Earth Mover's Distances for Rooted Labaled Unordered Trees based on Tai Mapping Hierarchy

Taiga Kawaguchi and Kouichi Hirata
Kyushu Institute of Technology, Kawazu 680-4, Iizuka 820-8502, Japan

Keywords: Earth Mover'S Distance, Rooted Labeled Unordered Tree, Tai Mapping Hierarchy, Tree Edit Distance.

Abstract: In this paper, we introduce earth mover's distances (EMDs, for short) for rooted labeled trees based on Tai mapping hierarchy. First, by focusing on the restricted mappings in the Tai mapping hierarchy providing the tractable variations of the tree edit distance, we formulate the EMDs whose signatures are all of the pairs of a complete subtree and its frequency and whose ground distances are the tractable variations. Then, we compare the EMDs with their ground distances, which are tractable variations.

1 INTRODUCTION

Comparing tree-structured data such as HTML and XML data for web mining or DNA and glycan data for bioinformatics is one of the important tasks for data mining. The most famous distance measure between *rooted labeled unordered trees* (*trees*, for short) is the *edit distance* (Tai, 1979). The edit distance is formulated as the minimum cost of *edit operations*, consisting of a *substitution*, a *deletion* and an *insertion*, applied to transform from a tree to another tree. Whereas the edit distance is a metric, the problem of computing the edit distance is MAX SNP-hard even if trees are binary (Hirata et al., 2011; Zhang and Jiang, 1994).

As constant-factor lower bounding distances of the edit distance, several *histogram distances* based on local information (Aratsu et al., 2009; Kailing et al., 2004; Li et al., 2013) have introduced. Whereas we can compute them more efficiently than the edit distance, none of them is a metric.

On the other hand, an *earth mover's distance* (*EMD*, for short) has originally developed to compare with two images in image retrieval and pattern recognition (Rubner et al., 2007) and is formulated as the solution of the transportation problem between the distributions of features in signatures in two images. It is known that the EMD is a metric if so is the ground distance between single features.

Gollapudi and Panigrahy (Gollapudi and Panigrahy, 2008) have extended the EMD to that between two leaf-labeled trees with the same height, where a tree is leaf-labeled if all of the labels are assigned to just leaves. However, it is difficult for the EMD to extend to be applicable to standard two trees, that is, labels are assigned to all the nodes and having possible different height as follows. In the EMD, first, by comparing each pair of leaves (that is, the nodes with height 1), we set the value 1 if both leaves have the same label and 0 otherwise. Then, by using the information between the pair of nodes in the height $k-1$, we solve the transportation problem of the pair of nodes in the height k. Hence, in order to apply such a recursion to trees, the trees are necessary to have the same height and have no internal nodes with labels.

Kawaguchi and Hirata (Kawaguchi and Hirata, 2017) have introduced another EMD based on complete subtrees. The EMD is formulated by the histograms consisting of either complete subtrees, cocomplete subtree or both and their frequencies as signatures and the L_1-distance between the histograms as ground distances, so we can apply the EMD to rooted labeled trees. Also the EMD is a metric and tractable. On the other hand, there exist trees that the EMD cannot reflect intuitive similarity.

Since the edit distance between trees is corresponding to a Tai mapping (Tai, 1979), many variations of the edit distance have developed as more structurally sensitive distances obtained by restricting the Tai mapping, that is, a top-down distance (Chawathe, 1999; Selkow, 1977), an LCA- and root-preserving distance (Yoshino and Hirata, 2017), an LCA-preserving distance (Zhang et al., 1996), an accordant distance (Kuboyama, 2007), an isolated-subtree (or a constrained) distance (Zhang, 1995; Zhang, 1996) and an alignment distance (Jiang et al.,

Kawaguchi, T. and Hirata, K.
Earth Mover's Distances for Rooted Labaled Unordered Trees based on Tai Mapping Hierarchy.
DOI: 10.5220/0006633701590168
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 159-168
ISBN: 978-989-758-276-9

1995). Almost variations are metrics except an alignment distance (Jiang et al., 1995). Also, whereas the problem of computing the edit distance or the alignment distance between trees is MAX SNP-hard (Hirata et al., 2011; Jiang et al., 1995; Zhang and Jiang, 1994), the problem of computing the other variations is tractable.

The reason why these variations are tractable is that the maximum weight bipartite matching problem can be applied to computing the variations after decomposing trees from the root (Yamamoto et al., 2014). In contrast, it cannot be applied to computing the edit distance and the alignment distance, because computing them is necessary to compare the decomposed trees and the remained trees after decomposing trees from the root.

Since we can regard the minimum weighted bipartite problem as a special case of the transportation problem in EMDs, in this paper, we formulate new EMDs based on the Tai mapping hierarchy whose signatures are pairs of a complete subtree and the ratio of frequencies occurring in a whole tree and whose ground distances are the tractable variations of the edit distance. Then, we show that the EMDs are always metrics and tractable. Finally, we give experimental results to evaluate the EMDs to compare them with their ground distances and investigate the properties of the EMDs.

2 PRELIMINARIES

A *tree* T is a connected graph (V,E) without cycles, where V is the set of vertices and E is the set of edges. We denote V and E by $V(T)$ and $E(T)$. The *size* of T is $|V|$ and denoted by $|T|$. We sometime denote $v \in V(T)$ by $v \in T$. We denote an empty tree $(\emptyset,\emptyset)$ by $\emptyset$. A *rooted tree* is a tree with one node r chosen as its *root*. We denote the root of a rooted tree T by $r(T)$.

For each node v in a rooted tree with the root r, let $UP_r(v)$ be the unique path from v to r. The *parent* of $v(\neq r)$, which we denote by $par(v)$, is its adjacent node on $UP_r(v)$ and the *ancestors* of $v(\neq r)$ are the nodes on $UP_r(v) - \{v\}$. We denote the set of all ancestors of v by $anc(v)$. We say that u is a *child* of v if v is the parent of u and u is a *descendant* of v if v is an ancestor of u. We use the ancestor orders $<$ and $\leq$, that is, $u < v$ if v is an ancestor of u and $u \leq v$ if $u < v$ or $u = v$. We say that w is the *least common ancestor* of u and v, denoted by $u \sqcup v$, if $u \leq w$, $v \leq w$ and there exists no w' such that $w' \leq w$, $u \leq w'$ and $v \leq w'$. is the number of children of v. The *degree* of a rooted tree T, denoted by $d(T)$, is the maximum number of $d(v)$ for every $v \in T$.

For nodes $u, v \in T$, u is *to the left of* v if $pre(u) \leq pre(v)$ for the preorder number *pre* and $post(u) \leq post(v)$ for the postorder number *post*. We say that a rooted tree is *ordered* if a left-to-right order among siblings is given; *unordered* otherwise. We say that a rooted tree is *labeled* if each node is assigned a symbol from a fixed finite alphabet Σ. For a node v, we denote the label of v by $l(v)$, and sometimes identify v with $l(v)$. In this paper, we call a rooted labeled unordered tree a *tree* simply.

Let T be a tree (V,E) and v a node in T. A *complete subtree of T at v*, denoted by $T[v]$, is a tree $T' = (V',E')$ such that $r(T') = v$, $V' = \{u \in V \mid u \leq v\}$ and $E' = \{(u,w) \in E \mid u,w \in V'\}$. We denote the (multi)set $\{T[v] \mid v \in T\}$ of all the complete subtrees in T by $cs(T)$. For a complete subtree S in T, we denote the frequency of the occurrences of S in T by $f(S,T)$.

Next, we introduce an *edit distance* and a *Tai mapping*.

Definition 1 (Edit operations (Tai, 1979))**.** The *edit operations* of a tree T are defined as follows. (Figure 1).

1. *Substitution*: Change the label of the node v in T.
2. *Deletion*: Delete a node v in T with parent v', making the children of v become the children of v'. The children are inserted in the place of v as a subset of the children of v'. In particular, if v is the root in T, then the result applying the deletion is a forest consisting of the children of the root.
3. *Insertion*: The complement of deletion. Insert a node v as a child of v' in T making v the parent of a subset of the children of v'.

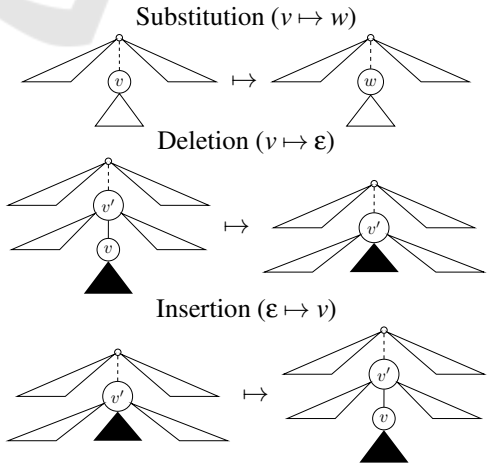


Figure 1: Edit operations for trees.

Let $\varepsilon \notin \Sigma$ denote a special *blank* symbol and define $\Sigma_\varepsilon = \Sigma \cup \{\varepsilon\}$. Then, we represent each edit operation

by $(l_1 \mapsto l_2)$, where $(l_1, l_2) \in (\Sigma_\varepsilon \times \Sigma_\varepsilon - \{(\varepsilon, \varepsilon)\})$. The operation is a substitution if $l_1 \neq \varepsilon$ and $l_2 \neq \varepsilon$, a deletion if $l_2 = \varepsilon$, and an insertion if $l_1 = \varepsilon$. For nodes u and v, we also denote $(l(u) \mapsto l(v))$ by $(u \mapsto v)$. We define a *cost function* $\gamma\colon (\Sigma_\varepsilon \times \Sigma_\varepsilon - \{(\varepsilon, \varepsilon)\}) \mapsto \mathbf{R}^+$ on pairs of labels. We often constrain a cost function γ to be a *metric*, that is, $\gamma(l_1, l_2) \geq 0$, $\gamma(l_1, l_2) = 0$ iff $l_1 = l_2$, $\gamma(l_1, l_2) = \gamma(l_2, l_1)$ and $\gamma(l_1, l_3) \leq \gamma(l_1, l_2) + \gamma(l_2, l_3)$. In particular, we call the cost function that $\gamma(l_1, l_2) = 1$ if $l_1 \neq l_2$ a *unit cost function*.

Definition 2 (Edit distance (Tai, 1979))**.** For a cost function γ, the *cost* of an edit operation $e = l_1 \mapsto l_2$ is given by $\gamma(e) = \gamma(l_1, l_2)$. The *cost* of a sequence $E = e_1, \ldots, e_k$ of edit operations is given by $\gamma(E) = \sum_{i=1}^{k} \gamma(e_i)$. Then, an *edit distance* $\tau_{\mathrm{TAI}}(T_1, T_2)$ between trees T_1 and T_2 is defined as follows:

$$\tau_{\mathrm{TAI}}(T_1, T_2) = \min \left\{ \gamma(E) \;\middle|\; \begin{array}{l} E \text{ is a sequence} \\ \text{of edit operations} \\ \text{transforming } T_1 \text{ to } T_2 \end{array} \right\}.$$

Definition 3 (Tai mapping (Tai, 1979))**.** Let T_1 and T_2 be trees. We say that a triple (M, T_1, T_2) is an *unordered Tai mapping* (a *mapping*, for short) from T_1 to T_2 if $M \subseteq V(T_1) \times V(T_2)$ and every pair (u_1, v_1) and (u_2, v_2) in M satisfies that (1) $u_1 = u_2$ iff $v_1 = v_2$ (one-to-one condition) and (2) $u_1 \leq u_2$ iff $v_1 \leq v_2$ (ancestor condition). We will use M instead of (M, T_1, T_2) when there is no confusion denote it by $M \in \mathcal{M}_{\mathrm{TAI}}(T_1, T_2)$.

Let M be a mapping from T_1 to T_2. Let I_M and J_M be the sets of nodes in T_1 and T_2 but not in M, that is, $I_M = \{u \in T_1 \mid (u, v) \notin M\}$ and $J_M = \{v \in T_2 \mid (u, v) \notin M\}$. Then, the *cost* $\gamma(M)$ of M is given as follows.

$$\gamma(M) = \sum_{(u,v) \in M} \gamma(u, v) + \sum_{u \in I_M} \gamma(u, \varepsilon) + \sum_{v \in J_M} \gamma(\varepsilon, v).$$

Theorem 1. *The following statement holds (Tai, 1979).*

$$\tau_{\mathrm{TAI}}(T_1, T_2) = \min\{\gamma(M) \mid M \in \mathcal{M}_{\mathrm{TAI}}(T_1, T_2)\}.$$

Unfortunately, the following theorem holds for computing τ_{TAI} between unordered trees.

Theorem 2. *For unordered trees T_1 and T_2, the problem of computing $\tau_{\mathrm{TAI}}(T_1, T_2)$ is* MAX SNP*-hard (Zhang and Jiang, 1994). This statement also holds even if both T_1 and T_2 are binary (Hirata et al., 2011).*

Finally, we introduce the variations of a Tai mapping and an edit distance.

Definition 4 (Variations of Tai mapping)**.** Let T_1 and T_2 be trees and $M \in \mathcal{M}_{\mathrm{TAI}}(T_1, T_2)$. We denote $M \setminus \{(r(T_1), r(T_2))\}$ by M^-.

1. We say that M is an *isolated-subtree mapping* (Zhang, 1995; Zhang, 1996), denoted by $M \in \mathcal{M}_{\mathrm{ILST}}(T_1, T_2)$, if M satisfies the following condition.
$$\forall (u_1, v_1)(u_2, v_2)(u_3, v_3) \in M \\ (u_3 < u_1 \sqcup u_2 \iff v_3 < v_1 \sqcup v_2).$$
2. We say that M is an *accordant mapping* (Kuboyama, 2007), denoted by $M \in \mathcal{M}_{\mathrm{ACC}}(T_1, T_2)$, if M satisfies the following condition.
$$\forall (u_1, v_1)(u_2, v_2)(u_3, v_3) \in M \\ (u_1 \sqcup u_2 = u_1 \sqcup u_3 \iff v_1 \sqcup v_2 = v_1 \sqcup v_3).$$
3. We say that M is an *LCA-preserving mapping* (Zhang et al., 1996), denoted by $M \in \mathcal{M}_{\mathrm{LCA}}(T_1, T_2)$, if M satisfies the following condition.
$$\forall (u_1, v_1)(u_2, v_2) \in M\,((u_1 \sqcup u_2, v_1 \sqcup v_2) \in M).$$
4. We say that M is an *LCA- and root-preserving mapping* (Yoshino and Hirata, 2017), denoted by $M \in \mathcal{M}_{\mathrm{LCART}}(T_1, T_2)$, if $M \in \mathcal{M}_{\mathrm{LCA}}(T_1, T_2)$ and $(r(T_1), r(T_2)) \in M$.
5. We say that M is a *Top-down mapping* (Chawathe, 1999; Selkow, 1977), denoted by $M \in \mathcal{M}_{\mathrm{TOP}}(T_1, T_2)$, if M satisfies the following condition.
$$\forall (u, v) \in M^-\,((par(u), par(v)) \in M).$$

The above variation of Tai mapping provides the following hierarchy (Kuboyama, 2007; Yoshino and Hirata, 2017).

$$\mathcal{M}_{\mathrm{TOP}}(T_1, T_2) \subseteq \mathcal{M}_{\mathrm{LCART}}(T_1, T_2) \subseteq \mathcal{M}_{\mathrm{LCA}}(T_1, T_2) \\ \subseteq \mathcal{M}_{\mathrm{ACC}}(T_1, T_2) \subseteq \mathcal{M}_{\mathrm{ILST}}(T_1, T_2) \subseteq \mathcal{M}_{\mathrm{TAI}}(T_1, T_2).$$

Definition 5 (Variations of edit distance)**.** For every $A \in \{\mathrm{ILST}, \mathrm{ACC}, \mathrm{LCA}, \mathrm{LCART}, \mathrm{TOP}\}$, we define the distance $\tau_{\mathrm{A}}(T_1, T_2)$ as follows.

$$\tau_{\mathrm{A}}(T_1, T_2) = \min\{\gamma(M) \mid M \in \mathcal{M}_{\mathrm{A}}(T_1, T_2)\}.$$

Here we call τ_{ILST} an *isolated-subtree distance* (Zhang, 1995; Zhang, 1996), τ_{ACC} an *accordant distance* (Kuboyama, 2007), τ_{LCA} an *LCA-preserving distance* (Zhang et al., 1996), τ_{LCART} an *LCA- and root-preserving distance* (Yoshino and Hirata, 2017), and τ_{TOP} a *top-down distance* (Chawathe, 1999; Selkow, 1977). By the Tai mapping hierarchy, the following inequality for the variation of edit distance holds.

$$\tau_{\mathrm{TAI}}(T_1, T_2) \leq \tau_{\mathrm{ILST}}(T_1, T_2) \leq \tau_{\mathrm{ACC}}(T_1, T_2) \leq \\ \tau_{\mathrm{LCA}}(T_1, T_2) \leq \tau_{\mathrm{LCART}}(T_1, T_2) \leq \tau_{\mathrm{TOP}}(T_1, T_2).$$

Furthermore, for all the above variations, the following theorem holds.

Theorem 3 (*cf.*, (Yamamoto et al., 2014; Yoshino and Hirata, 2017; Zhang et al., 1996))**.** For every $A \in \{\textsc{Ilst}, \textsc{Acc}, \textsc{Lca}, \textsc{LcaRt}, \textsc{Top}\}$, we can compute $\tau_A(T_1, T_2)$ in $O(n^2 d)$ time, where $n = \max(|T_1|, |T_2|)$ and $d = \min\{d(T_1), d(T_2)\}$.

3 EARTH MOVER'S DISTANCE FOR TREES

In this section, we first introduce an *earth mover's distance* (Rubner et al., 2007) and then extend to that for trees based on Tai mapping hierarchy.

We call the set of pairs of a feature p_i and its weight w_i a *signature* and denote it by $P = \{(p_i, w_i)\}$. For a feature p_i such that $(p_i, w_i) \in P$, we denote $p_i \in P$ simply. An *earth mover's distance* (*EMD*, for short) between two signatures is given as the minimum cost of the transportation problem from a signature to another signature.

Let $P = \{(p_i, u_i)\}$ and $Q = \{(q_j, v_j)\}$ be signatures. We call a distance between p_i and q_j a *ground distance* and denote it by $gd(p_i, q_i)$. Also we denote the *flow* from p_i to q_j by f_{ij}. When the *cost* of the flow from p_i to q_j is given by $gd(p_i, q_j) f_{ij}$, the overall cost of the flows from P to Q is defined as follows.

$$\sum_{p_i \in P} \sum_{q_j \in Q} gd(p_i, q_j) f_{ij}.$$

Then, find the minimum cost flow f^*_{ij} subject to the following constraints:

1. $f_{ij} \geq 0$,
2. $\sum_{p_i \in P} f_{ij} \leq u_i$,
3. $\sum_{q_j \in Q} f_{ij} \leq v_j$,
4. $\sum_{p_i \in P} \sum_{q_j \in Q} f_{ij} = \min\left(\sum_{p_i \in P} u_i, \sum_{q_j \in Q} v_j\right)$.

The constraint (1) allows moving "supplies" from P to Q and not vice versa. The constraints (2) and (3) limit the amount of supplies within the weight. The constraint (4) forces to move the maximum amount of supplies possible.

Let f^*_{ij} be the optimum flow of the transportation problem. Then, we define the EMD between two signatures P and Q as follows.

$$EMD_{gd}(P, Q) = \frac{\sum_{p_i \in P} \sum_{q_j \in Q} gd(p_i, q_j) f^*_{ij}}{\sum_{p_i \in P} \sum_{q_j \in Q} f^*_{ij}} = \frac{\sum_{p_i \in P} \sum_{q_j \in Q} gd(p_i, q_j) f^*_{ij}}{\min\left(\sum_{p_i \in P} u_i, \sum_{q_j \in Q} v_j\right)}.$$

Note that the EMD allows for *partial matches* when the total weight of a signature is different from that of another signature, which is important for image retrieval applications (Rubner et al., 2007). We can realize the partial match to transport from a signature whose total weight is smaller than a part of another signature. Also the following theorem holds for the EMD.

Theorem 4. *Suppose that two signatures have the same total weight. If a ground distance is a metric, then so is the EMD. Furthermore, we can compute the EMD in $O(n^3 \log n)$ time, where $n = \max\{|P|, |Q|\}$ (Rubner et al., 2007).*

Next, we formulate the EMD for trees based on Tai mapping hierarchy.

It is necessary for the EMD to introduce a signature and a ground distance between features. In order to formulate the EMD for trees, we transform from a tree to a signature. In this paper, we adopt the following signature $s(T)$ for a tree T.

$$s(T) = \left\{ (S, w) \,\middle|\, S \in cs(T), w = \frac{f(S, T)}{|T|} \right\}.$$

The features of $s(T)$ are complete subtrees of T and the weight of $s(T)$ is the ratio of the occurrences of complete subtrees. Hence, the total weight of s is 1. Since this signature contains T itself, we can transform T to $s(T)$ uniquely. On the other hand, as a ground distance between trees, we adopt 5 tractable variations of the edit distance, that is, $\tau_{\textsc{Top}}$, $\tau_{\textsc{LcaRt}}$, $\tau_{\textsc{Lca}}$, $\tau_{\textsc{Acc}}$ and $\tau_{\textsc{Ilst}}$.

Hence, by combining signatures and ground distances, we formalize the following 5 kinds of an *EMD for trees*. In the following, we assume that $A \in \{\textsc{Ilst}, \textsc{Acc}, \textsc{Lca}, \textsc{LcaRt}, \textsc{Top}\}$.

Definition 6 (EMD for trees)**.** We define an *EMD for trees* as $EMD_{\tau_A}(s(T_1), s(T_2))$ between signatures $s(T_1)$ and $s(T_2)$ for a ground distance τ_A and denote it by $EMD_A(T_1, T_2)$.

Corollary 1. *$EMD_A(T_1, T_2)$ is a metric.*

Proof. It is straightforward since a ground distance τ_A is a metric and the total weight of signatures is 1 and by Theorem 4. □

Theorem 5. *We can compute $EMD_A(T_1, T_2)$ in $O(n^3 \log n)$ time, where $n = \max\{|T_1|, |T_2|\}$.*

Proof. By using $s(T_1)$, $s(T_2)$ and $\{\tau_A(T_1[u], T_2[v]) \mid (u, v) \in T_1 \times T_2\}$, we can design the following algorithm to compute $EMD_A(T_1, T_2)$.

1. Construct $s(T_1)$ and $s(T_2)$ from T_1 and T_2.
2. Compute $G = \{\tau_A(T_1[u], T_2[v]) \mid (u, v) \in T_1 \times T_2\}$.
3. Compute $EMD_A(T_1, T_2)$ from G.

It is obvious that the running time of Step 1 is $O(n)$. For Step 2, since the algorithm of computing $\tau_{\mathrm{A}}(T_1,T_2)$ can store the value of $\tau_{\mathrm{A}}(T_1[u],T_2[v])$ for every $(u,v) \in T_1 \times T_2$ and by Theorem 3, we can compute G in $O(n^2 d)$ time, where $d = \min\{d(T_1), d(T_2)\}$. Since $|s(T_1)| = |s(T_2)| = O(n)$ and by Theorem 4, the running time of Step 3 is $O(n^3 \log n)$. Hence, we can compute $EMD_{\mathrm{A}}(T_1,T_2)$ in $O(n) + O(n^2 d) + O(n^3 \log n) = O(n^3 \log n)$ time. □

4 EXPERIMENTAL RESULTS

In this section, we give experimental results to evaluate EMD_{A} to compare EMD_{A} with τ_{A} and investigates the properties of EMD_{A}. Here, we assume that a cost function is a unit cost function.

In this section, we use two kinds of data; One is N-glycan data provided from KEGG[1] as real data. Another is 6 data of randomly generated trees by using the algorithm PTC (Luke and Panait, 2001). We call them R_i $(1 \leq i \leq 6)$, where the number of nodes in R_i is $50 \times i$. Furthermore, we use the computer environment that CPU is Intel Xeon E51650 v3 (3.50GHz), RAM is 1GB and OS is Ubuntsu Linux 14.04 (64bit).

Table 1 illustrates the details of data, that is, the number of data (#), the average number of nodes (n), the average degree (d) and the average height (h).

Table 1: The details of data.

data	#	n	d	h
N-glycan	2142	11.07	2.07	6.20
R_1	100	50.00	2.00	8.75
R_2	100	100.00	2.00	10.69
R_3	100	150.00	2.00	12.12
R_4	100	200.00	2.00	12.75
R_5	100	250.00	2.00	13.81
R_6	100	300.00	2.00	14.24

4.1 Running Time

First, we compare the running time to compute EMD_{A} and τ_{A} for N-glycan data and randomly generated trees in Table 1. Table 2 illustrates the running time to compute such distances.

Tables 1 and 2 show that the running time of both EMD_{A} and τ_{A} is increasing when the number of nodes is increasing and the ratio of increasing for EMD_{A} is larger than that for τ_{A}.

[1]Kyoto Encyclopedia of Genes and Genomes. http://www.kegg.jp/

Table 2: The running time to compute the distances (sec.).

distance	N-glycan	R_1	R_2
τ_{ILST}	1580.95	69.72	289.48
τ_{ACC}	1386.33	60.18	285.78
τ_{LCA}	1129.97	49.13	201.78
τ_{LCART}	1109.80	49.64	203.96
τ_{TOP}	485.42	20.71	83.56
EMD_{ILST}	1592.32	77.00	351.81
EMD_{ACC}	1399.14	66.23	307.31
EMD_{LCA}	1133.82	55.17	261.24
EMD_{LCART}	1128.05	55.08	261.36
EMD_{TOP}	509.49	26.45	138.04

distance	R_3	R_4	R_5	R_6
τ_{ILST}	665.12	1186.53	1874.17	2722.80
τ_{ACC}	578.79	1013.98	1597.39	2308.71
τ_{LCA}	461.58	824.09	1298.07	1873.32
τ_{LCART}	467.38	834.47	1313.06	1891.92
τ_{TOP}	189.58	336.86	527.42	760.66
EMD_{ILST}	894.53	1802.92	3073.49	4832.80
EMD_{ACC}	790.38	1583.50	2763.23	4376.64
EMD_{LCA}	687.20	1401.26	2474.14	3965.60
EMD_{LCART}	687.29	1414.33	2474.22	3961.98
EMD_{TOP}	397.84	875.98	1637.42	2759.20

Table 3 illustrates the ratio ($EMD_{\mathrm{A}}/\tau_{\mathrm{A}}$) of the running time of computing the EMDs (EMD_{A}) for that of computing the ground distances (τ_{A}) in Table 2. Here, we call it the ratio of EMD_{A} for τ_{A} simply.

Table 3: The ratio ($EMD_{\mathrm{A}}/\tau_{\mathrm{A}}$) of the running time of computing the EMDs (EMD_{A}) for that of computing the ground distances (τ_{A}) in Table 2.

A	N-glycan	R_1	R_2	R_3	R_4	R_5	R_6
ILST	1.01	1.10	1.22	1.34	1.52	1.64	1.77
ACC	1.01	1.10	1.08	1.37	1.56	1.73	1.90
LCA	1.00	1.12	1.29	1.49	1.70	1.91	2.12
LCART	1.02	1.11	1.28	1.47	1.69	1.88	2.09
TOP	1.05	1.28	1.65	2.10	2.60	3.10	3.63

Table 3 shows that, whereas the ratio of EMD_{A} for τ_{A} is between 1.00 and 1.05 for N-glycan data, the ratio of EMD_{TOP} for τ_{TOP} is over 3 for the data R_6. On the other hand, smaller distance in the inequality for the variations ($\tau_{\mathrm{ILST}} \leq \tau_{\mathrm{ACC}} \leq \tau_{\mathrm{LCA}} \leq \tau_{\mathrm{LCART}} \leq \tau_{\mathrm{TOP}}$) tends to give smaller ratio of EMD_{A} for τ_{A} except LCA and LCART; The ratio of EMD_{LCA} for τ_{LCA} is greater than the ratio of EMD_{LCART} for τ_{LCART}.

Furthermore, whereas the ratio of EMD_{A} for τ_{A} is $O(n \log n/d)$ in theoretical by Theorems 3 and 5, the ratio is at most 4 in experimental. Then, the problems

of computing EMDs are efficient for trees with at least 300 nodes and small degree.

4.2 Comparing EMDs with Ground Distances

Next, we investigate the relationship between the EMD EMD_{A} and its ground distance τ_{A} for N-glycan data.

Figure 2 illustrates the distributions of EMDs (upper) and ground distances (lower). Here, the x-axis is the value of the distance and the y-axis is the percentage of pairs with the distance pointed by the x-axis.

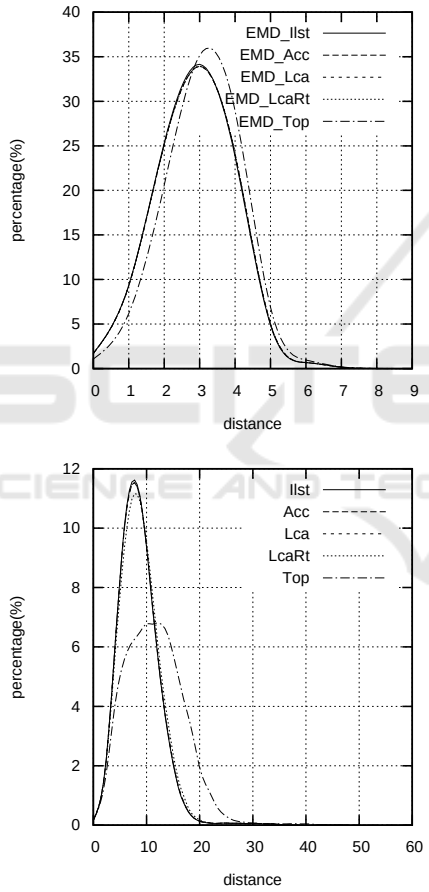


Figure 2: The distributions of EMDs (upper) and ground distances (lower) for N-glycan data.

Figure 2 shows that both EMDs and ground distances are near to normal distribution. Also the distributions of EMD_{TOP} and τ_{TOP} are right to other EMD_{A} and τ_{A} ($A \in \{\mathrm{ILST}, \mathrm{ACC}, \mathrm{LCA}, \mathrm{LCART}\}$), respectively. Whereas the peak of the distribution of EMD_{TOP} is larger than that of other distributions of EMD_{A}, the peak of the distribution of τ_{TOP} is smaller than that of other distributions of τ_{A}.

Figures 3 and 4 illustrate the scatter charts between the number of pairs of trees with τ_{A} pointed at the x-axis and that with EMD_{A} pointed at the y-axis for N-glycan data whose number of total pairs is 2,293,011. Here, the diameter and the color represent the number of pairs of trees such that longer diameter and deeper color are larger number. Also, Figures 3 and 4 represent the cases that $A \in \{\mathrm{ILST}, \mathrm{ACC}\}$ and $A \in \{\mathrm{LCA}, \mathrm{LCART}, \mathrm{TOP}\}$, respectively.

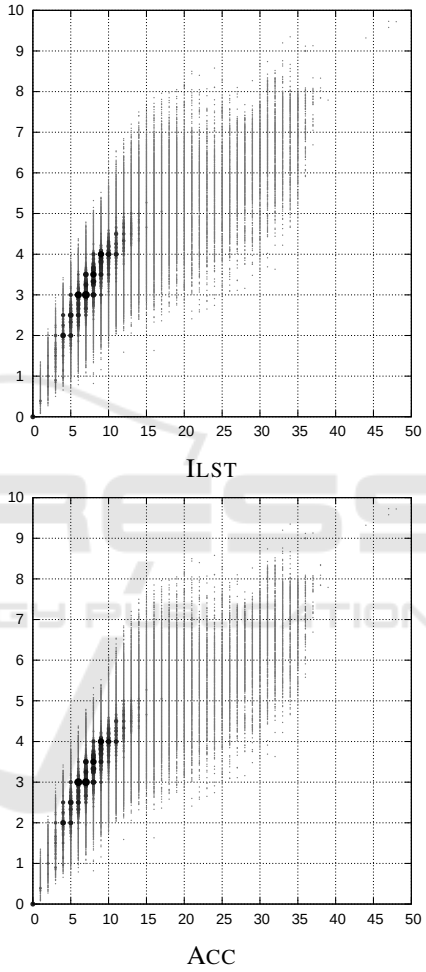
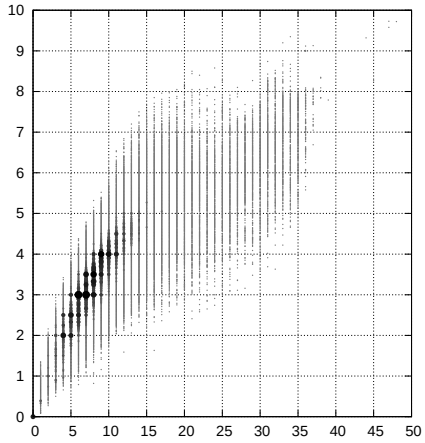


Figure 3: The scatter charts between the number of pairs of trees with τ_{A} pointed at the x-axis and that with EMD_{A} pointed at the y-axis for $A \in \{\mathrm{ILST}, \mathrm{ACC}\}$.

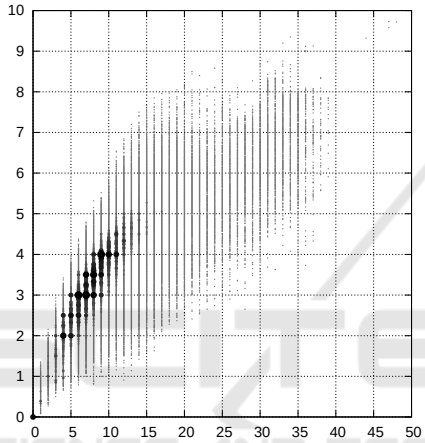
Figures 3 and 4 show that EMD_{A} is relative to τ_{A} and almost values of τ_{A} are larger than those of EMD_{A}. Also the plots of TOP vary more widely than others.

4.3 Typical Cases

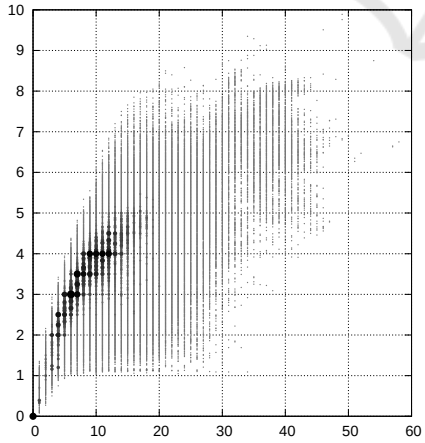
In the following, we point out the typical cases of trees with different values between of τ_{A} and EMD_{A}.



LCA



LCART



TOP

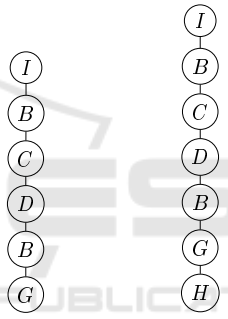
Figure 4: The scatter charts between the number of pairs of trees with τ_A pointed at the x-axis and that with EMD_A pointed at the y-axis for $A \in \{\text{LCA}, \text{LCART}, \text{TOP}\}$.

Here, let u_i be a node in T_1 such that $pre(u_i) = i$ and v_i a node in T_2 such that $pre(v_i) = i$.

Example 1. Consider trees T_1 and T_2 illustrated in Figure 5, that is, one tree (T_1) is obtained by deleting leaves to another tree (T_2). In this case, it holds that $\tau_A(T_1,T_2) \leq EMD_A(T_1,T_2)$. For the trees T_1 and T_2 in Figure 5, it holds that $\tau_A(T_1,T_2) = 1$ and $EMD_A(T_1,T_2) = 1.357$ for every $A \in \{\text{ILST}, \text{ACC}, \text{LCA}, \text{LCART}, \text{TOP}\}$.

It is obvious that $\tau_A(T_1,T_2) = 1$. On the other hand, it holds that $\tau_A(T_1[u_i],T_2[v_i]) = 1$ and $\tau_A(T_1[u_i],T_2[v_7]) = |T_1[u_i]|$ $(1 \leq i \leq 6)$. Since the weight of $T_1[u_i]$ (*resp.*, $T_2[v_i]$) is $1/6$ (*resp.*, $1/7$), the optimum flow consists of the 6 flows from $T_1[u_i]$ to $T_2[v_i]$ whose costs are $1/7$ and the 6 flows from $T_1[u_i]$ to $T_2[v_7]$ whose costs are $1/42$. Then, the cost of the optimum flow is $6(1/7)+(6+5+4+3+2+1)/42 = 57/42 = 1.357 = EMD_A(T_1,T_2)$.

Hence, whereas the ground distances are not sensitive to inserting leaves, the EMD is necessary to transport the remained weights for every node in one tree to an inserted leave in another tree.



$T_1 = \text{G01687}$ $T_2 = \text{G02836}$

Figure 5: Trees T_1 and T_2.

Example 2. Consider trees T_1 and T_2 illustrated in Figure 6, that is, just a label of the root in one tree (T_1) is different from that in another tree (T_2). In this case, it holds that $EMD_A(T_1,T_2) \leq \tau_A(T_1,T_2)$. For the trees T_1 and T_2 in Figure 6, it holds that $\tau_A(T_1,T_2) = 1$ and $EMD_A(T_1,T_2) = 0.083$ for every $A \in \{\text{ILST}, \text{ACC}, \text{LCA}, \text{LCART}, \text{TOP}\}$.

It is obvious that $\tau_A(T_1,T_2) = 1$. On the other hand, the signature containing $r(T_1)$ (*resp.*, $r(T_2)$) is just T_1 (*resp.*, T_2) itself. Since $\tau_A(T_1[u_i],T_2[v_i]) = 0$ for $2 \leq i \leq 12$, the cost of the flow from $T_1[u_i]$ to $T_2[v_i]$ is 0. Since the weight of $T_1[u_i]$ and $T_2[v_i]$ is $1/12$ and $\tau_A(T_1[u_1],T_2[v_1]) = 1$, the cost of the optimum flow is $1/12 + 11(0/12) = 0.083 = EMD_A(T_1,T_2)$.

Hence, the difference near to the root is more sensitive to the ground distances rather than the EMDs. Furthermore, in this case, the EMDs is much smaller than the ground distance.

Example 3. Consider trees T_1 and T_2 illustrated in Figure 7 and T_3 and T_4 illustrated in Figure 8, that

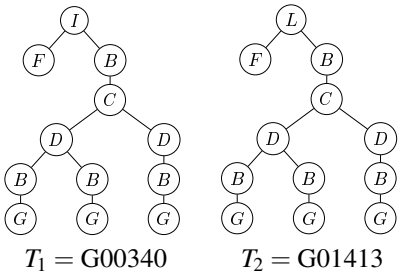


Figure 6: Trees T_1 and T_2.

is, one tree (T_1 or T_3) is obtained by deleting the root of another tree (T_2 or T_4). For these cases, it holds that $EMD_{\text{LCARt}}(T_1,T_2) \leq \tau_{\text{LCARt}}(T_1,T_2)$ and $EMD_{\text{Top}}(T_3,T_4) \leq \tau_{\text{Top}}(T_3,T_4)$. For the trees T_1 and T_2 in Figure 7, $\tau_{\text{A}}(T_1,T_2)$ and $EMD_{\text{A}}(T_1,T_2)$ are:

A	τ_{A}	EMD_{A}
Lca	2	0.841
LcaRt	12	0.917
Top	17	1.512

For the trees T_3 and T_4 in Figure 8, $\tau_{\text{A}}(T_3,T_4)$ and $EMD_{\text{A}}(T_3,T_4)$ are:

A	τ_{A}	EMD_{A}
Lca	2	0.810
LcaRt	4	0.813
Top	34	1.092

Here, we also illustrate the minimum cost mapping in $\mathcal{M}_{\text{A}}$ in Figures 7 and 8, where the corresponding node is denoted by $\circ$ and the non-corresponding node is denoted by $\bullet$, which implies τ_{A}.

The reason is that the structural difference near to the root is much sensitive to τ_{LCARt} and τ_{Top}, whose values tend to be large, but the EMDs are not.

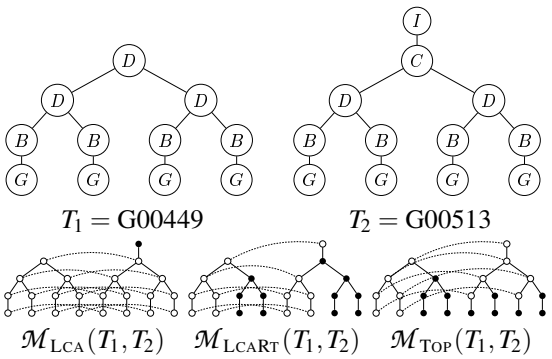


Figure 7: Trees T_1 and T_2.

Example 4. Consider trees T_1 and T_2 illustrated in Figure 9, that is, subtrees in one tree (T_1) frequently occur in another tree (T_2). In this case, it holds

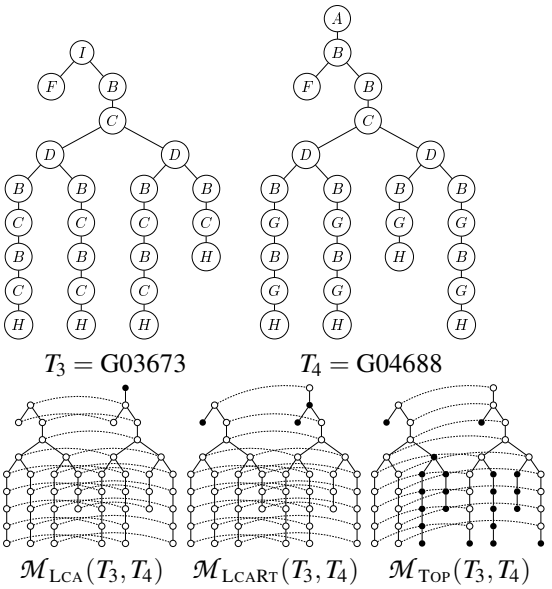


Figure 8: Trees T_3 and T_4.

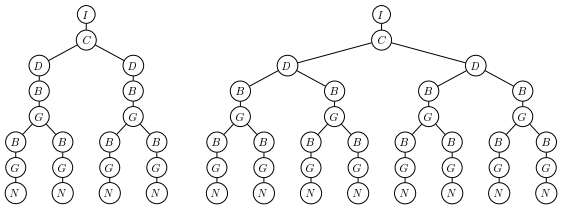
that $EMD_{\text{A}}(T_1,T_2)$ is much smaller than $\tau_{\text{A}}(T_1,T_2)$. For the trees T_1 and T_2 in Figure 6, it holds that $\tau_{\text{A}}(T_1,T_2) = 16$ and $EMD_{\text{A}}(T_1,T_2) = 1.63$ for every $A \in \{\text{Ilst}, \text{Acc}, \text{Lca}, \text{LcaRt}, \text{Top}\}$. Since T_2 is obtained by inserting 16 nodes to T_1, it holds that $\tau_{\text{A}}(T_1,T_2) = 16$.

The weight of $T_1[u]$ (*resp.*, $T_2[v]$) is $1/20$ (*resp.*, $1/36$). Then, $T_1[u_4]$, $T_1[u_{13}]$, $T_2[v_4]$, $T_2[v_{12}]$, $T_2[v_{21}]$ and $T_2[v_{29}]$ are isomorphic and $T_1[u_6]$, $T_1[u_9]$, $T_1[u_{15}]$, $T_1[u_{18}]$, $T_2[v_6]$, $T_2[v_9]$, $T_1[u_{14}]$, $T_1[u_{17}]$, $T_2[v_{23}]$, $T_2[v_{26}]$, $T_1[u_{31}]$ and $T_1[u_{34}]$ are isomorphic, so the weights of $T_1[u_4]$, $T_2[v_4]$, $T_1[u_6]$ and $T_2[u_6]$ as features are $2/20$, $4/36$, $4/20$ and $8/36$, respectively. Since these weights are preserved in the subtrees of them, the total weight of features consisting of $T_1[u_4]$ and its subtrees in T_1 is $2/20+2/20+4/20+4/20 = 16/20$ and that of $T_2[v_4]$ and its subtrees in T_2 is $4/36+4/36+8/36+8/36 = 32/36$. Hence, the cost of flows in these isomorphic subtrees from T_1 to T_2 is 0, because $\tau_{\text{A}}(T_1[u_4],T_2[v_4]) = 0$, for example. Since these flows move all the weight $16/20$ of $T_1[u_4]$, $T_2[v_4]$ and its subtrees can receive the weight $32/36 - 16/20 = 4/45$.

For the remained features in T_2, the weights of $T_2[v_1]$, $T_2[v_2]$ and $T_2[v_3]$ as features are $1/36$, $1/36$ and $2/36$, respectively. Furthermore, as $T_2[v_4]$ and its subtrees receive the weights, it is necessary to consider the ground distances between $T_1[u_3]$ and $T_2[v_i]$ ($4 \leq i \leq 8$). The ground distances necessary to compute $EMD_{\text{A}}(T_1,T_2)$ are given as follows.

$\tau_A(T_1[u_1], T_2[v_1]) = 16$, $\tau_A(T_1[u_2], T_2[v_2]) = 16$,
$\tau_A(T_1[u_3], T_2[v_3]) = 8$, $\tau_A(T_1[u_1], T_2[v_3]) = 3$,
$\tau_A(T_1[u_2], T_2[v_3]) = 4$, $\tau_A(T_1[u_3], T_2[v_4]) = 1$,
$\tau_A(T_1[u_3], T_2[v_5]) = 2$, $\tau_A(T_1[u_3], T_2[v_6]) = 6$,
$\tau_A(T_1[u_3], T_2[v_7]) = 7$, $\tau_A(T_1[u_3], T_2[v_8]) = 8$.

Hence, by computing the optimum flow to receive the weight $4/45 + 4/36 = 1/5$ in T_2, we can obtain $EMD_A(T_1, T_2)$ as $16(1/36) + 16(1/36) + 8(1/90) + 3(1/45) + 4(1/45) + 1(1/90) + 2(1/90) + 6(1/45) + 7(1/45) + 8(1/45) = 49/30 = 1.633$.



$T_1 = \text{G03824}$ $T_2 = \text{G04045}$

Figure 9: Trees T_1 and T_2.

4.4 Properties of EMDs for Trees

Finally, we investigate the properties of the EMDs for trees by summarizing the typical cases in Section 4.3.

1. Concerned with Example 1, just the case that one tree is obtained by deleting leaves to another tree implies that $\tau_A(T_1, T_2) \leq EMD_A(T_1, T_2)$ for N-glycan data. Whereas the trees T_1 and T_2 in Example 1 are paths, the statement holds when some internal nodes have some leaves as children.

   As another case concerned with Example 1, consider trees T_i $(1 \leq i \leq 6)$ in Figure 10. Then, it holds that $\tau_A(T_1, T_i) = 1$ for every i $(2 \leq i \leq 6)$ but $EMD_A(T_1, T_2) = 0.2$, $EMD_A(T_1, T_3) = 0.4$, $EMD_A(T_1, T_4) = 0.6$, $EMD_A(T_1, T_5) = 0.8$ and $EMD_A(T_1, T_6) = 1$. The reason is that the farther node with a different label from the root makes more different signatures.

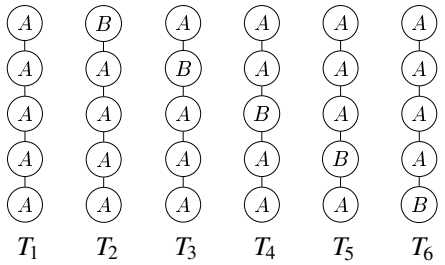


Figure 10: Trees T_i $(1 \leq i \leq 6)$.

2. Concerned with Examples 2 and 3, consider complete binary trees T_1 and T_2 with 15 nodes and a tree T_3 adding the root to T_1 illustrated in Figure 11. Then, for $A \in \{\text{ILST}, \text{TOP}\}$, $EMD_A(T_1, T_i)$ and $\tau_A(T_1, T_i)$ are as follows.

T_i	T_2	T_3
$EMD_{\text{ILST}}(T_1, T_i)$	0.067	0.796
$EMD_{\text{TOP}}(T_1, T_i)$	0.067	1.07
$\tau_{\text{ILST}}(T_1, T_i)$	1	1
$\tau_{\text{TOP}}(T_1, T_i)$	1	23

Hence, the difference of both labels and structures near to the root is more sensitive to τ_{TOP} than EMD_{TOP}. On the other hand, for the difference of labels near to the root, EMD_A is much smaller than τ_A. As stated in Examples 2 and 3, there also exists a case that LCATOP is sensitive to the difference of both labels and structures near to the root.

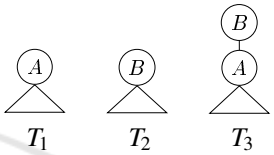


Figure 11: Trees T_1, T_2 and T_3.

3. Concerned with Example 4, consider a tree T_1 with 10 nodes and trees T_i $(2 \leq i \leq 5)$ containing T_1 as subtrees illustrated in Figure 12. Then, $EMD_A(T_1, T_i)$ and $\tau_A(T_1, T_i)$ are as follows.

T_i	T_2	T_3	T_4	T_5
$EMD_A(T_1, T_i)$	0.5	0.738	0.822	0.866
$\tau_A(T_1, T_i)$	1	11	21	31

In this case, whereas the ground distances are necessary to insert new nodes, the EMDs tend to absorb the influence of isomorphic subtrees.

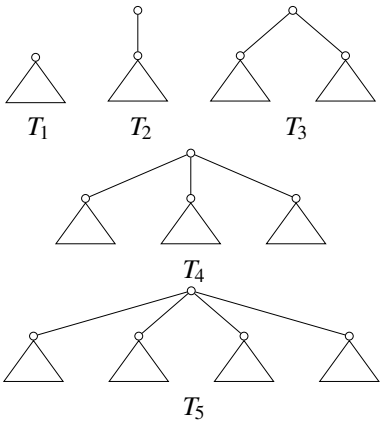


Figure 12: Trees T_i $(1 \leq i \leq 5)$.

5 CONCLUSION

In this paper, for the variations of edit distance τ_A for $A \in \{\text{ILST}, \text{ACC}, \text{LCA}, \text{LCART}, \text{TOP}\}$, we have formulated the earth mover's distances EMD_A based on τ_A. Then, we have given experimental results to evaluate EMD_A comparing with τ_A. As a result, we have investigated the properties of EMD_A.

It is a future work to give experimental results for more large data (with large degrees) to analyze the theoretical ratio $O(n \log n/d)$ in Section 4.1 in experimental. Also it is a future work to formulate EMDs to other tractable variations in Tai mapping hierarchy (Yoshino and Hirata, 2017).

Concerned with Example 1 in Section 4.3 and Stetement 1 in Section 4.4, we have found no trees T_1 and T_2 such that $\tau_A(T_1,T_2) < EMD_A(T_1,T_2)$ except the case that T_1 is obtained by deleting leaves to T_2. Then, it is a future work to determine whether or not there exist other cases satisfying that $\tau_A(T_1,T_2) < EMD_A(T_1,T_2)$.

It is a future work to analyze the properties of EMDs in Section 4.4 in more detail and investigate how data are appropriate for EMDs. In particular, since it is possible that the number of the signature is too small to formulate EMDs for trees, it is an important future work to investigate appropriate signatures for EMDs for trees.

ACKNOWLEDGEMENTS

This work is partially supported by Grant-in-Aid for Scientific Research 17H00762, 16H02870, 16H01743 and 15K12102 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

Aratsu, T., Hirata, K., and Kuboyama, T. (2009). Sibling distance for rooted labeled trees. In *JSAI PAKDD'08 Post-Workshop Proc. (LNAI 5433)*, pages 99–110.

Chawathe, S. S. (1999). Comparing hierarchical data in external memory. In *Proc. VLDB'99*, pages 90–101.

Gollapudi, S. and Panigrahy, R. (2008). The power of two min-hashes for similarity search among hierarchical data objects. In *Proc. PODS'08*, pages 211–219.

Hirata, K., Yamamoto, Y., and Kuboyama, T. (2011). Improved MAX SNP-hard results for finding an edit distance between unordered trees. In *Proc. CPM'11 (LNCS 6661)*, pages 402–415.

Jiang, T., Wang, L., and Zhang, K. (1995). Alignment of trees – an alternative to tree edit. *Theoret. Comput. Sci.*, 143:137–148.

Kailing, K., Kriegel, H.-P., Schönaur, S., and Seidl, T. (2004). Efficient similarity search for hierarchical data in large databases. In *Proc. EDBT'04*, pages 676–693.

Kawaguchi, T. and Hirata, K. (2017). On earth mover's distance based on complete subtrees for rooted labeled trees. In *Proc. SISA'17*, pages 225–228.

Kuboyama, T. (2007). *Matching and learning in trees*. Ph.D thesis, University of Tokyo.

Li, F., Wang, H., Li, J., and Gao, H. (2013). A survey on tree edit distance lower bound estimation techniques for similarity join on XML data. *SIGMOD Record*, 43:29–39.

Luke, S. and Panait, L. (2001). A survey and comparison of tree generation algorithms. In *Proc. GECCO'01*, pages 81–88.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2007). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40:99–121.

Selkow, S. M. (1977). The tree-to-tree editing problem. *Inform. Process. Lett.*, 6:184–186.

Tai, K.-C. (1979). The tree-to-tree correction problem. *J. ACM*, 26:422–433.

Yamamoto, Y., Hirata, K., and Kuboyama, T. (2014). Tractable and intractable variations of unordered tree edit distance. *Internat. J. Found. Comput. Sci.*, 25:307–329.

Yoshino, T. and Hirata, K. (2017). Tai mapping hierarchy for rooted labeled trees through common subforest. *Theory of Comput. Sys.*, 60:769–787.

Zhang, K. (1995). Algorithms for the constrained editing distance between ordered labeled trees and related problems. *Pattern Recog.*, 28:463–474.

Zhang, K. (1996). A constrained edit distance between unordered labeled trees. *Algorithmica*, 15:205–222.

Zhang, K. and Jiang, T. (1994). Some MAX SNP-hard results concerning unordered labeled trees. *Inform. Process. Lett.*, 49:249–254.

Zhang, K., Wang, J., and Shasha, D. (1996). On the editing distance between undirected acyclic graphs. *Internat. J. Found. Comput. Sci.*, 7:43–58.