the tree of Fig. 10d, failing to detect a reticulation event, but with edge lengths we could return the network of Fig. 10e—although that network clearly violates time coexistence. Finally, Fig. 11 shows three possible ways of augmenting the network to satisfy time coexistence, each network "detecting" one or two of the possible missing taxa. Of these, only the networks of Figs. 11a and 11b also satisfy the constraint on edges in networks, and only that of Fig. 11a can be given a consistent assignment of edge lengths.

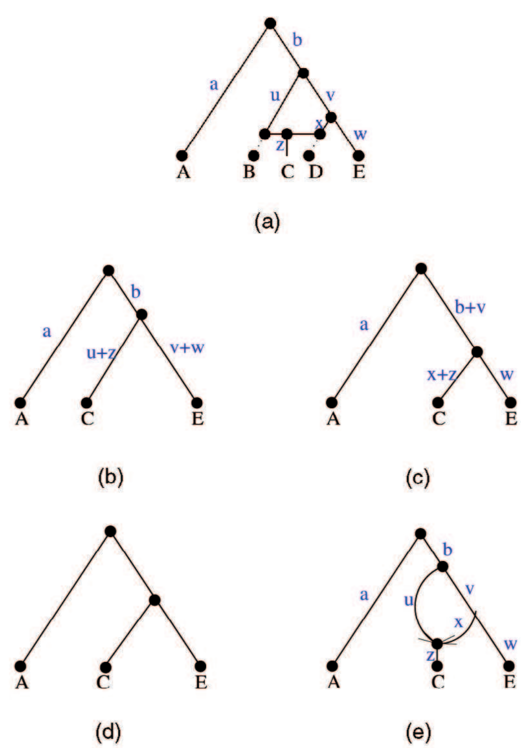

Fig. 10. A model network with (a) two extinct taxa, (b) and (c) its two induced gene trees, and (d) and (e) their tree and network reconciliations. Edges are labeled with their length. (a) Model network, (b) first gene tree, (c) second gene tree, (d) reconciliation, no lengths, and (e) reconciliation, lengths.

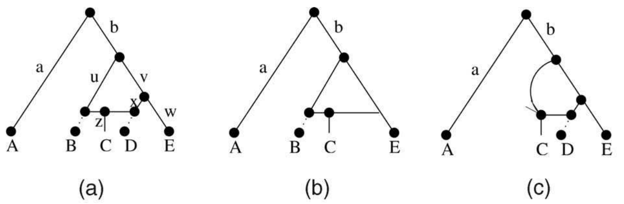

Fig. 11. Three possible augmentations of the network of Fig. 10e. (a) Lengths, (b) no lengths, and (c) violates coexistence.

Finally, successive hybrid speciation events with conjoined extinctions can yield situations where the order of hybrid speciation events cannot be inferred. Consider Fig. 12. Fig. 12a shows the model network; on all paths to the leaf labeled X, we have three consecutive reticulation events, but all parent lineages are lost, with only the final hybrid surviving. In that case, the manner in which the four parent lineages are all hybridized into the final lineage X is no longer reconstructible: Fig. 12b shows a network in which the hybrid speciation events take place in another order, but, absent the extinct lineages, cannot be distinguished from the network of Fig. 12a. In such a case, we have the same problem as with polytomies and should, in fact, represent the situation with the equivalent, for hybridizations, of a polytomy, namely, a network node of indegree higher than 2, as shown in Fig. 12c, denoting the fact that eight lineages are getting hybridized into one, but in some unknown and unreconstructible manner.

4.3 A Definition of Reconstructible Networks

We have seen two types of problems in reconstructing phylogenetic networks. First, slow evolution may give rise to edges so short that they cannot be reconstructed, leading to polytomies. This problem cannot be resolved within the DAG framework, so we must relax the constraints on the outdegree of tree nodes. Second, missing data may lead methods to reconstruct networks that violate indegree constraints or time coexistence. In such cases, we can postprocess the reconstructed network to restore compliance with most of the constraints in the three simple steps of Fig. 13. The resulting network is consistent with the original reconstruction, but now satisfies the outdegree requirement for network nodes, obeys time coexistence (the introduction of tree edges on the paths to the network node allow arbitrary time delays), and no longer violates the requirement that at least one endpoint of each edge be a tree node. Moreover, this postprocessing is unique and quite simple.

1) For each network node w with outdegree larger than 1, say with edges (w, v_1), ..., (w, v_k), add a new tree node u with edge (w, u) and, for each i, $1 \leq i \leq k$, replace edge (w, v_i) by edge (u, v_i).
2) For each network node w whose parents u and v violate time coexistence, add two tree nodes w_u and w_v and replace the two network edges (u, w) and (v, w) by four edges: the two tree edges (u, w_u) and (v, w_v) and the two network edges (w_u, w) and (w_v, w).
3) For each edge (u, v) where both u and v are network nodes, add a new tree node w and replace the edge (u, v) by the two edges (u, w) and (w, v).

Fig. 13. Three simple steps to restore compliance with most network constraints.

We can thus define a reconstructible network in terms similar to a model network.

Definition 5. *A* reconstructible phylogenetic network *is a rooted DAG obeying the following constraints:*

1. *Every node has indegree and outdegree defined by one of the three (indegree, outdegree) combinations $(0,x)$, $(1,y)$, or $(z,1)$, for $x \geq 1$, $y \geq 0$, and $z \geq 2$—corresponding to, respectively, root, other tree nodes (internal nodes and leaves), and network nodes.*
2. *If two nodes u and v cannot coexist in time, then there does not exist a network node w with edges (u,w) and (v,w).*
3. *Given any edge of the network, at least one of its endpoints must be a tree node.*

4.4 Distinguishability

As is common in phylogenetic tree reconstruction, algorithms for reconstructing phylogenetic networks could produce several different networks. In devising a measure to assess the quality of reconstruction, we must start by recognizing when two such networks are equally correct reconstructions from the data.

Before we define distinguishability, we first define the notion of isomorphism between two phylogenetic networks.

Definition 6. *Let $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ be two phylogenetic networks leaf-labeled by the same set L of taxa. We say that N_1 and N_2 are* isomorphic, *denoted $N_1 = N_2$, if there is a bijection $f : V_1 \rightarrow V_2$ such that the following two conditions hold:*

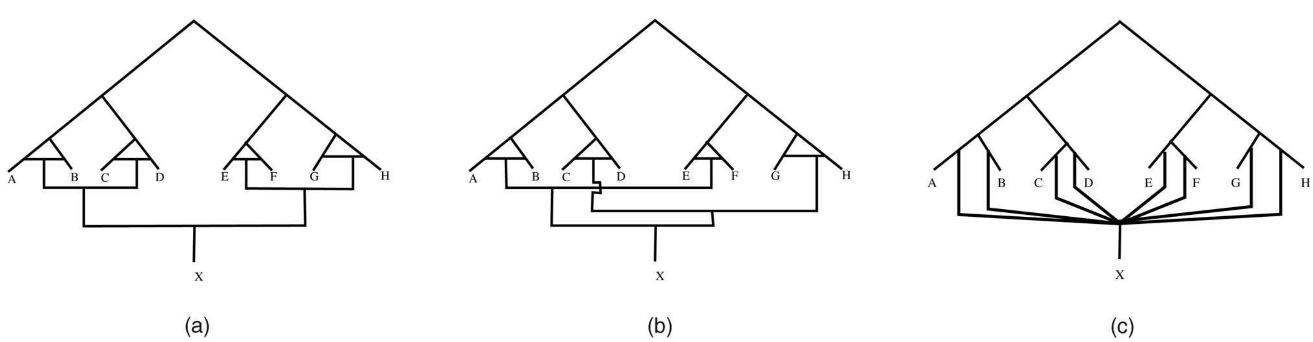

Fig. 12. Successive hybrid specification events cannot be ordered in the presence of matching extinctions. (a) First network, (b) second network, and (c) after collapsing incoming paths, from either (a) or (b).

0. Find all maximal subtrees of leaves (that include no network node—those are known as *clades* in phylogenetic parlance); for each such subtree t rooted at r_t, create a new node h_t and an edge (h_t, r_t). Let L be the set composed of all such h_t nodes—one can think of "condensing" the maximal subtrees and using h_t as a symbolic leaf representing the entire subtree attached to it.

Repeat the following two steps until no change occurs:

1. For each maximal convergent set U with leaf set $L_U \subseteq L$, remove all vertices and edges on the paths from a vertex in U to a leaf in L_U (including all vertices of U, but keeping all vertices of L_U). For any vertex x that is the tail of an edge with its head in the deleted set, say edge (x, v), replace it by the set $\{(x, l) \mid l \in L_U$ is reachable from $v\}$.
2. For each network node w of in- and outdegree 1, say with edges (u, w) and (w, v), replace these two edges with single edge (u, v) and remove any duplicate edges. Repeat until no such vertex can be found.

Fig. 14. A three-step procedure to simplify reconstructible networks.

1. $(u, v) \in E_1$ *if and only if* $(f(u), f(v)) \in E_2$.
2. *If leaf* $v \in V_1$ *is labeled by* $l \in L$, *then leaf* $f(v) \in V_2$ *is also labeled by* l.

Basically, two networks are isomorphic if they are identical up to relabeling of internal nodes. This definition is appropriate for comparing two model networks.

Definition 7. *Two model phylogenetic networks* $N_1 = (V_1, E_1)$ *and* $N_2 = (V_2, E_2)$ *are* indistinguishable *whenever they are isomorphic.*

However, this definition is much too strict for reconstructible phylogenetic networks. In that case, we simplify the networks as much as possible (at the risk of violating time coexistence), so as to reduce the problem back to one of isomorphism. The simplification procedure is in part the reverse of that described in Fig. 13. We collapse all consecutive hybridizations into single hybridization nodes with high indegrees, in the style of Fig. 12c, and remove all vertices of indegree and outdegree 1.

We have seen an example in Fig. 12 of two distinct networks that both match the available data; this is but one example of a large class of network structures that cannot be distinguished from the available data. This class is characterized by the presence of convergent paths from a collection of internal nodes to a collection of leaves.

Definition 8. *A subset* U *of internal nodes in a network is* convergent *if and only if it satisfies the following properties:*

1. $|U| \geq 2$ *and*
2. *every leaf reachable from some node in* U *is reachable from all nodes in* U.

A convergent subset is maximal *if it cannot be augmented by another vertex without losing the convergence property.*

A few observations are in order. First, note that this definition generalizes the example of Fig. 12 in that it allows similar results without requiring that hybridization events

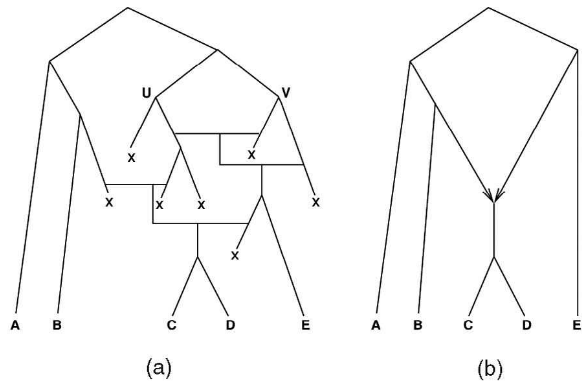

Fig. 15. A network with many extinctions (each identified with an X) and hybridizations and a reduced equivalent; set $\{U, V\}$ is a convergent set with leaf set $\{C, D, E\}$. (a) The network and (b) a reduced equivalent.

occur consecutively—they could be separated by tree nodes, as long as all lineages are made to reconverge to the collection of leaves. Second, note that, in the absence of any polyploidization event, paths with different numbers of hybridization nodes on the way from an internal node in the convergent set to a leaf are not reconstructible: We have no way to tell how much of each parent genome appears in the hybrid and so can reconstruct any scenario; and if any polyploidization events occurred, then the leaves they lead to will be different from the leaves at the end of paths that contain no such event, since they will have a different number of chromosomes. Finally, note that, if a network N contains a convergent set U reaching leaf set L, then we can derive equivalent (that is, indistinguishable from N) networks by any operations that alter the paths from vertices in U to leaves in L while preserving U as a convergent set. Convergent sets thus "hide" the hybridization scenario that takes place between them and the leaf set they reach.

These observations suggest that we simply replace a convergent set and its paths to the leaves with a complete bipartite graph directed from the parent nodes of the convergent set to the set of leaves reached from the convergent set; if any other path reaches some of these leaves (because it joins a path from the convergent set to the leaf set), then we simply replace the edge from the tree node outside the convergent set to the path from the convergent set with a collection of edges directly from that node to the leaves it can reach (another polytomy). The complete procedure is formalized in Fig. 14, while Fig. 15 shows a reconstructible network (with convergent set $\{U, V\}$) and its simplified form.

If we start with some reconstructible network, N, we denote by $R(N)$ the result of the steps of Fig. 14—it is a *reduced* version of N. Note that every node of $R(N)$ corresponds to a node of N, although its indegree and outdegree may have changed.

Definition 9. *Two reconstructible phylogenetic networks* $N_1 = (V_1, E_1)$ *and* $N_2 = (V_2, E_2)$ *are* indistinguishable *if they have isomorphic reduced versions.*

The notion of indistinguishability is closely tied to that of trees.

Theorem 2. *If two networks induce the same set of trees (up to vertices of indegree and outdegree 1), they are indistinguishable.*

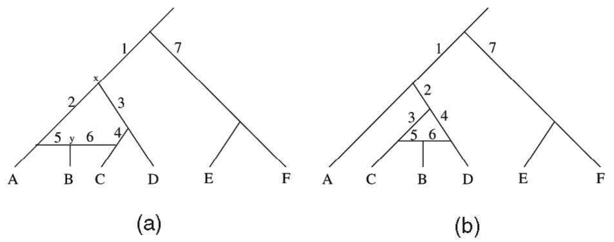

Fig. 16. The two networks used in the illustration of the tripartition-based distance measure.

Proof. First note that, if N_1 and N_2 induce the same set of trees, then so do $R(N_1)$ and $R(N_2)$. We show that the reduced versions must be isomorphic by simple construction. We already have a bijection between leaves and we can trivially match the root r_1 of $R(N_1)$ with the root r_2 of $R(N_2)$. Now, the outdegree of each root is the same in each network (or else one network induces at least one tree not present in the other). Let the children (both tree and network nodes) of r_1 be denoted $v_1, w_1, \ldots, z_1$, and let $L_v, L_w, \ldots, L_z$ denote the subsets of leaves reachable from these children. Because $R(N_1)$ is reduced, it no longer contains convergent sets and so $L_v, L_w, \ldots, L_z$ are all distinct. But, then, each leaf subset has a unique equivalent in $R(N_2)$, so that we can map each child of r_1 to the proper child of r_2. The process is then repeated recursively down the network until the mapping is fully defined. Note that, when edges from two subtree roots point to the same internal node, they define the same leaf subset, so that the mapping is well-defined. The mapping preserves both tree and network edges by construction and defines an isomorphism. □

5 A Network Error Metric

5.1 Desired Properties of Measures

In earlier work [16], we introduced a measure of distance between two phylogenetic networks and presented experimental results showing that our measure behaved in a qualitatively satisfying manner. We briefly review this measure in the framework of phylogenetic networks that we have established here; we assume that our phylogenetic networks are reconstructible networks in which all nodes with indegree and outdegree 1 have been removed, with their two edges merged. (We remove such vertices for simplicity of exposition: All our results hold even when such nodes exist in either network, but the definitions become more elaborate and the arguments start requiring detailed case analyses.)

We want a measure $m(N_1, N_2)$ of the distance between two networks N_1 and N_2, such that m is symmetric and nonnegative, and satisfies the following three conditions:

1. C1: If N_1 and N_2 are two trees, then we have $m(N_1, N_2) = RF(N_1, N_2)$, where $RF(N_1, N_2)$ denotes the Robinson-Foulds distance between trees.
2. C2: If we have $N_1 = N_2$, then we must have $m(N_1, N_2) = 0$.
3. C3: If we have $m(N_1, N_2) = 0$, then we must have $N_1 = N_2$.

5.2 The Tripartition-Based Measure

The measure we introduced is based on the tripartition that is induced by each edge of the network. Let N be a phylogenetic network, leaf-labeled by set S, and let $e = (u, v)$ be an edge of N. Edge e induces a tripartition of S, defined by the sets

- $A(e) = \{s \in S \mid s$ is reachable from the root of N only via $v\}$.
- $B(e) = \{s \in S \mid s$ is reachable from the root of N via at least one path passing through v and one path not passing through $v\}$.
- $C(e) = \{s \in S \mid s$ is not reachable from the root of N via $v\}$.

We denote by $\theta(e)$ the tripartition of S induced by edge e. Two tripartitions, $\theta(e_1)$ and $\theta(e_2)$, are *equivalent*, denoted by $\theta(e_1) \equiv \theta(e_2)$, whenever we have $A(e_1) = A(e_2)$, $B(e_1) = B(e_2)$, and $C(e_1) = C(e_2)$. (Note that we used a weighting scheme in the original formulation [16], but dropped it here.)

In [16], we defined two edges e_1 and e_2 to be *compatible*, denoted by $e_1 \equiv e_2$, whenever we had $\theta(e_1) \equiv \theta(e_2)$. Then, we defined the *false negative rate* (FN) and *false positive rate* (FP) between two networks N_1 and N_2 as follows:

$$FN(N_1, N_2) = \frac{|\{e_1 \in E(N_1) \mid \nexists e_2 \in E(N_2),\ e_1 \equiv e_2\}|}{|E(N_1)|}$$

$$FP(N_1, N_2) = \frac{|\{e_2 \in E(N_2) \mid \nexists e_1 \in E(N_1),\ e_1 \equiv e_2\}|}{|E(N_2)|}.$$

Definition 10. *The error rate between N_1 and N_2, denoted $m^{tri}(N_1, N_2)$, is $(FN(N_1, N_2) + FP(N_1, N_2))/2$.*

While satisfying conditions C1 and C2 for all networks, this measure fails to satisfy condition C3 for certain networks. For example, although the two networks N_1 and N_2 in Fig. 12 depict two different evolutionary histories, we have $m^{tri}(N_1, N_2) = 0$. The crucial observation to make here, however, is that these two networks are indistinguishable! Thus, we rewrite conditions C2 and C3 to read

1. C2: If N_1 and N_2 are indistinguishable, then we have $m(N_1, N_2) = 0$.
2. C3: If we have $m(N_1, N_2) = 0$, then N_1 and N_2 are indistinguishable.

With this correction, the measure meets condition C3 on the networks of Fig. 12.

This error measure can be computed in time polynomial in the size (number of nodes) of the two networks. We implemented this measure and reported preliminary experimental results in [16]. We illustrate the tripartition-based measure between the two networks N_1 and N_2 of Fig. 16. Table 1 lists the nontrivial tripartitions (i.e., tripartitions associated with internal edges) of N_1 and N_2. Based on the tripartitions induced by the edges of those networks, there are two false negative edges (edges 2 and 3 in network N_1 do not occur in N_2) and two false positive edges (edges 2 and 4 in network N_2 do not occur in N_1). Therefore, we have $FN(N_1, N_2) = FP(N_1, N_2) = \frac{2}{7}$ and $m^{tri}(N_1, N_2) = \frac{2}{7}$.

5.3 The Tripartition Measure is a Metric

The measure $m^{tri}(\cdot, \cdot)$ clearly satisfies condition C1 for all networks. In this section, we prove that the measure also

TABLE 1
The Tripartitions Induced by the Edges of Networks N_1 and N_2 of Fig. 16

Network N_1		Network N_2	
Edge	Tripartition	Edge	Tripartition
1	$\langle\{A,B,C,D\},\ \emptyset,\ \{E,F\}\rangle$	1	$\langle\{A,B,C,D\},\ \emptyset,\ \{E,F\}\rangle$
2	$\langle\{A\},\ \{B\},\ \{C,D,E,F\}\rangle$	2	$\langle\{B,C,D\},\ \emptyset,\ \{A,E,F\}\rangle$
3	$\langle\{C,D\},\ \{B\},\ \{A,E,F\}\rangle$	3	$\langle\{C\},\ \{B\},\ \{A,D,E,F\}\rangle$
4	$\langle\{C\},\ \{B\},\ \{A,D,E,F\}\rangle$	4	$\langle\{D\},\ \{B\},\ \{A,C,E,F\}\rangle$
5	$\langle\emptyset,\ \{B\},\ \{A,C,E,F\}\rangle$	5	$\langle\emptyset,\ \{B\},\ \{A,C,E,F\}\rangle$
6	$\langle\emptyset,\ \{B\},\ \{A,C,E,F\}\rangle$	6	$\langle\emptyset,\ \{B\},\ \{A,C,E,F\}\rangle$
7	$\langle\{E,F\},\ \emptyset,\ \{A,B,C,D\}\rangle$	7	$\langle\{E,F\},\ \emptyset,\ \{A,B,C,D\}\rangle$

satisfies conditions C2 and C3 and prove that it is a metric on the space of reduced reconstructible phylogenetic networks.

Theorem 3. *The tripartition measure is a metric on the space of reduced reconstructible phylogenetic networks on n leaves.*

Before proving that theorem, however, we need a series of results, some of independent interest. Since the tripartition induced by an edge is defined by the head of that edge, we have the following observation.

Observation 1. *Let $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ be two edges of a phylogenetic network. If we have $v_1 = v_2$, then we have $\theta(e_1) \equiv \theta(e_2)$.*

Because the tripartitions are induced by paths to leaves, sets of vertices that share the same sets of paths to leaves yield pairs of equivalent tripartitions.

Observation 2. *Let U be a convergent set of nodes in a network N. Then, for every pair (e_1, e_2) of edges in $\mathcal{E} = \{(v,u) \mid u \in U\}$, we have $\theta(e_1) \equiv \theta(e_2)$.*

The observation simply stems from the fact that every node $u \in U$ reaches exactly the same set of leaves.

We now prove a theorem that classifies all equivalent edges in a network.

Theorem 4. *Let N be a (not necessarily reduced) reconstructible phylogenetic network and let $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ be two edges of N. We have $\theta(e_1) \equiv \theta(e_2)$ if and only if we have $v_1 = v_2$ or $\{v_1, v_2\}$ forms a convergent set.*

Proof. The "if" part follows immediately from Observations 1 and 2. For the "only if" part, assume $\theta(e_1) \equiv \theta(e_2)$. The equivalence can occur simply because we have $v_1 = v_2$, in which case we are done. Otherwise, since the two edges are equivalent, we have $A(e_1) = A(e_2)$ and $B(e_1) = B(e_2)$, so that we can write $A(e_1) \cup B(e_1) = A(e_2) \cup B(e_2)$, showing that the subset of leaves reachable from the head of edge e_1 is the same as that reachable from the head of edge e_2; hence, $\{v_1, v_2\}$ is a convergent set. □

We are in position to prove that $m^{tri}(N_1, N_2)$ satisfies condition C3, and not just on reduced networks, but on any reconstructible networks.

Lemma 1. *Let $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ be two reconstructible phylogenetic networks with $m^{tri}(N_1, N_2) = 0$. Then, N_1 and N_2 are indistinguishable.*

Proof. We demonstrate a bijection between the vertices of $R(N_1)$ and those of $R(N_2)$ that induces an isomorphism between the two reduced networks. Obviously, we match each leaf of $R(N_1)$ to its matching counterpart in $R(N_2)$ and match the root of $R(N_1)$ to that of $R(N_2)$. Let v be some internal node of $R(N_1)$; note that v is also an internal node of N_1 and, if $e = (u, v)$ is an edge of N_1, the definition of $m^{tri}(\cdot,\cdot)$ implies

$$\exists e' = (u', v') \in E(N_2),\ e \equiv e'.$$

We would like to set $f(v) = v'$, but v' may not exist in $R(N_2)$; moreover, we need to show that this assignment would be well defined even if v' was in the reduced network. We begin with the second: Assume there are two vertices v' and v'' in N_2 (and also present in $R(N_2)$) that meet the requirements; then we have edges $e' = (u', v')$ and $e'' = (u'', v'')$ in N_2 with $e \equiv e'$ and $e \equiv e''$. By Observation 2, it follows that v' and v'' are equal—and, thus, our mapping is well defined. (If not equal, then v' and v'' must have been in a convergent set, but then they could not be present in the reduced sets.) If v' does not belong to $R(N_2)$, it was eliminated by the procedure of Fig. 14, which means that v' is a vertex of indegree and outdegree 1 in N_2, in which case we simply map v to the first descendant of v' in N_2 that does belong to $R(N_2)$. (The other vertices eliminated by the procedure are those in a convergent set or a along a path from a convergent set to the leaves; but if v' belongs to a convergent set in N_2, then v must belong to a convergent set in N_1 in order for e and e' to be equivalent, contradicting our choice of v in $R(N_1)$.) The mapping is thus well-defined; since it is built using paths in N_2 that stop at the first vertex that lies in $R(N_2)$, it preserves edges (whether tree edges or network edges) and defines an isomorphism. □

Lemma 2. *If N_1 and N_2 are two indistinguishable networks, we have*

$$m^{tri}(R(N_1), R(N_2)) = 0.$$

Proof. If N_1 and N_2 are two indistinguishable networks, then $R(N_1)$ and $R(N_2)$ are isomorphic. The conclusion follows trivially. □

We note that, if N_1 and N_2 are indistinguishable networks, we could have $m^{tri}(N_1, N_2) \neq 0$, because of topological differences within the sets of paths from a convergent set to the leaves.

We are finally in a position to prove Theorem 3. Recall that in the context of this theorem all phylogenetic networks are reduced reconstructible networks.

Proof (of Theorem 3). We prove the following:

1. $m^{tri}(N_1, N_2) \geq 0$ and $m^{tri}(N_1, N_2) = 0$ iff N_1 and N_2 are indistinguishable.
2. $m^{tri}(N_1, N_2) = m^{tri}(N_2, N_1)$.
3. $m^{tri}(N_1, N_3) \leq m^{tri}(N_1, N_2) + m^{tri}(N_2, N_3)$.

The measure is nonnegative and symmetric by definition; Lemma 1 proves the first part of the "iff" and Lemma 2 the second. Finally, the measure is based on the symmetric difference of edge sets, suitably normalized; since the symmetric difference of sets satisfies the triangle inequality, we can verify that so does m^{tri}, thereby satisfying Property 3.

Denote the set of edges in $E(N_i)$ but not in $E(N_j)$ by $E(i, j)$; we need to verify

$$\frac{|E(1,3)|}{|E(N_1)|} + \frac{|E(3,1)|}{|E(N_3)|} \leq \frac{|E(1,2)|}{|E(N_1)|} + \frac{|E(2,3)|}{|E(N_2)|} + \frac{|E(3,2)|}{|E(N_3)|} + \frac{|E(2,1)|}{|E(N_2)|}.$$

(Of course, this is immediate if we have $|E(N_1)| = |E(N_2)| = |E(N_3)|$.) Consider just one half of this inequality: We need to prove

$$\frac{|E(1,3)|}{|E(N_1)|} \leq \frac{|E(1,2)|}{|E(N_1)|} + \frac{|E(2,3)|}{|E(N_2)|}.$$

If we now consider the full Venn diagram of the three sets of edges $E(N_1)$, $E(N_2)$, and $E(N_3)$, we get seven disjoint subsets; if we rewrite the inequality above in terms of these seven subsets, it is trivial, if slightly tedious, to verify its correctness. □

6 Conclusions and Future Work

In this paper, we have described a mathematical model of phylogenetic networks in terms of DAGs, and distinguished between model networks and reconstructible ones based on the properties they possess. We have also shown that phylogenetic networks, combined with our tripartition-based distance measure, define a metric space. Thus, in addition to the good experimental attributes we observed [16], our framework for the study of phylogenetic networks is provably well founded. We thus now have a framework in which to design, analyze, test, and compare reconstruction algorithms for phylogenetic networks. Current methods for network reconstruction are limited to NeighborNet [4], T-Rex [14], the software of Addario-Berry et al. [1], and SplitsTree [12], none of which is targeted at hybrid speciation and most of which do not reconstruct phylogenetic networks. Thus, our next step will be to develop new reconstruction algorithms based on our findings and test them within our framework. The availability of a testing framework means that approaches for network reconstruction can now be tested, compared, and analyzed in some detail, a prerequisite to further progress in the area.

Acknowledgments

The authors want to thank the anonymous referees for many detailed and helpful comments, as well as for their insistence on clarity: Their reviews led to substantial improvements in the presentation. This work is supported in part by the US National Science Foundation grants DEB 01-20709 (Linder, Moret, and Warnow), IIS 01-13095 (Moret), IIS 01-13654 (Warnow), IIS 01-21377 (Moret), IIS 01-21680 (Linder and Warnow), ANI 02-03584 (Moret), EF 03-31453 (Warnow), EF 03-31654 (Moret), by the David and Lucile Packard Foundation (Warnow), by the Radcliffe Institute for Advanced Study (Warnow), by the Program for Evolutionary Dynamics at Harvard (Warnow), and by the Institute for Cellular and Molecular Biology at the University of Texas at Austin (Warnow).

References

[1] L. Addario-Berry, M.T. Hallett, and J. Lagergren, "Toward Identifying Lateral Gene Transfer Events," *Proc. Eighth Pacific Symp. Biocomputing (PSB '03)*, pp. 279-290, 2003.

[2] H.J. Bandelt and A.W.M. Dress, "Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data," *Molecular Phylogenetics and Evolution*, vol. 1, pp. 242-252, 1992.

[3] H.J. Bandelt, V. Macaulay, and M.B. Richards, "Median Networks: Speedy Construction and Greedy Reduction, One Simulation, and Two Case Studies from Human mtDNA," *Molecular Phylogenetics and Evolution*, vol. 16, pp. 8-28, 2000.

[4] D. Bryant and V. Moulton, "NeighborNet: An Agglomerative Method for the Construction of Planar Phylogenetic Networks," *Proc. Second Int'l Workshop Algorithms in Bioinformatics (WABI '02)*, pp. 375-391, 2002.

[5] R.C. Griffiths and P. Marjoram, "Ancestral Inference from Samples of DNA Sequences with Recombination," *J. Computational Biology*, vol. 3, pp. 479-502, 1996.

[6] M.T. Hallett and J. Lagergren, "New Algorithms for the Duplication-Loss Model," *Proc. Fourth Ann. Int'l Conf. Computational Molecular Biology (RECOMB '00)*, pp. 138-146, 2000.

[7] M.T. Hallett and J. Lagergren, "Efficient Algorithms for Lateral Gene Transfer Problems," *Proc. Fifth Ann. Int'l Conf. Computational Molecular Biology (RECOMB '01)*, pp. 149-156, 2001.

[8] M.T. Hallett, J. Lagergren, and A. Tofigh, "Simultaneous Identification of Duplications and Lateral Transfers," *Proc. Eighth Ann. Int'l Conf. Computational Molecular Biology (RECOMB '04)*, pp. 347-356, 2004.

[9] D.M. Hillis, J.J. Bull, M.E. White, M.R. Badgett, and I.J. Molineux, "Experimental Approaches to Phylogenetic Analysis," *Systematic Biology*, vol. 42, pp. 90-92, 1993.

[10] D.M. Hillis and J.P. Huelsenbeck, "Signal, Noise, and Reliability in Molecular Phylogenetic Analyses," *J. Heredity*, vol. 83, pp. 189-195, 1992.

[11] R.R. Hudson, "Properties of the Neutral Allele Model with Intergenic Recombination," *Theoretical Population Biology*, vol. 23, pp. 183-201, 1983.

[12] D.H. Huson, "SplitsTree: A Program for Analyzing and Visualizing Evolutionary Data," *Bioinformatics*, vol. 14, no. 1, pp. 68-73, 1998.

[13] W.P. Maddison, "Gene Trees in Species Trees," *Systematic Biology*, vol. 46, no. 3, pp. 523-536, 1997.

[14] V. Makarenkov, "T-REX: Reconstructing and Visualizing Phylogenetic Trees and Reticulation Networks," *Bioinformatics*, vol. 17, no. 7, pp. 664-668, 2001.

[15] V. Makarenkov and P. Legendre, "From a Phylogenetic Tree to a Reticulated Network," *J. Computational Biology*, vol. 11, no. 1, pp. 195-212, 2004.

[16] L. Nakhleh, J. Sun, T. Warnow, R. Linder, B.M.E. Moret, and A. Tholse, "Towards the Development of Computational Tools for Evaluating Phylogenetic Network Reconstruction Methods," *Proc. Eighth Pacific Symp. Biocomputing (PSB '03)*, pp. 315-326, 2003.

[17] L. Nakhleh, T. Warnow, and C.R. Linder, "Reconstructing Reticulate Evolution in Species—Theory and Practice," *Proc. Eighth Ann. Int'l Conf. Computational Molecular Biology (RECOMB '04)*, pp. 337-346, 2004.

[18] A. Piccolboni and D. Gusfield, "On the Complexity of Fundamental Computational Problems in Pedigree Analysis," Technical Report CSE-99-8, UC Davis, Dept. of Computer Science, 1999.

[19] D. Posada and K.A. Crandall, "The Effect of Recombination on the Accuracy of Phylogeny Estimation," *J. Molecular Evolution*, vol. 54, no. 3, pp. 396-402, 2002.

[20] D. Posada, K.A. Crandall, and E.C. Holmes, "Recombination in Evolutionary Genomics," *Ann. Rev. Genetics,* vol. 36, pp. 75-97, 2002.
[21] L.H. Rieseberg, "The Role of Hybridization in Evolution—Old Wine in New Skins," *Am. J. Botany,* vol. 82, no. 7, pp. 944-953, 1995.
[22] L.H. Rieseberg, S.J.E. Baird, and K.A. Gardner, "Hybridization, Introgression, and Linkage Evolution," *Plant Molecular Biology,* vol. 42, no. 1, pp. 205-224, 2000.
[23] L.H. Rieseberg and R.D. Noyes, "Genetic Map-Based Studies of Reticulate Evolution in Plants," *Trends in Plant Science,* vol. 3, no. 7, pp. 254-259, 1998.
[24] D.R. Robinson and L.R. Foulds, "Comparison of Phylogenetic Trees," *Math. Biosciences,* vol. 53, pp. 131-147, 1981.
[25] F.J. Rohlf, "Phylogenetic Models and Reticulations," *J. Classification,* vol. 17, no. 2, pp. 185-189, 2000.
[26] K. Strimmer and V. Moulton, "Likelihood Analysis of Phylogenetic Networks Using Directed Graphical Models," *Molecular Biology and Evolution,* vol. 17, pp. 875-881, 2000.
[27] K. Strimmer, C. Wiuf, and V. Moulton, "Recombination Analysis Using Directed Graphical Models," *Molecular Biology and Evolution,* vol. 18, no. 1, pp. 97-99, 2001.
[28] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis, "Phylogenetic Inference," *Molecular Systematics,* D.M. Hillis, B.K. Mable, and C. Moritz, eds., pp. 407-514, Sinauer Assoc., 1996.

Bernard M.E. Moret is a professor of computer science and of electrical and computer engineering at the University of New Mexico. His research centers on algorithm engineering—the design and application of high-performance algorithms to specific classes of problems. For the last five years, he has been applying algorithm engineering techniques to problems in computational molecular biology, with particular emphasis on phylogenetic reconstruction and comparative genomics. He was an editor-in-chief and founding editor of the *ACM Journal of Experimental Algorithmics* and serves as an associate editor of *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. He directs the CIPRES project (www.phylo.org), funded by a large ITR grant from the US National Science Foundation to develop a national cyber infrastructure to support the reconstruction of the Tree of Life.

Luay Nakhleh received the BSc degree in computer science in 1996 from the Technion—Israel Institute of Technology. He received the Master's degree in computer science from Texas A&M University in 1998. He received the PhD degree in computer science from The University of Texas at Austin, working under the supervision of Tandy Warnow. His research interests fall in the general areas of computational biology and bioinformatics; in particular, he works on computational phylogenetics in biology and historical linguistics. He will join the faculty of Rice University in August 2004, as an assistant professor in the Department of Computer Science.

Tandy Warnow is a professor of computer sciences at the University of Texas at Austin, and an Emeline Bigelow Conland Fellow at the Radcliffe Institute for Advanced Studies. Her research combines mathematics, computer science, and statistics to develop improved models and algorithms for reconstructing complex and large-scale evolutionary histories in both biology and historical linguistics. She is on the board of directors of the International Society for Computational Biology, and previously was the codirector of the Center for Computational Biology and Bioinformatics at the University of Texas at Austin. Tandy received a US National Science Foundation Young Investigator Award in 1994 and a David and Lucile Packard Foundation Award in Science and Engineering in 1996. She is currently focusing her efforts on the CIPRES project (www.phylo.org), funded by a large ITR grant from the US National Science Foundation to develop a national cyber infrastructure to support the reconstruction of the Tree of Life.

C. Randal Linder is an associate professor in the Section of Integrative Biology at the University of Texas at Austin. Dr. Linder's research currently focuses on two areas in evolutionary biology: reconstruction of reticulate evolutionary networks, especially at the phylogenetic level, and the adaptive evolution of plants, especially seed oil composition.

Anna Tholse received the BS and MS degrees in computer science from the University of New Mexico in 2001 and 2003, respectively. From 2001 to 2003, she was a research assistant in the Moret-Bader Laboratory, working on modeling and assessment issues in the reconstruction of phylogenetic networks. In August 2003, she joined the staff of Microsoft in Redmond, Washington, as a software design engineer.

Anneke Padolina received the BS degree in botany in 2000 from the University of Florida. She is a PhD candidate in biology at the University of Texas, Austin, where she studied under a US National Science Foundation graduate fellowship. She is working on hybrid evolution in a phylogenetic context and reconstructing the phylogeny of the orchid tribe Phalaenopsis, including horticultural hybrids (of known pedigrees) to test her methods.

Jerry Sun received the BS degree in computer science in 2003 from the University of Texas, Austin, where he worked in Tandy Warnow's lab on algorithms for phylogenetic reconstruction. He is a PhD candidate in computer science at the University of Colorado, Boulder, where he works on educational software as part of the Reading Tutor project.

Ruth Timme received the BS degree in botany in 1999 and the MA degree in biology in 2001, both from San Francisco State University. She is a PhD candidate in biology at the University of Texas, Austin, funded by a US National Science Foundation IGERT fellowship and, now, by a US National Science Foundation Dissertation Improvement grant. She is working on a molecular phylogeny of the sunflowers (Helianthus), a group in which hybridization is common.