A Distance Metric for a Class of Tree-Sibling Phylogenetic Networks (Supplementary Material)

Gabriel Cardona¹, Mercè Llabrés¹, Francesc Rosselló¹, and Gabriel Valiente²

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

ABSTRACT
In this Supplementary Material we give:

- Proofs we have not included in the main paper.
- Counterexamples to show that the technical conditions imposed are necessary to single out phylogenetic networks by means of their µ-representation, and that the definition given is strictly more general than the one given by Nakhleh in his PhD thesis.
- Some computational experiments we have made.

1 PROOFS

PROOF OF PROPOSITION 2. The result is obvious if $n \leq 2$, since then $N$ is a tree.

Assume that $n \geq 3$ and that the result is proved for networks with less than $n$ leaves. Let $M$ be the set of internal nodes with maximum time assignment, and let $M_h$ (respectively, $M_b$) be the set of tree nodes (respectively, hybrid nodes) in $M$. Notice that $M_h$ is non-empty, because if a hybrid node has maximum time assignment, its two parents have the same time assignment and, therefore, are in $M_h$. Consider the following different situations:

1. If some node $u$ in $M_h$ has two (or more) children leaves, let $N'$ be the sTSTC network obtained by removing one of these leaves and eventually collapsing the created elementary path into a single arc. Then the number of leaves, hybrid nodes and internal tree nodes in $N'$ is

   $n' = n - 1, \quad h' = h, \quad t' = t - \epsilon$,

   with $\epsilon = 0$ if the out-degree of $u$ in $N$ is greater than two, and $\epsilon = 1$ otherwise. Now, from the induction hypothesis we get

   $h = h' \leq 2n' - 4 = 2(n - 2) - 4 < 2n - 4$,

   $t = t' + \epsilon \leq 3n' - 6 + \epsilon = 3n - 9 + \epsilon < 3n - 6$.

2. If (1) does not hold, but every node in $M_h$ has one child leaf, let $N'$ be the sTSTC network obtained by removing all the nodes in $M_h$, together with their respective children leaves (say $k = |M_h|$), and collapsing the created elementary paths into single arcs. In this case we have that

   $n' = n - k, \quad h' = h - k, \quad t' = t - \tilde{k}$,

   where $\tilde{k} \leq 2k$ is the number of elementary paths that have been removed. Now, also from the induction hypothesis we get

   $h = h' + k \leq 2n' - 4 + k = 2n - 2k - 4 + k = 2n - 4 - 2k < 2n - 4$,

   $t = t' + \tilde{k} \leq 3n' - 6 + \tilde{k} = 3n - 3k - 6 + \tilde{k} < 3n - 6$.

3. If neither (1) nor (2) hold, then there exists a node $u \in M_h$ such that all its children, say $v_1, \ldots, v_h$ ($k \geq 2$), are in $M_h$. Let $N'$ be the sTSTC network obtained by removing all nodes $v_1, \ldots, v_h$ together with their respective children leaves, and collapsing the created elementary paths into single arcs. Notice that the node $u$ is no longer an internal tree node, but a leaf of $N'$. Then, the number of nodes in $N'$ is

   $n' = n - k + 1, \quad h' = h - k, \quad t' = t - \tilde{k} - 1$,

   where $\tilde{k} \leq k$ is the number of elementary paths that have been removed. Now, the induction hypothesis yields

   $h = h' + k \leq 2n' - 4 + k = 2n - 2k + 2 - 4 + k = 2n - k - 2 \leq 2n - 4$,

   $t = t' + \tilde{k} + 1 \leq 3n' - 6 + \tilde{k} + 1 = 3n - 3 - 3k + \tilde{k} \leq 3n - 2 - 2k \leq 3n - 6$.

Hence, in all cases, the result follows.

PROOF OF PROPOSITION 7. The assertion for $n = 2$ is straightforward: there is only one sTSTC phylogenetic network with one leaf and one sTSTC phylogenetic network with two leaves. As far as the assertion for $n = 3$ goes, it can be easily checked by means of the direct computation of all pairs of distances: the largest distance is 9, and it is reached (up to permutations of labels) only by the pair of networks depicted in Fig. 1.
Finally, in the case $n \geq 4$, we know that a sbTSTC phylogenetic network with $n$ leaves has at most $3(n-2)$ internal tree nodes and $2(n-2)$ hybrid nodes, which gives an upper bound of $5(n-2)$ for the total number of internal nodes. Now, the $\mu$-vector of the leaf $i$ is the same in any sbTSTC phylogenetic network, and therefore the $\mu$-distance between two sbTSTC phylogenetic networks is upper bounded by the sum of their numbers of internal nodes.

Combining these two upper bounds, we have that, for every pair of sbTSTC phylogenetic networks with $n$ leaves $N$ and $N'$,

$$d_\mu(N, N') \leq 2 \cdot 5(n-2) = 10(n-2).$$

It remains to display a pair of sbTSTC phylogenetic networks with $n$ leaves whose $\mu$-distance reaches this equality. Such a pair must consist of two sbTSTC phylogenetic networks with $3(n-2)$ internal tree nodes and $2(n-2)$ hybrid nodes each, and with disjoint sets of $\mu$-vectors of internal nodes.

One such pair is given by the network $N_n$ described in Example 1 (in the main paper) and the network $N'_n$ obtained from $N_n$ by interchanging on the one hand the labels 1 and 3 and on the other hand the labels 2 and 3. Fig. 2 depicts $N'_n$ side by side with $N_n$ to ease to spot the differences between these networks.

![Fig. 2. Two sbTSTC phylogenetic networks with 5 leaves at maximum $\mu$-distance.](image)

To prove that $N_n$ and $N'_n$ have disjoint sets of $\mu$-vectors of internal nodes, let us start by studying the clusters (that is, the sets of descendant leaves) of their internal nodes. We shall denote the cluster of a node $v$ in a network $N$ by $C_N(v)$, and we shall say that such a cluster is internal when $v$ is internal. Note that if two nodes have different clusters, then they must have different $\mu$-vectors.

The construction of $N_n$ from $N_{n-1}$ changes its set of internal clusters in the following way. On the one hand, every internal node of $N_{n-1}$ survives in $N_n$ and its cluster is modified as follows:

- If $1 \in C_{N_{n-1}}(v)$, then 2 is added to $C_{N_n}(v)$.
- If $2 \in C_{N_{n-1}}(v)$, then 1 is added to $C_{N_n}(v)$.
- If $n-1 \in C_{N_{n-1}}(v)$, then $n$ is added to $C_{N_n}(v)$.
- No other leaf is added to any cluster of an internal node.

On the other hand, this construction adds five new internal nodes with clusters

$$\{1, 2\}, \{2, 2\}, \{2, n\}, \{n\}, \{n-1, n\}.$$

Starting with the family of internal clusters of $N_3$ and using these rules, it is easy to prove by induction that the family of internal clusters of $N_n$ is (up to repetitions)

$$\{1, 2, 3, 4, \ldots, n\}, \{2, 3, 4, \ldots, n\}, \{3, 4, \ldots, n\}, \{4, \ldots, n\}, \ldots, \{n-1, n\}, \{n\}.$$

Now, $N'_n$ is obtained from $N_n$ by interchanging 1 with $n$ and 2 with 3, and therefore the clusters of its internal nodes can be obtained from the clusters of $N_n$ by applying this permutation. We conclude that the family of internal clusters of $N'_n$ is (again, up to repetitions)

$$\{1, 2, 3, 4, \ldots, n\}, \{1, 2, 3, 4, \ldots, n-1\}, \{1, 2, 3, 4, \ldots, n-2\}, \ldots, \{1, 2, 4, \ldots, n-1\}, \{1, 2, 4, \ldots, n-2\}, \ldots, \{1\}, \{1, 3, n\}, \ldots, \{1, 3, n-1\}, \ldots, \{1, 3\}, \ldots, \{1, 3, 5, 6, \ldots, n-1\}, \ldots, \{1, 3, 5, 6, \ldots, n\}.\]$$

A simple inspection shows that only one cluster appears in both lists: the whole $\{1, \ldots, n\}$. (Indeed, all internal clusters of $N_n$ contain the leaf $n$, except $\{1, 2\}$ and $\{2\}$. Now, on the one hand, the latter are not internal clusters of $N'_n$ and, on the other hand, every internal cluster in $N'_n$ containing $n$ also contains 1,3, while no internal cluster of $N_n$ other than $\{1, 2, 3, \ldots, n\}$ contains 1,3.)

So, if a pair of internal nodes of $N_n$ and $N'_n$ have the same $\mu$-vector, their clusters must be equal to $\{1, \ldots, n\}$. Now, both $N_n$ and $N'_n$ have exactly two nodes with cluster $\{1, \ldots, n\}$: the root and its out-degree three child $a$. The $\mu$-vectors of $a$ or $r$ in $N_n$ are different from the $\mu$-vectors of $a$ or $r$ in $N'_n$: in $N_n$, there is only one path from $r$ and from $a$ to 1, while in $N'_n$ it is clear that there is more than one such path (the parent of 1 in $N'_n$ is a hybrid node, and its two parents are descendants of both $a$ and $r$).

Therefore, $N_n$ and $N'_n$ have disjoint sets of $\mu$-vectors of internal nodes and their $\mu$-distance is $10(n-2)$.

## 2 COUNTEREXAMPLES

We have mentioned in the introduction of the main paper that the class of semi-binary tree-sibling time consistent phylogenetic networks generalizes those introduced in Nakleb’s thesis. Namely, the latter are obtained from a phylogenetic tree by repeating the following procedure:

- If $1 \in C_{N_{n-1}}(v)$, then 2 is added to $C_{N_n}(v)$.
- If $2 \in C_{N_{n-1}}(v)$, then 1 is added to $C_{N_n}(v)$.
- If $n-1 \in C_{N_{n-1}}(v)$, then $n$ is added to $C_{N_n}(v)$.
- No other leaf is added to any cluster of an internal node.

On the other hand, this construction adds five new internal nodes with clusters

$$\{1, 2\}, \{2, 2\}, \{2, n\}, \{n\}, \{n-1, n\}.$$
1. choose a pair of arcs \((u_1, v_1)\) and \((u_2, v_2)\) in the tree;
2. split these arcs by introducing intermediate nodes \(u_1\) (that will become a tree node) and \(u_2\) (that will become a hybrid node), respectively;
3. add a new arc \((w_1, w_2)\).

Each hybrid node introduced, \(w_2\) in the notations above, has a tree sibling, namely \(v_1\). Hence, the networks obtained by this procedure are sbTSTC networks. However, the sbTSTC network in Fig. 3 cannot be obtained by the procedure above from a tree \(T\). Indeed, the described procedure cannot introduce tree nodes with out-degree greater that 2; hence node \(a\) in the phylogenetic network should also be a node of \(T\), and the out-degree of \(r\) in \(T\) would be 1, yielding a contradiction.

Fig. 3. A sbTSTC phylogenetic network that cannot be obtained by Nakhleh’s procedure.

When, in the main paper, we have defined the class of sbTSTC phylogenetic networks, we have remarked that the conditions imposed are necessary in order to single out networks by means of their \(\mu\)-representation. We give here examples of pairs of more general, non-isomorphic networks but with the same \(\mu\)-representation.

In Fig. 4 we give an example of a pair of semi-binary time consistent networks not satisfying the tree-sibling property, and having the same \(\mu\)-representation.

Consider the phylogenetic networks depicted in Fig. 5. They are binary and tree-sibling; however, they do not satisfy the time consistency condition. As it can be easily checked, both networks have the same \(\mu\)-representation.

Also the semi-binarity is a necessary condition, since the first network in Fig. 6 is time consistent and tree-sibling, but not semi-binary, and has the same \(\mu\)-representation as the second one, which is a sbTSTC network.

Fig. 4. Non tree-sibling, semi-binary time consistent networks with the same \(\mu\)-representation.

Fig. 5. Non time consistent tree-sibling networks with the same \(\mu\)-representation.

Fig. 6. Non semi-binary, tree sibling, time consistent networks with the same \(\mu\)-representation.

Fig. 7. Networks with hybrid children of hybrid nodes and the same \(\mu\)-representation.

To conclude with this series of counterexamples, the condition that the single child of a hybrid node is a tree node is also necessary, as the networks in Fig. 7, both with the same \(\mu\)-representation, show.

### 3 COMPUTATIONAL EXPERIMENTS

As a previous step to releasing the next version of our Perl package for computing with phylogenetic networks, we have made a Python script to generate all the sbTSTC phylogenetic networks with up to 5 leaves.

We give in Table 1, for \(n = 3, 4, 5\), the number of networks with \(n\) leaves, the CPU time \(T_1\) to load them from a file in eNewick format and compute their \(\mu\)-representations, the number of pairs of networks to compute their \(\mu\)-distances, and the CPU time \(T_2\) to make this computation. These computations have been made on an Intel Core2 Duo @ 2.33Ghz.

| \(n\) | \# networks | \(T_1\) | \# pairs | \(T_2\) |
|---|---|---|---|---|
| 3 | 10 | <0.1 sec | 45 | <0.1 sec |
| 4 | 444 | 3.4 sec | 98 346 | 1.2 sec |
| 5 | 61 176 | 11 min 49 sec | 1 871 220 900 | 7 h 56 min |

Table 1. Number of sbTSTC networks and computation times to load them and compute all-against-all distances.
We show in Fig. 8 the distribution of (non-normalized) $\mu$-distances on the set of all sbTSTC phylogenetic networks with 4 and 5 leaves.

The following files are available in the Supplementary Material web page:

- `{3,4,5}-sbTSTC.dat`: File with the eNewick strings of the sbTSTC phylogenetic networks with 3, 4 and 5 leaves.
- `{3,4,5}-sbTSTC-stats.dat`: File with the number of pairs of networks with 3, 4 and 5 leaves at each of the possible distances.

**Fig. 8.** Distribution of $\mu$-distances on the set of all sbTSTC phylogenetic networks with 4 (top) and 5 (bottom) leaves.