ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/366655276

Generalization of Phylogenetic Matching Metrics with Experimental Tests of Practical Advantages

Article *in* Journal of Computational Biology: a Journal of Computational Molecular Cell Biology · December 2022

DOI: 10.1089/cmb.2022.0090

CITATIONS
0

READS
57

2 authors:

Damian Bogdanowicz
Gdansk University of Technology
15 PUBLICATIONS **285** CITATIONS
SEE PROFILE

Krzysztof Giaro
Gdansk University of Technology
40 PUBLICATIONS **863** CITATIONS
SEE PROFILE

All content following this page was uploaded by Damian Bogdanowicz on 11 March 2023.

The user has requested enhancement of the downloaded file.

This is the accepted version of the following article:
Bogdanowicz D., Giaro K., Generalization of Phylogenetic Matching Metrics with Experimental Tests of Practical Advantages, JOURNAL OF COMPUTATIONAL BIOLOGY Vol. 30, iss. 3 (2023), pp. 261-276, which has now been formally published in final form at **JOURNAL OF COMPUTATIONAL BIOLOGY** at **https://dx.doi.org/10.1089/cmb.2022.0090**. This original submission version of the article may be used for non-commercial purposes in accordance with the Mary Ann Liebert, Inc., publishers' self-archiving terms and conditions.

Generalizations of Phylogenetic Matching Metrics with Experimental Tests of Practical Advantages

Damian Bogdanowicz,[1*] Krzysztof Giaro,[1]

[1]Department of Algorithms and System Modeling,

Faculty of Electronics, Telecommunications and Informatics,

Gdansk University of Technology,

Narutowicza 11/12, 80-233 Gdansk, Poland

[*]To whom correspondence should be addressed;

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

October 2, 2022

Keywords: comparison of phylogenetic trees, matching metrics, phylogenetic tree distance

Abstract: The ability to quantify a dissimilarity of different phylogenetic trees is required in various types of phylogenetic studies, e.g., such metrics are used to assess the quality of phylogeny construction methods and to define optimization criteria in supertree building algorithms. In this article, starting from the already described concept of matching metrics, we define three new metrics for rooted phylogenetic trees. One of them, MPJ, is still purely topological, but we now utilize the Jaccard index set dissimilarity measure in its

MOST WIEDZY Downloaded from mostwiedzy.pl

construction. This modification substantially changes the structural features of metric space. In particular, we investigate the properties of the previously known MCJ and the new MPJ metrics, such as the asymptotic behavior of their expected distance between two random trees, the space diameter and the change of a distance after a single leaf relocation. The other two metrics, MCW and MCJW, are the first propositions of generalization of matching metrics designed for rooted phylogenies with branch lengths. The experimental tests of the practical utility of the phylogenetic metrics show the superiority of MCJ, MPJ over the previous best tree comparison method. In order to define the MCW and MCJW metrics, we introduce a general method for constructing matching metrics for weighted rooted phylogenetic trees.

1 Introduction

The necessity of defining a dissimilarity measure of phylogenetic trees, i.e., introducing a structure of metric space[1] in the set of phylogenetic trees, appears in various types of phylogenetic studies. For example, in the simulation of phylogeny-inference algorithms, we analyze how close to the true tree are the trees recovered by different algorithms. Quantifying the similarities between phylogenies is also useful in the analysis and visualization of a group of phylogenetic trees (Hillis et al., 2005). Furthermore, phylogenetic tree metrics are often used to define optimization criteria in supertree building algorithms, see e.g. Bansal et al. (2010); Whidden et al. (2014).

Although there are many various phylogenetic metrics known in the liter-

[1] A metric space (X, d) is a pair consisting of the set X and the function $d : X \times X \to \mathbb{R}$ (the *metric over* X) such that (i) $\forall_{x,y \in X} d(x, y) = 0 \Leftrightarrow x = y$, (ii) $\forall_{x,y \in X} d(x, y) = d(y, x)$, (iii) $\forall_{x,y,z \in X} d(x, y) + d(y, z) \geq d(x, z)$ – the *triangle inequality*.

MOST WIEDZY Downloaded from mostwiedzy.pl

ature, it is difficult to state unambiguously which one is the best. The comparative tests described in the literature focus mainly on the tests' suitability for particular applications (e.g. the ones mentioned above), which substantially impacts the obtained rankings of their usefulness. The noticeable exception is a recent work (Kuhner and Yamato, 2015) analyzing the practical properties of some currently available metrics. In particular, the authors of Kuhner and Yamato (2015) designed two interesting experiments allowing for the comparison of metric properties for binary rooted phylogenetic trees according to their phylogenetic reliability and practical usefulness. They are designed to examine the general, application-independent expectations of dissimilarity tree measure behavior during gradual, successive modification or when inferring trees on the basis of poorer biological data.

The task in the first experiment, called "n-away", distinguished which of two trees is separated by a smaller recombinational distance from the same tree. The best performance in this experiment was achieved by the *Alignment* metric, which was the distance defined on the basis of a similarity measure proposed by Nye et al. in Nye et al. (2006).

The second experiment, called "bullseye", tested the ability to distinguish trees inferred with a lower versus higher quality of input data. In this experiment, branch-length versions of the Robinson–Foulds metric performed best.

Here, we want to investigate how to define metrics to achieve an even better performance in at least one of the above-mentioned tests. In Bogdanowicz and Giaro (2012, 2013, 2017) we defined the general framework for defining phylogenetic metrics for rooted and unrooted trees, and argued that a customization of an element of this template leads us to metrics with properties better than the currently known methods. In particular, any change of function h, (see Definition 2) which is an internal element of this template, results in a new

MOST WIEDZY Downloaded from mostwiedzy.pl

dissimilarity measure with completely different properties.

In our earlier works Bogdanowicz and Giaro (2012, 2013, 2017), we used, to put it in simple terms, dissimilarity measures based on a simple size of sets' symmetric difference to compare variously defined internal elements of tree description. This is a common approach, encountered in the case of classic metrics (e.g. RF Robinson and Foulds (1981), Triples Critchlow et al. (1996), Quartet Estabrook et al. (1985), MAST Finden and Gordon (1985); Goddard et al. (1994) metrics), as it seems natural to assume that a comparison of large, partially overlapping clades should result in a larger difference value than in the case of small ones. The described experimental results unexpectedly support the opposite hypothesis. In this paper, we establish new metrics based on the Jaccard index, which evaluates the inconsistency of species groups "relative" to their sizes. This change, along with the appropriate use of the framework allows obtaining metrics that behave as well as or even outperform the metrics evaluated in Kuhner and Yamato (2015).

It is obvious that apart from the shape of the tree, the lengths of its branches contain valuable phylogenetic information. In particular, in Kuhner and Yamato (2015), metrics that took into account branch lengths performed better in the "bullseye" experiment than the purely topological ones. Therefore, we introduce a generalization of the earlier matching metrics paradigm (Definition 13) to formulate branch-length-aware matching metrics. Furthermore, using that approach, we defined weight versions of the MC and MCJ metrics, namely MCW and MCJW. In one of the experiments, the MCJW metric performed almost as well as the best in this case weight versions of RF distance.

MOST WIEDZY Downloaded from mostwiedzy.pl

2 Basic definitions and notation

For sets A, B let $A \oplus B = (A \setminus B) \cup (B \setminus A)$ be their symmetric difference, $|A|$ denotes the cardinality of set A. By $\mathcal{P}(A)$ we denote the family of all subsets of A (i.e. the *power set of* A), and $\mathcal{F}(A, B)$ is the set of all functions $f : A \to B$. Moreover, if B is a numbers' set $B \subseteq \mathbb{R}$, then the support of a function $f \in \mathcal{F}(A, B)$ consists of arguments with nonzero values $supp(f) = \{x \in A : f(x) \neq 0\} = f^{-1}(\mathbb{R} \setminus \{0\})$ and $\mathcal{F}_{fin}(A, B) \subseteq \mathcal{F}(A, B)$ contains functions with $|supp(f)| < \infty$. Similarly, $\mathcal{P}_{fin}(A)$ contains finite subsets, i.e., $\mathcal{P}_{fin}(A) = \{X \in \mathcal{P}(A) : |X| < \infty\}$ (in particular, $\mathcal{P}_{fin}(A) = \mathcal{P}(A)$ if A is finite). For finite sets A, B with $A \cup B \neq \emptyset$, the Jaccard distance $JC(A, B)$ is defined as $JC(A, B) = 1 - |A \cap B|/|A \cup B| = |A \oplus B|/|A \cup B| \in [0, 1]$ (moreover, $JC(\emptyset, \emptyset) = 0$). It is commonly known that functions $(A, B) \to |A \oplus B|$ and $(A, B) \to JC(A, B)$ introduce the metric space in every family of finite sets (see e.g., Kosub (2019) for the case of the Jaccard distance).

Let $G = (V, E)$ be a *graph* with a set of vertices V and a set of edges E. A *bipartite graph* $G(V_1 \cup V_2, E)$ 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 *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).

MOST WIEDZY Downloaded from mostwiedzy.pl

A *tree* is a connected acyclic graph. A graph $T = (V, E)$ is a *rooted phylogenetic tree* if it is a tree whose *leaves*, that is, vertices (nodes) of degree one, are labeled bijectively by the elements of a finite set L (representing the species), there is exactly one distinguished non-leaf vertex $r(T) \in V$ called the *root* and none of the vertices of $V \setminus \{r(T)\}$ has degree two. For the sake of simplicity, we can identify the leaves with their labels, i.e., for a phylogenetic tree T by $L(T) \subseteq V$, we denote the set of leaves of T or the set of labels of those leaves. The phylogenetic interpretation is as follows: present-day species under examination form the finite set L and are represented by the 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.

A *rooted binary phylogenetic tree* is a rooted phylogenetic tree such that the root has degree two and all other internal vertices have degree three. 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. A rooted tree T defines a partial order relation of being a descendant (and ancestor) on its vertices, for $a, b \in V(T)$ 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. We can assign to every vertex v 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 $c(r(T)) = L(T)$), and all other clusters are *non-trivial*. By $\sigma(T)$ and $\sigma_*(T)$ we denote the families of all clusters of T and all non-trivial clusters of T, respectively. Hence $|\sigma_*(T)| \leq |L(T)| - 2$ and the equality holds exactly 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 (see Semple and Steel (2003) section 3.5). By $L^{(2)}$ we denote the set of all unordered pairs of leaves, i.e., $L^{(2)} = \{\{x, y\} : x, y \in L, x \neq y\}$ and

MOST WIEDZY Downloaded from mostwiedzy.pl

$|L^{(2)}| = |L|(|L|-1)/2$.

The *lowest common ancestor* (LCA), also called the *most recent common ancestor*, of a pair of leaves u, v of a rooted tree T, $lca(u, v)$ 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 $c^{(2)}(v)$ for which v is the lowest common ancestor, i.e., $c^{(2)}(v) = \{\{x, y\} \in L^{(2)} : lca(x, y) = v\}$. We will call the set $c^{(2)}(v)$ the *pair set* of v. By $\sigma_*^{(2)}(T)$ we denote the family of all pair sets of T, so $\sigma_*^{(2)}(T)$ is a partition of the set $L^{(2)}$ into the non-empty disjoint sets determined by T. Note that a rooted phylogenetic tree T is uniquely described by a set $\sigma_*^{(2)}(T)$ because $\sigma_*^{(2)}(T)$ determines $\sigma_*(T)$ and we have $c(v) = \bigcup_{z \in c^{(2)}(v)} z$ for $v \in V \setminus L$.

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*

$$\begin{aligned} d_{RF}(T_1, T_2) &= \frac{1}{2}|\sigma(T_1) \oplus \sigma(T_2)| \\ &= \frac{1}{2}|\sigma_*(T_1) \oplus \sigma_*(T_2)|. \end{aligned} \tag{1}$$

3 Matching metrics

We recall the general construction of matching metrics presented in Bogdanowicz and Giaro (2012).

Definition 2. *There are given a set D, an element $O \notin D$ and a metric h on $D \cup \{O\}$. We define a metric $d_h : \mathcal{P}_{fin}(D) \times \mathcal{P}_{fin}(D) \to \mathbb{R}_{\geq 0}$, where the*

MOST WIEDZY Downloaded from mostwiedzy.pl

distance between $A, B \in \mathcal{P}_{fin}(D)$ $d_h(A, B)$ *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 an 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 vertex 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 1 (Bogdanowicz and Giaro (2012))**.** *The function* d_h *is a metric and the value of* $d_h(A, B)$ *does not depend on* s *and* t *(when* $s - t = |A| - |B|$*).*

In particular for $|A| = |B|$, we may assume $s = t = 0$ and $|V_1| = |V_2| = |A|$. Moreover, in Bogdanowicz and Giaro (2012), we showed that

$$d_h(A, B) = d_h(A \setminus B, B \setminus A) \tag{2}$$

as every matching consisting of edges connecting vertices with equal labels can be extended to a minimum-weight perfect matching in graph $G(V_1, V_2, E)$ from Definition 2.

MOST WIEDZY Downloaded from mostwiedzy.pl

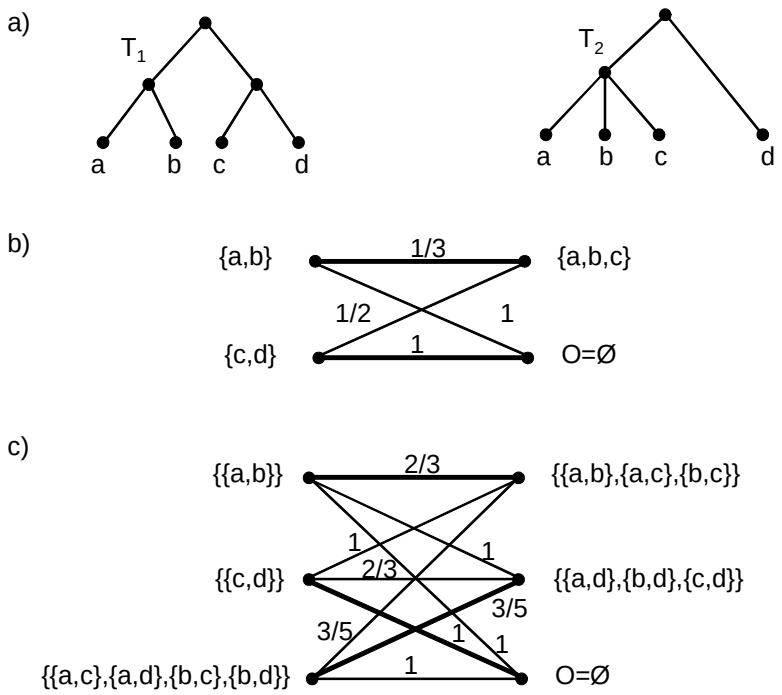

Figure 1: Computation of matching distances: a) example of trees, b) computation of MCJ distance, $d_{MCJ}(T_1, T_2) = 4/3 \approx 1.34$, c) computation of MPJ distance, $d_{MPJ}(T_1, T_2) = 34/15 \approx 2.27$. Values of some classic phylogenetic metrics (see Table 1) for these trees: $d_{RF}(T_1, T_2) = 1.5$, $d_{TT}(T_1, T_2) = 3$, $d_{NS}(T_1, T_2) \approx 2.65$, $d_{CPH}(T_1, T_2) = 2$.

In our earlier works Bogdanowicz and Giaro (2012, 2013, 2017), we considered features of only integer-valued (or half-integer) phylogenetic metrics for rooted (Matching Cluster and Matching Pair distances, i.e., MC, MP) or unrooted (Matching Split distance, MS) trees based on Definition 2, where the h function used a cardinality of symmetric difference of some sets describing trees.

However, we will see that a metric with normalized (Jaccard) values gives better experimental results.

Here, we recall a definition of some variant of matching metrics proposed in Böcker et al. (2013) and called the Jaccard-Robinson-Foulds metric there. In

MOST WIEDZY Downloaded from mostwiedzy.pl

our opinion, this method is more similar to the Matching Cluster distance, so further in the article we will be referring to it as the Matching Cluster Jaccard distance.

Definition 3. *Let $T_1, T_2 \in R_L$ be rooted phylogenetic trees, $D = \mathcal{P}(L) \setminus \{\emptyset\}$, and $O = \emptyset$, $h_{JC} : \mathcal{P}(L) \times \mathcal{P}(L) \to \mathbb{R}_{\geq 0}$ be such that $h_{JC}(A, B) = JC(A, B)$. According to Definition 2 we define the Matching Cluster Jaccard (MCJ) distance $d_{MCJ} : R_L \times R_L \to \mathbb{R}_{\geq 0}$ as*

$$\begin{aligned} d_{MCJ}(T_1, T_2) &= d_{h_{JC}}(\sigma(T_1), \sigma(T_2)) \\ &= d_{h_{JC}}(\sigma_*(T_1), \sigma_*(T_2)). \end{aligned} \tag{3}$$

The second equality follows from (2), since all trivial clusters belong to $\sigma(T_1) \cap \sigma(T_2)$. Comparing this to Matching Cluster distance of Bogdanowicz and Giaro (2013) note, that the function $h_C(A, B) = |A \oplus B|$ was applied in the original MC definition instead of h_{JC}.

Definition 4. *Let $T_1, T_2 \in R_L$ be rooted phylogenetic trees, $D = \mathcal{P}(L^{(2)}) \setminus \{\emptyset\}$, and $O = \emptyset$, $h_{JP} : \mathcal{P}(L^{(2)}) \times \mathcal{P}(L^{(2)}) \to \mathbb{R}_{\geq 0}$ be such that $h_{JP}(A, B) = JC(A, B)$. According to Definition 2, we define the Matching Pair Jaccard (MPJ) distance $d_{MPJ} : R_L \times R_L \to \mathbb{R}_{\geq 0}$ as*

$$d_{MPJ}(T_1, T_2) = d_{h_{JP}}(\sigma_*^{(2)}(T_1), \sigma_*^{(2)}(T_2)). \tag{4}$$

In the definition of Matching Pair distance in Bogdanowicz and Giaro (2017) $h_P(A, B) = |A \oplus B|/2$ appeared originally in place of h_{JP}. For a graphic illustration of the MCJ and MPJ metrics computation, see Fig. 1.

MOST WIEDZY Downloaded from mostwiedzy.pl

Now we investigate the asymptotic behavior of the expected distance between two random binary trees from R_L^B in the MCJ and MPJ metrics under the most popular models of phylogenetic tree generation. In the uniform model, all binary phylogenetic trees in R_L^B are equally likely. Another very popular (considered as more biologically plausible) model of phylogenetic tree generations is the Yule model (see McKenzie and Steel (2000)), where trees are constructed in iterative discrete random process:

1. Choose uniformly at random two different taxa $v_1, v_2 \in L$, create the only rooted tree with leaf set $\{v_1, v_2\}$,
2. For $i = 3, \ldots, |L|$ repeat:
   - Select randomly the already used species $u \in \{v_1, \ldots, v_{i-1}\}$ and unused $v_i \in L \setminus \{v_1, \ldots, v_{i-1}\}$. Both choices should follow uniform distributions.
   - Extend the tree by attaching the leaf v_i using a new edge to the "middle point" of the pendant edge connected with u.

Lemma 2. *For sets $\emptyset \neq A \neq B$, $|B| \in \{1, 2\}$, we have $JC(A, B) \geq 1/3$.*

Proof. If $A \subseteq B$, then $|A| = 1$, $|B| = 2$ and $JC(A, B) = 1/2$. Otherwise, $JC(A, B) = 1 - |A \cap B|/|A \cup B| \geq 1 - |B|/(|B| + 1) \geq 1/3$. □

Theorem 1. *The diameters of metric spaces R_L, $n = |L|$ with d_{MCJ} and d_{MPJ} fulfill*

$$\max_{T_{1_n}, T_{2_n} \in R_L} d_{MCJ}(T_{1_n}, T_{2_n}) = \Theta(n), \tag{5}$$

$$\max_{T_{1_n}, T_{2_n} \in R_L} d_{MPJ}(T_{1_n}, T_{2_n}) = \Theta(n). \tag{6}$$

Moreover, for rooted trees $T_{1_n}, T_{2_n} \in R_L^B$, $n = |L|$ generated independently at random according to the Yule model or the uniform model, their expected distances are

MOST WIEDZY Downloaded from mostwiedzy.pl

$$\mathbb{E}_{T_{1_n},T_{2_n}}[d_{MCJ}(T_{1_n},T_{2_n})] = \Theta(n), \tag{7}$$

$$\mathbb{E}_{T_{1_n},T_{2_n}}[d_{MPJ}(T_{1_n},T_{2_n})] = \Theta(n). \tag{8}$$

Proof. The expected distance is obviously upper bounded by the diameter, which is $O(n)$ as $JC \leq 1$, hence it is enough to show the lower bound $\Omega(n)$ part of (7) and (8). As was shown in Steel and Penny (1993), the expected RF-distance of random trees tends to the RF-diameter of R_L^B, $|L| = n$ in both the Yule and uniform models, i.e., $\lim_{n\to\infty} \mathbb{E}[d_{RF}(T_{1_n},T_{2_n})]/(n-2) = 1$. But $d_{RF}(T_{1_n},T_{2_n})] = (|\sigma_*(T_{1_n})| + |\sigma_*(T_{2_n})|)/2 - |\sigma_*(T_{1_n}) \cap \sigma_*(T_{2_n})| = n - 2 - |\sigma_*(T_{1_n}) \cap \sigma_*(T_{2_n})|$, hence the number of common non-trivial clusters in both trees has expected value $o(n)$:

$$\lim_{n\to\infty} \mathbb{E}_{T_{1_n},T_{2_n}}[|\sigma_*(T_{1_n}) \cap \sigma_*(T_{2_n})|]/n = 0. \tag{9}$$

A cherry in tree $T \in R_L^B$ is a vertex v with only two descendant leaves, equivalently $|c(v)| = 2$ and $|c^{(2)}(v)| = 1$. According to McKenzie and Steel (2000), the expected number of cherries in the T_{1_n} is $n/3 \pm O(1)$ in the Yule model and $n/4 \pm O(1)$ in a uniform model. So the expected number of cherries v in T_{1_n}, which are not present in T_{2_n}, i.e., with $c(v) \notin \sigma(T_{2_n})$ due to (9) is greater than $n/5$ for big enough n in both random tree models. According to Lemma 2, for each such cherry v we have $\forall_{c\in\sigma(T_{2_n})} JC(c(v),c) \geq 1/3$, and $\forall_{p\in\sigma_*^{(2)}(T_{2_n})} JC(c^{(2)}(v),p) \geq 1/3$. Considering the bipartite graph $G = (V_1, V_2, E)$ from definition 2 used for evaluation of $d_{MCJ}(T_{1_n},T_{2_n})$ or $d_{MPJ}(T_{1_n},T_{2_n})$ (we assume $s = t = 0$ here), we have that the weight of every edge in G incident with the vertex corresponding to such cherry v is not smaller than $1/3$. Summing up, for big enough n the expected number of vertices in G with weights of all edges incident to them equal to $1/3$ or more is not

MOST WIEDZY Downloaded from mostwiedzy.pl

smaller than $n/5$, so the expected weight of a minimum weight perfect matching is at least $n/15$. □

Now consider a binary tree $T_1 \in R_L^B$ and its "small modification" $T_2 \in R_L^B$ obtained by a *single leaf $x \in L$ relocation*, i.e., a transformation consisting of two steps (it is a special case of a more general *rooted subtree prune and regraft operation*, see Semple and Steel (2003)):

1. *Pruning of x*: let the edge $\{x, u\} \in E(T_1)$. If $u = r(T_1)$, make the second neighbor of u the new root and delete u. Otherwise, delete the edge $\{x, u\}$ and suppress the degree-two vertex u. So, we obtained two separate components: x and the rooted tree T_1'.

2. *Regrafting x*: subdivide an arbitrary edge e of T_1' putting a new vertex v on it and reconnect x with v, or connect the new vertex v with x and $r(T_1')$ and make v the new root.

Theorem 2. *Let $T, T_1, T_2 \in R_L^B$, $n = |L|$ and tree T_2 be a tree created from T_1 by relocation of a single leaf $x \in L$, then the following relations hold:*

$$|d_{MCJ}(T, T_1) - d_{MCJ}(T, T_2)| = O(\ln(n)),$$

$$|d_{MPJ}(T, T_1) - d_{MPJ}(T, T_2)| = O(\ln(n)).$$

Proof. Both d_{MCJ} and d_{MPJ} as metrics fulfill the triangle inequality, so it is enough to show $d_{MCJ}(T_1, T_2) = O(\ln(n))$, $d_{MPJ}(T_1, T_2) = O(\ln(n))$ and a leaf relocation may be considered as at most two subsequent "leaf move up or down" operations defined in Fig. 2. Therefore, we can assume that T_1, T_2 are related as in Fig. 2.

MOST WIEDZY Downloaded from mostwiedzy.pl

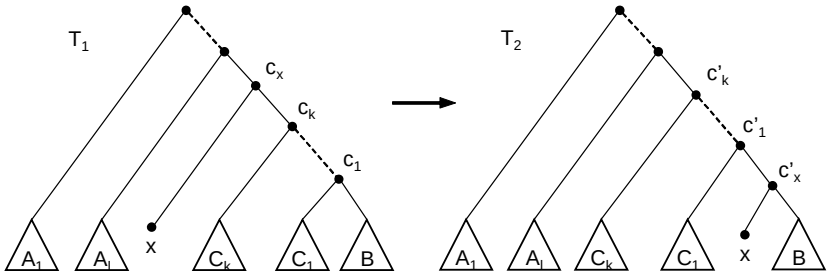

Figure 2: Relocation of a single leaf x. Capital letters denote leaf sets in the respective subtrees. It is possible that A_i-leaf subtrees are empty, B-leaf and C_i-leaf subtrees are non-empty, $k \geq 1$ and we have $c_x = \bigcup_{i=1}^{k} C_i \cup B \cup \{x\}$, $c_i = \bigcup_{j=1}^{i} C_j \cup B$, $c'_x = B \cup \{x\}$, $c'_i = \bigcup_{j=1}^{i} C_j \cup B \cup \{x\}$.

For d_{MCJ} we have $d_{MCJ}(T_1, T_2) \leq JC(c_x, c'_k) + JC(c_k, c'_x) + \sum_{i=1}^{k-1} JC(c_i, c'_i) \leq 1 + \sum_{i=1}^{k-1} 1/|c'_i|$. Note that $3 \leq |c'_i| < |c'_j| \leq n$ for $1 \leq i < j \leq k$. Since for partial sums of the harmonic series, we have $\sum_{i=1}^{k} 1/i \leq 1 + \ln(k)$, then $d_{MCJ}(T_1, T_2) \leq 1 + \sum_{i=3}^{n-1} 1/i < 1/2 + \ln(n-1)$.

Let $c_i^{(2)} = c^{(2)}(v)$ for the only vertex $v \in V(T_1)$ with $c(v) = c_i$, and we use a similar notation $c_x^{(2)}$ (and $c_i'^{(2)}$, $c_x'^{(2)}$ for T_2). In the case of d_{MPJ}, we have $d_{MPJ}(T_1, T_2) \leq JC(c_x^{(2)}, c_k'^{(2)}) + JC(c_k^{(2)}, c_x'^{(2)}) + \sum_{i=1}^{k-1} JC(c_i^{(2)}, c_i'^{(2)}) \leq 2 + \sum_{i=1}^{k-1} JC(c_i^{(2)}, c_i'^{(2)})$. Note that $JC(c_i^{(2)}, c_i'^{(2)}) = 1/|c'_{i-1}|$ for $1 < i < k$ and $JC(c_1^{(2)}, c_1'^{(2)}) = 1/|c'_x|$. Since $2 \leq |c'_x| < |c'_i| < |c'_j| \leq n$ for $1 \leq i < j \leq k$, then $d_{MPJ}(T_1, T_2) \leq 2 + 1/|c'_x| + \sum_{i=2}^{k-1} 1/|c'_{i-1}| \leq 2 + \ln(n-1)$. □

One of the main disadvantages of RF is its overestimation of small changes in tree topology (Bogdanowicz and Giaro, 2012). In the case of RF, a displacement of only one leaf may create a rooted tree distanced from the original by as much as $|L| - 2$, which is the maximum possible distance in this metric. Despite the minor change, these trees seem to be very distant. Metrics MCJ and MPJ are not misleading in these situations. Conducting a fixed number $k = const$ of leaf displacements may create a tree distanced by $O(\ln |L|)$, but the spaces diameters are $\Theta(|L|)$. Such a feature is a quite general property of matching

MOST WIEDZY Downloaded from mostwiedzy.pl

metrics, see analogical facts for the MS, MC and MP metrics (Bogdanowicz and Giaro, 2012, 2013, 2017).

4 Metrics for trees with edge lengths

Definition 5. *A weighted rooted phylogenetic tree over the set of leaves L is a pair $(T(V,E),l)$, where $T(V,E) \in R_L$ and l is a function $l : E \to \mathbb{R}_{>0}$.*

We denote by R_L^w the set of all weighted rooted phylogenetic trees over the set of leaves L. There is a relatively small number of popular metrics for weighted trees. Metrics that compare only the topology of the trees do not usually have a natural extension to the weighted case.

The exceptions are splitted nodal distances (Cardona et al., 2010), where two vectors containing the distances between all pairs u, v and their $lca(u,v)$ are compared using L^1 or L^2 norms and cophenetic metrics (Cardona et al., 2013), where the same norms are used to compare the *cophenetic vectors* of the trees, i.e., vectors consisting of depths of the lowest common ancestors of all pairs of taxa and the depths of all taxa. However, it is hard to find an intuitive phylogenetic interpretation of a metric defined in such a manner. Another popular method for comparing trees with weights is the weighted RF distance (Robinson and Foulds, 1979).

The weighted tree $(T,l) \in R_L^w$ can be represented by the length of the edges $\sigma_w(T,l) : \mathcal{P}(L) \setminus \{\emptyset\} \to \mathbb{R}_{\geq 0}$ such that $supp(\sigma_w(T,l)) = \sigma(T) \setminus \{L\}$ and $\sigma_w(T,l)(c) = l(e)$ for an edge e going up in T from a vertex with cluster c. The weighted variant of RF that is described in Robinson and Foulds (1979) concerns unrooted trees. Since in this study we are interested in rooted phylogenies, we define an analog of the RF metric for trees from R_L^w. For $(T_1,l_1),(T_2,l_2) \in R_L^w$,

MOST WIEDZY Downloaded from mostwiedzy.pl

the weighted RF metric is defined by

$$d^w_{RF} = \frac{1}{2} \sum_{c \in \mathcal{P}(L) \setminus \{\emptyset\}} |\sigma_w(T_1, l_1)(c) - \sigma_w(T_2, l_2)(c)|. \tag{10}$$

An interesting representation of R^w_L as a topological space of Euclidean regions (corresponding to different binary tree shapes) glued on boundaries leads to the definition of *geodesic distance* d_{Geo} (Billera et al., 2001) on R^w_L. The idea of this metric is elegant and natural. However, the efficient algorithm for computing it had been unknown for many years and was finally invented in Owen and Provan (2011).

In this section, we propose a method of introducing a metric space in R^w_L, which can be regarded as continuous analogs of definition 2. Note that comparing weighted trees can be reduced to defining a metric in a set of functions σ_w perceived as a representation of the weighted trees.

For numbers x, y, we define $inc(x, y) = \max\{0, x - y\}$, so at least one of $inc(x, y)$ and $inc(y, x)$ equals 0. Following notions of D, O, h from Definition 2 at first we propose an auxiliary distance measure h' for elements $d_1, d_2 \in D$ taken with some numerical "sizes" $r_1, r_2 > 0$. After that we will incorporate h' to the Definition 2, obtaining a metric for finite support numerical functions of the domain D. The idea of the first step is as follows: distance $h(d_1, d_2)$ should be taken with the coeficient appropriate to the common "amounts" of both elements, i.e., $\min\{r_1, r_2\}$, plus the remaining $(\max\{r_1, r_2\} - \min\{r_1, r_2\})$-size part of the "bigger" one among d_1, d_2 should be compared with O. This informal concept may be specified by the symmetric distance measure $h' : D' \times D' \to \mathbb{R}_{\geq 0}$

MOST WIEDZY Downloaded from mostwiedzy.pl

defined for $D' = (D \times \mathbb{R}_{>0}) \cup \{O\}$ by the following formulas: $h'(O,O) = 0$ and

$$\begin{aligned} h'((d_1,r_1),(d_2,r_2)) = \quad & \min\{r_1,r_2\}h(d_1,d_2) \\ & + inc(r_1,r_2)h(d_1,O) \\ & + inc(r_2,r_1)h(d_2,O), \end{aligned} \tag{11}$$

$$h'((d,r),O) = h'(O,(d,r)) = r \cdot h(d,O). \tag{12}$$

Lemma 3. *If h is a metric on $D \cup \{O\}$, then h' fulfils metric axioms on $D' = (D \times \mathbb{R}_{>0}) \cup \{O\}$.*

Proof. The only non-trivial part is the triangle inequality. In order to reduce the number of subcases under consideration it is convenient to examine the modified measure h'' defined by the right side of (11) only, but over $(D \cup \{O\}) \times \mathbb{R}_{>0}$ (i.e. we temporarily[2] accept arguments of the form (O, r_O), $r_O > 0$), as $h'((d,r),O)$ given by (12) equals to $h''((d,r),(O,r_O))$ from (11) with any $r_O > 0$. So, we consider three pairs $p_1 = (d_1,r_1)$, $p_2 = (d_2,r_2)$, $p_3 = (d_3,r_3)$, where $d_i \in D \cup \{O\}$, $r_i > 0$ and let $\overline{r}_1 \leq \overline{r}_2 \leq \overline{r}_3$ correspond to r_1, r_2, r_3 sorted in non-decreasing order. Moreover, to reformulate (11) we also need

$$d_{i,s} = \begin{cases} d_i & \text{if } \ r_i \leq \overline{r}_s \\ O & \text{otherwise} \end{cases}$$

for $i, s = 1, 2, 3$. Now it is enough to observe that the distance (11) between

[2]We show here, that h'' is a pseudometric over $(D \cup \{O\}) \times \mathbb{R}_{>0}$. It gives $h''((O,x),(O,y)) = 0$ for all $x, y > 0$, and it is easy to see, that this is the only case, where different elements have zero h''-distance. The metric h' is obtained form h'' by gluing all such pairs (O,x) into a single point O.

MOST WIEDZY Downloaded from mostwiedzy.pl

pairs p_i, p_j $(i, j = 1, 2, 3)$ can be expressed in the form

$$h''((d_i, r_i), (d_j, r_j)) = \overline{r}_1 \cdot h(d_{i,1}, d_{j,1}) + (\overline{r}_2 - \overline{r}_1) \cdot h(d_{i,2}, d_{j,2})$$

$$+(\overline{r}_3 - \overline{r}_2) \cdot h(d_{i,3}, d_{j,3})$$

and the triangle inequality follows from the same axiom for h. □

Definition 6. *Given are (not necessarily finite) a set D, an element $O \notin D$ and a metric h on $D \cup \{O\}$. For the given functions $f, g \in \mathcal{F}_{fin}(D, \mathbb{R}_{\geq 0})$, let $A = supp(f) = \{a_1, \ldots, a_{|A|}\}$, $B = supp(g) = \{b_1, \ldots, b_{|B|}\}$, and s, t be arbitrary numbers such that $s - t = |A| - |B|$. We construct a complete bipartite graph $G(V_1, V_2, E)$ with $|V_1| = |V_2|$ defined as follows:*

- *the sets of vertices are*

  $V_1 = \{v_1, \ldots, v_{|A|}, v_{|A|+1}, \ldots, v_{|A|+t}\}$,

  $V_2 = \{u_1, \ldots, u_{|B|}, u_{|B|+1}, \ldots, u_{|B|+s}\}$,

- *the weights of the edges are defined as*

$$w(\{v_i, u_j\}) = \begin{cases} \min\{f(a_i), g(b_j)\}h(a_i, b_j) & \\ +inc(f(a_i), g(b_j))h(a_i, O) & \text{if } i \leq |A|, j \leq |B| \\ +inc(g(b_j), f(a_i))h(b_j, O) & \\ f(a_i)h(a_i, O) & \text{if } i \leq |A|, j > |B| \\ g(b_j)h(b_j, O) & \text{if } i > |A|, j \leq |B| \\ 0 & \text{if } i > |A|, j > |B| \end{cases} \tag{13}$$

Now we define $\bar{d}_h(f, g)$ as the weight of a minimum-weight perfect matching in G.

Lemma 4. *The value of $\bar{d}_h(f, g)$ does not depend on the numbers s and t (while fulfilling the above condition) and is a metric on $\mathcal{F}_{fin}(D, \mathbb{R}_{\geq 0})$.*

MOST WIEDZY Downloaded from mostwiedzy.pl

Proof. We can injectively encode any finite support numerical function $f \in \mathcal{F}_{fin}(D, \mathbb{R}_{\geq 0})$ by a set of all its non-zero valued argument-value pairs $\kappa(f) = \{(x, f(x)) : f(x) \neq 0\}$. Then, however, $\bar{d}_h(f, g)$ appears to be a metric specified according to the definition 2 with the use of h', i.e., (11), (12) since we have $\bar{d}_h(f, g) = d_{h'}(\kappa(f), \kappa(g))$. □

Note that if the functions f, g are $\{0, 1\}$-valued, then the construction from definition 6 agrees with definition 2, hence $\bar{d}_h$ can also be regarded as an extension of the metric d_h.

Lemma 5. *If $A, B \in \mathcal{P}_{fin}(D)$, then $d_h(A, B) = \bar{d}_h(\delta_A, \delta_B)$, where indicator functions $\delta_A, \delta_B : D \to \{0, 1\}$ fulfill $\delta_A^{-1}(1) = A$, $\delta_B^{-1}(1) = B$.*

For example, d^w_{RF} is a special case of $\bar{d}_h$: we take $D = \mathcal{P}(L) \setminus \{\emptyset\}$, $O = \emptyset$; for non-empty sets A, B, a metric h_{RF} fulfills $h_{RF}(A, B) = 1 \Leftrightarrow A \neq B$ and $h_{RF}(A, \emptyset) = 0.5$, then for $(T_1, l_1), (T_2, l_2) \in R^w_L$ we have

$$d^w_{RF}((T_1, l_1), (T_2, l_2)) = \bar{d}_{h_{RF}}(\sigma_w(T_1, l_1), \sigma_w(T_2, l_2)).$$

Similarly, we define a weighted generalizations of the Matching Cluster distance (see Bogdanowicz and Giaro (2013)) and d_{MCJ}.

Definition 7. *Let $(T_1, l_1), (T_2, l_2) \in R^w_L$, and $D = \mathcal{P}(L) \setminus \{\emptyset\}$, $O = \emptyset$, $h_C : \mathcal{P}(L) \times \mathcal{P}(L) \to \mathbb{Z}_{\geq 0}$ be such that $h_C(A, B) = |A \oplus B|$. Then $d^w_{MC} : R^w_L \times R^w_L \to R_{\geq 0}$ is a metric MCW defined as*

$$d^w_{MC}((T_1, l_1), (T_2, l_2)) = \bar{d}_{h_C}(\sigma_w(T_1, l_1), \sigma_w(T_2, l_2)).$$

Definition 8. *Let $(T_1, l_1), (T_2, l_2) \in R^w_L$ and D, O, h_{JC} be like in definition 3.*

MOST WIEDZY Downloaded from mostwiedzy.pl

Then $d^w_{MCJ} : R^w_L \times R^w_L \to R_{\geq 0}$ *is a metric MCJW defined as*

$$d^w_{MCJ}((T_1, l_1), (T_2, l_2)) = \bar{d}_{h_{JC}}(\sigma_w(T_1, l_1), \sigma_w(T_2, l_2)).$$

If we treat trees of R_L as weighted trees with unit weights, then according to Lemma 5, d^w_{MCJ} is an extension of d_{MCJ}:

$$d_{MCJ}(T_1, T_2) = d^w_{MCJ}((T_1, \mathbf{1}), (T_2, \mathbf{1})).$$

For a graphic illustration of the MCW and MCJW metrics computation, see Fig. 3.

5 Experimental results

We analyzed the properties of fifteen metrics (see Table 1 for the detailed list) according to the test procedures designed by Kuhner and Yamato in Kuhner and Yamato (2015).

5.1 "N-away" experiment

Here, we want to check how well the tested metrics recognize trees at different rearrangement distances. For the experiment, we used a data file kindly provided to us by the authors of Kuhner and Yamato (2015). The data set consists of local trees estimated on the basis of the simulated ancestral recombination graphs (ARGs) of long chromosomal regions. The simulation was performed with the use of *ms* program (Hudson, 2002) with the θ parameter set to 100. It generated 5000 random ARGs consisting of 20 leaves with sequences of length 40,000 bp for further tests, see Kuhner and Yamato (2015) for details.

MOST WIEDZY Downloaded from mostwiedzy.pl

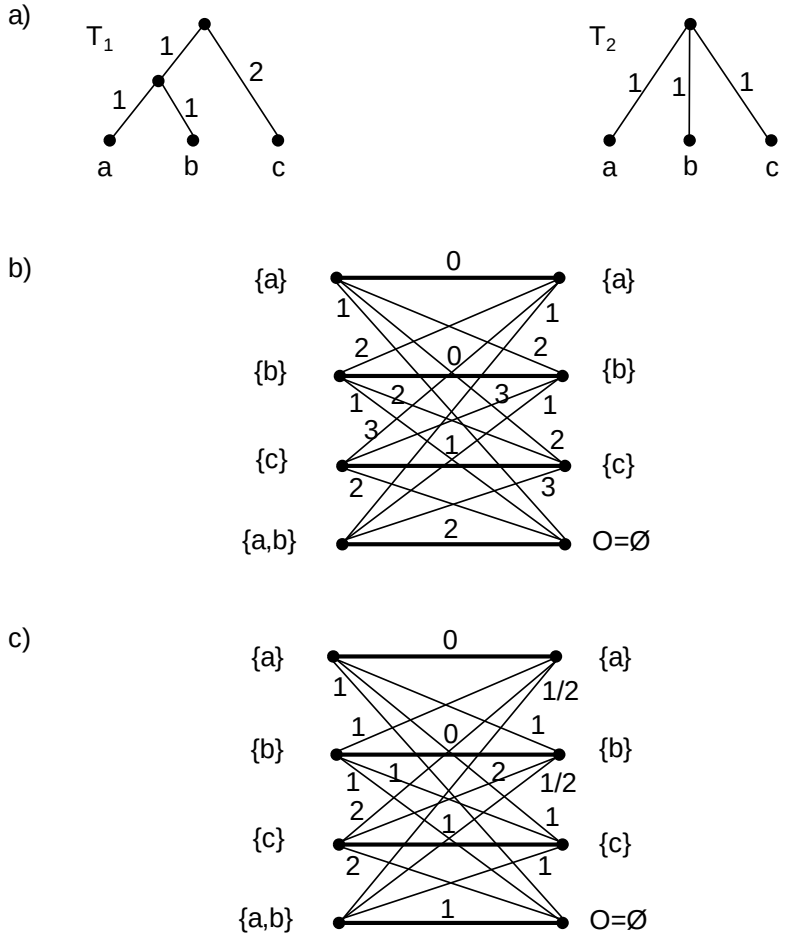

Figure 3: Computation of matching distances for trees with weights: a) example of trees, b) computation of MCW distance, $d_{MCW}(T_1, T_2) = 2$; observe, that $\sigma_w(T_1, l_1)(\{c\}) = 2 \neq \sigma_w(T_2, l_2)(\{c\}) = 1$, c) computation of MCJW distance, $d_{MCJW}(T_1, T_2) = 2$. Values of some classic phylogenetic metrics (see Table 1) for these trees: $d_{RFW}(T_1, T_2) = 1$, $d_{RFW085}(T_1, T_2) = 1$, $d_{Geo}(T_1, T_2) \approx 1.41$.

MOST WIEDZY Downloaded from mostwiedzy.pl

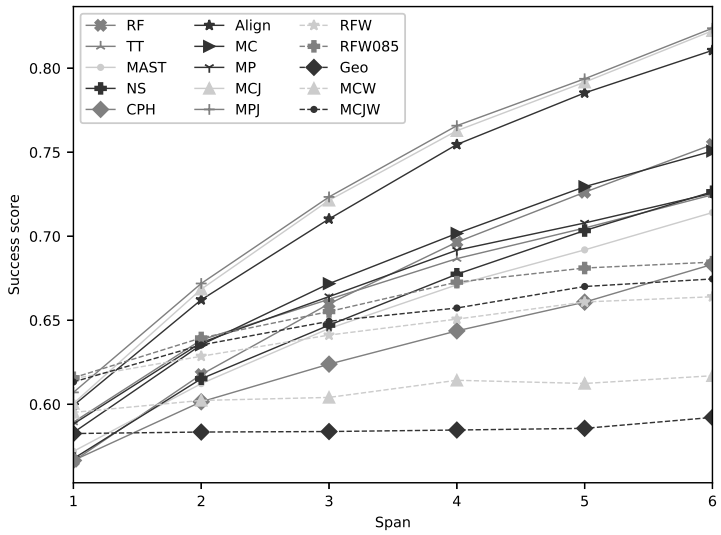

Figure 4: Success score for trees at magnitude 13 in the "n-away" experiment.

According to Kuhner and Yamato (2015), adjacent local trees in a sequence obtained from a single ARG "almost always have a recombination distance of 0 or 1". Therefore, for given local trees T_i, T_j, T_k in a sequence, such that $i < j < k$, we check the relation between distances $d(T_i, T_j)$ and $d(T_i, T_k)$ for d being each of the abovementioned metrics. Similarly to Kuhner and Yamato (2015), we decided to use the following scoring schema: $score_d(T_i, T_j, T_k)$ is 1 if $d(T_i, T_j) < d(T_i, T_k)$, the value equals 0.5 if $d(T_i, T_j) = d(T_i, T_k)$, and becomes 0 otherwise. There are various strategies for choosing indexes i, j, k. We adopted one of the methodologies mentioned in Kuhner and Yamato (2015) and analyzed the distances for a fixed *magnitude* defined as the value of $j - i$ and a variable *span*, which is by definition equal to $k - j$. We also set i to 1, which corresponds to choosing the first tree in a sequence.

In Fig. 4 we present the average $score_d(T_1, T_{14}, T_{14+span})$ over the 5000 tree sequences for a magnitude of 13. Success score presented on the vertical axis in

MOST WIEDZY Downloaded from mostwiedzy.pl

the figure corresponds to the average score computed for each of the analyzed metrics. We are especially interested in metrics with the highest success score, because higher score means that the particular metric more often orders analyzed trees in the expected way, i.e., trees at larger recombination distance are also more distant to each other according to the particular metric.

In the case of a very small span equal to 1, metrics using branch lengths, i.e., RFW, RFW085, MCJW receive a slightly higher score than any other purely topological ones. This can be explained by the fact that many of trees at span of 1 can have the same topology, so they are indistinguishable for purely topological metrics. As the span grows, the situation changes, so the metrics which ignore branch lengths get the highest scores. In particular, we confirmed the observation made by Kuhner and Yamato in Kuhner and Yamato (2015) that the Align metric performs very well in this task, simultaneously we were able to find two metrics that perform even better, namely MPJ and MCJ.

Based on the results presented in Fig. 4 we can distinguish three groups. The metrics in the first group receive the highest score consecutively for all spans greater than 1, i.e., starting from the best: MPJ, MCJ and Align. All the three metrics are purely topological and their score increases as span grows.

The second group is formed by metrics with moderate success score. This group includes metrics that do not take into account branch lengths, i.e: MC, RF, MP, NS, TT, MAST, CPH as well as metrics for weighted trees: RFW085, MCJW, RFW. Similarly as for the metrics in the first group, their score increases as span grows. We can notice that from all five metrics designed for weighted trees the MCJW metric (defined in this study) receives the second best score, right after RFW085.

The last two metrics having the lowest score (below 0.62) for spans of size 3 and higher, i.e., MCW and Geo form the third group. Both the metrics take

MOST WIEDZY Downloaded from mostwiedzy.pl

branch lengths into account and their score is almost constant across the spans.

Another interesting observation is that the original matching metrics, MP and MC received a considerably lower scores that Align, which may indicate that normalization in branch (which takes place for Align) or cluster (used in MPJ and MCJ) comparison has a positive impact on the general performance of the particular methods in this task. A similar observation is also valid for metrics that use branch lengths, where the Jaccard variant of MCW performs much better than the original MCW distance.

5.2 "Bullseye" experiment

As described in Kuhner and Yamato (2015), in this experiment we wanted to check how well the tested metrics distinguish between a better and worse reconstruction of the tree. For the experiment, we generated 1000 random trees with 20 leaves using the *rantree.c* program (Kuhner and Yamato, 2015) under the Yule (branching) process. Next, we generated random DNA sequences of length 2000 bp on these trees with the *rectreedna.c* application (Kuhner and Yamato, 2015) using a Kimura two-parameter model with a transition/transversion ratio of 2. Both programs have been archived on Dryad[3] by the authors of Kuhner and Yamato (2015).

Then we reduced the data sets by 200 bp to produce additional sets consisting of sequences of length 1800, 1600,..., 200 bp. Next, for each of the data sets, we inferred the maximum likelihood tree (breaking ties arbitrarily by choosing the first tree) with the PAUP* 4 application (Swofford, 2003) under the molecular-clock assumption and using the same substitution model that was used for the simulation of the data. The rationale of the experiment is that we expect that, on average, trees inferred from longer sequences will be more accurate (Kuhner and Yamato, 2015), so their distance to the true tree should be smaller. Similarly

[3] https://datadryad.org/resource/doi:10.5061/dryad.g9089

MOST WIEDZY Downloaded from mostwiedzy.pl

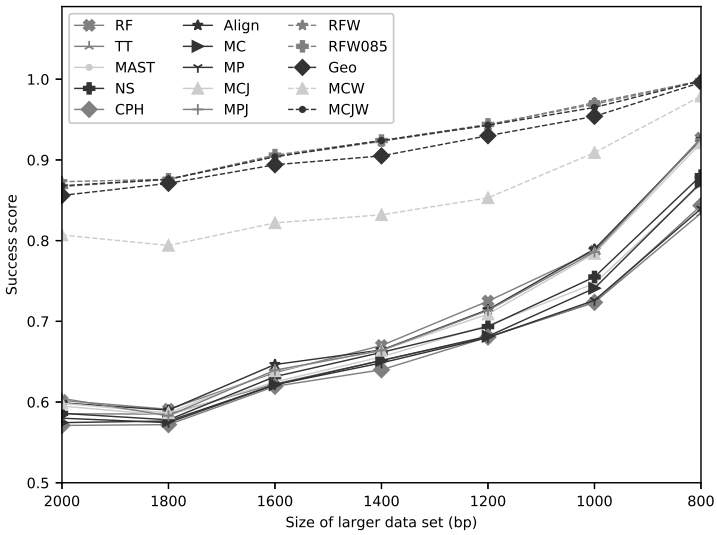

Figure 5: Success score for trees in the "bullseye" experiment. The difference in the sequence length between the longer and shorter data sets equals 600 bp.

as in the first experiment, a tested metric d receives a score of 1 if the distance according to d between the true tree T^* and a tree T_{long} inferred on the basis of longer sequences is lower than the distance between a tree T_{short} inferred using shorter sequences and T^*, d receives a score of 0.5 if $d(T^*, T_{long}) = d(T^*, T_{short})$, otherwise it obtains a score of 0.

Similarly, as described in Kuhner and Yamato (2015), we analyzed the distances for trees inferred from sequences differing by 600 bp, e.g., if T_{long} was inferred from 2000 bp, then a 1200 bp dataset was used during the inference of T_{short}. In Fig. 5, we present the average score received by the analyzed metrics based on 1000 random true trees. Similarly to the situation in the previous experiment, success score presented on the vertical axis in this figure corresponds to average score computed for each of the analyzed metrics. Here, we also are interested in metrics with the highest success score. The higher the score is, the

MOST WIEDZY Downloaded from mostwiedzy.pl

more often the particular metric indicates that the tree reconstructed based on a larger amount of data is more similar to the true tree than the tree created using less data, what seems to be an intuitive and natural situation.

We can easily observe that all branch-length-aware metrics perform much better than any of the purely topological distances. Based on the overall performance in this experiment we can clearly split the analyzed fifteen metrics into three groups. The first group consists of the three best-performing metrics: RFW, RFW085, MCJW and the fourth best metric Geo. The two RFW, RFW085 metrics received slightly better scores than MCJW; however, their difference from MCJW did not exceed 0.006.

The second group consists of only one metric, i.e.: MCW, which achieved moderate score in the experiment. This score is clearly lower than the score of any other metric for weighted trees, but is also considerably higher than the results received by any of the purely topological distances.

The third group (with the score lower than the two previous ones) contains all and only the metrics which do not use branch lengths. The best results in this group, especially visible for shorter sequences, can be observed for four metrics: MPJ, MCJ, Align and RF.

It can be also noticed that the best results in all the cases appear for the shortest data sets. This can be explained by the fact that in this case the ratio between the length of the data used to construct tested trees (i.e.: T_{long} and T_{short}) is the highest reaching $4 = 800/200$ (so it is easier to detect impact of the data size on the quality of the reconstructed trees), while for the case of longer sequences (where 2000 bp is used for T_{long}) the respective factor is approximately equal to only $1.67 \approx 2000/1200$.

Furthermore, similarly as in the "n-away" experiment, we can observe that MCJW always performs much better than the original MCW distance (the

MOST WIEDZY Downloaded from mostwiedzy.pl

difference between scores varying from 0.019 to 0.092). The same relation can be observed for the MCJ, MC pair (difference from 0.010 to 0.049) and MPJ, MP (difference from 0.0085 to 0.0865).

6 Discussion

The vast majority of phylogenetic metrics prevailing in the literature fall into two groups:

- possible to compute efficiently in polynomial time but based on topological and graph features devoid of clear biological and evolutional interpretation (e.g.: Nodal Distance (Bluis and Shin, 2003), Path Difference (Steel and Penny, 1993));
- biologically perspicuous but with NP-hard evaluation, therefore difficult to determine in a satisfactory time span (i.e. NNI (DasGupta et al., 1997), SPR, TBR (Allen and Steel, 2001)).

The matching metrics method and its particular implementations (Bogdanowicz and Giaro, 2012, 2013, 2017) give hope to overcome this frustrating dichotomy, as we obtained metrics showing valuable properties:

- computational efficiency, i.e. polynomial time evaluation algorithms,
- clear output interpretation: a side product of the distance computing is the pairing that illustrates the corresponding structure fragments in both compared trees (e.g., similar clades, pair sets),
- minor sensitivity to small topological changes in comparison to the space diameter or an expected distance between random trees.

Both the MCJ and MPJ metrics preserve the above advantages, but compared to MC, MP, MS and some classic measures, they lose one convenient

MOST WIEDZY Downloaded from mostwiedzy.pl

feature, that is an (half-)integer valuation. Integer expression of the differences between purely combinatorial objects, such as trees without numerical weights, seems to be natural, since it facilitates the insight into the metric space structure (i.e. examining the nearest neighborhood of a tree, differentiation between "compact" tree space and those scattered into distant "islands").

The need to define metrics for trees with numerical weights at the edges has provoked a search for generalizations of matching metrics. It is natural to construct definitions of dissimilarity measures as extensions of their purely topological counterparts, i.e., retaining same values on trees with unit edge lengths. The Definition 13 and the MCW and MCJW measures derived from it meet this need. We also examine a different approach to extending of the matching concept on weighted trees. However, the properties of these measures are still hard to predict, and they need an algorithmic implementation and repetitions of the described experiments.

A general comparison of the quality of phylogenetic metrics seems unfeasible, as each performs better or worse depending on the application. For example, the experiments of Bogdanowicz and Giaro (2013) prove increased efficiency of the heuristic for the RF-supertree problem (Bansal et al., 2010) equipped with MC instead of RF, and in Bogdanowicz and Giaro (2017) MP used as a guide metric overcame the wide set of dissimilarity measures in tests of the heuristic for the rSPR-distance evaluation proposed in Boc et al. (2010); Bordewich et al. (2009). However, both measures performed much worse than their new Jaccard based analogs, MCJ and MPJ in the experiments described here. The evaluation of the quality of the metrics proposed by Kuhner and Yamato Kuhner and Yamato (2015) is important as it focuses on the generally desirable characteristics of metrics that are not limited to application in a specific heuristic. Due to the potentially important applications, it is also important to examine the suitability

MOST WIEDZY Downloaded from mostwiedzy.pl

of metrics for weighted trees in the monitoring of convergence of the Bayesian phylogenetics analysis process (Nylander et al., 2008).

7 Conclusion

We defined a new metric MPJ for the comparison of rooted phylogenetic trees by modifying the already known MP (Bogdanowicz and Giaro, 2017) distance. We also analyzed the properties of a previously known MCJ metric which can be regarded as a Jaccard-based version of the MC (Bogdanowicz and Giaro, 2013) distance. In both cases, the modification concerns the h function that is responsible for assessing the distance between basic elements, i.e., clusters (for MC) and pair sets (for MP), and relies on replacing the original function with its normalized (Jaccard) version. In the "n-away" experiment, the modification achieves results superior to Align, the previous best method. We showed that the diameter and the expected distance between random trees in both of these metrics grow linearly with respect to the number of leaves n, but a relocation of a constant number of leaves can change the distance by no greater than $O(\ln(n))$.

As shown in Kuhner and Yamato (2015), metrics that take into account branch lengths perform better in the "bullseye" experiment than the purely topological ones. Therefore, we presented a general method (Definition 13) of defining branch-length-aware matching metrics, extending our earlier matching metrics paradigm. Furthermore, using that approach, we defined weight versions of the MC, and MCJ metrics, namely MCW and MCJW. In the "bullseye" experiment, the MCJW metric performed almost as well as the weight versions of RF distance.

An interesting fact is that introducing normalization through applying the Jaccard distance resulted in increasing the performance of a suitable version of

MOST WIEDZY Downloaded from mostwiedzy.pl

metrics in both experiments. This observation motivates further research in that area, which may include, for instance, expanding the experimental framework by adding other types of tests or conducting the same experiments on larger and differently prepared test sets. Finally, although we present some preliminary results of the theoretical properties of the MCJ, MPJ metrics, there is still a place for more formal analyses of the properties of the MCW and MCJW distances.

The discussed metrics MCJ, MPJ, MCW and MCJW (among many others) are implemented in TreeCmp 1.7 application freely available at `https://github.com/dbogdanowicz/TreeCmp-weighted-metrics`.

Acknowledgments

We would like to thank Mary K. Kuhner for the helpful discussion and the dataset of trees used in the "n-away" experiment.

References

B. Allen, and M. Steel, "Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees", *Ann. Comb.* vol. 5, no. 1. pp. 1-–15, Jun. 2001, doi:10.1007/s00026-001-8006-8.

M. S. Bansal, J. G. Burleigh, O.Eulenstein, and D. Fernández-Baca, "Robinson-Foulds Supertrees", *Algorithms Mol. Biol.*, vol. 5, Feb. 2010, Art. no. 18, doi: 10.1186/1748-7188-5-18 [Online].

L. Billera, S. Holmes, and K. Vogtmann, "Geometry of the Space of Phylogenetic Trees", *Adv. Appl. Math.*, vol. 27, no. 4, pp. 733–767, Nov. 2001, doi: 10.1006/aama.2001.0759.

MOST WIEDZY Downloaded from mostwiedzy.pl

J. Bluis, and D.-G. Shin, "Nodal distance algorithm: calculating a phylogenetic tree comparison metric", in *Proc. BIBE*, Bethesda, MA, USA, 2003, pp. 87–94, doi: 10.1109/BIBE.2003.1188933.

A. Boc, H. Philippe, and V. Makarenkov, "Inferring and validating horizontal gene transfer events using bipartition dissimilarity", *Syst. Biol.*, vol. 59, no. 2, pp. 195–211, Mar. 2010. doi: 10.1093/sysbio/syp103.

D. Bogdanowicz, and K. Giaro, "Matching split distance for unrooted binary phylogenetic trees", *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 1, pp. 150–160, Jan.-Feb. 2012, doi: 10.1109/TCBB.2011.48.

D. Bogdanowicz, and K. Giaro, "On a matching distance between rooted phylogenetic trees", *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 3, pp. 669–684, Sep. 2013, doi: 10.2478/amcs-2013-0050.

D. Bogdanowicz, and K. Giaro, "Comparing Phylogenetic Trees by Matching Nodes Using the Transfer Distance Between Partitions", *J. Comput. Biol.*, vol. 24, no 5, pp. 422–435, May 2017, doi: 10.1089/cmb.2016.0204.

M. Bordewich, O. Gascuel, K.T. Huber, and V. Moulton, "Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference", *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 6, no. 1, pp. 110–117, Jan.-Mar. 2009, doi: 10.1109/TCBB.2008.37.

S. Böcker, S. Canzar, and G. W. Klau. "The Generalized Robinson-Foulds Metric", in *Proc. WABI*, Sophia Antipolis, France, 2013, pp. 156–169, doi: 10.1007/978-3-642-40453-5_13.

G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, "Nodal distances for rooted phylogenetic trees", *J. Math. Biol.*, vol. 61, no 2, pp. 253–276, Aug. 2010, doi: 10.1007/s00285-009-0295-2.

MOST WIEDZY Downloaded from mostwiedzy.pl

G. Cardona, A. Mir, F. Rosselló, L. Rotger, and D. Sanchez, "Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf", *BMC Bioinf.*, vol. 14, Jan. 2013, Art. no. 3, doi: 10.1186/1471-2105-14-3 [Online].

D. E. Critchlow, D. K. Pearl, and C. Qian, "The Triples Distance for Rooted Bifurcating Phylogenetic Trees", *Syst. Biol.*, vol. 45, no. 3, pp. 323–334, Sep. 1996, doi: 10.1093/sysbio/45.3.323.

B. Dasgupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang, "On distances between phylogenetic trees", in *Proc. SODA*, New Orleans, LA, USA, 1997, pp. 427–436.

G. F. Estabrook, F. R. McMorris, and C. A. Meacham, "Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units", *Syst. Biol.*, vol. 34, no. 2, pp. 193–200, Jun. 1985, doi: 10.2307/sysbio/34.2.193

C. R. Finden, and A. D. Gordon, "Obtaining common pruned trees", *J. Classif.* vol. 2, no. 1, pp. 255–276, Dec. 1985, doi: 10.1007/BF01908078.

H. N. Gabow, and R. E. Tarjan, "Faster scaling algorithms for network problems", *SIAM J. Comput.*, vol. 18, no. 5, pp. 1013–1036, Oct. 1989, doi: 10.1137/0218069.

W. Goddard, E. Kubicka, G. Kubicki, and F. R. McMorris, "The agreement metric for labeled binary trees", *Math. Biosci.*, vol. 123, no. 2, pp. 215–226, Oct. 1994, doi: 10.1016/0025-5564(94)90012-4.

D. M. Hillis, T. A. Heath, and K. St. John, "Analysis and Visualization of Tree Space", *Syst. Biol.*, vol. 54, no. 3, pp. 471–482, Jun. 2005, doi: 10.1080/10635150590946961.

MOST WIEDZY Downloaded from mostwiedzy.pl

R. R. Hudson, "Generating samples under a Wright–Fisher neutral model of genetic variation", *Bioinformatics*, vol. 18, no. 2 pp. 337–338, Feb. 2002, doi: 10.1093/bioinformatics/18.2.337.

S. Kosub, "A note on the triangle inequality for the Jaccard distance", *Pattern Recognit. Lett.*, vol. 120, pp. 36–38, Apr. 2019, doi: 10.1016/j.patrec.2018.12.007.

M. K. Kuhner, and J. Yamato, "Practical performance of tree comparison metrics", *Syst. Biol.*, vol. 64, no. 2, pp. 205–214, Mar. 2015, doi: 10.1093/sysbio/syu085.

A. McKenzie, and M. Steel, "Distributions of cherries for two models of trees", *Math. Biosci.*, vol. 164, no. 1, pp. 81–92, Mar. 2000, doi: 10.1016/s0025-5564(99)00060-7.

T. M. W. Nye, P. Lio, and W. R. Gilks, "A novel algorithm and web-based tool for comparing two alternative phylogenetic trees", *Bioinformatics*, vol. 22, no. 1, pp. 117–119, Jan. 2006, doi: 10.1093/bioinformatics/bti720.

J. A. A. Nylander, J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford, "AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics", *Bioinformatics*, vol. 24, no. 4, pp. 581-–583, Feb. 2008, doi: 10.1093/bioinformatics/btm388.

J. B. Orlin, and R. K. Ahuja, "New scaling algorithms for the assignment and minimum mean cycle problems", *Math. Program.*, vol. 54, pp. 41–56, Feb. 1992, doi: 10.1007/BF01586040.

M. Owen, and J. S. Provan, "A Fast Algorithm for Computing Geodesic Distances in Tree Space", *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 1, pp. 2–13, Jan.-Feb. 2011, doi: 10.1109/TCBB.2010.3.

MOST WIEDZY Downloaded from mostwiedzy.pl

J. Ramon, and M. Bruynooghe, "A polynomial time computable metric between point sets", *Acta Inform.*, vol. 37, no. 10, pp. 765–780, Jul. 2001, doi: 10.1007/PL00013304.

D. F. Robinson, and L. R. Foulds, "Comparison of weighted labelled trees", *Lect. Notes Math.*, vol. 748, pp. 119–126, 1979, doi: 10.1007/BFB0102690.

D. F. Robinson, and L. R Foulds, "Comparison of phylogenetic trees", *Math. Biosci.*, vol. 53, pp. 131–147, Feb. 1981, doi: 10.1016/0025-5564(81)90043-2.

C. Semple, and M. Steel, *"Phylogenetics"*, 1st ed., New York, NY, USA: Oxford University Press OUP, 2003, pp. 50–53.

M. A. Steel, and D. Penny, "Distributions of tree comparison metrics – some new results", *Systematic Biology Syst. Biol.*, vol. 42, no. 2, pp. 126–141, Jun. 1993, doi: 10.1093/sysbio/42.2.126.

D. L. Swofford, "PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4", Sunderland, MA, USA: Sinauer Associates, Jan. 2003.

C. Whidden, N. Zeh, and R. G. Beiko, "Supertrees based on the subtree prune-and-regraft distance", *Syst. Biol.*, vol. 63, no. 4, pp. 566–581, Jul. 2014, doi: 10.1093/sysbio/syu023.

MOST WIEDZY Downloaded from mostwiedzy.pl

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 1: Phylogenetic tree metrics used in the experiments. The fist ten are topological metrics, i.e, are defined on R_L or R_L^B, whereas the next five use both topology and branch lengths.

Abbreviation	Citation	Basis
RF	(Robinson and Foulds, 1981)	The simplest and popular phylogenetic metric, it counts the number of clades appearing in only one of the compared trees.
TT	(Critchlow et al., 1996)	For each 3-element subset of L, there are three possible binary and one non-binary tree shapes with these leaves. TT checks for differences in shapes of the subtrees induced by the triple of leaves in the compared trees, and finally counts them over all such 3-sized leaf sets.
MAST	(Finden and Gordon, 1985; Goddard et al., 1994)	The *agreement subtree* of the trees is a subtree induced by subset of $\|L\|$ and present in both trees. The metric counts the leaves not included in the agreement subtree with the largest possible leaf set.
NS	(Cardona et al., 2010)	For each ordered pair of leaves u, v, we find a length of a path in the tree leading from u to the lowest common ancestor of u, v. These lengths for all leaf pairs can be arranged into a numerical vector in $\|L\|(\|L\|-1)$-dimensional real space. The metric value is a distance between these vectors of both trees.
CPH	(Cardona et al., 2013)	Similar to NS, but we consider the vectors of the depths (relative to the root) of the lowest common ancestors of u, v over all unordered pairs u, v and singletons (i.e. $u = v$).
Align	(Nye et al., 2006)	The cost of the appropriately defined assignment between branch splits of both trees.
MC	(Bogdanowicz and Giaro, 2013)	The weight of the matching between clusters, i.e. consider Def. 3, but with $\|A \oplus B\|$ instead of $J(A, B)$.
MP	(Bogdanowicz and Giaro, 2017)	The weight of the matching between pair sets, i.e. consider Def. 4, but with $\|A \oplus B\|$ instead of $J(A, B)$.
MCJ	(Böcker et al., 2013)	MC with a normalized inter-cluster distance, i.e. see Def. 3.
MPJ	this study	MP with a normalized inter-pair set distance, i.e. see Def. 4.
RFW	(Robinson and Foulds, 1979)	The generalization of RF for weighted trees, but instead of simply counting clades, it sums the differences of the lengths of their edges in both trees, assuming a branch length of 0 for absent clades.
RFW085	(Kuhner and Yamato, 2015)	RFW with differences raised to the power 0.85.
Geo	(Billera et al., 2001)	Special topological *tree space* was defined in Billera et al. (2001), with points corresponding to weighted trees. The metrics value is the length of a geodesic path in this space connecting the compared trees.
MCW	this study	weighted version of MC, i.e. see Def. 7.
MCJW	this study	MCW with a normalized inter-cluster distance, i.e. see Def. 8.

View publication stats