New metrics for phylogenies

Łukasz Bolikowski
Interdisciplinary Center for Mathematical and Computational Modeling, Warsaw University

Anna Gambin
Institute of Informatics, Warsaw University

March 23, 2006

Abstract

In this paper we propose two new metrics on the space of phylogenetic trees. The problem of determining how distant two trees are from each other is crucial because many various methods have been proposed for reconstructing phylogenetic trees from data. These techniques (in fact often heuristics) applied to the same data results in different trees. We investigate the basic properties of new metrics and present an efficient algorithms approximating the distance between two trees. Experiments, which has been performed for randomly generated trees justify the applicability of our algorithms.

1 Introduction and preliminaries

The phylogenetic trees are used to illustrate the evolutionary history of current species. In molecular phylogenetic trees are reconstructed from molecular data, i.e. fragments of sequenced genomes. There are various methods proposed for tree reconstruction problem (see e.g. (10)). Also various mathematical models have been proposed to describe molecular sequence evolution (9). Hence, often for the same set of species there exist several non-identical evolutionary histories resulting from different evolutionary models and tree construction algorithms.

In this setting the question of similarity measure for trees arises naturally. However, there is no one, widely accepted approach. A number of tree comparison metrics have been proposed (8; 11; 13; 14). Most of them are computationally inefficient, i.e. computing the distance is proved to be NP-hard problem (see (4) for the NP-hardness proof for Nearest Neigbor Interchange distance). In other cases, like Subtree Prunning and Regrafting (SPR) metric, the question about the computational complexity of computing the distance remains open (2).

Results. In this paper we propose new metrics on the space of phylogenetic trees. They can be considered as a generalization of Maximum Agreement

Subtree (7) and Maximum Agreement Forest (12) metrics. The usefulness of our approach can be argued as follows:

- it can be used to estimate the divergence inside the family of trees e.g. by calculating all-pairs distances.
- it can provide consensus measure that summarizes the evolutionary history common to multiple alternative trees, i.e. it can be generalized to a discrepancy measure for a set of trees.
- the algorithm for distance approximation can be effectively parallelized.

Organization of the paper. In this section we introduce the notation and prove several facts necesary for definition of new metrics. Section 2 is devoted to a description of metrics and efficient algorithms approximating the distance between trees. Section 3 discuss the outcomes of several computational experiments perfomed on randomly generated trees.

Preliminaries. In the sequel, we will consider rooted binary phylogenetic trees. It is worth to mention that all the results could be restated for unrooted trees. A tree labeled by elements of Σ is defined as an arbitrary set $T \subseteq 2^{\Sigma}$ such that:

(i) $\emptyset \in T,\ \Sigma \in T$

(ii) $\forall_{\sigma \in \Sigma}\ \{\sigma\} \in T$

(iii) $t_1, t_2 \in T \Rightarrow (t_1 \cap t_2 = \emptyset) \vee (t_1 \subseteq t_2) \vee (t_2 \subseteq t_1)$

We call the elements of $T \setminus \emptyset$, *nodes* of the tree. The node t_2 is *ancestor* of the node t_1 iff $t_1 \subseteq t_2$.

While comparing trees we always assume that they have common set of leaves. We denote by $L(T)$ the set of leaves of the tree T.

Induced subtrees. We define the subtree of T induced by the subset of leaves L as follows:

$$T_{|L} = \{t \cap L \mid t \in T\}$$

The induced subtree $T_{|L}$ are also called the **restriction** of T on L and the **homeomorphic subtree** of T on L. Define also the **contraction** of T as a tree obtained by deleting the internal edge and identifying its endpoints. We say that the tree T_B **extends** the tree T_A, denoted $T_A \trianglelefteq T_B$, if T_A is an induced subtree of a contraction of T_B.

Agreement partitions. Our metric is based on the notion of *agreement partition.* This is a partition on the set of leaves such that the subtrees induced by its blocks coincide in both trees.

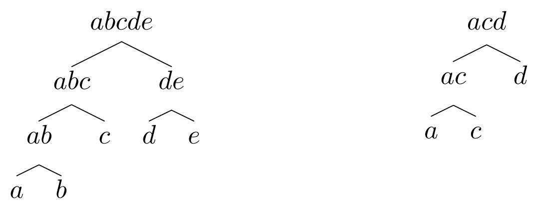

Figure 1: The subtree of T (on the left) induced by a set $L = \{a, c, d\}$ $T_{|L}$ (on the right)

Definition 1 *For given two trees T_A and T_B the subset of leaves L is called an* **agreement set** *if*

$$T_{A|L} = T_{B|L}$$

The subtree induced (in both trees) by the agreement set is called an **agreement subtree**.

The Maximum agreement subtree (MAST), i.e. the agreement subtree induced by the agreement set of maximal cardinality, is widely used as a consensus measure for phylogenetic trees.

Definition 2 *For given two trees T_A and T_B (on the same set of leaves Σ), the set $\mathcal{P} = \{L_1, \ldots, L_k\}$, where $L_i \subseteq \Sigma$ for $i \in \{1, \ldots k\}$, is an* **agreement partition** *of T_A and T_B, if the following are fulfilled:*

(i) $i \neq j \Rightarrow L_i \cap L_j = \emptyset$

(ii) $\bigcup_i L_i = \Sigma$

(iii) $\forall_i \; T_{A|L_i} = T_{B|L_i}$

Definition 3 *For two partitions (not necessary agreement partitions) $\mathcal{P}_1$ and $\mathcal{P}_2$ a* **product partition** *is defined as*

$$\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2 = \{L_1 \cap L_2 \mid L_1 \in \mathcal{P}_1 \wedge L_2 \in \mathcal{P}_2\}$$

Notice, that if at least one of the partition $\mathcal{P}_1$ or $\mathcal{P}_2$ is an agreement partition then $\mathcal{P}_1 \times \mathcal{P}_2$ is also agreement partition.

Rooted triples. We introduce the notion of **rooted triples**. The notation used is adopted from (3).

Definition 4 *A* **rooted triple** *is an arbitrary binary phylogenetic 3-leaf tree.*

Hence, each rooted triple is in one of the following forms:

$$r = \{\{a, b, c\}, \{a, b\}, \{a\}, \{b\}, \{c\}, \emptyset\}$$

For the simplicity reason we use also the natural notation $(a, b \,|\, c)$ or $ab|c$. We denote by $r(T)$ the set of all rooted triples induced by the tree T ($|r(T)| = \binom{|L(T)|}{3}$). For a given set of rooted triples R let

$$L(R) = \bigcup_{r \in R} L(r)$$

Similarly for a given partition $\mathcal{P}$ let

$$L(\mathcal{P}) = \bigcup_{L' \in \mathcal{P}} L'$$

Definition 5 *A set of rooted triples is called* **inconsistent** *if there is no tree T such that $R \subseteq r(T)$.*

The following facts will be useful in the next Section.

Fact 1 *For two trees T_A, T_B (on the same leaf set Σ) we have:*

(i) $r(T_A) = r(T_B)$ *iff* $T_A = T_B$.

(ii) *if* $T_A \trianglelefteq T_B$ *then* $r(T_A) \subseteq r(T_B)$.

(iii) *if for* $L \subseteq \Sigma$ *we have* $T_{A|L} = T_{B|L}$, *then* $r(T_{A|L}) = r(T_{B|L}) \subseteq r(T_A) \cap r(T_B)$.

Proof: *(i)*($\Rightarrow$) Let $\Sigma = L(T_A) = L(T_B)$. If $|\Sigma| \leq 2$, then the implication holds since there is exactly one tree T such that $L(T) = \Sigma$. When $|\Sigma| \geq 3$, then suppose $r(T_A) \neq r(T_B)$. Without loss of generality, let's choose r' such that $r' \in r(T_A)$ and $r' \notin r(T_B)$. We have $T_{A|L(r')} \neq T_{B|L(r')}$, so $T_A \neq T_B$.

($\Leftarrow$) If $T_A = T_B$, then for each L (in particular for $|L| = 3$) we have $T_{A|L} = T_{B|L}$, so $r(T_A) = r(T_B)$.
(ii) Let $r' \in r(T_A)$. Since $L(r') \subseteq L(T_A)$ and $T_A = T_{A|L(T_A)} = T_{B|L(T_A)}$, then also $T_{A|L(r')} = T_{B|L(r')}$. Therefore $r' \in r(T_B)$.
(iii) Assume that we have $r' \in r(T_{A|L}) = r(T_{B|L})$ such that $r' \notin r(T_A) \cap r(T_B)$. Further, without loss of generality, let's assume that $r' \notin r(T_A)$. On the other hand, $r' \in r(T_{A|L}) \subseteq r(T_A)$. We have reached a contradiction, so the first assumption was wrong. Therefore,

$$\forall_{r'}\ r' \in r(T_{A|L}) = r(T_{B|L}) \Rightarrow r' \in r(T_A) \cap r(T_B)$$

$\square$

The following corollary is an immediate conclusion of Fact 1 (iii).

Corollary 1 *For any agreement partition $\mathcal{P}$ of trees T_A and T_B holds*

$$r(\mathcal{P}) \subseteq r(T_A) \cap r(T_B)$$

2 Metric induced by partitions

For given two phylogenetic trees T_A and T_B let $\delta(T_A, T_B)$ be the cardinality of their **Minimal Agreement Partition** (MAP), i.e. the smallest number of blocks in the agreement partition. To ensure the triangle inequality we define the **partition metric** as:

$$\lceil \log_2 \delta(T_A, T_B) \rceil$$

where $\lceil x \rceil$ for any real number x denote the greatest integer which is less than or equal to x.

We have studied also the second metric. It can be defined as a path metric in the graph $\mathcal{G}_\Sigma$ having as vertices all phylogenetic trees with leaves labeled bijectively by elements of Σ. Two trees T_A and T_B are incident to each other iff there exists for these trees an agreement partition with two blocks. The distance between two trees can be defined as the length of the shortest path connecting them in the graph $\mathcal{G}_\Sigma$. The obvious fact that the graph $\mathcal{G}_\Sigma$ is connected guarantees the correctness of this definition. We call this metric a **path metrics**, and denote it by γ.

Theorem 1 *For any two trees T_A and T_B, on the same leaf set the following inequality holds: $\delta(T_A, T_B) \leq \gamma(T_A, T_B)$*

Proof: Let $(T_A = T_0, T_1, T_2, \ldots, T_k = T_B)$ be the shortest path in graph $\mathcal{G}_\Sigma$ connecting T_A and T_B. Let $L_1, \ldots, L_k$ be a series of two-block agreement partitions such that L_i is a two-block agreement partition for T_{i-1} and T_i. Then the partition $L = L_1 \times L_2 \times \cdots \times L_k$ is an agreement partition for T_A and T_B. We have:

$$|L| \leq |L_1| \times |L_2| \times \cdots \times |L_k| = 2 \times 2 \times \cdots \times 2 = 2^k$$

so $\lceil \log_2 |L| \rceil \leq k$ and therefore $\delta(T_A, T_B) \leq k = \gamma(T_A, T_B)$. □

As a conjecture we postulate the equivalence of two metrics and therefore in the sequel we concentrate on the **partition metric**.

Recall, that the algoritm computing the Maximum Agreement Forest of two trees selects the minimal subset of edges in both trees that have to be cut for obtained forests to coincide (see (12) for details).

While calculating **partition metric** we are looking for the minimal partition on the leaf-set such that the subtrees induced by blocks, in both trees, form isomorphic forests. Intuitively, the blocks in the agreement partition for two trees group together the species sharing the same topology of evolutionary history. Smaller number of blocks correspond to less divergent trees. The following inequalities hold:

$$1 \geq \delta(T_A, T_B) \geq \lceil \frac{n}{2} \rceil$$

where $\delta(T_A, T_B) = 1$ iff $T_A = T_B$.

In this Section we present efficient algorithms giving the lower and upper bounds for the MAP metric. The interesting open question is the computational complexity of the MAP distance. However, experiments show that our approximations for partition metric perform much better than 3-approximation algorithm for Maximum Agreement Forest (12). Very often for the pair of phylogenies lower and upper bound provided by our algorithms almost coincide.

Lower bound for MAP.

Form Corollary 1 we know that $R = r(T_A) \cap r(T_B)$ contains all informations sufficient to reconstruct any agreement partition for trees T_A and T_B.

The algorithm LabelledOneTree below is a modified version of Aho et al. algorithm OneTree from (1). It takes as an input two arguments: the set of rooted triples R and the set of leaf labels Σ, checks whether R is **consistent** and calculates the lower bound for $\delta(T_A, T_B)$, i.e. the lower bound for the minimum number of blocks in the agreement partition of T_A and T_B.

LabelledOneTree(R, Σ)

if $|\Sigma| \leq 2$ **then return** 1
$G_{R,\Sigma} = \langle \Sigma, E \rangle$ such that $\Sigma^2 \ni (\sigma_1, \sigma_2) \in E \iff \exists_{\sigma_3 \in \Sigma}\ \sigma_1\sigma_2|\sigma_3 \in R$
if $G_{R,\Sigma}$ is connected **then return** Inconsistent
$lbsum = 0$; $lbmax = 0$
foreach connected component G' of the graph $G_{R,\Sigma}$
    $lb =$ LabelledOneTree$(R, vertexes(G'))$
    if lb = Inconsistent **then return** Inconsistent
    if $lbmax < lb$ **then** $lbmax = lb$
    $lbsum = lbsum + lb$
return $\max(\lceil lbsum/2 \rceil, lbmax)$

We use this algorithm to compute the lower bound for the minimal number of blocks in the agreement partition for trees T_A and T_B as follows:

LowerBound(T_A, T_B) = LabelledOneTree$(r(T_A) \cap r(T_B), L(T_A) \cap L(T_B))$

Example. This example illustrates how the algorithm LowerBound works. Consider two trees T_A and T_B with leaves labeled by $\Sigma = \{a, b, c, d, e, f, g, h\}$ from the Fig 2. The set of rooted triples R that belong to both trees looks as follows:

$$
\begin{aligned}
R = r(T_A) \cap r(T_B) \quad = \quad & \{ac|b, de|a, de|c, df|a, df|b, df|c, ef|a, ef|c, de|g, df|g, \\
& ef|g, dh|a, dh|b, eh|a, fh|a, fh|b, dh|g, eh|g, fh|g\}
\end{aligned}
$$

The graph $G_{R,\Sigma}$ (see Fig. 3) has four connected components: $\{a, c\}$, $\{b\}$, $\{d, e, f, h\}$, $\{g\}$. The only one non-trivial connected component is $\{d, e, f, h\}$. We run re-

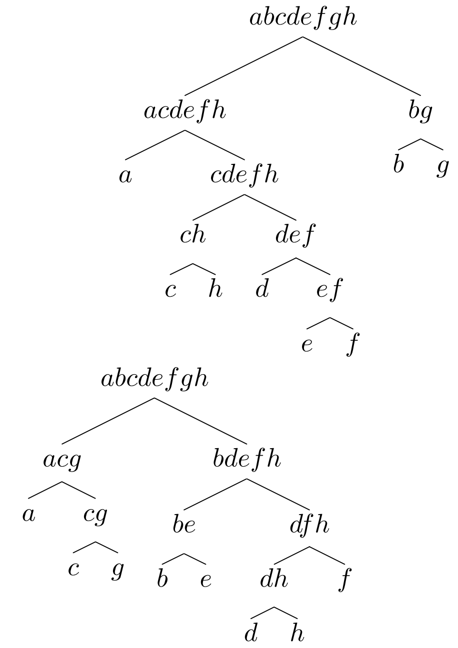

Figure 2: Algorithm for lower bound: an example

cursively the algorithm for this component and we obtain the graph $G_{R,\{d,e,f,h\}}$ consisting of isolated vertexes (i.e. having 4 connected components). Hence, the estimation of the number of blocks in MAP is: $s_{defh} = 2$, $s_{ac} = 1$, $s_b = 1$, $s_g = 1$, and finally: $s_{abcdefgh} = \lceil \frac{2+1+1+1}{2} \rceil = 3$. The following theorem summarizes the applicability of the algorithm LowerBound:

Theorem 2 *For rooted binary phylogenetic trees T_A, T_B holds*

$$\textsc{LowerBound}(T_A, T_B) \leq \delta(T_A, T_B)$$

To prove our Theorem we need following two Lemmas:

Lemma 1 *For any R and Σ, such that $|\Sigma| \geq 3$, if the graph $G_{R,\Sigma}$ constructed by the algorithm LabelledOneTree is connected, then the set of rooted triples R is inconsistent.*

Proof: Assume that R is consistent. Then there is a tree T such that $R \in r(T)$ and $L(R) = L(T)$. Let t_1 and t_2 be two different children of the root of T. Note that for the subtrees $T_{|t_1}$ and $T_{|t_2}$ the following property holds:

$$\forall_{a \in t_1} \forall_{b \in t_2} \neg \exists_{c \in L(T)} \; ab|c \in r(T)$$

Therefore t_1 and t_2 are located in different connected components of the graph $G_{r(T),L(T)}$, hence the graph is not connected. Since $R \subseteq r(T)$ then also the original graph $G_{R,\Sigma}$ is not connected. $\square$

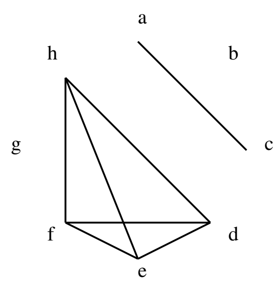

Figure 3: Graph $G_{R,\Sigma}$ for $R = r(T_A) \cap r(T_B)$

Lemma 2 *For any R and Σ, if each of the leaves $a, b, c \in \Sigma$ belongs to different connected component of the graph $G_{R,\Sigma}$, then there is no rooted triple $r \in R$ such that $L(r) = \{a, b, c\}$.*

Proof: Let's assume there exists a rooted triple $r \in R$ such that $L(r) = \{a, b, c\}$. Without loss of generality let $r = ab|c$. Hence, from the definition of the graph $G_{R,\Sigma}$ in line 2 of LabelledOneTree, a and b are connected by an edge, so these two vertices are located in the same connected component of the graph $G_{R,\Sigma}$. This is a contradiction, so there is no rooted triple $r \in R$ such that $L(r) = \{a, b, c\}$. □

Proof of the Theorem: We can deduce that does not exist a block $B \subseteq \Sigma$ of an agreement partition, which has elements belonging to three different connected components of the graph $G_{\cdot,B}$. To obtain the lower bound estimation for the minimal number of blocks in an agreement partition we make an optimistic assumption that, if for arbitrary two connected components G_1 and G_2 of the graph $G_{\cdot,\cdot}$ we have two blocks B_1 and B_2 in some agreement partition such that $B_1 \subseteq vertices(G_1)$ and $B_2 \subseteq vertices(G_2)$, then $B_1 \cap B_2$ also forms a block in some agreement partition.

The algorithm is based on this assumption: for each connected component of $G_{R,\Sigma}$ we calculate the minimal number of blocks in some agreement partition and we join the blocks belonging to different connected components.

Assume that our graph $G_{R,\Sigma}$ consists of k connected components. We run the algorithm recursively for all of them and obtain the sequence $(s_1, \ldots, s_k)$ of estimates of the number of blocks.

If there exists i such that $s_i > \sum_{j \neq i} s_j$ then the minimal number of blocks in some agreement partition is bounded below by $s = s_i$. If there is no such i that $s_i > \sum_{j \neq i} s_j$ then all blocks can be paired except one (depending on the parity of $\sum_i s_i$). In this case we estimate the minimal number of blocks as: $s = \lceil \frac{1}{2} \sum_i s_i \rceil$. □

Upper bound for MAP. We are also interested in the upper bound for the minimum number of blocks in the agreement partition for two trees T_A and

T_B. As a subroutine in our heuristics we use an algorithm that finds for two trees the maximum agreement subtree (MAST), i.e the agreement subtree of the maximal size. MAST approach belongs to the methods of reconciling different evolutionary trees for the same set of species (leaves). For two binary phylogenetic trees their MAST can be find in the time polynomial w.r.t. the size of the tree (see (15) for details).

Here we present an algorithm that for a given set of rooted triples $R = r(T_A) \cap r(T_B)$ seeks for the maximum size tree T such that $r(T) \subseteq R$, i.e.: $T = T_{A|L(T)} = T_{B|L(T)}$ is the MAST of T_A and T_B.

```
MAST*(a, b, R)
    if a = b then return {{a}, ∅}
    A = {x | (a, x | b) ∈ R} ∪ {a}
    B = {y | (b, y | a) ∈ R} ∪ {b}
    T1 = max_{x∈A} MAST*(a, x, R)
    T2 = max_{y∈B} MAST*(b, y, R)
    return T1 ∪ T2 ∪ {L(T1) ∪ L(T2)}
```

The procedure MAST* from (3) finds the largest agreement subtree T having the property that the leaves a and b belong to different subtrees (i.e. $|\{t \in T \mid \{a,b\} \subseteq t\}| = 1$). The MAST for T_A and T_B can be found by running the algorithm MAST* for all pairs of leaves from $L = L(T_A) \cap L(T_B)$.

```
MAST(T_A, T_B)
    R = r(T_A) ∩ r(T_B)
    L = L(T_A) ∩ L(T_B)
    mast = ∅
    foreach a ∈ L
        foreach b ∈ L
            if |mast| < |MAST*(a, b, R)| then mast =
            MAST*(a, b, R)
    return mast
```

The algorithm UpperBound(T_A, T_B) finds some agreement partition for two input trees T_A and T_B. It uses the greedy approach by iterating the procedure that finds MAST for two trees.

UpperBound(T_A, T_B)

$L = L(T_A) \cap L(T_B)$
$\mathcal{P} = \emptyset$
while $L \neq \emptyset$
  $M = \mathrm{MAST}(T_{A|L}, T_{B|L})$
  $\mathcal{P} = \mathcal{P} \cup \{M\}$
  $L = L \setminus L(M)$
return $\mathcal{P}$

It is easy to see that the algorithm UpperBound(T_A, T_B) approximates the partition metric from above by finding agreement partition with blocks being MASTs. Unfortunately this algorithm does not calculate the metric exactly. The counterexample is straightforward to construct.

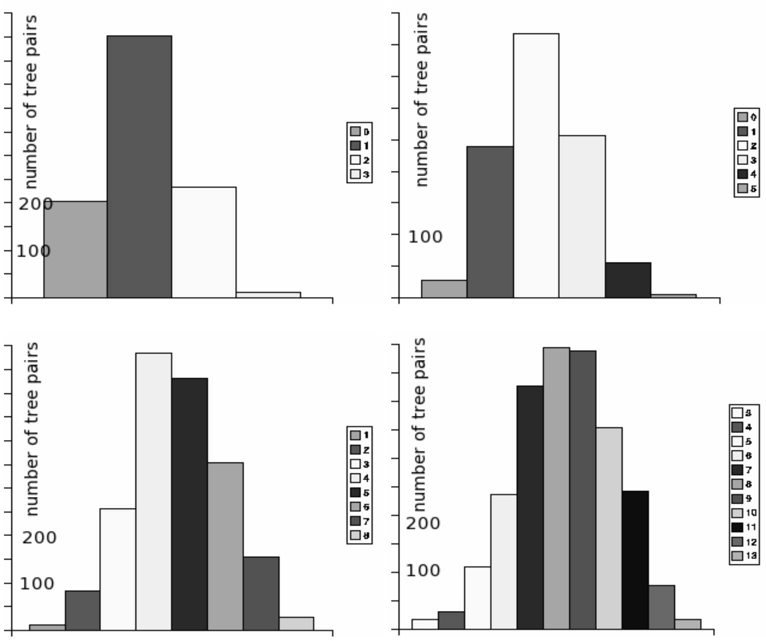

Figure 4: Histograms of differences between lower and upper bounds for MAP metric for 1000 pairs of random trees with 20, 40, 100 and 200 leaves

3 Experimental results

In this Section we present the estimates given by the algorithms LowerBound and UpperBound for the partition metric.

Instead of the log-transformed metric value we give here the estimates for the minimal number of blocks in the agreement partition. Recall that if k is the minimal number of blocks in some agreement partition for two given trees then $\lceil \log_2 k \rceil$ is the distance between these trees under the partition metrics. We denote by (xy) the pairs of lower (x) and upper (y) bounds for the number of blocks.

We have run both algorithms for 1000 pairs of trees generated in Yule model of size 20, 40, 100 and 200 (c.f. Fig. 4). Notice that for small but still reasonable size of trees (i.e. 20 and 40 leaves) our estimates almost coincide.

4 Conclusions and open problems

The outcomes of the algorithms calculating lower and upper bound of partition metric are very promising. For medium size trees (i.e. up to 40 leaves) our bounds almost coincide. This fact is especially significant comparing with the best approximation algorithm for Maximum Agreement Forest, which has the approximation factor 3.

The most challenging open problem is to determine the complexity of computing the partition distance. It would be also interesting to obtain the analytical estimations for the average distance between two random trees, analogous to the formula given for MAST in (6).

References

[1] Aho A. V. et. al (1981), Inferring a tree from lowest common ancestors with an application to the optimization o relational expressions, SIAM Journal of Computing, 10(3):405–421.

[2] Allen B. L., Steel M. (2001), Subtree Transfer Operations and their Induced Metrics on Evolutionary Trees *Annals of Combinatorics* 5:1–13.

[3] Bryant D. (1997), Building trees, hunting for trees, and comparing trees, *Ph.D. thesis, University of Canterbury, New Zealand.*

[4] DasGupta B. et al. (2000), On computing the nearest neighbor interchange distance, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, 55:125–143, Amer. Math. Soc., Providence, RI

[5] Gambin A., Lasota S., Szklarczyk R., Tiuryn J., Tyszkiewicz J. (2002), Contextual Alignment of Biological Sequences, *Proc. ECCB'02, Bioinformatics* 18:116–127, Oxford University Press.

[6] Gambin A., Slonimski P. P. (2005), Hierarchical clustering based upon contextual alignment of proteins: a different way to approach phylogeny, *C.R. Biologie Acad. Science Paris*, 328:11-22.

[7] Kubicka, E., Kubicki, G., McMorris, F.R. An algorithm to find agreement subtrees (1995), *Journal of Classification*,12:91-99.

[8] Li, M., Tromp, J., Zhang, L. (1995)On the nearest neighbour interchange distance between evolutionary trees. *J. Theor. Biol.* 182:463-467.

[9] Li, W-H. (1997)*Molecular Evolution*,Sinauer Associates.

[10] Page R. D. M., Holmes E. C. (1998), *Molecular Evolution: A phylogenetic approach*, Blackwell Scientific, Oxford.

[11] Robinson, D.F., Foulds, L.R. (1981) Comparison of phylogenetic trees, *Math. Biosci.*53:131-147.

[12] Rodrigues E. M., Sagot M-F. and Wakabayashi Y.(2001) Some approximation results for the maximum agreement forest problem, *in Proc. APPROX and RANDOM'01* LNCS, vol. 2129:159–169.

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

[14] D. H. Huson and M. Steel. (2004) Distances that perfectly mislead, *Systematic Biology* 53(2): 327–332.

[15] Steel M., Warnow T. (1993), Kaikoura Tree Theorems: Computing the Maximum Agreement Subtree, *Inf. Process. Lett.* 48(2): 77–82.

[16] Tatusov R. L. et al (2001), The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 21(1):22–28.