BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 13 2008, pages 1481–1488
doi:10.1093/bioinformatics/btn231

Phylogenetics

A distance metric for a class of tree-sibling phylogenetic networks

Gabriel Cardona[1,*], Mercè Llabrés[1], Francesc Rosselló[1] and Gabriel Valiente[2]

[1]Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma de Mallorca and [2]Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain

Received on March 19, 2008; revised and accepted on May 11, 2008

Advance Access publication May 12, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The presence of reticulate evolutionary events in phylogenies turn phylogenetic trees into phylogenetic networks. These events imply in particular that there may exist multiple evolutionary paths from a non-extant species to an extant one, and this multiplicity makes the comparison of phylogenetic networks much more difficult than the comparison of phylogenetic trees. In fact, all attempts to define a sound distance measure on the class of all phylogenetic networks have failed so far. Thus, the only practical solutions have been either the use of rough estimates of similarity (based on comparison of the trees embedded in the networks), or narrowing the class of phylogenetic networks to a certain class where such a distance is known and can be efficiently computed. The first approach has the problem that one may identify two networks as equivalent, when they are not; the second one has the drawback that there may not exist algorithms to reconstruct such networks from biological sequences.

Results: We present in this article a distance measure on the class of *semi-binary tree-sibling time consistent* phylogenetic networks, which generalize tree-child time consistent phylogenetic networks, and thus also galled-trees. The practical interest of this distance measure is 2-fold: it can be computed in polynomial time by means of simple algorithms, and there also exist polynomial-time algorithms for reconstructing networks of this class from DNA sequence data.

Availability: The Perl package `Bio::PhyloNetwork`, included in the BioPerl bundle, implements many algorithms on phylogenetic networks, including the computation of the distance presented in this article.

Contact: gabriel.cardona@uib.es

Supplementary information: Some counterexamples, proofs of the results not included in this article, and some computational experiments are available at *Bioinformatics* online.

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 (Pagel, 1999). Although phylogenetic trees have been used since the early days of phylogenetics (Burkhardt and Smith, 1987) to represent evolutionary histories under mutation, it is currently well known that the existance of genetic recombinations, hybridizations and lateral gene transfers makes species evolve more in a reticulate way than in a simple, arborescent way (Doolittle, 1999).

Now, as it happens in the case of phylogenetic trees, given a set of operational taxonomic units, different reconstruction algorithms, or different sets of sampled data, may lead to different reticulate evolutionary histories. Thus, a well-defined distance measure for phylogenetic networks becomes necessary.

In a completely general setting, a phylogenetic network is simply a directed acyclic graph whose leaves (nodes without outgoing edges) are labeled by the species they represent (Strimmer and Moulton, 2000; Strimmer *et al.*, 2001). However, this situation is so general that even the problem of deciding when two such graphs are isomorphic is computationally hard. Hence, one has to put additional constraints to narrow down the class of phylogenetic networks. There have been different approaches to this problem in the literature, giving rise to different definitions of phylogenetic network (Bandelt, 1994; Huson, 2006, 2007; Linder *et al.*, 2003; Semple, 2007; Strimmer and Moulton, 2000; Strimmer *et al.*, 2001).

In this article, we give a distance measure on the class of *semi-binary tree-sibling time consistent* phylogenetic networks. This class first appeared in Nakhleh's thesis (Nakhleh, 2004), and it is of special interest because there exist algorithms to reconstruct phylogenetic networks of this class from the analysis of biological sequences (Jin *et al.*, 2006, 2007a). However, all previous attempts to provide a sound distance measure on this class of networks have failed (Cardona *et al.*, 2008b).

2 SEMI-BINARY TREE-SIBLING TIME CONSISTENT PHYLOGENETIC NETWORKS

Let $N=(V,E)$ be a directed acyclic graph, or DAG for short. We will say that a node u is a *tree node* if $\mathrm{indeg}(u)\leqslant 1$; moreover, if $\mathrm{indeg}(u)=0$, we will say that u is a *root* of N. If a single root exists, we will say that the DAG is *rooted*. We will say that a node u is a *hybrid node* if $\mathrm{indeg}(u)\geqslant 2$. A node u is a *leaf* if $\mathrm{outdeg}(u)=0$.

In 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 roots of the graph.

Whenever there exists a directed path (including the trivial path) from a node u to v, we will say that v is a *descendant* of u, or that u is an *ancestor* of v.

[*]To whom correspondence should be addressed.

A *phylogenetic network* on a set S of labels is a rooted DAG such that:

- No tree node has out-degree 1.
- Every hybrid node has out-degree 1, and its single child is a tree node.
- Its leaves are bijectively labeled by S.

Moreover, if all hybrid nodes have in-degree equal to 2, we will say that it is a *semi-binary phylogenetic network*. Note that semi-binarity does not impose any further condition on the out-degree of tree nodes.

We will say that two phylogenetic networks are *isomorphic* if there exists an isomorphism of DAGs between them that preserves the labels of the leaves.

The underlying motivation for such definitions is that tree nodes represent species, the leaves corresponding to extant ones, and the internal tree nodes to ancestral ones. Hybrid nodes model reticulate evolutionary events, where the parents of a hybrid node correspond to the species involved in this process, and its single child corresponds to the resulting species. Hence, the semi-binarity condition means that these events always involve two, and only two, ancestral species. This is enough to cope with most types of non-tree-like evolutionary processes: horizontal gene transfers, where a piece of DNA of an organism is transfered into the genome of another organism, hybridizations between pairs of lineages, and recombinations of genetic material between pairs of individuals within a lineage. As a matter of fact, and as Semple (2007) points out, 'vertices with indegree more than two do not represent a simultaneous exchange of genetic information between several parents but rather an uncertainty of the exact order of "hybridization"'. Moreover, most reconstruction algorithms for phylogenetic networks produce semi-binary networks in this sense (Bandelt and Dress, 1986; Bereg and Zhang, 2005; Diday and Bertrand, 1986; Jin *et al.*, 2006, 2007a).

Several other definitions of phylogenetic networks appear in the literature. In particular, some authors model as a single node the hybridization process and the resulting species; this approach leads one to consider hybrid nodes with out-degree greater than one. Notice, however, that one can easily move from one model to the other: Given a phylogenetic network according to the approach we follow here, by simply collapsing the arc from a hybrid node to its single child into a single node we get the corresponding phylogenetic network in the other model. Conversely, given a phylogenetic network with hybrid nodes of out-degree greater than one, we can split each hybrid node into a pair of nodes connected by an arc; the first one will hold all the parents of the original node (and it will be a hybrid node) and the other one will hold all its children (and it will be a tree node).

Although in real applications of phylogenetic networks, the set S labeling the leaves would correspond to a given set of taxa of extant species, for the sake of simplicity we will hereafter assume that the set of labels is simply $S=\{1,\dots,n\}$.

We will say that two nodes u and v are *siblings* of each other if they share a parent. Note that the relation of being siblings is reflexive and symmetric, but not transitive.

We will say that a tree node v is *quasi-sibling* of another tree node u if the parent of v is a hybrid node that is also a sibling of u (Fig. 1).[1] The relation of being quasi-siblings is neither reflexive nor symmetric.

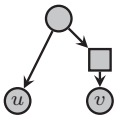

Fig. 1. Node v is quasi-sibling of u.

We will say that a phylogenetic network is *tree-sibling* if each hybrid node has at least one sibling that is a tree node.

Biologically, this condition means that for each of the reticulation events, at least one of the species involved in it has also some descendant through mutation.

A *time assignment* on a network $N=(V,E)$ is a mapping $\tau: V\to\mathbb{N}$ such that:

(1) $\tau(r)=0$, where r is the root of N.

(2) If v is a hybrid node and $(u,v)\in E$, then $\tau(u)=\tau(v)$.

(3) If v is a tree node and $(u,v)\in E$, then $\tau(u)<\tau(v)$.

We will say that a network is *time consistent* if it admits a time assignment (Baroni *et al.*, 2006).

From a biological point of view, a time assignment represents the time when a certain species exists, or a certain hybridization process occurs. Note that whenever such a process takes place, the species involved must coexist; this is what the time-consistency property ensures.

By a *sbTSTC network* we will mean a semi-binary tree-sibling, time consistent phylogenetic network, and this will be the class of phylogenetic networks that we will consider in the rest of the article.

Remark. In our previous paper (Cardona *et al.*, 2007) we considered the class of *tree-child* phylogenetic networks, that is, those networks where every internal node has at least one tree child. The tree-child condition is clearly more restrictive than the tree-sibling one. For instance, the network in Figure 2 is not tree-child, since all the children of v are hybrid, while it is tree-sibling, since the hybrid nodes A and B have respective sibling nodes 1 and 4. However, the additional condition of time consistency that we impose here makes that neither of the two considered classes (tree-child and sbTSTC) is contained in the other. For instance, the network in (Cardona *et al.*, 2007, Fig. 4) is tree-child (and hence tree-sibling) but it is not time consistent.

Remark. Besides the biological considerations we have made while presenting our assumptions on phylogenetic networks, these are also motivated by the fact that we want to single out phylogenetic networks by means of their μ-representation (see Section 3 below). In the Supplementary Material we give examples showing that the technical conditions imposed on phylogenetic networks are necessary to achieve this goal. We also show there how the definition of sbTSTC phylogenetic network we have given is related to that given in (Nakhleh, 2004), showing that the latter is more restrictive than the former.

The following result ensures the existence of sibling or quasi-sibling leaves in sbTSTC networks.

[1] Henceforth, in graphical representations of phylogenetic networks, hybrid nodes are represented by squares, tree nodes by circles and indeterminate nodes (that is, that can be either tree or hybrid nodes) by hexagons.

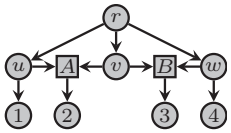

Fig. 2. A sbTSTC phylogenetic network.

LEMMA 1. *Let N be a sbTSTC network with more than one leaf. Then, there exists at least one pair of leaves that are either siblings or quasi-siblings.*

PROOF. Let M be the set of internal nodes of N with maximal time assignment, and note that M is non-empty, since otherwise N would be reduced to a single leaf.

If no node of M is hybrid, let $u \in M$ be any tree node. Then, all its children are leaves: indeed, if a child of u were an internal tree node, then its time assignment would be strictly greater than that of u, against our assumption; also, if a child of u were a hybrid node, then its time assignment would be the same as that of u, and hence M would contain a hybrid node. Therefore, since we do not allow out-degree 1 tree nodes, the node u has at least two children that are leaves, and these leaves are siblings.

If M contains a hybrid node v, then its parents are tree nodes u, u' with the same time assignment as that of v, and at least one of them must have a tree child because of the tree-sibling property. Say that u has a tree child; the same argument as before proves that this child must be a leaf i. Moreover, the single child of v must be a tree node, hence also a leaf j. In this situation, j is a quasi-sibling of i. ■

We give now tight bounds for the number of hybrid and internal tree nodes of a sbTSTC phylogenetic network, depending on its number of leaves. The existence of such bounds implies, in particular, that there exists a finite number of sbTSTC phylogenetic networks on a given set of taxa up to isomorphisms. Nevertheless, we have not yet been able to find a closed expression for this number of networks depending only on the number of leaves. Table 1 shows the experimental results we have found in this direction using the procedure described in Section 6.

PROPOSITION 2. *Let N be a sbTSTC network. Let n, h, t be, respectively, the number of leaves, the number of hybrid nodes and the number of internal tree nodes of N. If $n \leqslant 2$, then $h = 0$ and $t = n-1$. Otherwise, $h \leqslant 2n-4$ and $t \leqslant 3n-6$.*

PROOF. See the Supplementary Material. ■

The bounds in the proposition above are tight, as the following example shows.

Example 1. Consider the family of sbTSTC phylogenetic networks $(N_n)_{n \geqslant 3}$ defined recursively in the following way:

- N_3 is the first phylogenetic network depicted in Figure 4.
- The network N_{n+1} is obtained from N_n by applying the transformation described in Figure 3. Figure 4 depicts also N_4 and N_5, where we label the internal nodes in these networks to ease understanding of the construction.

Note that all networks N_n are semi-binary and tree-sibling by construction. Also, the time consistency property can be easily verified: when constructing N_{n+1} from N_n, we can assign to each of the internal nodes introduced the maximum of the times that the leaves $1, 2, n$ have in N_n, and reassign to the leaves $1, 2, n, n+1$ this maximum plus one.

Now, N_3 has three internal tree nodes and two hybrid nodes, and the construction of N_{n+1} from N_n adds three internal tree nodes and two hybrid nodes. It is evident, then, that each N_n has $3(n-2)$ internal tree nodes and $2(n-2)$ hybrid nodes.

Table 1. Number of sbTSTC networks for small number n of leaves

n	1	2	3	4	5
Number of networks	1	1	10	444	61 176

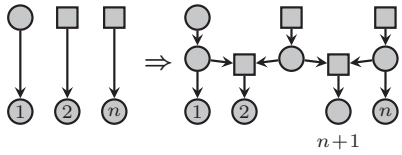

Fig. 3. The transformation that produces N_{n+1} from N_n.

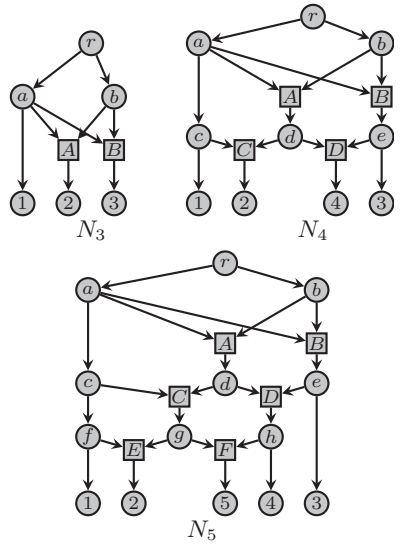

Fig. 4. sbTSTC phylogenetic networks with maximum number of nodes, with 3, 4 and 5 leaves.

3 THE μ-REPRESENTATION

In Cardona *et al.* (2007) we introduced the μ-representation for the class of tree-child phylogenetic networks. In this section we review the definition of the μ-representation of phylogenetic networks, and we will prove later that this representation characterizes a sbTSTC phylogenetic network, up to isomorphism.

Let $N = (V, E)$ be a phylogenetic network on the set $S = \{1, \ldots, n\}$. For each node u of N, we consider its *μ-vector*,

$$\mu(u) = (m_1(u), \ldots, m_n(u)),$$

where $m_i(u)$ is the number of different paths from u to the leaf i. Moreover, we define the μ-representation of N, $\mu(N)$, as the multiset

$$\mu(N)=\{\mu(u)\,|\,u\in V\},$$

with each element appearing as many times as the number of different nodes having it as its μ-vector. On $\mu(N)$ we will consider the ordering induced by the product partial order on $\mathbb{N}^n$; that is, $(m_1,\ldots,m_n)\geqslant(m'_1,\ldots,m'_n)$ if and only if, $m_i\geqslant m'_i$ for each $i=1,\ldots,n$.

For each leaf i, we have that its μ-vector is $\mu(i)=\delta(i)$, with $\delta(i)$ the vector with 0 at each position, except at its i-th position, where it is 1. If u is not a leaf, we have that $\mu(u)=\sum_{v_k}\mu(v_k)$, where the sum ranges over the set of children of u (Cardona *et al.*, 2007, Lemma 4). This property allows for the computation of $\mu(N)$ in polynomial time (see Section 6 below).

Example 2. Consider the sbTSTC phylogenetic network in Figure 2. In Table 2 we give its μ-representation, except for the leaves, whose μ-vector is trivial.

In the next section we will introduce a set of reduction procedures for sbTSTC phylogenetic networks. It will turn out that the application conditions for these procedures can be read from the μ-representation of the network.

LEMMA 3. *Let N be a sbTSTC phylogenetic network, i,j a pair of leaves, and let u be the parent of i. Then j is sibling or quasi-sibling of i if, and only if:*

1. $\mu(u)$ is the least element in the set
$$M=\{\mu\in\mu(N)\,|\,\mu\geqslant\delta(i)+\delta(j)\}.$$
2. The multiset
$$M_i=\{\mu\in\mu(N)\,|\,\mu(u)>\mu\geqslant\delta(i)\}$$
is equal to $\{\delta(i)\}$.
3. The multiset
$$M_j=\{\mu\in\mu(N)\,|\,\mu(u)>\mu\geqslant\delta(j)\}$$
is equal to $\{\delta(j)\}$ (when j is sibling of i) or to $\{\delta(j),\delta(j)\}$ (when j is quasi-sibling of i).

PROOF. Let us assume that j is sibling or quasi-sibling of i, and note that u, the parent of i, is a tree node. In either case, both i and j are descendants of u, so that $\mu(u)\in M$. Now, for any other node w with $\mu(w)\in M$, we have that $w\neq i$ and it is an ancestor of i, hence it is also an ancestor of u, and therefore $\mu(w)\geqslant\mu(u)$; hence, $\mu(u)$ is the least element in M. Moreover, the only μ-vector in M_i is $\delta(i)$, with multiplicity 1, because the only ancestor of i that is a non-trivial descendant of u is the leaf i itself. The situation for M_j is analogous, taking into account that M_j contains a second copy of $\delta(j)$ in the case that the parent of j is hybrid.

As for the converse, let us assume that for a node w, its μ-vector is the least element in M. Note that, since a hybrid node and its single child (a tree node) have the same μ-vector, we can assume that w is a tree node. Because of the definition of M, we have that w is an ancestor of both i and j. Now, if some child v of w were an ancestor of both i and j, we would have that $\mu(w)>\mu(v)\geqslant\delta(i)+\delta(j)$, against our assumption on the minimality of $\mu(w)$ in M. Therefore, w has two children v_i,v_j such that v_i is ancestor of i (but not of j) and v_j is ancestor of j (but not of i). Then, $\mu(v_i)\in M_i$ and, by the uniqueness of the element in M_i, we have that $v_i=i$, and it follows that w is the parent of i, that is, $w=u$. Symmetrically, we have that $v_j\in M_j$. Now, two situations may arise: first, if the multiplicity of $\delta(j)$ in M_j is one, then $v_j=j$ and j is a sibling of i; second, if this multiplicity is two, then v_j must be a hybrid node whose single child is j, hence j is quasi-sibling of i. ■

Table 2. μ-representation of the network in Figure 2

node	μ-vector
r	$(1,2,2,1)$
u	$(1,1,0,0)$
v	$(0,1,1,0)$
w	$(0,0,1,1)$
A	$(0,1,0,0)$
B	$(0,0,1,0)$

LEMMA 4. *Let N be a sbTSTC phylogenetic network. Let j be a leaf sibling or quasi-sibling of another leaf i, and let u be the parent of i. Then,* $\mathrm{outdeg}(u)=2$ *if, and only if,* $\mu(u)=\delta(i)+\delta(j)$.

PROOF. Note that with the assumptions made, and by the previous lemma, we have that $\mu(u)\geqslant\delta(i)+\delta(j)$. Now, the equality holds if, and only if, u has no other children apart from i and j (in case that j is sibling of i) or the hybrid parent of j (in case that j is quasi-sibling of i). ■

For future reference, we gather these last results into the following proposition.

PROPOSITION 5. *Let N be a sbTSTC phylogenetic network. The following properties can be decided from the knowledge of $\mu(N)$:*

1. Two leaves are siblings, or not.
2. A leaf is quasi-sibling of another one, or not.
3. A leaf is sibling or quasi-sibling of another leaf, and the parent of the latter has out-degree 2, or greater than 2.

4 THE REDUCTION PROCEDURES

We now introduce four reduction procedures that decrease either the number of leaves or of hybrid nodes in a sbTSTC phylogenetic network. Despite the name, the reductions we introduce here have nothing to do with the reduction process introduced in Moret *et al.* (2004). As we have already mentioned, our reductions decrease either the number of leaves or hybrid nodes and they turn out to be reversible, allowing the recursive construction of all sbTSTC phylogenetic networks, while those in Moret *et al.* (2004) simplify the internal structure of the network without removing any leaf and are not reversible.

The T reduction. Let N be a sbTSTC phylogenetic network on S, i,j two sibling leaves, u their common parent, and assume that $\mathrm{outdeg}(u)>2$. The DAG $N_{T(i,j)}$ is obtained by removing from N the leaf j and its incoming arc (Fig. 5).

It is easy to check that the obtained DAG is a sbTSTC phylogenetic network on $S\setminus\{j\}$. Indeed, if the removed node j were a sibling of some hybrid node x, then i would still be a tree node

sibling of x in $N_{T(i,j)}$, hence the tree-sibling condition is preserved. Also, the time consistency and semi-binarity conditions are trivially preserved.

Note that, given $N_{T(i,j)}$, we can reconstruct N, up to isomorphism, by simply adding the leaf j and an arc from the parent of i to j.

Note also that the μ-representation of $N_{T(i,j)}$ can be easily obtained from that of N. Indeed, for any node u [except for the deleted leaf, which implies removing $\delta(j)$ from $\mu(N)$] we have that its μ-vector in the reduced network is the same as in the original network but with the j-th component removed.

The TR reduction. Let N be a sbTSTC phylogenetic network on S, i,j two sibling leaves, u their common parent, and assume that $\mathrm{outdeg}(u)=2$. Suppose also that N is not a tree with two leaves, which is equivalent to have that u is not the root of N. The DAG $N_{TR(i,j)}$ is obtained by removing from N the leaf j and its incoming arc, and collapsing the created elementary path into a single arc (Fig. 6).

As in the previous case, the resulting network is a sbTSTC phylogenetic network on $S\setminus\{j\}$. Indeed, if the node u in N is sibling of a hybrid node w, then in the obtained network $N_{TR(i,j)}$ the leaf i is a sibling of w.

Analogously to the previous case, given $N_{TR(i,j)}$, we can reconstruct N up to isomorphism by simply adding the leaf j, splitting the arc with head i by introducing an intermediate node u, and adding an arc from u to j.

Moreover, the μ-representation of $N_{TR(i,j)}$ can be easily obtained from that of N. The procedure is analogous to the previous case, taking into account that we have also to remove from $\mu(N)$ a node with μ-vector equal to $\delta(i)+\delta(j)$.

The H reduction. Let N be a sbTSTC phylogenetic network on S, j a leaf quasi-sibling of another leaf i, u the parent of i, v the parent of j, and assume that $\mathrm{outdeg}(u)>2$. The DAG $N_{H(i,j)}$ is obtained by removing from N the arc (u,v) and collapsing the resulting elementary path with intermediate node v into a single arc (Fig. 7).

Since we have only removed a hybrid node of N, when collapsing the elementary path, it is straightforward to check that the obtained DAG is a sbTSTC phylogenetic network on S.

Now, given $N_{H(i,j)}$, we can reconstruct N up to isomorphism by simply splitting the arc with head j by introducing an intermediate node v, and adding an arc from the parent of i to v.

Note that the μ-representation of $N_{H(i,j)}$ can be easily obtained from that of N. Namely, for every node x [except for the removed hybrid node, which implies removing one copy of $\delta(j)$ from $\mu(N)$] we have that if $\mu_N(x)=(m_1(x),\ldots,m_n(x))$, then $\mu_{N_{H(i,j)}}(x)=(m'_1(x),\ldots,m'_n(x))$ with

$$m'_k(x)=\begin{cases} m_k(x) & \text{if } k\neq j,\\ m_j(x)-m_i(x) & \text{if } k=j.\end{cases}$$

This follows from the fact that we have only removed the paths $x\rightsquigarrow j$ that pass through the parent of i, which are in bijection with the paths $x\rightsquigarrow i$.

The HR reduction. Let N be a sbTSTC phylogenetic network on S, j a leaf quasi-sibling of another leaf i, u the parent of i, v the parent of j, and assume that $\mathrm{outdeg}(u)=2$. The DAG $N_{HR(i,j)}$ is obtained by removing from N the arc (u,v) and collapsing the created elementary paths with respective intermediate nodes u and v into single arcs (Fig. 8).

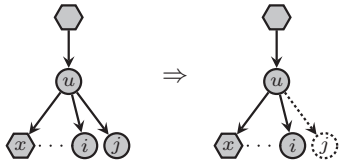

Fig. 5. The T reduction.

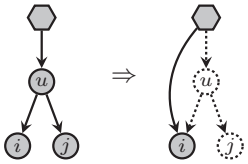

Fig. 6. The TR reduction.

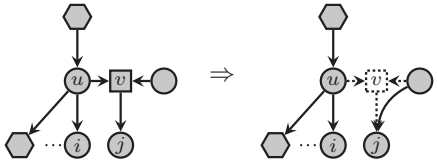

Fig. 7. The H reduction.

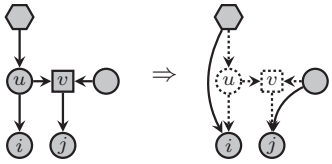

Fig. 8. The HR reduction.

The fact that the obtained DAG is a sbTSTC phylogenetic network on S follows as in the previous cases.

Also, given $N_{HR(i,j)}$, we can reconstruct N by simply splitting the arcs with respective heads i,j by introducing intermediate nodes u,v, and adding an arc from u to v.

Moreover, the μ-representation of $N_{HR(i,j)}$ can be also obtained from that of N. The procedure is the same as in the last case, taking into account that we have also to remove from $\mu(N)$ a node with μ-vector equal to $\delta(i)+\delta(j)$.

Remark. Notice that given a sbTSTC phylogenetic network, it makes sense to apply to it the inverse of any of the reductions introduced, simply following the procedure to recover a network from its reduction. However, the resulting DAG may not be a sbTSTC phylogenetic network.

Example 3. In Figure 9 we show a sequence of reduction processes that, applied to the network in Figure 2, reduce it to a tree with two leaves.

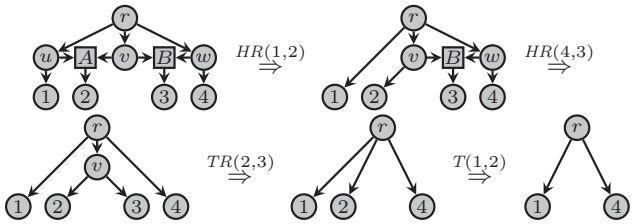

Fig. 9. Reduction processes for network in Figure 2.

Remark. The construction given in Example 1 for the networks with maximal number of nodes can also be described in terms of the reductions (or rather their inverses) we have defined. Indeed, N_{n+1} can also be described as the network obtained from N_n by application of the inverses of the reductions $TR(2,n+1)$, $HR(1,2)$ and $HR(n,n+1)$ (in this order).

5 THE μ-DISTANCE

For any pair of phylogenetic networks N_1,N_2 on the same set of leaves, let

$$d_\mu(N_1,N_2)=|\mu(N_1)\triangle\mu(N_2)|,$$

where both the symmetric difference and the cardinality operator refer to multisets.

Our main result in this article is that this mapping d_μ gives a distance on the class of sbTSTC phylogenetic networks on a given set S of taxa. We remark that d_μ is also a distance on the set of tree-child phylogenetic networks on S and, in particular, on phylogenetic trees, where it coincides with the Robinson–Foulds distance (Cardona *et al.*, 2007).

THEOREM 6. *Let N_1,N_2,N_3 be sbTSTC phylogenetic networks on the same set of taxa. Then:*

1. $d_\mu(N_1,N_2)\geqslant 0$,
2. $d_\mu(N_1,N_2)=0$ if, and only if, $N_1\cong N_2$,
3. $d_\mu(N_1,N_2)=d_\mu(N_2,N_1)$,
4. $d_\mu(N_1,N_3)\leqslant d_\mu(N_1,N_2)+d_\mu(N_2,N_3)$.

PROOF. Except for the second statement, the result follows from the properties of the symmetric difference of multisets.

Also, if N_1 and N_2 are isomorphic, it follows from the definition of the μ-representation that $\mu(N_1)$ and $\mu(N_2)$ are equal as multisets.

We will prove the separation property ($d_\mu(N_1,N_2)=0$ implies that $N_1\cong N_2$) by induction on the number n of leaves and the number h of hybrid nodes.

If $n\leqslant 2$, which implies that $h=0$, the result is obvious, since there exist only two such sbTSTC phylogenetic networks, namely the rooted trees with 1 and 2 leaves. Also, when $h=0$, the networks are, in fact, trees and the separation property of the Robinson–Foulds distance implies that $N_1\cong N_2$.

Let us assume that the result is proved for sbTSTC networks with at most $n-1\geqslant 2$ leaves, and with n leaves and at most $h-1\geqslant 0$ hybrid nodes. Let N_1, N_2 be sbTSTC phylogenetic networks with n leaves and h hybrid nodes. Because of Lemma 1 there exists a pair of leaves i,j such that j is a sibling of i (respectively, j is quasi-sibling of i) in N_1. Now since $\mu(N_1)=\mu(N_2)$, we can apply Proposition 5 to get that j is also a sibling (respectively, quasi-sibling) of i in N_2. Moreover, also from Proposition 5 it follows that the out-degree of the parent of i in N_1 is equal to 2 if, and only if, the out-degree of the parent of i in N_2 is equal to 2. From this, it follows that we can apply the same reduction to both networks; let N_1',N_2' be the networks obtained from N_1,N_2 using this reduction. Since the μ-representation of the reductions depends only on the μ-representation of the original network and the reduction procedure applied, we get that $\mu(N_1')=\mu(N_2')$. Since now N_1' and N_2' have fewer leaves or hybrid nodes than N_1 and N_2, it follows from the induction hypothesis that $N_1'\cong N_2'$. Finally, since we can recover up to isomorphisms the original networks from their reduced networks and the reductions applied, we conclude that $N_1\cong N_2$. ■

The tight bounds found in Section 2 for the number of internal nodes in a sbTSTC phylogenetic network allow us to find the *diameter* of this class of phylogenetic networks with respect to the μ-distance, that is, the maximum of the distances between two networks in this class. The interest of having a closed expression for the diameter is that it allows to normalize the μ-distance in order to take values in the unit interval $[0,1]$ of real numbers.

PROPOSITION 7. *The diameter of the class of sbTSTC phylogenetic networks with respect to d_μ is 0 when $n\leqslant 2$, 9 when $n=3$, and $10(n-2)$ when $n\geqslant 4$.*

PROOF. See the Supplementary Material. ■

As discussed before, we can now define the *normalized μ-distance* as

$$\bar{d}_\mu(N_1,N_2)=\frac{1}{10(n-2)}d_\mu(N_1,N_2)$$

if the involved networks have $n>3$ leaves, or $\bar{d}_\mu(N_1,N_2)=\frac{1}{9}d_\mu(N_1,N_2)$ if $n=3$. This way, $\bar{d}_\mu$ takes values in the interval $[0,1]$, and there exist pairs of networks at maximum normalized distance 1 for every number of leaves.

Example 4. Consider now the phylogenetic networks in Figure 10. The two networks N_1,N_2 are adapted from networks (a) and (b) in Jin *et al.* (2007b) (Fig. 10) (where we have substituted the actual names of the species by integers identifying them); we remark that the third one in the aforementioned paper and figure is isomorphic to the first one. The phylogenetic tree T depicted above is the underlying tree from which both networks are obtained by adding edges corresponding to horizontal gene transfer events. Both networks are binary and time consistent; however, the first one is tree-child (hence tree-sibling) while the second one is not tree-child, but it is tree-sibling. Also, the tree can be considered a binary tree-sibling time consistent phylogenetic network. Hence, we can compute their μ-distances, obtaining that the two networks are more similar to the underlying phylogenetic tree than to each other:

$$d_\mu(T,N_1)=22, \qquad \bar{d}_\mu(T,N_1)\approx 0.169,$$
$$d_\mu(T,N_2)=32, \qquad \bar{d}_\mu(T,N_2)\approx 0.246,$$
$$d_\mu(N_1,N_2)=38, \qquad \bar{d}_\mu(N_1,N_2)\approx 0.292.$$

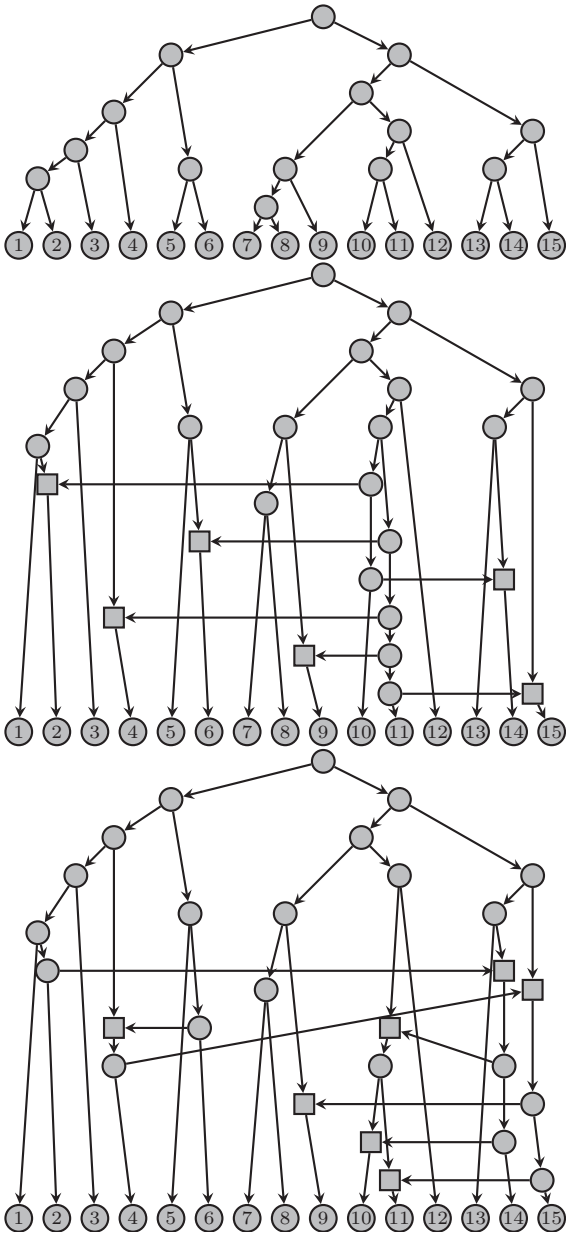

Fig. 10. Tree T (top) and networks N_1 (middle), N_2 (bottom) from (Jin *et al.*, 2007b, Fig. 10).

6 COMPUTATIONAL ASPECTS

We have already mentioned in Section 3 that the μ-representation of a phylogenetic network can be efficiently computed by means of a simple bottom-up technique. Indeed, if we define the *height* of a node as the length of the longest path starting in this node, we get a stratification of nodes. The nodes with height 0 are the leaves, and their μ-vectors are trivially computed. Assuming that we have computed the μ-vectors of nodes up to a given height h, we can compute the μ-vector of a node at height $h+1$ by simply adding up the μ-vectors of its children, which are already computed. If the network has n leaves, m nodes and the out-degree of tree nodes is bounded by $k<m$, the cost of this computation is $O(kmn)=O(m^2n)$. In order to improve the efficiency of the computations of distances below, the μ-representation of the network is stored with the μ-vectors sorted in any total order, for instance the lexicographic order; note that the computational cost of sorting the μ-representation is $O(nm\log m)$; hence, the total cost of the computation and sorting is still $O(m^2n)$.

Also, given two networks and their μ-representations, their μ-distance can be computed efficiently. We can assume that the μ-vectors of each network are sorted as explained above. Then, a simultaneous traversal of the μ-representation of both networks allows the computation of their μ-distance in $O(n(m_1+m_2))$, where m_1, m_2 are the number of nodes of each of the networks.

We have implemented the computation of the μ-representation of networks and the μ-distance between them in a Perl package (Cardona *et al.*, 2008a), part of the BioPerl bundle (Stajich *et al.*, 2002).

Note also that the reduction procedures introduced in Section 4 allow for the construction of all sbTSTC phylogenetic networks on a given set of taxa. Let $\mathcal{N}_n$ be the set of isomorphism classes of sbTSTC networks on $\{1,\ldots,n\}$, and let $\mathcal{N}_{n,h}$ be the subset of those with h hybrid nodes, so that $\mathcal{N}_n=\mathcal{N}_{n,0}\cup\ldots\cup\mathcal{N}_{n,2n-4}$, if $n>2$, or $\mathcal{N}_n=\mathcal{N}_{n,0}$ otherwise (see Lemma 2). The computation of $\mathcal{N}_1,\mathcal{N}_2$ is trivial, since both of them hold a single network. To compute $\mathcal{N}_{n,h}$, assume that $\mathcal{N}_{n-1,h}$ and $\mathcal{N}_{n,h-1}$ are already computed and initialize $\mathcal{N}_{n,h}:=\emptyset$; then:

- For each network $N'\in\mathcal{N}_{n-1,h}$, and for each leaf i of N' ($i\leqslant n-1$), construct the network N obtained by reversing the reduction $T(i,n)$ and check whether it is a sbTSTC network. If so, for each $k=1,\ldots,n-1$, let N_k be the network obtained by interchanging the leaves k and n. Set $\mathcal{N}_{n,h}:=\mathcal{N}_{n,h}\cup\{N_1,\ldots,N_n\}$.
- Repeat the step above with the reduction TR instead of T.
- For each network $N'\in\mathcal{N}_{n,h-1}$, and for each ordered pair of leaves i,j of N' ($1\leqslant i,j\leqslant n$), construct the network N obtained by reversing the reduction $H(i,j)$ and check that it is a sbTSTC network; if so, set $\mathcal{N}_{n,h}:=\mathcal{N}_{n,h}\cup\{N\}$.
- Repeat the last step with the reduction HR instead of H.

Note that since the sequence of reduction steps that reduce a given phylogenetic network to a tree with two leaves is far from being unique, repetitions within the set $\mathcal{N}_{n,h}$ must be deleted.

PROPOSITION 8. *The procedure above generates the set $\mathcal{N}_{n,h}$ of all sbTSTC phylogenetic networks (up to isomorphism) with n leaves and h hybrid nodes.*

PROOF. It is straightforward to check that each of the generated networks is a sbTSTC phylogenetic network with n leaves and h hybrid nodes. Let N be a sbTSTC phylogenetic network with $n\geqslant 2$ leaves and h hybrid nodes. Two different, non-exclusive, situations may arise:

(1) A reduction $H(i,j)$ [or $HR(i,j)$] can be applied to N, obtaining a sbTSTC network with the same set of leaves and $h-1$ hybrid nodes. Then, N is obtained (in the third or fourth step of the procedure) from a network in $\mathcal{N}_{n,h-1}$.

(2) A reduction $T(i,j)$ [or $TR(i,j)$] can be applied to N. Let N_σ be the network obtained from N by permuting the leaves j and n. Then, the reduction $T(i,n)$ [or $TR(i,n)$] can be applied

to N_σ, obtaining a network with set of leaves $\{1,\ldots,n-1\}$ and h hybrid nodes. Then, N_σ is obtained (in the first or second step of the procedure) from a network in $\mathcal{N}_{n-1,h}$. Now, interchanging the leaves j and n in N_σ we get the network N, hence it is also obtained by the described procedure. ■

The aforementioned Perl package contains a module to construct all *tree-child* phylogenetic networks on a given set of leaves. We are working on a module that generates all sbTSTC phylogenetic networks, which will be incorporated in the next release of the package. Some experimental computations are given in the Supplementary Material.

7 CONCLUSIONS

While there exist in the literature some algorithms to reconstruct sbTSTC phylogenetic networks from biological sequences, no distance metric was known on this class that is both mathematically consistent and computationally efficient. The μ-distance we have defined fulfills these two requirements, and is already implemented in a package included in the BioPerl bundle.

This μ-distance is based on the μ-representation of networks: a multiset of vectors of natural numbers, each of them associated to a node. This μ-representation could also be used to define alignments between phylogenetic networks (Cardona *et al.*, 2007, Section VI), which are useful in order to display at a glance the differences between alternative evolutionary histories of a set of species. Some results in this direction will be shortly published elsewhere.

As a by-product, we have also obtained a procedure to generate all the sbTSTC networks on a given set of taxa up to isomorphism. We are working in an efficient implementation for their generation, in order to include it in a forthcoming release of BioPerl.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions on the article.

Funding: The research described in this article has been partially supported by Spanish DGI projects MTM2006-07773 COMGRIO and MTM2006-15038-C02-01.

Conflict of Interest: none declared.

REFERENCES

Bandelt,H.-J. and Dress,A. (1986) Weak hierarchies associated with similarity measures—an additive clustering technique. *Bull. Math. Biol.*, **51**, 133–166.

Bandelt,H.-J. (1994) Phylogenetic networks. *Verh. Naturwiss. Ver. Hambg.*, **34**, 51–71.

Baroni,M. *et al.* (2006) Hybrids in real time. *Syst. Biol.*, **55**, 46–56.

Bereg,S. and Zhang,Y. (2005) Phylogenetic networks based on the molecular clock hypothesis. In *Proceeding of the Fifth IEEE International Symposium on Bioinformatic and Bioengineering (BIBE 2005)*. IEEE Computer Society, Minneapolis, MN, pp. 320–323.

Burkhardt,F. and Smith,S. (eds) (1987) *The Correspondence of Charles Darwin*. Vol. 2. Cambridge University Press, Cambridge.

Cardona,G. *et al.* (2007) Comparison of tree-child phylogenetic networks. *IEEE T. Comput. Biol.* http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70270.

Cardona,G. *et al.* (2008a) A perl package and an alignment tool for phylogenetic networks. *BMC Bioinformatics*, **9**, 175.

Cardona,G. *et al.* (2008b) Tripartitions do not always discriminate phylogenetic networks. *Math. Biosci.*, **211**, 356–370.

Diday,E. and Bertrand,P. (1986) An extension of hierarchical clustering: the pyramidal representation. In Gelsema,E. and Kanal,L. (eds) *Pattern Recognition in Practice*. NorthHolland, Amsterdam, pp. 411–424.

Doolittle,W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2128.

Huson,D.H. (2006) GCB 2006 – Tutorial: Introduction to phylogenetic networks. Tutorial presented at the German Conference on Bioinformatics (GCB'06). Available at http://www-ab.informatik.uni-tuebingen.de/research/phylonets/GCB2006.pdf (last accessed date June 1, 2008).

Huson,D.H. (2007) Split networks and reticulate networks. In Gascuel,O. and Steel,M.A. (eds) *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford University Press, Oxford, pp. 247–276.

Jin,G. *et al.* (2006) Maximum likelihood of phylogenetic networks. *Bioinformatics*, **22**, 2604–2611.

Jin,G. *et al.* (2007a) Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, **23**, 123–128.

Jin,G. *et al.* (2007b) Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Mol. Biol. Evol.*, **24**, 324–337.

Linder,C.R. *et al.* (2003) Network (reticulate) evolution: biology, models and algorithms. Tutorial presented at the Ninth Pacific Symposium on Biocomputing. Available at http://www.cs.rice.edu/~nakhleh/Papers/psb04.pdf (last accessed date June 1, 2008).

Moret,B.M.E. *et al.* (2004) Phylogenetic Networks: modeling, reconstructibility, and accuracy. *IEEE T. Comput. Biol.*, **1**, 13–23.

Nakhleh,L. (2004) *Phylogenetic Networks*. PhD Thesis. University of Texas, Austin.

Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Semple,C. (2007) Hybridization networks. In Gascuel,O. and Steel,M. (eds), *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford University Press, Oxford, pp. 277–314.

Stajich,J.E. *et al.* (2002) The BioPerl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618. http://www.bioperl.org/ (last accessed date June 1, 2008).

Strimmer,K. and Moulton,V. (2000) Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, **17**, 875–881.

Strimmer,K. *et al.* (2001) Recombination analysis using directed graphical models. *Mol. Biol. Evol.*, **18**, 97–99.