VERSITA

Int. J. Appl. Math. Comput. Sci., 2013, Vol. 23, No. 3, 669–684
DOI: 10.2478/amcs-2013-0050

amcs

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 (Critchlow *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 grater 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 $w : E \to \mathbb{Z}_{\geq 0}$ 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 $d : X \times X \to \mathbb{R}_{\geq 0}$ (the *metric over* X) such that (i) $\forall_{x,y \in X} d(x, y) = 0 \Leftrightarrow x = y$, (ii) $\forall_{x,y \in X} d(x, y) = d(y, x)$, and (iii) $\forall_{x,y,z \in X} d(x, y) + d(y, z) \geq d(x, z)$—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_L^B 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 $\leq_T$. For $a, b \in V(T)$, we have $a \leq_T b$ (that is, a is a descendant of b) if the path in T from a to $r(T)$ contains b. In particular, $v \leq_T r(T)$ and $v \leq_T v$ 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 $\sigma_*(T)$ 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 $|\sigma(T)| \leq 2|L| - 1$, $|\sigma_*(T)| \leq |L| - 2$, and both inequalities are tight for binary trees. A rooted phylogenetic tree T is uniquely described by a set $\sigma_*(T)$ 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 [1]

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

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 $L = \{1, \ldots, n\}$, we 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 $\beta_*(T)$. 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 1, 2 and Definition 2 here correspond to Lemmas 3.3, 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 $l : V_1 \cup V_2 \to X$. 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 $1 \leq i \leq k \leq n$. 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, an element $O \notin D$ 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 $V_1 = \{a_1, \ldots, a_{|A|}, a_{|A|+1}, \ldots, a_{|A|+t}\}$, $V_2 = \{b_1, \ldots, b_{|B|}, b_{|B|+1}, \ldots, b_{|B|+s}\}$ 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 $x \in D$ are unnecessary.

The presented method gives a convenient way to define metrics for phylogenetic trees (i.e., introducing a metric space in R_L and U_L) based on any metric h on subsets of L and $Splits(L) \cup \{O\}$ 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 $Splits(L) \cup \{O\}$.

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

$$h_S(A_1|B_1, A_2|B_2) = \min\{|A_1 \oplus A_2|, |A_1 \oplus B_2|\}, \quad (2)$$

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

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

$$d_{MS}(T_1, T_2) = d_{h_S}(\beta(T_1), \beta(T_2)) = d_{h_S}(\beta_*(T_1), \beta_*(T_2)). \quad (3)$$

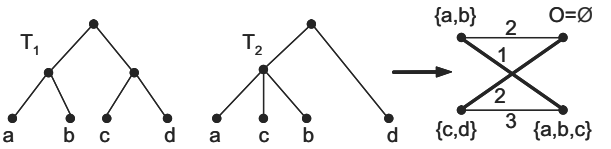

Fig. 1. Calculation of the MC distance between trees T_1 and T_2. The bipartite graph of their non-trivial clusters has a perfect matching of minimum weight equal to 3.

Thus customization of the function for comparing splits h leads us to more utile 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 analogy 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 $\Theta(n^2)$

(Theorem 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 *et al*. (2003) and Nye *et al*. (2006); such an analysis was also adopted to phylogenetic networks (Cardona *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.

4.1. Conservation of ancestor–descendant relations. 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 $M_1 = \{\{s_{i+1}, t_i\}\}_{i=1,\ldots,n-3} \cup \{\{s_1, t_{n-2}\}\}$ and $M_2 = \{\{s_i, t_i\}\}_{i=1,\ldots,n-2}$ with the same minimal cost $2n-4$. 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 $\dot{V}_i = V(T_i) \setminus (L \cup \{r(T_i)\})$ for $i = 1, 2$. There exists an injection $f : \dot{V}_1 \to \dot{V}_2$ with the cost $\sum_{v \in \dot{V}_1} |c(v) \oplus c(f(v))| + \sum_{u \in \dot{V}_2 \setminus f(\dot{V}_1)} |c(u)| = d_{MC}(T_1, T_2)$ such that all internal non-root vertices $a, b \in V(T_1)$ fulfill the following conditions:*

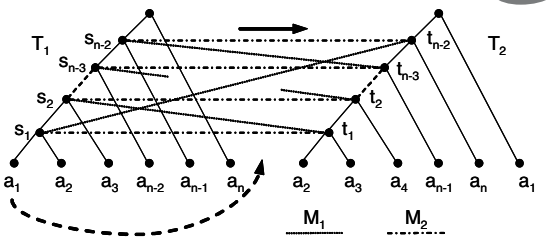

Fig. 2. Two rooted caterpillars such that the tree T_2 was obtained from T_1 by placing the leaf a_1 behind a_n and connecting it to the root. Mappings M_1, M_2 of their internal vertices have equal costs.

1. *If $a \leq_{T_1} b$, then their related vertices in T_2 fulfill $f(a) \leq_{T_2} f(b)$ or they are $\leq_{T_2}$-incomparable.*
2. *If $f(a) \leq_{T_2} f(b)$, then $a \leq_{T_1} b$ or they are $\leq_{T_1}$-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 $f(b) \leq_{T_2} f(a)$.

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 $A = c(a) \subsetneq c(b) = B = X \oplus A$ and $B' = c(b') \subsetneq c(a') = A' = B' \oplus X'$. We have $|A \oplus A'| + |B \oplus B'| = |A \oplus B'| + |B \oplus A'| + 2|X \cap X'|$; hence, after making modifications to f so that $f(a) := b'$, $f(b) := a'$, the equality $\sum_v |c(v) \oplus c(f(v))| + \sum_u |c(u)| = d_{MC}(T_1, T_2)$ still holds. But $|A||B'| + |B||A'| = |A||A'| + |B||B'| + |X||X'|$, so after the modification the value of the parameter $\sum_v |c(v)||c(f(v))|$ 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 $T_1, T_2 \in R_L^B$, 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 $T_1 \in R_L^B$, 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. 4).*
4. *There exist $T_1, T_2 \in R_n^B$, such that $d_{RF}(T_1, T_2) = 1$ and $d_{MC}(T_1, T_2) = n - 1$ (see Fig. 5).*
5. *If $T_1, T_2 \in R_n^B$ 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.*
6. *For a tree $T_1 \in R_n^B$, the number of trees $T_2 \in R_n^B$ such that $d_{MC}(T_1, T_2) = 2$ may vary between 0 and $n - 1$ (rooted caterpillar case).*

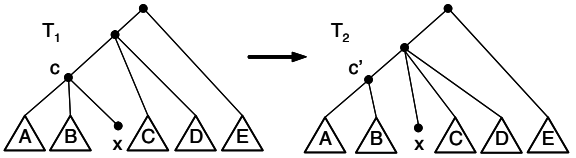

Fig. 3. MC distance between the trees T_1 and T_2 equals 1.

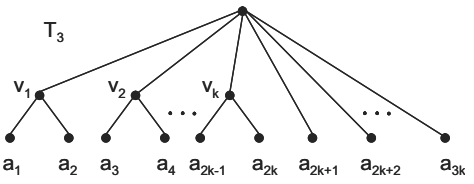

Fig. 4. Tree T with $3k$ leaves and k^2 trees at the MC distance 1. Removing leaves $a_{2k+1}, \ldots, a_{3k}$ from the root and reattaching them to the appropriate internal vertices results in a tree T' with $d_{RF}(T, T') = d_{MS}(T, T') = k$.

In summary, observe that the MC-metric space of binary trees seems to be less "regular" than in the RF case, where the number of the closest possible (i.e., distanced by 1) points is always $2n - 4$. However, there are no "isolated regions" in the MC-metric space, since analysis

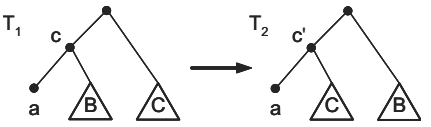

Fig. 5. MC distance between the trees T_1 and T_2 equals $|L| - 1$.

analogous to Theorem 5.2 of Bogdanowicz and Giaro (2012) gives the following result.

Theorem 3. *There are given two trees $T_a, T_b \in R_L$.*

1. *There exists a sequence of trees $T_a = T_1, T_2, \ldots, T_{k-1}, T_k = T_b$, $T_i \in R_L$ for $i = 1, \ldots, k$ such that $d_{MC}(T_j, T_{j+1}) \leq 4$, where $j = 1, \ldots, k - 1$.*
2. *If $T_a, T_b \in R_L^B$, 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 $T_a, T_b \in R_L^B$, 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 only one leaf may create an unrooted tree distanced 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 distanced 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).

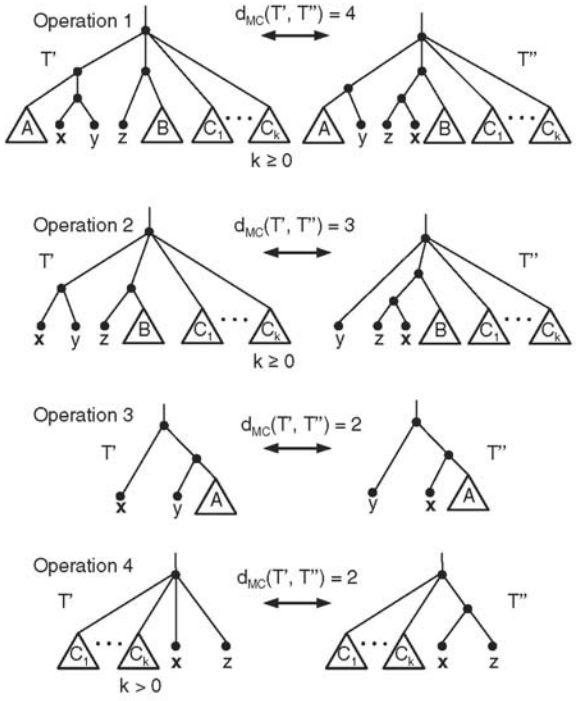

Fig. 6. Local modifications of subtrees which can connect every pair of trees from R_L.

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 $c \in \sigma_*(T) \setminus \sigma_*(T_e)$. By Lemma 1 we have $d_{MC}(T, T_e) = h_C(c, O) = |c| \leq n - 1$. ■

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 $T_1, T_2 \in R_L$, $|L| = n$, $A \subsetneq L$ and $|A| = n - 1$. Then*

$$d_{MC}(T_1, T_2) \leq d_{MC}(T_{1|A}, T_{2|A}) + 2n - 3,$$
$$d_{MC}(T_1, T_2) \geq d_{MC}(T_{1|A}, T_{2|A}) - n + 1.$$

Proof. See Appendix. ■

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

$$\begin{aligned} &\frac{n^2 - 4 - (n \mod 2)}{2} \\ &\leq \max_{T_1, T_2 \in R_n^B} d_{MC}(T_1, T_2) \\ &\leq \max_{T_1, T_2 \in R_n} d_{MC}(T_1, T_2) \\ &\leq n^2 - 2n. \end{aligned}$$

Proof. For the lower bound, consider two binary rooted caterpillars $T_1, T_2 \in R_n^B$ 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 $d_{MC}(T_1, T_2) \geq 2\sum_{i=1}^{\lfloor n/2 \rfloor - 1}(i+1) + 2\sum_{i=1}^{\lceil n/2 \rceil - 1} i \geq (n^2 - 4 - (n \mod 2))/2$. Since $|\sigma_*(T)| \leq n - 2$ 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 $\frac{3}{8}n^2 \pm O(n)$ (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 $T_1, T_2 \in U_L^B$. Let $T_1', T_2' \in R_L^B$ 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. ■

The above inequality provokes an interesting interpretation of the distance $d_{MC}(T_1', T_2')$ for binary rooted trees. The value consists of a component $d_{MS}(T_1, T_2)$ that quantifies the difference between the topologies, i.e., the unrooted equivalents, T_1, T_2, and a component $d_{MC}(T_1', T_2') - d_{MS}(T_1, T_2)$ quantifying the additional difference related to the direction of a time flow along edges introduced into the trees T_1', T_2' during a rooting operation. The second component can have big values, even close to the diameter, e.g., the same unrooted binary caterpillars (at a distance 0 in MS) rooted

in the opposite ends take the distance of the order of $\sim \frac{1}{2}n^2$ in MC (see the proof of Theorem 6). On the other hand, for given trees $T_1, T_2 \in U_L^B$ 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 $T_1, T_2 \in U_n^B$. 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. 7, where $d_{MS}(T_1, T_2) = 2$ while $d_{MC}(T_1', T_2') = 1$.

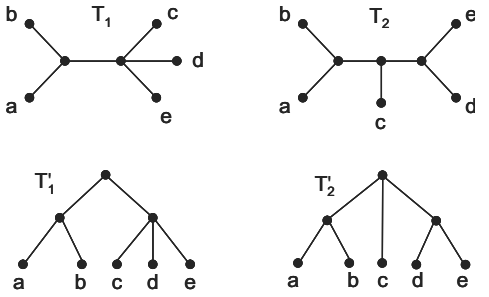

Fig. 7. Example of trees having $d_{MS}(T_1, T_2) > d_{MC}(T_1', T_2')$.

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.

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

# taxa	Avg.	Max.	$\text{Avg}_{T_1,T_2} \min d_{MC}(T_1', T_2')$
4	0	0	3.360
5	0	0	5.971
6	0.032	1	9.105
7	0.054	1	12.822
8	0.029	2	16.922
10	0.058	2	26.570
20	0.333	4	97.205
30	0.691	4	198.785
40	0.965	5	325.445
50	1.173	5	474.145

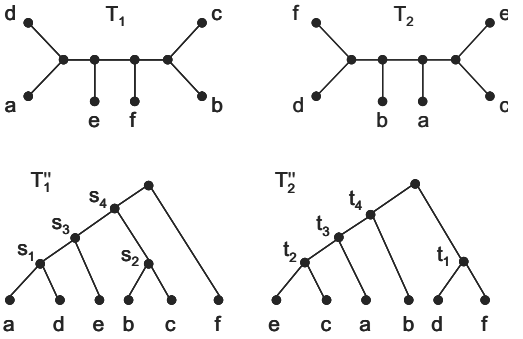

Fig. 8. Trees having $d_{MS}(T_1, T_2) = 6$ for which $\min d_{MC}(T_1', T_2')$ taken over all the possible rooted versions of T_1 and T_2 equals $d_{MC}(T_1'', T_2'') = 7$.

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 $\sum_{c \in \sigma(T)} |c|$, if $T \in R_L^B$ and $\sum_{s \in \beta(T)} \min(s)$ for $T \in U_L^B$. 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 $T \in R_n^B$ 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_n^B. 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 ($\mathcal{A}$) 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_n}, T_{2_n} chosen independently uniformly at random from R_n^B their expected distance is $\mathbb{E}[d_{MC}(T_{1_n}, T_{2_n})] = \Theta(n^{3/2})$.*
2. *For unrooted trees T'_{1_n}, T'_{2_n} chosen independently uniformly at random from U_n^B their expected distance is $\mathbb{E}[d_{MS}(T'_{1_n}, T'_{2_n})] = \Theta(n^{3/2})$.*

Proof. Let T_1, T_2 be chosen independently uniformly at random from R_n^B, 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^B_{n+1} (see Semple and Steel, 2003, Proposition 2.2.3). We define a perfect matching of their splits as $M' = \{(A|\{1,\dots,n+1\} \setminus A, B|\{1,\dots,n+1\} \setminus B) : (A,B) \in M \vee A = B = \{1,\dots,n\}\}$ and another perfect matching M'' accomplishing $d_{MS}(T_1', T_2')$. Consequently,

$$\begin{aligned}
&|S(T_1') - S(T_2')| \\
&\quad \le \sum_{(s_1,s_2)\in M''} |\min(s_1) - \min(s_2)| \\
&\quad \le \sum_{(s_1,s_2)\in M''} h_S(s_1,s_2) = d_{MS}(T_1', T_2') \\
&\quad \le \sum_{(s_1,s_2)\in M'} h_S(s_1,s_2) \le d_{MC}(T_1,T_2) \\
&\quad \le \sum_{(A,B)\in M} (|A| + |B|) = S(T_1) + S(T_2).
\end{aligned}$$

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

$$S(T_1') \le S(T_1). \tag{5}$$

Consider an unrooted n-leaf binary tree $T \in U^B_L$. 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 the sum of three binary trees T_{u_1}, T_{u_3}, 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^B_n. By (5) we have $\lim_{n\to\infty} \mathbb{E}[S(T)/n^{3/2}] \le \sqrt{\pi}$. Thus for sufficiently large values of n (such that $\mathbb{E}[S(T)/n^{3/2}] < 2$), by Markov's inequality we obtain

$$\begin{aligned}
\Pr(S(T) \le 4n^{3/2}) &\ge 1 - \Pr(S(T) \\
&\ge 2\mathbb{E}[S(T)]) \ge \frac{1}{2}.
\end{aligned} \tag{6}$$

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

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

Combining (6) with (7) we obtain that, for two independently randomly drawn 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. 9 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 `http://www.kaims.pl/~dambo/mcdist`. 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 $(T_1,\dots,T_k)$. For a given profile P, we define a *supertree* $T^* \in R^B_{L^*}$ on

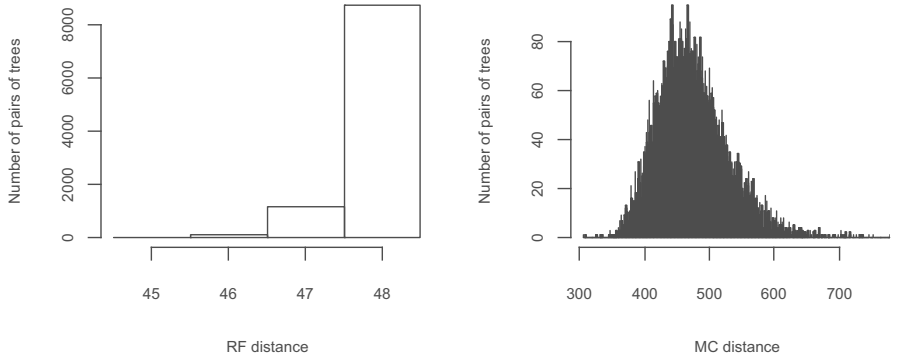

Fig. 9. Histograms of distances in the RF and MC metrics based on 10000 randomly generated pairs of binary trees with 50 leaves according to the uniform model.

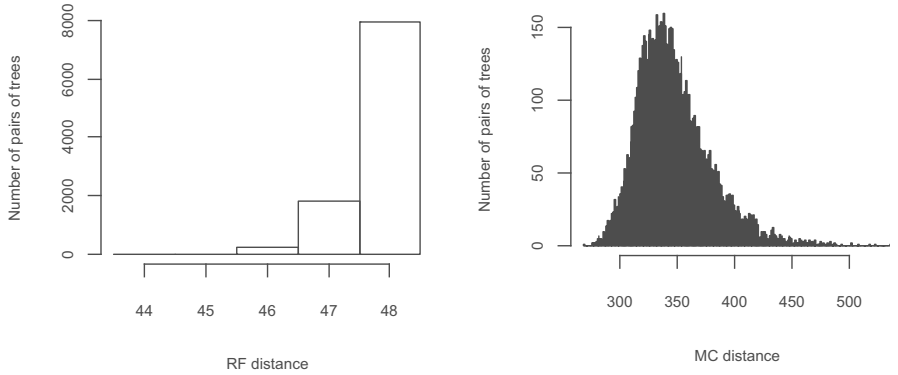

Fig. 10. Histograms of distances in the RF and MC metrics based on 10000 randomly generated pairs of binary trees with 50 leaves according to the Yule model.

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

Property	RF (unrooted)	RF (rooted)	MS	MC
Minimal positive distance	1	1	2	2
Number of trees at the minimal positive distance from a given tree	$2n-6$	$2n-4$	$0 \le x \le n-1$	$0 \le x \le n-1$
Maximal distance	$n-3$	$n-2$	$\frac{3}{8}n^2+O(n)$	$\frac{n^2-4-(n \mod 2)}{2} \le x \le \frac{3}{4}n^2+O(n)$
Distance of caterpillar trees constructed as in Fig. 2	$n-3$	$n-2$	$n-2$	$2n-4$
Average distance of random trees (in the uniform model)	$\Theta(n)$	$\Theta(n)$	$\Theta(n^{3/2})$	$\Theta(n^{3/2})$

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:

$$d^*(T^*, P) = \sum_{i=1}^{k} d(T^*_{|L(T_i)}, T_i), \tag{8}$$

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 (8) 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 supertees could be found.

The properties of the best found supertrees are presented in Table 4. The parsimony scores presented in Tables 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.

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

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 $Splits(L)$ (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

Table 4. Properties of the best supertrees received in particular trials. The number of rSPRs operations describes the number of moves performed from the particular starting tree after which no better supertree could be found.

Starting supertree	RF-supertree				MC-supertree			
	RF	MC	Pars. Score	# rSPRs	RF	MC	Pars. score	# rSPRs
1	80.5	718	394	83	72.5	183	253	123
2	88.5	787	411	76	54.5	155	237	119
3	75.5	686	353	79	68.5	182	261	106
4	94.5	744	456	57	73.5	209	265	100
5	**44.5**	194	**228**	25	60.5	168	245	31
6	104.5	900	481	63	50.5	**154**	230	116
7	77.5	575	370	68	71.5	180	252	102
Avg.	80.79	657.71	384.71	64.43	**64.50**	**175.86**	**249.00**	99.57

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

Alberich, R., Cardona, G., Rosselló, F. and Valiente, G. (2009). An algebraic metric for phylogenetic trees, *Applied Mathematics Letters* **22**(9): 1320–1324.

Aldous, D.J. (1991). The continuum random tree II: An overview, *in* M.T. Barlow and N. H. Bingham (Eds.), *Stochastic Analysis*, Cambridge University Press, Cambridge, pp. 23–70.

Bansal, M., Burleigh, J.G., Eulenstein, O. and Fernández-Baca, D. (2010). Robinson–Foulds supertrees, *Algorithms for Molecular Biology* **5**(1): 18.

Bansal, M.S., Dong, J. and Fernández-Baca, D. (2011). Comparing and aggregating partially resolved trees, *Theoretical Computer Science* **412**(48): 6634–6652.

Biedrzycki, R. and Arabas, J. (2012). KIS: An automated attribute induction method for classification of DNA sequences, *International Journal of Applied Mathematics and Computer Science* **22**(3): 711–721, DOI: 10.2478/v10006-012-0053-2.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R. D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. and Purvis, A. (2007). The delayed rise of present-day mammals, *Nature* **446**(7135): 507–512.

Blum, M.G.B., François, O. and Janson, S. (2006). The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance, *The Annals of Applied Probability* **16**(4): 2195–2214.

Boc, A., Philippe, H. and Makarenkov, V. (2010). Inferring and validating horizontal gene transfer events using bipartition dissimilarity, *Systematic Biology* **59**(2): 195–211.

Bogdanowicz, D. and Giaro, K. (2012). Matching split distance for unrooted binary phylogenetic trees, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**(1): 150–160.

Bogdanowicz, D., Giaro, K. and Wróbel, B. (2012). TreeCmp: Comparison of trees in polynomial time, *Evolutionary Bioinformatics* **8**: 475–487.

Bolikowski, L. and Gambin, A. (2007). New metrics for phylogenies, *Fundamenta Informaticae* **78**(2): 199–216.

Boorman, S.A. and Olivier, D.C. (1973). Metrics on spaces of finite trees, *Journal of Mathematical Psychology* **10**(1): 26–59.

Bordewich, M. and Semple, C. (2005). On the computational complexity of the rooted subtree prune and regraft distance, *Annals of Combinatorics* **8**(4): 409–423.

Brinkmeyer, M., Griebel, T. and Böcker, S. (2011). Polynomial supertree methods revisited, *Advances in Bioinformatics* **2011**: 524182.

Bryant, D. (1997). *Building Trees, Hunting for Trees, and Comparing Trees—Theory and Methods in Phylogenetic Analysis*, Ph.D. thesis, University of Canterbury, Christchurch.

Cardona, G., Llabrés, M., Rosselló, F. and Valiente, G. (2009a). Metrics for phylogenetic networks I: Generalizations of the Robinson–Foulds metric, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **6**(1): 46–61.

Cardona, G., Llabrés, M., Rosselló, F. and Valiente, G. (2009b). Metrics for phylogenetic networks II: Nodal and triplets

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

Cardona, G., Rossello, F. and Valiente, G. (2009). Comparison of tree-child phylogenetic networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **6**(4): 552–569.

Cardona, G., Llabrés, M., Rosselló, F. and Valiente, G. (2010). Nodal distances for rooted phylogenetic trees, *Journal of Mathematical Biology* **61**(2): 253–276.

Critchlow, D.E., Pearl, D.K. and Qian, C. (1996). The triples distance for rooted bifurcating phylogenetic trees, *Systematic Biology* **45**(3): 323–334.

Darlu, P. and Guénoche, A. (2011). TreeOfTrees method to evaluate the congruence between gene trees, *Journal of Classification* **28**(3): 390–403.

Davies, T.J., Barraclough, T.G., Chase, M.W., Soltis, P.S., Soltis, D.E. and Savolainen, V. (2004). Darwin's abominable mystery: Insights from a supertree of the angiosperms, *Proceedings of the National Academy of Sciences of the United States of America* **101**(7): 1904–1909.

Felsenstein, J. (2003). *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA.

Frąckiewicz, M. and Palus, H. (2011). KHM clustering technique as a segmentation method for endoscopic colour images, *International Journal of Applied Mathematics and Computer Science* **21**(1): 203–209, DOI: 10.2478/v10006-011-0015-0.

Gabow, H.N. and Tarjan, R.E. (1989). Faster scaling algorithms for network problems, *SIAM Journal on Computing* **18**(5): 1013–1036.

Górecki, P. and Eulenstein, O. (2012). A Robinson–Foulds measure to compare unrooted trees with rooted trees, *in* L. Bleris, I. Mandoiu, R. Schwartz and J. Wang (Eds.), *Bioinformatics Research and Applications*, Lecture Notes in Computer Science, Vol. 7292, Springer, Berlin/Heidelberg, pp. 115–126.

Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees, *Networks* **21**(1): 19–28.

Hayes, M., Walenstein, A. and Lakhotia, A. (2009). Evaluation of malware phylogeny modelling systems using automated variant generation, *Journal in Computer Virology* **5**(4): 335–343.

Hillis, D.M., Heath, T.A. and John, K.S. (2005). Analysis and visualization of tree space, *Systematic Biology* **54**(3): 471–482.

Kennedy, M., Page, R. D.M. and Prum, R. (2002). Seabird supertrees: Combining partial estimates of procellariiform phylogeny, *The Auk* **119**(1): 88–108.

Lin, Y., Rajan, V. and Moret, B.M.E. (2012). A metric for phylogenetic trees based on matching, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**(4): 1014–1022.

Ma, B., Li, M. and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses, *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology, RECOMB'98, New York, NY, USA*, pp. 182–191.

McKenzie, A. and Steel, M. (2000). Distributions of cherries for two models of trees, *Mathematical Biosciences* **164**(1): 81–92.

Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L. and Zhou, Y. (2003). TreeJuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility, *ACM Transactions on Graphics* **22**(3): 453–462.

Nguyen, N., Mirarab, S. and Warnow, T. (2012). MRL and SuperFine+MRL: New supertree methods, *Algorithms for Molecular Biology* **7**(1): 3.

Nye, T.M., Liò, P. and Gilks, W.R. (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees, *Bioinformatics* **22**(1): 117–119.

Orlin, J.B. and Ahuja, R.K. (1992). New scaling algorithms for the assignment and minimum mean cycle problems, *Mathematical Programming* **54**(1–3): 41–56.

Penny, D., Watson, E.E. and Steel, M.A. (1993). Trees from languages and genes are very similar, *Systematic Biology* **42**(3): 382–384.

Pompei, S., Loreto, V. and Tria, F. (2011). On the accuracy of language trees, *PLoS ONE* **6**(6): e20109.

Restrepo, G., Héber, M. and Llanos, E.J. (2007). Three dissimilarity measures to contrast dendrograms, *Journal of Chemical Information and Modeling* **47**(3): 761–770.

Robinson, D.F. and Foulds, L.R. (1981). Comparison of phylogenetic trees, *Mathematical Biosciences* **53**(1–2): 131–147.

Sackin, M.J. (1972). "Good" and "bad" phenograms, *Systematic Zoology* **21**(2): 225–226.

Semple, C. and Steel, M. (2003). *Phylogenetics*, Oxford University Press, Oxford.

Shao, K.-T. and Sokal, R.R. (1990). Tree balance, *Systematic Zoology* **39**(3): 266–276.

Steel, M. A. and Penny, D. (1993). Distributions of tree comparison metrics—some new results, *Systematic Biology* **42**(2): 126–141.

Stockham, C., Wang, L.-S. and Warnow, T. (2002). Statistically based postprocessing of phylogenetic analysis by clustering, *Bioinformatics* **18**(suppl 1): S285–S293.

Suri, R. and Warnow, T. (2010). Spruce, Website, `http://www.cs.utexas.edu/~phylo/software/spruce/`.

Swenson, M.S., Suri, R., Linder, C.R. and Warnow, T. (2011). An experimental study of Quartets MaxCut and other supertree methods, *Algorithms for Molecular Biology* **6**(1): 7.

Swofford, D. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4,* Sinauer Associates, Sunderland, MA.

Wang, J.T., Shan, H., Shasha, D. and Piel, W.H. (2005). Fast structural search in phylogenetic databases, *Evolutionary Bioinformatics Online* **1**: 37–46.

Williams, W.T. and Clifford, H.T. (1971). On the comparison of two classifications of the same set of elements, *Taxon* **20**(4): 519–522.

Damian Bogdanowicz received the M.Sc. and Ph.D. degrees in computer science from the Gdańsk University of Technology in 2006 and 2012, respectively. His research interests are bioinformatics, computational biology and discrete optimization.

Krzysztof Giaro received the M.Sc., Ph.D. and D.Sc. degrees (all in computer science) from the Gdańsk University of Technology in 1997, 1999 and 2004, respectively, and an M.Sc. in mathematics from the University of Gdańsk in 1998. Since 1997 he has been with the Gdańsk University of Technology, where he is presently an associate professor at the Department of Algorithms and System Modeling. He is also the head of the Department of Informatics at the Kotarbinski Academy of Informatics and Management in Olsztyn. He is the author or coauthor of dozens of papers and a book on *Introduction to Quantum Algorithms* (in Polish). His research interest is discrete optimization, algorithmic graph theory, task scheduling, quantum computing and bioinformatics.

Appendix

A1. Proof of Theorem 2

Proof.

1. The discussion is analogous to $d_{RF}(T_1,T_2) \leq d_{MS}(T_1,T_2) \leq |L|\, d_{RF}(T_1,T_2)$ for $T_1,T_2 \in U_L$ described by Bogdanowicz and Giaro (2012). Note that the complete bipartite graphs used for computing $d_{RF}(T_1,T_2)$ and $d_{MC}(T_1,T_2)$ differ only in the weights of the edges, and for clusters $c_1,c_2 \subsetneq L$ we have $h_{RF}(c_1,c_2) \leq h_C(c_1,c_2) \leq |L| h_{RF}(c_1,c_2) \leq 2(|L|-1)h_{RF}(c_1,c_2)$; similarly $\frac{1}{2} = h_{RF}(c,O) < h_C(c,O) = |c| \leq 2(|L|-1)h_{RF}(c,O)$ for cluster $c \subsetneq L$.

2. For the lower bound we use a similar result concerning binary unrooted trees and the MS distance (see Bogdanowicz and Giaro, 2012, Theorem 4.4). Let $T \in R^B_{L'}$ be an arbitrary tree, where $L \cap L' = \emptyset$, $|L'| = |L|$. Based on T_1, T_2 and T we construct two unrooted trees $T_1', T_2' \in U^B_{L \cup L'}$ by connecting the roots of T_1, T_2 with $r(T)$. We know that $d_{RF}(T_1',T_2') + 1 \leq d_{MS}(T_1',T_2')$. Finally, it is easy to observe that $d_{RF}(T_1',T_2') = d_{RF}(T_1,T_2)$ and $d_{MS}(T_1',T_2') = d_{MC}(T_1,T_2)$. For the upper bound note that clusters c_1, c_2 of the same tree fulfill the compatibility condition (c.c.), i.e., $c_1 \subseteq c_2$ or $c_2 \subseteq c_1$ or $c_1 \cap c_2 = \emptyset$. Let $A_1, A_2, \ldots, A_{d_{RF}(T_1,T_2)}$ form the set $\sigma_*(T_1) \setminus \sigma_*(T_2)$ and $|A_i| \leq |A_j|$ for $i < j$. We similarly sort the elements $B_1, B_2, \ldots, B_{d_{RF}(T_1,T_2)}$ of the set $\sigma_*(T_2) \setminus \sigma_*(T_1)$. We show that $\sum_{i=1}^{d_{RF}(T_1,T_2)} |A_i \oplus B_i| \leq (|L|-1)d_{RF}(T_1,T_2)$. If $A_i \neq L \setminus B_i$ for all i, then $|A_i \oplus B_i| < |L|$ and the result follows. Otherwise, let i be the smallest index such that $A_i = L \setminus B_i$ and without loss of generality we can assume that $|B_i| \geq |L|/2$. We consider two cases:

Case 1. $i < d_{RF}(T_1,T_2)$, then for $j > i$ by (c.c.) $B_i \subsetneq B_j$ and either $A_i \subsetneq A_j$ or $A_j \subsetneq B_i \subsetneq B_j$, so $|A_j \oplus B_j| \leq |L| - 2$. Since $|A_j \oplus B_j| \leq |L| - 1$ for $j < i$, the result holds.

Case 2. $i = d_{RF}(T_1,T_2)$. Suppose that there exists $B \in \sigma_*(T_2)$ such that $B_i \subsetneq B$. Then $B \in \sigma_*(T_1)$, but B is incompatible with A_i. Therefore, such B cannot exist. Hence, $L \setminus B_i = A_i \in \sigma_*(T_2)$ and we have a contradiction. ■

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 $T_{1|A}$. A cluster $s \in \sigma_*(T_{1|A})$ transforms into $s' \in \sigma_*(T_1)$, 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 $T_{1|A}$, then one *new cluster* $s_{new} \in \sigma_*(T_1)$, which is not a corresponding cluster, appears. The case of transformation of $T_{2|A}$ 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 $h_C(s,t) \leq h_C(s',t') \leq h_C(s,t) + 1$. Consider a pairing $\{(s_i,t_i) : i = 1,\ldots,k\}$, where $s_i \in \sigma_*(T_{1|A}) \cup \{O\}, t_i \in \sigma_*(T_{2|A}) \cup \{O\}$, $k = \max_{i=1,2} |\sigma_*(T_{i|A})| \leq n-3$ and $\sum_{i=1}^k h_C(s_i,t_i) = d_{MC}(T_{1|A},T_{2|A})$. 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 $\sigma_*(T_{1|A}) \cup \{O\}$ and $\sigma_*(T_{2|A}) \cup \{O\}$ 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 $s_{new} \in \sigma_*(T_1)$, $t_{new} \in \sigma_*(T_2)$ and $(s_{new}, t_{new}) \in M$. Then

$$\begin{aligned} d_{MC}(T_1, T_2) &= \sum_{(s,t)\in M} h_C(s,t) \\ &\geq \sum_{(s,t)\in M'} h_C(s,t) \\ &\geq d_{MC}(T_{1|A}, T_{2|A}). \end{aligned}$$

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 $s_{new} \in \sigma_*(T_1)$, $t_{new} \in \sigma_*(T_2)$ in M, i.e., $(s_{new}, x_2), (x_1, t_{new}) \in M$. Then we extend M' by a pair (x_1', x_2'), where x_i' is an element that corresponds to x_i; hence

$$\begin{aligned} d_{MC}(T_1, T_2) &= \sum_{(s,t)\in M} h_C(s,t) \\ &\geq \sum_{(s,t)\in M'} h_C(s,t) - h_C(x_1', x_2') \\ &\geq d_{MC}(T_{1|A}, T_{2|A}) - n + 1. \end{aligned}$$

Case 4. There is only one new cluster; assume that it is $s_{new} \in \sigma_*(T_1)$ and $(s_{new}, x) \in M$. Then we extend M' by a pair (O, x'), where x' is an element that corresponds to x; hence

$$\begin{aligned} d_{MC}(T_1, T_2) &= \sum_{(s,t)\in M} h_C(s,t) \\ &\geq \sum_{(s,t)\in M'} h_C(s,t) - h_C(x', O) \\ &\geq d_{MC}(T_{1|A}, T_{2|A}) - n + 1. \end{aligned}$$

■

A3. Proof of Theorem 7

Proof. Consider $T_1, T_2 \in U_L^B$ 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 is exactly one corresponding cluster in $\sigma(T_1') \setminus \{A, B, L\}$ such that the equivalent of a cluster X is a split $X|L \setminus X$. An analogous relation may be introduced between clusters $Y \in \sigma(T_2') \setminus \{C, D, L\}$ and the splits $Y|L\setminus Y \in \beta(T_2)\setminus\{C|D\}$. Moreover, $|X \oplus Y| \geq h_S(X|L \setminus X, Y|L \setminus Y)$.

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 $\sum_{(c,c')\in M'} |c \oplus c'| = d_{MC}(T_1', T_2')$ 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 $\sum_{(s,s')\in M} h_S(s,s') \leq d_{MC}(T_1', T_2')$. 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 $(X_1|L \setminus X_1, X_2|L \setminus X_2)$. Depending on the rest of M', i.e., pairs with clusters A, B, C, D in M', only the following three situations are possible:

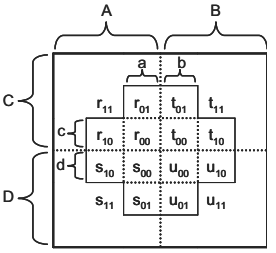

Fig. A1. Venn diagram of possible intersections of the sets A, B, C, D, a, b, c, d in the proof of Theorem 7.

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: $(A|B, Y|L \setminus Y)$ and $(X|L \setminus X, C|D)$.

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: $m' = \{(A, c), (a, D), (B, d), (b, C)\} \subseteq M'$ or symmetric variant $\{(A, d), (a, C), (B, c), (b, D)\} \subseteq M'$, where $a \subsetneq A$, $b \subsetneq B$, $c \subsetneq C$ and $d \subsetneq D$. We consider the first case, the second one is analogous. We now construct two pairings of splits:

$$\begin{aligned} m_1 = \{&(a|L \setminus a, c|L \setminus c), (b|L \setminus b, C|D), \\ &(A|B, d|L \setminus d)\}, \\ m_2 = \{&(a|L \setminus a, C|D), (b|L \setminus b, d|L \setminus d), \\ &(A|B, c|L \setminus c)\}, \end{aligned}$$

and two pairings of clusters:

$$\begin{aligned} m_1' &= \{(a, c), (b, C), (B, d)\}, \\ m_2' &= \{(a, D), (b, d), (A, c)\}. \end{aligned}$$

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

$$\sum_{(s,s')\in m_i} h_S(s,s') \leq \sum_{(c,c')\in m_i'} h_C(c,c').$$

Based on calculations of the total costs of pairings m', 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.

Set	Definition	m'	m'_1	m'_2	$\Delta m'_1$	$\Delta m'_2$
r_{00}	$a \cap c$	2	1	1	-1	-1
r_{01}	$(a \cap C) \setminus (a \cap c)$	3	2	2	-1	-1
r_{10}	$(A \cap c) \setminus (a \cap c)$	1	2	0	**1**	-1
r_{11}	$(A \cap C) \setminus (a \cup c)$	2	1	1	-1	-1
s_{00}	$a \cap d$	2	2	2	0	0
s_{01}	$(a \cap D) \setminus (a \cap d)$	1	1	1	0	0
s_{10}	$(A \cap d) \setminus (a \cap d)$	3	1	3	-2	0
s_{11}	$(A \cap D) \setminus (a \cup d)$	2	0	2	-2	0
t_{00}	$b \cap c$	2	2	2	0	0
t_{01}	$(b \cap C) \setminus (b \cap c)$	1	1	1	0	0
t_{10}	$(B \cap c) \setminus (b \cap c)$	3	3	1	0	-2
t_{11}	$(B \cap C) \setminus (b \cup c)$	2	2	0	0	-2
u_{00}	$b \cap d$	2	1	1	-1	-1
u_{01}	$(b \cap D) \setminus (b \cap d)$	3	2	2	-1	-1
u_{10}	$(B \cap d) \setminus (b \cap d)$	1	0	2	-1	**1**
u_{11}	$(B \cap D) \setminus (b \cup d)$	2	1	1	-1	-1

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