JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 5, 2017
© Mary Ann Liebert, Inc.
Pp. 422–435
DOI: 10.1089/cmb.2016.0204

Comparing Phylogenetic Trees by Matching Nodes Using the Transfer Distance Between Partitions

DAMIAN BOGDANOWICZ and KRZYSZTOF GIARO

ABSTRACT

Ability to quantify dissimilarity of different phylogenetic trees describing the relationship between the same group of taxa is required in various types of phylogenetic studies. For example, such metrics are used to assess the quality of phylogeny construction methods, to define optimization criteria in supertree building algorithms, or to find horizontal gene transfer (HGT) events. Among the set of metrics described so far in the literature, the most commonly used seems to be the Robinson–Foulds distance. In this article, we define a new metric for rooted trees—the Matching Pair (MP) distance. The MP metric uses the concept of the minimum-weight perfect matching in a complete bipartite graph constructed from partitions of all pairs of leaves of the compared phylogenetic trees. We analyze the properties of the MP metric and present computational experiments showing its potential applicability in tasks related to finding the HGT events.

Keywords: matching pair distance, minimum-weight perfect matching, phylogenetic tree comparison, phylogenetic tree metric.

1. INTRODUCTION

Ability to quantify dissimilarity of different phylogenetic trees describing the relationship between the same group of taxa is required in various types of phylogenetic studies (Boc et al., 2010; Koonin et al., 2011; Smith et al., 2013; Whidden et al., 2014; Chaudhary et al., 2013). The most common application of such distances is measuring the difference of trees inferred from the same data but using different methods, for example, the distance between the ‘‘true’’ tree and a tree reconstructed using a particular method can be used as an indicator of the particular method's accuracy [see Bogdanowicz et al. (2012)].

Metrics are also used directly in some of phylogenetic tree construction methods, for example, in super trees construction, whereby the sum of distances in a particular metric between input trees, whose leaf sets overlap partially, and a super tree (containing all the taxa) is an object of minimization [see Lin et al. (2009); Bansal et al. (2010); Chaudhary et al. (2013); Whidden et al. (2014)]. Quantifying similarities between phylogenies is also useful in an analysis and visualization of a group of phylogenetic trees (Hillis et al., 2005).

The values of some metrics are defined as a minimum number of operations of a particular type needed to transform one phylogenetic tree into another, for example, nearest neighbor interchange, subtree prune

Department of Algorithms and System Modeling, Gdansk University of Technology, Gdansk, Poland.

and regraft (SPR), and tree bisection and reconnection operations. In particular, SPR metric can be used in modeling and analyzing horizontal gene transfers (HGTs) (Section 5). Unfortunately, the computation of the value of all three metrics is proven to be NP-hard (DasGupta et al., 1997; Allen and Steel, 2001; Bordewich and Semple, 2005; Hickey et al., 2008). Fortunately, there are methods for the analysis of HGTs that incorporate other polynomially computable distances (Boc et al., 2010). As metrics are important tools in phylogenetics, properties of the metric spaces are often subjects of ongoing research (Gordon et al., 2013; Humphries and Wu, 2013).

Recently (Bogdanowicz and Giaro, 2012), we described a general framework for defining phylogenetic metrics, which is based on computing a minimum-weight perfect matching in bipartite graphs. Each partition of such a graph describes one of the compared trees. The weights of the edges of the graph represent the distance between the elements (e.g., splits and clusters) used to describe analyzed trees. Using this framework, we described two metrics: the Matching Split (MS) distance for unrooted trees (Bogdanowicz and Giaro, 2012) and the Matching Cluster (MC) distance for rooted trees (Bogdanowicz and Giaro, 2013). Both metrics are polynomially computable and can be regarded as generalizations of the Robinson–Foulds (RF) metric (Robinson and Foulds, 1981). Recently, another generalization of RF metric was formulated (Böcker et al., 2013). This metric is also based on the concept of matchings similar to MC, but with additional assumptions related to the phylogenetic information that, unfortunately, makes computation NP-hard (Böcker et al., 2013).

Although the concept of matching metrics is relatively new, the MS distance constructed based on this idea has already been found useful in biological studies (Smith et al., 2013). Both the MS and MC metrics have some advantages in comparison with the RF metric (despite the fact that they operate on the same tree representation, i.e., splits or clusters). They are more discriminative. Both metrics have a property that a relocation of a small subtree results in a relatively small distance increase (Bogdanowicz and Giaro, 2012, 2013). Note that the RF metrics do not have such a feature (Bogdanowicz and Giaro, 2012), because it is possible for two trees to have a maximum RF distance although they only differ in the position of a single leaf. It should also be noted that there is no simple answer to the question about ''the best'' phylogenetic metric in general. Depending on application, some of the phylogenetic metric's features can be more or less desired, for example, if the main criterion is computational efficiency, then most probably the RF distance will be the most suitable, but if discriminative power is considered, then other metrics (e.g., MS) might be a better choice.

In this article, we define and analyze the properties of a new phylogenetic metric based on the already mentioned matching concept—the Matching Pair (MP) distance (Section 3). We show that the MP distance has many interesting features (Section 4), for example, similarly to MC, a small change in a tree structure results in a small change in the MP distance, which does not hold for instance for the RF metric, but in contrast to MC (and MS), average MP distance between random trees grows as fast as its diameter. Moreover, MP performs very well in comparison with five other metrics, known in the literature, when applied to the detection of HGTs (Section 5).

2. BASIC DEFINITIONS AND NOTATION

For sets A, B let $A \oplus B = (A \backslash B) \cup (B \backslash A)$ be their symmetric difference, $|A|$ denote the cardinality of set A, and $2^A = \{B : B \subseteq A\}$. Let $G = (V, E)$ be a *graph* with a set of vertices V and a set of edges E. A family of nonempty sets $A_1, \ldots, A_k$ such that $\bigcup_{i=1}^{k} A_i = A$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ is a *partition* of the set A. A *bipartite graph* $G(V_1, V_2, E)$ has vertices partitioned 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 *matching* $M \subseteq E$ in a graph $G = (V, E)$ is a set of pairwise nonadjacent 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 *rooted phylogenetic tree* $T = (V, E)$ is a tree whose *leaves*, that is, vertices (nodes) of degree 1, are labeled bijectively by the elements of a finite set L (representing the species), there is exactly one distinguished nonleaf vertex $r(T) \in V \backslash L$ called the *root* and none of the vertices of $V \backslash \{r(T)\}$ has degree 2.

Present-day species under examination form the finite set L and are represented by leaves of a tree. Internal vertices, that is, members of $V\backslash L$, 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 the leaves with their labels, that is, for a phylogenetic tree T, by $L(T)$ we denote the set of leaves of T or the set of labels of those leaves. By $L^{(2)}$ we denote the set of all unordered pairs of leaves, that is, $L^{(2)}=\{\{x,y\}:x,y\in L, x\neq y\}$ and $|L^{(2)}|=|L|(|L|-1)/2$.

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. 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 notations 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 $a\leq_T b$ (i.e., a is a descendant of b and b is an ancestor of a) 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* $c(v)\subseteq L$, that is, 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 other clusters are *nontrivial*. By $\sigma(T)$ and $\sigma_*(T)$, we denote families of all clusters of T and all nontrivial clusters of T, respectively. 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].

Let $A\subseteq L$, and let $T(A)$ be a minimal subgraph of T that connects leaves of A and choose its root as 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. The tree $T_{|A}$ contains all phylogenetic information about evolutionary history of taxa from A, which is represented by T. A tree $T\in R_A$, $A\subseteq L$, is an *agreement subtree* for trees $T_1,T_2\in R_L$ if $T=T_{1|A}=T_{2|A}$. An agreement subtree having maximum number of leaves is called a *Maximum Agreement Subtree* (MAST) (Finden and Gordon, 1985).

The *lowest common ancestor* (*LCA*), also called the *most recent common ancestor*, of a pair of leaves u, v of a rooted tree T, $l(u,v)$, is the closest vertex to $r(T)$ on the path connecting u and v in T. To every internal vertex v of $T\in R_L$, we can assign a set of pairs of leaves $lp(v)$ for which v is the LCA, that is, $lp(v)=\{\{x,y\}\in L^{(2)}:l(x,y)=v\}$. We will call the set $lp(v)$ the *pair set* of v. By $\gamma(T)$, we denote the family of all pair sets of T, so $\gamma(T)$ is a partition of the set $L^{(2)}$ determined by T. Note that a rooted phylogenetic tree T is uniquely described by a set $\gamma(T)$ because $\gamma(T)$ determines $\sigma_*(T)$ and we have $c(v)=\bigcup_{z\in lp(v)} z$ for $v\in V\backslash L$.

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 RF distance (Robinson and Foulds, 1981) based on clusters.

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

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

3. MATCHING METHOD

Before we proceed to the formal definition of a new metric, we recall the general construction of matching metrics presented in Bogdanowicz and Giaro (2012).

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:*

*A version of this definition without factor 1/2 is also used in the literature.

- *for arbitrary $s, t \in \mathbb{Z}_{\geq 0}$ 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 $w(\{a_i, b_j\}) = h(l(a_i), l(b_j))$.*

The function d_h is a metric on 2^D (i.e., for any $A, B, C \in 2^D$ holds $d_h(A, B) = 0 \Leftrightarrow A = B$, $d_h(A, B) = d_h(B, A)$, $d_h(A, B) + d_h(B, C) \geq d_h(A, C)$) and the value of $d_h(A, B)$ does not depend on s and t (when $s - t = |A| - |B|$). Hence, we can always assume that $\min\{s, t\} = 0$ and $\max\{s, t\} = ||A| - |B||$. Moreover, $d_h(A, B) = d_h(A \backslash B, B \backslash A)$ (Bogdanowicz and Giaro, 2012).

In two recent articles (Bogdanowicz and Giaro, 2012, 2013), we described properties of phylogenetic metrics created according to the shown schema, that is, the MS distance for unrooted trees and the MC distance for rooted tree. In the case of the matching metric for rooted trees, the dissimilarity between two clades A, B can be measured as the number of elements that appear in one of the clade but not in the other, that is, the cardinality of the set $A \oplus B$. As the cardinality of $A \oplus B$ introduces a metric space structure in an arbitrary family of finite sets, we obtain (Bogdanowicz and Giaro, 2013) the following:

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

$$d_{MC}(T_1, T_2) = d_{h_C}(\sigma(T_1), \sigma(T_2)) = d_{h_C}(\sigma_*(T_1), \sigma_*(T_2)). \tag{2}$$

In this work, we want to recall the definition of a metric introduced in Bogdanowicz (2008), which, similarly to MC, is based on the schema presented in Definition 2. The rest of the article contains a detailed study of properties of the metric.

Definition 4. *Let $T_1, T_2 \in R_L$ be rooted phylogenetic trees, $h_P : 2^{L^{(2)}} \times 2^{L^{(2)}} \to \mathbb{Z}_{\geq 0}$ be such that $h_P(A, B) = |A \oplus B|$, and let $O = \emptyset$. According to Definition* 2 *we define the MP distance $d_{MP} : R_L \times R_L \to \mathbb{Z}_{\geq 0}$ as*

$$d_{MP}(T_1, T_2) = \frac{1}{2} d_{h_P}(\gamma(T_1), \gamma(T_2)). \tag{3}$$

Observe that the MP distance between phylogenetic trees $T_1, T_2 \in R_L$ is equal to the transfer distance between partitions $\gamma(T_1)$ and $\gamma(T_2)$ of the set $L^{(2)}$. *The transfer distance* between two partitions of a finite set S, equal to the minimum number of elements that must be moved from one set of a partition to another (possibly empty) to turn first partition into the second (Guénoche, 2011) has been introduced by Régnier (1965). The equivalence between the value of a minimum-weight perfect matching in a bipartite graph constructed according to Definitions 2 and 4 and the value of the transfer distance has been proven by Day (1981), for more details about the relation, see also Charon et al. (2006); Denœud (2008); Gusfield (2002); and Konovalov et al. (2005). An equivalent definition of the transfer distance between two partitions P and P' of a set S, which can be found in Gusfield (2002), says that the distance is equal to ''the minimum number of elements that must be deleted from S, so that the two induced partitions (P and P' restricted to the remaining elements) are identical.''

For example, we calculate the MP distance between trees as shown in Figure 1. The weight of a minimum-weight perfect matching in a bipartite graph shown in Figure 1 is equal to 6, so $d_{MP}(T_1, T_2) = 3$. The MP metric between two trees $T_1, T_2 \in R_n$ can be computed in time $O(n^{5/2} \log(n))$ using scaling algorithms for finding minimum-weight perfect matchings (Gabow and Tarjan, 1989; Orlin and Ahuja, 1992).

Note that, although both the MP and MC metrics are defined as the weight of a minimum-weight perfect matching in a bipartite graph, there is no direct relationship between d_{MC} and d_{MP}. In other words, if we know $d_{MC}(T, T')$, we cannot compute easily $d_{MP}(T, T')$.

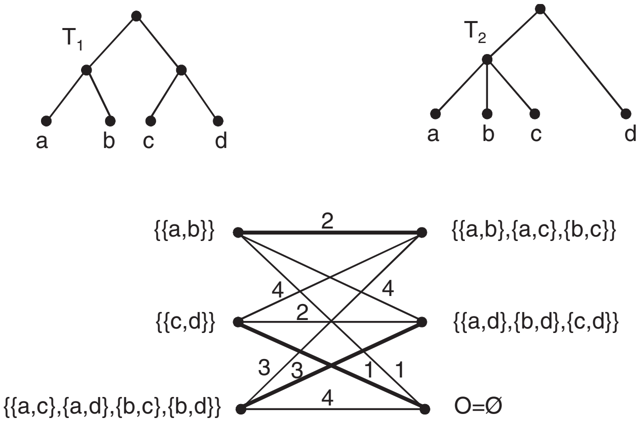

FIG. 1. Calculation of the MP distance between trees T_1 and T_2. The bipartite graph of their partitions has a perfect matching of minimum weight equal to 6. MP, matching pair.

4. PROPERTIES OF THE MATCHING PAIR DISTANCE FOR ROOTED PHYLOGENETIC TREES

4.1. Small topological transformations

In the case of the MC metric (as well as for the MS metric for unrooted trees), it was shown that adding a new leaf to compared trees can decrease their distance (Bogdanowicz and Giaro, 2012, 2013). Regarding this feature, the MP metric behaves more similarly to RF metric, which is expressed in Theorem 1 and Corollary 2. Before we present the theorem, we give two lemmas that we then use in the proof of the theorem.

Lemma 1. *Let $T_1, T_2 \in R_n^B$ be rooted binary phylogenetic trees constructed as presented in Figure 2, then $d_{MP}(T_1, T_2) = (1/2)n^2 - (3/2)n + 1$.*

Proof. See Supplementary Materials. ■

Lemma 2. *Let $T_3, T_4 \in R_n^B$ be rooted binary phylogenetic trees constructed as presented in Figure 3, then $d_{MP}(T_3, T_4) \geq (1/2)n^2 - (3/2)n + 2$.*

Proof. See Supplementary Materials. ■

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

$$d_{MP}(T_1, T_2) \geq d_{MP}(T_{1|A}, T_{2|A}), \tag{4}$$

$$d_{MP}(T_1, T_2) \leq d_{MP}(T_{1|A}, T_{2|A}) + n - 1, \tag{5}$$

and both inequalities can be tight.

Proof. Let $L = A \cup \{z\}$ and $A_z = \{\{x, z\} : x \in A\}$. We have ■

$$\gamma(T_{i|A}) = \{S \backslash A_z : S \in \gamma(T_i)\} \backslash \{\emptyset\}, i = 1, 2. \tag{6}$$

Let $M = \bigcup_{i=1}^{k} \{(g_i, h_i)\}$, where $g_i \in \gamma(T_1) \cup \{\emptyset\}$, $h_i \in \gamma(T_2) \cup \{\emptyset\}$, be an optimal pairing that defines the distance $d_{MP}(T_1, T_2)$. Let M' be a pairing formed based on M by removing from it all leaf pairs containing z,

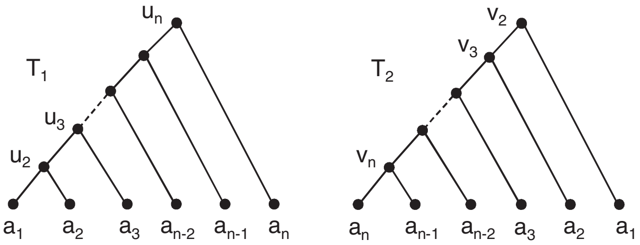

FIG. 2. Two caterpillar trees labeled in opposite directions are examples of very distant trees in the MP metric.

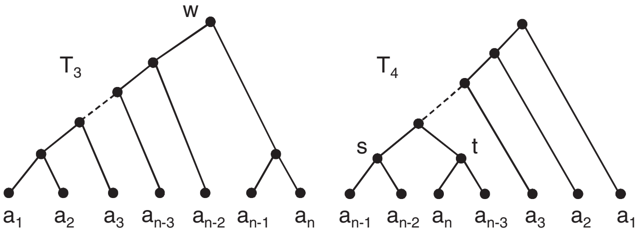

FIG. 3. An example of trees at a distance of $(1/2)n^2 - (3/2)n + 2$ in the MP metric.

that is, $M' = \bigcup_{i=1}^{k}\{(g'_i, h'_i)\}$, where $g'_i = g_i \backslash A_z$ and $h'_i = h_i \backslash A_z$. Note that M' is a pairing of elements of $\gamma(T_{1|A}) \cup \{\emptyset\}$ with sets of $\gamma(T_{2|A}) \cup \{\emptyset\}$. Moreover, each pair set of $\gamma(T_{1|A})$, as well as each element of $\gamma(T_{2|A})$, appears in $M\prime$ exactly once. Hence, we have

$$\begin{aligned} d_{MP}(T_1, T_2) &= \frac{1}{2}\sum_{i=1}^{k} |g_i \oplus h_i| \\ &= \frac{1}{2}\sum_{i=1}^{k}(|g\prime_i \oplus h\prime_i| + |(g_i \cap A_z) \oplus (h_i \cap A_z)|) \\ &\geq \frac{1}{2}\sum_{i=1}^{k} |g\prime_i \oplus h'_i| \geq d_{MP}(T_{1|A}, T_{2|A}). \end{aligned}$$

Tightness of this inequality is straightforward, which can be seen, for example, for trees defined so that $T_1 = T_2 \in R_L$ and $T_{1|A} = T_{2|A}$.

The second inequality follows directly from the fact that the MP distance is a special case of the transfer distance and Equation (6), because $|A_z| = n - 1$.

Let $L = \{a_1, a_2, \ldots, a_n\}$ and $a_n = z$, hence $A = L\backslash\{a_n\}$. For a tight example of the inequality, consider trees $T'_1, T'_2 \in R_A^B$ constructed as presented in Figure 2 but for $n-1$ leaves. Note that trees $T_3, T_4 \in R_L^B$ (Fig. 3) can be constructed by adding leaf a_n to trees T'_1, T'_2, respectively. Hence, by Lemmas 1 and 2, we obtain that $d_{MP}(T_3, T_4) - d_{MP}(T'_1, T'_2) \geq n - 1$. ■

Corollary 1. *Let $T_3, T_4 \in R_n^B$ be rooted binary phylogenetic trees constructed as presented in Figure 3, then $d_{MP}(T_3, T_4) = (1/2)n^2 - (3/2)n + 2$.*

Proof. The proof follows from Lemmas 1 and 2 and Equation (5). ■

Corollary 2. *For $A \subset L$ and $T_1, T_2 \in R_L$ the following inequality holds:*

$$d_{MP}(T_1, T_2) \geq d_{MP}(T_{1|A}, T_{2|A}).$$

Proof. The proof follows from repeated application of Theorem 1. ■

It seems to be a very natural property, a similar inequality holds for other classical metrics for phylogenetic trees, for example, RF and Triples metrics (Critchlow et al., 1996), but it is not fulfilled, for example, for MC (Bogdanowicz and Giaro, 2013).

Corollary 3. *Let $T_1, T_2 \in R_L$ and N be the number of leaves in MAST for trees T_1, T_2. The following inequality holds:*

$$d_{MP}(T_1, T_2) \leq \frac{1}{2}(|L|^2 - |L| - N^2 + N).$$

Proof. Let $X \subset L$, $|L| = n$, $|X| = N$, and $T_* \in R_X$ be a MAST for trees $T_1, T_2 \in R_L$, so $T_* = T_{1|X} = T_{2|X}$ and T_I (as well as T_2) can be obtained from T_* by a sequence of additions of leaves from $L\backslash X$. Based on Equation (5) from Theorem 1, it is clearly seen that by adding a single leaf to T_* we can obtain trees at the MP distance of at most N. Therefore, by adding $n - N$ leaves to T_* we can obtain trees at the MP distance less than or equal to $N + (N+1) + \cdots + (n-2) + (n-1) = (n^2 - n - N^2 + N)/2$. ■

The metric space determined by the MC distance has an interesting feature, which we can call ''regularity'' or ''lack of isolated islands'' (Bogdanowicz and Giaro, 2013), that is, any two trees of R_L can be connected by a sequence of trees defined so that any two subsequent elements of it are at a distance of 4 at the most. The MP metric, however, does not have this property.

Lemma 3. *For each $s \in \mathbb{Z}_{>0}$ and a finite set L with $|L| > 8s$, there exists a partition $S \cup S' = R_L$, $S \cap S' = \emptyset$ such that*

$$\forall_{T \in S} \forall_{T' \in S'} d_{MP}(T, T') > s.$$

Proof. See Supplementary Materials. ■

Note that there are metrics, for example, Quartet metric (QT) (Estabrook et al., 1985), whereby even the minimal distance between different trees grows together with the number of leaves. In the case of MP, such a situation does not hold. Moreover, for any tree $T \in R_L$, there exists a tree $T' \neq T \in R_L$ close to T, that is, $d_{MP}(T, T') \leq 4$. The same remains true for binary trees, that is, if $T \in R_L^B$ then T' can be chosen from R_L^B. Such a tree can be constructed based on T using one of the operations presented in Figure 4. The operations 1 and 2 are feasible within a family of binary trees. The MP distance between trees after performing operation 1 is equal to 4, the distance between trees taking part in operations 2 and 3 equals 1. Therefore, operations 2 and 3 lead to the creation of a *neighboring* tree (a tree being at the minimal positive distance) in MP.

4.2. Matching pair metric space diameter

One of the important and frequently studied properties of any phylogenetic metric space is its diameter, that is, the maximum distance between n-leaf trees. Diameters of phylogenetic metrics indicate the range, in which a particular metric defines differences between trees. Usually, more discriminative metrics have larger diameters, provided that the minimum distance between distinct trees is constant or grows much slower than the diameter. In this section we present bounds on the maximal distance in the MP metric.

Theorem 2 (Charon et al., 2006). *Let P, Q be partitions of a set X having the cardinality at most p and q, $p \leq q$, respectively. The maximum transfer distance between such partitions is equal to $|X| - \lceil |X|/q \rceil$.*

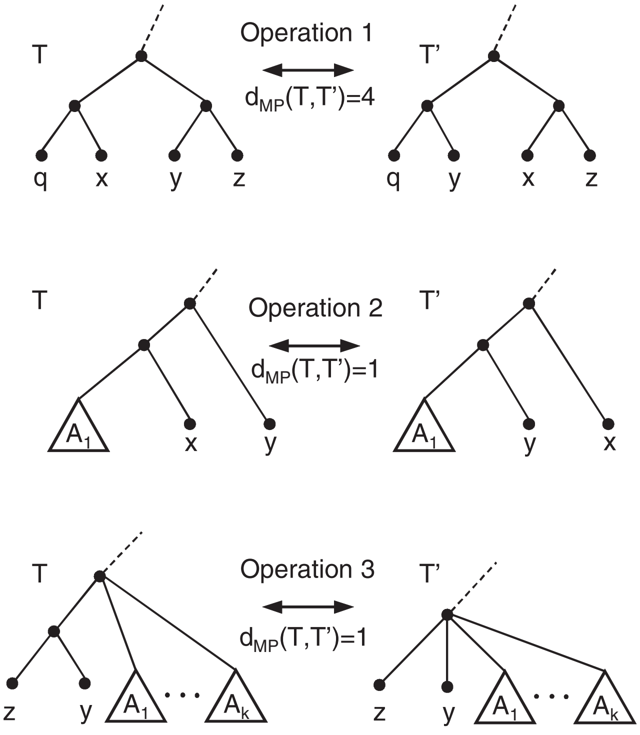

FIG. 4. Operations that create close trees. By letters A_i, $i = 1, \ldots, k$ we denote single leaves or rooted subtrees.

Theorem 3. *The maximal distance in the MP metric between trees of R_n can be characterized as*

$$\frac{1}{2}n^2 - \frac{3}{2}n + 2 \leq \max_{T_1, T_2 \in R_n^B} d_{MP}(T_1, T_2) \tag{7}$$

$$\leq \max_{T_1, T_2 \in R_n} d_{MP}(T_1, T_2) \leq \frac{1}{2}n^2 - n. \tag{8}$$

Proof. By Corollary 1, the lower bound is realized by the trees shown in Figure 3. The upper bound can be obtained using Theorem 2 by taking $P = \gamma(T_1)$, $Q = \gamma(T_2)$, where $|X| = n(n-1)/2$, $q < n$. ■

Construction of distant trees in the MP metric is presented in Figure 3. The trees are at maximum possible MP distance in R_n for $3 \leq n \leq 6$. We suspect that Figure 3 shows trees at maximum possible MP distance also for $n \geq 7$ and Equation (7) is tight.

4.3. Distances of random trees

To fully interpret the level of dissimilarity of two trees using the value of the distance between them in a particular metric, a reference value is usually needed. In most cases, the average distance between random trees generated according to a particular model can be used for such purposes. In this section, we consider the distance in the MP metric between binary phylogenetic trees drawn independently from any label-invariant model. A model of random phylogenetic trees is *label invariant* if the probability of a tree remains constant under an arbitrary permutation of its taxa labels (Steel and Penny, 1993). This is a natural assumption stating that the probability of a particular phylogenetic tree depends on its shape and does not depend on labeling of its leaves. It is easy to observe that the following lemma holds.

Lemma 4. *Let $A \subseteq L$ and T be a tree chosen randomly from R_L^B according to any label-invariant model. Then, the probability distribution of trees $T_{|A}$ over R_A^B is also label invariant.*

Two models of random phylogenetic trees can be distinguished as those that are most often used in the literature: the *uniform model*, whereby equal probability is assigned to each possible tree, and the *Yule model*, whereby trees are constructed iteratively: starting from three random taxa, new taxa (chosen randomly) are added to a branch connected to a leaf (also chosen uniformly at random). Both models are label invariant [see McKenzie and Steel (2000) for more information].

The main result in this section, expressed in Theorem 4, concerns estimation of an asymptotic behavior of the expected MP distance between two random binary trees. To prove this fact, we utilized a combinatorial tool called *Steiner triple system* [see Colbourn and Dinitz (2006)], which is a pair $(V, \mathcal{B})$ of a set V together with a collection $\mathcal{B}$ of three-element subsets (called *blocks* or *triples*) of V with the property that every two-element subset of V occurs in exactly one block $B \in \mathcal{B}$.

Lemma 5 (Kirkman, 1847). *If V is a set of cardinality $n = |V| > 1$ such that $|V| \equiv 1 \bmod 6$ or $|V| \equiv 3 \bmod 6$, then there exists a Seiner triple system, in which $|\mathcal{B}| = n(n-1)/6$.*

A term *migrated pair* has been introduced by Chen (2012) in a general context of partition distance. For two partitions P_1 and P_2 of a fine set V, a set $\{s, t\} \in V^{(2)}$ is a migrated pair if s and t are in the same set in one partition and in different sets in the other. It is easy to see from Chen (2012) that the cardinality of any set $S \subseteq V^{(2)}$ of disjoint migrated pairs is a lower bound on a partition distance between P_1 and P_2, because at least one element from any migrated pair has to be deleted to make considered partitions identical.

Theorem 4. *For rooted trees $T_{1_n}, T_{2_n} \in R_L^B$, $n = |L|$ generated independently at random according to any label-invariant model, for example, the Yule model or the uniform model, their expected distance is*

$$\mathbb{E}[d_{MP}(T_{1_n}, T_{2_n})] = \Theta(n^2). \tag{9}$$

Proof. The upper bound $O(n^2)$ on the expected distance follows from Theorem 3. Therefore, we have to prove that the lower bound is $\Omega(n^2)$ and by Theorem 1, Equation (4), and Lemma 4 we may also assume that $n = 6k + 1$, $k \in Z_{>0}$. We consider a tree induced by unordered triples $\{x, y, z\} \subseteq L$ of leaves. There are

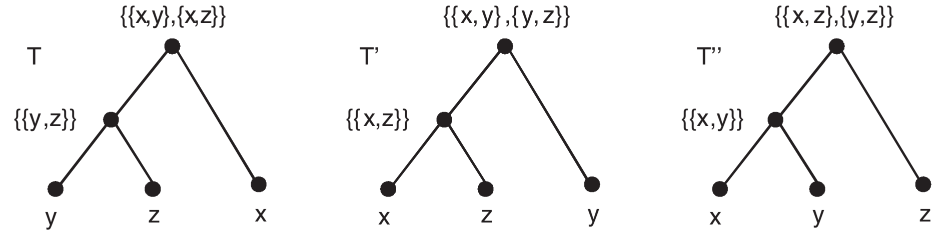

FIG. 5. The three possible trees having $\{x, y, z\}$ leaf set.

only three possible subtrees (Figure 5). Let $L^{(3)} = \{\{x, y, z\} : x, y, z \in L, x \neq y \neq z \neq x\}$ be a family of all unordered triples of leaves of L. We define the indicator function $I : L^{(3)} \times R_L^B \times R_L^B \to \{0, 1\}$ such that $I(\{x, y, z\}, \hat{T}, \check{T}) = 1$ if and only if induced trees $\hat{T}_{|\{x,y,z\}}$, $\check{T}_{|\{x,y,z\}}$are different.

Let $(L, \mathcal{B})$, $|L| = n$ be a Steiner triple system. Such a system exists based on Lemma 5 and the fact that $n = 6k + 1$, $k \in Z_{>0}$. Consider a triple $B = \{x, y, z\} \in \mathcal{B}$. If the set B induces different subtrees in the compared trees $T_{1_n}, T_{2_n} \in R_L^B$, then note that at least one pair of leaves from B has to be transferred between pair sets of $\gamma(T_{1_n})$ to transform it to $\gamma(T_{2_n})$ (Fig. 5). Hence, if $I(B, T_{1_n}, T_{2_n}) = 1$, then $B^{(2)}$ contains a migrated pair for partitions $\gamma(T_{1_n})$, $\gamma(T_{2_n})$ of the set $L^{(2)}$. For example, a pair $\{s = \{x, y\}, t = \{x, z\}\}$ is a migrated pair for partitions induced by trees T and T' in Figure 5. Moreover, sets $B^{(2)}$ are disjoint for different elements B of a Steiner system $(L, \mathcal{B})$, hence we obtain the following relation:

$$d_{MP}(T_{1_n}, T_{2_n}) \geq \sum_{B \in \mathcal{B}} I(B, T_{1_n}, T_{2_n}). \tag{10}$$

Thus by taking the expected value according to random choices of T_{1_n}, T_{2_n} from both sides of the inequality, we obtain that

$$\mathbb{E}[d_{MP}(T_{1_n}, T_{2_n})] \geq \mathbb{E}\left[\sum_{B \in \mathcal{B}} I(B, T_{1_n}, T_{2_n})\right] = \sum_{B \in \mathcal{B}} \mathbb{E}[I(B, T_{1_n}, T_{2_n})]. \tag{11}$$

As the trees $T_{1_n}, T_{2_n} \in R_L^B$ are drawn independently according to some label-invariant model, from Lemma 4 $\mathbb{E}[I(B, T_{1_n}, T_{2_n})] = (9-3)/9 = 2/3$. Consequently, as $|\mathcal{B}| = n(n-1)/6$, by Equation (11) we obtain the lower bound on $\mathbb{E}[d_{MP}(T_{1_n}, T_{2_n})]$, which is at least $n(n-1)/9$. ■

One of the main advantages of MC and MS metrics based on Definition 2 in comparison with RF metric is the fact that relocation of $O(1)$ leaves in compared trees $T_1, T_2 \in R_n$ can change their distance by at most $O(n)$, which is a relatively small value when referring to the maximum distance being $\Theta(n^2)$ for MC and MS metrics (Bogdanowicz and Giaro, 2012, 2013). We can say that a little change in phylogenetic trees cannot cause large (considering the whole space R_n) changes in their distance. Theorems 1 and 3 show that the mentioned statements are valid also for MP. It is worth mentioning that the expected MP distance between random trees is $\Theta(n^2)$, so the value is $\Omega(n)$ times more than the possible MP distance change, this being the result of a relocation of a bounded number of leaves. Considering this property further, we can notice that the MP metric behaves more naturally than the MC metric or the MS metric, for which the expected value grows asymptotically slower than the diameter $\Theta(n^{3/2})$ in the uniform model (Bogdanowicz and Giaro, 2012, 2013) and only $O(n \log n)$ in the Yule model.[†] In Table 1 we summarized the main properties of MP metric discussed in this section.

In Supplemental Materials, we present comparisons of distributions of various metrics, that is, RF metric, the Triples distance (TT) (Critchlow et al., 1996), the Nodal Splitted metric with L^2 norm (NS) (Cardona et al., 2010), MC, MP, and the cophenetic metric with L^2 norm (CPH) (Cardona et al., 2013), between 1000 random pairs of trees having 100 leaves generated according to the two most popular models, that is, the Yule model and the uniform model.

[†] Let $S(T)$ for $T \in R_n^B$ be defined as $\sum_{c \in \sigma(T)} |c|$. Note that the value of $S(T)$ is closely related to the Sackin's index $S_{ind}(T)$ used to measure the tree balance (Sackin, 1972; Shao and Sokal, 1990). We have $S(T) = S_{ind}(T) + n$. The $O(n \log n)$ upper bound on the average MC-distance in the Yule model can be obtained using the fact that for trees $T_1, T_2 \in R_n^B$, $d_{MC}(T_1, T_2) \leq S(T_1) + S(T_2)$ (Bogdanowicz and Giaro, 2013) and the fact that $\mathbb{E}_{T \in R_n^B}(S_{ind}(T)) = 2n(\sum_{i=1}^{n} \frac{1}{i} - 1) = \Theta(n \log n)$ (Kirkpatrick and Slatkin, 1993).

5. HEURISTICS FOR ROOTED SUBTREE PRUNE AND REGRAFT DISTANCE: AN APPLICATION OF THE MATCHING PAIR DISTANCE IN DETECTING HORIZONTAL GENE TRANSFER

In this section, we provide a preliminary study of the usefulness of the MP distance in identifying HGT. HGT is a direct transfer of genetic material from one lineage to another (Boc et al., 2010). HGT events lead to SPR transformations (Wu, 2009; Boc et al., 2010), appearing in rooted phylogenetic trees constructed based on respective DNA regions.

There are HGT detection methods present in literature (Wu, 2009; Boc et al., 2010), whose main concept is based on determining the SPR distance between phylogenetic trees T_1 and T_2, that is, the minimum number of SPR (or rooted SPR [rSPR] for rooted trees) operations needed to transform tree T_1 into T_2. In this section, we present two simple heuristic approaches to the problem of computing the rSPR distance, which make use of the polynomially computable comparison metrics.

Unfortunately, computing the rSPR distance is NP-hard (Bordewich and Semple, 2005), so some approximation or heuristic algorithms have to be applied to large instances of the problem. In the next part of this section, we present the results of computational experiments on the effectiveness of heuristics for estimation of the rSPR distance between analyzed trees. The tests have been performed based on randomly generated pairs of trees of various sizes.

For small instances, that is, when the compared trees are rather similar, exact exponential algorithms work reasonably fast. We have tested the SPRDist application (Wu, 2009), which was able to compute all our test cases (each test case consists of 1000 comparisons between randomly generated pair of trees) of size up to 20 leaves, and UltraNet (Chen et al., 2015) able to compute test cases up to 40 leaves. For test cases with 50 leaves UltraNet (which worked much faster than SPRDist) was able to compute only 25 distances in about 5 days on a desktop computer, so we decided to interrupt further computation of the exact rSPR distance. Finally, we manged to compute exact distances for all test cases on 50 leaves using rSPR 1.2.0 software (Whidden et al., 2014) (http://kiwi.cs.dal.ca/Software/RSPR).

All of the trees used in the experiments were generated according to the uniform model. In our experiments, we included most of the known metrics for rooted trees, that is, RF metric, the Triples distance (TT) (Critchlow et al., 1996), the NS metric (Cardona et al., 2010), MC and MP metrics, and the CPH metric (Cardona et al., 2013). It turns out that the MP distance performs better than any other metric in both experiments.

5.1. Single metric algorithm

Here, we present a simple heuristic that uses the value of an arbitrary phylogenetic trees metric as an indicator for consequent rSPR moves. The idea of such type of heuristics has been introduced by Boc et al. (2010). The method presented hereunder, in contrast to that described in Boc et al. (2010), uses only metrics for rooted phylogenetic trees and does not check any evolutionary constraints on possible rSPR moves.

Let $T_a, T_b \in R_L^B$ be the input trees. In the first step, we compute the distance in a selected polynomially computable metric d between each tree T_c of $1 -$ rSPR neighborhood of T_a and the target tree T_b. We call the trees within $1 -$ rSPR neighborhood as *candidate trees*. Let T_c' be an arbitrary chosen candidate tree whose distance in the metric d to T_b is minimal. In the next step, we compute the $1 -$ rSPR neighborhood of T_c' and again choose from it an arbitrary candidate tree of minimal distance to T_b. We then repeat the steps

Table 1. Comparison of Selected Properties of Analyzed Metrics for Binary Trees

Property	*RF*	*MC*	*MP*
Minimal positive distance	1	2	1
Single-rooted NNI operation distance	1	$\geq 2, \leq n-1$	$O(n^2)$
Maximal distance	$n-2$	$\geq \frac{1}{2}n^2 - O(n), \leq \frac{3}{4}n^2 + O(n)$	$\frac{1}{2}n^2 - O(n)$
Average distance (the uniform model)	$\Theta(n)$	$\Theta(n^{3/2})$	$\Theta(n^2)$
Average distance (the Yule model)	$\Theta(n)$	$O(n \log n)$	$\Theta(n^2)$

MC, matching cluster; MP, matching pair; NNI, nearest neighbor interchange; RF, Robinson–Foulds.

Table 2. Performance Comparison of a Single Metric Rooted Subtree Prune and Regraft Heuristic—Average Distance and Relative Performance

	RF		*TT*		*MP*		
N	*y*	*y/rSPR*	*y*	*y/rSPR*	*y*	*y/rSPR*	*rSPR*
10	5.958	1.169	5.552	1.090	**5.458**	1.071	5.103
15	10.499	1.209	9.774	1.125	**9.590**	1.104	8.701
20	15.212	1.230	14.265	1.152	**13.901**	1.122	12.396
30	24.886	1.236	23.972	1.190	**23.098**	1.147	20.152
40	34.617	1.233	34.143	1.216	**32.583**	1.160	28.099
50	44.355	1.226	44.396	1.227	**42.327**	1.170	36.184

By y we denote the average number of rSPR operations obtained using a particular heuristic. SPR, subtree prune and regraft; rSPR, rooted SPR. Bold indicates the minimum value among the analyzed heuristics.

until the tree T_b is reached and the number of steps is an obtained estimation (upper bound) on rSPR distance between T_a and T_b.

Note that, to be useful in such a procedure, a phylogenetic tree metric d should guarantee that for any two trees T_a, $T_b \in R_L^B$, $T_a \neq T_b$, there exists $T \in R_L^B$ obtained from T_a by a single rSPR operation such that $d(T_a, T_b) > d(T, T_b)$. The fact that RF and QT metrics for unrooted trees fulfill the mentioned condition (where unrooted SPR instead of rSPR is considered) has been proved in Bordewich et al. (2009). During our experiments, we found that such a property does not hold for the NS, MC, and CPH metrics, but we did not find any counterexamples for metrics TT and MP, which suggests that both these metrics have this property.[‡]

In Table 2 we present the results, that is, the number of rSPR operations estimated using a particular metric, true rSPR distance (the last column), and the ratio of the estimated distance to the exact value of rSPR metric. Each row contains average values from 1000 test cases. Table 2 in Supplementary Materials presents the fraction of the test cases, in which a particular method gave the best result (note that more than one method can give the best result in a single test case). Based on the data, we can see that the best results for all analyzed sizes of trees have been obtained using the MP metric.

5.2. Guided Robinson–Foulds metric algorithm

Here, we consider a modified version of the mentioned procedure, in which two metrics are used. First, the RF distances between candidate trees and the destined tree are computed. Next, ties in a subset of candidate trees of the minimum RF distance in a particular iteration are resolved using any other metric. Note that any metric (or even only a dissimilarity measure) can be chosen as a tiebreaker because using the RF metric as the based one guarantees that the procedure finishes in a number of steps not greater than the RF distance between compared trees [see footnote on this page and similar analysis for unrooted trees for RF and QT metrics described in Bordewich et al. (2009)].

The results of the experiment are presented in Table 3 and in Supplementary Materials, where we present the fraction of test cases in which a particular method gave the best result. Similarly to the previous experiment, Table 3 contains average values from 1000 test cases for the estimated number of rSPR operations, true rSPR distance, and the ratio of the estimated distance to the exact value of rSPR metric. The MP metric performs better that the other distances in almost all cases.

Based on the results from both experiments, we obtain that the strategy incorporating two metrics (e.g., RF + MP), where the second metric is used as a tiebreaker, is worth further development and can lead to better results than using a single metric. A similar observation has been recently reported by Whidden et al. (2014), where the RF distance is used as a tiebreaker for the SPR supertree method.

[‡] We can formally prove that for $T_a \neq T_b \in R_L^B$, there exists a tree $T \in R_L^B$ obtained from T_a by a single rSPR operation such that $d_{MP}(T, T_b) \leq d_{MP}(T_a, T_b)$ and $d_{RF}(T, T_b) < d_{RF}(T_a, T_b)$. Hence, a modification of the heuristic so that in each step a tree T_c with lexicographically the lowest pair $(d_{MP}(T_c, T_b), d_{RF}(T_c, T_b))$ of candidate trees is chosen assures convergence. However, during our experiments, we did not encounter the need for considering the second element $d_{RF}(T_c, T_b)$ of a pair.

TABLE 3. PERFORMANCE COMPARISON OF A GUIDED ROBINSON–FOULDS METRIC ROOTED SUBTREE PRUNE AND REGRAFT HEURISTIC—AVERAGE DISTANCE AND RELATIVE PERFORMANCE

	RF+CPH		*RF+NS*		*RF+MC*		*RF+TT*		*RF+MP*		
n	*y*	*y/rSPR*	*y*	*y/rSPR*	*y*	*y/rSPR*	*y*	*y/rSPR*	*y*	*y/rSPR*	*rSPR*
10	5.553	1.089	5.473	1.073	5.450	1.068	5.429	1.065	**5.406**	1.060	5.103
15	9.757	1.124	9.622	1.108	9.569	1.102	**9.459**	1.089	9.475	1.091	8.701
20	14.014	1.132	13.849	1.119	13.811	1.115	13.624	1.100	**13.573**	1.096	12.396
30	23.177	1.151	22.890	1.137	22.846	1.135	22.481	1.116	**22.421**	1.114	20.152
40	32.584	1.160	32.128	1.144	32.234	1.148	31.546	1.123	**31.386**	1.117	28.099
50	41.952	1.160	41.409	1.145	41.485	1.147	40.742	1.126	**40.621**	1.123	36.184

By *y* we denote the average number of rSPR operations obtained using a particular heuristic. Bold indicates the minimum value among the analyzed heuristics.

6. CONCLUSION

By using a general matching method (Bogdanowicz, 2008; Bogdanowicz and Giaro, 2012), we defined a new phylogenetic metric that can be used for general (binary and nonbinary) rooted phylogenetic trees and analyzed its main properties (such as minimum positive distance, diameter, and average distance between random trees). The interesting fact is that the MP metric becomes the transfer distance between partitions of the set of unordered leaf pairs determined by compared trees. As the transfer distance has already been found useful in bioinformatics (Konovalov et al., 2005), we believe that the MP distance will also become a valuable supplement to the current set of computational tools for phylogenies.

Among the analyzed properties of the MP distance, at least the following two deserve more attention, the quadratic (with respect to the number of leaves) diameter and the quadratic distance between random trees, and a relatively low, that is, $O(n)$ distance increase after relocation of a bounded number of leaves in comparison with the maximum or average distance. Presented computational experiments show potential area in which the MP metric can be successfully applied, that is, HGT analysis according to the method introduced in Boc et al. (2010). Application of MP metric as a tiebreaker metric in the procedure of searching for SPR supertree (Whidden et al., 2014) might also be interesting and requires further research.

The MP metric (among many others) is implemented in TreeCmp 1.1 application freely available at (http://kaims.eti.pg.gda.pl/~dambo/treecmp/).

ACKNOWLEDGMENT

This project has been partially supported by Narodowe Centrum Nauki under contract DEC-2011/02/A/ST6/00201.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Allen, B.L., and Steel, M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Comb.* 5, 1–15.

Bansal, M.S., Burleigh, J.G., Eulenstein, O., et al. 2010. Robinson-Foulds Supertrees. *Algorithms Mol. Biol.* 5, 18.

Boc, A., Philippe, H., and Makarenkov, V. 2010. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst. Biol.* 59, 195–211.

Böcker, S., Canzar, S., and Klau, G.W. 2013. The generalized Robinson-Foulds metric, 156–169. *In Algorithms in Bioinformatics*. LNCS 8126. Springer, Berlin, Heidelberg.

Bogdanowicz, D. 2008. Comparing phylogenetic trees using a minimum weight perfect matching. Proc. 1st International Conference on Information Technology, 451–454. Gdansk, Poland.

Bogdanowicz, D., and Giaro, K. 2012. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 150–160.

Bogdanowicz, D., and Giaro, K. 2013. On a matching distance between rooted phylogenetic trees. *Int. J. Appl. Math. Comput. Sci.* 23, 669–684.

Bogdanowicz, D., Giaro K., and Wróbel, B. 2012. TreeCmp: Comparison of trees in polynomial time. *Evol. Bioinform.* 8, 475–487.

Bordewich, M., Gascuel, O., Huber, K.T., et al. 2009. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6, 110–117.

Bordewich, M., and Semple, C. 2005. On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Comb.* 8, 409–423.

Cardona, G., Llabrés, M., Rosselló, F., et al. 2010. Nodal distances for rooted phylogenetic trees. *J. Math. Biol.* 61, 253–276.

Cardona, G., Mir, A., Rosselló, F., et al. 2013. Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinformat*. 14, 3.

Charon, I., Denœud, L., Guénoche, A., et al. 2006. Maximum transfer distance between partitions. *J. Classification* 23, 103–121.

Chaudhary, R., Burleigh, J.G., and Fernández-Baca, D. 2013. Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol. Biol.* 8, 28.

Chen, Y.H. 2012. The k partition-distance problem. *J. Comp. Biol.* 19, 404–417.

Chen, Z.-Z., Fan, Y., and Wang, L. 2015. Faster exact computation of rSPR distance. *J. Comb. Optim.* 29, 605–635.

Colbourn, C.J., and Dinitz, J.H. 2006. *Handbook of Combinatorial Designs*. Second Edition, Chapman and Hall/CRC. Boca Raton, FL, USA.

Critchlow, D.E., Pearl, D.K., and Qian, C. 1996. The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol.* 45, 323–334.

DasGupta, B., He, X., Jiang, T., et al. 1997. On distances between phylogenetic trees. Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 427–436.

Day, W.H.E. 1981. The complexity of computing metric distances between partitions. *Math. Soc. Sci.* 1, 269–287.

Denœud, L. 2008. Transfer distance between partitions. *Adv. Data Anal. Classification* 2, 279–294.

Estabrook, G.F., McMorris, F.R., and Meacham, C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst. Biol.* 34, 193–200.

Finden, C.R., and Gordon, A.D. 1985. Obtaining common pruned trees. *J. Classification* 2, 255–276.

Gabow, H.N., and Tarjan, R.E. 1989. Faster scaling algorithms for network problems. *SIAM J. Sci. Comput.* 18, 1013–1036.

Gordon, K., Ford, E., and St. John, K. 2013. Hamiltonian Walks of Phylogenetic Treespaces. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 1076–1079.

Guénoche, A. 2011. Consensus of partitions: a constructive approach. *Adv. Data Anal. Classif.* 5, 215–229.

Gusfield, D. 2002. Partition-distance: A problem and class of perfect graphs arising in clustering. *Inf. Process. Lett.* 82, 159–164.

Hickey, G., Dehne, F., Rau-Chaplin, A., et al. 2008. SPR distance computation for unrooted trees. *Evol. Bioinform.* 4, 17–27.

Hillis, D.M., Heath, T.A., and St. John, K. 2005. Analysis and visualization of tree space. *Syst. Biol.* 54, 471–482.

Humphries, P.J., and Wu, T. 2013. On the neighborhoods of trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 721–728.

Kirkman, T.P. 1847. On a problem in combinatorics. *Cambridge Dublin Math. J.* 2, 191–204.

Kirkpatrick, M., and Slatkin, M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47, 1171–118.

Konovalov, D.A., Litow, B., and Bajema, N. 2005. Partition-distance via the assignment problem. *Bioinformatics* 21, 2463–2468.

Koonin, E.V., Puigbò, P., and Wolf, Y.I. 2011. Comparison of phylogenetic trees and search for a central trend in the ‘‘forest of life.’’ *J. Comp. Biol.* 18, 917–924.

Lin, H.T., Burleigh, J.G., and Eulenstein, O. 2009. Triplet supertree heuristics for the tree of life. *BMC Bioinformat.* 10(Suppl 1), S8.

McKenzie, A., and Steel, M. 2000. Distributions of cherries for two models of trees. *Math. Biosci.* 164, 81–92.

Orlin, J.B., and Ahuja, R.K. 1992. New scaling algorithms for the assignment and minimum mean cycle problems. *Math. Program.* 54, 41–56.

Régnier, S. 1965. Sur quelques aspects mathématiques des problèmes de classification automatique. *I.C.C. Bulletin* 4, 175–191. Reprint in *Mathématiques et Sciences Humaines* 82, 13–29, 1983.

Robinson D.F., and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Sackin, M.J. 1972. ‘‘Good’’ and ‘‘bad’’ phenograms. *Syst. Zool.* 21, 225–226.

Semple, C., and Steel, M. 2003. *Phylogenetics*. Oxford University Press. Oxford, United Kingdom.

Shao, K.-T., and Sokal, R.R. 1990. Tree balance. *Syst. Zool.* 39, 266–276.

Smith, K.C., Castro-Nallar, E., Fisher, J.N.B., et al. 2013. Phage cluster relationships identified through single gene analysis. *BMC Genomics* 14, 410.

Steel, M.A., and Penny, D. 1993. Distributions of tree comparison metrics–some new results. *Syst. Biol.* 42, 126–141.

Whidden, C., Zeh, N., and Beiko, R.G. 2014. Supertrees based on the subtree prune-and-regraft distance. *Syst. Biol.* 63, 566–581.

Wu, Y. 2009. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics* 25, 190–196.

Address correspondence to:
Dr. Damian Bogdanowicz
Department of Algorithms and System Modeling
Gdansk University of Technology
Narutowicza 11/12
80-233 Gdansk
Poland

E-mail: damian.bogdanowicz@eti.pg.gda.pl