ON A MATCHING DISTANCE BETWEEN ROOTED PHYLOGENETIC TREES

DAMIAN BOGDANOWICZ, KRZYSZTOF GIARO

Department of Algorithms and System Modeling, Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland
e-mail: {Damian.Bogdanowicz,giaro}@eti.pg.gda.pl

The Robinson–Foulds (RF) distance is the most popular method of evaluating the dissimilarity between phylogenetic trees. In this paper, we define and explore in detail properties of the Matching Cluster (MC) distance, which can be regarded as a refinement of the RF metric for rooted trees. Similarly to RF, MC operates on clusters of compared trees, but the distance evaluation is more complex. Using the graph theoretic approach based on a minimum-weight perfect matching in bipartite graphs, the values of similarity between clusters are transformed to the final MC-score of the dissimilarity of trees. The analyzed properties give insight into the structure of the metric space generated by MC, its relations with the Matching Split (MS) distance of unrooted trees and asymptotic behavior of the expected distance between binary $n$-leaf trees selected uniformly in both MC and MS ($\Theta(n^{3/2})$).

Keywords: phylogenetic tree, phylogenetic tree metric, phylogenetic tree comparison, matching cluster distance, matching split distance.

1. Introduction

Phylogenetic trees (phylogenies) are widely used in research related to evolution. Such trees (sometimes also called evolutionary trees) represent the historical evolutionary relationships among different species. Present-day organisms correspond to the labels on the leaves of those trees, while ancestral species are represented by the remaining (unlabeled) vertices. A phylogenetic analysis often starts with finding an unrooted tree that describes evolutionary relationships in a group of taxa under study. This is a natural consequence of commonly used models of the evolution of DNA or amino acid sequences in the form of a time-reversible process, e.g., the reversible Markov chain. However, a complete solution to the phylogeny problem requires finding a rooted tree, which gives information about the location of a common ancestor of the group of taxa under study and defines the time flow along branches (from the root to leaves).

Many popular methods for constructing phylogenetic trees (e.g., the distance, parsimony, maximum likelihood, Bayesian approaches; see the work of Felsenstein (2003) for a review) often result in different trees for the same input data, and an important problem is to determine how distant are two reconstructed trees from each other. This is the most common application of phylogenetic metrics, but phylogenetic tree distances are used not only for simple comparison of slightly different results. There are many other applications of the distances, e.g., mining phylogenetic information databases (Wang et al., 2005), defining the consensus and median point of trees (Bryant, 1997), the postprocessing of Bayesian phylogenetic analysis results with clustering techniques (Stockham et al., 2002) or its visualizations (Hillis et al., 2005), analyzing sets of gene trees using the TreeOfTrees method (Darlu and Guénoche, 2011). It is worth noting that analyzing data using clustering and other grouping methods based on some measure of distance or similarity is a common approach used across many areas in bioinformatics, not only in phylogenetics (see, e.g., Frąckiewicz and Palus, 2011; Biedrzycki and Arabas, 2012). Moreover, a relatively new application concerns using polynomially computable tree comparison metrics in constructing heuristic algorithms for detecting Horizontal Gene Transfers (HGTs) (for details, see Boc et al., 2010). Phylogenetic tree distances are also useful in different branches of science, e.g., computer science—analysis of malware evolution (Hayes et al., 2009), chemistry (Restrepo et al., 2007) or linguistics (Penny et al., 1993; Pompei et al., 2011).

For rooted trees, there is a deficiency of metrics in
the literature. Therefore new definitions appear constantly, e.g., the triples distance (Crichtlow et al., 1996; Bansal et al., 2011), the symmetric duplication cost (Ma et al., 1998), the Minimal Agreement Partition (MAP) metric (Bolikowski and Gambin, 2007), the transposition distance (Alberich et al., 2009), or nodal (Williams and Clifford, 1971) and splitted nodal metrics (Cardona et al., 2010).

In a recent work (Bogdanowicz and Giaro, 2012) we presented a general method for creating matching metrics for unrooted phylogenetic trees (not necessarily binary) and an example of a new metric constructed using the method—the Matching Split (MS) distance with interesting properties. Moreover, the usability and desirable properties of the MS distance have been confirmed by a recent work of Lin et al. (2012). In particular, the authors showed that the MS metric performs significantly better than the well-known Robinson–Foulds distance (Robinson and Foulds, 1981) when applied to the clustering of phylogenetic trees.

The main idea of those matching distances is based on comparing splits that correspond to the edges of the analyzed trees using an arbitrary metric $h$. This approach is extended to entire phylogenetic trees in such a manner that the value of the distance is equal to the weight of a minimum-weight perfect matching in a complete bipartite graph constructed based on the trees and the function $h$.

In this article we want to show that the described approach can be generalized or transferred to more complex structures, i.e., rooted trees. As a case study we consider in detail the properties of the simplest metrics defined by this method, i.e., the Matching Cluster (MC) distance.

The proposed definitions can be regarded as an extension of the most popular method of comparing phylogenetic trees—the Robinson–Foulds distance (Robinson and Foulds, 1981). The values of the proposed distances can be effectively computed using algorithms with polynomial time complexity.

2. Definitions and notation

For sets $A, B$, let $A \oplus B = (A \setminus B) \cup (B \setminus A)$ be their symmetric difference. Let $|A|$ denote the cardinality of set $A$. By $2^A$ we denote the family of all subsets of $A$. Let $G = (V,E)$ be a graph with a set of vertices $V$ and a set of edges $E$. A bipartite graph $G(V_1,V_2,E)$ has vertices decomposed into two disjoint sets $V_1 \cup V_2 = V$ such that no two vertices within the same set are adjacent. A bipartite graph is complete if every two vertices $v_1 \in V_1$ and $v_2 \in V_2$ are adjacent. A tree is a connected acyclic graph. A path is a tree with two vertices of degree 1 and all the others of degree 2. A caterpillar is a graph whose subgraph induced by vertices of degree greater than 1 is a path.

A matching $M \subseteq E$ in a graph $G = (V,E)$ is a set of pairwise non-adjacent edges; that is, no two edges share a common vertex. A perfect matching covers all vertices of the graph. If we assign a weight function $w : E \rightarrow \mathbb{Z}_{\geq 0}$ to the edges of $G$, then a minimum-weight perfect matching is defined as a perfect matching where the sum of the weights of its edges has a minimum value. Minimum-weight perfect matchings in bipartite graphs can be computed efficiently in time $O(|E| \sqrt{|V|} \log(|V| \max_{e \in E} w(e)))$ (Gabow and Tarjan, 1989; Orlin and Ahuja, 1992).

A metric defined over an arbitrary set is often used to quantify a difference or a distance between any two elements of the set. A metric space is a pair $(X,d)$ consisting of a set $X$ and a function $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$ (the metric over $X$) such that (i) $d_{x,y} \leq 0 \iff x = y$, (ii) $d_{x,y} = d_{y,x}$, and (iii) $d_{x,y,z} + d_{y,z} \geq d_{x,z}$—the triangle inequality.

Besides this common notation, the phylogenetic literature uses also a set of specific terms, a part of which we recall in this section (see Bryant, 1997; Semple and Steel, 2003).

A rooted phylogenetic tree $T = (V,E)$ is a tree whose leaves, that is, vertices of degree one are labeled bijectively by the elements of a finite set $L$ (representing the species), each non-leaf vertex is unlabeled, there is exactly one distinguished non-leaf vertex $r(T) \in V \setminus L$ called the root and none of the vertices of $V \setminus \{r(T)\}$ has degree two.

Present-day species under examination form the finite set $L$ and are represented by leaves of a tree. Internal vertices, i.e., members of $V \setminus L$; represent hypothetical ancestors of the taxa of $L$, in particular, $r(T)$ is the ancestor of all species under study. For the sake of simplicity, we can identify leaves with their labels, i.e., for a phylogenetic tree $T$, by $L(T)$ we denote a set of leaves of $T$ or a set of labels of those leaves. This general definition includes trees for which the sequence of speciation events may not be fully resolved. Such a situation, called multifurcation, concerns the vertex (node) in a tree that is incident to more than three edges (branches). A multifurcation may represent the lack of resolution due to insufficient data available for inferring the phylogeny. Multifurcations can also occur in consensus trees (see the work of Bryant (1997) for a review of consensus methods) that present information common to a set of partially contradictory trees obtained from, e.g., maximum parsimony. The most informative are binary (fully resolved) phylogenetic trees.

A rooted binary phylogenetic tree is a rooted phylogenetic tree such that the root has degree 2 and all other internal vertices have degree 3. In each rooted binary phylogenetic tree $T$ over the set of leaves $L$, there are $|L| - 2$ internal edges (i.e., not pendant) and $|L| - 1$
internal vertices. In a non-binary tree these numbers are smaller. By $R_L$ and $R_B^L$ we denote the sets of all rooted phylogenetic trees and all rooted binary phylogenetic trees over the set of leaves $L$, respectively. For $L = \{1, \ldots, n\}$, we use the notation $R_n$ and $R_n^B$. A rooted tree $T$ defines a partial order relation of being descendant (and ancestor) on its vertices, denoted by $\preceq_T$. For $a, b \in V(T)$, we have $a \preceq_T b$ (that is, $a$ is a descendant of $b$) if the path in $T$ from $a$ to $r(T)$ contains $b$. In particular, $v \preceq_T r(T)$ and $v \preceq_T v$ for any $v \in V(T)$. To every vertex $v$ we can assign its cluster $c(v) \subseteq L$, i.e., the set of leaves (labels) that are descendants of $v$. There are $|L| + 1$ trivial clusters in a tree $T$ that are related to leaves $u$ (where $c(u) = \{u\}$) and to the root (where $c(r(T)) = L(T)$); all the other clusters are non-trivial. By $\sigma(T)$ and $\sigma_r(T)$ we denote families of all clusters of $T$ and all non-trivial clusters of $T$, respectively. Therefore for a tree $T \in R_L$ we have $|\sigma(T)| \leq 2|L| - 1$, $|\sigma_r(T)| \leq |L| - 2$, and both inequalities are tight for binary trees. A rooted phylogenetic tree $T$ is uniquely described by a set $\sigma_r(T)$ and the translation between these two descriptions can be performed efficiently in linear time (see Semple and Steel, 2003, Section 3.5).

In order to compare phylogenetic histories represented by trees $T_1, T_2 \in R_L$, the structure of a metric space in the set $R_L$ is introduced. One of the most widely used metrics on a set $R_L$ is the Robinson–Foulds distance (Robinson and Foulds, 1981) based on clusters:

**Definition 1.** The *Robinson–Foulds (RF) distance* between two rooted trees $T_1, T_2 \in R_L$ is defined as

$$d_{RF}(T_1, T_2) = \frac{1}{2} |\sigma(T_1) \oplus \sigma(T_2)|.$$  

(1)

The relationship among the species without the knowledge of the location of the common ancestor nor about the time flow along edges is illustrated by an unrooted phylogenetic tree. An *unrooted phylogenetic tree* is a tree whose leaves (vertices of degree one) are labeled bijectively by the elements of a finite set $L$ (species) and no vertex has degree 2. An unrooted phylogenetic tree is *binary* if each non-leaf vertex has degree 3.

Let $U_L$ and $U_L^B$ denote the set of all unrooted phylogenetic trees and the set of all unrooted binary trees over the set of leaves $L$, respectively. For $L = \{1, \ldots, n\}$, use the notation $U_n$ and $U_n^B$. In each $T \in U_L^B$ there are $|L| - 2$ internal vertices and $|L| - 3$ internal edges. In an unrooted tree there is no descendant–ancestor relation, but, analogically to the correspondence between vertices and clusters in the rooted case, we now have a similar relation between edges and *splits*. A *split* $A|B$ of a set $L$ is an unordered pair (i.e., $A|B = B|A$) of its nonempty subsets such that $L = A \cup B$ and $A \cap B = \emptyset$. Let $\min(A|B) = \min(|A|, |B|)$ and, if $\min(A|B) = 1$, then $A|B$ is trivial, otherwise it is non-trivial (Bryant, 1997). A set of all splits of a finite set $L$ is denoted as $Splits(L)$. For $T \in U_L$ and an edge $e$ of $T$, removing $e$ divides $T$ into two components. Let $A$ be the set of leaves in one of them and $B$ in the other. Then $A|B$ is a split of $L(T)$ corresponding to $e$. The set of splits corresponding to edges in $T \in U_L$ is denoted by $\beta(T)$ (Bryant, 1997), thus $|\beta(T)| \leq 2|L| - 3$ and there are $|L|$ trivial splits. The remaining (non-trivial) splits form the set $\beta_r(T)$. Similarly as in the rooted case, there is a linear time algorithm for reconstructing a tree from the set of its splits (Gusfield, 1991).

A well-known metric used for comparing unrooted phylogenetic trees is the RF distance (Robinson and Foulds, 1981), defined analogically to (1), where $\beta(T)$ is used instead of $\sigma(T)$.

Inferring the root of an unrooted phylogenetic tree from $U_L$ is usually called *rooting*. The root is usually introduced in an internal vertex or on an edge by adding a new vertex dividing this edge, so we obtain a tree from $R_L$. *Unrooting* is the opposite operation transforming a rooted tree into an unrooted one.

Finally, consider trees that contain less phylogenetic information than $T \in R_L$. For an internal edge $e = \{u, v\}$ of $T$ we define contracting an edge $e$ as an operation that transforms $T$ into the tree $T_e \in R_L$, in which $e$ is removed and the vertices $u$ and $v$ are identified. Let $A \subseteq L$, and let $T(A)$ be a minimal subgraph of $T$ that connects leaves of $A$ and choose as its root the vertex closest to $r(T)$. The *subtree of $T$ induced by $A$* is a tree $T_{|A} \in R_A$ obtained from $T(A)$ by successively removing all vertices of degree 2 (with exception of the root) and identifying their adjacent edges. Hence $T_{|A}$ is the tree containing the whole phylogenetic information from $T$, but only concerning the species from $A$ (these definitions are often used in mathematical phylogenetics; for details, see the work of Bryant (1997)).

### 3. Matching method for rooted and unrooted phylogenetic trees

The crucial point for the proposed method is describing a way for defining the distance between finite subsets of a given metric space (see Bogdanowicz and Giaro, 2012). Lemmas 3.2 and Definition 3.1 here correspond to Lemmas 3.3.2 and Definition 3.1 in the above-mentioned work, respectively, where we considered the sets of splits.

**Lemma 1.** There are given a metric space $(X, d)$, a complete bipartite graph $G(V_1, V_2, E)$, $|V_1| = |V_2| = n$ and a labeling $l : V_1 \cup V_2 \rightarrow X$. We assign weights to the edges of $G$ so that $w(\{a, b\}) = d(l(a), l(b))$ for $a \in V_1$, $b \in V_2$. Let $a_1, \ldots, a_k \in V_1$, $b_1, \ldots, b_k \in V_2$. If $l(a_i) = l(b_i)$ for $1 \leq i \leq k$ then there exists a minimum-weight perfect matching $M \subseteq E$ satisfying $\{a_i, b_i\} \in M$ for $1 \leq i \leq k$. 

---

1 A version of this definition without factor $1/2$ is also used in the literature.
Definition 2. There are given a finite set $D$ and a metric $h$ on $D \cup \{O\}$. We define a metric $d_h : 2^D \times 2^D \to \mathbb{R}_{\geq 0}$, where the distance between $A, B \in 2^D$, $d_h(A, B)$, is equal to the value of a minimum-weight perfect matching in a complete bipartite graph $G = (V_1, V_2, E)$ defined as follows:

- for arbitrary $s, t$ such that $s - t = |A| - |B|$, we define the sets $V_1 = \{a_1, \ldots, a_{|A|}, b_{|A|}, \ldots, b_{|A|+t}\}$, $V_2 = \{b_1, \ldots, b_{|B|}, b_{|B|+1}, \ldots, b_{|B|+s}\}$, as the vertices partitions of the graph $G(V_1, V_2, E)$ and vertex labeling $l : V_1 \cup V_2 \to D \cup \{O\}$, so that $A = \{l(a_i) : 1 \leq i \leq |A|\}$, $B = \{l(b_j) : 1 \leq j \leq |B|\}$ and $l(a_i) = l(b_j) = O$ for $|A|+1 \leq i \leq |A|+t$, $|B|+1 \leq j \leq |B|+s$;

- the weights of the edges are defined using the metric $h$ as $w((a_i, b_j)) = h(l(a_i), l(b_j))$.

Lemma 2. The function $d_h$ is a metric and the value of $d_h(A, B)$ does not depend on $s$ or $t$ (when $s - t = |A| - |B|$).

Hence, we can always assume that $\min\{s, t\} = 0$ and $\max\{s, t\} = |A| - |B|$. The distance $d_h(A, B)$ can be interpreted as the total cost of the most accurate pairing between the elements of $A$ and $B$. The value $h(O, x)$ is the cost of leaving the element $x$ unmatched. Note that, if $|A| = |B|$, then considering the element $O$ and defining its distance $h(x, O)$ for $x \in D$ is unnecessary.

The presented method gives a convenient way to define metrics for phylogenetic trees (i.e., introducing a metric space in $2^L$ and $U_L$) based on any metric $h$ on subsets of $L$ and $\text{Splits}(L) \cup \{O\}$ appropriately. First note that the classical RF metric defined by the formula \[1\] can be expressed using Definition \[2\] and Lemma \[2\] by taking $D = 2^L \setminus \{\emptyset\}$, $O = \emptyset$ and the following simple function $h_{RF} : 2^L \times 2^L \to \{0, 0.5, 1\}$ for clusters comparison: $h_{RF}(c_1, c_2) = 1$ if $c_1 \neq c_2$ and $c_1, c_2 \neq \emptyset$, and $h_{RF}(c_1, c_2) = 0.5$ if exactly one of $c_1, c_2$ is an empty set. Hence we obtain $d_{RF}(T_1, T_2) = d_{h_{RF}}(\sigma(T_1), \sigma(T_2)) = d_{h_{RF}}(\sigma^*(T_1), \sigma^*(T_2))$. An analogous equality for unrooted trees can be derived using a similar $\{0, 0.5, 1\}$-valued metric on $\text{Splits}(L) \cup \{O\}$.

Using a similar approach in our earlier work (Bogdanowicz and Giaro, 2012), we described a more complex metric on $\text{Splits}(L) \cup \{O\}$:

\[ h_S(A | B_1, B_2) = \min\{|A_1 \cap A_2|, |A_1 \cap B_2|\}, \quad (2) \]

\[ h_S(A | B, O) = \min\{|A|, |B|\}, \]

and used it to define the matching split distance between unrooted trees $T_1, T_2 \in U_L$ as

\[ d_{MS}(T_1, T_2) = d_{h_S}(\beta(T_1), \beta(T_2)) = d_{h_S}(\beta^*(T_1), \beta^*(T_2)). \quad (3) \]

Thus customization of the function for comparing $h$ leads us to more useful metrics than the traditional RF. We now apply a similar approach for comparison of rooted trees introducing the matching cluster distance that can be regarded as an analog of MS for the rooted case.

The most natural manner for quantifying an amount of phylogenetic information represented by an internal vertex of a rooted tree is the size of its clade, i.e., the cardinality of the cluster related to it. Dissimilarity between two clades $A, B$ can be measured as the number of elements that appear in one of the clades but not in the other, i.e., the cardinality of the set $A \oplus B$. Since the cardinality of $A \oplus B$ introduces a metric space structure in an arbitrary family of finite sets, we obtain what follows.

Definition 3. Let $T_1, T_2 \in R_L$ be rooted phylogenetic trees, $h_C : 2^L \times 2^L \to \mathbb{Z}_{\geq 0}$ such that $h_C(A, B) = |A \oplus B|$, and let $O = \emptyset$. According to Definition \[2\] we define the matching cluster distance $d_{MC} : R_L \times R_L \to \mathbb{Z}_{\geq 0}$ as

\[ d_{MC}(T_1, T_2) = d_{h_C}(\sigma(T_1), \sigma(T_2)) = d_{h_C}(\sigma^*(T_1), \sigma^*(T_2)). \quad (4) \]

For example, we calculate the matching cluster distance between trees in Fig. \[1\] We have the following non-trivial clusters for $T_1$: $\{a, b\}$, $\{c, d\}$ and for $T_2$: $\{a, b, c\}$. Using the function $h_C$ we calculate the distances between them: $h_C(\{a, b\}, \emptyset) = 2$, $h_C(\{a, b\}, \{a, b, c\}) = 1$; $h_C(\{c, d\}, \emptyset) = 2$, $h_C(\{c, d\}, \{a, b, c\}) = 3$. The weight of a minimum-weight perfect matching in a bipartite graph shown in Fig. \[1\] is equal to 3, so $d_{MC}(T_1, T_2) = 3$.

The MC distance can be computed in time $O(|L|^{2.5} \log |L|)$ with the already mentioned weighted matching algorithms (Gabow and Tarjan, 1989; Orlin and Ahuja, 1992).

We show that the advantages of MS (Bogdanowicz and Giaro, 2012) are retained for matching metrics on rooted trees, e.g., for MC, and can be shortly summarized as follows:

- MC takes into account not only the identity of clusters, but also more subtle similarities allowing enhanced diversification;

- the maximal distance in $R_n$ for the RF metric is only $n - 2$, while in the case of MC it is $\Theta(n^2)$.
On a matching distance between rooted phylogenetic trees

(Section 6). A wider range of distance values than that of RF is also observed for other phylogenetic metrics, e.g., for triplet distance, but in this case the interpretation of the distance is not so obvious;

• the changes corresponding to edges placed near large clades are recognized as more significant than those corresponding to edges placed closer to leaves;

• an important side effect of computing matching distances is an injective mapping between internal non-root vertices in both trees, where \(|V(T_1)| - |V(T_2)|\) internal vertices of the “bigger” tree have no pair. In the particular case of binary trees, the mapping is bijective. Moreover, the mapping can be regarded as a suggestion about similar clades in both trees and can be helpful in understanding the structural difference between the analyzed trees.

Methods for determining some kind of “tree alignments” and numerical measures of trees similarity were proposed, e.g., by Munzner \textit{et al.} (2003) and Nye \textit{et al.} (2006); such an analysis was also adopted to phylogenetic networks (Cardona \textit{et al.}, 2009). A similar approach to the comparison of tree-like structures, but not in phylogenetic context, can be found in the work of Boorman and Olivier (1973).

4. Properties of the MC distance for rooted phylogenetic trees

In this section we analyze in detail the properties of one of the simplest metrics for rooted trees.


An important byproduct of calculating \(d_{MC}(T_1,T_2)\) is a “tree alignment” described by a minimum-weight matching and a corresponding pairing of internal nodes from \(T_1\) and \(T_2\). As presented in Fig. 2, such a minimum-cost mapping is not in general unique. In this case we can create two matchings \(M_1 = \{(s_{i+1},t_i)\}_{i=1}^{n-3} \cup \{(s_{1},t_{n-2})\}\) and \(M_2 = \{(s_{1},t_{1})\}_{i=1}^{n-2}\) with the same minimal cost 2\(n - 4\). Nevertheless, it is always possible to obtain a mapping that in some sense agrees with the ancestor–descendant relations between the nodes in both trees (here it is the matching \(M_2\)). No analogy of this feature is known for the MS distance of unrooted trees.

**Theorem 1.** Let \(T_1, T_2 \in R_L\), \(|V(T_1)| \leq |V(T_2)|\), and denote non-root internal vertices of these trees \(V_i = V(T_i) \setminus (L \cup \{r(T_i)\})\) for \(i = 1, 2\). There exists an injection \(f : V_1 \rightarrow V_2\) with the cost \(\sum_{v \in V_1} c(v) + c(f(v))\) such that all internal non-root vertices \(a, b \in V(T_1)\) fulfill the following conditions:

1. If \(a \leq T_1, b\), then their related vertices in \(T_2\) fulfill:
   \[f(a) \leq f(b)\]
   \(f(a) \neq f(b)\) if \(a \leq T_1, b\) or they are \(\leq T_2\)-incomparable.

2. If \(f(a) \leq f(b)\), then \(a \leq T_1, b\) or they are \(\leq T_2\)-incomparable.

In other words, there is always a minimum-cost mapping \(f\) such that it is impossible to have two different vertices \(a, b\) with contradictory relations \(a \leq T_1, b\) and \(f(b) \leq T_2\).

**Proof.** Let \(a \leq T_1, b\), \(a \neq b\) and \(b' = f(b) \leq T_2\), \(a' = f(a)\), where \(f\) corresponds to an arbitrary pairing of clusters that realizes \(d_{MC}(T_1,T_2)\). Then \(A = c(a) \not\subset c(b) = B = X \oplus A\) and \(B' = c(b') \not\subset c(a') = A' = B' \oplus X'\). We have \(|B'| + |A' X'| = |B| \oplus |A'| + 2X \cap X'|\); hence, after making modifications to \(f\) so that \(f(a) := b', f(b) := a'\), the equality \(\sum_{v} c(v) + c(f(v))\) increases. Therefore at most \(O(|L|^3)\) described consequent operations are possible, after which we obtain \(f\) that fulfills the first part of the theorem. The second part follows directly from the first one.

4.2. Structure of the MC metric space.

We now present some basic properties of MC and its relations with the most popular phylogenetic metric, i.e., RF. Some of the described properties are similar to MS, with differences in the coefficients. Thus for the completeness of the paper we only list them, skipping the proofs if they are analogous to the discussion presented for the MS distance by Bogdanowicz and Giaro (2012).

**Theorem 2.** Let \(T_1 \neq T_2 \in R_L\). We have

1. \(d_{RF}(T_1,T_2) \leq d_{MC}(T_1,T_2)\)
   \[\leq 2(|L| - 1) d_{RF}(T_1,T_2);\]

2. if \(T_1, T_2 \in R^R_L\), then
   \[d_{RF}(T_1,T_2) + 1 \leq d_{MC}(T_1,T_2)\]
   \[\leq (|L| - 1) d_{RF}(T_1,T_2).\]
Proof. See Appendix.

We list basic extreme cases of inequalities in Theorem 2. First, observe that $d_{MC}(T_1, T_2) = 1$ is possible only when $\sigma(T_1)$ and $\sigma(T_2)$ differ by only one pair of clusters $c_1 \in \sigma(T_1)$, $c_2 \in \sigma(T_2)$, $c_1 \neq c_2$ and, additionally, $c_1$ and $c_2$ differ by only one element of $L$, so their vertices must be multifurcations (see Fig. 3). Hence, we obtain the following result.

**Corollary 1.**

1. If $T_1 \in R^P_R$, then there is no $T_2 \in R_L$ such that $d_{MC}(T_1, T_2) = 1$.

2. Let $T_1 \in R_L$, $|L| = n$. Then the number of trees $T_2 \in R_L$ such that $d_{MC}(T_1, T_2) = 1$ can be estimated as $O(n^2)$ (see Fig. 4).

3. The equality $d_{RF}(T_1, T_2) = d_{MC}(T_1, T_2)$ may hold for arbitrary values of $d_{RF}(T_1, T_2)$ (see Fig. 5).

4. If $T_1, T_2 \in R^P_R$ and $d_{MC}(T_1, T_2) = 1$ and $d_{MC}(T_1, T_2) = 2$, then $d_{RF}(T_1, T_2) = 1$ and these trees differ by one operation that swaps two leaves neighboring to opposite ends of an internal edge.

5. For a tree $T_1 \in R^P_R$, the number of trees $T_2 \in R^P_R$ such that $d_{MC}(T_1, T_2) = 2$ may vary between 0 and $n - 1$ (rooted caterpillar case).

6. There exists a sequence of trees $T_1, T_2 \in R^P_R$, such that $d_{MC}(T_1, T_2) = 1$.

**Theorem 3.** There are given two trees $\tilde{T}_a, \tilde{T}_b \in R_L$.

1. There exists a sequence of trees $\tilde{T}_a = T_1, T_2, \ldots, T_{k-1}, T_k = \tilde{T}_b$, $\tilde{T}_i \in R_L$ for $i = 1, \ldots, k$ such that $d_{MC}(T_j, T_{j+1}) \leq 4$, where $j = 1, \ldots, k - 1$.

2. If $\tilde{T}_a, \tilde{T}_b \in R^P_R$, then the trees $T_i$ are binary as well.

**Proof.** (Sketch) Note that any tree can be transformed into a rooted caterpillar using a series of operations presented in Fig. 6. Two caterpillars can then be connected by a series of Operations 3. However, if $\tilde{T}_a, \tilde{T}_b \in R^P_R$, then Operation 4 is unnecessary. Trees after Operations 1 or 2 are at a distance of 4 or 3, respectively. Operations 3 and 4 create trees at a distance of 2.

It is worth noting that “isolated regions” appear for other metrics, e.g., for the triples distance (TT) and the splitted nodal metric with $L^2$ norm (SN). In both these cases the star tree, i.e., $T_n \in R_n$ with $n + 1$ vertices, is an example of such an “isolated region” because the distance between $T_n$ and any other tree $T' \in R_n \setminus \{T_n\}$ grows with the number of taxa, i.e., $d_{SN}(T, T') \geq \sqrt{2(n - 2)}$ and $d_{TT}(T, T') \geq n - 2$.

**4.3. Small topological transformations and the MC-space diameter.** One of the main advantages of MS over RF (Bogdanowicz and Giaro, 2012) is its insensitivity to small changes in the tree topology. In the RF case, the displacement of one leaf may create an unrooted tree distant from the original one by as much as $|L| - 3$, and it is the maximum possible distance in this metric. Despite the minor change, these trees seem to be very distant (in the RF metric). MS is not misleading in these situations. We will see (Theorems 4 and 5) that conducting a fixed number $k = const$ of leaf displacements or edge contractions may create a tree displaced by $O(|L|)$, but the MC-space diameter is $\Theta(|L|^2)$ (see Theorem 6). Hence, this fundamental advantage is maintained also in the MC metric. An extreme example of this property is shown in Fig. 2 where after a single modification the distance in RF increases to the maximum possible value, whereas in MC it reaches $2|L| - 4$, which is far less than the maximum value in MC ($\Theta(|L|^2)$, Theorem 6).
On a matching distance between rooted phylogenetic trees

Theorem 4. Let $T \in R_n$ and let $e$ be an internal edge of $T$. Then $d_{MC}(T, T_e) \leq n - 1$.

Proof. Note that $T_e$ has one cluster less than $T$. Let $e = c \in \sigma_{A}(T) \setminus \sigma_{A}(T_e)$. By Lemma 1 we have $d_{MC}(T, T_e) = h_{c}(e, O) = |c| \leq n - 1$.

In our earlier work (Bogdanowicz and Giaro, 2012) we studied the effects of attaching or removing a leaf on the minimum-weight perfect matching in the MS case. Here, for the MC distance we obtain the following results.

Theorem 5. Let $T_1, T_2 \in R_n$, $|L| = n$, $A \not\subseteq L$ and $|A| = n - 1$. Then

\[
\begin{align*}
    d_{MC}(T_1, T_2) &\leq d_{MC}(T_{1|A}, T_{2|A}) + 2n - 3, \\
    d_{MC}(T_1, T_2) &\geq d_{MC}(T_{1|A}, T_{2|A}) - n + 1.
\end{align*}
\]

Proof. See Appendix.

Theorem 6. The maximal distance in the MC metric can be characterized as

\[
\frac{n^2 - 4 - (n \mod 2)}{2} \leq \max_{T_1, T_2 \in R^n} d_{MC}(T_1, T_2) \leq \max_{T_1, T_2 \in R_n} d_{MC}(T_1, T_2) \leq n^2 - 2n.
\]

Proof. For the lower bound, consider two binary rooted caterpillars $T_1, T_2 \in R^n_{\text{c}}$ created from the same unrooted caterpillar by rooting it in the middle of two most distant edges. In this case there is only one pairing fulfilling Theorem 1 and we obtain $d_{MC}(T_1, T_2) \geq 2 \sum_{i=1}^{\lfloor n/2 \rfloor} (i + 1) + 2 \sum_{i=1}^{\lfloor n/2 \rfloor - 1} i \geq (n^2 - 4 - (n \mod 2))/2$. Since $|\sigma_{e}(T)| \leq n - 2$ for $T \in R_n$, we immediately obtain the upper bound.

The diameter of the space of $n$-leaf rooted trees in the MC distance is greater than in the unrooted trees case for the MS distance, which equals $\frac{3}{4}n^2 \pm O(n)$ (Bogdanowicz and Giaro, 2012). In fact, using the method of clusters pairing in the order appropriate to their non-decreasing sizes we can strengthen the upper bound to the form

\[
\max_{T_1, T_2 \in R_n} d_{MC}(T_1, T_2) \leq \frac{3}{4}n^2 + O(n).
\]

However, we suspect that the diameter is even smaller:

Conjecture. The diameter in the MC-metric space can be expressed as follows: $\max_{T_1, T_2 \in R_n} d_{MC}(T_1, T_2) = \frac{1}{2}n^2 \pm O(n)$.

4.4. MS-component of the MC distance. The discussed metrics (MS and MC) are defined on particular types of phylogenetic trees, i.e., either unrooted or rooted. However, there are cases in phylogenetic analysis where the comparison of an unrooted tree with a rooted one is necessary, e.g., when we compare gene trees with species trees (Górecki and Eulenstein, 2012). In this subsection we present an interesting relation between these two matching metrics.

Theorem 7. There are given trees $T_1, T_2 \in U^{\text{c}}_n$. Let $T_1, T_2 \in R^n_{\text{c}}$ be trees obtained from $T_1$ and $T_2$, respectively, as a result of a rooting operation. Then

\[
d_{MC}(T_1, T_2) \geq d_{MS}(T_1, T_2).
\]

Proof. See Appendix.
in the opposite ends take the distance of the order of \( \frac{1}{4} n^2 \) in MC (see the proof of Theorem 9). On the other hand, for given trees \( T_1, T_2 \in U^B_n \) it is usually possible to find a time flow, i.e., the location of the root such that the component \( d_{MC}(T_1', T_2') - d_{MS}(T_1, T_2) \) is small. Computer simulations using unrooted random binary trees (each tree is equally likely to appear) indicate that the value of \( \Delta_{MC}(T_1, T_2) = \min d_{MC}(T_1', T_2') - d_{MS}(T_1, T_2) \), where the minimum is taken over all possible positions for introducing roots (on the edges of \( T_1 \) and \( T_2 \)), is usually very small (compared to the expected value of the MC distance); see Table 1. For trees with up to 8 leaves, the values are computed based on all possible pairs \( T_1, T_2 \in U^B_n \). In the case of bigger trees, the presented results (in each row) come from the analysis of 10000 pairs of random trees.

Observe that Theorem 7 is no longer valid for arbitrary phylogenetic trees. As a counterexample, consider the trees in Fig. 4. For trees with \( 1 \leq n \leq 6 \) and \( d_{MS}(T_1, T_2) = 2 \) while \( d_{MC}(T_1', T_2') = 1 \).

4.5. Distances of random trees. A reference point is usually needed to interpret the level of dissimilarity of two trees (based on the value of the distance between them in a particular metric). In most cases the average distance between random trees generated according to a particular model can be used for such purposes.

![Fig. 7. Example of trees having \( d_{MS}(T_1, T_2) > d_{MC}(T_1', T_2') \).](image)

**Table 1. Values of the \( \Delta_{MC}(T_1, T_2) \) parameter.**

<table>
<thead>
<tr>
<th># taxa</th>
<th>Avg.</th>
<th>Max.</th>
<th>Avg(\Delta_{MC}(T_1', T_2'))</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3.360</td>
</tr>
<tr>
<td>5</td>
<td>0.032</td>
<td>1</td>
<td>5.971</td>
</tr>
<tr>
<td>6</td>
<td>0.054</td>
<td>1</td>
<td>9.105</td>
</tr>
<tr>
<td>7</td>
<td>0.029</td>
<td>2</td>
<td>12.822</td>
</tr>
<tr>
<td>8</td>
<td>0.058</td>
<td>2</td>
<td>16.922</td>
</tr>
<tr>
<td>10</td>
<td>0.333</td>
<td>4</td>
<td>26.570</td>
</tr>
<tr>
<td>20</td>
<td>0.691</td>
<td>4</td>
<td>97.205</td>
</tr>
<tr>
<td>30</td>
<td>0.965</td>
<td>5</td>
<td>198.785</td>
</tr>
<tr>
<td>40</td>
<td>1.173</td>
<td>5</td>
<td>325.445</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td>474.145</td>
</tr>
</tbody>
</table>

We investigate the asymptotic behavior of the expected distance between two random binary trees in the MC and MS metrics under one of the most popular models of phylogenetic tree generation—the uniform model. In this model all binary phylogenetic trees are equally likely. This process is not an explicit model of evolution, but it is biologically motivated as it arises from a random sample of species from a large group of species generated by a conditioned branching process (see Aldous, 1991; Blum et al., 2006; McKenzie and Steel, 2000). Let \( S(T) \) be defined as \( \sum_{s \in \sigma(T)}|\beta_s| \), if \( T \in R^B_n \) and \( \sum_{s \in \sigma(T)}\min(s) \) for \( T \in U^B_n \). In fact, \( S(T) \) for a rooted tree \( T \) is an equivalent to Sackin’s index \( S_{ind}(T) \) used to measure the tree balance (Sackin, 1972; Shao and Sokal, 1990). Particularly, for a tree \( T \in R^B_n \) we have \( S(T) = S_{ind}(T) + n \). We use a strong result (Theorem 8) given by Blum et al. (2006).

**Theorem 8.** Let \( T_n \) be a tree chosen uniformly at random from \( R^B_n \). Then a cumulative distribution function of the random variable \( S(n)/n^{3/2} \) converges pointwise to a cumulative distribution function of the Airy distribution (A) and \( \lim_{n \to \infty} \mathbb{E}[S(n)]/n^{3/2} = \sqrt{\pi} \).

The following theorem solves the problem regarding the asymptotic behavior of the expected distance in MS stated by Bogdanowicz and Giaro (2012).

**Theorem 9.**

1. **For rooted trees** \( T_1, T_2 \) chosen independently uniformly at random from \( R^B_n \) their expected distance is \( \mathbb{E}[d_{MC}(T_1, T_2)] = \Theta(n^{3/2}) \).

2. **For unrooted trees** \( T'_1, T'_2 \) chosen independently uniformly at random from \( R^B_n \) their expected distance is \( \mathbb{E}[d_{MS}(T'_1, T'_2)] = \Theta(n^{3/2}) \).

**Proof.** Let \( T_1, T_2 \) be chosen independently uniformly at random from \( R^B_n \), and let \( M \) be a pairing of their clusters such that \( \sum_{(A,B) \in M} |A \oplus B| = d_{MC}(T_1, T_2) \).
Then the unrooted trees $T'_1$, $T'_2$, created from $T_1$ and $T_2$ by connecting a new leaf $n + 1$ to their roots, are uniformly drawn from $U_{n+1}^R$ (see Semple and Steel, 2003, Proposition 2.2.3). We define a perfect matching of their splits as $M' = \{(A|1, \ldots, n + 1) \setminus A, B|1, \ldots, n + 1) \setminus B\}$ : $(A, B) \in M \land A = B = \{1, \ldots, n\}$ and another perfect matching $M''$ accomplishing $d_{MS}(T'_1, T'_2)$. Consequently,

$$|S(T'_1) - S(T'_2)| \leq \sum_{(s_1, s_2) \in M''} |\min(s_1) - \min(s_2)| \leq \sum_{(s_1, s_2) \in M''} h_S(s_1, s_2) = d_{MS}(T'_1, T'_2) \leq \sum_{(s_1, s_2) \in M''} h_S(s_1, s_2) \leq d_{MC}(T_1, T_2) \leq \sum_{(A, B) \in M} (|A| + |B|) = S(T_1) + S(T_2).$$

By Theorem\[8\] the proof of the upper bounds is completed in both cases. It remains to prove that $\mathbb{E}[|S(T'_1) - S(T'_2)|] = \Omega(n^{3/2})$. Additionally, we have

$$S(T'_1) \leq S(T_1). \quad (5)$$

Consider an unrooted $n$-leaf binary tree $T \in U_n^R$. We introduce an orientation of the edges of $T$ in the direction of smaller partitions of splits (in this way at most one edge does not receive any orientation). The input degree of each node is 0 or 1. Two situations are possible:

**Case 1.** Exactly one edge $\{u_1, u_2\}$ has been left undirected, so $T$ may be regarded as two binary trees $T_{u_1}$ and $T_{u_2}$, rooted in $u_1$ and $u_2$, respectively, where $|L(T_1)| = |L(T_2)| = n/2$. In this case we say that $T$ is assigned to the split $\{L(T_{u_1}), L(T_{u_2})\}$ of the set $L$.

**Case 2.** All edges have been directed. Then there exists one vertex $u$ of indegree 0, and $T$ may be considered as the sum of three binary trees $T_{u_1}, T_{u_2}, T_{u_3}$ rooted in neighbors of $u$, i.e., $u_1, u_2, u_3$, respectively. We then say that $T$ is assigned to the 3-split $\{L(T_{u_1}), L(T_{u_2}), L(T_{u_3})\}$ treated as an unordered triple.

Now let $T$ be uniformly and randomly chosen from $U_n^R$. By \[5\] we have $\lim_{n \to \infty} \mathbb{E}[S(T)/n^{3/2}] \leq \sqrt{\pi}$. Thus for sufficiently large values of $n$ (such that $\mathbb{E}[S(T)/n^{3/2}] < 2$), by Markov’s inequality we obtain

$$\Pr(S(T) \leq 4n^{3/2}) \geq 1 - \Pr(S(T) \geq 2\mathbb{E}[S(T)]) \geq \frac{1}{2}. \quad (6)$$

Let $p > 0$ be less than the probability that the random variable of the Airy distribution $\lambda$ has the value greater than $5 \cdot 3^{1/2}$. The randomly chosen tree $T$ is assigned to exactly one 2- or 3-split of $L$ with the largest partition $A \subseteq L (|A| \geq n/3)$. Let $T_A \in R_n^B$ be the rooted subtree of $T$ corresponding to $A$. For sufficiently large $n$, the rooted tree $T_A$ (where $|A| \geq n/3$) fulfills $S(T_A)/|A|^{1/2} \geq 5 \cdot 3^{1/2}$ with a probability greater than $p$. Therefore, $S(T_A) \geq 5n^{3/2}$ and, finally,

$$\Pr(S(T) \geq 5n^{3/2}) \geq p. \quad (7)$$

Combining \[5\] with \[7\] we obtain that, for two independently randomly chosen trees $T'_1$ and $T'_2$ with probability at least $p$, one of them fulfills the condition \[6\] and the other the condition \[7\], thus $\mathbb{E}[|S(T'_1) - S(T'_2)|] = \Omega(n^{3/2})$.

Not only the maximum value, but also the expected value between two random binary trees in the MC and MS metrics grows faster than the diameter of the RF distance, hence more subtle MC-dissimilarity evaluation results in a greater range. Moreover, the expected value is asymptotically smaller than the diameters of MC and MS, while in the case of RF both these parameters grow equally fast, i.e., as $\Theta(n)$ (Steel and Penny, 1993).

Another very popular model of phylogenetic tree generation is the Yule model, where trees are constructed iteratively: starting from three random taxa, new taxa (chosen randomly) are added to a branch connected to a leaf (chosen uniformly randomly as well) (McKenzie and Steel, 2000). We observed that the distributions of the RF distance between random trees (in both models: uniform and Yule) are highly asymmetrical compared to the MC metric (see Figs. \[8\] and \[10\]).

Detailed statistical results (e.g., average distances, standard deviation, quantiles) concerning the MC distance computed on the basis of the analysis of 10000 pairs of random trees (in both models) having between 10 and 1000 leaves are available at \text{http://www.kaiims.pl/\sim dambro/mcdlist}. The MS and MC metrics, as well as many other distances, are implemented in the freely available TreeCmp application (Bogdanowicz et al., 2012).

The most important properties regarding the MC distance for binary trees discussed in this section are summarized in Table \[2\].

## 5. MC distance in supertree construction

Supertree methods allow constructing tress that combine phylogenetic information represented by a set of smaller trees with partially overlapped taxa. Such analyses play an important role in phylogenetic research, e.g., they allowed constructions of the first family-level phylogeny of flowering plants (Davies et al., 2004) and the first species-level phylogeny of nearly all extant mammal species (Bininda-Emonds et al., 2007).

Let profile $P$ be a tuple of rooted trees $(T_1, \ldots, T_k)$. For a given profile $P$, we define a supertree $T^P \in R_n^B$, on
Fig. 9. Histograms of distances in the RF and MC metrics based on 10000 randomly generated pairs of binary trees with 50 leaves according to the uniform model.

Fig. 10. Histograms of distances in the RF and MC metrics based on 10000 randomly generated pairs of binary trees with 50 leaves according to the Yule model.

Table 2. Comparison of selected properties of analyzed metrics for binary trees.

<table>
<thead>
<tr>
<th>Property</th>
<th>RF (unrooted)</th>
<th>RF (rooted)</th>
<th>MS</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal positive distance</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Number of trees at the minimal positive distance from a given tree</td>
<td>$2n - 6$</td>
<td>$2n - 4$</td>
<td>$0 \leq x \leq n - 1$</td>
<td>$0 \leq x \leq n - 1$</td>
</tr>
<tr>
<td>Maximal distance</td>
<td>$n - 3$</td>
<td>$n - 2$</td>
<td>$\frac{3}{2}n^2 + O(n)$</td>
<td>$\frac{n^2 - 4 - (n \mod 2)}{2} \leq x \leq \frac{3}{2}n^2 + O(n)$</td>
</tr>
<tr>
<td>Distance of caterpillar trees constructed as in Fig 4</td>
<td>$n - 3$</td>
<td>$n - 2$</td>
<td>$\Theta(n)$</td>
<td>$\Theta(n^{3/2})$</td>
</tr>
<tr>
<td>Average distance of random trees (in the uniform model)</td>
<td>$\Theta(n)$</td>
<td>$\Theta(n)$</td>
<td>$\Theta(n^{3/2})$</td>
<td>$\Theta(n^{3/2})$</td>
</tr>
</tbody>
</table>

$P$ to be a binary rooted tree such that $L^* = \bigcup_{i=1}^{k} L(T_i)$. There are various methods and approaches to the problem of finding the most suitable supertrees for a given profile (for a comparison and a review, see, e.g., Brinkmeyer et al., 2011; Bansal et al., 2010; Swenson et al., 2011; Nguyen et al., 2012).

Here we are interested in methods based on distances between trees that used clusters during the calculation of the dissimilarity value. The classical approach that belongs to the described group is the RF-supertree method (Bansal et al., 2010). In a similar manner, we can introduce a new procedure based on the MC distance (MC-supertree method). Let us define the distance of a tree $T^* \in R^{B}_L$ to an arbitrary profile $P = (T_1, \ldots, T_k)$, $L^* = \bigcup_{i=1}^{k} L(T_i)$ as follows:

$$d^*(T^*, P) = \sum_{i=1}^{k} d(T^*_{|L(T_i)}, T_i),$$

where $d$ can be an arbitrary metric defined for rooted phylogenetic trees. In the remaining part of this section, as $d$ we consider the two metrics: RF and MC.

Now we present some preliminary experimental results based on the biological data concerning the properties of the MC-supertree method. We used a data set of seabirds (121 taxa, the profile of 7 source trees; see the work of Kennedy et al. (2002)) and evaluated the topological accuracy of a simple supertree search...
heuristic based on the two above mentioned metrics. For each of the metrics, 7 tests were performed, during which 7 different binary trees on 121 taxa were chosen as starting supertrees. Each of these trees was constructed based on a particular source tree by adding missing taxa and solving multifurcations (if there was such a need) randomly, i.e., the same starting supertree was used for both RF and MC based tests having the same test number. The properties of chosen starting supertrees are presented in Table 3. Searching for a supertree was performed according to the local search hill climbing technique as follows. In each iteration all trees in the rooted Subtree-Prune-and-Regraft (rSPR) neighborhood (Bordewich and Semple, 2005) of the currently best tree were analyzed. A tree with the lowest $d^*$ value to source profile was chosen as the best tree in the next iteration. Ties were resolved arbitrarily. The procedure ended if no better supertrees could be found.

The properties of the best found supertrees are presented in Table 4. The parsimony scores presented in Table 3 and 4 were calculated using the PAUP application (Swofford, 2002) and the spurce Python package (Suri and Warnow, 2010). The best RF-supertree is at the RF distance of 44.5 and the MC distance of 194. The best MC-supertree is at the RF distance of 50.5 and at the MC distance of 154. Both these trees have similar parsimony scores equal to 228 in the case of the RF-supertree and 230 for the MC-supertree. We can observe that the average distance (MC and also RF) of MC-supertrees to source trees is lower than the average distance of RF-supertrees. Moreover, MC-supertrees have a better, i.e., lower), average parsimony score than RF-supertrees. It seems that, on the average, searching using the MC metric can give better results than that employing the RF distance.

Table 3. Distances and parsimony scores of starting supertrees used in the experiment. The column with the number of taxa contains the size of the source tree that was used to build the particular starting supertree.

<table>
<thead>
<tr>
<th>Test</th>
<th># taxa</th>
<th>RF</th>
<th>MC</th>
<th>Pars. Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>184.5</td>
<td>1309</td>
<td>852</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>183.5</td>
<td>1267</td>
<td>840</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>170.5</td>
<td>1170</td>
<td>822</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>167.5</td>
<td>1125</td>
<td>788</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>93.5</td>
<td>367</td>
<td>337</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>179.5</td>
<td>1286</td>
<td>847</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>171.5</td>
<td>1095</td>
<td>758</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td>164.36</td>
<td>1088.43</td>
<td>749.14</td>
</tr>
</tbody>
</table>

6. Conclusions

The RF distance has been the most popular phylogenetic metric for years. Despite its incontrovertible advantages (simplicity, effective computation) it has some shortcomings, e.g., similarities between the parts of compared trees are quantified binarily (i.e, clusters are considered equal or not without any similarity evaluation). Moreover, its distribution is highly concentrated around the diameter. In this work and earlier (Bogdanowicz and Giaro, 2012) we proposed a method for defining phylogenetic distances based on any metric on clusters or splits that generalizes the RF distance. The simplest metrics (MS, MC) defined using that approach were analyzed. The described metrics have many interesting properties, e.g., small sensitivity to modifications being the result of a bounded number of leaf relocation or edge contraction, the average distance that grows slower than the diameter.

Despite worse time complexity (compared to RF), the MC metric can be still used in practical applications, since the computation of the MC distance between two random trees with 1000 leaves takes only about 2 s on a desktop computer with an Intel Core2 1.66 GHz processor, even using not very time-efficient Java implementation.

In Section 5 we presented a proof-of-concept study of the usefulness of the MC distance for constructing supertrees. Based on the experiments it seems that the MC-supertree method can be a promising approach and might be an interesting complement for the set of known supertree methods.

Finally, observe that metrics constructed according to Definition 2 allow us to compare arbitrary subsets of $2^L$ or $\text{Splits}(L)$ (even not compatible, i.e., not forming a tree) based on any metric $h$ on these sets analogously to $h_S$ in (3) and $h_C$ in (4). This fact provokes suggestions on other potential areas of matching metrics applications in phylogenetics, different than simple tree comparison.

Biological events such as coalescences of separate species lines or gene transfer result in the increasing popularity of describing the evolution by more general and flexible structures than trees—phylogenetic networks. These are directed acyclic graphs on the set of leaves $L$ corresponding to present-day species and various restrictions imposed on the network structure. Recently, methods of comparing phylogenetic networks through introducing a structure of a metric space in a family of networks (on a given leaf set $L$) have been proposed in the literature. Some of them are based on generalizations of known metrics for trees (such as RF or nodal distances (Cardona et al., 2009a; 2009b). We believe that it is a promising idea to define matching metrics for phylogenetic networks in all models, where the network is unambiguously described by the family of clusters related to its internal nodes, e.g., for tree-child time consistent networks (Cardona et al., 2009b).

Some phylogenetic reconstruction methods (e.g., Bayesian approach) generate large sets of high reliability
trees. This type of data can be post-processed by clustering algorithms in order to identify subsets of similar trees, describing alternative solutions. However, an application of popular algorithms for clustering point sets in $\mathbb{R}^n$ (e.g., $k$-means algorithm) is not obvious, because there is no natural method of finding a “mass center” or an “average tree” of a set of incompatible trees forming a cluster. Stockham et al. (2002) proposed an adaptation of the $k$-means algorithm, where trees are compared using the RF distance and as the center of a family of trees acts an appropriately chosen subset of splits (not necessarily compatible) of the analyzed trees. This approach can be in a natural manner applied to the metrics constructed according to Lemma 2. The verification of the effectiveness of this approach requires further experimental research.

Acknowledgment

The authors would like to thank the anonymous reviewers for their interesting and constructive comments on the manuscript. This work was partially supported by the Polish National Science Centre (decision number DEC-2011/02/A/ST6/00201).

References


metrics, IEEE/ACM Transactions on Computational Biology and Bioinformatics 6(3): 454–469.


A2. Proof of Theorem 5

Proof. Let \( L = A \cup \{x\} \). Consider the changes in the set of non-trivial clusters after adding a leaf \( x \) to \( T_{1 '\cup A} \). A cluster \( s \in \sigma_*(T_{1 '\cup A}) \) transforms to \( s' \in \sigma_*(T_1) \), where \( s' = s \) or \( s' = s \cup \{x\} \) if \( s \notin \sigma(T_1) \). We call \( s' \) and \( s \) corresponding clusters. Additionally, we consider the special element \( O = \emptyset \) as corresponding to itself. If \( x \) is attached to the middle of an edge of \( T_{1 '\cup A} \), then one new cluster \( s_{new} \in \sigma_*(T_1) \), which is not a corresponding cluster, appears. The case of transformation of \( T_{2 \cup A} \) into \( T_2 \) is analogical. Thus, if \( s, t \subseteq A \) and \( s', t' \subseteq L \) are clusters corresponding to \( s, t \) in \( T_1 \) and \( T_2 \), respectively, then \( h_{RC}(s, t) \leq h_{RC}(s', t') \leq h_{RC}(s, t) + 1 \). Consider a pairing \( \{(s_i, t_i) : i = 1, \ldots, k\} \), where \( s_i \in \sigma_*(T_{1 \cup A}) \cup \{O\}, t_i \in \sigma_*(T_{2 \cup A}) \cup \{O\}, k = \max_{i=1,2}|\sigma_*(T_{1 \cup A})| \leq n - 3 \) and \( \sum_{i=1}^k h_{RC}(s_i, t_i) = d_{MC}(T_{1 \cup A}, T_{2 \cup A}). \) We create a pairing that consists of pairs of corresponding elements \( s_i' \in \sigma_*(T_1) \cup \{O\} \) and \( t_i' \in \sigma_*(T_2) \cup \{O\} \) (and if necessary) a pair \((s_{new}, t_{new})\) if both sets \( \sigma_*(T_1) \) and \( \sigma_*(T_2) \) contain new clusters or \((s_{new}, O)\) alternatively \((O, t_{new})\) if exactly one set of \( \sigma_*(T_1) \), \( \sigma_*(T_2) \) contains a new cluster. Thus the first inequality follows.

Now consider a pairing \( M \) of elements of \( \sigma_*(T_1) \cup \{O\} \) with elements of \( \sigma_*(T_2) \cup \{O\} \) analogous to a minimum-weight perfect matching defining the MC distance between \( T_1 \) and \( T_2 \). We transform \( M \) into a pairing \( M' \) between the sets \( \sigma_*(T_{1 \cup A}) \cup \{O\} \) and \( \sigma_*(T_{2 \cup A}) \cup \{O\} \) by changing all elements in these pairs to their corresponding elements where it is possible (i.e., in pairs without new clusters) and remove the other pairs (i.e., containing new clusters). The following four cases remain to be considered:
Case 1. There are new clusters $s_{\text{new}} \in \sigma_*(T_1)$, $t_{\text{new}} \in \sigma_*(T_2)$ and $(s_{\text{new}}, t_{\text{new}}) \in M$. Then
\[
d_{MC}(T_1, T_2) = \sum_{(s,t) \in M} h_C(s, t) \geq \sum_{(s,t) \in M'} h_C(s, t) - h_C(x_1', x_2') \geq d_{MC}(T_1|A, T_2|A) - n + 1.
\]

Case 2. There is no new cluster in $\sigma_*(T_1)$ nor in $\sigma_*(T_2)$. Then $d_{MC}(T_1, T_2) \geq d_{MC}(T_1|A, T_2|A)$ as in the above case.

Case 3. There are unpaired new clusters $s_{\text{new}} \in \sigma_*(T_1)$, $t_{\text{new}} \in \sigma_*(T_2)$ in $M$, i.e., $(s_{\text{new}}, x_2), (x_1, t_{\text{new}}) \in M$. Then we extend $M'$ by a pair $(x_1', x_2')$, where $x_1'$ is an element that corresponds to $x_1$; hence
\[
d_{MC}(T_1, T_2) = \sum_{(s,t) \in M} h_C(s, t) \geq d_{MC}(T_1|A, T_2|A) - n + 1.
\]

Case 4. There is only one new cluster; assume that it is $s_{\text{new}} \in \sigma_*(T_1)$ and $(s_{\text{new}}, x) \in M$. Then we extend $M'$ by a pair $(O, x')$, where $x'$ is an element that corresponds to $x$; hence
\[
d_{MC}(T_1, T_2) = \sum_{(s,t) \in M} h_C(s, t) \geq d_{MC}(T_1|A, T_2|A) - n + 1.
\]

\[\square\]

A3. Proof of Theorem 7

Proof. Consider $T_1, T_2 \in U^B$ and their rootings in the middle of the edges $e_1, e_2$ with corresponding splits $A\{B\}$ and $C\{D\}$, respectively. Notice that for each $e \in E(T_1) \setminus \{e_1\}$ there exists exactly one corresponding cluster in $\sigma_*(T_1) \setminus \{A, B, L\}$ such that the equivalent of a cluster $X$ is a split $X \setminus L \setminus X$. An analogous relation may be introduced between clusters $Y \in \sigma_*(T_2) \setminus \{C, D, L\}$ and the splits $Y \setminus L \setminus \beta_{T_2}(Y)$ of $\{C, D\}$. Moreover, $|X \oplus Y| \geq h_S(X \setminus L \setminus X, Y \setminus L \setminus Y)$.

Consider a pairing $M'$ between clusters of $\sigma_*(T_1) \setminus \{L\}$ and $\sigma_*(T_2) \setminus \{L\}$ that realizes the minimum-weight perfect matching $\sum_{(c,c') \in M'} h_C(c, c') = d_{MC}(T_1, T_2)$ and fulfills the conditions from Theorem 1. Our aim is to construct a perfect matching $M$ of the splits of $\beta(T_1)$ with the splits of $\beta(T_2)$ having the weight $\sum_{(s,t) \in M} h_S(s, t) \leq d_{MC}(T_1', T_2')$. First, for each pair $(X_1, X_2) \in M'$ such that $X_1 \notin \{A, B\}$ and $X_2 \notin \{C, D\}$ we add to $M$ a pair consisting of the corresponding splits $(X_1 \setminus L \setminus X_2) \setminus L \setminus X_2$. Depending on the rest of $M'$, i.e., pairs with clusters $A, B, C, D$ in $M'$, only the following three situations are possible:

Case 1. Each of the clusters $A, B$ is matched to some cluster of $\{C, D\}$. We add a pair $(A|B, C|D)$ to $M$, so the proof is complete.

Case 2. Exactly one cluster of $\{A, B\}$ is matched to some of $\{C, D\}$, e.g., $(A, C) \in M'$. Therefore we have $(B, Y), (X, D) \in M'$, where $X \in \sigma(T_1), Y \in \sigma(T_2)$. In this situation we only need to add to $M$ two pairs: $(A|B, Y\setminus L \setminus Y)$ and $(X \setminus L \setminus X, C|D)$.

Case 3. None of $A, B$ is matched to any of $\{C, D\}$. Since $M'$ conserves the ancestor-descendant relations, only the following two pairings are possible: $m_1 = \{(A, c), (a, d), (B, d), (b, C)\} \subseteq M'$ or symmetric variant $\{(A, d), (a, C), (b, c), (B, d)\} \subseteq M'$, where $a \not\subset A$, $b \not\subset B$, $c \not\subset C$ and $d \not\subset D$. We consider the first case, the second one is analogous. We now construct two pairings of splits:

\[
m_1 = \{(a\setminus L \setminus a, c|L \setminus c), (b\setminus L \setminus b, c|D), (A|B, d|L \setminus d)\},
\]

\[
m_2 = \{(a\setminus L \setminus a, c|D), (b\setminus L \setminus b, d|L \setminus d), (A|B, c|L \setminus c)\},
\]

and two pairings of clusters:

\[
m_1' = \{(a, c), (b, C), (B, d)\},
\]

\[
m_2' = \{(a, d), (b, d), (A, c)\}.
\]

Note that, for $i = 1, 2$,

\[
\sum_{(s,t) \in M_i} h_S(s, t) \leq \sum_{(c,c') \in M_i'} h_C(c, c').
\]

Based on calculations of the total costs of pairings $m_1, m_1', m_2'$ presented in Table A1 (see also the Venn
Table A1. Computation of the total cost of the pairings $m', m'_1, m'_2$. The numbers in rows represent a contribution of the cardinality of a particular set into the total cost, e.g., 2 in row $r_{00}$, column $m'$ means that the summand $2|_{r_{00}}|$ appears in the cost of $m'$.

The columns $\Delta m'_1$ and $\Delta m'_2$ contain the difference between the values $m'_1$ or $m'_2$ and $m'$, respectively.

<table>
<thead>
<tr>
<th>Set</th>
<th>Definition</th>
<th>$m'$</th>
<th>$m'_1$</th>
<th>$m'_2$</th>
<th>$\Delta m'_1$</th>
<th>$\Delta m'_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{00}$</td>
<td>$a \cap c$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$r_{01}$</td>
<td>$(a \cap c) \setminus (a \cap c)$</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$r_{10}$</td>
<td>$(A \cap c) \setminus (a \cap c)$</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$r_{11}$</td>
<td>$(A \cap c) \setminus (a \cup c)$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$s_{00}$</td>
<td>$a \cap d$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$s_{01}$</td>
<td>$(a \cap d) \setminus (a \cap d)$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$s_{10}$</td>
<td>$(A \cap d) \setminus (a \cap d)$</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>$s_{11}$</td>
<td>$(A \cap d) \setminus (a \cup d)$</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>$t_{00}$</td>
<td>$b \cap c$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$t_{01}$</td>
<td>$(b \cap c) \setminus (b \cap c)$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$t_{10}$</td>
<td>$(B \cap c) \setminus (b \cap c)$</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>$t_{11}$</td>
<td>$(B \cap c) \setminus (b \cup c)$</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>$u_{00}$</td>
<td>$b \cap d$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_{01}$</td>
<td>$(b \cap d) \setminus (b \cap d)$</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_{10}$</td>
<td>$(B \cap d) \setminus (b \cap d)$</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>$u_{11}$</td>
<td>$(B \cap d) \setminus (b \cup d)$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

Diagram in Fig. A1, we obtain that at least one of the two inequalities, $\sum_{(c,c') \in m'_i} |c \oplus c'| \leq \sum_{(c,c') \in m} |c \oplus c'|$, $i = 1, 2$, is always valid, because the only positive value (row $r_{10}$) in column $\Delta m'_1$ is negative in column $\Delta m'_2$ and vice versa (row $u_{10}$). Therefore, we extend $M$ by pairs of splits of the least expensive of the two pairings $m_1$ or $m_2$. 

Received: 27 January 2012
Revised: 3 April 2013