APPLICATIONES MATHEMATICAE
38,1 (2011), pp. 1–16

Damian Bogdanowicz (Gdańsk)

ANALYZING SETS OF PHYLOGENETIC TREES USING METRICS

Abstract. The reconstruction of evolutionary trees is one of the primary objectives in phylogenetics. Such a tree represents historical evolutionary relationships between different species or organisms. Tree comparisons are used for multiple purposes, from unveiling the history of species to deciphering evolutionary associations among organisms and geographical areas.

In this paper, we describe a general method for comparing phylogenetic trees and give some basic properties of the Matching Split metric, which is a special case of a general definition. We focus on four metrics for binary unrooted trees. We present results of a computational experiment concerning an application of those metrics to estimating the quality of a phylogenetic signal.

1. Introduction. A phylogenetic tree (also called an evolutionary tree) represents historical evolutionary relationships between different species or organisms. Typically a set of extant (present day) species labels the leaves of the tree and the remaining vertices represent ancestral species. If a root vertex is present (the tree is rooted), then it corresponds to the oldest ancestor of the species under study. Often, there is insufficient information to determine the root and the tree is left unrooted. Unrooted trees still provide a notion of evolutionary relationships between organisms even if the direction of descendants remains unknown [9].

There are many methods for constructing phylogenetic trees, e.g. Distance, Parsimony, Maximum Likelihood, Bayesian approach. Applying those techniques usually results in different trees for the same input data. An im-

2010 *Mathematics Subject Classification*: Primary 05C05, 92D15; Secondary 68R10, 05C90.
Key words and phrases: phylogenetic tree, phylogenetic tree metric, splits, minimum weight perfect matching, Matching Split distance.

DOI: 10.4064/am38-1-1

portant problem is to determine how distant two trees reconstructed in such a way are from each other. In comparison to others, the Bayesian methodology, which we have used during the experiments presented in Section 6, is relatively new and becomes more and more popular [17]. Bayesian analysis of phylogenies is similar to Maximum Likelihood in that the user postulates a probabilistic model of evolution and the program searches for the best trees that are consistent with both the model and the data. Unlike ML, which seeks the single most likely tree, Bayesian analysis searches for the best set of trees. In order to obtain useful, reliable biological information based on those sets of trees, methods of postprocessing and visualization are required. Various methods of computing a consensus tree of a set of trees have been developed. One of such methods that uses phylogenetic metrics, is based on finding a median tree. A median tree can be regarded as a manner of extracting common biological information from the analysis of a set of slightly different trees. Penny et al. [16] propose defining a consensus of a collection P of binary trees based on the same leaves as a median tree of P in the space of binary trees with metric d. That is, given a set (also called *profile*) $P = \{T_1, \ldots, T_k\}$ of arbitrary phylogenetic trees with the same sets of leaves, *a median tree* for P is a tree T which minimizes the expression $D(T, P) = \sum_{i=1}^{k} d(T, T_i)$.

Another newer approach of postpocessing suggested by Stockham et al. [22] uses clustering technique. A method of visualization has been investigated by Hillis et al. [10], where multidimensional scaling of a phylogenetic tree space is considered.

Phylogenetic trees have taken on a great importance in evolutionary biology and tree comparisons are used for multiple purposes, from unveiling the history of species to deciphering evolutionary associations among organisms and geographical areas [24]. Phylogenetic tree distances can be used as a tool in studies of host-parasite associations [24]. Comparing phylogenetic trees is also very useful in mining phylogenetic information databases [25].

In this paper, we focus on methods of analyzing sets of trees using various phylogenetic metrics. We describe a general method for creating phylogenetic tree metrics (Section 2). In Section 4, we present a concrete definition of a new distance for unrooted trees and investigate some basic properties of that metric.

We also report some results of analyzing data from the EMBL-ALIGN database using our software tool that implements some methods of comparing phylogenetic trees (Section 6). The results indicate that there is a relation between concentration of trees (average values of distances between trees in the phylogenetic metrics under study) that are produced during a Bayesian Markov Chain Monte Carlo (MCMC) process and the phylogenetic

signal quality in the input sequences. A *phylogenetic signal* is described as a tendency for related species to resemble each other more than they resemble species drawn at random from the phylogenetic tree [2], [13].

2. A general method for creating metrics. In this section, we suggest a general *metric* (a function that satisfies non-negativity, identity of indiscernibles, symmetry and triangle inequality) between phylogenetic trees. This is a generalization and extension of the approach suggested by Nye et al. [15]. Nye et al. describe a distance between phylogenetic trees that is based on finding a bijection between branches in both trees being compared that maximizes a sum of scores for related branches. Although the authors of [15] claim that the procedure can be applied to any kind of phylogenetic trees, i.e. rooted, unrooted and not necessarily bifurcating, they do not explicitly show how to compare trees with different numbers of edges. Here, we overcome that problem by introducing an artificial element O. We use a slightly different approach and concentrate on minimizing (not maximizing as in [15]) a sum of scores. Secondly, we do not limit the methodology to scoring a matching between branches in phylogenetic trees, but introduce a general concept of description elements, members of a set which we denote by $D \setminus \{O\}$. A *description element* is an abstraction of a piece of information about the topology of a tree, for example it can be a branch, split or cluster in some phylogenetic tree description. A set of description elements which we consider should explicitly identify a tree. An example of description elements that are related to internal nodes in phylogenetic trees is described in Section 5.

Given a graph $G = (V, E)$, a *matching* M in G is a set of pairwise non-adjacent edges; that is, no two edges in M share a common vertex. A *perfect matching* is a matching which covers all vertices of the graph. If we assign weights to edges in G, then a *minimum weight perfect matching* is defined to be a perfect matching for which the sum of the weights of the edges has a minimal value. A *bipartite graph* is a graph whose vertices can be decomposed into two disjoint sets such that no two vertices within the same set are adjacent.

The main idea behind our definition of a general metric consists of three steps. In the first step we use a function f in order to transform both trees into sets of description elements.

In the second step we construct a bipartite graph G whose vertices correspond to those sets of description elements such that all vertices of each part correspond to description elements of one of the trees. Weights of edges in G depend on the value of a metric function h on the set of description elements. In the last step we compute the value of a minimum weight perfect matching in the graph G. The value is a topological distance between the trees under consideration.

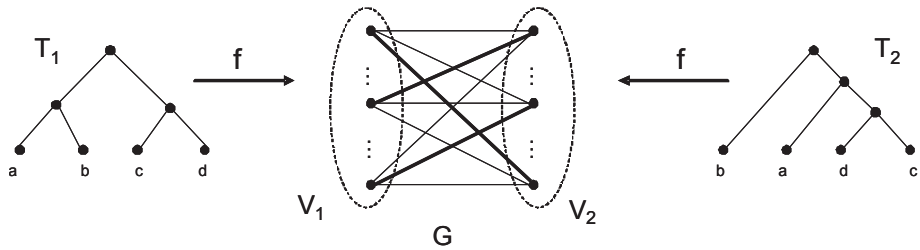

Fig. 1. Illustration for Definition 2.1

Let D be a fixed set with a distinguished element O. Then the set $D\setminus\{O\}$ will be called *the set of description elements*.

DEFINITION 2.1. Let T_1 and T_2 be phylogenetic trees such that $T_1, T_2 \in \mathcal{T}_L$, where $\mathcal{T}_L$ is a family of phylogenetic trees on a set L of species (leaves). Let $f : \mathcal{T}_L \to 2^{D\setminus\{O\}}$ be an injective function that assigns a finite set of description elements (elements of $D\setminus\{O\}$) to a tree. Let $h : D\times D \to \mathbb{R}_+\cup\{0\}$ be a metric on D. Finally, let $G = (V_1, V_2, E)$ be a complete bipartite graph constructed as follows:

(1) vertices in V_1, V_2 correspond to elements of $f(T_1)$, $f(T_2)$;
(2) if $f(T_1), f(T_2)$ do not have the same number of elements, we extend the smaller part of G by the missing number of dummy vertices, which all correspond to the element $O \in D$;
(3) we connect every $u \in V_1$ to every $w \in V_2$ by an edge $\{u, w\}$ with weight $w(\{u, w\}) = h(f_u(T_1), f_w(T_2))$, where $f_v(T_i)$ denotes the element of $f(T_i) \cup \{O\}$ which corresponds to the vertex v.

We define

$$d_{f,h}(T_1, T_2) = \min_M \Big(\sum_{e\in M\subseteq E(G)} w(e) \Big), \tag{2.1}$$

where the minimum is taken over all perfect matchings M in G.

2.1. Correctness of the definition

THEOREM 2.2. *The function $d_{f,h}$ is a metric on $\mathcal{T}_L$.*

Proof. It is easy to observe that if $T_1, T_2 \in \mathcal{T}_L$, then $d_{f,h}(T_1, T_2) \geq 0$, and $d_{f,h}(T_1, T_2) = 0 \Leftrightarrow T_1 = T_2$. For every $T_1, T_2 \in \mathcal{T}_L$ we have $d_{f,h}(T_1, T_2) = d_{f,h}(T_2, T_1)$. We have to prove the triangle inequality: if $T_1, T_2, T_3 \in \mathcal{T}_L$, then $d_{f,h}(T_1, T_2) + d_{f,h}(T_2, T_3) \geq d_{f,h}(T_1, T_3)$.

Let $G_{ij} = (V_i, V_j, E_{ij})$ be the bipartite graph corresponding to the pair of trees T_i, T_j and M_{ij} be a minimum weight perfect matching in this graph. Note that adding dummy vertices to both parts in a graph G does not change the weight of a minimum weight perfect matching in this graph because h is a metric, and therefore $h(O, O) = 0$ (compare Lemma 2.3). Hence, without

loss of generality, we can assume that both parts in the graphs G_{12}, G_{23}, G_{13} have the same number of vertices equal to N.

Let $V_1 = \{a_1, \ldots, a_N\}$, V_2 be the set $\{b_1, \ldots, b_N\}$ such that $\{a_l, b_l\} \in M_{12}$, and V_3 be the set $\{c_1, \ldots, c_N\}$ such that $\{b_l, c_l\} \in M_{23}$ for $1 \le l \le N$. Let $w_{12}(l) = w(\{a_l, b_l\})$, $w_{23}(l) = w(\{b_l, c_l\})$ be the weights of the edges in M_{12} and M_{23} respectively (see Fig. 2). Since weights of edges correspond to values of the metric h, we have

$$d_{f,h}(T_1, T_2) + d_{f,h}(T_2, T_3) = \sum_{l=1}^{N} (w_{12}(l) + w_{23}(l)) \ge \sum_{l=1}^{N} w_{123}(l),$$

where $w_{123}(l) = w(\{a_l, c_l\})$ is the weight of the edge in G_{13} that forms a triangle with the edges $\{a_l, b_l\} \in M_{12}$ and $\{b_l, c_l\} \in M_{23}$ (see Fig. 2). It is easy to see that $\sum_{l=1}^{N} w_{123}(l) \ge d_{f,h}(T_1, T_3)$, thus finally $d_{f,h}(T_1, T_2) + d_{f,h}(T_2, T_3) \ge d_{f,h}(T_1, T_3)$. ∎

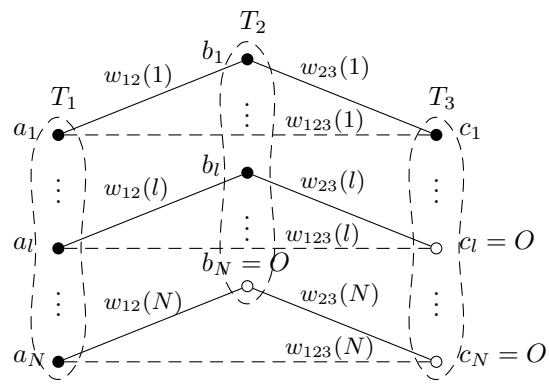

Fig. 2. Illustration of the proof

The computational complexity of the metric depends on three aspects: computation of the functions f and h, and computation of a minimum weight perfect matching in bipartite graphs, which can be done very efficiently, in time $O(|E|\sqrt{|V|}\log(|V|\max_{e\in E} w(e)))$ [12]. Weighted bipartite matching algorithms can be implemented efficiently, and can be applied to graphs of reasonably large size (about 100000 vertices). In this paper, we consider only fully resolved (binary) trees and suggest two special cases of the definition. In both cases considered the cardinalities of the parts of the graph G are equal (we do not need to perform the operation described in the second item of the definition); therefore, the dummy element O is unnecessary.

LEMMA 2.3. *Let $d_{f,h}$ be the metric defined by* (2.1) *and let $k \in f(T_1) \cup \{O\}$, $l \in f(T_2) \cup \{O\}$. If $k = l$, then there exists a minimum weight perfect matching M that contains an edge whose ends correspond to k and l and whose weight is* 0.

Proof. For simplicity we denote by $\{u, v\}$ the edge between the vertices in G that correspond to description elements u and v. Suppose that there exist elements k and l such that $k \in f(T_1) \cup \{O\}$, $l \in f(T_2) \cup \{O\}$, $k = l$ and the edge $\{k, l\}$ does not belong to a minimum weight perfect matching M. Then there exist x, y such that $x \in f(T_1) \cup \{O\}$, $y \in f(T_2) \cup \{O\}$ and the edges $\{x, l\}$, $\{y, k\}$ belong to M. We can always create a new perfect matching $M' = M \setminus \{\{x, l\}, \{y, k\}\} \cup \{\{k, l\}, \{x, y\}\}$. Since weights of edges correspond to values of the metric h, we have

$$w(\{x, l\}) + w(\{y, k\}) \geq w(\{x, y\}) = w(\{x, y\}) + w(\{k, l\}).$$

Therefore, the weight of the matching M' is less than or equal to the weight of M. Since M is a minimum weight perfect matching, so is M', and in addition it fulfills all conditions of the lemma. ∎

The property described in Lemma 2.3 can be used to improve the calculation time of the metric. Namely, during the calculation, the description elements that are present in both trees can be omitted without changing the value of a minimum weight perfect matching.

3. Classical phylogenetic tree metrics for unrooted trees. In this section, we present three definitions of well-known phylogenetic metrics for fully resolved (binary) unrooted trees, which have been used during the experiment described in Section 6. An *unrooted phylogenetic tree* is defined as an acyclic connected graph with no vertices of degree two and every *leaf* (vertex of degree one) labeled uniquely. In a *binary unrooted phylogenetic tree* every internal (i.e. non-leaf) vertex has degree three.

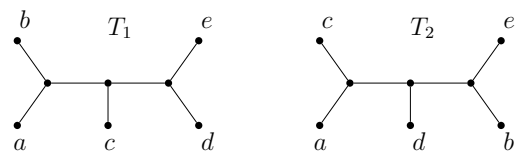

Fig. 3. Examples of binary unrooted trees with five leaves

Definition 3.1 ([5]). Let T be a tree over a set L of leaves and let e be an edge of T. If we remove e, then we divide T into two components. Let A be the set of leaves in one component and B be the set of leaves in the other component. This partition of leaves of T is denoted by $A|B$ and called the *split* corresponding to the edge e.

The split $A|B$ is an unordered pair (i.e. $A|B = B|A$). If $|A| = 1$ or $|B| = 1$, then $A|B$ is *trivial*, otherwise it is *non-trivial* [5]. The set of splits corresponding to all edges in T is called the *set of splits* of T and is denoted $\beta(T)$ [5]. Let $L(T)$ be the leaf set of T and set $|L(T)| = n$.

DEFINITION 3.2. The *Robinson–Foulds distance* (RF) [18] between two trees T_1 and T_2 is defined as

$$d_{RF}(T_1, T_2) = \tfrac{1}{2}|\beta(T_1) \setminus \beta(T_2)| + \tfrac{1}{2}|\beta(T_2) \setminus \beta(T_1)|.$$

The Robinson–Foulds distance between the trees in Fig. 3 is equal to 2. The RF distance is one of the best known and widely used methods of comparison. Our general method can easily yield the RF distance if the functions f and h are defined as follows:

$$f(T) = \beta(T) \quad \text{and} \quad h(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 = s_2, \\ 1 & \text{otherwise.} \end{cases}$$

DEFINITION 3.3 ([5]). Given a split $A|B$. We define the *quartet set* of the split to be

$$q(A|B) = \{ab|cd : a, b \in A,\ c, d \in B,\ a \neq b,\ c \neq d\}.$$

Let T be a phylogenetic tree, and set $q(T) = \bigcup_{A|B \in \beta(T)} q(A|B)$ [5].

DEFINITION 3.4. The *Quartet distance* [21] between two trees T_1 and T_2 is defined as

$$d_{Qt}(T_1, T_2) = \tfrac{1}{2}|q(T_1) \setminus q(T_2)| + \tfrac{1}{2}|q(T_2) \setminus q(T_1)|.$$

The Quartet distance between the trees in Fig. 3 is equal to 4. Our general method yields the Quartet distance if the functions f and h are defined as follows:

$$f(T) = q(T) \quad \text{and} \quad h(q_1, q_2) = \begin{cases} 0 & \text{if } q_1 = q_2, \\ 1 & \text{otherwise.} \end{cases}$$

Let $v(T)$ be a vector of non-negative integers, defined as follows. Each entry in $v(T)$ gives the number of edges in a path between two leaves in T. Let $\mathrm{len}(v)$ be the number of elements in v. Then

$$\mathrm{len}(v(T)) = |L(T)|\frac{|L(T)| - 1}{2}.$$

DEFINITION 3.5. The *Nodal distance* [3] between two trees T_1 and T_2 is defined as

$$d_{Nd}(T_1, T_2) = \sum_{i=1}^{\mathrm{len}(v(T_1))} |v(T_1)(i) - v(T_2)(i)|,$$

where the vectors $v(T_1)$ and $v(T_2)$ are ordered so that the values on the same positions correspond to the same pair of leaves in both vectors.

The Nodal distance between the trees in Fig. 3 is equal to 10.

4. The Matching Split distance. Based on the method described in the second section we can define a new metric that is also based on splits. We will call it *the Matching Split* distance (MS).

DEFINITION 4.1. Let f and h be functions defined as follows:

$$f(T) = \beta(T) \quad \text{and} \quad h(A_1|B_1, A_2|B_2) = \min\{|A_1 \ominus A_2|, |A_1 \ominus B_2|\},$$

where $A_1|B_1$, $A_2|B_2$ are splits corresponding to edges in T_1, T_2 and $A \ominus B = (A \setminus B) \cup (B \setminus A)$. We then define *the Matching Split* (MS) distance as

$$d_{MS}(T_1, T_2) = d_{f,h}(T_1, T_2).$$

For example, we calculate the Matching Split distance between the trees in Fig. 3. Note that we do not need to consider trivial splits for binary unrooted trees because they are the same in each tree. We have the following non-trivial splits for T_1: $ab|cde$, $abc|de$, and for T_2: $ac|bde$, $acd|be$. Using the function h we calculate the distances between those splits: $h(ab|cde, ac|bde) = 2$, $h(ab|cde, acd|be) = 2$; $h(abc|de, ac|bde) = 1$, $h(abc|de, acd|be) = 2$. The weight of a minimum weight perfect matching is 3, so $d_{MS}(T_1, T_2) = 3$.

DEFINITION 4.2 ([1]). Any internal edge of an unrooted binary tree has four subtrees attached to it. A *nearest neighbor interchange* (NNI) occurs when one subtree on one side of an internal edge is swapped with a subtree on the other side of the edge.

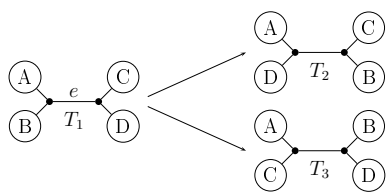

Fig. 4. The trees T_2 and T_2 are results of two possible NNI operations about edge e in T_1. The circles represent any binary subtrees over sets of leaves A, B, C and D.

LEMMA 4.3. *Let $T_1, T_2 \in \mathcal{T}_L$. If $d_{RF}(T_1, T_2) = 1$ then $2 \leq d_{MS}(T_1, T_2) \leq \lfloor n/2 \rfloor$, where $|L| = n$.*

Proof. Let A, B, C, D denote the sets of leaves in the subtrees in Fig. 4. First, notice that if $d_{RF}(T_1, T_2) = 1$, then the trees T_1 and T_2 are at a distance of exactly one NNI operation [4]. Therefore, we have $d_{MS}(T_1, T_2) = \min\{|B|+|D|, |A|+|C|\}$. It is easy to observe that the function h has maximal value equal to $\lfloor n/2 \rfloor$. The distance is minimal and equal to 2 if $|B| = 1$ and $|D| = 1$ or if $|A| = 1$ and $|C| = 1$. Depending on the number of leaves in the subtrees, the distance between the trees T_1 and T_2 can vary between 2 and $\lfloor n/2 \rfloor$. ∎

4.1. Differences between the MS and RF metrics. Although the MS and RF distances seem similar because both use splits, there are substantial differences between them.

Let us consider trees that are as close as possible in the RF metric. The trees T_1 and T_2 (or T_1 and T_3) in Fig. 4 are examples of such trees. For example, for $n = 8$ based on Lemma 4.3 we can always find a pair of totally balanced trees that are at distance $\lfloor n/2 \rfloor$ in the MS metric, which gives a distance of 4 in this case, whereas in the case of the RF metric it is always 1. We can say that the MS metric is sensitive to the sizes of the subtrees that are swapped during an NNI operation; the RF metric is insensitive to it.

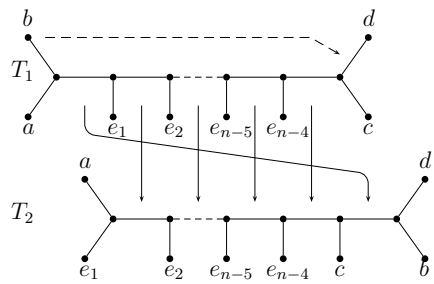

Fig. 5. Differences between the RF and MS distances. Solid arrows indicate connections that appear when computing the MS metric in a minimum weight perfect matching.

The next case considered in this section is important because it illustrates an interesting, not necessarily desired, property of the RF metric.

Consider the trees in Fig. 5. T_2 is obtained from T_1 as a result of removing the leaf b and attaching it to the edge of the leaf d. In spite of the fact that we have done only one modification, we obtain two trees that are as far as possible in the RF metric, i.e. $d_{RF}(T_1, T_2) = n - 3$. Considering the MS distance we obtain $d_{MS}(T_1, T_2) = n - 2$, which is much smaller than the maximal possible distance in this metric, e.g. for $n = 8$ we obtain $d_{MS}(T_1, T_2) = 6$, whereas the maximal possible value is 16. The quotient of $d_{MS}(T_1, T_2)$ by the maximal distance in the MS metric for $n = 8$ equals $3/8$, while in the case of the RF metric it is constant and equal to 1 because the maximal distance in that metric is $n - 3$.

Another argument for substantial similarity of those two trees comes from an agreement subtree approach. Let T be an unrooted tree, and let A be a subset of its leaf set $L(T)$. Consider the minimal subgraph $T(A)$ of T that connects elements of A. Let $T_{|A}$ denote an unrooted tree that is obtained by deleting all vertices of degree two in $T(A)$ and identifying their adjacent edges. $T_{|A}$ is called the *subtree of* T *induced by* A.

Definition 4.4. A tree T with a leaf set $X \subseteq L$ is called an *agreement subtree* of trees $T_1, T_2 \in \mathcal{T}_L$ if $T = T_{1|X} = T_{2|X}$. A *Maximum Agreement Subtree* or MAST is an agreement subtree with the maximum number of leaves.

MAST is one of several methods for extracting information common to trees. In the case of our trees T_1 and T_2 the number of leaves in $\mathrm{MAST}(T_1, T_2)$ is equal to $n-1$ (almost all leaves), which means that those two tress describe evolutionary relationships very similarly. This similarity is not noticed by the RF distance, where the trees are classified as 100% dissimilar, but it is taken into consideration by the MS distance, which describes the dissimilarity as only 37.5% (for $n = 8$).

4.2. Distribution of the MS distance for random trees. In the last part of this section we would like to consider average distances and distributions of the MS metric in the case where the trees are randomly generated. In order to generate random trees we have used the Evolver application from the PAML 4 package [26]. Since tree comparisons are often performed by testing the null hypothesis that the trees are not more congruent (topologically similar) than expected by chance [24], it is important to know what the distribution of distances between random trees for a particular metric looks like. Knowing the distribution of a particular metric on pairs of trees generated by some random process is also useful to interpret the significance of a measured value of the metric on a pair of phylogenetic trees [20].

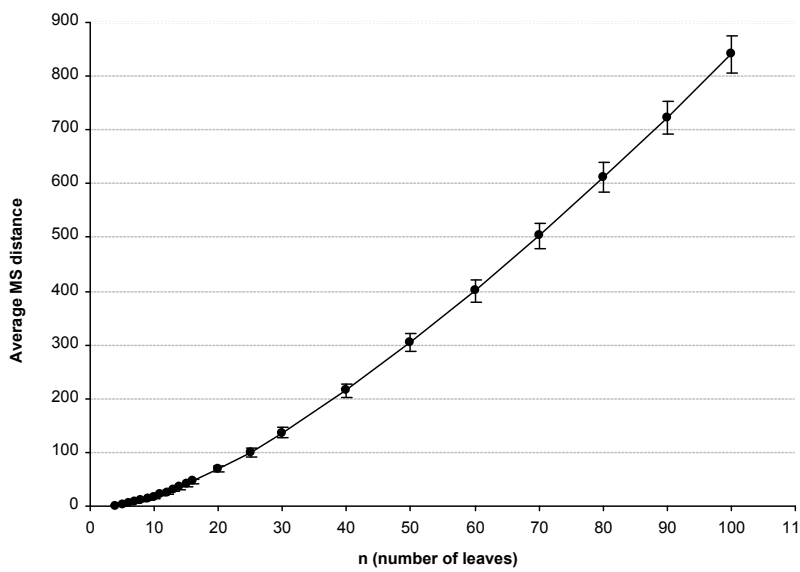

Fig. 6. Average distances in the MS metric for 10000 randomly generated pairs of trees

In Fig. 6 we present the average distance between random trees as a function of the number of leaves in those trees. The character of the relation

seems to be subquadratic; however, we do not have any analytical argument for this. Approximation using the least squares method based on randomly generated data gives the following relation: $\mathrm{AVG}_{MS}(n) = 2.4359n \log_2 n - 7.9490n + 15.7946$ with $R^2 = 0.999983$.

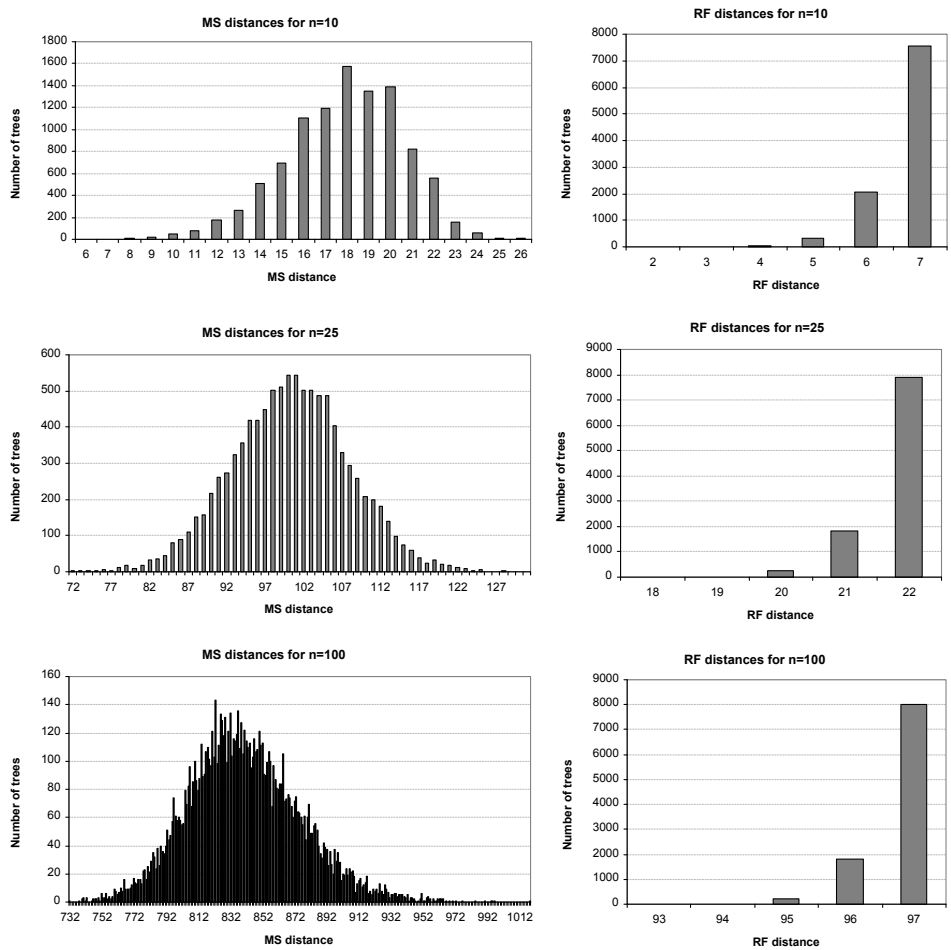

Fig. 7. Distributions of the MS and RF metrics for 10000 randomly generated pairs of trees

In Fig. 7 we present the distributions of the MS and RF metrics for 10000 random pairs of trees with 10, 25 and 100 leaves. Comparing these distributions, we see that the MS metric has a much larger range so it is more discriminating than the RF distance. We can also observe that the shape of the MS distribution is similar to a normal distribution, whereas, as shown by Steel and Penny in [20], the distribution of the RF distance is described asymptotically by a Poisson distribution.

5. A new metric for rooted trees. In this section, we give an example of a new metric for binary rooted trees, the *Matching Pair* distance (MP). A *rooted binary phylogenetic tree* is defined similar to unrooted binary tree, except that one internal vertex, which has degree two, is distinguished and called the *root*. In biological context, the root corresponds to the oldest ancestor of the species under study.

In this case the function f assigns a set of unordered pairs of leaves to the internal vertex that is the least common ancestor for every pair in the set. As a result, f transforms a phylogenetic tree into a family of sets of pairs. As a measure of distance between sets of pairs we can use symmetric difference, therefore $h(A, B) = 0.5|A \ominus B|$, where A, B are sets of unordered pairs (for example see Fig. 8).

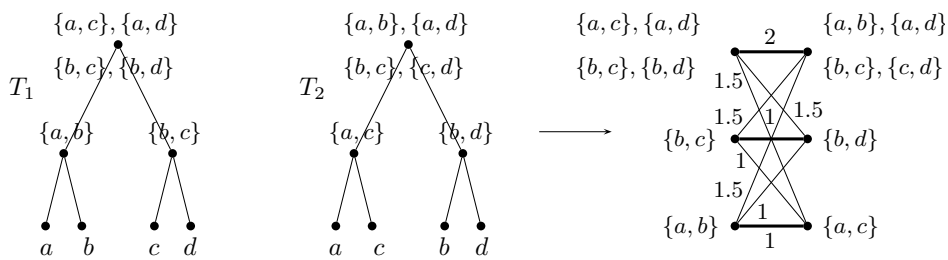

Fig. 8. The MP distance between the trees is equal to 4

Because the computational experiments have been performed for unrooted trees only, we do not explore the rooted trees further in this paper.

6. Experimental results. In this section, we want to investigate whether distances between trees that are created during a Bayesian MCMC (Markov Chain Monte Carlo) process are related to the quality of a phylogenetic signal in the input data. Bayesian inference of phylogeny is based on a quantity called the posterior probability of a tree. Bayes's theorem is used to combine the prior probability of a phylogeny (Pr[Tree]) with the likelihood (Pr[Data | Tree]) to produce a posterior probability distribution on trees (Pr[Tree | Data]) [11]. The posterior probability of a tree can be interpreted as the probability that the tree is correct. For the phylogeny problem, the MCMC algorithm involves two steps. In the first step a new tree is proposed by stochastically perturbing the current tree. In the second step this tree is either accepted or rejected with a probability described by Metropolis et al. [14] and Hastings [8]. If the new tree is accepted, then it is subject to a further perturbation [11]. It turns out that for a properly constructed and appropriately run Markov chain, the proportion of the time that any tree is visited is a valid approximation of the posterior probability of that tree [23], [11].

The Bayesian MCMC method searches a landscape of possible trees, it moves from point to point seeking higher points (more likely trees). The method allows a search to occasionally leap a valley that would otherwise trap it on a suboptimal hill. The final product is a set of trees that the program has repeatedly visited.

During the experiment we analyze MrBayes [19] output files generated based on 16 data sets. Each data set contains an alignment of 8 DNA sequences obtained from the EMBL database. The total number of MCMC generations was set to 10 million. We have omitted the first million of generations ("burn-in" phase). Next, we sequentially computed the distances between trees in resolution of 50000 generations in the four metrics considered above for unrooted trees. In Table 1 we present the average values of 180 distances computed between 181 trees.

Table 1. Average distances derived from eight-taxon data sets obtained from the EMBL-ALIGN database.

No	EMBL-ALIGN accession number	Sequences included	Length	Average values			
				RF	Nodal	Quartet	MS
1	ALIGN_000002	1,2,3,4,5,6,7,8	1632	1.16	13.42	7.44	2.42
2	ALIGN_000521	1,2,3,4,5,6,7,8	1325	0.70	9.77	5.83	1.92
3	ALIGN_000832	2,3,4,5,6,7,9,10	1185	0.79	10.62	6.63	2.04
4	ALIGN_000205	2,3,4,6,8,10,11,12	1386	0.62	7.47	3.11	1.24
5	ALIGN_000297	2,3,4,6,15,16,17,19	1167	2.27	22.32	14.71	4.63
6	ALIGN_000397	2,3,4,6,7,8,9,10	1662	2.01	20.38	13.56	4.13
7	ALIGN_000398	1,2,3,4,5,6,7,8	1656	0.60	7.09	3.31	1.24
8	ALIGN_000623	2,3,4,5,6,10,11,12	1312	0.13	1.69	0.76	0.31
9	ALIGN_000628	2,3,4,5,7,13,17,31	1385	0.01	0.13	0.06	0.02
10	ALIGN_000767	2,3,4,5,6,7,8,10	1386	0.66	7.93	3.31	1.32
11	ALIGN_000771	1,2,3,4,5,6,7,8	4547	1.29	13.84	6.47	2.59
12	ALIGN_000788	2,3,4,5,6,7,12,14	1629	1.70	19.04	12.18	3.60
13	ALIGN_000853	1,2,3,4,5,6,7,8	5307	1.25	14.22	8.22	2.97
14	ALIGN_000930	2,3,4,5,6,7,8,9	1321	3.67	36.09	36.48	8.49
15	ALIGN_000931	2,3,5,14,15,16,19,21	1231	4.63	42.71	45.49	10.81
16	ALIGN_000984	2,3,4,6,7,11,12,13	1139	3.29	31.79	33.59	7.48

In order to compare and analyze the above values, in Table 2 we present results of analyzing the same sets of sequences obtained by Czarna et al. [6], where the authors considered different methods for testing a phylogenetic signal. The last five columns in Table 2 contain results of using those methods, the lower the score the better the phylogenetic signal in data.

It is not hard to notice that there is a relation between the quality of a phylogenetic signal and the average values of distances between trees in MCMC data. Data sets with small distances (for example data sets 7, 8 and 9) correspond to a high percentage of resolved quartets, which indicates a good phylogenetic signal. A poor phylogenetic signal corresponds to larger values of the distances (for example data sets 14, 15 and 16).

Table 2. The percentage of four-taxon subset for which the star topology was the ML solution (unresolved quartets) was calculated using TreePuzzle. The last four columns show the number of trees out of possible 10395 included in the 0.95 confidence set using: expected likelihood weights (ELW) test, Shimodaira-Hasegawa (SH) test, generalised least squares (GLS) test, and weighted least squares test (WLS) [6].

No	EMBL-ALIGN accession number	Unresolved quartets (%)	Number of trees in the 0.95 confidence set			
			SH	ELW	GLS	WLS
1	ALIGN_000002	22.9	141	14	9	135
2	ALIGN_000521	5.7	135	11	105	107
3	ALIGN_000832	10	327	50	49	315
4	ALIGN_000205	4.3	15	6	18	9
5	ALIGN_000297	31.4	315	258	315	315
6	ALIGN_000397	24.3	2745	328	10395	10206
7	ALIGN_000398	0	477	20	815	77
8	ALIGN_000623	0	380	20	10395	33
9	ALIGN_000628	0	141	5	117	21
10	ALIGN_000767	4.3	15	6	9	9
11	ALIGN_000771	14.3	81	9	10	15
12	ALIGN_000788	24.3	945	80	225	135
13	ALIGN_000853	12.9	225	20	225	45
14	ALIGN_000930	78.6	10395	8925	10395	10393
15	ALIGN_000931	100	10395	9876	10395	10395
16	ALIGN_000984	45.7	10395	2344	10395	10391

7. Conclusion. In this paper, we have presented a general method for creating phylogenetic metrics. We have considered only binary (fully resolved) trees; however, the method can be used for comparing non-binary trees as well. Another advantage of the method is its time complexity, which is polynomial (under the condition that both functions f and h can be computed in polynomial time) in contrast with, for example, most metrics based on edit operations like NNI or TBR [7], [1]. Since the method is based on matching, it can also be used for visualizing similarities and differences between phylogenetic trees (similar to the method described in [15]).

We have defined and investigated some basic properties and advantages of a special case of the general definition, called the Matching Split metric. We have also indicated basic differences between the MS distance and the well-known RF metric.

We have performed computational experiments, whose results indicate the possibility of application of some of the metrics to estimating a phylogenetic signal quality. The experiments have been carried out for data sets containing 8-taxon trees. The question arises if the results can be generalized to bigger trees. Another question concerns the possibility of normalizing the values to make them independent of the number of leaves.

Acknowledgements. We thank Krzysztof Giaro for many fruitful comments. We also thank Aleksandra Czarna and Borys Wróbel for providing us with data for testing our metrics and for discussions concerning the connections between the values obtained and the phylogenetic signal quality. The anonymous reviewer gave several very useful comments that led to improvements in the article.

This paper was partially presented at the XVth National Conference on Applications of Mathematics in Biology and Medicine.

References

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

[2] S. P. Blomberg and T. Garland, *Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods*, J. Evolutionary Biol. 15 (2002), 899–910.

[3] J. Bluis and D.-G. Shin, *Nodal distance algorithm: calculating a phylogenetic tree comparison metric*, in: BIBE 03: Proc. 3rd IEEE Symposium on BioInformatics and BioEngineering, 2003, 87–94.

[4] D. Bryant, *The splits in the neighborhood tree*, Ann. Combin. 8 (2004), 1–11.

[5] —, *Building trees, hunting for trees, and comparing trees – theory and methods in phylogenetic analysis*, Ph.D. Thesis, Department of Mathematics, Univ. of Canterbury, 1997.

[6] A. Czarna, R. Sanjuán, F. González-Candelas, and B. Wróbel, *Topology testing of phylogenies using least squares methods*, BMC Evolutionary Biol. 6 (2006), 105.

[7] B. Dasgupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang, *On computing the nearest neighbor interchange distance*, in: Proc. DIMACS Workshop on Discrete Problems with Medical Applications, 2000, 125–143.

[8] W. K. Hastings, *Monte Carlo sampling methods using Markov chains*, Biometrika 57 (1970), 97–109.

[9] G. Hickey, F. Dehne, A. Rau-Chaplin, and C. Blouin, *SPR distance computation for unrooted trees*, Evolutionary Bioinformatics 4 (2008), 17–27.

[10] D. M. Hillis, T. A. Heath, and K. St. John, *Analysis and visualization of tree space*, Systematic Biol. 54 (2005), 471–482.

[11] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback, *Bayesian inference of phylogeny and its impact on evolutionary biology*, Science 294 (2001), 2310–2314.

[12] M.-Y. Kao, T. W. Lam, W. K. Sung, and H. F. Ting, *All-cavity maximum matchings*, in: ISAAC 97: Proc. 8th International Symposium on Algorithms and Computation, 1997, 364–373.

[13] J. B. Losos, *Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species*, Ecology Lett. 11 (2008), 995–1007.

[14] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys. 21 (1953), 1087–1092.

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

[16] D. Penny, L. R. Foulds, M. D. Hendy, *Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences*, Nature 297 (1982), 197–200.

[17] B. Rannala, *Identifiability of parameters in MCMC Bayesian inference of phylogeny*, Systematic Biol. 51 (2002),754–760.

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

[19] F. Ronquist and J. P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models*, Bioinformatics 19 (2003), 1572–1574.

[20] M. A. Steel and D. Penny, *Distributions of tree comparison metrics—some new results*, Systematic Biol. 42 (1993), 126–141.

[21] M. Stissing, C. N. S. Pedersen, T. Mailund, G. S. Brodal, and R. Fagerberg, *Computing the quartet distance between evolutionary trees of bounded degree*, in: Proc. 5th Asia-Pacific Bioinformatics Conference, 2007, 101–110.

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

[23] L. Tierney, *Markov chains for exploring posterior distributions*, Ann. Statist. 22 (1994), 1701-1728.

[24] D. M. de Vienne, T. Giraud, and O. C. Martin, *A congruence index for testing topological similarity between trees*, Bioinformatics 23 (2007), 3119–3124.

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

[26] Z. Yang, *PAML 4: Phylogenetic analysis by maximum likelihood*, Molecular Biol. Evolution 24 (2007), 1586–1591.

Damian Bogdanowicz
Department of Algorithms and System Modeling
Gdańsk University of Technology
Narutowicza 11/12
80-233 Gdańsk, Poland
E-mail: Damian.Bogdanowicz@eti.pg.gda.pl

Received on 17.5.2010;
revised version on 21.10.2010 (2049)