Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7534258, 9 pages
http://dx.doi.org/10.1155/2016/7534258

Research Article

A Metric on the Space of Partly Reduced Phylogenetic Networks

Juan Wang

School of Computer Science, Inner Mongolia University, Hohhot 010021, China

Correspondence should be addressed to Juan Wang; wangjuanangle@hit.edu.cn

Received 30 March 2016; Accepted 23 May 2016

Academic Editor: Dariusz Mrozek

Phylogenetic networks are a generalization of phylogenetic trees that allow for the representation of evolutionary events acting at the population level, such as recombination between genes, hybridization between lineages, and horizontal gene transfer. The researchers have designed several measures for computing the dissimilarity between two phylogenetic networks, and each measure has been proven to be a metric on a special kind of phylogenetic networks. However, none of the existing measures is a metric on the space of partly reduced phylogenetic networks. In this paper, we provide a metric, d_e-distance, on the space of partly reduced phylogenetic networks, which is polynomial-time computable.

1. Introduction

Phylogenies reveal the history of evolutionary events of a group of species, and they are central to comparative analysis methods for testing hypotheses in evolutionary biology [1]. Computing the distance between a pair of phylogenies is very important for understanding the evolutionary history of species.

A metric d on a space S satisfies four properties for all $a, b, c \in S$:

(I) $d(a, b) \geq 0$ (nonnegative property);

(II) $d(a, b) = 0$ if and only if $a = b$ (separation property);

(III) $d(a, b) = d(b, a)$ (symmetry property);

(IV) $d(a, b) + d(b, c) \geq d(a, c)$ (triangle inequality).

Phylogenetic network can represent reticulate evolutionary events, such as recombinations between genes, hybridization between lineages, and horizontal gene transfer [2–5]. For the comparison of phylogenetic networks, there are many metrics on the restricted subclasses of networks including the tripartition metric on the space of tree-child phylogenetic networks [6–9], the μ-distance on the space of tree-sibling phylogenetic networks [10], and the m-distance on the space of reduced phylogenetic networks [11]. Later the m-distance was also proved to be a metric on the space of tree-child phylogenetic networks, semibinary tree-sibling time consistent phylogenetic networks, and multilabeled phylogenetic trees [12–15].

For any rooted phylogenetic network N, we can obtain its reduced version by removing all nodes in maximal convergent sets (will be discussed in the following) and all the nodes, with indegree 1 and outdegree 1, from N. The reduced versions of all rooted phylogenetic networks form the space of reduced phylogenetic networks (m-distance, defined by Nakhleh, is on this space). In this paper, we will discuss the partly reduced version of a phylogenetic network by removing the nodes in a part of the convergent sets and all the nodes, with indegree 1 and outdegree 1, from the phylogenetic network. The partly reduced versions of all rooted phylogenetic networks form the space of partly reduced phylogenetic networks. Then we will introduce a novel metric on the space of partly reduced phylogenetic networks. The space is not the space of rooted phylogenetic networks, but it is the largest space on which a polynomial-time computable metric has been defined so for. The papers [16, 17] have proved that the isomorphism for rooted phylogenetic networks is graph isomorphism-complete. Unless the graph isomorphism problem belongs to P, there is no hope of defining a polynomial-time computable metric on the space of all rooted phylogenetic networks. However, our paper's aim is mainly to find a larger space on which a polynomial-time computable metric can be defined such that the space is closer to the space of rooted phylogenetic networks.

2. Preliminaries

Let $N = (V, E)$ be a directed acyclic graph, or DAG for short. We denote the indegree of a node u as indeg(u) and the outdegree of u as outdeg(u). We will say that a node u is a *tree node* if indeg$(u) \le 1$. Particularly, u is a *root* of N if indeg$(u) = 0$ of N. If a single root exists, we will say that the DAG is *rooted*. We will say that a node u is a *reticulate node* if indeg$(u) \ge 2$. A tree node u is a *leaf* if outdeg$(u) = 0$. A node is called an *internal node* if its outdeg ≥ 1. For a DAG $N = (V, E)$, we will say that v is a *child* of u if $(u, v) \in E$; in this case, we will also say that u is a *parent* of v. Note that any tree node has a single parent, except for the root of the graph. Whenever there is a directed path from a node u to v, we will say that v is a *descendant* of u or u is an *ancestor* of v.

The *height* of a node is the length of a longest path starting at the node and ending in a leaf. The absence of cycles implies that the nodes of a DAG N can be stratified by means of their heights: the nodes of height 0 are the leaves; if a node has height $a > 0$, then all its children have heights that are smaller than a and at least one of them has height exactly $a - 1$.

The *depth* of a node is the length of a longest path starting at the root and ending in the node. Similarly, the absence of cycles implies that the nodes of a DAG N can also be stratified according to their depths: the node of depth 0 is the root; if a node has depth $b > 0$, then all its parents have depths that are smaller than b and at least one of them has depth exactly $b - 1$.

Let $\mathcal{X}$ be a set of taxa. A rooted phylogenetic network N on $\mathcal{X}$ is a rooted DAG such that

(i) no tree node has outdeg 1;

(ii) its leaves are labeled by $\mathcal{X}$ by a bijective mapping f.

We use the notation $N = ((V, E), f)$ (or $N = (V, E)$) for the rooted phylogenetic network N and the notation V_N for its leaf set.

Definition 1. Two rooted phylogenetic networks $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ are isomorphic if and only if there is a bijection G from V_1 to V_2 such that

(i) (u, v) is an edge in E_1 if and only if $(G(u), G(v))$ is an edge in E_2;

(ii) $f_1(w) = f_2(G(w))$ for all $w \in V_{N_1}$.

Moret et al. (2004) discussed the concept of reduced phylogenetic networks from a reconstruction standpoint. Subsequently, we briefly review the concept of reduced phylogenetic networks and introduce a new definition of partly reduced phylogenetic networks. In the following section, we present a metric on the space of all partly reduced phylogenetic networks. First we review the concept of a maximal convergent set that has been given in [7, 11].

Definition 2. Given a network $N = (V, E)$, we say that a set U of internal nodes in V is convergent if $|U| \ge 2$ and

> every leaf reachable from some node in U is reachable from all nodes in U.

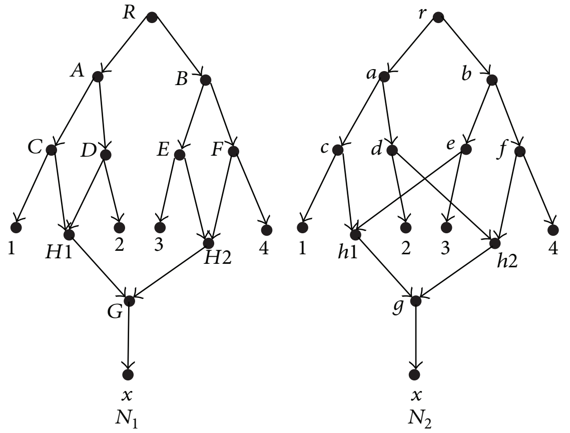

FIGURE 1: Networks N_1 and N_2 from refinements (1) and (2) in Table 1 in [11]. $H1$ and $H2$ (resp., $h1$ and $h2$) are the reticulate nodes, A~G (resp., a~g); $H1$ and $H2$ (resp., $h1$ and $h2$) as well as the root R (resp., r) are the internal tree nodes in network N_1 (resp., N_2).

If there is no convergent set U_0 containing U except U itself, we say that U is a maximal convergent set.

Here the leaf set reachable from the nodes in a convergent set U is called the leaf set of U.

We will take Figure 1 as an example in the following. The two networks N_1, N_2 on $\{1, 2, 3, 4, x\}$ are adapted from refinements (1) and (2) in Table 1 in [11].

Example 3. Consider the networks in Figure 1. The set $\{H1, H2, G\}$ is the only maximal convergent set of N_1 and the set $\{h1, h2, g\}$ is the only maximal convergent set of N_2.

For a phylogenetic network $N = ((V, E), f)$ on $\mathcal{X}$, the reduced version of N can be obtained by the following reduction procedures:

(1) For each maximal pendant subtree (i.e., the maximal clade that includes no reticulate nodes) t, rooted at node r_t, create a new node h_t and an edge (p_t, h_t), where p_t is the parent of r_t, delete the edge (p_t, r_t) and the subtree t, and label h_t as t. Then we denote the resulting network as N_0.

(2) Repeat the following two steps on N_0 until no change occurs:

(I) For each maximal convergent set U with leaf set $L_U \subseteq V_{N_0}$, remove all nodes and edges on the paths from a node in U to the parent of leaf in L_U, including all nodes in U and excluding the parent of leaf in L_U. For each edge (p, v), where p lies outside the deleted set and v lies inside the deleted set, replace it with a set of edges $\{(p, q)$: q is the parent of leaf in $L_U\}$.

(II) For each node w in the network, with indeg(w) = outdeg(w) = 1, remove the edges (u, w), (w, v) and the node w, add an edge (u, v), where u is the parent of w and v is the child of w. Repeat this step until no such node can be removed.

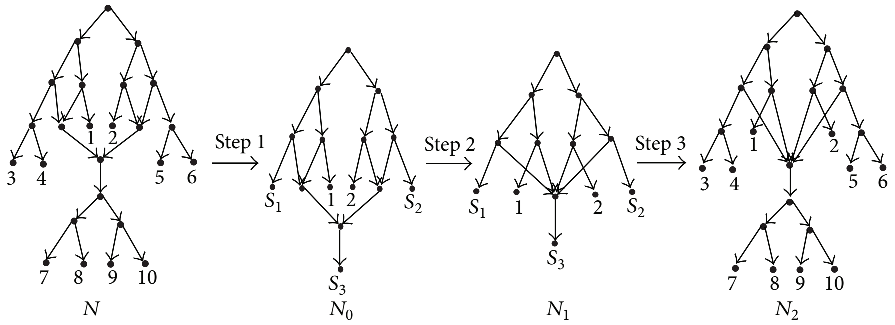

Figure 2: The rooted phylogenetic network N is on $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. N_0, N_1, and N_2 are the networks obtained by applying each one of the three reduction procedures to N, respectively.

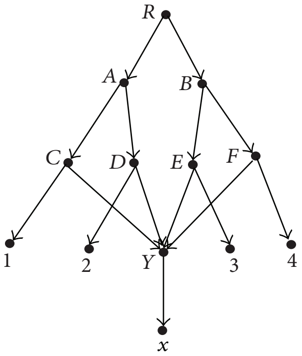

Figure 3: The reduced version of the networks in Figure 1.

(3) Replace each leaf labeled by the subtree t by its root r_t.

Figure 2 shows the results of applying the reduction procedures to the network N. For the networks in Figure 1, their reduced versions are the same (see Figure 3). The reduced versions of all rooted phylogenetic networks form the space of reduced phylogenetic networks. Nakhleh has introduced a polynomial-time computable metric on this space [11]. In order to enlarge the space in which a polynomial-time computable metric can be defined, we will introduce a new metric and a new space that contains the space of reduced phylogenetic networks.

Definition 4. Given a network $N = (V, E)$, let $\mathscr{P}(v)$ be the set of parents of a node v in V. We say that $U \subset V$ is a super convergent set, if

(i) U is a convergent set;

(ii) $\mathscr{P}(u_1) = \mathscr{P}(u_2)$ for any two nodes $u_1, u_2 \in U$;

(iii) $\mathscr{P}(u)$ is a convergent set for a node $u \in U$, if $|\mathscr{P}(u)| \geq 2$.

Example 5. The set $\{H, J\}$ is the only superconvergent set for any one network in Figure 4, while the networks in Figure 1 have no superconvergent set.

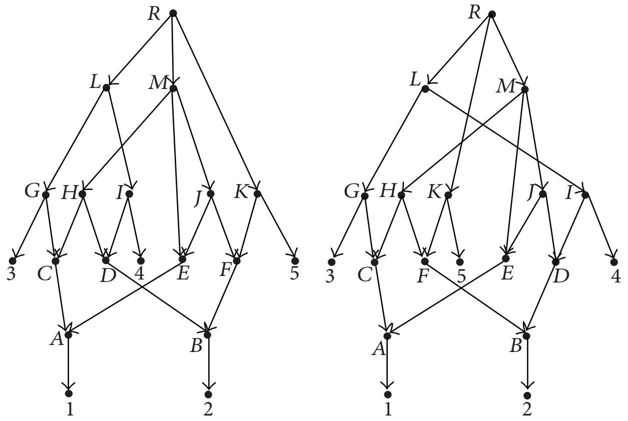

Figure 4: Networks N_1 and N_2 are not isomorphic.

We will obtain the new reduction procedures, called partial reduction procedures, from the above reduction procedures by just processing superconvergent sets rather than maximal convergent sets in step (I) of step (2). After applying the partial reduction procedures to a rooted phylogenetic network N, the partly reduced version of N is obtained. The partly reduced versions of all rooted phylogenetic networks form the space of partly reduced phylogenetic networks. This space contains the space of reduced phylogenetic networks, but they are not identical. Next we will introduce a polynomial-time computable metric for the partly reduced phylogenetic networks.

We begin with the notion of node semiequivalence. For the sake of simplicity, we will hereafter refer to the rooted phylogenetic networks as the networks.

3. A Metric

Definition 6. Given a network $N = ((V, E), f)$, we say that two nodes $u, v \in V$ (not necessarily different) are semiequivalent, denoted by $u \mathrel{\hat{=}} v$, if

(i) $u, v \in V_N$ and $f(u) = f(v)$ or

(ii) node u has k (≥ 1) children $u_1, u_2, \ldots, u_k$; node v has k children $v_1, v_2, \ldots, v_k$, and $u_i \mathrel{\hat{=}} v_i$ for $1 \leq i \leq k$.

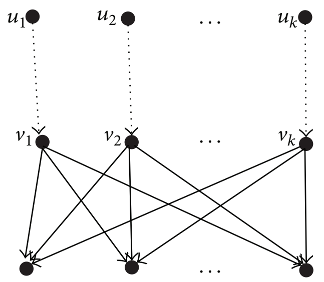

Figure 5: The topology relation of semiequivalent nodes.

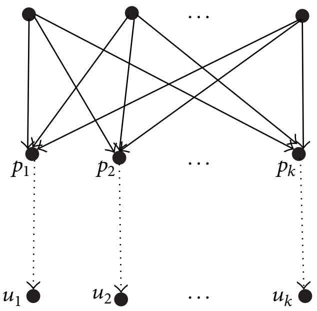

Figure 6: The topology relation of equivalent nodes.

By the definition, it follows that the semiequivalence of nodes is an equivalence relation; that is, it is reflexive, symmetric, and transitive, and the semiequivalent nodes must have the same height.

Example 7. Consider the network N_1 in Figure 1. For any node $u \in V_1 \setminus \{H1, H2\}$, u is only semiequivalent to u itself, while the nodes $H1$ and $H2$ are semiequivalent.

Property 1. If $u_1, u_2, \ldots, u_k$ are semiequivalent from the network $N = ((V, E), f)$, then $u_1, u_2, \ldots, u_k$ are the same nodes or there are the nodes v_1 (u_1 or a descendant of u_1), v_2 (u_2 or a descendant of u_2), ..., v_k (u_k or a descendant of u_k) such that $v_1, v_2, \ldots, v_k$ have the same children. See Figure 5.

Proof. We use induction on the height a of u_1 to prove it. If $a = 0$, obviously $u_1, u_2, \ldots, u_k$ are the only leaf. Thus, in this case, the property holds. We assume that the result is tenable when $a \leq n$, and let $a = n + 1$. Then the children of $u_1, u_2, \ldots, u_k$ are semiequivalent, respectively (let the children of u_i be $a_{i1}, a_{i2}, \ldots, a_{il}$ for $1 \leq i \leq k$; then $a_{1j}, a_{2j}, \ldots, a_{kj}$ are semiequivalent for $1 \leq j \leq l$), and their height is at most n by the property of node height. By the induction hypothesis, the children of $u_1, u_2, \ldots, u_k$ satisfy the property. The descendants of children of $u_1, u_2, \ldots, u_k$ are the descendants of $u_1, u_2, \ldots, u_k$. Thus, the property holds. □

Definition 8. Given a network $N = (V, E)$, we say that two nodes $u, v \in V$ (not necessarily different) are equivalent, denoted by $u \equiv v$, if $u \triangleq v$, and

(i) u, v are the root or

(ii) node u has l (≥ 1) parents $u_1, u_2, \ldots, u_l$; node v has l parents $v_1, v_2, \ldots, v_l$, and $u_i \equiv v_i$ for $1 \leq i \leq l$.

For any node u in N, it is equivalent to itself. The equivalence of nodes is also an equivalence relation. The equivalent nodes have the same height and depth.

Example 9. Consider the network N_1 in Figure 1. For any node $u \in V_1$, it is equivalent to itself. Consider the network N_1 in Figure 4. For any node $u \in V_1 \setminus \{H, J\}$, it is equivalent to itself, while the nodes H and J are equivalent to each other.

Property 2. If $u_1, u_2, \ldots, u_k$ are equivalent in the network $N = ((V, E), f)$, then $u_1, u_2, \ldots, u_k$ are the same nodes or there are the nodes p_1 (u_1 or an ancestor of u_1), p_2 (u_2 or an ancestor of u_2), ..., p_k (u_k or an ancestor of u_k) such that $p_1, p_2, \ldots, p_k$ have the same parents. See Figure 6.

Proof. We use induction on the depth b of u_1 to prove it. If $b = 0$, then $u_1, u_2, \ldots, u_k$ are the unique root node. Thus, in this case, the property holds. We assume that the result is tenable when $b \leq n$, and let $b = n + 1$. Then the parents of $u_1, u_2, \ldots, u_k$ are equivalent, respectively (let the parents of u_i be $a_{i1}, a_{i2}, \ldots, a_{il}$ for $1 \leq i \leq k$; then $a_{1j}, a_{2j}, \ldots, a_{kj}$ are equivalent for $1 \leq j \leq l$), and their depth is at most n by the property of node depth. By the induction hypothesis, the parents of $u_1, u_2, \ldots, u_k$ satisfy the property. The ancestors of the parents of $u_1, u_2, \ldots, u_k$ are the ancestors of $u_1, u_2, \ldots, u_k$. Thus, the property holds. □

In this paper, we are mainly concerned with comparing networks; the notion of node semiequivalence and equivalence will be extended to nodes from two different networks, as established in the semiequivalence and equivalence mapping of Definitions 10 and 13, respectively.

Given a set V, we use $P(V)$ to denote the set of all subsets of V.

Definition 10. Let $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ be two networks on $\mathcal{X}$. We define the semiequivalence mapping between N_1 and N_2, $h : V_1 \rightarrow P(V_2)$, such that $v \in h(u)$, for $u \in V_1$ and $v \in V_2$, if

(i) $u \in V_{N_1}$, $v \in V_{N_2}$, and $f_1(u) = f_2(v)$ or

(ii) node u has k (≥ 1) children $u_1, u_2, \ldots, u_k$; node v has k children $v_1, v_2, \ldots, v_k$, and $v_i \in h(u_i)$ for $1 \leq i \leq k$.

Further, while inequation $|h(u_1)| \leq 1$ holds in phylogenetic trees, it is not always the case for general phylogenetic networks.

Example 11. Consider the networks in Figure 1. h is a semiequivalence mapping between N_1 and N_2. For the reticulate nodes $H1$ and $H2$ in N_1, $h(H1) = \{h1, h2\}$ and $h(H2) = \{h1, h2\}$. For the other nodes in N_1, $h(A) = \{a\}$, $h(B) = \{b\}, \ldots, h(G) = \{g\}$, $h(1) = \{1\}, \ldots, h(4) = \{4\}$, $h(x) = \{x\}$, and $h(R) = \{r\}$.

Theorem 12. *Let $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ be two networks on $\mathcal{X}$, and let u_1, u_2 be two nodes in V_1 and h a semiequivalence mapping between N_1 and N_2. Assume that*

$h(u_1) \neq \emptyset$ and $h(u_2) \neq \emptyset$. Then, $u_1 \triangleq u_2$ if and only if $v_1 \triangleq v_2$, for $v_1 \in h(u_1)$ and $v_2 \in h(u_2)$.

Proof. For the "only if" direction, let $v_1 \in h(u_1)$, $v_2 \in h(u_2)$, and $u_1 \triangleq u_2$. Obviously, u_1, u_2, v_1, and v_2 have the same height a. Then, we use induction on such height a to prove $v_1 \triangleq v_2$. In particular, if $a = 0$, that is, $u_1, u_2 \in V_{N_1}$, and $f_1(u_1) = f_1(u_2)$, then $v_1, v_2 \in V_{N_2}$ and $f_2(v_1) = f_1(u_1) = f_1(u_2) = f_2(v_2)$. Thus, in this case, $v_1 \triangleq v_2$. We assume that the result is tenable when $a \leq n$, and let $a = n+1$. We assume that node u_1 has k children $p_1, p_2, \ldots, p_k$. Due to $u_1 \triangleq u_2$, it follows that node u_2 has k children $q_1, q_2, \ldots, q_k$, and $p_i \triangleq q_i$ $(1 \leq i \leq k)$. Due to $v_1 \in h(u_1)$ and $v_2 \in h(u_2)$, it follows that v_1 has k children $w_1, w_2, \ldots, w_k$, and $w_i \in h(p_i)$ $(1 \leq i \leq k)$, v_2 has k children $y_1, y_2, \ldots, y_k$, and $y_i \in h(q_i)$ $(1 \leq i \leq k)$. The height of p_i, q_i, w_i, and y_i is at most n. By the induction hypothesis, $w_i \triangleq y_i$. Thus, $v_1 \triangleq v_2$.

For the "if" direction, let $v_1 \in h(u_1)$, $v_2 \in h(u_2)$, and $v_1 \triangleq v_2$. Similarly, we also use induction on the same height a of u_1, u_2, v_1, and v_2 to prove $u_1 \triangleq u_2$. If $a = 0$, that is, $v_1, v_2 \in V_{N_2}$, and $f_2(v_1) = f_2(v_2)$, then $u_1, u_2 \in V_{N_1}$ and $f_1(u_1) = f_2(v_1) = f_2(v_2) = f_1(u_2)$. Thus, in this case, $u_1 \triangleq u_2$. We assume that the result is tenable when $a \leq n$, and let $a = n+1$. We assume that node v_1 has k children $w_1, w_2, \ldots, w_k$. Since $v_1 \triangleq v_2$, node v_2 has k children $y_1, y_2, \ldots, y_k$, and $w_i \triangleq y_i$ $(1 \leq i \leq k)$. Since $v_1 \in h(u_1)$ and $v_2 \in h(u_2)$, u_1 has k children $p_1, p_2, \ldots, p_k$, and $w_i \in h(p_i)$ $(1 \leq i \leq k)$, u_2 has k children $q_1, q_2, \ldots, q_k$, and $y_i \in h(q_i)$ $(1 \leq i \leq k)$. The height of p_i, q_i, w_i, and y_i is at most n by the property of node height. By the induction hypothesis, $p_i \triangleq q_i$. Thus, $u_1 \triangleq u_2$. □

Theorem 12 tells us that the semiequivalence mapping keeps the semiequivalence of nodes. Thus, all nodes in $h(u)$ are semiequivalent. Sometimes we use $h(u)$ to denote an arbitrary node in the set. We say that the nodes in $h(u)$ are semiequivalent with u.

Definition 13. Let $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ be two networks on $\mathscr{X}$. We define the equivalence mapping between N_1 and N_2, $g : V_1 \rightarrow P(V_2)$, such that $v \in g(u)$, for $u \in V_1$ and $v \in V_2$, if $v \in h(u)$, and

(i) u, v are the roots or

(ii) node u has l (≥ 1) parents $u_1, u_2, \ldots, u_l$; node v has l parents $v_1, v_2, \ldots, v_l$, and $v_i \in g(u_i)$, for $1 \leq i \leq l$,

where h is a semiequivalence mapping between N_1 and N_2.

Example 14. Consider the networks in Figure 1. h is the semiequivalence mapping between N_1 and N_2 discussed in Example 11. g is an equivalence mapping between N_1 and N_2 defined in Definition 13. For any node $u \in V_1 \setminus \{H1, H2, G \text{ and } x\}$, $g(u) = h(u)$, while $g(v) = \emptyset$ when $v \in \{H1, H2, G \text{ and } x\}$.

Theorem 15. *Let $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ be two networks on $\mathscr{X}$, and let u_1, u_2 be two nodes in V_1. g is an equivalence mapping between N_1 and N_2. Assume that $g(u_1) \neq \emptyset$ and $g(u_2) \neq \emptyset$. Then, $u_1 \equiv u_2$ if and only if $v_1 \equiv v_2$, for $v_1 \in g(u_1)$ and $v_2 \in g(u_2)$.*

Proof. Let $v_1 \in g(u_1)$, $v_2 \in g(u_2)$. Then $v_1 \in h(u_1)$, $v_2 \in h(u_2)$ based on Definition 13. For the "only if" direction, let $u_1 \equiv u_2$. We can deduce that $v_1 \triangleq v_2$ according to Theorem 12, and u_1, u_2 and v_1 and v_2 have the same depth b. Then, we use induction on b to prove that $v_1 \equiv v_2$. If $b = 0$, that is, u_1, u_2 are the unique root node of N_1, then v_1, v_2 are the unique root node of N_2. Thus, in this case, $v_1 \equiv v_2$. We assume that the result is tenable when $b \leq n$, and let $b = n+1$. We assume that node u_1 has l parents $p_1, p_2, \ldots, p_l$. Due to $u_1 \equiv u_2$, node u_2 has l parents $q_1, q_2, \ldots, q_l$, and $p_i \equiv q_i$ $(1 \leq i \leq l)$. Due to $v_1 \in g(u_1)$ and $v_2 \in g(u_2)$, v_1 has l parents $w_1, w_2, \ldots, w_l$, and $w_i \in g(p_i)$ $(1 \leq i \leq l)$, v_2 has l parents $y_1, y_2, \ldots, y_l$, and $y_i \in g(q_i)$ $(1 \leq i \leq l)$. The depth of p_i, q_i, w_i, and y_i is at most n by the property of node depth. By the induction hypothesis, $w_i \equiv y_i$. Thus, $v_1 \equiv v_2$.

For the "if" direction, let $v_1 \in g(u_1)$, $v_2 \in g(u_2)$, and $v_1 \equiv v_2$. We can deduce first that $u_1 \triangleq u_2$ according to Theorem 12. Similarly, we also use induction on the same depth b of u_1, u_2 and v_1, v_2 to prove that $u_1 \equiv u_2$. If $b = 0$, that is, v_1, v_2 are the unique root node of N_2, then u_1, u_2 are the unique root node of N_1. Thus, in this case, $u_1 \equiv u_2$. We assume that the result is tenable when $b \leq n$, and let $b = n+1$. We assume that node v_1 has l parents $w_1, w_2, \ldots, w_l$. Due to $v_1 \equiv v_2$, node v_2 has l parents $y_1, y_2, \ldots, y_l$, and $w_i \equiv y_i$ $(1 \leq i \leq l)$. Due to $v_1 \in g(u_1)$ and $v_2 \in g(u_2)$, u_1 has l parents $p_1, p_2, \ldots, p_l$, and $w_i \in g(p_i)$ $(1 \leq i \leq l)$, u_2 has l parents $q_1, q_2, \ldots, q_l$, and $y_i \in g(q_i)$ $(1 \leq i \leq l)$. The depth of p_i, q_i, w_i, and y_i is at most n. So, by the induction hypothesis, $p_i \equiv q_i$. Thus, $u_1 \equiv u_2$. □

Theorem 15 tells us that the equivalence mapping keeps the equivalence of nodes. Thus, all nodes in $g(u)$ are equivalent. Sometimes we use $g(u)$ to denote an arbitrary node in the set. We say that the nodes in $g(u)$ are equivalent to u.

Lemma 16. *Let $N = ((V, E), f)$ be a network and $u, v \in V$ two equivalent nodes. Then u, v belong to a superconvergent set.*

Proof. This lemma is obtained easily from Properties 1 and 2. □

Lemma 17. *Let $N = ((V, E), f)$ be a partly reduced phylogenetic network. Then $u_1 \not\equiv u_2$ for any two nodes $u_1, u_2 \in V$.*

Proof. From the partial reduction procedures of the network, we have that all superconvergent sets in a partly reduced network have been deleted. □

Given two networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$, assume that $V_1 = \{v_1, v_2, \ldots, v_p\}$. The unique nodes of N_1, denoted by $L(N_1)$, is defined by the following processes. First let $L(N_1) = \emptyset$. Then for each one node $u \in V_1$, if there exists no node $u' \in L(N_1)$ such that $u' \equiv u$, add u to $L(N_1)$. We define $L(N_2)$ in a similar way. Further for each node $v_i \in L(N_1)$, we define $e_{N_1}(v_i) = |\{v \in V_1 : v \equiv v_i\}|$ and $e_{N_2}(u_i)$ similarly for each node $u_i \in V_2$. We define $e(\emptyset) = 0$ for any network N. When the context is clear, we drop the subscript of e. We are now in a position to define the measure on pairs of partly reduced phylogenetic networks.

```
input: nodes $u$ and $v$
if the outdeg of $u$ and the outdeg of $v$ are not equal then
    return
end if
if $u$ and $v$ are leaves and $f_1(u) = f_1(v)$ (or $f_1(u) = f_2(v)$ i.e., $u$ and $v$ are from two networks) then
    add $v$ to the ISE of $u$
    add $u$ to the ISE of $v$
else
    flag := false
    for each child $a$ of $u$ do
        for each child $b$ of $v$ do
            if $b.label$ = true then
                continue
            end if
            if the ISE of $a$ has $b$ then
                flag = true
                $b.label = true$
            end if
        end for
        if flag = false then
            return
        else
            flag = false
        end if
    end for
    add $v$ to the ISE of $u$
    add $u$ to the ISE of $v$
end if
```

Algorithm 1: Deciding semiequivalence for two nodes u and v.

Definition 18. Let $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ be two phylogenetic networks on $\mathcal{X}$. Then $d_e(N_1, N_2)$ equals

$$\frac{1}{2}\left[\sum_{v\in L(N_1)} \max\left\{0, e(v) - e(v')\right\} + \sum_{u\in L(N_2)} \max\left\{0, e(u) - e(u')\right\}\right], \tag{1}$$

where $v'(u')$ is a node in $L(N_2)(L(N_1))$ that is equivalent to $v(u)$, and if no such equivalent node exists, then $v'(u') = \emptyset$.

Lemma 19. *If $d_e(N_1, N_2) = 0$ for two networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$, then $|V_1| = |V_2|$.*

Proof. Let $g_1 : V_1 \to P(V_2)$ and $g_2 : V_2 \to P(V_1)$ be two equivalence mappings from Definition 13. Since $d_e(N_1, N_2) = 0$, it follows that $e(v_1) = e(g_1(v_1))$ (where $g_1(v_1)$ denotes a node u, which is equivalent to $g_1(v_1)$ and in $L(N_2)$) along with $|g_1(v_1)| > 0$ for all $v_1 \in L(N_1)$ and $e(v_2) = e(g_2(v_2))$ (where $g_2(v_2)$ denotes a node u, which is equivalent to $g_2(v_2)$ and in $L(N_1)$) along with $|g_2(v_2)| > 0$ for all $v_2 \in L(N_2)$. From this and Theorem 15, we have that $|V_1| = \sum_{v_1\in L(N_1)} e(v_1) = \sum_{v_1\in L(N_1)} e(g_1(v_1)) \le |V_2|$ (due to $g_1(v_1) \in V_2$) and $|V_2| = \sum_{v_2\in L(N_2)} e(v_2) = \sum_{v_2\in L(N_2)} e(g_2(v_2)) \le |V_1|$ (due to $g_2(v_2) \in V_1$). Thus $|V_1| = |V_2|$. □

Theorem 20. *Let $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ be two partly reduced networks. Then, N_1 and N_2 are isomorphic if and only if $d_e(N_1, N_2) = 0$.*

Proof. Let $g : V_1 \to P(V_2)$ be an equivalence mapping, as given in Definition 13. From Lemma 19, it follows that $|V_1| = |V_2|$ and $e(v) = e(g(v))$ for all $v \in L(N_1)$. From Lemmas 16 and 17, we have that $g(v_1)$ is defined and unique for each $v_1 \in V_1$. We now prove that if $(u, v) \in E_1$, then $(u_0, v_0) \in E_2$, where $v_0 = g(v)$ and $u_0 = g(u)$. Given that $v_0 = g(v)$, that is, v and v_0 are equivalent, this implies that v_0 and v have equivalent parents. Since $u_0 = g(u)$ is defined and unique, u_0 is a parent of v_0. Thus, $(u_0, v_0) \in E_2$. It shows that the mapping g is bijective, which also preserves the labels of the leaves and the edges of networks. Thus, N_1 and N_2 are isomorphic.

The converse implication is obvious. □

From the definition of the measure, the symmetry property follows immediately.

Lemma 21. *For any pair networks N_1 and N_2, one has $d_e(N_1, N_2) = d_e(N_2, N_1)$.*

The measure $d_e(N_1, N_2)$ can be viewed as half of the symmetric difference of two multisets on the same set of elements, where the multiplicity of element u in N_1 is $e_{N_1}(u)$ and similarly for N_2. Since the symmetric difference defines a metric on multisets [12], we have the following triangle inequality.

```
input: nodes $u$ and $v$
if the indeg of $u$ and the indeg of $v$ are not equal, or the ISE of $u$ doesn't have $v$ (the ESE of $u$
    doesn't have $v$ i.e., $u$ and $v$ are from two networks) then
    return
end if
if $u$ and $v$ are roots then
    add $v$ to the IE of $u$
    add $u$ to the IE of $v$
else
    flag := false
    for each parent $a$ of $u$ do
        for each parent $b$ of $v$ do
            if $b.label$ = true then
                continue
            end if
            if the IE of $a$ has $b$ then
                flag = true
                $b.label$ = true
            end if
        end for
        if flag = false then
            return
        else
            flag = false
        end if
    end for
    add $v$ to the IE of $u$
    add $u$ to the IE of $v$
end if
```

Algorithm 2: Deciding equivalence for two nodes u and v.

Lemma 22. *Let N_1, N_2, and N_3 be three networks. Then, $d_e(N_1, N_2) + d_e(N_2, N_3) \geq d_e(N_1, N_3)$.*

From Theorem 20 and Lemmas 21 and 22, we have the following main result.

Theorem 23. *The measure d_e is a metric on the space of partly reduced phylogenetic networks.*

Proof. It follows from Theorem 20 and Lemmas 21 and 22 and the fact that $\max\{0, e(v) - e(v')\} \geq 0$. □

Let $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ be two phylogenetic networks. For a node u in N_1, we refer to its semiequivalent nodes from N_1 as internal semiequivalence (equivalence) nodes and its semiequivalent (equivalence) nodes from N_2 as external semiequivalence (equivalence) nodes. When computing the distance between two networks, we first compute internal and external equivalence nodes for every node in the two networks; subsequently by formula (1) we obtain the distance between the two considered networks. The maximum of measure $d_e(N_1, N_2)$ is $(|V_1|+|V_2|)/2.0$, when any node in N_1 and in N_2 has no external equivalence nodes.

In order to show the results of the distance computed by formula (1), we give an example as follows.

Example 24. Consider the networks in Figure 1. N_1, N_2 are two different networks on $\{1, 2, 3, 4, x\}$. However, in [11], they are indistinguishable and their m-distance [11] is 0. Now, we compute the d_e-distance between them: $d_e(N_1, N_2) = 4$ (see Example 14).

4. Computational Aspects

From the definition of semiequivalent nodes, whether in the same network or in two different networks, we have that the semiequivalent nodes can be computed by means of a bottom-up technique. Similarly, the equivalent nodes can be computed by means of a top-down technique. Let $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ be two phylogenetic networks. For a pair of nodes u and v, whether in the same network or in different networks, the following shows the pseudocode (Algorithm 1) that decides whether they are internal semiequivalent to each other, the pseudocode (Algorithm 2) that decides whether they are internal equivalent to each other, and the pseudocode (Algorithm 3) that computes the d_e-distance for a pair of networks (where ISE is the abbreviation for the set of internal semiequivalent nodes, ESE is the abbreviation for the set of external semiequivalent nodes, IE is the abbreviation for the set of internal equivalent nodes, and EE is the abbreviation for the set of external equivalent nodes). If two nodes u and v from the same network are semiequivalent, then we add u to the ISE of v and add v to the ISE of u. Obviously, this decision costs at

```
input: networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$
output: $d_e$-distance
for each pair of nodes $u$ and $v$ in $V_1$ do
    decide semi-equivalence and equivalence for them
end for
for each pair of nodes $u$ and $v$ in $V_2$ do
    decide semi-equivalence and equivalence for them
end for
for each pair of nodes $u$ in $V_1$ and $v$ in $V_2$ do
    decide semi-equivalence and equivalence for them
end for
$L(N_1) = \emptyset$; $L(N_2) = \emptyset$
flag1 = false; flag2 = false
for each node $u$ in $V_1$ do
    for each node $v$ in $L(N_1)$ do
        if the IE of $v$ contains $u$ then
            flag1 = true
        end if
    end for
    if flag1 = false then
        add $u$ to $L(N_1)$
    end if
end for
for each node $u$ in $V_2$ do
    for each node $v$ in $L(N_2)$ do
        if the IE of $v$ contains $u$ then
            flag2 = true
        end if
    end for
    if flag2 = false  then
        add $u$ to $L(N_2)$
    end if
end for
$d = 0$
for each node $u$ in $L(N_1)$ do
    $c = |\mathrm{IE}| - |\mathrm{EE}|$
    if$c > 0$ then
        $d = d + c$
    end if
end for
for each node $u$ in $L(N_2)$ do
    $c = |\mathrm{IE}| - |\mathrm{EE}|$
    if $c > 0$ then
        $d = d + c$
    end if
end for
return $d = d/2$
```

Algorithm 3: Computing the d_e-distance for $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$.

most $O(n^3)$ time, where $n = \max(|V_1|, |V_2|)$. So, it takes totally $O(n^5)$ time to find out all internal and external semiequivalent nodes for every node in the two networks. In a similar way, we have that it also takes $O(n^5)$ time to find out all internal and external equivalent nodes for every node in the two networks. Subsequently we spend $O(n)$ time computing the formula (1). In conclusion, it costs totally $O(n^5)$ time to compute the distance between two networks, where n is the maximum between their node numbers.

5. Conclusion

In [11], Nakhleh introduced a polynomial-time computable m-distance in the space of reduced phylogenetic networks. In order to enlarge the space of phylogenetic networks we can compare, we devised a polynomial-time computable d_e-distance on the space of partly reduced phylogenetic networks, which can be viewed as half of the symmetric difference of two multisets on the same set of elements. To our knowledge, the space is the largest space that has a polynomial-time computable metric. d_e-distance is also a metric on the space of reduced phylogenetic networks which is included in the space of partly reduced phylogenetic networks. In general, for two phylogenetic networks, their d_e-distance is larger than their m-distance. From [12], we have that the d_e-distance is also a metric on the space of tree-child phylogenetic networks, semibinary tree-sibling time consistent phylogenetic networks, and multilabeled phylogenetic trees. However, the d_e-distance is not a metric on the space of all rooted phylogenetic networks; for example, in the two phylogenetic networks in Figure 4, their d_e-distance is 0, but they are not isomorphic.

d_e-distance can also apply to computing the dissimilarity for other types of networks, such as spiking neural networks [18–20], which will be a direction of further research.

Competing Interests

The author declares that they have no competing interests.

Acknowledgments

This work was supported by the Natural Science Foundation of Inner Mongolia province of China (2015BS0601).

References

[1] M. Pagel, "Inferring the historical patterns of biological evolution," *Nature*, vol. 401, no. 6756, pp. 877–884, 1999.

[2] J. Wang, M. Guo, X. Liu et al., "Lnetwork: an efficient and effective method for constructing phylogenetic networks," *Bioinformatics*, vol. 29, no. 18, pp. 2269–2276, 2013.

[3] J. Wang, M. Guo, L. Xing, K. Che, X. Liu, and C. Wang, "BIMLR: a method for constructing rooted phylogenetic networks from rooted phylogenetic trees," *Gene*, vol. 527, no. 1, pp. 344–351, 2013.

[4] J. Wang, "A new algorithm to construct phylogenetic networks from trees," *Genetics and Molecular Research*, vol. 13, no. 1, pp. 1456–1464, 2014.

[5] J. Wang, M.-Z. Guo, and L. L. Xing, "FastJoin, an improved neighbor-joining algorithm," *Genetics and Molecular Research*, vol. 11, no. 3, pp. 1909–1922, 2012.

[6] L. Nakhleh, J. S. T. Warnow, C. R. Linder, B. M. Moret, and A. Tholse, "Towards the development of computational tools for evaluating phylogenetic network reconstruction methods," in *Proceedings of the 18th Pacific Symposium on Biocomputing*, Kauai, Hawaii, USA, January 2003.

[7] B. M. E. Moret, L. Nakhleh, T. Warnow et al., "Phylogenetic networks: modeling, reconstructibility, and accuracy," *IEEE/ACM*

Transactions on Computational Biology and Bioinformatics, vol. 1, no. 1, pp. 13–23, 2004.

[8] M. Baroni, C. Semple, and M. Steel, “A framework for representing reticulate evolution,” *Annals of Combinatorics*, vol. 8, no. 4, pp. 391–408, 2004.

[9] G. Cardona, F. Rosselló, and G. Valiente, “Tripartitions do not always discriminate phylogenetic networks,” *Mathematical Biosciences*, vol. 211, no. 2, pp. 356–370, 2008.

[10] G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, “A distance metric for a class of tree-sibling phylogenetic networks,” *Bioinformatics*, vol. 24, no. 13, pp. 1481–1488, 2008.

[11] L. Nakhleh, “A metric on the space of reduced phylogenetic networks,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 218–222, 2010.

[12] G. Cardona, M. Llabrès, F. Rossellò, and G. Valiente, “On Nakhleh’s metric for reduced phylogenetic networks,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 4, pp. 629–638, 2009.

[13] Q. Zou, Q. Hu, M. Guo, and G. Wang, “HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy,” *Bioinformatics*, vol. 31, no. 15, pp. 2475–2481, 2015.

[14] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, “Similarity computation strategies in the microRNA-disease network: a survey,” *Briefings in Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2016.

[15] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, “Survey of MapReduce frame operation in bioinformatics,” *Briefings in Bioinformatics*, vol. 15, no. 4, Article ID bbs088, pp. 637–647, 2014.

[16] G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, “The comparison of tree-sibling time consistent phylogenetic networks is graph isomorphism-complete,” *The Scientific World Journal*, vol. 2014, Article ID 254279, 6 pages, 2014.

[17] K. S. Booth and C. J. Colbourn, “Problems polynomially equivalent to graph isomorphism,” http://cs.uwaterloo.ca/research/tr/1977/CS-77-04.pdf.

[18] S. Chowhan, U. V. Kulkarni, and G. N. Shinde, “Iris recognition using modified fuzzy hypersphere neural network with different distance measures,” *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, 2011.

[19] A. Van Schaik, “Building blocks for electronic spiking neural networks,” *Neural Networks*, vol. 14, no. 6-7, pp. 617–628, 2001.

[20] B. J. Graham and D. P. M. Northmore, “A spiking neural network model of midbrain visuomotor mechanisms that avoids objects by estimating size and distance monocularly,” *Neurocomputing*, vol. 70, no. 10–12, pp. 1983–1987, 2007.