LCA Histogram Distance for Rooted Labeled Caterpillars

Takuya Yoshino, Kohei Muraka and Kouichi Hirata
Kyushu Institute of Technology, Kawazu 680-4, Iizuka 820-8502, Japan

Keywords: LCA Histogram Distance, Rooted Labeled Caterpillars, Path Histogram Distance, Complete Subtree Histogram Distance.

Abstract: An *LCA histogram distance* is an L_1-distance between histograms consisting of triples of two nodes and their least common ancestor (LCA) in two trees. In this paper, we show that the LCA histogram distance for caterpillars is always a metric, whereas that for trees is not. Then, we give experimental results for computing the LCA histogram distance by comparing with the path histogram distance and the complete subtree histogram distance for caterpillars.

1 INTRODUCTION

Comparing tree-structured data such as HTML and XML data for web mining or RNA and glycan data for bioinformatics is one of the important tasks for data mining. Then, we deal with them as *rooted labeled unordered trees*, (*trees*, for short). In particular, a *caterpillar* (*cf.* (Gallian, 2007)) is a tree transformed to a path after removing all the leaves in it. Whereas the caterpillars are very restricted and simple, there are some cases containing many caterpillars in real dataset, see Table 3 in Section 4.

The *edit distance* (Tai, 1979) is the most famous distance measure between trees. It is formulated as the minimum cost of *edit operations*, consisting of a *substitution*, a *deletion* and an *insertion*, applied to transform a tree to another tree and is always a metric. Recently, Muraka *et al.* (Muraka et al., 2018) have designed the algorithm to compute the edit distance between two caterpillars in $O(\lambda^2 h^2)$ time, where λ and h are the maximum number of leaves and the maximum height in two caterpillars, respectively. Then, this algorithm runs in $O(n^4)$ time, where n is the maximum number of vertices in two caterpillars.

A *local frequency distance* (Aratsu et al., 2009; Kailing et al., 2004; Li et al., 2013) is formulated as an L_1-distance between histograms concerned with local information. Whereas we can compute the local frequency distance efficiently and they sometimes provide the constant factor lower bound of the edit distance, almost all of them is not a metric. In order to compare caterpillars efficiently by using a metric, a *path histogram distance* (Kawaguchi et al., 2018b) and a *complete subtree histogram distance* (Akutsu et al., 2013) are appropriate local frequency distances for caterpillars.

A *path histogram distance* is an L_1-distance between histograms consisting of paths from the root to leaves in two trees (Kawaguchi et al., 2018b). It is computable in linear time, always a metric for caterpillars, which is not a metric for trees in general, and incomparable with the edit distance (Kawaguchi et al., 2018b). On the other hand, as an extreme case, for two paths with the same length such that every label in one path is *a* and that in another path is *b*, the edit distance between them is the number of vertices in a path but the path histogram distance is one.

A *complete subtree histogram distance* is an L_1-distance between histograms consisting of complete subtrees in two trees (Akutsu et al., 2013). It is computable in linear time, always a metric for trees and greater than or equal to the edit distance (Akutsu et al., 2013). On the other hand, as an extreme case, for two paths with the same length such that the labels of leaves are different, the edit distance between them is one but the complete subtree histogram distance is the number of vertices in two paths, which is the maximum value.

In this paper, we focus on an *LCA histogram distance* (Tatikonda and Parthasarathy, 2010), which is an L_1-distance between histograms consisting of triples of two vertices and the LCA of them with their depth. Whereas Tatikonda and Parthasarathy (Tatikonda and Parthasarathy, 2010) have claimed that the LCA histogram distance is a metric for trees, in this paper, we give a counterex-

Yoshino, T., Muraka, K. and Hirata, K.
LCA Histogram Distance for Rooted Labeled Caterpillars.
DOI: 10.5220/0006951603070314
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 307-314
ISBN: 978-989-758-330-8

ample that their claim is false, even if the information of depth is given, which is not well-known. On the other hand, we show that the LCA histogram distance is a metric for caterpillars. By using the LCA histogram distance, we can avoid not only the above extreme cases but also the case that both the path histogram distance and the complete subtree histogram distance are their maximum values but the edit distance is not. We can compute the LCA histogram distance in quadratic time.

Then, by using caterpillars in real data in Table 3 in Section 4, we give experimental results of computing the LCA histogram distance comparing with the path histogram distance and the complete subtree histogram distance. Note that the maximum values of the path histogram distance, the complete subtree histogram distance and the LCA histogram distance are different. Then, by normalizing the distances to compare them as experimental results, we compare the running time, distributions and scatter charts of the three distances.

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 vertex r chosen as its *root*. We denote the root of a rooted tree T by $r(T)$.

Let T be a rooted tree such that $r = r(T)$ and $u,v,w \in T$. We denote the unique path from r to v, that is, the tree (V',E') such that $V' = \{v_1,\ldots,v_k\}$, $v_1 = r$, $v_k = v$ and $(v_i,v_{i+1}) \in E'$ for every i $(1 \leq i \leq k-1)$, by $UP_r(v)$. The *depth* of v, denoted by $d(v)$, is the number of edges in $UP_r(v)$.

The *parent* of $v(\neq r)$, which we denote by $par(v)$, is its adjacent vertex on $UP_r(v)$ and the *ancestors* of $v(\neq r)$ are the vertices on $UP_r(v) - \{v\}$. 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 call a vertex with no children a *leaf* and denote the set of all the leaves in T by $lv(T)$.

We denote the set of all the children of v in T by $ch(v)$. The *degree* of v, denoted by $g(v)$, is the number of children of v, that is, $|ch(v)|$, and the *degree* of T, denoted by $g(T)$, is $\max\{g(v) \mid v \in T\}$. The *height* of v, denoted by $h(v)$, is $\max\{|UP_v(w)| \mid w \in lv(T[v])\}$, and the *height* of T, denoted by $h(T)$, is $\max\{h(v) \mid v \in T\}$.

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* (*LCA*, for short) of u and v, denoted by $u \sqcup v$, if $u \leq w$, $v \leq w$ and there exists no vertex $w' \in T$ such that $w' \leq w$, $u \leq w'$ and $v \leq w'$.

Let T be a rooted tree (V,E) and v a vertex in T. A *complete subtree of T at v*, denoted by $T[v]$, is a rooted tree $T' = (V',E')$ such that $r(T') = v$, $V' = \{u \in V \mid u \leq v\}$ and $E' = \{(u,w) \in E \mid u,w \in V'\}$. For a tree T', we say that *T' occurs in T at v* if $T' = T[v]$.

For a vertex $v \in T$, we call the occurrence number of v in the preorder (*resp.*, postorder) traversal on T the *preorder* (*resp.*, *postorder*) *number* of v and denote it by $pre(v)$ (*resp.*, $post(v)$). We say that u is *to the left of* v in T if $pre(u) \leq pre(v)$ and $post(u) \leq post(v)$. 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 vertex is assigned a symbol from a fixed finite alphabet Σ. For a vertex 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.

As the restricted form of trees, we introduce a *rooted labeled caterpillar* (a *caterpillar*, for short) as follows, which this paper mainly deals with.

Definition 1 (Caterpillar (*cf.*, (Gallian, 2007))). We say that a tree is a *caterpillar* if it is transformed to a path after removing all the leaves in it. For a caterpillar C, we call the remained path a *backbone* of C and denote it by $bb(C)$.

Next, we introduce an *edit distance* for trees.

Definition 2 (Edit operations (Tai, 1979)). The *edit operations* of a tree T are defined as follows.

1. *Substitution*: Change the label of the vertex v in T.
2. *Deletion*: Delete a non-root vertex 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'.
3. *Insertion*: The complement of deletion. Insert a vertex v as a child of v' in T making v the parent of a subset of the children of v'.

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 vertices v and w, we also denote $(l(v) \mapsto l(w))$ by $(v \mapsto w)$. We define a *cost function* $\gamma : (\Sigma_\varepsilon \times \Sigma_\varepsilon \setminus \{(\varepsilon,\varepsilon)\}) \mapsto \mathbf{R}^+$ 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 3 (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:

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

For $n_i = |T_i|$ $(i=1,2)$, it holds that $0 \leq \tau_{\mathrm{TAI}}(T_1,T_2) \leq n_1+n_2-1$.

Unfortunately, the problem of computing the edit distance between trees is MAX SNP-hard (Zhang and Jiang, 1994). On the other hand, Muraka *et al.* (Muraka et al., 2018) have recently shown the following theorem for caterpillars.

Theorem 1 ((Muraka et al., 2018))**.** *For caterpillars C_1 and C_2, we can compute $\tau_{\mathrm{TAI}}(C_1,C_2)$ in $O(\lambda^2h^2)$ time, where $\lambda=\max\{|lv(C_1)|,|lv(C_2)|\}$ and $h=\max\{h(C_1),h(C_2)\}$.*

As the previous local frequency distances to compare caterpillars, we introduce the *path histogram distance* (Kawaguchi et al., 2018b) and the *complete subtree histogram distance* (Akutsu et al., 2013).

Let T be a tree such that $r=r(T)$. Then, for $v \in lv(T)$, we regard the path $P=UP_r(v)$ such that $V(P)=\{v_1,\ldots,v_k\}$, $v_1=r$, $v_k=v$ and $(v_i,v_{i+1}) \in E(P)$ for every i $(1 \leq i \leq k-1)$ as a string $l(v_1)\cdots l(v_k)$ on Σ and denote it by $s(r,v)$. Also we say that a string $s \in \Sigma^*$ *occurs* in T if there exists a leaf $v \in lv(T)$ such that $s=s(r,v)$ and denote the number of occurrences of s in T by $f(s,T)$. Furthermore, we define $\mathcal{S}(T)$ as $\{s(r,v) \mid r=r(T), v \in lv(T)\}$.

Definition 4 (Path histogram distance (Kawaguchi et al., 2018b))**.** For a tree T, a *path histogram* $\mathcal{H}_{\mathrm{P}}(T)$ of T consists of pairs $\langle s, f(s,T)\rangle$ for every $s \in \mathcal{S}(T)$.

For trees T_1 and T_2, a *path histogram distance* $\delta_{\mathrm{P}}(T_1,T_2)$ between T_1 and T_2 is defined as an L_1-distance between $\mathcal{H}_{\mathrm{P}}(T_1)$ and $\mathcal{H}_{\mathrm{P}}(T_2)$:

$$\delta_{\mathrm{P}}(T_1,T_2)=\sum_{s\in\mathcal{S}(T_1)\cup\mathcal{S}(T_2)}|f(s,T_1)-f(s,T_2)|.$$

For $\lambda=|lv(T)|$, it is obvious that $|\mathcal{H}_{\mathrm{P}}(T)| \leq \lambda$ and $\sum_{s\in\mathcal{S}(T)} f(s,T)=\lambda$.

We denote the set $\{T[v] \mid v \in T\}$ of all the complete subtrees in T by $\mathcal{C}(T)$. For $c \in \mathcal{C}(T)$, the number of occurrences of c in T by $f(c,T)$.

Definition 5 (Complete subtree histogram distance (Akutsu et al., 2013))**.** For a tree T, a *complete subtree histogram* $\mathcal{H}_{\mathrm{CS}}(T)$ consists of pairs $\langle s, f(s,T)\rangle$ for every $s \in \mathcal{C}(T)$.

For trees T_1 and T_2, a *complete subtree histogram distance* $\delta_{\mathrm{CS}}(T_1,T_2)$ between trees T_1 and T_2 is defined as an L_1-distance between $\mathcal{H}_{\mathrm{CS}}(T_1)$ and $\mathcal{H}_{\mathrm{CS}}(T_2)$:

$$\delta_{\mathrm{CS}}(T_1,T_2)=\sum_{c\in\mathcal{C}(T_1)\cup\mathcal{C}(T_2)}|f(c,T_1)-f(c,T_2)|.$$

For $n=|T|$, it is obvious that $|\mathcal{H}_{\mathrm{CS}}(T)| \leq n$ and $\sum_{c\in\mathcal{C}(T)} f(c,T)=n$.

We summarize the properties of δ_{P} and δ_{CS} as follows (Akutsu et al., 2013; Kawaguchi et al., 2018a; Kawaguchi et al., 2018b).

Theorem 2. *Let C_1 and C_2 be caterpillars such that $n=\max\{|C_1|,|C_2|\}$ and $\lambda=\max\{|lv(C_1)|,|lv(C_2)|\}$.*

1. *δ_{P} is a metric for caterpillars but not a metric for trees in general.*
2. *δ_{CS} is a metric for trees, so is for caterpillars.*
3. *We can compute $\delta_{\mathrm{P}}(C_1,C_2)$ and $\delta_{\mathrm{CS}}(C_1,C_2)$ in $O(n)$ time.*
4. *$\tau_{\mathrm{TAI}}(C_1,C_2) \leq \delta_{\mathrm{CS}}(C_1,C_2)$.*
5. *There exist C_1 and C_2 such that $\tau_{\mathrm{TAI}}(C_1,C_2)=\delta_{\mathrm{CS}}(C_1,C_2)=1$ but $\delta_{\mathrm{P}}(C_1,C_2)=O(\lambda)$.*
6. *There exist C_1 and C_2 such that $\delta_{\mathrm{P}}(C_1,C_2)=2$ but $\tau_{\mathrm{TAI}}(C_1,C_2)=\delta_{\mathrm{CS}}(C_1,C_2)=O(n)$.*

3 LCA HISTOGRAM DISTANCE

Let T be a tree. Then, we say that $p=((l_1,d_1):\{(l_2,d_2),(l_3,d_3)\})$ is an *LCA pivot* in T if there exist mutually distinct vertices v and w in T such that $l_1=l(v\sqcup w)$, $d_1=d(v\sqcup w)$, $l_2=l(v)$, $d_2=d(v)$, $l_3=l(w)$ and $d_3=d(w)$, respectively. We denote p by a 6-tuple $(l_1d_1:l_2d_2\sqcup l_3d_3)$ simply. In this case, we also say that p *occurs in* T and denote p by $p(v,w)$. We denote the number of the occurrences of p in T by $f(p,T)$. Furthermore, we denote the set of all the LCA pivots in T by $\mathcal{P}(T)$, that is, $\mathcal{P}(T)=\{p(v,w) \mid (v,w) \in T\times T, v\neq w\}$.

Definition 6 (LCA histogram distance)**.** For a tree T, an *LCA histogram* $\mathcal{H}_{\mathrm{LCA}}(T)$ of T consists of a pair $\langle p, f(p,T)\rangle$ for every $p \in \mathcal{P}(T)$.

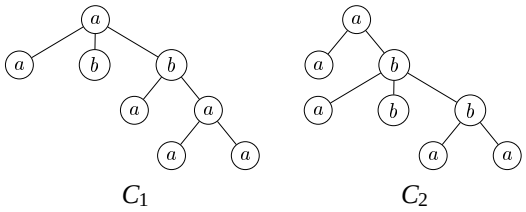
For two trees T_1 and T_2, an *LCA histogram distance* $\delta_{\mathrm{LCA}}(T_1,T_2)$ between T_1 and T_2 is defined as an L_1-distance between $\mathcal{H}_{\mathrm{LCA}}(T_1)$ and $\mathcal{H}_{\mathrm{LCA}}(T_2)$:

$$\delta_{\mathrm{LCA}}(T_1,T_2)=\sum_{p\in\mathcal{P}(T_1)\cup\mathcal{P}(T_2)}|f(p,T_1)-f(p,T_2)|.$$

For $n=|T|$, it is obvious that $|\mathcal{H}_{\mathrm{LCA}}(T)| \leq n(n-1)/2$ and $\sum_{p\in\mathcal{P}(T)} f(p,T)=n(n-1)/2$.

Example 1. Let C_1 and C_2 be caterpillars illustrated in Figure 1.

Then, we obtain the histograms $\mathcal{H}_{\mathrm{LCA}}(C_1)$ and $\mathcal{H}_{\mathrm{LCA}}(C_2)$ illustrated in Table 1. Note



C_1 C_2

Figure 1: The caterpillars C_1 and C_2 in Example 1.

that, since $|C_1| = |C_2| = 8$, it holds that $\sum_{p \in \mathcal{P}(C_i)} f(p, C_i) = {}_8C_2 = 28$ for $i = 1, 2$. Also, the bold faces illustrate the LCA pivots occurring in either $\mathcal{H}_{\mathrm{LCA}}(C_1)$ or $\mathcal{H}_{\mathrm{LCA}}(C_2)$ and its frequency, or the frequencies of the LCA pivot if they are different in $\mathcal{H}_{\mathrm{LCA}}(C_1)$ and $\mathcal{H}_{\mathrm{LCA}}(C_2)$.

Table 1: The histograms $\mathcal{H}_{\mathrm{LCA}}(C_1)$ and $\mathcal{H}_{\mathrm{LCA}}(C_2)$.

$\mathcal{H}_{\mathrm{LCA}}(C_1)$

LCA pivots	freq.
$(a0 : a1 \sqcup b1)$	**2**
$(a0 : a1 \sqcup a2)$	**2**
$(a0 : a1 \sqcup a3)$	2
$(\mathbf{a0} : \mathbf{b1} \sqcup \mathbf{b1})$	**1**
$(\mathbf{a0} : \mathbf{b1} \sqcup \mathbf{a2})$	**2**
$(\mathbf{a0} : \mathbf{b1} \sqcup \mathbf{a3})$	**2**
$(a0 : a0 \sqcup a1)$	1
$(a0 : a0 \sqcup b1)$	**2**
$(a0 : a0 \sqcup a2)$	**2**
$(a0 : a0 \sqcup a3)$	2
$(\mathbf{b1} : \mathbf{a2} \sqcup \mathbf{a2})$	**1**
$(b1 : a2 \sqcup a3)$	2
$(b1 : b1 \sqcup a2)$	**2**
$(b1 : b1 \sqcup a3)$	2
$(\mathbf{a2} : \mathbf{a3} \sqcup \mathbf{a3})$	**1**
$(\mathbf{a2} : \mathbf{a2} \sqcup \mathbf{a3})$	**2**

$\mathcal{H}_{\mathrm{LCA}}(C_2)$

LCA pivots	freq.
$(a0 : a1 \sqcup b1)$	**1**
$(a0 : a1 \sqcup a2)$	**1**
$(\mathbf{a0} : \mathbf{a1} \sqcup \mathbf{b2})$	**2**
$(a0 : a1 \sqcup a3)$	2
$(a0 : a0 \sqcup a1)$	1
$(a0 : a0 \sqcup b1)$	**1**
$(a0 : a0 \sqcup a2)$	**1**
$(\mathbf{a0} : \mathbf{a0} \sqcup \mathbf{b2})$	**2**
$(a0 : a0 \sqcup a3)$	2
$(\mathbf{b1} : \mathbf{a2} \sqcup \mathbf{b2})$	**2**
$(b1 : a2 \sqcup a3)$	2
$(\mathbf{b1} : \mathbf{b2} \sqcup \mathbf{b2})$	**1**
$(\mathbf{b1} : \mathbf{b2} \sqcup \mathbf{a3})$	**2**
$(b1 : b1 \sqcup a2)$	**1**
$(\mathbf{b1} : \mathbf{b1} \sqcup \mathbf{b2})$	**2**
$(b1 : b1 \sqcup a3)$	2
$(\mathbf{b2} : \mathbf{a3} \sqcup \mathbf{a3})$	**1**
$(\mathbf{b2} : \mathbf{b2} \sqcup \mathbf{a3})$	**2**

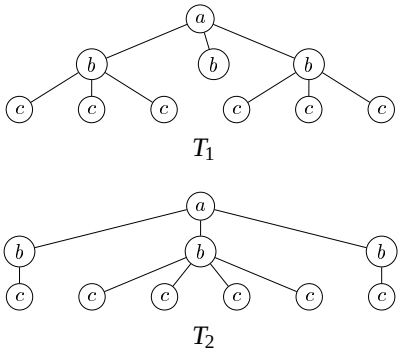
Hence, it holds that:

$$\begin{aligned} & \delta_{\mathrm{LCA}}(C_1, C_2) \\ = & \sum_{p \in \mathcal{P}(C_1) \cup \mathcal{P}(C_2)} |f(p, C_1) - f(p, C_2)| \\ = & \sum_{p \in \mathcal{P}(C_1) \setminus \mathcal{P}(C_2)} f(p, C_1) + \sum_{p \in \mathcal{P}(C_2) \setminus \mathcal{P}(C_1)} f(p, C_2) \\ & + \sum_{p \in \mathcal{P}(C_1) \cap \mathcal{P}(C_2)} |f(p, C_1) - f(p, C_2)| \\ = & \ 9 + 14 + 5 = 28. \end{aligned}$$

Whereas the LCA histogram distance seems to be a metric, for example, Theorem 3.2 in (Tatikonda and Parthasarathy, 2010), we show that it is not a metric for trees as follows.

Theorem 3. *There exist trees T_1 and T_2 such that $\mathcal{H}_{\mathrm{LCA}}(T_1) = \mathcal{H}_{\mathrm{LCA}}(T_2)$ but $T_1 \neq T_2$. Hence, the LCA histogram distance is not a metric for trees in general.*

Proof. Consider the trees T_1 and T_2 in Figure 2.



T_1

T_2

Figure 2: Trees T_1 and T_2.

Then, we obtain the histogram $\mathcal{H}_{\mathrm{LCA}}(T_1)(= \mathcal{H}_{\mathrm{LCA}}(T_2))$ illustrated in Table 2.

Table 2: The histogram $\mathcal{H}_{\mathrm{LCA}}(T_1)(= \mathcal{H}_{\mathrm{LCA}}(T_2))$.

LCA pivots	freq.
$(a0 : c2 \sqcup c2)$	9
$(a0 : b1 \sqcup c2)$	12
$(a0 : b1 \sqcup b1)$	3
$(a0 : a0 \sqcup b1)$	3

LCA pivots	freq.
$(a0 : a0 \sqcup c2)$	6
$(b1 : c2 \sqcup c2)$	6
$(b1 : c2 \sqcup b1)$	6

Here, since $|T_1| = |T_2| = 10$, it holds that $\sum_{p \in \mathcal{P}(T_i)} f(p, T_i) = {}_{10}C_2 = 45$ for $i = 1, 2$. Furthermore, since the labels are not essential, this statement also holds for unlabeled trees. □

On the other hand, note that neither T_1 nor T_2 in Figure 2 is a caterpillar. In the remainder of this section, we discuss the LCA histogram distance between caterpillars.

For caterpillars, the following lemma is obvious.

Lemma 1. *Let $p(v, w) = (l_1 d_1 : l_2 d_2 \sqcup l_3 d_3) \in \mathcal{P}(C)$ an LCA pivot in C. Then, the following statements hold.*

1. *It holds that $v \sqcup w \in bb(C)$.*
2. *If $v, w \in lv(C)$, then it holds that $d_1 = \min\{d_2, d_3\} - 1$. Also it holds that $v \sqcup w = par(v)$ if $d_2 < d_3$, $v \sqcup w = par(w)$ if $d_3 < d_2$ and $v \sqcup w = par(v) = par(w)$ if $d_2 = d_3$.*
3. *If $v, w \in bb(C)$, then it holds that $d_1 = \min\{d_2, d_3\}$. Also it holds that $v \sqcup w = v$ if $d_2 < d_3$ and $v \sqcup w = w$ if $d_3 < d_2$.*

4. *Suppose that $v \in lv(C)$ and $w \in bb(C)$. If $d_2 < d_3$, then it holds that $d_1 = d_3$ and $v \sqcup w = w$. Otherwise, that is, $d_2 \geq d_3$, it holds that $d_1 = d_2 - 1$ and $v \sqcup w = par(v)$.*

Then, the following theorem holds.

Theorem 4. *For caterpillars, the LCA histogram distance is a metric.*

Proof. By the definition, it is sufficient to show that two caterpillars C_1 and C_2 are isomorphic iff $\delta_{\mathrm{LCA}}(C_1,C_2) = 0$. In other words, it is sufficient to show that we can transform a caterpillar C from $\mathcal{H}_{\mathrm{LCA}}(C)$ uniquely.

By Lemma 1.1, we can uniquely determine $bb(C)$ from $\mathcal{P}(C)$ because of l_1 and d_1 in $p(v,w)$. Since $lv(C) = C \setminus bb(C)$, we can determine $lv(C)$. Then, by Lemma 1.2, we can determine the set of leaves with depth i for every i ($1 \leq i \leq d(C)$). □

For $\lambda_i = |lv(C_i)|$ and $n_i = |C_i|$ ($i = 1,2$), it holds that $0 \leq \delta_{\mathrm{P}}(C_1,C_2) \leq \lambda_1 + \lambda_2$, $0 \leq \delta_{\mathrm{CS}}(C_1,C_2) \leq n_1 + n_2$ and $0 \leq \delta_{\mathrm{LCA}}(C_1,C_2) \leq (n_1(n_1-1) + n_2(n_2-1))/2$. Then, consider the extreme cases in Section 1.

Example 2. Let C_1 and C_2 be paths with length n.

Suppose that every vertex in C_1 is labeled by a and that in C_2 by b. Then, it holds that $\tau_{\mathrm{TAI}}(C_1,C_2) = n$, $\delta_{\mathrm{P}}(C_1,C_2) = 1$, $\delta_{\mathrm{CS}}(C_1,C_2) = 2n$ and $\delta_{\mathrm{LCA}}(C_1,C_2) = 2n(n-1)$. Note that $\delta_{\mathrm{P}}(C_1,C_2)$, $\delta_{\mathrm{CS}}(C_1,C_2)$ and $\delta_{\mathrm{LCA}}(C_1,C_2)$ are their maximum values.

Suppose that every vertex in C_1 and every non-leaf vertex in C_2 is labeled by a and the leaf of C_2 is labeled by b. Then, it holds that $\tau_{\mathrm{TAI}}(C_1,C_2) = 1$, $\delta_{\mathrm{P}}(C_1,C_2) = 1$, $\delta_{\mathrm{CS}}(C_1,C_2) = 2n$ and $\delta_{\mathrm{LCA}}(C_1,C_2) = 2(n-1)$. Note that $\delta_{\mathrm{P}}(C_1,C_2)$ and $\delta_{\mathrm{CS}}(C_1,C_2)$ are their maximum values but $\delta_{\mathrm{LCA}}(C_1,C_2)$ is not.

In particular, $\delta_{\mathrm{P}}(C_1,C_2)$ cannot distinguish the difference of labels between two paths C_1 and C_2.

Furthermore, the following theorem holds.

Theorem 5. *There exist caterpillars C_1 and C_2 satisfying the following statements.*

1. *$\tau_{\mathrm{TAI}}(C_1,C_2) = \delta_{\mathrm{CS}}(C_1,C_2) = 1$ but $\delta_{\mathrm{P}}(C_1,C_2)$ and $\delta_{\mathrm{LCA}}(C_1,C_2)$ are their maximum values.*
2. *$\delta_{\mathrm{P}}(C_1,C_2)$ and $\delta_{\mathrm{CS}}(C_1,C_2)$ are their maximum values but $\tau_{\mathrm{TAI}}(C_1,C_2)$ and $\delta_{\mathrm{LCA}}(C_1,C_2)$ are not.*

Proof. 1. Let C_1 and C_2 be stars, that is, $|bb(C_1)| = |bb(C_2)| = 1$, such that $r(C_1) = r_1$, $r(C_2) = r_2$, $l(r_1) \neq l(r_2)$, $ch(r_1) = ch(r_2)$ and $|ch(r_1)| = |ch(r_2)| = n-1$. Then, it is obvious that $\tau_{\mathrm{TAI}}(C_1,C_2) = \delta_{\mathrm{CS}}(C_1,C_2) = 1$ and $\delta_{\mathrm{P}}(C_1,C_2) = 2(n-1)$. Also, since $\mathcal{P}(C_1) \cap \mathcal{P}(C_2) = \emptyset$, it holds that $\delta_{\mathrm{LCA}}(C_1,C_2) = 2n(n-1)$.

2. Let C_1 and C_2 be caterpillars obtained by connecting λ leaves to the leaves of paths with length h, where every vertex in C_1 and in a path in C_2 is labeled by a and every leaf in C_2 by b. Then, $|C_1| = |C_2| = h + \lambda = n$. It is obvious that $\delta_{\mathrm{P}}(C_1,C_2) = 2\lambda$ and $\delta_{\mathrm{CS}}(C_1,C_2) = 2(\lambda + h) = 2n$, so they are the maximum values. On the other hand, it holds that $\tau_{\mathrm{TAI}}(C_1,C_2) = \lambda$ and $\delta_{\mathrm{LCA}}(C_1,C_2) = 2\lambda(n-1)$, where their maximum values are $2n-1$ and $2n(n-1)$. □

By selecting every pair of vertices in two caterpillars, we can compute $\delta_{\mathrm{LCA}}(C_1,C_2)$ in $O(n^2)$ time, because $\mathcal{H}_{\mathrm{LCA}}(C) = O(n^2)$.

Note that the inequality that $\delta_{\mathrm{P}} < \delta_{\mathrm{CS}} < \delta_{\mathrm{LCA}}$ tends to hold by the values of δ_{P}, δ_{CS} and δ_{LCA}. Then, we normalize δ_{P}, δ_{CS} and δ_{LCA} by dividing their maximum values when comparing distances. We denote the normalized distances of δ_{P}, δ_{CS} and δ_{LCA} by δ^*_{P}, δ^*_{CS} and δ^*_{LCA}, respectively. Then, the following example shows that the inequality that $\delta^*_{\mathrm{P}} < \delta^*_{\mathrm{CS}} < \delta^*_{\mathrm{LCA}}$ does not always hold.

Example 3. Consider caterpillars C_1, C_2 and C_3 in Figure 3.

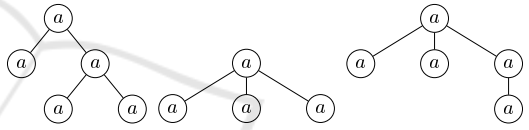


Figure 3: Caterpillars C_1, C_2 and C_3 in Example 3.

Then, we obtain $\delta_{\mathrm{P}}(C_i,C_j)$, $\delta_{\mathrm{CS}}(C_i,C_j)$, $\delta_{\mathrm{LCA}}(C_i,C_j)$, $\delta^*_{\mathrm{P}}(C_i,C_j)$, $\delta^*_{\mathrm{CS}}(C_i,C_j)$ and $\delta^*_{\mathrm{LCA}}(C_i,C_j)$ for $(i,j) = (1,2),(1,3),(2,3)$ as follows.

(i,j)	δ_{P}	δ_{CS}	δ_{LCA}	δ^*_{P}	δ^*_{CS}	δ^*_{LCA}
$(1,2)$	4	3	10	$2/3$	$1/3$	$5/8$
$(1,3)$	2	4	6	$1/3$	$2/5$	$3/10$
$(2,3)$	2	3	4	$1/3$	$1/3$	$1/4$

Hence, the following statements hold:

$\delta^*_{\mathrm{CS}}(C_1,C_2) < \delta^*_{\mathrm{LCA}}(C_1,C_2) < \delta^*_{\mathrm{P}}(C_1,C_2)$,
$\delta^*_{\mathrm{LCA}}(C_1,C_3) < \delta^*_{\mathrm{P}}(C_1,C_3) < \delta^*_{\mathrm{CS}}(C_1,C_3)$,
$\delta^*_{\mathrm{LCA}}(C_2,C_3) < \delta^*_{\mathrm{P}}(C_2,C_3) = \delta^*_{\mathrm{CS}}(C_2,C_3)$.

4 EXPERIMENTAL RESULTS

Table 3 illustrates the number (#cat) of caterpillars in the datasets in N-glycans and all of the glycans from KEGG[1], CSLOGS[2] and dblp[3] datasets, whose number of data is denoted by #data.

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

[2]http://www.cs.rpi.edu/~zaki/www-new/pmwiki.php/Software/Software

[3]http://dblp.uni-trier.de/

Table 3: The number of caterpillars in N-glycans and all-glycans from KEGG, CSLOGS and dblp datasets.

dataset	#cat	#data	%
N-glycans	514	2,142	23.996
all-glycans	8,005	10,704	74.785
CSLOGS	41,592	59,691	69.679
dblp	5,154,295	5,154,530	99.995

We deal with caterpillars for N-glycans, all-glycans, CSLOGS and the selected 50,000 caterpillars in dblp (we refer to dblp$^-$). Table 4 illustrates the information of such caterpillars. Here, $([a,b];c)$ means that a, b and c are the minimum, the maximum and the average number.

In the remainder of this section, we compare the LCA histogram distance with the path histogram distance and the complete subtree histogram distance for caterpillars.

Table 5 illustrates the running time of computing δ^*_{P}, δ^*_{CS} and δ^*_{LCA} for N-glycans, all-glycans, CSLOGS and dblp$^-$.

Table 5 shows that, whereas we compute δ^*_{P} and δ^*_{CS} in linear time and δ^*_{LCA} in quadratic time in theoretical, the running time of computing δ^*_{LCA} is within twice for N-glycans and all-glycans, within thrice for CSLOGS and about seven times for dblp$^-$, respectively, of computing δ^*_{CS} in experimental. The reason why the running time of computing δ^*_{LCA} is not so large is that the number of $|\mathcal{H}_{\mathrm{LCA}}|$ is not so large except dblp$^-$; For dblp$^-$, $|\mathcal{H}_{\mathrm{LCA}}|$ is larger than others because the number of leaves is large but the height is small in Table 4.

Figure 4 illustrates the distributions of δ^*_{P}, δ^*_{CS} and δ^*_{LCA} for N-glycans, all-glycans, CSLOGS and dblp$^-$.

Figure 4 shows that almost of the distributions concentrate near to 1, in particular, CSLOGS and dblp$^-$. On the other hand, for dblp$^-$, the distributions appear near to 0. For N-glycans and all-glycans, δ^*_{P} is larger than δ^*_{CS} and δ^*_{CS} is larger than δ^*_{LCA}.

Figure 5 illustrates the detailed distributions of δ^*_{P}, δ^*_{CS} and δ^*_{LCA} for N-glycans, all-glycans, CSLOGS and dblp$^-$, where the scopes of the distances of N-glycans, all-glycans, CSLOGS and dblp$^-$ are $[0.8,1]$, $[0.9,1]$, $[0.995,1]$ and $[0.99.1]$, respectively.

Note that, for dblp$^-$, since the maximum value of δ^*_{CS} is 0.992308 and the frequency is low, the distribution is just of δ^*_{P} and δ^*_{LCA}. Figure 5 shows that, near to 1 and for N-glycans, all-glycans and CSLOGS, the inequality of $\delta^*_{\mathrm{P}} < \delta^*_{\mathrm{CS}} < \delta^*_{\mathrm{LCA}}$ holds.

Figure 6 illustrates the scatter charts of δ^*_{P}, δ^*_{CS} and δ^*_{LCA} for N-glycans and all-glycans and Figure 7 illustrates those for CSLOGS and dblp$^-$, and their correlation coefficients (*cc*). Here, the representation of

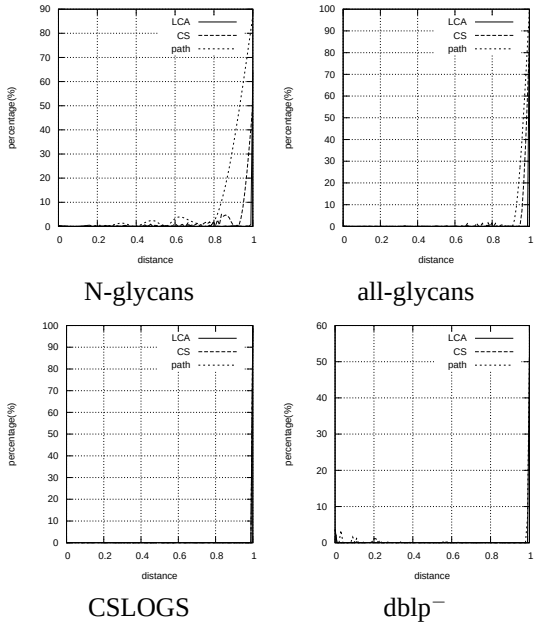


Figure 4: The distributions of δ^*_{P}, δ^*_{CS} and δ^*_{LCA} for N-glycans, all-glycans, CSLOGS and dblp$^-$.

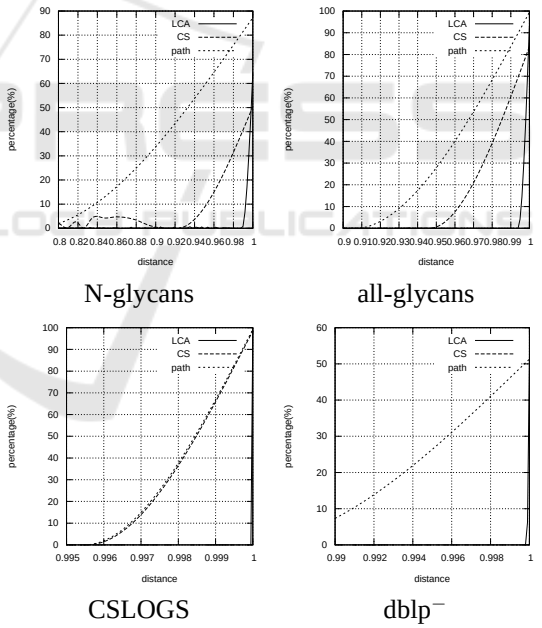


Figure 5: The detailed distributions of δ^*_{P}, δ^*_{CS} and δ^*_{LCA} for N-glycans, all-glycans, CSLOGS and dblp$^-$.

$\delta^*_{\mathrm{Y}}/\delta^*_{\mathrm{X}}$ means that the number of pairs of caterpillars with δ^*_{X} is pointed at the x-axis and that with δ^*_{Y} is pointed at the y-axis.

Figures 6 and 7 show that, the scatter charts for N-glycans and all-glycans in Figure 6 are more sparse than those for CSLOGS and dblp$^-$ in Figure 7, because the number of caterpillars in N-glycans and all-glycans is much smaller than that in CSLOGS

Table 4: The information of caterpillars in N-glycans, all-glycans, CSLOGS and dblp$^-$.

dataset	#vertices	degree	height	#leaves	#labels
N-glycans	([6,15];6.40)	([1,3];1.84)	([1,9];4.22)	([1,7];2.18)	([2,8];4.50)
all-glycans	([1,24];4.74)	([0,5];1.49)	([0,15];3.02)	([1,14];1.72)	([1,9];2.84)
CSLOGS	([2,404];5.84)	([1,403];3.05)	([1,70];2.20)	([1,403];3.64)	([2,168];5.18)
dblp$^-$	([7,244];11.96)	([6,243];10.94)	([1,3];1.02)	([6,243];10.94)	([7,13];9.86)

Table 5: The running time of computing δ_P^*, δ_{CS}^* and δ_{LCA}^* (msec.).

dataset	δ_P^*	δ_{CS}^*	δ_{LCA}^*
N-glycans	142	239	419
all-glycans	34,113	40,364	73,219
CSLOGS	1,017,730	1,361,343	3,439,560
dblp$^-$	1,980,062	3,534,120	24,633,812

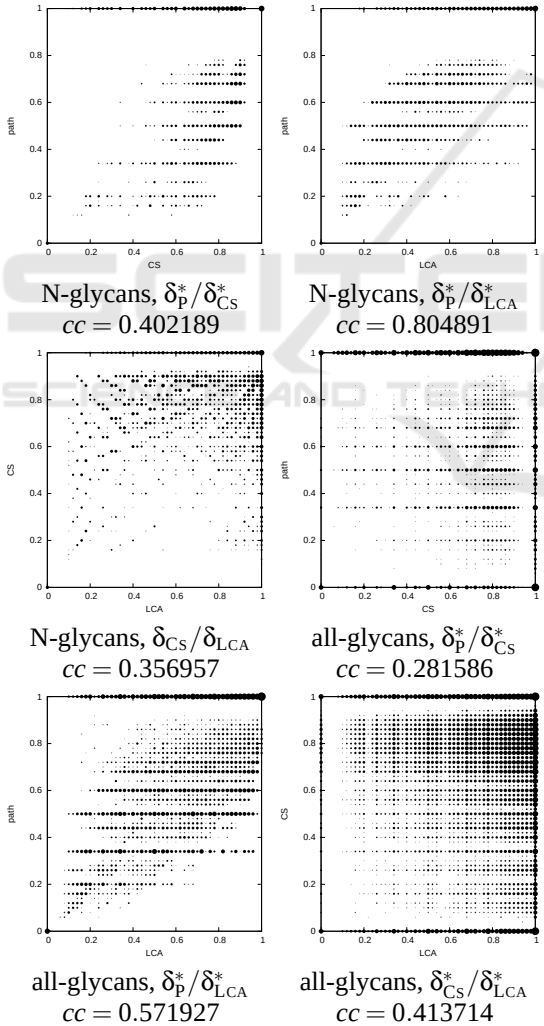


Figure 6: The scatter charts of δ_P^*, δ_{CS}^* and δ_{LCA}^* for N-glycans and all-glycans.

and dblp$^-$. Also for all datasets, the scatter chart for $\delta_{CS}^*/\delta_{LCA}^*$ spreads more widely than those for $\delta_P^*/\delta_{LCA}^*$ and δ_P^*/δ_{CS}^*.

For Figure 6, the scatter charts for N-glycans have the values on the line that $y = 1$ and, in particular, the scatter charts of $\delta_{CS}^*/\delta_{LCA}^*$ also have the values on the line that $x = 1$. On the other hand, the scatter charts for all-glycans have the values on the line that $y = 1$, those of δ_P^*/δ_{CS}^* and $\delta_{CS}^*/\delta_{LCA}^*$ the vales on the lines that $x = 1$ and $y = 0$.

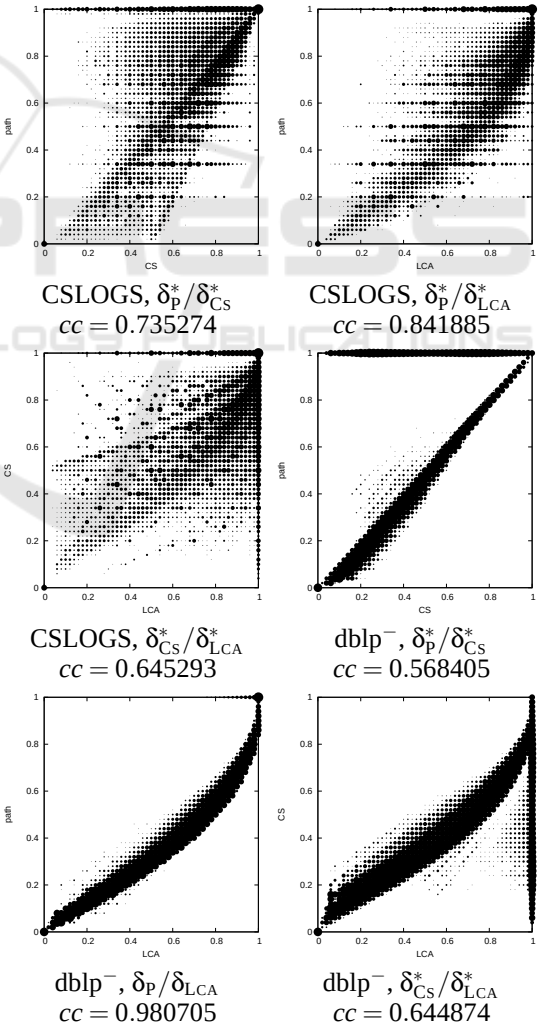


Figure 7: The scatter charts δ_P^*, δ_{CS}^* and δ_{LCA}^* for CSLOGS and dblp$^-$ and their correlation coefficients (cc).

For Figure 7, the scatter charts for CSLOGS have the values on the line that $y = 1$ and those of $\delta^*_{\mathrm{CS}}/\delta^*_{\mathrm{LCA}}$ have the values on the line that $x = 1$. On the other hand, the scatter charts of $\delta^*_{\mathrm{P}}/\delta^*_{\mathrm{CS}}$ for dblp$^-$ have the values on the line that $y = 1$ and those of $\delta^*_{\mathrm{CS}}/\delta^*_{\mathrm{LCA}}$ have the values on the line that $x = 1$. In particular, the scatter charts for dblp$^-$ constitutes at most two clusters, where one lies on the axis.

For correlation coefficients, which we denote by $cc(\delta^*_{\mathrm{Y}}/\delta^*_{\mathrm{X}})$, the value of $cc(\delta^*_{\mathrm{P}}/\delta^*_{\mathrm{LCA}})$ is highest for all the data. On the other hand, it holds that

$$cc(\delta^*_{\mathrm{CS}}/\delta^*_{\mathrm{LCA}}) < cc(\delta^*_{\mathrm{P}}/\delta^*_{\mathrm{CS}}) < cc(\delta^*_{\mathrm{P}}/\delta^*_{\mathrm{LCA}})$$

for N-glycans and CSLOGS, whereas it holds that

$$cc(\delta^*_{\mathrm{P}}/\delta^*_{\mathrm{CS}}) < cc(\delta^*_{\mathrm{CS}}/\delta^*_{\mathrm{LCA}}) < cc(\delta^*_{\mathrm{P}}/\delta^*_{\mathrm{LCA}})$$

for all-glycans and dblp$^-$. For the values of correlation coefficients, almost of the distances are related for CSLOGS and dblp$^-$, because $cc(\delta_{\mathrm{X}}/\delta_{\mathrm{Y}})$ is greater than 0.6, just δ^*_{LCA} is related with δ^*_{P} for N-glycans, and no distances are related for all-glycans. In particular, $cc(\delta^*_{\mathrm{P}}\delta^*_{\mathrm{LCA}})$ is greater than 0.8 for N-glycans, CSLOGS and dblp$^-$.

5 CONCLUSION

In this paper, we have introduced an LCA histogram distance δ_{LCA} between trees and shown that it is not a metric for trees but is a metric for caterpillars. Furthermore, we have given experimental results of computing δ_{LCA} for caterpillars, by comparing the path histogram distance δ_{P} and the complete subtree histogram distance δ_{CS} (or their normalized distances δ^*_{LCA}, δ^*_{P} and δ^*_{CS}).

It is a future work to design the algorithm to compute δ_{LCA} more efficiently, without constructing LCA histograms explicitly, for example. It is also a future work to analyze the relationship between δ_{LCA}, δ_{P} and δ_{CS} (or δ^*_{LCA}, δ^*_{P} and δ^*_{CS}) in more detail in experimental, in particular, as stated in Section 4, to analyze why the correlation coefficients of δ^*_{P} and δ^*_{LCA} have been high, and that in theoretical.

Furthermore, it is a future work to give experimental results for other data of caterpillars. Finally, it is an important future work to analyze the relationship between δ_{LCA} and τ_{TAI} (Muraka et al., 2018).

ACKNOWLEDGEMENTS

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

REFERENCES

Akutsu, T., Fukagawa, D., Halldórsson, M. M., Takasu, A., and Tanaka, K. (2013). Approximation and parameterized algorithms for common subtrees and edit distance between unordered trees. *Theoret. Comput. Sci.*, 470:10–22.

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.

Gallian, J. A. (2007). A dynamic survey of graph labeling. *Electorn. J. Combin.*, 14:DS6.

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., Yoshino, T., and Hirata, K. (2018a). Path histogram distance and complete subtree histogram distance for rooted labeled caterpillars. *(submitted)*.

Kawaguchi, T., Yoshino, T., and Hirata, K. (2018b). Path histogram distance for rooted labeled caterpillars. In *Proc. ACIIDS'18 (LNAI 10751)*, pages 276–286.

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.

Muraka, K., Yoshino, T., and Hirata, K. (2018). Computing edit distance between rooted labeled caterpillars. In *Proc. FedCSIS'18* (to appear).

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

Tatikonda, S. and Parthasarathy, S. (2010). Hashing tree-structured data: Methods and applications. In *Proc. ICDM'10*, pages 429–440.

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