SCIENTIFIC REPORTS

OPEN

A Metric on the Space of *k*th-order reduced Phylogenetic Networks

Juan Wang[1] & Maozu Guo[2]

Received: 5 August 2016
Accepted: 27 April 2017
Published online: 09 June 2017

Phylogenetic networks can be used to describe the evolutionary history of species which experience a certain number of reticulate events, and represent conflicts in phylogenetic trees that may be due to inadequacies of the evolutionary model used in the construction of the trees. Measuring the dissimilarity between two phylogenetic networks is at the heart of our understanding of the evolutionary history of species. This paper proposes a new metric, i.e. *k*th-distance, for the space of *k*th-order reduced phylogenetic networks that can be calculated in polynomial time in the size of the compared networks.

Phylogenetic networks play a vital role in the description of the evolutionary history of species, and are especially appropriate for datasets whose evolutions contain significant amounts of reticulate events caused by recombination, hybridization, horizontal gene transfer, gene duplication, gene conversion and loss[1–7]. Even for the species which have evolved based on a tree-like model of evolution, phylogenetic networks can be used to represent conflicts in phylogenetic trees that may be caused by inadequacies of an used evolutionary model. So far, there have been many algorithms and programs for constructing phylogenetic networks. The assessment of the algorithms for constructing phylogenetic networks is mainly by means of the comparison of the networks, for example, comparing the constructed network with simulate network or actual network. In addition, comparing two phylogenetic networks can help us to understand the evolutionary history of species. Recently, researchers have shown an increased interest in definition of metrics for computing the dissimilarity between a pair of phylogenetic networks.

A measure d is called a metric on a space S if it satisfies four properties: for any a, b, $c \in S$:

- $d(a, b) \geq 0$ (nonnegative);
- $d(a, b) = 0$ if and only if $a = b$ (i.e. a and b are isomorphic) (reflexivity);
- $d(a, b) = d(b, a)$ (symmetry);
- $d(a, b) + d(b, c) \geq d(a, c)$ (triangle inequality).

In general, it is much easier to prove a defined measure to satisfy the above-mentioned properties except the reflexivity. For a metric, if two phylogenetic networks are isomorphic, the distance between them computed by the metric is 0, otherwise it is 1; then we say that the metric is trivial. A trivial metric satisfies obviously above-mentioned properties, but it doesn't show other information about evolutionary history implied by the two phylogenetic networks. Accordingly, in addition to these four properties, it is desired that the metric can give us some information on the dissimilarity of the evolutionary histories expressed by the phylogenetic networks being compared[8–13].

Up to now, several metrics have been designed and proven that each one of them is a metric on a certain subspace of rooted phylogenetic networks, for example, μ-metric on the space of tree-sibling phylogenetic networks[14], the tripartition metric on the space of tree-child phylogenetic networks[15–18], the m-distance on the space of reduced phylogenetic networks[19], and the d_e-distance on the space of partly reduced phylogenetic networks[20]. The largest one among those subspace is the partly reduced phylogenetic networks, so the d_e-distance is also the metric on the subspaces of tree-child phylogenetic networks, tree-sibling phylogenetic networks and reduced phylogenetic networks. The paper will introduce a new metric, denoted by kth-distance, on space of kth-order reduced phylogenetic networks (will be discussed in the following sections), and the metric is polynomial-time computable. The space of kth-order reduced phylogenetic networks is larger subspace of rooted phylogenetic networks than any one subspace on which has been defined a metric. If no special instructions, the rest of paper will use the network to denote the rooted phylogenetic network.

[1]School of Computer Science, Inner Mongolia University, Hohhot, 010021, P.R. China. [2]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, 100044, P.R. China. Correspondence and requests for materials should be addressed to M.G. (email: guomaozu@bucea.edu.cn)

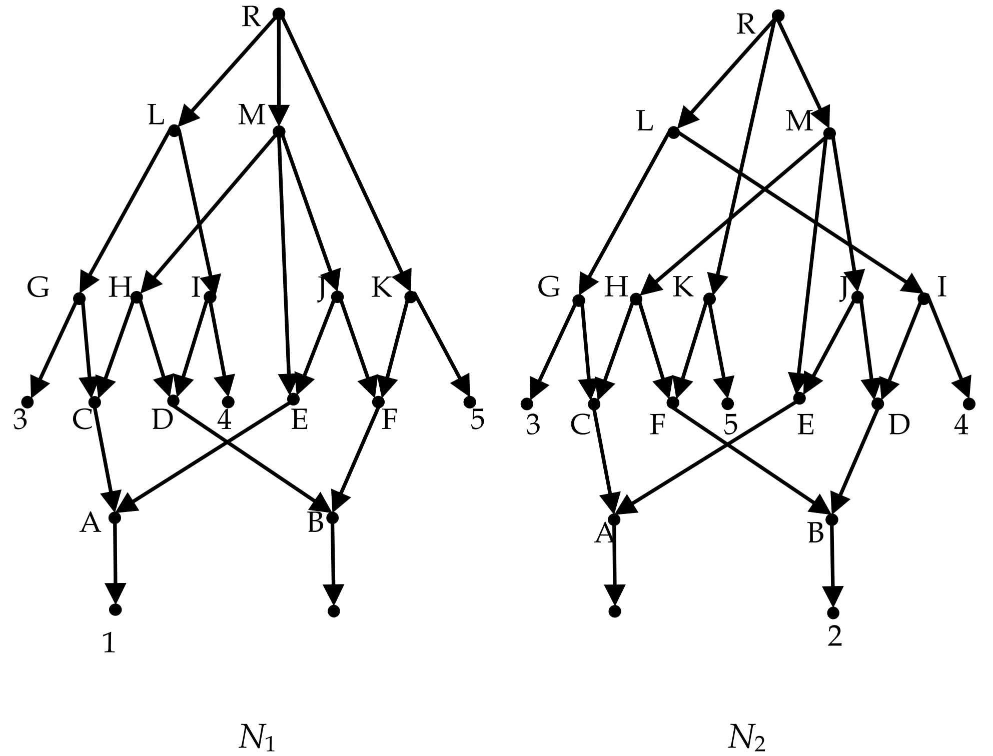

Figure 1. N_1 and N_2 are not isomorphic.

Preliminaries

Let $\mathcal{X}$ be a set of taxa. A rooted phylogenetic network $N = (V, E)$ on $\mathcal{X}$ is a directed acyclic graph (DAG for short), with one root node, and its leaves labelled as $\mathcal{X}$ by a bijection f.

For a network $N = (V, E)$ and a node $u \in V$, if:

- indeg$(u) = 0$, then u is the root;
- indeg$(u) \leq 1$, then u is a tree node;
- indeg$(u) \geq 2$, then u is a reticulate node;
- outdeg$(u) = 0$, then u is a leaf;
- outdeg$(u) \geq 1$, then u is an internal node.

Sometimes we use the notation $N = ((V, E), f)$ to denote the network N, and V_N to denote the leaf set of N. Given two nodes $u, v \in V$. If $(u, v) \in E$, then we say that v is a child of u or u is a parent of v. If there exists a directed path from u to v, then we say that v is a descendant of u or u is an ancestor of v.

The height of a node u is the length of a longest directed path beginning from u and ending with a leaf. The non-existence of cycles indicates that all nodes of N can be categorized by height: the nodes with height 0 are the leaves; for a node u with height $a > 0$, each child of u has height $m < a$ and there exists at least one child with height exactly $a - 1$.

The depth of a node v is the length of a longest directed path beginning from the root and ending with v. In the same way, the non-existence of cycles indicates that all nodes of N can be categorized by depth: the only node with depth 0 is the root; for a node v with depth $b > 0$, each parent of v has depth $m < b$ and there exists at least one parent with depth exactly $b - 1$.

Definition 1. For two networks $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$, they are isomorphic if and only if there exists a bijection H from V_1 to V_2 such that:

- (u, v) is an edge in E_1 if and only if $(H(u), H(v))$ is an edge in E_2;
- for each leaf $w \in V_1$, $f_1(w) = f_2(H(w))$.

Although the subspace defined by the d_e-distance is the largest one among all defined subspaces, there exist a large number of networks that aren't measured by the d_e-distance. For example, the two networks in Fig. 1 (from the paper[20]) are not isomorphic, while the d_e-distance between them is 0. Even for two non-isomorphic networks whose d_e-distance is not 0, the distance is usually maximal value 1. For example the networks in Fig. 2, there is a certain resemblance between them, so it is desired that the distance between them is less than 1. However, their d_e-distance is maximal value 1. On the other hand, for any two networks N_1 on $\mathcal{X}_1$ and N_2 on $\mathcal{X}_2$, the d_e-distance

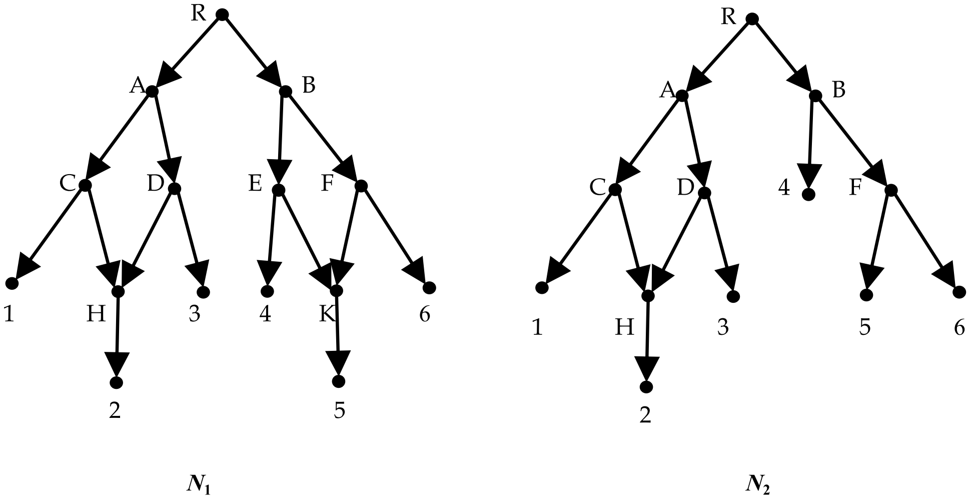

Figure 2. N_1 and N_2 on $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ are not isomorphic.

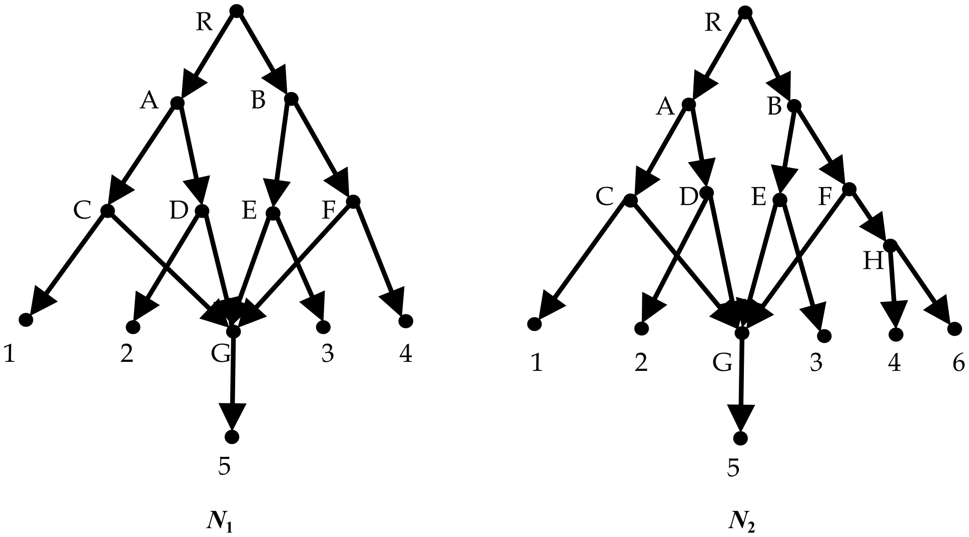

Figure 3. N_1 is on the $\mathcal{X}_1 = \{1, 2, 3, 4, 5\}$; N_2 is on the $\mathcal{X}_2 = \{1, 2, 3, 4, 5, 6\}$.

between them is 1 as long as $\mathcal{X}_1 \neq \mathcal{X}_2$. When $\mathcal{X}_1 \subset \mathcal{X}_2$, the two compared networks may share some information (see Fig. 3).

Methods

Let $N = ((V, E), f)$ be a network. Now we begin to give several definitions for the same network.

Definition 2. Two nodes $u, v \in V$ (not necessarily different) are called first-order equivalent, denoted by $u \equiv^1 v$, if

- $u, v \in V_N$ and $f(u) = f(v)$, or
- node u has $l(\geq 1)$ children $u_1, u_2, \cdots, u_l$, node v has l children $v_1, v_2, \cdots, v_l$, and $u_i \equiv^1 v_i$ for $1 \leq i \leq l$.

Example 1. Consider the network N_1 in Fig. 1. Each node of N_1 is first-order equivalent with itself, and $C \equiv^1 E$, $D \equiv^1 F$, $H \equiv^1 J$.

Definition 3. Given an even number $k \geq 2$. Two nodes $u, v \in V$ (not necessarily different) are called kth-order equivalent, denoted by $u \equiv^k v$, if $u \equiv^{k-1} v$, and:

- u, v are the root, or
- node u has $l(\geq 1)$ parents $u_1, u_2, \cdots, u_l$, node v has l parents $v_1, v_2, \cdots, v_l$, and $u_i \equiv^k v_i$ for $1 \leq i \leq l$.

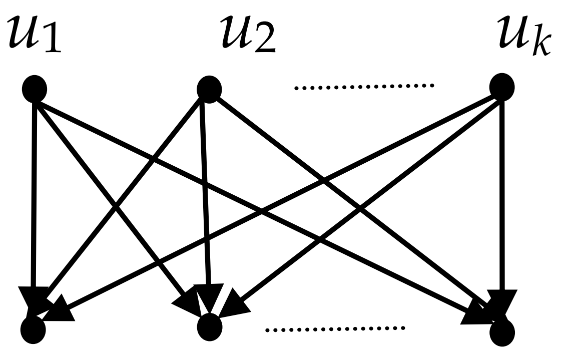

Figure 4. The topology relation of odd-order equivalent nodes.

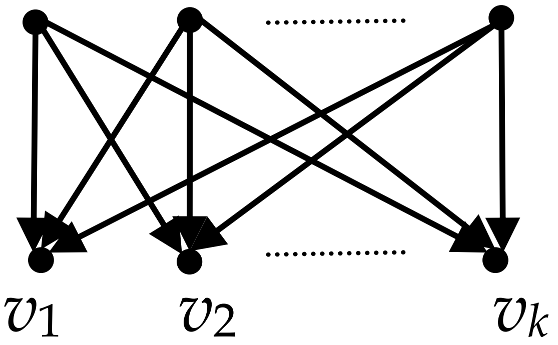

Figure 5. The topology relation of even-order equivalent nodes.

Definition 4. Given an odd number $k \geq 2$. Two nodes u, $v \in V$ (not necessarily different) are called kth-order equivalent, denoted by $u \equiv^{k} v$, if $u \equiv^{k-1} v$, and:

- u, $v \in V_N$, and $f(u) = f(v)$, or
- node u has $l(\geq 1)$ children $u_1, u_2, \cdots, u_l$, node v has l children $v_1, v_2, \cdots, v_l$, and $u_i \equiv^{k} v_i$ for $1 \leq i \leq l$.

Example 2. Consider the network N_1 in Fig. 1 again. Each node of N_1 is second-order equivalent with itself, and $\mathrm{H} \equiv^{2} \mathrm{J}$. Each node of N_1 is only kth-order equivalent with itself ($k \geq 3$).

Lemma 1. *Here k is an odd number. Given nodes $u_1, u_2, \cdots, u_s$ in a network, if each u_i has l children, and each child of u_i is only* kth-order *equivalent with itself ($1 \leq i \leq s$). Then $u_1 \equiv^{k} u_2 \equiv^{k} \cdots \equiv^{k} u_s$ if and only if $u_1, u_2, \cdots, u_s$ have the same children (refer to the* Fig. 4*).*

Lemma 2. *Here k is an even number. Given nodes $v_1, v_2, \cdots, v_s$ in a network, if each v_i has l parents, and each parent of v_i is only* kth-order *equivalent with itself. Then $v_1 \equiv^{k} v_2 \equiv^{k} \cdots \equiv^{k} v_s$ if and only if $v_1, v_2, \cdots, v_s$ have the same parents (refer to the* Fig. 5*).*

Lemma 3. *For all leaves, the root and the nodes with height 1 in a network, each of them is* kth*-order equivalent with itself (for any k).*

The proofs of Lemmas 1, 2 and 3 aren't listed here. It can be concluded from these definitions that each kth-order equivalence is an equivalence relation, i.e. it is transitive, reflexive and symmetric. It can be easily proved that all the first-order equivalent nodes have the same height and all the kth-order equivalent nodes ($k \geq 2$) have the same height and depth (refer to the literature[20]).

If a node u is kth-order equivalent with other nodes except itself, we say that u has non-trivial kth-order equivalent nodes. For a network, after deleting the non-trivial kth-order equivalent nodes of each node, as well as the nodes with indegree 1 and outdegree 1, the resulting network is called the kth-order reduced phylogenetic network. All the kth-order reduced phylogenetic networks form the space of kth-order reduced phylogenetic network. So a network N is in the space of kth-order reduced phylogenetic networks, if and only if each node of N is only kth-order equivalent with itself.

The space of first-order reduced phylogenetic networks is the space of reduced phylogenetic networks defined in the paper[19]. The space of second-order reduced phylogenetic networks is the space of partly reduced phylogenetic networks defined in the paper[20]. Figure 6 shows the relationship of these subspaces.

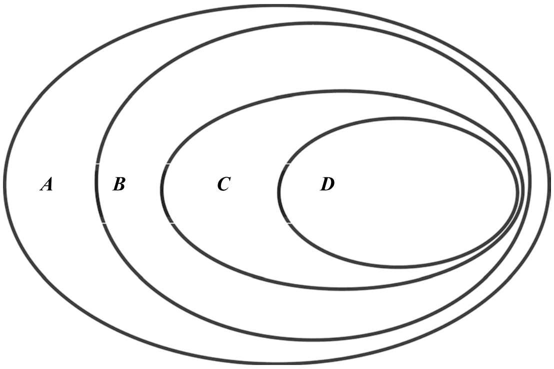

Figure 6. *A* is the space of rooted phylogenetic networks; *B* is the space of *k*th-order reduced phylogenetic networks ($k \geq 2$); *C* is the space of partly reduced phylogenetic networks; and *D* is the space of reduced phylogenetic networks.

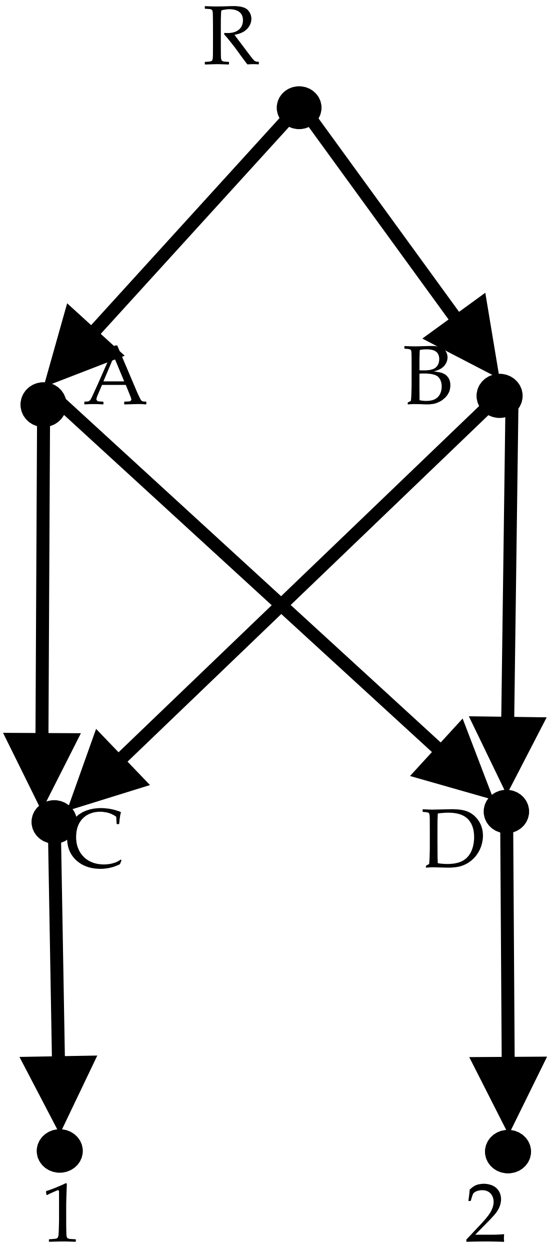

Figure 7. *N* is a rooted phylogenetic network.

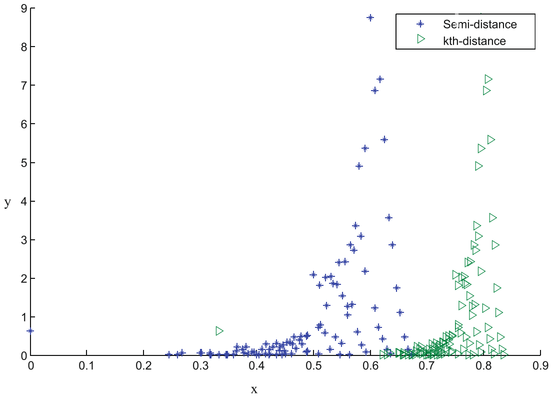

Figure 8. The results of *m*-distance and *k*th-distance.

The space of *k*th-order reduced phylogenetic networks is not equals to the space of rooted phylogenetic network. For example the network *N* in Fig. 7, for any *k*, each node of *N* is *k*th-order equivalent with itself, and $A \equiv^k B$. So *N* isn't the *k*th-order reduced phylogenetic network, i.e. not in the space of *k*th-order reduced phylogenetic networks.

In order to compute the dissimilarity of the networks, we will extend the above concepts defined in a network to two networks in the following sections. Let $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$ be two networks.

Definition 5. Two nodes $u \in V_1$, $v \in V_2$ are called first-order equivalent, denoted by $u \equiv^1 v$, if

- $u \in V_{N_1}$, $v \in V_{N_2}$, and $f_1(u) = f_2(v)$, or
- node u has $l(\geq 1)$ children $u_1, u_2, \cdots, u_l$, node v has l children $v_1, v_2, \cdots, v_l$, and $u_i \equiv^1 v_i$ for $1 \leq i \leq l$.

Definition 6. Given an even number $k \geq 2$. Two nodes $u \in V_1$, $v \in V_2$ are called *k*th-order equivalent, denoted by $u \equiv^k v$, if $u \equiv^{k-1} v$, and:

- u, v are the root, or
- node u has $l(\geq 1)$ parents $u_1, u_2, \cdots, u_l$, node v has l parents $v_1, v_2, \cdots, v_l$, and $u_i \equiv^k v_i$ for $1 \leq i \leq l$.

Definition 7. Given an odd number $k \geq 2$. Two nodes $u \in V_1$, $v \in V_2$ are called *k*th-order equivalent, denoted by $u \equiv^k v$, if $u \equiv^{k-1} v$, and:

- $u \in V_{N_1}$, $v \in V_{N_2}$ and $f_1(u) = f_2(v)$, or
- node u has $l(\geq 1)$ children $u_1, u_2, \cdots, u_l$, node v has l children $v_1, v_2, \cdots, v_l$, and $u_i \equiv^k v_i$ for $1 \leq i \leq l$.

Let u, u_0 be two nodes from two networks or the same network. From these definitions, it follows that if there exists a positive integer k_1, such that $u \not\equiv^{k_1} u_0$, then for any $k > k_1$, $u \not\equiv^k u_0$. Given two networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$. We use the following processes to compute the *k*th-order unique nodes of N_1, denoted by $L^k(N_1)$. First $L^k(N_1) = \varnothing$. Then for each node $u \in V_1$, if there has no node $u_0 \in L^k(N_1)$ such that $u \equiv^k u_0$, add u to $L^k(N_1)$. Similarly, we can compute $L^k(N_2)$. For each node $u \in L^k(N_1)$, $e^k_{N_1}(u)$ denotes the number of nodes which are *k*th-order equivalent with u, i.e. $e^k_{N_1}(u) = |\{v \in V_1: v \equiv^k u\}|$. Similarly, we can define $e^k_{N_2}(u)$ for each node $u \in L^k(N_2)$. For the sake of simplicity, we drop the subscript of e. Here $e^k(\varnothing) = 0$.

Lemma 4. *Given two networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$. For $u_1, u_2 \in V_1$, $v_1, v_2 \in V_2$, and $u_1 \equiv^k v_1$, $u_2 \equiv^k v_2$. Then, $u_1 \equiv^k u_2$ if and only if $v_1 \equiv^k v_2$.*

Proof. Refer to the proof of the Theorem 15 in the paper[20]. □

A Metric

Definition 8. For two networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$, the *k*th-distance $d_k(N_1, N_2)$ equals

$$\frac{1}{k(n_1 + n_2)}\left\{\sum_{i=1}^{k}\left[\sum_{v \in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u \in L^i(N_2)} max\{0, e^i(u) - e^i(u')\}\right]\right\} \tag{1}$$

where v' (or u') is a node in $L^i(N_2)$ (or $L^i(N_1)$) that is *i*th-order equivalent to v (or u), and if no such node exists, then $v' = \varnothing$ (or $u' = \varnothing$). n_1 and n_2 are the number of nodes in N_1 and N_2 respectively.

For each i $(1 \leq i \leq k)$, the maximal value of $\sum_{v \in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u \in L^i(N_2)} max\{0, e^i(u) - e^i(u')\}$ is $n_1 + n_2$, so the formulate 1 has maximal value 1 and minimal value 0. For a give i $(1 \leq i \leq k)$, if the value of

$\sum_{v\in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\}$ is d, then for any j $(i+1 \le j \le k)$, the value of $\sum_{v\in L^j(N_1)} max\{0, e^j(v) - e^j(v')\} + \sum_{u\in L^j(N_2)} max\{0, e^j(u) - e^j(u')\}$ is more than d.

From the definition 8, it follows that the 1st-distance is the m-distance defined in the space of reduced phylogenetic networks, and the 2nd-distance is the d_e-distance defined in the space of partly reduced phylogenetic networks.

Lemma 5. *If $d_k(N_1, N_2) = 0$. Then $|V_1| = |V_2|$, and there exists a node $v_0 \in L^i(V_2)$ for each node $v \in L^i(V_1)$, such that $v_0 \equiv^i v$ and $e^i(v_0) = e^i(v)$ $(1 \le i \le k)$.*

Proof. From $d_k(N_1, N_2) = 0$, it follows that $\sum_{v\in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} = 0$ and $\sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} = 0$ $(1 \le i \le k)$. So $max\{0, e^i(v) - e^i(v')\} = 0$ for each node $v \in L^i(N_1)$. Suppose that there exists a node $v \in L^i(N_1)$ such that $e^i(v) - e^i(v') < 0$, then $e^i(v') - e^i(v) > 0$. So $\sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} > 0$. It contradict $\sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} = 0$. Therefore, for each node $v \in L^i(N_1)$, we have $e^i(v) - e^i(v') = 0$, i.e. $e^i(v) = e^i(v')$. Similarly, for each node $u \in L^i(N_2)$, $e^i(u) = e^i(u')$. Accordingly, $|V_1| = |V_2|$. □

Lemma 6. *Given two* kth*-order reduced phylogenetic networks $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$. Then $d_k(N_1, N_2) = 0$ if and only if N_1 and N_2 are isomorphic.*

Proof. If N_1 and N_2 are isomorphic, obviously $d_k(N_1, N_2) = 0$. The converse conclusion will be proven as follows.

Lemma 5 tells us that $|V_1| = |V_2|$. From the property of the kth-order reduced phylogenetic networks, it follows that each node u in V_1 is just kth-order equivalent with itself and $u \in L^k(V_1)$. Similarly, each node v in V_2 is just kth-order equivalent with itself and $v \in L^k(V_2)$. Moreover, for each node $u \in V_1$, there exists the only one node $v \in V_2$ such that $u \equiv^k v$. So we define a mapping H from V_1 to V_2, for each node $u \in V_1$, $H(u) = u'$, where $u' \in V_2$ and $u' \equiv^k u$.

First we prove that the mapping H is a bijection. For any two different nodes $u_1, u_2 \in V_1$, there exist two nodes $u_1', u_2' \in V_2$, such that $H(u_1) = u_1'$ and $H(u_2) = u_2'$. Here u_1' and u_2' are not the same nodes. If not, then $u_1 \equiv^k u_2$. It contradict that each node $u \in V_1$ is just kth-order equivalent with itself. So H is injective. Due to $|V_1| = |V_2|$, we have that H is a surjection.

Then we prove that if $(u, v) \in E_1$, then $(H(u), H(v)) \in E_2$. Let $u_0 = H(u)$ and $v_0 = H(v)$, i.e. $u_0 \equiv^k u$ and $v_0 \equiv^k v$. If k is an odd number, then the children of u are kth-order equivalent with the children of u_0 respectively. Thus, v is kth-order equivalent with a child v' of u_0, i.e. $v' \equiv^k v \equiv^k v_0$. Since every node is only kth-order equivalent with itself, v' and v_0 are the same nodes, i.e. v_0 is a child of u_0. Therefore, $(u_0, v_0) \in E_2$. Similarly, we can come to the conclusion when k is an even number.

The mapping H also preserves the labels of the leaves from the definition of kth-order equivalence. In conclusion, N_1 and N_2 are isomorphic.

Lemma 7. *For any one pair of networks N_1 and N_2, $d_k(N_1, N_2) = d_k(N_2, N_1)$.*

The distance $d_k(N_1, N_2)$ can be viewed as the symmetric difference of the same set of elements $\bigcup_{i=1}^{k}\{L^i(N_1) \cup L^i(N_2)\}$. From the property of the symmetric difference[21], it follows that the following triangle inequality holds:

Lemma 8. *For any three networks N_1, N_2 and N_3, $d_k(N_1, N_2) + d_k(N_2, N_3) \ge d_k(N_1, N_3)$.*

From Lemmas 6, 7 and 8, we have the following result:

Theorem 9 *The* kth*-distance defined by the formula 1 is a metric on the space of* kth*-order reduced phylogenetic networks.*

Let $k = 3$ and n_j the number of nodes of network N_j $(j = 1, 2)$. Consider the two networks in Fig. 1. For $i = 1$ and 2, $\sum_{v\in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} = 0$. For $i = 3$, $\sum_{v\in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} = n_1 + n_2$. So the $d(N_1, N_2) = 1/3$.

Consider two networks in Fig. 2. The nodes R, B, E, F, K in V_1 don't exist first-order equivalent nodes in V_2, while the nodes R, B, F in V_2 don't exist first-order equivalent nodes in V_1. Everyone else has only one first-order equivalent node. So $\sum_{v\in L^1(N_1)} max\{0, e^1(v) - e^1(v')\} + \sum_{u\in L^1(N_2)} max\{0, e^1(u) - e^1(u')\} = 8$. For $i = 2$ and 3, every node in V_1 doesn't exist ith-order equivalent nodes in V_2. So $\sum_{v\in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} = n_1 + n_2 = 13 + 15 = 28$. Accordingly $d(N_1, N_2) = (8 + 28 + 28)/(3 \times 28) = 16/21$.

Consider two networks in Fig. 3. The nodes R, B, F in V_1 don't exist first-order equivalent nodes in V_2, and the nodes $R, B, F, H, 6$ in V_2 don't exist first-order equivalent nodes in V_1. Everyone else has only one first-order equivalent with node. So $\sum_{v\in L^1(N_1)} max\{0, e^1(v) - e^1(v')\} + \sum_{u\in L^1(N_2)} max\{0, e^1(u) - e^1(u')\} = 8$. For $i = 2$ and 3, every node in V_1 doesn't exist ith-order equivalent nodes in V_2. So $\sum_{v\in L^i(N_1)} max\{0, e^i(v) - e^i(v')\} + \sum_{u\in L^i(N_2)} max\{0, e^i(u) - e^i(u')\} = n_1 + n_2 = 13 + 15 = 28$. Accordingly $d(N_1, N_2) = (8 + 28 + 28)/(3 \times 28) = 16/21$.

Lemma 10. *If there is $d_k(N_1, N_2) = 0$ for all k. Then there exists a positive integer m, such that for any $m_0 \geq m$, we have that each node u in V_1 has a m_0th-order equivalent node u′ in V_2.*

Proof. Assume that the above conclusion does not hold, i.e. for any positive integer m, there exist $k_0 \geq m$ and a node $u \in V_1$, such that $u' \not\equiv^{k_0} u$ for any node $u' \in V_2$. So when $m = 1$, there exist k_1 and $u_1 \in V_1$, such that $u_1 \not\equiv^{k_1} u'$ for any node $u' \in V_2$. So $d_{k_1}(N_1, N_2) \neq 0$. This conclusion is in contradiction with $d_k(N_1, N_2) = 0$ for all k. □

Computational Aspects. For odd number k (or even number k), the kth-order equivalent nodes can be computed by a bottom-up (or top-down) approach, no matter whether the nodes are in the same network or two different networks. Given two networks $N_1 = ((V_1, E_1), f_1)$ and $N_2 = ((V_2, E_2), f_2)$. Algorithm 8 shows the pseudocode that decides whether two nodes are kth-order equivalent or not, where E(k) is the abbreviation for the set of kth-order equivalent nodes. This process will cost at most $O(n^3)$ time, where $n = \max(|V_1|, |V_2|)$. Therefore, it takes totally at most $O(n^5)$ time to find out all ith-order (where $1 \leq i \leq k$) equivalent nodes for each node of the two networks. Computing the formula 1 will costs $O(n)$ time. In conclusion, we will spend $O(n^5)$ time in computing the kth-distance between two networks, where n is the maximum of $|V_1|$ and $|V_2|$.

Results

We compared the kth-distance with m-distance on the space of reduced phylogenetic networks[19] and the d_e-distance on the space of partly reduced phylogenetic networks[20], by means of 100 networks constructed by the LNETWORK method[3]. Thus, each distance method can obtain a distance matrix with approximately 5000 values.

Algorithm 1: Deciding whether u and v are kth-order equivalent or not for an odd number k (or an even number k).

```
input: nodes $u$ and $v$
if outdegree of $u$ is not equals to that of $v$ (or indegree of $u$ is not equals to that of $v$) then
    return
end if
if $u$ and $v$ are leaves and they have the same labels (or $u$ and $v$ are the root) then
    add $v$ to E($k$) of $u$
    add $u$ to E($k$) of $v$
else
    flag := false
    if E($k-1$) of $u$ doesn't contain $v$ then
        return
    end if
    for each child $a$ of $u$ (or each parent $a$ of $u$) do
        for each child $b$ of $v$ (or each parent $b$ of $v$) do
            if $b.label$ = true then
                continue
            end if
            if the E($k$) 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 E($k$) of $u$
    add $u$ to E($k$) of $v$
end if
```

Figure 8 shows the distribution of the distance values, where the horizontal axis is the distance value and the vertical axis is the percent of the distance value in all values. Here the results of d_e-distance didn't show in Fig. 8, because it just has two distance values 1 and 0, and 99.38 percent and 0.62 percent respectively. The minima of m-distance and the d_e-distance are 0, while the minimum of kth-distance is 0.32.

From the results, we reached the following conclusions. First, almost all d_e-distance values are maximum values 1. Second, the kth-distance values are not 0 between the networks whose d_e-distance and the m-distance values are 0. Third, the kth-distance values are larger than the m-distance values for the same networks.

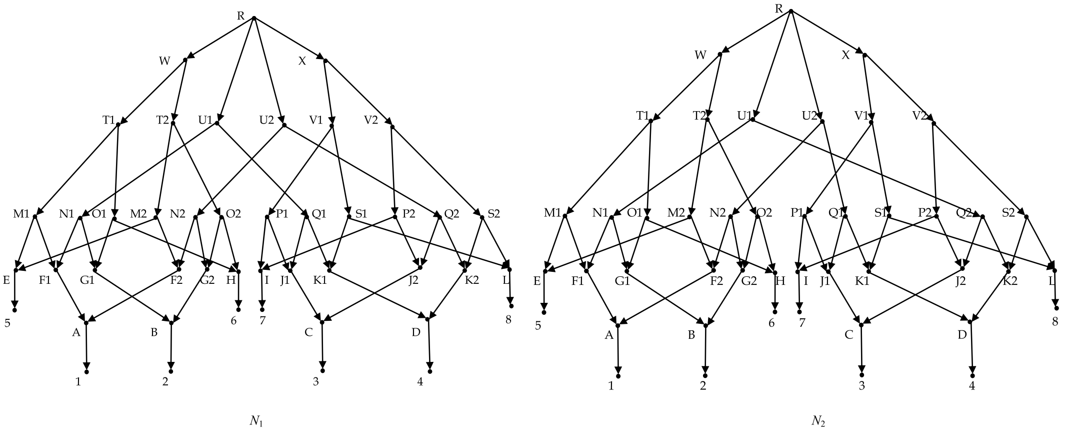

Figure 9. Two networks are not isomorphic.

Discussion

In order to compare dissimilarity for more phylogenetic networks, we define a polynomial-time computable metric on the space of *k*th-order reduced phylogenetic networks. Here the larger *k* is, the larger the space of *k*th-order reduced phylogenetic networks is. Moreover, the larger *k* is, the more precise the distance between two phylogenetic networks is. Take the non-isomorphism networks in Fig. 1 for example. When $k = 1$ or 2, the value computed by the formula 1 is 0, i.e. their *m*-distance and d_e-distance are 0. However, when $k = 3$, the value computed by the formula 1 is 1/3. So when $k = 1$ or 2, the value computed by the formula 1 doesn't indicate the real dissimilarity between the two networks. The choose of *k* in general is based on the desired precision of distance. Whatever *k* is, the *k*th-distance is not a metric on the space of all rooted phylogenetic networks. For example, the two phylogenetic networks in Fig. 9, their *k*th-distance is 0, but they are not isomorphic.

References

1. Pagel, M. Inferring the Historical Patterns of Biological Evolution. *Nature* **401**, 877–884 (1999).
2. Wang, J. A new algorithm to construct phylogenetic networks from trees. *Genetics and Molecular Research* **13**, 1456–1464 (2014).
3. Wang, J. *et al*. LNETWORK: an efficient and effective method for constructing phylogenetic networks. *Bioinformatics* **29**, 2269–2276 (2013).
4. Wang, J. *et al*. BIMLR: A Method for Constructing Rooted Phylogenetic Networks from Rooted Phylogenetic Trees. *Gene* **527**, 344–351 (2013).
5. Zou, Q. *et al*. Survey of MapReduce frame operation in bioinformatics. *Briefings in Bioinformatics* **15**, 637–647 (2013).
6. Zou, Q. *et al*. HAlign: Fast Multiple Similar DNA/RNA Sequence Alignment Based on the Centre Star Strategy. *Bioinformatics* **31**, 2475–2481 (2015).
7. Zou, Q. *et al*. Similarity computation strategies in the microRNA-disease network: A Survey. *Briefings in Functional Genomics* **15** (2015).
8. Wang, J. *et al*. FastJoin, an improved neighbor-joining algorithm. *Genetics and Molecular Research* **11**, 1909–1922 (2012).
9. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
10. Critchlow, D. E. *et al*. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology* **45**, 323–334 (1996).
11. Waterman, M. S. & Smith, T. F. On the similarity of dendograms. *J. Theor. Biol.* **73**, 789–800 (1978).
12. Bluis, J. & Shin, D. G. Nodal distance algorithm: Calculating a phylogenetic tree comparison metric, *in Proc. 3rd IEEE Symp. BioInformatics and BioEngineering*, pp. 87–94 (2003).
13. Huber, K. *et al*. Metrics on Multilabeled Trees: Interrelationships and Diameter Bounds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**, 1029–1040 (2011).
14. Cardona, G. *et al*. A Distance Metric for a Class of Tree-Sibling Phylogenetic Networks. *Bioinformatics* **24**, 1481–1488 (2008).
15. Nakhleh, L. *et al*. Towards the Development of Computational Tools for Evaluating Phylogenetic Network Reconstruction Methods, *Proc. Eighth Pacific Symp. Biocomputing*, pp. 315–326 (2003).
16. Moret, B. *et al*. Phylogenetic networks: modeling, reconstructibility and accuracy, *IEEE/ACM Trans. Computational Biology and Bioinformatics* **1**, 13–23 (2004).
17. Baroni, M. *et al*. A Frame work for Representing Reticulate Evolution. *Annals of Combinatorics* **8**, 391–408 (2004).
18. Cardona, G. *et al*. Tripartitions Do Not Always Discriminate Phylogenetic Networks. *Math. Biosciences* **211**, 356–370 (2008).
19. Nakhleh, L. A metric on the space of reduced phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**, 218–222 (2010).
20. Wang, J. A Metric on the Space of Partly Reduced Phylogenetic Networks. *BioMed Research International* 1–10 (2016).
21. Cardona, G. *et al*. On Nakhleh's Metric for Reduced Phylogenetic Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **6**, 629–638 (2009).

Acknowledgements

The work was supported by the Natural Science Foundation of Inner Mongolia province of China (2015BS0601); the National Natural Science Foundation of China (61661040, 61571163, 61532014, 61671189); the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902).

Author Contributions

J.W. devised the metric, proved it and wrote the paper. M.G. designed the experiments and revised the paper.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.