Comparison of Tree-Child Phylogenetic Networks

Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente

Abstract—Phylogenetic networks are a generalization of phylogenetic trees that allow for the representation of non-treelike evolutionary events, like recombination, hybridization, or lateral gene transfer. While much progress has been made to find practical algorithms for reconstructing a phylogenetic network from a set of sequences, all attempts to endorse a class of phylogenetic networks (strictly extending the class of phylogenetic trees) with a well-founded distance measure have, to the best of our knowledge and with the only exception of the bipartition distance on regular networks, failed so far. In this paper, we present and study a new meaningful class of phylogenetic networks, called tree-child phylogenetic networks, and we provide an injective representation of these networks as multisets of vectors of natural numbers, their path multiplicity vectors. We then use this representation to define a distance on this class that extends the well-known Robinson-Foulds distance for phylogenetic trees and to give an alignment method for pairs of networks in this class. Simple polynomial algorithms for reconstructing a tree-child phylogenetic network from its path multiplicity vectors, for computing the distance between two tree-child phylogenetic networks and for aligning a pair of tree-child phylogenetic networks, are provided. They have been implemented as a Perl package and a Java applet, which can be found at http://bioinfo.uib.es/~recerca/phylonetworks/mudistance/.

Index Terms—Phylogenetic network, tree-child phylogenetic network, phylogenetic tree, partition distance, network alignment.

1 INTRODUCTION

PHYLOGENETIC networks have been studied over the last years as a richer model of the evolutionary history of sets of organisms than phylogenetic trees, because they take not only mutation events but also recombination, hybridization, and lateral gene transfer events into account.

The problem of reconstructing a phylogenetic network with the least possible number of recombination events is NP-hard [7], [43], and much effort has been devoted to bounding the number of recombination events needed to explain the evolutionary history of a set of sequences [2], [28], [41]. On the other hand, much progress has been made to find practical algorithms for reconstructing a phylogenetic network from a set of sequences [12], [13], [25], [31], [32], [40].

Since different reconstruction methods applied to the same sequences, or a single method applied to different sequences, may yield different phylogenetic networks for a given set of species, a sound measure to compare phylogenetic networks becomes necessary [31]. The comparison of phylogenetic networks is also needed in the assessment of phylogenetic reconstruction methods [23], and it will be required to perform queries on the future databases of phylogenetic networks [35].

Many metrics for the comparison of phylogenetic trees are known, including the Robinson-Foulds metric [38], the nearest neighbor interchange metric [44], the subtree transfer distance [1], the triplet metric [10], and the nodal distance [6]. But, to our knowledge, beside the natural extension of the Robinson-Foulds metric to regular networks (whose nodes are singled out by their sets of descendant leaves) [4], only one metric (up to small variations) for phylogenetic networks has been proposed so far. It is the so-called error, or tripartition, metric, developed by Moret et al. in a series of papers devoted to the study of reconstructibility of phylogenetic networks [20], [21], [24], [25], [29], [30], [31], and which we recall in Section 2.4. Unfortunately, it turns out that, even in its strongest form [25], this error metric never distinguishes all pairs of phylogenetic networks that, according to its authors, are distinguishable: see [9] for a discussion of the error metric’s downsides.

The main goal of this paper is to introduce a metric on a restricted, but meaningful, class of phylogenetic networks: the tree-child phylogenetic networks. These are the phylogenetic networks where every nonextant species has some descendant through mutation. This is a slightly more restricted class of phylogenetic networks than the tree-sibling ones (see Section 2.3), where one of the versions of the error metric was defined. Tree-child phylogenetic networks include galled trees [12], [13] as a particular case, and they have been recently proposed as the class where meaningful phylogenetic networks should be searched [47].

We prove that each tree-child phylogenetic network with \( n \) leaves can be singled out, up to isomorphisms, among all tree-child phylogenetic networks with \( n \) leaves by means of a finite multisubset of \( \mathbb{N}^n \). This multiset of vectors consists of the path multiplicity vectors, or \( \mu \)-vectors, in short, \( \mu(v) \) of all nodes \( v \) of the network: For every node \( v \), \( \mu(v) \) is the vector listing the number of paths from \( v \) to each one of the leaves of the network. We present a simple polynomial time algorithm for reconstructing a tree-child phylogenetic network from the knowledge of this multiset.
This injective representation of tree-child phylogenetic networks as multisets of vectors of natural numbers allows us to define a metric on any class of tree-child phylogenetic networks with the same leaves as simply the symmetric difference of the path multiplicity vectors multisets. This metric, which we call $\mu$-distance, is a true distance on this class of networks, in the sense that it satisfies the axioms of distances, including the separation axiom (non-isomorphic phylogenetic networks are at nonzero distance) and the triangle inequality. Moreover, it extends the Robinson-Foulds metric for phylogenetic trees. Actually, and more in general, it extends to the whole class of tree-child phylogenetic networks the bipartition distance defined on the class of regular tree-child phylogenetic networks. This class of regular tree-child phylogenetic networks includes the tree-child phylogenetic networks without outdegree 1 nodes and where no node can hybridize with a descendant of it [9].

The properties of the path multiplicity representation of tree-child phylogenetic networks allow us also to define an alignment method for them. Our algorithm outputs an injective matching from the network with less nodes into the other network that minimizes in some specific sense the difference between the $\mu$-vectors of the matched nodes. Although several alignment methods for phylogenetic trees are known [27], [34], [35], [37], this is to our knowledge the first one that can be applied to a larger class of phylogenetic networks.

We have implemented our algorithms to recover a tree-child phylogenetic network from its path multiplicity representation and to compute the $\mu$-distance, together with other related algorithms (like, for instance, the systematic and efficient generation of all tree-child phylogenetic networks with a given number of leaves), in a Perl package, which can be found at http://bioinfo.uib.es/~recerca/phylonetworks/mudistance/. We have also implemented our alignment method as a Java applet, which can be run interactively at the aforementioned web page. The package and Web interface are both described in [8].

## 2 Preliminaries

### 2.1 Directed Acyclic Graphs (DAGs)

Let $N = (V, E)$ be a DAG. We denote by $di(u)$ and $do(u)$ the indegree and outdegree, respectively, of a node $u \in V$.

A node $u \in V$ is a leaf if $di(u) = 0$, internal if $di(u) > 0$, a root if $di(u) = 0$, a tree node if $di(u) \leq 1$, and a hybrid node if $di(u) > 1$. We denote by $V_L$, $V_T$, and $V_H$ the sets of leaves, of tree nodes, and of hybrid nodes of $N$, respectively. A DAG is said to be rooted when it has only one root.

Given an arc $(u, v) \in E$, we call the node $u$ its tail and the node $v$ its head. An arc $(u, v) \in E$ is a tree arc if $v$ is a tree node, and a hybridization arc if $v$ is hybrid. We denote by $E_T$ and $E_H$ the sets of tree arcs and of hybridization arcs, respectively.

A node $v \in V$ is a child of $u \in V$ if $(u, v) \in V$; we also say that $u$ is a parent of $v$. For every node $u \in V$, let child($u$) denote the set of its children. All children of the same node are said to be siblings of each other. The tree children of a node $u$ are its children that are tree nodes.

A DAG is binary when all its internal tree nodes have an outdegree 2, and all its hybrid nodes have indegree 2 and outdegree 1.

Let $S$ be any finite set of labels. We say that the DAG $N$ is labeled in $S$ or that it is an $S$-DAG, for short, when its leaves are bijectively labeled by elements of $S$. Two DAGs $N$, $N'$ labeled in $S$ are isomorphic, in symbols $N \cong N'$, when they are isomorphic as directed graphs and the isomorphism preserves the leaves’ labels.

In this paper, we shall always assume, usually without any further notice, that the DAGs appearing in it are labeled in some set $S$, and we shall always identify, usually without any further notice either, each leaf of a DAG with its label in $S$.

A path in $N$ is a sequence of nodes $(v_0, v_1, \ldots, v_k)$ such that $(v_{i-1}, v_i) \in E$ for all $i = 1, \ldots, k$. We say that such a path starts in $v_0$, passes through $v_1, \ldots, v_{k-1}$, and ends in $v_k$; consistently, we call $v_0$ the origin of the path, $v_1, \ldots, v_k$ its intermediate nodes and $v_k$ its end. The position of the node $v_i$ in the path $(v_0, v_1, \ldots, v_k)$ is $i + 1$. The length of the path $(v_0, v_1, \ldots, v_k)$ is $k$, and it is nontrivial if $k \geq 1$: A trivial path is, then, simply a node. We denote by $u \rightsquigarrow v$ any path with origin $u$ and end $v$.

The height of a node is the length of a longest path starting in the node and ending in a leaf.

We shall say that a path $u \rightsquigarrow v$ is contained in, or that it is a subpath of, a path $u' \rightsquigarrow v'$ when there exist paths $u' \rightsquigarrow u$ and $v \rightsquigarrow v'$ such that the path $u' \rightsquigarrow v'$ is the concatenation of the paths $u' \rightsquigarrow u$, $u \rightsquigarrow v$, and $v \rightsquigarrow v'$.

A path is elementary when its origin has outdegree 1 and all its intermediate nodes have in and outdegree 1. The relation $\geq$ on $V$ defined by

$$u \geq v \iff \text{there exists a path } u \rightsquigarrow v$$

is a partial order, called the path ordering on $N$. Whenever $u \geq v$, we shall say that $v$ is a descendant of $u$ and also that $u$ is an ancestor of $v$. For every node $u \in V$, we shall denote by $C(u)$ the set of all its descendants and by $C_L(u)$ the set of leaves that are descendants of $u$: we call $C_L(u)$ the cluster of $u$.

A node $v$ of $N$ is a strict descendant of a node $u$ if it is a descendant of it, and every path from a root of $N$ to $v$ contains the node $u$. In particular, we understand every node as a strict descendant of itself. For every node $u \in V$, we shall denote by $A(u)$ the set of all its strict descendants and by $A_L(u)$ the set of leaves that are strict descendants of $u$: we call $A_L(u)$ the strict cluster of $u$.

A tree path is a nontrivial path such that its end and all its intermediate nodes are tree nodes. A node $v$ is a tree descendant of a node $u$ when there exists a tree path from $u$ to $v$. For every node $u \in V$, we shall denote by $T(u)$ the set of all its tree descendants and by $T_L(u)$ the set of leaves that are tree descendants of $u$: we call $T_L(u)$ the tree cluster of $u$.

We recall from [9] the following easy results, which will be used several times in the next sections.

### Lemma 1

Let $u \rightsquigarrow v$ be a tree path. Then, for every other path $w \rightsquigarrow v$ ending in $v$, it is either contained in $u \rightsquigarrow v$ or it contains $u \rightsquigarrow v$. 

Corollary 1. If \( v \in T(u) \), then \( v \in A(u) \), and the path \( u \leadsto v \) is unique.

2.2 The Robinson-Foulds Metric on Phylogenetic Trees

A phylogenetic tree on a set \( S \) of taxa is a rooted tree without outdegree 1 nodes with its leaves labeled bijectively in \( S \), i.e., a rooted \( S \)-DAG with neither hybrid nodes nor outdegree 1 nodes.

Every arc \( e = (u, v) \) of a phylogenetic tree \( T = (V, E) \) on \( S \) defines a bijection on \( S \):

\[
\pi(e) = (C_L(v), S \setminus C_L(v)).
\]

Let \( \pi(T) \) denote the set of all of these bijections.

The Robinson-Foulds metric \([38]\) between two phylogenetic trees \( T \) and \( T' \) on the same set \( S \) of taxa is defined as

\[
d_{RF}(T, T') = |\pi(T) \triangle \pi(T')|,
\]

where \( \triangle \) denotes the symmetric difference of sets.

The Robinson-Foulds metric is a true distance for phylogenetic trees, in the sense that it satisfies the axioms of distances up to isomorphisms: For every phylogenetic trees \( T, T', T'' \) on the same set \( S \) of taxa, we have

1. Nonnegativity: \( d_{RF}(T, T') \geq 0 \).
2. Separation: \( d_{RF}(T, T') = 0 \) if and only if \( T \cong T' \).
3. Symmetry: \( d_{RF}(T, T') = d_{RF}(T', T) \).
4. Triangle inequality: \( d_{RF}(T, T') \leq d_{RF}(T, T'') + d_{RF}(T'', T') \).

2.3 Previous Definitions of Phylogenetic Networks

A natural model for describing an evolutionary history is a DAG whose arcs represent the parent-child relation. Such a DAG will satisfy some specific features depending on the nature and properties of this relation. For instance, if we assume the existence of a common ancestor of all individuals under consideration, then the DAG will be rooted: It will have only one root. If, moreover, the evolutionary history to be described is driven only by mutation events, and hence, every individual has only one parent, then the DAG will be a tree. In this line of thought, a phylogenetic network is defined formally as a rooted DAG with some specific features that are suited to model evolution under mutation and recombination, but the exact definition varies from paper to paper, see, for instance, \([3, 14, 15, 16, 21, 39, 41, 42]\).

For instance, Moret et al. have proposed several slightly different definitions of phylogenetic networks \([20, 21, 24, 25, 29, 30]\). To recall one of them, in \([20]\), a model phylogenetic network on a set \( S \) of taxa is defined as a rooted \( S \)-DAG \( N \), satisfying the following conditions:

1. The root and all internal tree nodes have outdegree 2. All hybrid nodes have outdegree 1, and they can only have indegree 2 (allopolyploid hybrid nodes) or 1 (autopolyploid hybrid nodes).
2. The child of a hybrid node is always a tree node.
3. There is time consistency. If \( x, y \) are two nodes for which there exists a sequence of nodes \( (v_0, v_1, \ldots, v_k) \) with \( v_0 = x \) and \( v_k = y \) such that:
   - for every \( i = 0, \ldots, k - 1 \), either \((v_i, v_{i+1})\) is an arc of \( N \), or \((v_i, v_{i+1})\) is a hybridization arc of \( N \), and
   - at least one pair \((v_i, v_{i+1})\) is a tree arc of \( N \),
   then \( x \) and \( y \) cannot have a hybrid child in common.

(6)

This time compatibility condition (condition 3) is equivalent to the existence of a temporal representation of the network \([5, 22]\): An assignment of times to the nodes of the network that strictly increases on tree arcs and so that the parents of each hybrid node coexist in time. See \([5, \text{Theorem 3}] \) or \([9, \text{Proposition 1}] \) for a proof of this equivalence.)

On the other hand, these authors define in \textit{loc. cit.} a reconstructible phylogenetic network as a rooted \( S \)-DAG, where the previous conditions are relaxed as follows: Tree nodes can have any outdegree greater than 1, hybrid nodes can have any indegree greater than 1 and any outdegree greater than 0, hybrid nodes can have hybrid children, and the time consistency need not hold any longer. Therefore, reconstructible phylogenetic networks in this sense are simply rooted DAGs with neither outdegree 1 tree nodes nor hybrid leaves. These model and reconstructible phylogenetic networks are used, for instance, in \([31]\).

A generalization of reconstructible phylogenetic networks are the hybrid phylogenies of \([4]\): rooted \( S \)-DAGs without outdegree 1 tree nodes. But, although outdegree 1 tree nodes cannot be reconstructed, they can be useful both from the biological point of view to include autopolyploidy in the model, as well as from the formal point of view, to restore time compatibility, and the impossibility of successive hybridizations in reconstructed phylogenetic networks \([25, \text{Fig. 13}] \).

In papers on phylogenetic networks, it is usual to impose extra assumptions to the structure of the network, in order to narrow the output space of reconstruction algorithms or to guarantee certain desired properties. For instance, Nakhleh imposes in his PhD dissertation \([29]\) the tree-sibling\(^1\) condition to the phylogenetic networks defined above: Every hybrid node must have at least one sibling that is a tree node. Although this condition is imposed therein to try to guarantee that the error metric considered in that work satisfies the separation axiom of distances (see Section 2.4), it has also appeared under a different characterization in some papers devoted to phylogenetic network reconstruction algorithms \([17, 18]\). Indeed, the phylogenetic networks considered in these papers are obtained by adding hybridization arcs to a phylogenetic tree by repeating the following procedure:

1. choose pairs of arcs \((u_1, v_1)\) and \((u_2, v_2)\) in the tree,
2. split the first into \((u_1, w_1)\) and \((w_1, v_1)\), with \(w_1\) a new (tree) node,
3. split the second one into \((u_2, w_2)\) and \((w_2, v_2)\), with \(w_2\) a new (hybrid) node, and
4. add a new arc \((w_1, w_2)\).

It is not difficult to prove that the phylogenetic networks obtained in this way are tree sibling.

An even stronger condition is the one imposed on galled trees \([12, 13, 43]\): No tree node has outdegree 1, all hybrid nodes have indegree 2, and no arc belongs to

\(^1\) Nakhleh uses the term \textit{class I} to refer to these networks, but for consistency with the notations, we introduce in the next section, we have renamed them here.
two recombination cycles. Here, by a recombination cycle, we mean a pair of paths with the same origin and end and no intermediate node in common. In the aforementioned papers, these galled trees need not satisfy the time compatibility condition, but in other works, they are imposed to satisfy it [29], [30], [32].

2.4 Previous Work on Metrics for Phylogenetic Networks

While many metrics for phylogenetic trees have been introduced and implemented in the literature (see, for instance, [11], [37], and the references therein), to our knowledge, only one true distance has been defined so far on a class of phylogenetic networks strictly extending the class of phylogenetic trees. This class is that of regular networks: phylogenetic networks where the mapping \( v \mapsto C(v) \) induces an isomorphism of posets between \((V, \leq)\) and \((C(v), \subseteq)\). It is clear that, on this class, the symmetric difference of the sets of clusters defines a metric that extends the Robinson-Foulds metric [4, Proposition 4.3]. A variation of this metric, on a class slightly extending a subclass of regular networks, was also presented in [9]; we shall return on it below (see Remark 5).

Besides this distance, and, again, to our knowledge, the only similarity measures for phylogenetic networks proposed so far are due to Moret et al. in the series of papers quoted in Section 2.3, where they are applied in the assessment of phylogenetic network reconstruction algorithms. We briefly recall these measures in this section.

The error, or tripartition, metric is a natural generalization to networks of the Robinson-Foulds metric for phylogenetic trees recalled in Section 2.2. The basis of this method is the representation of a network by means of the tripartitions associated to its arcs. For each arc \( e = (u, v) \) of a DAG \( N \) labeled in \( S \), the tripartition of \( S \) associated to \( e \) is

\[
\theta(e) = (A_L(v), C_L(v) \setminus A_L(v), S \setminus C_L(v)),
\]

where, moreover, each leaf \( s \) in \( A_L(v) \) and \( C_L(v) \setminus A_L(v) \) is weighted with the greatest number of hybrid nodes contained in a path from \( v \) to \( s \) (including \( v \) and \( s \) themselves). Let \( \Theta(N) \) denote the set of all these tripartition of arcs of \( N \).

In some of the aforementioned papers, the authors enrich these tripartitions with an extra piece of information. Namely, they define the reticulation scenario \( RS(v) \) of a hybrid node \( v \) with parents \( u_1, \ldots, u_k \) as the set of clusters of its parents:

\[
RS(v) = \{C_L(u_1), \ldots, C_L(u_k)\}.
\]

Then, the enriched tripartition \( \Psi(e) \) associated to an arc \( e \) is defined as \( \theta(e) \) if \( e \) is a tree arc and as the pair \( (\theta(e), RS(v)) \) if \( e \) is a hybridization arc with head \( v \). Let \( \Psi(N) \) denote the set of all these enriched tripartitions.

For every \( \Upsilon = \theta, \Psi \), the error, or tripartition, metric relative to \( \Upsilon \) between two DAGs \( N_1 = (V_1, E_1) \) and \( N_2 = (V_2, E_2) \) labeled in the same set \( S \) is defined by these authors as

\[
m_\Upsilon(N_1, N_2) = \frac{1}{2} \left( \frac{|\Upsilon(N_1) \setminus \Upsilon(N_2)|}{|E_1|} + \frac{|\Upsilon(N_2) \setminus \Upsilon(N_1)|}{|E_2|} \right).
\]

Unfortunately, and despite the word “metric,” this formula does not satisfy the separation axiom on any of the subclasses of phylogenetic networks where it is claimed to do so by the authors, and hence, it does not define a distance on them: For instance, \( m_\Upsilon \) does not satisfy the separation axiom on the class of tree-sibling model phylogenetic networks recalled above. Moreover, this metric does not even satisfy the triangle inequality on the class of all phylogenetic trees. See [9] for a detailed discussion of this issue.

Two other dissimilarity measures considered in [29], [30], and [31] are based on the representation of a rooted DAG by means of its induced subtrees: the phylogenetic trees with the same root and the same leaves as the network that are obtained by taking a spanning subtree of the network and then contracting elementary paths into nodes. For every rooted DAG \( N \), let \( T(N) \) denote the set of all its induced subtrees and \( C(N) \) the set of all clusters of nodes of these induced subtrees.

Then, for every two rooted DAGs \( N_1 = (V_1, E_1) \) and \( N_2 = (V_2, E_2) \) labeled in the same set \( S \), the authors define:

- \( m^{tree}(N_1, N_2) \) as the weight of a minimum weight edge cover of the complete bipartite graph with nodes \( T(N_1) \cup T(N_2) \) and edge weights the value of the Robinson-Foulds metric between the pairs of induced subtrees of \( N_1 \) and \( N_2 \) connected by each edge.
- \( m^{op}(N_1, N_2) \) as \( m_\Upsilon \), replacing \( \Upsilon \) by \( C \):

\[
m^{op}(N_1, N_2) = \frac{1}{2} \left( \frac{|C(N_1) \setminus C(N_2)|}{|E_1|} + \frac{|C(N_2) \setminus C(N_1)|}{|E_2|} \right).
\]

These measures do not satisfy the separation axiom on the class of tree-sibling phylogenetic networks, see, for instance, [29, Fig. 6.8]. On the positive side, Nakhleh et al. prove in [29, Section 6.4] and [30, Section 5] that they are distances on the subclass of time-consistent binary galled trees. But, it can be easily checked that on arbitrary galled trees they do not define distances either: see, for instance, Fig. 3.

3 Tree-Child Phylogenetic Networks

Since in this paper we are not interested in the reconstruction of networks, for the sake of generality, we assume the most general notion of phylogenetic network on a set \( S \) of taxa: any rooted \( S \)-DAG. Therefore, its hybrid nodes can have any indegree greater than one and any outdegree, and its tree-nodes can have any outdegree. In particular, they may contain hybrid leaves and outdegree 1 tree nodes.

We shall introduce two comparison methods on a specific subclass of such networks.

Definition 1. A phylogenetic network satisfies the tree-child condition, or it is a tree-child phylogenetic network, when every internal node has at least one tree child.

Tree-child phylogenetic networks can be understood thus as general models of reticulate evolution, where every species other that the extant ones, represented by the leaves, has some descendant through mutation. This slightly

2. Actually, Moret et al. consider also other variants of this definition, weighting only the nonstrict descendant leaves or not weighting any leaf, but for the sake of brevity and generality, we only recall here the most general version.
strengthens the condition imposed on phylogenetic networks in [24], where tree nodes had to have at least one tree child, because we also require internal hybrid nodes to have some tree child. Therefore, if hybrid nodes are further imposed to have exactly one child (for instance, in the definition of model phylogenetic network recalled in Section 2.3), this node must be a tree node: this corresponds to the interpretation of hybrid nodes not as individuals but as recombination events, producing a new individual represented by their only child. On the other hand, if hybrid nodes represent individuals, then a hybrid node with all its children hybrid corresponds to a hybrid individual that hybridizes before undergoing a speciation event, a scenario, according to [24], that “almost never arises in reality.”

The following result gives two other alternative characterizations of tree-child phylogenetic networks in terms of their strict and tree clusters.

**Lemma 2.** The following conditions are equivalent for every phylogenetic network \( N = (V, E) \):

a. \( N \) is tree-child.

b. \( T_L(v) \neq \emptyset \) for every node \( v \in V \setminus V_L \).

c. \( A_L(v) \neq \emptyset \) for every node \( v \in V \).

**Proof.** Condition a ⇒ condition b. Given any node \( v \) other than a leaf, we can construct a tree path by successively taking tree children. This path must necessarily end in a leaf that, by definition, belongs to \( T_L(v) \).

Condition b ⇒ condition c. If \( v \notin V_L \), then, by Corollary 1, \( \emptyset \neq T_L(v) \subseteq A_L(v) \), while if \( v \in V_L \), then, by definition, \( v \in A_L(v) \).

Condition c ⇒ condition a. Let \( v \) be any internal node. We want to prove that if \( A_L(v) \neq \emptyset \), then \( v \) has a tree child. Therefore, let \( s \in A_L(v) \) and consider the set \( W \) of children of \( v \) that are ancestors of \( s \). It is nonempty, because \( s \) must be a descendant of some child of \( v \). Let \( w \) be a maximal element of \( W \) with respect to the path ordering on \( N \). If \( w \) is a tree node, we are done. Otherwise, let \( v' \) be a parent of \( w \) different from \( v \). Let \( r \leadsto v' \) be any path from the root \( r \) to \( v' \). Concatenating this path with the arc \((v', w)\), any path \( w \leadsto s \) we get a path \( r \leadsto s \). Since \( s \in A_L(v) \), this path must contain \( v \), and then, since \( N \) is acyclic, \( v \) must be contained in the path \( r \leadsto v' \). Let \( w' \) be the node that follows \( v \) in this path. This node \( w' \) is a child of \( v \), and there exists a nontrivial path \( w' \leadsto w \) (through \( v' \)), which makes \( w' \) also an ancestor of \( s \). But then, \( w' \in W \) and \( w' \succ w \), which contradicts the maximality assumption on \( w \). □

The next lemma shows that tree-child phylogenetic networks are a more general model of evolution under mutation and recombination than galled trees.

**Lemma 3.** Every rooted galled tree is a tree-child phylogenetic network.

**Proof.** Let \( N = (V, E) \) be a galled tree. If \( N \) does not satisfy the tree-child condition, then it contains an internal node \( u \in V \) with all its children \( v_1, \ldots, v_k \in V \) hybrid.

The node \( u \) cannot be hybrid, because in galled trees a hybrid node cannot have any hybrid children. Indeed, assume that \( u \) has two parents \( a, b \), and let \( u' \) be the other parent of the child \( v_1 \) of \( u \). Let \( x \) be the least common ancestor of \( a \) and \( b \), and \( y \) the least common ancestor of \( b \) and \( u' \). Then, the recombination cycles defined by the paths \((x, \ldots, a, u)\) and \((x, \ldots, b, u)\), on one hand, and \((y, \ldots, b, u, v_1)\) and \((y, \ldots, u', v_1)\), on the other hand, share the arc \((b, u)\), contradicting the hypothesis that \( N \) is a galled tree; see Fig. 1a.

Thus, \( u \) is a tree node. In this case, \( k \geq 2 \), because galled trees cannot have outdegree 1 tree nodes. Now, if \( u \) is the root of \( N \), then \( A_L(u) = V_L \neq \emptyset \), and hence, by the proof of the implication condition c ⇒ condition a in Lemma 2, it has some tree child. If, on the contrary, \( u \) is not the root of \( N \), let \( w \) be its parent and \( u_1 \) and \( u_2 \) the parents other than \( u \) of \( v_1 \) and \( v_2 \), respectively. Let \( x_i \) be the least common ancestor of \( w \) and \( u_i \), and let \( x_2 \) be the least common ancestor of \( w \) and \( u_2 \). Then, the recombination cycles defined by the paths \((x_1, \ldots, u_1, v_1)\) and \((x_1, \ldots, w, u, v_1)\), on the one hand, and \((x_2, \ldots, u_2, v_2)\) and \((x_2, \ldots, w, u, v_2)\), on the other hand, share the arc \((w, u)\), contradicting again the hypothesis that \( N \) is a galled tree; see Fig. 1b. □

**Remark 1.** Not every tree-child phylogenetic network is a galled tree: see, for instance, the tree-child phylogenetic network in Fig. 4.

We provide now some upper bounds on the number of nodes in a tree-child phylogenetic network.

**Proposition 1.** Let \( N = (V, E) \) be a tree-child phylogenetic network with \( n \) leaves:

a. \( |V_H| \leq n - 1 \).

b. If \( N \) has no outdegree 1 tree node, then \( |V| \leq 2n - 1 + \sum_{v \in V} d_i(v) \).

c. If \( N \) has no outdegree 1 tree node and if \( m = \max\{d_i(v) | v \in V_H\} \), then \( |V| \leq m(n + 1) + (m - 1) \).

**Proof.**

a. Let \( r \) be the root of \( N \). Consider a mapping \( t : V \setminus V_L \rightarrow V_T \setminus \{r\} \) that assigns to every internal node one of its tree children; since tree nodes have a single parent, this mapping is injective. Then, \( |V| - |V_L| \leq |V_T| - 1 \), and since \( |V| = |V_H| + |V_T| \) and \( |V_L| = n \),

\[|V_H| = |V| - |V_T| \leq |V_L| - 1 = n - 1.\]
nodes is unbounded. Of course, if we do not forbid genetic networks with there are infinitely many nonisomorphic tree-child phylogenetic networks for which these inequalities are equalities: For 

\[ n \]

and let us fix a positive integer number \( m \geq 2 \).

For every \( i = 1, \ldots, n - 1 \), let us call \( v_{i,m} \) the parent of the leaf \( i \): To simplify the language, set \( v_{0,m} = n \). Notice that \( v_{1,m} \) is the root of the tree. Now, for every \( i = 1, \ldots, n - 1 \), split the arc \((v_{i,m}, i)\) into a path of length \( m \),

\[(v_{i,m}, v_{i,m-1}, \ldots, v_{i,1}, i),\]

and for every \( i = 1, \ldots, n - 1 \) and \( j = 1, \ldots, m - 1 \), add an arc \((v_{i,j}, h_{i+1})\). Fig. 2 displays this construction for \( n = 4 \) and \( m = 3 \).

The original binary tree had \( 2n - 1 \) nodes, and we have added \((m - 1)(n - 1)\) new tree nodes and \( n - 1 \) hybrid nodes (of indegree \( m \)). Therefore, the resulting tree-child phylogenetic network has \((m + 2)(n - 1) + 1\) nodes.

In the next sections, we define a distance on the class of all tree-child phylogenetic networks. It is convenient thus to remember here that the tripartition metrics \( m_\phi \) or \( m_\psi \) recalled in Section 2.4 do not define a distance on this class, because there exist pairs of nonisomorphic tree-child phylogenetic networks on the same set of taxa with the same sets of enriched tripartitions: for instance, the networks depicted in Figs. 4 and 8 (see [9] for details). As far as the metrics \( m_{\text{tree}} \) and \( m_{\text{spr}} \) also recalled in Section 2.4 goes, they do not define either distances on the class of all tree-child phylogenetic networks, because there also exist pairs of nonisomorphic tree-child phylogenetic networks on the same set of taxa with the same sets of induced subtrees. For instance, the tree and the galled tree depicted in Fig. 3 have the same sets of induced subtrees, namely, the tree itself; hence, the same sets of clusters of induced subtrees.

3. Henceforth, in graphical representations of phylogenetic networks, and of DAGs in general, hybrid nodes are represented by squares and tree nodes by circles.
4 THE μ-REPRESENTATION OF TREE-CHILD PHYLOGENETIC NETWORKS

Let us fix henceforth a set of labels $S = \{l_1, \ldots, l_n\}$: Unless otherwise stated, all DAGs appearing henceforth are assumed to be labeled in $S$, usually without any further notice.

Let $N = (V, E)$ be an S-DAG. For every node $u \in V$ and for every $i = 1, \ldots, n$, we denote by $m_i(u)$ the number of different paths from $u$ to the leaf $l_i$. We define the path-multiplicity vector, or simply μ-vector for short, of $u \in V$ as $\mu(u) = (m_1(u), \ldots, m_n(u))$.

To simplify the notations, we shall denote henceforth by $\delta_i^{(n)}$ the unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)$.

Lemma 4. Let $u \in V$ be any node of an S-DAG $N = (V, E)$.

a. If $u = l_i \in V_L$, then $\mu(u) = \delta_i^{(n)}$.

b. If $u \notin V_L$ and $\text{child}(u) = \{v_1, \ldots, v_k\}$, then $\mu(u) = \mu(v_1) + \cdots + \mu(v_k)$.

Proof. The statement for leaves is trivial. When $u \in V \setminus V_L$, by deleting or prepending $u$ we get, for every $i = 1, \ldots, n$, a bijection

$$\{\text{paths } u \leadsto l_i\} \leftrightarrow \bigcup_{1 \leq j \leq k} \{\text{paths } v_j \leadsto l_i\}$$

which clearly implies the statement in this case. \[\square\]

Remark 2. If $v \in \text{child}(u)$, then $\mu(u) = \mu(v)$ if and only if $v$ is the only child of $u$: Any other child would contribute something to $\mu(u)$.

Lemma 4 implies the simple Algorithm 1 to compute the μ-vectors of the nodes of an S-DAG in polynomial time. Since the height of the nodes can be computed in $O(n + |E|)$ time, it takes $O(n|E|)$ time to compute $\mu(N)$ on an S-DAG $N = (V, E)$ with $n$ leaves.

\[\text{Algorithm 1. Given } S\text{-DAG } N = (V, E), \text{ compute } \mu(N).\]

begin

for $i = 1, \ldots, n$ do

set $\mu(l_i) = \delta_i^{(n)}$

sort $V \setminus V_L$ increasingly on height

for each $x \in V \setminus V_L$ do

let $y_1, \ldots, y_k \in V$ be the children of $x$

set $\mu(x) = \mu(y_1) + \cdots + \mu(y_k)$

end

Example 2. Consider the tree-child phylogenetic network depicted in Fig. 4. Table 1 gives the μ-vectors of its nodes, sorted increasingly by their heights.

Example 3. Consider the phylogenetic network depicted in Fig. 5. Table 2 gives the μ-vectors of its nodes sorted increasingly by their heights.

Definition 2. The μ-representation of a DAG $N = (V, E)$ is the multiset $\mu(N)$ of μ-vectors of its nodes: Its elements are the vectors $\mu(u)$ with $u \in V$, and each one appears in $\mu(N)$ as many times as the number of nodes having it as μ-vector.

It turns out that a tree-child phylogenetic network can be singled out up to isomorphism among all tree-child phylogenetic networks by means of its μ-representation (Theorem 1). Before proceeding with the proof of this fact, we establish several auxiliary results.

The following lemma shows that the path ordering on a tree-child phylogenetic network is almost determined by its
\[ \begin{array}{c|c|c|c|c|c|c|c} \text{node} & \text{height} & \mu \text{-vector} & \text{node} & \text{height} & \mu \text{-vector} \\ \hline 1 & 0 & (1, 0, 0, 0, 0) & a & 2 & (1, 1, 0, 0, 0) \\ 2 & 0 & (0, 1, 0, 0, 0) & g & 2 & (0, 0, 1, 1, 0) \\ 3 & 0 & (0, 0, 1, 0, 0) & b & 3 & (1, 1, 1, 1, 0) \\ 4 & 0 & (0, 0, 0, 1, 0) & f & 3 & (0, 1, 1, 1, 0) \\ 5 & 0 & (0, 0, 0, 0, 1) & c & 4 & (1, 1, 1, 1, 0) \\ A & 1 & (1, 0, 0, 0, 0) & e & 4 & (1, 1, 1, 1, 0) \\ B & 1 & (0, 1, 0, 0, 0) & d & 5 & (1, 1, 1, 1, 1) \\ C & 1 & (0, 0, 1, 0, 0) & r & 6 & (2, 2, 2, 2, 1) \\ D & 1 & (0, 0, 0, 1, 0) & \end{array} \]

\(\mu\)-representation. In it, and henceforth, the order \(\succeq\) considered between \(\mu\)-vectors is the product partial order on \(\mathbb{N}^n\):

\[(m_1, \ldots, m_n) \succeq (m'_1, \ldots, m'_n) \iff m_i \geq m'_i \quad \text{for every } i = 1, \ldots, n.\]

**Lemma 5.** Let \(N = (V, E)\) be a tree-child phylogenetic network:

a. If there exists a path \(u \leadsto v\), then \(\mu(u) \geq \mu(v)\).

b. If \(\mu(u) > \mu(v)\), then there exists a path \(u \leadsto v\).

c. If \(\mu(u) = \mu(v)\), then \(u\) and \(v\) are connected by an elementary path.

**Proof.** Assertion a is a straightforward consequence of Lemma 4. As far as b and c go, let us assume for the moment that \(\mu(u) \geq \mu(v)\) and let \(l_i \in T_I(v)\) in particular, \(m_i(v) \geq 1\). Then, \(m_i(u) \geq m_i(v)\), and therefore, there exists also a path \(u \leadsto l_i\). Now, consider the tree path \(v \leadsto l_i\). By Lemma 1, it must happen that either the path \(u \leadsto l_i\) contains the path \(v \leadsto l_i\) or vice versa, and therefore, \(u\) and \(v\) are connected by a path.

If \(\mu(u) > \mu(v)\), by a, there cannot exist any path \(v \leadsto u\), and therefore, there exists a path \(u \leadsto v\); this proves b.

On the other hand, if \(\mu(u) = \mu(v)\), nothing prevents the existence of a path \(u \leadsto v\) or a path \(v \leadsto u\). To fix ideas, assume that there exists a path \(u \leadsto v\), say, \((u, v_1, \ldots, v_{k-1}, v)\). Since, by a,

\[\mu(u) = \mu(v_1) = \cdots = \mu(v_{k-1}) = \mu(v) = \mu(u),\]

we conclude that

\[\mu(u) = \mu(v_1) = \cdots = \mu(v_{k-1}) = \mu(v) = \mu(u).\]

As we noticed in Remark 2, this implies that each one of \(u, v_1, \ldots, v_{k-1}\) has only one child, the node that follows it in this path. Therefore, the path \(u \leadsto v\) is elementary. The same argument shows that if it is the path \(v \leadsto u\) that exists, then it is elementary. \(\square\)

**Remark 3.** Assertion a in the previous lemma holds for every DAG, but assertions b and c need not be true if the DAG \(N\) does not satisfy the tree-child condition, even if \(N\) is a binary tree-sibling phylogenetic network. Indeed, consider the phylogenetic network described in Fig. 5. We have in it that

\[\mu(c) = (1, 1, 1, 1, 0) = \mu(e),\]

\[\mu(d) = (1, 1, 1, 1, 1).\]

Then, \(c\) and \(e\) have the same \(\mu\)-vectors, but they are not connected by any path, yielding a counterexample to assertion c. And, \(\mu(d) > \mu(e)\), but there is no path \(d \leadsto e\), which yields a counterexample to assertion b.

The next lemmas show how to recover the children of a node in a tree-child phylogenetic network from the knowledge of the \(\mu\)-representation of the network.

**Lemma 6.** Let \(N\) be any DAG. Let \(u \in V\) be any internal node, and let

\[M_u = \{w \in V \mid u > w\}.\]

Then, \(M_u\) has maximal elements, and all of them are children of \(u\).

**Proof.** The set \(M_u\) is nonempty, since \(u\) is not a leaf, and every descendant of \(u\) is in \(M_u\). Since \(M_u\) is finite, it has maximal elements. Let \(v\) be any such a maximal element. Since \(u > v\), there exists a nontrivial path \(u \leadsto v\). If this path passes through some other node \(w\), then \(u > w > v\), against the assumption that \(v\) is maximal in \(M_u\). Therefore, the path \(u \leadsto v\) has length 1, and \(v\) is a child of \(u\). \(\square\)

The maximal elements of \(M_u\) are exactly the children of \(u\) such that the arc \((u, v)\) is the only path \(u \leadsto v\). This includes all tree children of \(u\) and all hybrid children \(v\) of \(u\) such that no other parent of \(v\) is a descendant of \(u\). In time consistent phylogenetic networks, this covers all children of \(u\). But, in arbitrary tree-child phylogenetic networks, this need not cover all of them. Consider, for instance, the right-hand side tree-child phylogenetic network in Fig. 3. In it, the only maximal element of \(M_b\) is \(a\), but \(A\) is also a child of \(b\).

**Lemma 7.** Let \(N\) be a tree-child phylogenetic network. Let \(u \in V\) be any internal node and \(v_1, \ldots, v_k\) some of its children:

a. If \(\mu(u) = \mu(v_1) + \cdots + \mu(v_k)\), then \(u\) has no other children.

b. If \(\mu(u) > \mu(v_1) + \cdots + \mu(v_k)\), let

\[M_{u,v_1,\ldots,v_k} = \{w \in V \mid u > w, \mu(u) - \mu(u) + \sum_{i=1}^{k} \mu(v_i)\}.\]

Then, \(M_{u,v_1,\ldots,v_k}\) has maximal elements, and all of them are children of \(u\) and different from \(v_1, \ldots, v_k\).

**Proof.** Let us assume that \(\mu(u) = \mu(v_1) + \cdots + \mu(v_k)\). Then, \(u\) cannot have other children, since if it has any other child \(w\), then

\[\mu(u) \geq \mu(v_1) + \cdots + \mu(v_k) + \mu(w) > \mu(v_1) + \cdots + \mu(v_k).\]

Assume now that \(\mu(u) > \mu(v_1) + \cdots + \mu(v_k)\). Then, by Lemma 4b, \(u\) has other children than \(v_1, \ldots, v_k\). Let \(N'\) be the DAG obtained by removing from \(N\) the arcs \((u, v_1), \ldots, (u, v_k)\). For any node \(s \in V\), let \(m'(s)\) be the number of different paths \(s \leadsto l_i\) in \(N'\) and set \(\mu'(s) = (m'_1(s), \ldots, m'_n(s))\). Then, \(\mu'(u) = \mu(u) - (\mu(v_1) + \cdots + \mu(v_k))\), because the paths \(u \leadsto l_i\) in \(N\) that are not in \(N'\) are exactly those whose first visited vertex is one of \(v_1, \ldots, v_k\). Moreover, if \(w\) is a
Then, we have that $\mu(w) = \mu'(w)$ for every $w \in M_{u,v_1,\ldots,v_k}$ and thus

$$M_{u,v_1,\ldots,v_k} = \{w \in V \mid u > w \in N \text{ and } \mu'(u) \geq \mu'(w)\}.$$ 

Now, it turns out that $w \in M_{u,v_1,\ldots,v_k}$ if and only if there exists a nontrivial path $u \rightarrow w$ in $N'$. Indeed, if there exists a nontrivial path $u \rightarrow w$ in $N'$, then there exists also the same path in $N$, and hence, $u > w$ in $N$, and moreover, by Lemma 5.a, $\mu'(u) \geq \mu'(w)$. Conversely, let $w$ be a descendant of $u$ in $N$ and assume that $\mu'(u) \geq \mu'(w) = \mu(w)$. If $l_i \in T_L(w)$, then $m_i'(u) = m_i(w) = 1$. Take the tree path $u \rightarrow l_i$ in $N$, which also exists in $N'$, and any path $u \rightarrow l_i$ in $N'$. By Lemma 1 (applied to $N'$), either $u \rightarrow l_i$ contains $w \rightarrow l_i'$ or vice versa. But, the existence of a nontrivial path $u \rightarrow w$ in $N$ prevents the existence of a path $w \rightarrow u$ in $N'$. Therefore, it is $u \rightarrow l_i$ that contains $w \rightarrow l_i'$ and in particular, there exists a path $u \rightarrow w$ also in $N'$.

So,

$$M_{u,v_1,\ldots,v_k} = \{w \in V \mid u > w \in N'\}.$$ 

Since at least one child of $u$ has survived in $N'$, the previous lemma implies that this set has maximal elements, and they are children of $u$ in $N'$, and hence, they are also children of $u$ in $N$, and they are different from $v_1, \ldots, v_k$.

As we have already mentioned, Lemma 6 applies to any DAG (and we make use of this fact in the proof of Lemma 7, because the DAG $N'$, we consider in it need not satisfy the tree-child condition and can have more than one root, as well as outdegree 1 tree nodes), but Lemma 7.b need not be true if $N$ does not satisfy the tree-child condition. Consider again, for instance, the tree-sibling phylogenetic network $N$ described in Fig. 5. In it, $c$ is a maximal element of $M_{r,d} = \{x \in V \mid r > x, \mu(v) \geq \mu(x) + \mu(d)\}$, but it is not a child of $r$.

We can prove now our main result.

**Theorem 1.** Let $N, N'$ be tree-child phylogenetic networks. Then, $N \cong N'$ if and only if $\mu(N) = \mu(N')$.

**Proof.** Let $N = (V, E)$ be a tree-child phylogenetic network labeled in $S$, and let $\mu(N)$ be its $\mu$-representation. Let $V_\mu \subseteq N \times N$ be the set consisting of the vectors of the form $(x, i)$ with $x \in \mu(N)$ and $i$ between 1 and the multiplicity of $x$ in $\mu(N)$. Consider on $V_\mu$ the partial order $\preceq$ defined by

$$(x, i) \preceq (y, j) \iff x > y \text{ with respect to the product partial order}, \text{ or } x = y \text{ and } i < j.$$ 

We know from Lemma 5 that if $x \in N^\mu$ belongs to $\mu(N)$ with multiplicity $m \geq 1$, then there exist $m$ nodes in $N$ with $\mu$-vector $x$, and that they form an elementary path. For every node $v \in V$, let $i_v$ be the position of $v$ in the elementary path formed by all nodes with the same $\mu$-vector as $v$. In particular, if $\mu(v)$ appears in $\mu(N)$ with multiplicity 1, then $i_v = 1$.

Lemma 5 implies then that the mapping

$$\tilde{\mu} : V \rightarrow V_\mu,$$

$$v \mapsto (\mu(v), i_v)$$

is an isomorphism of partially ordered sets between $V$ with the path ordering and $V_\mu$ with the partial order $\preceq$. Indeed, if $u > v$, then either $\mu(u) > \mu(v)$ or $\mu(u) = \mu(v)$, and $u$ appears before $v$ in the elementary path of all nodes with this $\mu$-vector, and hence, $i_u < i_v$. Conversely, if $\mu(u) > \mu(v)$ or if $\mu(u) = \mu(v)$ and $i_u < i_v$, then there exists a nontrivial path $u \rightarrow v$.

Therefore, we can rephrase the last two lemmas as follows:

1. For every $u \in V$ internal, the set

$$\mathcal{M}_u = \{\tilde{\mu}(w) \in \tilde{\mu}(V) \mid \tilde{\mu}(u) = \tilde{\mu}(w)\}$$

has maximal elements, and all of them are images under $\tilde{\mu}$ of children of $u$.

2. For every $u \in V$ and for every $v_1, \ldots, v_k \in \text{child}(u)$, we have the following:

   a. If $\mu(u) = \mu(v_1) + \cdots + \mu(v_k)$, then $u$ has no other children.

   b. If $\mu(u) > \mu(v_1) + \cdots + \mu(v_k)$, then the set

$$\mathcal{M}_{u,v_1,\ldots,v_k} = \{\tilde{\mu}(w) \in \tilde{\mu}(V) \mid \tilde{\mu}(u) \times \tilde{\mu}(w), \mu(u) \geq \mu(w) + \mu(v_1) + \cdots + \mu(v_k)\}$$

has maximal elements, and all of them are images under $\tilde{\mu}$ of children of $u$ other than $v_1, \ldots, v_k$.

We shall prove that we can recover the set $E$ of arcs in $N$ from $\mu(N)$. To do that, consider the set $E_\mu \subseteq V_\mu \times V_\mu$ obtained through the application of Algorithm 2:

**Algorithm 2.** Given $\mu(N)$, compute $E_\mu$.

begin

set $E_\mu = \emptyset$

sort $V_\mu$ decreasingly on the partial order $\preceq$

for each $(x, i) \in V_\mu$

set $m = x$

while $m > 0$

for each $(y, j) \in V_\mu$ such that $(x, i) \succ (y, j)$

if $m \geq y$

add the arc $((x, i), (y, j))$ to $E_\mu$

set $m = m - y$

end

end

Let us prove that $(u, v) \in E$ if and only if $(\tilde{\mu}(u), \tilde{\mu}(v)) \in E_\mu$. To do that, let $u \in V$ be an arbitrary node of $N$. If there is no arc in $E_\mu$ with tail $\tilde{\mu}(u)$, it can only be because there is no $v \in V$ such that $\mu(u) > \mu(v)$, and hence, $u$ is a leaf of $N$, and $E$ does not contain any arc with tail $u$, either. Otherwise, let $(\tilde{\mu}(u), \tilde{\mu}(v_1)), (\tilde{\mu}(u), \tilde{\mu}(v_2)), \ldots, (\tilde{\mu}(u), \tilde{\mu}(v_k))$ be the arcs contained in $E_\mu$, with tail $\tilde{\mu}(u)$, given in the order they
are added to $E_t$. This entails that $\mu(v_i)$ is a maximal element of

$$\{\mu(v) \in \mu(V) \mid \mu(u) > \mu(v)\} = M_u,$$

and, for each $i = 2, \ldots, l$, $\mu(v_i)$ is a maximal element of

$$\{\mu(v) \in \mu(V) \mid \mu(u) > \mu(v) \land \mu(u) - (\mu(v_1) + \cdots + \mu(v_{i-1})) \geq \mu(v_i)\} = \{v \in V \mid \mu(u) > \mu(v) \land \mu(v) \geq \mu(v_1) + \mu(v_2) + \cdots + \mu(v_{i-1})\} = \bar{M}_{u,v_1,v_2,\ldots,v_{i-1}}.$$

Therefore, as we have recalled in points 1 and 2.b, the nodes $v_1, \ldots, v_l$ are children of $u$ in $N$, that is, $(u, v_1), (u, v_2), \ldots, (u, v_l) \in E$. On the other hand, the algorithm adds arcs $(\mu(u), \mu(v_i))$ to $E_t$ until it happens either that $\mu(u) = \mu(v_1) + \cdots + \mu(v_l)$, in which case, by 2.a, $v_1, \ldots, v_l$ are exactly the children of $u$ in $N$ or that the set of nodes is exhausted and $\mu(u) > \mu(v_1) + \cdots + \mu(v_l)$: but the latter cannot happen, because $\mu(u)$ must be the sum of the $\mu$-vectors of its children in $N$. Thus, in summary, child$(u) = \{v_1, \ldots, v_l\}$ in $N$ and thus, $(u, v_1), (u, v_2), \ldots, (u, v_l)$ are also all the arcs contained in $E$ with tail $u$.

This proves that $E = \{(u, v) \mid (\mu(u), \mu(v)) \in E_t\}$, as we claimed. Now, if $N = (V, E)$ and $N' = (V', E')$ are two tree-child phylogenetic networks such that $\mu(N) = \mu(N')$, we have bijections:

$$V \mapsto V', V' \mapsto V.$$

Renaming in $V$ the nodes of $N'$ through this bijection $V' \mapsto V$, we obtain a phylogenetic network $N'' = (V, E'')$ isomorphic to $N$ and such that $\mu(N'') = \mu(N)$. Let $E''_t$ and $E''_p$ be, respectively, the sets or arcs obtained by applying the previous algorithm to $N$ and $N''$. Since $\mu(N) = \mu(N'')$, we have that $E''_t = E''_p$, and hence,

$$E = \{(u, v) \mid (\mu(u), \mu(v)) \in E_t\} = \{(u, v) \mid (\mu(u), \mu(v)) \in E''_t\} = E''.$$

This implies that $N \cong N'$ and, therefore, $N \cong N'$ as DAGs. Now, this isomorphism clearly preserves the leaves’ labels, because it preserves $\mu$-vectors. Therefore, $N \cong N'$ also as $S$-DAGs.

This proves the "if" implication in the statement. Of course, the "only if" implication is obvious. □

To recover, up to isomorphism, a tree-child phylogenetic network $N$ from its $\mu$-representation $\mu(N)$, it is enough to compute the set $V_t$ associated to the multiset $\mu(N)$, then to apply Algorithm 2 to compute the set of arcs $E_t$, and finally to label each leaf of the resulting DAG, which will have the form $(e_i^{(n)}, m_i)$ with $m_i$ the multiplicity of $\delta_i^{(n)}$ in $\mu(N)$, with the corresponding label $l_i$.

**Example 4.** Let us apply this procedure to the $\mu$-representation of the tree-child phylogenetic network $N$ depicted in Fig. 4. From the multiset $\mu(N)$ described in Table 1, we obtain the following set $V_t$, which we give sorted decreasingly on $\succ$ (and, to simplify the description of the application of the algorithm, we give names $x_i$ to its elements):

$$V_t = \{x_1 = ((1, 2, 3, 4, 1), 1), x_2 = ((0, 1, 2, 3, 1), 1), x_3 = ((0, 1, 1, 2, 1), 1), x_4 = ((0, 1, 1, 1, 1), 1), x_5 = ((1, 1, 1, 1, 0), 1), x_6 = ((0, 1, 1, 1, 0), 1), x_7 = ((0, 1, 1, 1, 0), 2), x_8 = ((0, 0, 1, 1, 0), 1), x_9 = ((0, 0, 1, 1, 0), 2), x_{10} = (1, 0, 0, 0, 0), 1), x_{11} = (1, 0, 0, 0, 0), 1), x_{12} = (0, 0, 1, 0, 0), 1), x_{13} = (0, 0, 0, 1, 0), 1), x_{14} = (0, 0, 0, 1, 0), 2), x_{15} = (0, 0, 0, 0, 1), 1)\}.

We shall denote the first and the second component of each $x_i$ by $\mu_i$ and $\nu_i$, respectively.

We begin with an empty set of arcs:

$$E_t = \emptyset.$$

Then, we proceed with the for each in Algorithm 2, visiting all elements of $V_t$ in the given order:

- $(x_1)$. We set $m = (1, 2, 3, 4, 1)$. Then, since $m \geq \mu_2$, we add $(x_1, x_2)$ to $E_t$ and we set $m = m - \mu_2 = (1, 1, 1, 0)$. The first next element $x_3$ of $V_t$ with $\mu_3 \leq m$ is $x_3$. Then, we add $(x_1, x_3)$ to $E_t$ and we set $m = m - \mu_3 = (0, 0, 0, 0, 0)$. This makes us decide to stop with $x_1$. At the end of this step, we have

$$E_t = \{(x_1, x_2), (x_1, x_3)\}.

- $(x_2)$. We set $m = (0, 1, 2, 3, 1)$. Since $m \geq \mu_3$, we add $(x_2, x_3)$ to $E_t$ and we set $m = m - \mu_3 = (0, 0, 1, 1, 0)$. The first next element $x_4$ of $V_t$ with $\mu_4 \leq m$ is $x_4$. Then, we add $(x_2, x_3)$ to $E_t$ and we set $m = m - \mu_4 = (0, 0, 0, 0, 0)$, and we stop. Therefore, at the end of this step, we have

$$E_t = \{(x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_4)\}.

- $(x_3)$. We set $m = (0, 1, 1, 2, 1)$. Then, since $m \geq \mu_4$, we add $(x_3, x_4)$ to $E_t$ and we set $m = m - \mu_4 = (0, 0, 0, 1, 0)$. The first next element $x_5$ of $V_t$ with $\mu_5 \leq m$ is $x_5$. Then, we add $(x_3, x_4)$ to $E_t$ and we set $m = m - \mu_5 = (0, 0, 0, 0, 0)$, and we stop. At the end of this step, we have

$$E_t = \{(x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_4), (x_3, x_4), (x_3, x_5)\}.

- $(x_4)$. We set $m = (0, 1, 1, 1, 1)$. The first element $x_6$ with $\mu_6 \leq m$ is $x_6$, and therefore, we add $(x_4, x_6)$ to $E_t$ and we set $m = m - \mu_6 = (0, 0, 0, 0, 1)$. The first next element $x_7$ of $V_t$ with $\mu_7 \leq m$ is $x_7$. Then, we add $(x_4, x_6)$ to $E_t$ and we set $m = m - \mu_7 = (0, 0, 0, 0, 0)$, and we stop. At the end of this step, we have

$$E_t = \{(x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_4), (x_3, x_4), (x_3, x_5), (x_4, x_6), (x_4, x_7)\}.$$

- $(x_5)$. We set $m = (1, 1, 1, 1, 0)$. Since $m \geq \mu_6$, we add $(x_5, x_6)$ to $E_t$ and we set $m = m - \mu_6 = (1, 0, 0, 0, 0)$. The first next element $x_7$ of $V_t$ with
\( \mu_k \leq m \) is \( x_{10} \). Then, we add \( (x_5, x_{10}) \) to \( E_\mu \). Now, 
\( m = m - x_{10} = (0, 0, 0, 0, 0) \), and we stop. At the end of this step, we have

\[
E_\mu = \{ (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_8), (x_3, x_4), (x_3, x_{13}), (x_4, x_6), (x_4, x_{15}), (x_5, x_6), (x_5, x_{10}), (x_6, x_7) \}.
\]

- \( (x_6) \). We set \( m = (0, 1, 1, 1, 0) \). Since \( m \geq \mu_7 \), we add \( (x_7, x_8) \) to \( E_\mu \), and we set \( m = m - \mu_7 = (0, 0, 0, 0, 0) \), and we stop. At the end of this step, we have

\[
E_\mu = \{ (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_8), (x_3, x_4), (x_3, x_{13}), (x_4, x_6), (x_4, x_{15}), (x_5, x_6), (x_5, x_{10}), (x_6, x_7) \}.
\]

- \( (x_7) \). We set \( (again) m = (0, 1, 1, 1, 0) \). Since \( m \geq \mu_6 \), we add \( (x_7, x_8) \) to \( E_\mu \), and we set \( m = m - \mu_6 = (0, 1, 0, 0, 0) \). The first next element \( x_8 \) of \( V_\mu \) with \( \mu_k \leq m \) is \( x_{11} \). Then, we add \( (x_7, x_{11}) \) to \( E_\mu \). Now, \( m = m - \mu_{11} = (0, 0, 0, 0, 0) \), and we stop. At the end of this step, we have

\[
E_\mu = \{ (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_8), (x_3, x_4), (x_3, x_{13}), (x_4, x_6), (x_4, x_{15}), (x_5, x_6), (x_5, x_{10}), (x_6, x_7), (x_7, x_8), (x_7, x_{11}) \}.
\]

- \( (x_8) \). We set \( m = (0, 0, 1, 1, 0) \). Since \( m \geq \mu_9 \), we add \( (x_8, x_9) \) to \( E_\mu \), we set \( m = m - \mu_9 = (0, 0, 0, 0, 0) \), and we stop. At the end of this step, we have

\[
E_\mu = \{ (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_8), (x_3, x_4), (x_3, x_{13}), (x_4, x_6), (x_4, x_{15}), (x_5, x_6), (x_5, x_{10}), (x_6, x_7), (x_7, x_8), (x_7, x_{11}), (x_8, x_9) \}.
\]

- \( (x_9) \). We set \( m = (0, 0, 1, 1, 0) \). The first element \( x_k \) in \( V_\mu \) with \( \mu_k \leq m \) is \( x_{12} \), and then we add \( (x_9, x_{12}) \) to \( E_\mu \), and we set \( m = m - \mu_{12} = (0, 0, 0, 1, 0) \). The first next element \( x_k \) in \( V_\mu \) with \( \mu_k \leq m \) is \( x_{13} \). Then, we add \( (x_9, x_{13}) \) to \( E_\mu \). Now, \( m = m - \mu_{13} = (0, 0, 0, 0, 0) \), and we stop. At the end of this step, we have

\[
E_\mu = \{ (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_8), (x_3, x_4), (x_3, x_{13}), (x_4, x_6), (x_4, x_{15}), (x_5, x_6), (x_5, x_{10}), (x_6, x_7), (x_7, x_8), (x_7, x_{11}), (x_8, x_9), (x_9, x_{12}), (x_9, x_{13}) \}.
\]

- \( (x_{10}) \). Since \( m = (1, 0, 0, 0, 0) \) has only one nonzero entry, and \( x_{10} \) is the only element of \( V_\mu \) with \( \mu_k = m \), the algorithm does not find any arc in \( E_\mu \) with head \( x_{10} \). The same happens with \( x_{11}, x_{12}, \) and \( x_{15} \).

\[
\text{Fig. 6. The DAG recovered from the } \mu \text{-representation of the phylogenetic network in Fig. 4.}
\]

- \( (x_{13}) \). We set \( m = (0, 0, 0, 1, 0) \). Since \( m \geq \mu_{14} \), we add \( (x_{13}, x_{14}) \) to \( E_\mu \), we set \( m = m - \mu_{14} = (0, 0, 0, 0, 0) \), and we stop. At the end of this step, we have

\[
E_\mu = \{ (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_8), (x_3, x_4), (x_3, x_{13}), (x_4, x_6), (x_4, x_{15}), (x_5, x_6), (x_5, x_{10}), (x_6, x_7), (x_7, x_8), (x_7, x_{11}), (x_8, x_{9}), (x_9, x_{12}), (x_9, x_{13}), (x_{13}, x_{14}) \}.
\]

- \( (x_{14}) \). Since \( m = (0, 0, 0, 1, 0) \) has only one nonzero entry, and no other element of \( V_\mu \) after \( x_{14} \) has this first component, the algorithm does not find any arc in \( E_\mu \) with head \( x_{14} \).

The DAG \( (V_\mu, E_\mu) \) obtained up to now is depicted in Fig. 6. Finally, we would label 1, 2, 3, 4, and 5 the nodes \( x_{10}, x_{11}, x_{12}, x_{14}, \) and \( x_{15} \), respectively. The resulting DAG labeled in \( \{1, \ldots, 5\} \) is clearly isomorphic to the tree-child phylogenetic network \( N \) in Fig. 4.

\textbf{Remark 4.} The thesis of Theorem 1 need not hold if \( N_1 \) and \( N_2 \) do not satisfy the tree-child condition. Indeed, it is not difficult to check that the tree-sibling phylogenetic network given in Fig. 7 has the same \( \mu \)-representation, as the one given in Fig. 5, but they are not isomorphic as S-DAGs.

\section{The \( \mu \)-Distance for Tree-Child Phylogenetic Networks}

For every pair of DAGs \( N_1 \) and \( N_2 \) labeled in the same set \( S \), let

\[
d_\mu(N_1, N_2) = |\mu(N_1) \triangle \mu(N_2)|,
\]

where the symmetric difference \( \triangle \) refers to multisets: If a vector belongs to \( \mu(N_1) \) with multiplicity \( a \) and to \( \mu(N_2) \) with multiplicity \( b \), then it belongs to \( \mu(N_1) \triangle \mu(N_2) \) with multiplicity \( |a - b| \), and hence, it contributes \( |a - b| \) to \( |\mu(N_1) \triangle \mu(N_2)| \).
Theorem 2. Let $N_1, N_2, N_3$ be tree-child phylogenetic networks on the same set of taxa. Then,

a. $d_\mu(N_1, N_2) \geq 0$,

b. $d_\mu(N_1, N_2) = 0$ if and only if $N_1 \cong N_2$,

c. $d_\mu(N_1, N_2) = d_\mu(N_2, N_1)$, and

d. $d_\mu(N_1, N_3) \leq d_\mu(N_1, N_2) + d_\mu(N_2, N_3)$.

Proof. Points a, c, and d are direct consequences of the properties of the symmetric difference, and point b is a consequence of Theorem 1.

Therefore, $d_\mu$ defines a distance on the class of all tree-child phylogenetic networks: We shall call it the $\mu$-distance.

We have shown in Section 4 that the $\mu$-representation of an S-DAG can be computed in polynomial time. Now, given two S-DAGs $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ with $n$ leaves together with their $\mu$-representations $\mu(N_1)$ and $\mu(N_2)$, the simple Algorithm 3 performs a simultaneous traversal of the internal nodes of $N_1$ and $N_2$, sorted by their $\mu$-vectors, in order to compute the $\mu$-distance $d_\mu(N_1, N_2)$ in $O(n|V|)$ time, where $|V| = \max(|V_1|, |V_2|)$.

Algorithm 3. Given the $\mu$-representations $\mu(N_1)$ and $\mu(N_2)$ of two S-DAGs $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$, compute $d_\mu(N_1, N_2)$.

begin
let $X_1$ be the set of internal nodes of $N_1$
let $X_2$ be the set of internal nodes of $N_2$
sort $X_1$ and $X_2$ increasingly according to the lex ordering of the $\mu$-vectors
set $d = 0$
while $X_1 \neq \emptyset$ and $X_2 \neq \emptyset$ do
let $x_1$ and $x_2$ be the first element of $X_1$ and $X_2$, respectively
\textbf{case} $\mu(x_1) < \mu(x_2)$
set $X_1 = X_1 \setminus \{x_1\}$
set $d = d + 1$
\textbf{case} $\mu(x_1) > \mu(x_2)$
set $X_2 = X_2 \setminus \{x_2\}$
set $d = d + 1$
\textbf{otherwise}
set $X_1 = X_1 \setminus \{x_1\}$
return $d + |X_1| + |X_2|$
end

Example 5. Consider the tree $T$ and the network $N$ in Fig. 3. Their $\mu$-representations are

$\mu(T) = \{(0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 1, 1)\}$,

$\mu(N) = \{(0, 0, 1), (0, 1, 0), (0, 1, 1), (0, 2, 1), (1, 0, 0), (1, 2, 1)\}$,

and therefore,

$\mu(T) \triangle \mu(N) = \{(0, 1, 0), (0, 2, 1), (1, 1, 1), (1, 2, 1)\}$

from which we obtain $d_\mu(T, N) = 4$.

Example 6. It was shown in [9] that the tree-child phylogenetic network $N$ depicted in Fig. 4 could not be distinguished from the tree-child phylogenetic network $N'$ depicted in Fig. 8 using the tripartition metric. We have already given in Table 1 the $\mu$-vectors of the nodes of $N$. In Table 3, we give the $\mu$-vectors of the nodes of $N'$. From these tables, we get that

$|\mu(N) \triangle \mu(N')| = \{(0, 1, 1, 2, 1), (0, 1, 2, 2, 1)\}$,

which implies that $d_\mu(N, N') = 2$.

Example 7. Let $N$ be any tree-child phylogenetic network labeled in $S$, and let $v$ be any internal node of it. Let $N'$ be a tree-child phylogenetic network obtained by adding to

\begin{table}[h]
\centering
\caption{\mu-vectors of the Nodes of the Network Depicted in Fig. 8}
\begin{tabular}{|c|c|c|c|c|}
\hline
\text{node} & \text{height} & \mu-vector & \text{node} & \text{height} & \mu-vector \\
\hline
1 & 0 & (1,0,0,0,0) & e & 4 & (0,1,1,1,0) \\
2 & 0 & (0,1,0,0,0) & A & 5 & (0,1,1,0) \\
3 & 0 & (0,1,1,0,0) & a & 6 & (1,1,1,1) \\
4 & 0 & (0,0,1,1,0) & d & 6 & (0,1,1,1) \\
5 & 0 & (0,0,0,1,0) & c & 7 & (0,1,2,2,1) \\
C & 1 & (0,0,1,0,0) & b & 8 & (0,1,2,3,1) \\
f & 2 & (0,0,1,1,0) & r & 9 & (1,2,3,4,1) \\
B & 3 & (0,0,1,1,0) & & & \\
\hline
\end{tabular}
\end{table}
There exist 66 pairwise nonisomorphic binary 
-distance 1 \( \Delta \) representations are

\[ \mu(N) = \{ \{0, 0, 1\}, \{0, 1, 0\}, \{1, 0, 0\}, \{1, 1, 0\}, \{1, 1, 1\}, \{2, 2, 1\} \}, \]

\[ \mu(N') = \{ \{0, 0, 1\}, \{0, 1, 0\}, \{1, 0, 0\}, \{1, 1, 0\}, \{1, 1, 1\}, \{2, 2, 1\} \}, \]

and thus \( d_{\mu}(N, N') = 1 \).

It should also be noticed that, against what happens in the tree case, contracting an arc in a tree-child phylogenetic network \( N \) into a node (that is, given an arc \( (v, v') \), removing \( v' \) and this arc, and replacing every other arc with tail or head \( v' \) by a new arc with tail or head, respectively), \( v \) need not produce a network at \( -distance 1 \) of \( N \), for instance, if \( v \) and \( v' \) hybridize in \( N \) or if \( v' \) is a nonstrict descendant of \( v \). We leave to the interested reader to draw specific counterexamples.

Example 8. There exist 66 pairwise nonisomorphic binary tree-child phylogenetic networks with three leaves. All of them have an even number of internal nodes, and therefore, the \( \mu \)-distance between two of them is always an even number. In Proposition 2, we shall see that this \( \mu \)-distance is smaller than or equal to 12. Fig. 10a shows the distribution of distances between unordered pairs of such networks.

In a similar way, there exist 4,059 pairwise nonisomorphic binary tree-child phylogenetic networks with four leaves. Again, all of them have an even number of internal nodes, and therefore, the \( \mu \)-distance between two of them is always even, and in Proposition 2, we shall see that it is smaller than or equal to 18. Fig. 10b shows the corresponding distribution of distances. See the supplementary material for more details, which can be found at http://bioinfo.uib.es/~recerca/phylonetworks/mudistance/.

Every phylogenetic tree is a tree-child phylogenetic network, and as we have already mentioned, it turns out that the restricted of this \( \mu \)-distance to the class of phylogenetic trees is the Robinson-Foulds metric.

Theorem 3. For every phylogenetic trees \( T_1, T_2 \) on the same set of taxa \( S \),

\[ d_{\mu}(T_1, T_2) = d_{RF}(T_1, T_2). \]

Proof. Let \( S = \{l_1, \ldots, l_n\} \). The uniqueness of paths in trees implies that, if \( T = (V, E) \) is any phylogenetic tree labeled in \( S \), then, for every \( u \in V \),

\[ m_u(u) = \begin{cases} 1, & \text{if } l \in C_L(u), \\ 0, & \text{if } l \notin C_L(u). \end{cases} \]

Therefore, the \( \mu \)-vector of a node \( u \) of a phylogenetic tree labeled in \( S \) is the image of its cluster \( C_L(u) \) under the bijection between the powerset \( \mathcal{P}(S) \) of \( S \) and \( \{0, 1\}^n \) that sends each subset \( A \) of \( S \) to its characteristic vector:

\[ \chi_A = (\chi_A(l_1), \ldots, \chi_A(l_n)), \quad \text{with } \chi_A(l_i) = \begin{cases} 1, & \text{if } l_i \in A, \\ 0, & \text{if } l_i \notin A. \end{cases} \]

Then, given two phylogenetic trees \( T_1 \) and \( T_2 \) on the set of taxa \( S \), this bijection \( \mathcal{P}(S) \rightarrow \{0, 1\}^n \) transforms the sets \( C_L(T_1) \) and \( C_L(T_2) \) of clusters of their nodes into their \( \mu \)-representations \( \mu(T_1) \) and \( \mu(T_2) \), respectively, and hence, the symmetric difference of the former into the symmetric difference of the latter. Therefore,

\[ |\pi(T_1) \triangle \pi(T_2)| = |C_L(T_1) \triangle C_L(T_2)| = |\mu(T_1) \triangle \mu(T_2)|, \]

as we claimed. \( \square \)

Remark 5. The same proof shows that if \( T_1 \) and \( T_2 \) are regular tree-child phylogenetic networks, then \( d_{\mu}(T_1, T_2) \) is equal to the symmetric difference of their sets of clusters. In [9], we proved that the tree-child phylogenetic networks without outdegree 1 tree nodes and where no parent of a hybrid node is a descendant of another parent of the same hybrid node are regular.

The \( \mu \)-distance \( d_{\mu} \) takes integer values. Its smallest nonzero value is 1, but it can be arbitrarily large. If we bound the indegree of the hybrid nodes of the networks, then we can compute the diameter of the resulting subclass of phylogenetic networks.
Let $TCN_{n,m}$ be henceforth the class of all tree-child phylogenetic networks on a fixed set of taxa of $n$ elements, without outdegree 1 tree nodes and with all their hybrid nodes of indegree at most $m$.

**Proposition 2.** For every $N_1, N_2 \in TCN_{n,m}$,
\[
d_m(N_1, N_2) \leq 2(m+1)(n-1),
\]
and there exist pairs of networks in $TCN_{n,m}$ at $\mu$-distance $2(m+1)(n-1)$.

**Proof.** Let $N_1, N_2 \in TCN_{n,m}$. By Proposition 1.c, each one of $\mu(N_1), \mu(N_2)$ has at most $(m+1)(n-1)+1$ elements, from which at least the $n$ $\mu$-vectors corresponding to the leaves will appear in both sets. Therefore,
\[
d_m(N_1, N_2) \leq 2(m+2)(n-1)+1-2n = 2(m+1)(n-1).
\]

To find a pair of networks in $TCN_{n,m}$ at distance $2(m+1)(n-1)$, let $N$ be the tree-child phylogenetic network with $n$ leaves and $n-1$ hybrid nodes of indegree $m$ described in Example 1. A simple argument by induction shows that
\[
\mu(h_n) = (0, \ldots, 0, 0, 0, 1),
\]
\[
\mu(v_{n-1,k}) = (0, \ldots, 0, 0, 1, k) \text{ for every } k = 1, \ldots, m,
\]
\[
\mu(h_{n-1}) = (0, \ldots, 0, 0, 1, m),
\]
\[
\mu(v_{n-2,k}) = (0, \ldots, 0, 1, k, km) \text{ for every } k = 1, \ldots, m,
\]
\[
\mu(h_{n-2}) = (0, \ldots, 0, 1, m, m^2),
\]
\[
\mu(v_{n-3,k}) = (0, \ldots, 1, k, km, km^2) \text{ for every } k = 1, \ldots, m
\]
\ldots,
\]
and, in general,
\[
\mu(h_{n-j}) = (0, \ldots, 0, 1, m, m^2, \ldots, m^j), \quad j = 0, \ldots, n-2,
\]
\[
\mu(v_{n-j,k}) = (0, \ldots, 0, 1, k, km, \ldots, km^j), \quad k = 1, \ldots, m,
\]
\[
\mu(h_{n-j-1}) = (0, \ldots, 0, 1, m, m^2, \ldots, m^j), \quad j = 1, \ldots, n-1.
\]
Therefore, $\mu(N)$ contains, beside the $\mu$-vectors $\delta^{(n)}_i$ of leaves, all vectors of the form
\[
(0, \ldots, 0, 1, k, km, \ldots, km^j),
\]
\[
k = 1, \ldots, m-1, \quad j = 1, \ldots, n-1
\]
with multiplicity 1, and all vectors of the form
\[
(0, \ldots, 0, 1, m, m^2, \ldots, m^j), \quad j = 1, \ldots, n-2
\]
with multiplicity 2.

Now, let $N'$ be the tree-child phylogenetic network in $TCN_{n,m}$ obtained by performing the same construction starting with the binary phylogenetic tree described by the Newick string
\[
(n, (n-1, (n-2, \ldots, (2, 1) \ldots)).
\]
The same argument shows that $\mu(N')$ contains, again, besides the $\mu$-vectors $\delta^{(n)}_i$ of leaves, all vectors of the form
\[
(km', \ldots, km, k, 1, 0, \ldots, 0), \quad k = 1, \ldots, m-1, \quad j = 1, \ldots, n-1
\]
with multiplicity 1, and all vectors of the form
\[
(m', \ldots, m^2, m, 1, 0, \ldots, 0), \quad j = 1, \ldots, n-2
\]
with multiplicity 2.

Then, $\mu(N)$ and $\mu(N')$ have no $\mu$-vector of internal node in common, and since each one has $(m+1)(n-1)$ internal nodes, this implies that $d_m(N, N') = 2(m+1)(n-1)$. \qed

This result allows us to normalize the $\mu$-distance on $TCN_{n,m}$.

**Corollary 2.** The mapping
\[
d'_m : TCN_{n,m} \times TCN_{n,m} \rightarrow \mathbb{R}
\]
\[
(N_1, N_2) \mapsto \frac{1}{2(m+1)(n-1)} d_m(N_1, N_2)
\]
is a distance on $TCN_{n,m}$ that takes values in the unit interval $[0, 1]$.

## 6 Alignment of Tree-Child Phylogenetic Networks

Let $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$ be two tree-child phylogenetic networks labeled in the same set $S = \{l_1, \ldots, l_n\}$. For simplicity, we assume that they do not have outdegree 1 tree nodes, and therefore, if two nodes in one of these networks have the same $\mu$-vector, then they must be a hybrid node and its only child.

For every $v_1 \in V_1$ and $v_2 \in V_2$, let
\[
H(v_1, v_2) = \sum_{i=1}^{n} |m_i(v_1) - m_i(v_2)|,
\]
\[
\chi(v_1, v_2) = \begin{cases} 0, & \text{if } v_1, v_2 \text{ are of the same type} \\ 1, & \text{if } v_1, v_2 \text{ are of different type} \end{cases}
\]
Notice that $H(v_1, v_2)$ is the Manhattan, or $L_1$, distance between $\mu(v_1)$ and $\mu(v_2)$. The advantage of this distance over the euclidean distance is that it takes integer values on $\mathbb{N}^n$.

Define finally the weight of the pair $(v_1, v_2)$ as
\[
w(v_1, v_2) = H(v_1, v_2) + \frac{\chi(v_1, v_2)}{2n}.
\]
To fix ideas, assume that $|V_1| \leq |V_2|$. Then, given a matching between $N_1$ and $N_2$, that is, an injective mapping $M : V_1 \rightarrow V_2$ that preserves leaves and their labels, its total weight is defined as
\[
w(M) = \sum_{v \in V_1} w(v, M(v)).
\]
An optimal alignment between $N_1$ and $N_2$ is a matching with the smallest total weight. Such an optimal alignment can be computed in time $O((|V_1| + |V_2|)^2)$ using the Hungarian algorithm [19], [26].
Proposition 3. A matching $M$ between $N_1$ and $N_2$ is an optimal alignment if and only if it minimizes the sum
\[ \sum_{v \in V(N_1)} H(v, M(v)), \]
and among those matchings minimizing this sum, it maximizes the number of nodes that are sent to nodes of the same type.

Proof. Let $M : V_1 \rightarrow V_2$ be any matching. Then,
\[ w(M) = \sum_{v \in V_1} w(v, M(v)) \]
\[ = \sum_{v \in V_1} H(v, M(v)) + \frac{1}{2n} \sum_{v \in V_1} \chi(v, M(v)). \]

The first addend is a positive integer, while the second addend is strictly smaller than 1, because by Proposition 1.a both $N_1$ and $N_2$ have at most $n-1$ hybrid nodes, and therefore, $\sum_{v \in V_1} \chi(v, M(v)) \leq 2(n-1)$. Therefore, $\sum_{v \in V_1} H(v, M(v))$ is the integer part of $w(M)$. This implies that $w(M) < w(M')$ if and only if
\[ \sum_{v \in V_1} H(v, M(v)) < \sum_{v \in V_1} H(v, M'(v)), \]
or the latter are equal, and then,
\[ \frac{1}{2n} \sum_{v \in V_1} \chi(v, M(v)) < \frac{1}{2n} \sum_{v \in V_1} \chi(v, M'(v)), \]
from where the statement clearly follows.

Remark 6. If we restrict this alignment method to phylogenetic trees, the weight of a pair of nodes $(v_1, v_2)$ is simply $|C_L(v_1) \triangle C_L(v_2)|$. This can be seen as an unnormalized version of the score used in TreeJuxtaposer [27].

Remark 7. Let $N = (V, E)$ and $N' = (V', E')$ be two tree-child phylogenetic networks without outdegree 1 tree nodes. If they are isomorphic, the isomorphism between them is an optimal alignment of total weight 0. The converse implication is clearly false in general: a matching of total weight 0 need not be an isomorphism. Consider, for instance, the optimal alignment between the phylogenetic trees described by the Newick strings $(1, 2, 3)$ and $(1, 2, 3)$.

But, if there exists an alignment $M$ (obviously optimal) between $N$ and $N'$ of total weight 0 and if $|V| = |V'|$, then $M$ is a bijection between $V$ and $V'$ that preserves the $\mu$-vectors, because
\[ \sum_{v \in V} |\mu(v) - \mu(M(v))| \leq w(M) = 0 \Rightarrow \mu(v) = \mu(M(v)) \]
for every $v \in V$, and therefore, $\mu(N) = \mu(N')$, which implies, by Theorem 1, $N \cong N'$.

Given two $S$-DAGs $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$, together with their $\mu$-representations $\mu(N_1)$ and $\mu(N_2)$, the simple Algorithm 4 computes an optimal alignment between $N_1$ and $N_2$, together with its weight, in $O((|V_1| + |V_2|)^3)$ time.

Algorithm 4. Given the $\mu$-representations $\mu(N_1)$ and $\mu(N_2)$ of two $S$-DAGs $N_1 = (V_1, E_1)$ and $N_2 = (V_2, E_2)$, compute an optimal alignment between $N_1$ and $N_2$, together with its weight.

begin
let $X_1$ be the set of internal nodes of $N_1$
let $X_2$ be the set of internal nodes of $N_2$
let $G = (X_1 \cup X_2, X_1 \times X_2)$ be the complete bipartite graph on $X_1 \cup X_2$
for each $x_1 \in X_1$ and $x_2 \in X_2$ do
set weight[$x_1, x_2$] = abs($\mu_1(x_1) - \mu_2(x_2)$)
if $x_1$ and $x_2$ are not both tree nodes or both hybrid
then add $1/2n$ to weight[$x_1, x_2$]
let $M : X_1 \rightarrow X_2$ be a minimum-weight bipartite matching of $G$
extend $M$ to $V_1$ by sending each leaf in $N_1$
to the leaf of $N_2$ with the same label
set $w = \sum_{v \in X_1} \text{weight}[^v, M(v)]$
return $(M, w)$
end

Example 9. Consider the tree $T$ and the galled tree $N$ depicted in Fig. 3. The total weight of the matching between $T$ and $N$ that sends the root of $T$ to the root of $N$ and the node $u$ of $T$ to the tree node $a$ of $N$ is
\[ w(r, r) + w(u, a) + w(1, 1) + w(2, 2) + w(3, 3) \]
\[ = 0 + 1 + 0 + 0 + 1 = 1, \]
and hence, since $\mu(T) \not\cong \mu(N)$, it is an optimal alignment, see Fig. 11.

Example 10. Consider the tree-child phylogenetic networks $N$ and $N'$ labeled in $\{1, \ldots, 5\}$ given in Fig. 12. The $\mu$-vectors of their internal nodes are given in Table 4. Table 5 gives the values of $w(x, y)$ for every internal node $x$ of $N$ and every internal node $y$ of $N'$. From this table, the optimal alignment marked in boldface in the table and depicted in Fig. 13 (where to simplify the picture, the arrows joining each leaf of $N$ to the homonymous leaf in $N'$ are omitted) is deduced: its total weight is 8. It is the only optimal alignment between these networks.

A Web tool that computes an optimal alignment of two tree-child phylogenetic networks with the same leaves and without outdegree 1 tree nodes is available at http://bioinfo.ub.es/~recerca/phylonetworks/mudistance/.

Fig. 11. An optimal alignment between the tree and the galled tree in Fig. 3.
7 Conclusion

In this paper, we have presented two methods for comparing pairs of tree-child phylogenetic networks: a metric and an alignment algorithm. While the former quantifies the similarity between two networks in a way that allows to soundly establish whether a network is more similar to a second one than to a third one, the latter allows the interactive visualization of the differences between two networks. They are respectively the first true distance and the first alignment method defined on a meaningful class of phylogenetic networks strictly extending the class of phylogenetic trees, let alone the natural extension of the bipartition metric to regular networks. Since the class of tree-child phylogenetic networks includes the galled trees, this distance and this alignment can be used to compare the latter.

Tree-child phylogenetic networks with no redundant arcs (for which there is already a path from the tail to the head of the arc), known as normal networks, have been recently proposed by S.J. Willson as the class of networks where to look for meaningful phylogenies [47], and polynomial time algorithms are known for reconstructing a normal network from distances [45], [46], [48], but for the moment no reconstruction algorithm for arbitrary tree-child phylogenetic networks has been developed. Therefore, it remains an interesting open question to characterize the sets of sequences whose evolution can be explained by means of a tree-child network and to provide an algorithm to reconstruct this network, as well as to characterize the computational complexity of these problems.

On the other hand, several reconstruction methods for time-consistent tree-sibling phylogenetic networks have been proposed by Nakhleh and collaborators. Since no true distance for these networks is known so far, it is an interesting open question whether our distance and alignment method can be extended to these networks or not.

The Supplementary Material referenced in the paper is available at http://bioinfo.uib.es/~recerca/phylonet works/mudistance/.

Acknowledgments

The research described in this paper has been partially supported by the Spanish CICYT project TIN 2004-07925-C03-01 GRAMMARS and by Spanish DGI projects MTM2006-07773 COMGRIIO and MTM2006-15038-C02-01. The authors would like to acknowledge with thanks the anonymous referees, whose comments and criticism have led to a substantial improvement of this paper.

References

G. Cardona is an associate professor in the Department of Mathematics and Computer Science, University of the Balearic Islands. He is a member of the Computational Biology and Bioinformatics Research Group, University of the Balearic Islands, and of the Number Theory Research Group, Technical University of Catalonia. His research interests are split between number theory (mainly, arithmetical properties of curves of genus $2$) and computational biology (mainly, mathematical models of evolution).
Francesc Rosselló is an associate professor in the Department of Mathematics and Computer Science, University of Balearic Islands, and the main researcher in the Computational Biology and Bioinformatics Research Group, Research Institute on Health Sciences (IUNICS), and the University of the Balearic Islands. His current research interests include computational and systems biology, with emphasis on the development of new mathematical models in these areas and the comparison of graphs and other nonlinear structures arising in them.

Gabriel Valiente is an associate professor in the Department of Software and a member of the Algorithms, Bioinformatics, Complexity, and Formal Methods Research Group, Technical University of Catalonia, Barcelona. He is also a member of the Computational Biology and Bioinformatics Research Group, University of the Balearic Islands, and an associated member of the Bioinformatics Research Centre, University of Glasgow. His research interests include computational and systems biology, with emphasis on algorithms in bioinformatics and mathematical models in computational and systems biology.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.