An algebraic metric for phylogenetic trees*

Ricardo Alberich, Gabriel Cardona, Francesc Rosselló
Department of Mathematics and Computer Science, University of the Balearic Islands,
E-07122 Palma de Mallorca (Spain)
E-mail: {r.alberich,gabriel.cardona,cesc.rossello}@uib.es

Gabriel Valiente
Department of Software, Technical University of Catalonia
E-08034 Barcelona (Spain)
E-mail: valiente@lsi.upc.edu

Abstract

The definition of similarity measures for phylogenetic trees has been motivated by the computation of consensus trees, the search by similarity in databases, and the assessment of phylogenetic reconstruction methods. The transposition distance for fully resolved trees is a recent addition to the extensive collection of available metrics for comparing phylogenetic trees. In this paper, we generalize the transposition metric from fully resolved to arbitrary phylogenetic trees, through a construction that involves an embedding of the set of phylogenetic trees (up to isomorphisms) with a fixed number of labeled leaves into a symmetric group. We also show that this transposition distance can be computed in linear time and we establish some of its basic properties.

Keywords: Comparison of phylogenetic trees, permutations, linear time algorithms

1 Introduction

The need for comparing phylogenetic trees arises when alternative phylogenies are obtained using different phylogenetic methods or sequences of different genes for a given set of species. The comparison of phylogenetic trees is also used to assess the stability of reconstruction methods as well as in the comparative analysis of clustering results obtained using different methods or different distance matrices, and it is also essential to performing phylogenetic queries on databases. Many metrics for phylogenetic tree comparison have been proposed so far: among others, the Robinson-Foulds metric, the nearest-neighbor interchange metric, the subtree transfer distance, the triples metric, and several nodal distances. One of the most recently proposed such distances is the transposition distance for fully resolved, or binary, phylogenetic trees [9].

In this paper, we propose a new metric between phylogenetic trees, which generalizes the aforementioned transposition distance for fully resolved trees, and that we consistently call hence the *transposition distance*. This distance is induced by the canonical distance for permutations through an embedding of the set of isomorphism classes phylogenetic trees with leaves bijectively labeled in a set S into a certain symmetric group of permutations, and it is directly inspired on

*This work has been partially supported by the Spanish DGI projects MTM2006-07773 COMGRIO and MTM2006-15038-C02-01.

the one hand by the matching representation of fully resolved phylogenetic trees [3] and on the other hand by the *involution metric* for RNA contact structures [6].

We establish some basic properties of our transposition distance, like for instance its diameter, and in particular we show that it can be computed in linear time, and thus it is one of the only two known linear time metrics for arbitrary phylogenetic trees, together with the Robinson-Foulds metric [7]. But, against what happens with the latter [2], the linear time computation of the transposition distance does not need the use of sophisticated algorithms and data structures.

We have implemented in Python the algorithms for the transposition distance, as well as other distances, and have made some computational experimentations, see section 4 for details.

2 Matching Representation of Phylogenetic Trees

Throughout this paper, by a *phylogenetic tree* on a set S of *taxa* we mean a rooted tree without out-degree 1 nodes and with its leaves bijectively labeled in S. We shall use the following terminology: the *children* of a node v in a phylogenetic tree $T = (V, E)$ are those nodes $w \in V$ such that $(v, w) \in E$; the set of leaves of T is denoted by $\mathcal{L}(T)$; the nodes of T that are not leaves are called *internal*; the *height* of a node v in a tree T is the length of a longest directed path from v to a leaf.

We consider the set S ordered, and although in applications it can be any set of extant species, in this paper we shall always take $S = \{1, \ldots, n\}$, ordered in the usual way. We shall denote by $\mathcal{T}_n$ the set of all phylogenetic trees with n leaves labeled $1, \ldots, n$ (up to label-preserving isomorphisms of rooted trees).

Definition 1. The *bottom-up ordering* (cf. [3, 8]) of a phylogenetic tree $T = (V, E) \in \mathcal{T}_n$ is the injective mapping $\ell : V \to \{1, \ldots, |V|\}$ defined by the following properties: (a) If $v \in \mathcal{L}(T)$, then $\ell(v)$ is its label; (b) If $\mathit{height}(u) < \mathit{height}(v)$, then $\ell(u) < \ell(v)$; (c) If $0 < \mathit{height}(u) = \mathit{height}(v)$ and

$$\min\{\ell(x) \mid x \in \text{children}(u)\} < \min\{\ell(x) \mid x \in \text{children}(v)\},$$

then $\ell(u) < \ell(v)$.

It is easy to notice that this bottom-up ordering is unique, and it can be computed in time linear in the size of the tree, and hence linear in n, by bottom-up tree traversal techniques [8, 9]. First, the leaves of T are labeled by their labels in $\{1, \ldots, n\}$. Then, the height 1 nodes are labeled from $n + 1$ on in the order given by the smallest label of their children: i.e., the height 1 node with the smallest child label is assigned the label $n + 1$, the height 1 node with the next-smallest child label is assigned the label $n + 2$, etc. And this procedure is continued for consecutively increasing heights: see Fig. 1 for an example.

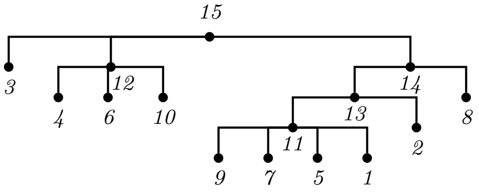

Figure 1: A bottom-up ordered phylogenetic tree.

The next definition generalizes the matching representation of fully resolved trees [3].

Definition 2. Let $T = (V, E)$ be a phylogenetic tree with n leaves labeled $1, \ldots, n$, and let $\ell : V \to \{1, \ldots, |V|\}$ be its bottom-up ordering. The *matching representation* $M(T)$ of T is the partition of $\{1, \ldots, |V| - 1\}$ defined as follows:

$$M(T) = \{\ell(\mathit{children}(u)) \mid u \in V \setminus \mathcal{L}(T)\}.$$

It is clear that, once the bottom-up ordering of T has been obtained, the partition $M(T)$ can be produced in linear time.

Example 3. The matching representation of the tree in Fig. 1 is the partition of $\{1, \ldots, 14\}$ given by

$$\Big\{\{1, 5, 7, 9\}, \{4, 6, 10\}, \{2, 11\}, \{8, 13\}, \{3, 12, 14\}\Big\}.$$

The following result establishes that the matching representations single out phylogenetic trees. We leave its easy proof to the reader.

Proposition 4. *For every $T_1, T_2 \in \mathcal{T}_n$, if $M(T_1) = M(T_2)$, then $T_1 = T_2$.* □

3 The Transposition Distance

For every $m \geqslant 1$, let $\mathcal{S}_m$ denote the symmetric group on $\{1, \ldots, m\}$. The *cycle associated to a subset* $X = \{i_1, \ldots, i_k\}$, with $i_1 < \cdots < i_k$ and $k \geqslant 2$, of $\{1, \ldots, m\}$, is $\kappa(X) := (i_1, i_2, \ldots, i_k) \in \mathcal{S}_m$. The *length* of a cycle $(i_1, i_2, \ldots, i_k)$ is the number k of elements it moves.

Definition 5. The *matching permutation* $\pi(T)$ associated to a phylogenetic tree $T = (V, E) \in \mathcal{T}_n$ is the permutation of $\{1, \ldots, |V| - 1\}$ defined by the product of the cycles associated to the members of its matching representation:

$$\pi(T) = \prod_{u \in V \setminus \mathcal{L}(T)} \kappa(\ell(\mathit{children}(u)))\,.$$

Remark 6. If $u, v \in V \setminus \mathcal{L}(T)$ are two different internal nodes of T, then $\ell(\mathit{children}(u)) \cap \ell(\mathit{children}(v)) = \emptyset$. Therefore, all cycles $\kappa(\ell(\mathit{children}(u)))$ appearing in the product defining $\pi(T)$ are disjoint to each other, and hence they commute with each other. This implies that the product t yielding $\pi(T)$ is well defined.

Example 7. The matching permutation associated to the tree in Fig. 1 is the product of cycles

$$(1, 5, 7, 9)(4, 6, 10)(2, 11)(8, 13)(3, 12, 14) \in \mathcal{S}_{14}.$$

No element in $\{1, \ldots, |V| - 1\}$ remains fixed under $\pi(T)$, because every $\ell(\mathit{children}(u))$, with u internal, has at least two elements and every element in $\{1, \ldots, |V| - 1\}$ is the bottom-up ordering label of a child of some internal node. Now, if $T = (V, E)$ is a phylogenetic tree with n leaves, then $|V| \leqslant 2n - 1$, the equality holding if and only if T is binary. To be able to compare matching permutations of phylogenetic trees with the same number of leaves n but different numbers of internal nodes, we shall understand henceforth that the matching permutation $\pi(T)$ belongs to $\mathcal{S}_{2n-2}$, leaving fixed the elements $|V|, \ldots, 2n - 2$.

The next result is a direct consequence of the fact that the matching representation of a phylogenetic tree uniquely determines it (Proposition 4) and every permutation has a unique decomposition as a product of disjoint cycles of length $\geqslant 2$.

Proposition 8. *For every $T_1, T_2 \in \mathcal{T}_n$, if $\pi(T_1) = \pi(T_2)$, then $T_1 \cong T_2$.* □

Since the mapping $\pi : \mathcal{T}_n \to \mathcal{S}_{2n-2}$ that sends every $T \in \mathcal{T}_n$ to its matching permutation $\pi(T)$ is injective, any metric on $\mathcal{S}_{2n-2}$ induces a metric on $\mathcal{T}_n$ through it. Using with this purpose the metric that associates to each pair of permutations (π_1, π_2) the least number of transpositions necessary to represent $\pi_2^{-1} \cdot \pi_1$ and arguing as in [6, Cor. 1], we have the following result.

Theorem 9. *The mapping that associates to every pair (T_1, T_2) of phylogenetic trees with n leaves labeled in $\{1, \ldots, n\}$, the least number $d'_{tr}(T_1, T_2)$ of transpositions necessary to represent the permutation $\pi(T_2)^{-1}\pi(T_1)$, is a metric on $\mathcal{T}_n$.* □

Remark 10. Recall that the least number of transpositions required to represent a cycle of length k is $k-1$ and that the least number of transpositions required to represent a product of *disjoint* cycles is the sum of the least numbers of transpositions each cycle decomposes into, and hence the sum of the cycles' lengths minus the number of cycles.

Proposition 11. *For every $T_1, T_2 \in \mathcal{T}_n$, $d'_{tr}(T_1, T_2)$ is an even integer.*

Proof. If each T_i, for $i = 1, 2$, has m_i internal nodes, then $\pi(T_i)$ decomposes into m_i disjoint cycles: say $\pi(T_i) = C_{i,1} \cdots C_{i,m_i}$, with each $C_{i,j}$ of length $k_{i,j}$. Then, by Remark 10, $\pi(T_i)$ has a decomposition into $\sum_{j=1}^{m_i}(k_{i,j} - 1) = \sum_{j=1}^{m_i} k_{i,j} - m_i = n + m_i - 1 - m_i = n - 1$ transpositions. But then $\pi(T_2)^{-1}\pi(T_1)$ admits a decomposition into $2(n-1)$ transpositions. This entails that *every* decomposition of this permutation into a product of transpositions must involve an even number of them, and therefore that $d'_{tr}(T_1, T_2)$ is an even integer. □

In other words, this metric d'_{tr} has a 'redundant' 2 factor.

Definition 12. The *transposition distance* on $\mathcal{T}_n$ is

$$\begin{array}{rlcl} d_{tr} & : \mathcal{T}_n \times \mathcal{T}_n & \to & \mathbb{N} \\ & (T_1, T_2) & \mapsto & \frac{1}{2} d'_{tr}(T_1, T_2) \end{array}$$

The transposition distance $d_{tr}(T_1, T_2)$ between two phylogenetic trees $T_1, T_2 \in \mathcal{T}_n$ can be easily calculated in linear time, by first computing the tables of values of $\pi(T_1)$ and $\pi(T_2)^{-1}$ from their decompositions into disjoint cycles (that is, from the matching representations of T_1 and T_2), then computing the composition of these permutations, and finally decomposing the resulting permutation into the product of disjoint cycles and then applying Remark 10.

Proposition 13. *For every $n \geqslant 3$, the diameter of $\mathcal{T}_n$ under d_{tr} is $n-2$.*

Proof. Let us prove first that $d_{tr}(T_1, T_2) \leqslant n - 2$ for every $T_1, T_2 \in \mathcal{T}_n$. Indeed, the permutation $\pi(T_2)^{-1}\pi(T_1)$ belongs to $\mathcal{S}_{2n-2}$, and therefore, by Remark 10, a minimal decomposition of this permutation into transpositions will involve at most $(2n-2) - 1$ transpositions. Therefore, $d'_{tr}(T_1, T_2) \leqslant 2n - 3$, and since $d'_{tr}(T_1, T_2)$ is an even number, $d'_{tr}(T_1, T_2) \leqslant 2n - 4$, and hence $d_{tr}(T_1, T_2) \leqslant n - 2$.

It remains to show a pair of phylogenetic trees in $\mathcal{T}_n$ at transposition distance $n-2$. Let $T_1, T_2 \in \mathcal{T}_n$ be the binary phylogenetic trees described by the Newick strings

$$T_1 : ((\ldots((((1,2),3),4),5)\ldots,n-1),n), \quad T_2 : ((\ldots((((2,3),4),5),6)\ldots,n),1)$$

Their matching permutations are

$$\pi(T_1) = (1,2)(3,n+1)(4,n+2)\cdots(n,2n-2), \ \pi(T_2) = (2,3)(4,n+1)(5,n+2)\cdots(1,2n-2)$$

and therefore

$$\pi(T_2)^{-1}\pi(T_1) = (1,3,5,7,\ldots,2n-3)(2n-2,2n-4,2n-6,\ldots,2)$$

which shows that $d_{tr}(T_1, T_2) = \frac{1}{2}(2n-2-2) = n-2$. □

In the Introduction we mentioned that the transposition distance defined in this paper generalizes the transposition distance for fully resolved phylogenetic trees introduced in [9]. This will be a direct consequence of [9, Thm. 1] and the following result; the example given in the proof of the last proposition is a special case of its proof.

Proposition 14. *For every pair of fully resolved phylogenetic trees $T_1, T_2 \in \mathcal{T}_n$, let $G = (V, E)$ be the undirected multigraph with $V = \{1, \ldots, 2n-2\}$ and $E = M(T_1) \sqcup M(T_2)$, and let κ be the number of connected components of G. Then, $d_{tr}(T_1, T_2) = n - 1 - \kappa$.*

Proof. If T_1 and T_2 are fully resolved, then $\pi(T_1)$ and $\pi(T_2) = \pi(T_2)^{-1}$ are products of disjoint transpositions and have not fixed point. Then, every connected component of G corresponds to 2 disjoint cycles in the decomposition of $\pi(T_2) \cdot \pi(T_1)$ and therefore, by Remark 10, $d_{tr}(T_1, T_2) = \frac{1}{2}(2n - 2 - 2\kappa) = n - 1 - \kappa$. □

4 Computational Experiments

We have implemented all the algorithms described in this paper in `PhyloNetwork.py`, a Python package which also deals with phylogenetic networks and computes, among other things, the Robinson-Foulds [7] and splitted nodal [1] distances. We have also implemented the algorithm described in [5] for the generation of uniformly distributed random trees with a given set of taxa, as well as an adaptation of it for the sequential generation of all trees (which only makes sense for an small number of leaves), in the Python package `TreeGenerator.py`. Both packages will be shortly available to the public domain.

Using the aforementioned packages we have generated all trees with up to 7 taxa, and random samples (each one with approximately 20 000 trees) of trees with from 8 to 14 taxa. For each pair with the same taxa, we have computed their transposition, Robinson-Foulds, and splitted nodal distances. In Figure 2 we give histograms of the distributions of these three distances for $n = 7$ and $n = 14$ leaves (which are the most significative ones where we have generated, respectively, the set of all trees and a random sample). In the supplementary material webpage `http:/bioinfo.uib.es/~recerca/phylotrees/transdist/` we provide the data for the remaining cases.

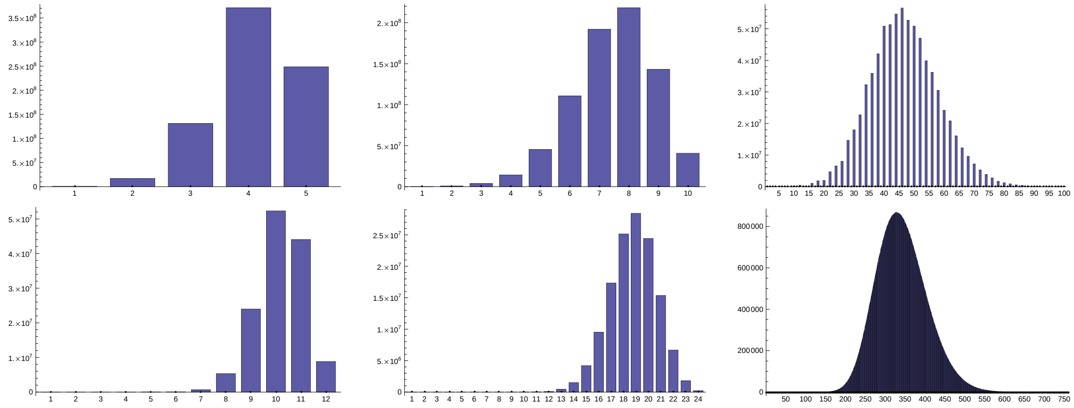

Figure 2: Histograms of distributions of the transposition (left), Robinson-Foulds (center) and splitted nodal (right) distances for trees with 7 leaves (top) and 14 leaves (bottom).

A parameter that shows how two different distances within the same set are related is the Spearman's rank correlation. In Table 1 we give the correlations between the Robinson-Foulds, the splitted nodal and the transposition distances, for different values of the number of leaves. Since the correlations of the transposition distance with the other ones are small enough, the distance we have defined is not related (from an ordinal point of view) to the other known ones.

Table 1: Spearman's rank correlation for the transposition (TR), Robinson Foulds (RF), and splitted nodal (SN) distances for small number n of leaves.

n	TR/RF	TR/SN	RF/SN
3	1.0000	1.0000	1.0000
4	0.4757	0.5258	0.8247
5	0.4651	0.4485	0.7419
6	0.4520	0.3900	0.6658
7	0.4480	0.3481	0.6039
8	0.4509	0.3213	0.5589
9	0.4536	0.2955	0.5152
10	0.4629	0.2882	0.4968
11	0.4705	0.2740	0.4723
12	0.4728	0.2616	0.4516
13	0.4747	0.2506	0.4344
14	0.4834	0.2464	0.4238

5 Conclusions

In this paper we have defined and analyzed a metric for arbitrary phylogenetic trees on a given set of taxa that generalizes the transposition distance for fully resolved phylogenetic trees and that can be computed in linear time. This metric adds to the number of other metrics for phylogenetic trees defined so far. As Moulton, Zuker *et al* claimed in the context of RNA secondary structure comparison, "[...] generally speaking, it is probably safest to try as many metrics as possible" [4, p. 290].

References

[1] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, Nodal metrics for rooted phylogenetic trees, submitted (2008).

[2] W. Day, Optimal algorithms for comparing trees with labeled leaves, Journal of Classification 2 (1) (1985) 7–28.

[3] P. W. Diaconis, S. P. Holmes, Matchings and phylogenetic trees, Proc. Natl. Acad. Sci. USA 95 (1998) 14600–14602.

[4] V. Moulton, M. Zuker, M. Steel, R. Pointon, D. Penny, Metrics on RNA secondary structures, Journal of Computational Biology 7 (2000) 277–292.

[5] N. L. Oden, K. Shao, An algorithm to equiprobably generate all directed trees with k labeled terminal nodes and unlabeled interior nodes, Bulletin of Mathematical Biology 46 (3) (1984) 379–387.

[6] C. Reidys, P. F. Stadler, Bio-molecular shapes and algebraic structures, Computers & Chemistry 20 (1) (1996) 85–94.

[7] D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees, Mathematical Biosciences 53 (1/2) (1981) 131–147.

[8] G. Valiente, Algorithms on Trees and Graphs, Springer, 2002.

[9] G. Valiente, A fast algorithmic technique for comparing large phylogenetic trees, in: Proc. 12th Int. Symp. String Processing and Information Retrieval, Vol. 3772 of Lecture Notes in Computer Science, Springer, Berlin, 2005, pp. 370–375.