RAIRO-Inf. Theor. Appl. **40** (2006) 593-609
DOI: 10.1051/ita:2006043

EDIT DISTANCE BETWEEN UNLABELED ORDERED TREES

Anne Micheli[1] and Dominique Rossin[1]

Abstract. There exists a bijection between one-stack sortable permutations (permutations which avoid the pattern (231)) and rooted plane trees. We define an edit distance between permutations which is consistent with the standard edit distance between trees. This one-to-one correspondence yields a polynomial algorithm for the subpermutation problem for (231) pattern-avoiding permutations.

Moreover, we obtain the generating function of the edit distance between ordered unlabeled trees and some special ones. For the general case we show that the mean edit distance between a rooted plane tree and all other rooted plane trees is at least $n/ln(n)$.

Some results can be extended to labeled trees considering colored Dyck paths or, equivalently, colored one-stack sortable permutations.

Mathematics Subject Classification. 05C12, 05C05, 05A05, 05A15.

1. Introduction

The edit distance between two labeled trees is the minimal number of edit operations necessary to transform one tree into the other. The edit operations are deletion (edge contraction), insertion of an edge and relabeling of a vertex.

The main problem is to find efficient algorithms to compute this distance between ordered labeled trees. Many algorithms have been proposed [6, 14]. The basic idea of all these dynamic algorithms arises from the paper of Zhang and Shasha [14]. Further improvements have been made [6].

Comparing the structure of molecules and finding the preserved ones during a genetic mutation can be seen as an edit distance problem. The application

Keywords and phrases. Edit distance, trees.

[1] CNRS, LIAFA, Université Paris 7, 2 Place Jussieu, 75251 Paris Cedex 05, France;
amicheli,rossin@liafa.jussieu.fr

field of this problem is not restricted to biology: in computer vision, objects are represented by their skeletons, which are trees, and in computer science, edit distance is used to compare structural similarities between XML documents [5].

But no combinatorial interpretation has yet been found. In this article, we introduce one-stack sortable permutations [2, 13]. These one-stack sortable permutations are (231) pattern-avoiding permutations and we show that they are in one-to-one correspondence with ordered trees.

Moreover the edit operations on unlabeled trees can be easily described in terms of one-stack sortable permutations. This leads to a purely combinatorial explanation of the edit distance between unlabeled trees. For labeled ones, a similar correspondence can be given and is briefly discussed in the conclusion.

Some polynomial algorithms are known to compute the edit distance between trees [14]. By our correspondence, we show that computing the greatest common pattern between two (231)-avoiding permutations is also polynomial whereas it is NP-complete for general permutations [1].

In Section 2, we recall basic definitions on trees and permutations. In the main Section 3, we first transpose the edit operations on one-stack sortable permutations and then characterize the edit distance on permutations in terms of permutation patterns. The Section 4 gives a lower bound on the average edit distance between random trees. The last section is devoted to the study of the distribution of the edit distance between the path (resp. the star) and a random tree.

2. Definitions

2.1. One-stack sortable permutations

We describe in this section an encoding for rooted plane trees. A *rooted plane tree* is en embedding of a tree in the plane with a distinguished vertex called the *root*. A rooted plane tree is also called an *ordered tree* because the children of each vertex are linearly ordered from left to right by the embedding. We number the edges of the tree by a postfix traversal and then read the permutation by a prefix traversal. The permutations so obtained are called one-stack sortable permutations [2, 13]. An alternate definition is the following:

Definition 2.1. For any $n \in \mathbb{N}$, a one-stack sortable permutation on $\{1 \dots n\}$ is a permutation σ such that $\sigma = InJ$ where I and J are one-stack sortable permutations on $\{1 \dots p\}$ and $\{p+1 \dots n-1\}$ respectively. Note that I or J could be empty.

Note that in the sequel, permutations are seen as words.

Theorem 2.2 (see Fig. 3)**.** *One-stack sortable permutations are in one-to-one correspondence with ordered trees.*

Proof. Given a tree T with n edges, number the edges by a postfix Depth First Search Traversal (DFS). Read it again by a prefix DFS. Note that the DFS,

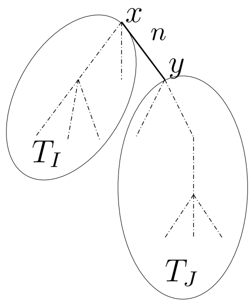


Figure 1. Coding a tree with a one-stack sortable permutation.

whether prefix or postfix, respects the order on the edges hanging on each vertex. It is clear that the permutation so obtained is of the form InJ. Moreover I corresponds to the encoding by a postfix DFS of the left subtree T_I as shown in Figure 1. The same goes for J but its numbers are shifted by $|I|$.

Conversely, take a one-stack sortable permutation $\sigma = InJ$.

- If $\sigma = k$ then the corresponding tree is a single edge.
- If $\sigma = InJ$ then the corresponding tree $\mathcal{T}(\sigma)$ is the tree obtained by taking an edge $e = (xy)$ (corresponding to n) where x is the root of $\mathcal{T}(\sigma)$. Since I and J are also one-stack sortable permutations, we can recursively build the corresponding trees T_I and T_J. Put them at each end of the edge e, *i.e.* T_I is hung on x such that e is the rightmost edge of x, and T_J on y.

This construction is unique. □

If σ is a one-stack sortable permutation, let $\mathcal{T}(\sigma)$ denote the tree associated to σ. Conversely, if T is a tree, its associated one-stack sortable permutation is denoted by $\Theta(T)$. Moreover, in the sequel, σ_k will either denote the k-th letter of the word σ or the corresponding edge in $\mathcal{T}(\sigma)$.

Definition 2.3. A *subsequence* of a permutation $\sigma = \sigma_1 \dots \sigma_n$ is a word $\sigma' = \sigma_{i_1} \dots \sigma_{i_k}$ where $i_1, \dots, i_k$ is an increasing sequence of elements of $\{1, \dots, n\}$.

Let Φ be the bijective mapping of $\{\sigma_{i_1}, \sigma_{i_2}, \dots, \sigma_{i_k}\}$ onto $\{1, \dots, k\}$ preserving the order on σ_{i_l}.

$\Phi(\sigma')$ is defined to be the *normalized* subsequence (pattern) $\hat{\sigma'}$.

Remark 2.4. The one-stack sortable permutations are the permutations avoiding the normalized subsequence (pattern) 231 [7].

2.2. Edit distance

We briefly recall the definition of the edit distance between unlabeled trees. Given two unlabeled trees, the *edit distance* is the minimal number of operations necessary to transform one into the other. The operations are (see Fig. 2):

- **Deletion**: this is the contraction of an edge; two vertices are merged.
- **Insertion**: this is the converse operation of deletion.

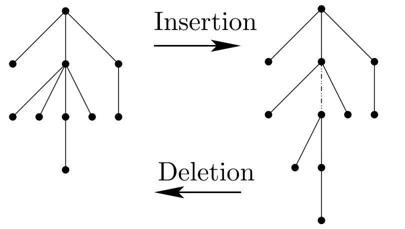
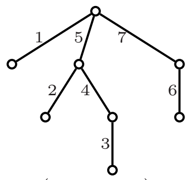


Figure 2. Insertion and Deletion operations on a tree.



(1524376)

Figure 3. Tree associated with $\sigma = (1524376)$.

A cost can be given to each operation. In this article we define the cost of each operation to be 1.

3. Distance on one-stack sortable permutations

Since one-stack sortable permutations are in one-to-one correspondence with rooted plane trees, we define similar edit operations between one-stack sortable permutations and show that these definitions are consistent with the edit distance between trees. Moreover, we give a combinatorial interpretation of the distance.

A *factor* of a permutation $\sigma = \sigma_1\sigma_2\ldots\sigma_n$ is a *factor* of the word $\sigma_1\sigma_2\ldots\sigma_n$ *i.e.* a word of the form $\sigma_k\sigma_{k+1}\ldots\sigma_{k+l}$.

A factor f is *compact* if it is a permutation of an interval of $\mathbb{N}$.

A compact factor f of σ is *complete* if no non-empty factor g of σ verifies both:

(1) fg is compact where fg is the concatenation of the words f and g;
(2) the greatest element of fg is equal to the greatest element of f.

Take for example the one-stack sortable permutation $\sigma = (1524376)$ (see Fig. 3). The complete factors of σ are $\{1\}$, $\{15243\}$, $\{1524376\}$, $\{5243\}$, $\{524376\}$, $\{2\}$, $\{243\}$, $\{43\}$, $\{3\}$, $\{76\}$, $\{6\}$. The compact factors of σ that are not complete are $\{4\},\{5\}$ and $\{7\}$.

Let $T = (V(T), E(T))$ be a rooted plane tree, where $V(T)$ denotes its set of vertices and $E(T)$ its set of edges. A *subtree* $T' = (V(T'), E(T'))$ of T rooted at $v \in V(T')$ is a tree such that $E(T') \subseteq E(T)$, $V(T') \subseteq V(T)$ and the graph $G = (V(T) \setminus V(T') \bigcup \{v\}, E(T) \setminus E(T'))$ is connected.

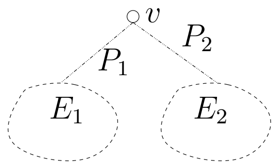


FIGURE 4. Compact factors are connected components.

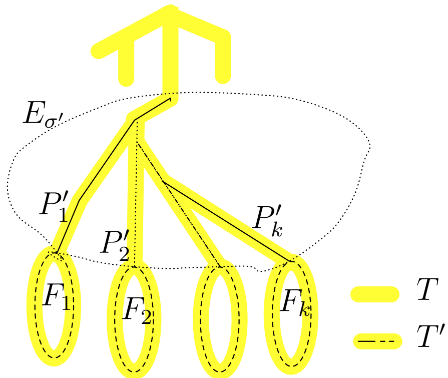


FIGURE 5. Subtree of T induced by $E_{\sigma'}$.

Lemma 3.1. *The set of compact factors of σ is in one-to-one correspondance with the union of two sets: the set of subtrees of $T = \mathcal{T}(\sigma)$ and the set of the internal paths P in T such that each internal vertex of P is of degree 2 in T and P does not end at a leaf (P can be reduced to an internal edge).*

Proof. First let us prove that the subset of edges corresponding to a compact factor is connected.

Let σ' be a compact factor of $\sigma = \Theta(T)$. Let $E_{\sigma'}$ be the set of edges corresponding to σ' in T. Suppose that $E_{\sigma'}$ is not connected. Let E_1 and E_2 be two connected components and v be the nearest common ancestor of E_1 and E_2. Let P_1 (resp. P_2) be the path starting from v and ending at the first vertex of E_1 (resp. E_2). Note that we can choose E_1 and E_2 such that the edges of P_1 and P_2 are not in $E_{\sigma'}$. Suppose that P_1 is at the left of P_2 (see Fig. 4). In the prefix DFS of T, the edges of P_2 are visited between those of E_1 and E_2. Thus they should appear in σ', hence $P_2 = \varnothing$. Thus $v \in E_2$ so that P_1 links E_2 and E_1. In the postfix DFS, the edges of P_1 have labels greater than those of E_1 and less than E_2. If $P_1 \neq \varnothing$, it implies that σ' is not compact. Thus $E_{\sigma'}$ is connected.

Consider the subtree T' of T induced by $E_{\sigma'}$. It consists of $E_{\sigma'}$ plus all the edges of T whose endpoints have an ancestor in $E_{\sigma'}$ as shown in Figure 5.

$E_{\sigma'}$ can be decomposed into $k \geq 1$ edge-disjoint paths P'_i thanks to the prefix DFS (see Fig. 5). F_i is the subtree hanging on P'_i which can be empty.

The prefix DFS of T', which is a factor of σ, gives the associated permutation $\Theta(P'_1)\ \Theta(F_1)\ \Theta(P'_2)\ \Theta(F_2)\dots\ \Theta(P'_k)\ \Theta(F_k)$. So $\sigma' = \Theta(P'_1)\ \Theta(F_1)\ \Theta(P'_2)\ \Theta(F_2)\dots\Theta(P'_k)$, hence $F_i = \varnothing, \forall i < k$.

- Suppose that $F_k \neq \varnothing$. If $k > 1$, then the edges of F_k are visited after at least one edge of P'_1, and before the edges of P'_k in the postfix DFS. Since σ' is compact, it implies that $k = 1$.
- If $F_k = \varnothing$, then $E_{\sigma'}$ is a subtree.

The converse is straightforward. □

Proposition 3.2. *The set of complete factors of σ corresponds to the set of subtrees of the associated tree.*

Proof. Let T' be a subtree of T and $\sigma = \Theta(T)$. The edges of T' are visited consecutively by the postfix (resp. prefix) DFS of T. Thus the sequence of edges of T' is a compact factor $\sigma_k\sigma_{k+1}\dots\sigma_{k+l}$ of σ. σ_{k+l+1} is an edge which is visited after all edges of T' by the prefix DFS. Thus it is the first time this edge is visited by the traversal. Hence, its label is greater than those of T'. Thus $\sigma_k\sigma_{k+1}\dots\sigma_{k+l}$ is complete.

Conversely, let σ' be a complete factor. As σ' is compact, by Lemma 3.1, it corresponds either to a subtree or to an internal path P with a subtree F hanging on P. In addition, $\Theta(P)\Theta(F) = \sigma'\Theta(F)$ is also a compact factor of σ and it has the same maximum as σ' which contradicts the completeness of σ'. □

Remark 3.3. Let σ be a one-stack sortable permutation and $\sigma_k = (p(v_k)v_k)$ an edge, where $p(v_k)$ denote the parent of v_k. Let σ' be the shortest complete factor of σ such that $\sigma' = \sigma_k\sigma_{k+1}\dots\sigma_{k+l}$, where $\sigma_i = (p(v_i)v_i)$. By the previous proposition $\mathcal{T}(\sigma')$ is a subtree of $\mathcal{T}(\sigma)$. The children of v_k are the vertices v_{k+i} such that $i \leq l$ and $\sigma_k > \sigma_{k+i} > \sigma_{k+j}, 1 \leq j \leq i-1$.

Let $\sigma = \sigma_1\dots\sigma_k$ be a word of $\{1\dots n\}$ and a be a letter of $\{1\dots n\}$. We denotes by $[\sigma]_a$ the word $\sigma'_1\dots\sigma'_k$ where

$$\sigma'_i = \begin{cases} \sigma_i \text{ if } \sigma_i < a \\ \sigma_i + 1 \text{ otherwise.} \end{cases}$$

Definition 3.4. We define two operations on permutations which correspond to the standard definition on trees [14]:

(1) *Deletion*: let $1 \leq k \leq n$. The deletion $(\sigma_k \to \Lambda)$ is the removal of σ_k in a permutation σ and the renormalization on S_{n-1} of the result. We will either talk about the deletion of the edge σ_k or the deletion of the vertex v such that σ_k is the edge $(p(v)v)$.

(2) *Insertion (see Fig. 6)*: $(\Lambda \to \varnothing)$ corresponds to the transformation of the permutation $\sigma = \varnothing$ into $\sigma' = (1)$. If $\sigma \neq \varnothing$, let f be a complete factor of σ. Then, $\sigma = ufv$ with u, v factors of σ.

  (a) $(\Lambda \to f)$: the resulting permutation is $\sigma' = [u]_a af[v]_a$, $a = \max\{f\} + 1$. This corresponds to the insertion of an inner vertex with $\mathcal{T}(f)$ as subtree.

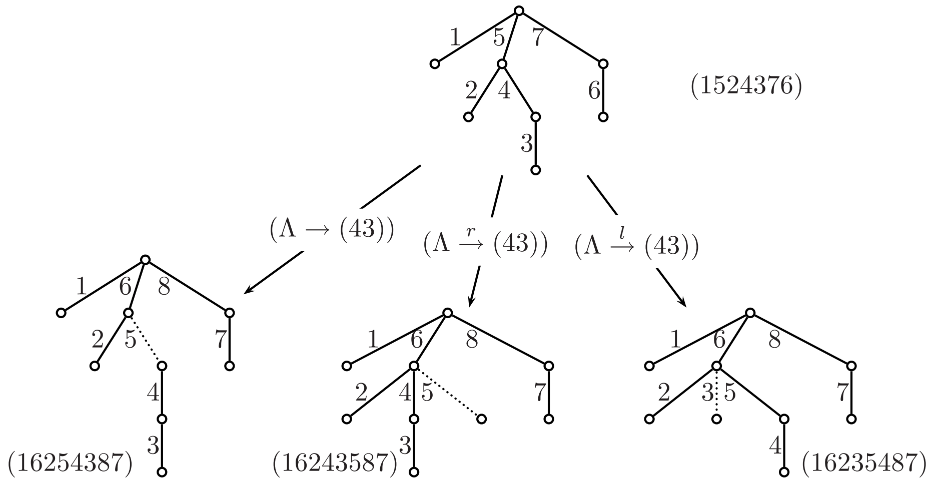


FIGURE 6. Insertion operations for $f = (43)$.

(b) $(\Lambda \xrightarrow{r} f)$: the resulting permutation is $\sigma' = [u]_a f a [v]_a$, $a = \max\{f\} + 1$. This corresponds to the insertion of a leaf as the right sibling of $\mathcal{T}(f)$.

(c) $(\Lambda \xrightarrow{l} f)$: the resulting permutation is $\sigma' = [u]_a a [f]_a [v]_a$, $a = \min\{f\}$. This corresponds to the insertion of a leaf as the left sibling of $\mathcal{T}(f)$.

The array of Figure 7 gives all the permutations that can be obtained with a single insertion in $\sigma = (1524376)$.

We now prove that the operations (deletion and insertion) defined on one-stack sortable permutations are in fact internal operators for one-stack sortable permutations. Moreover, these operators define an edit distance between permutations consistent with the usual edit distance between trees.

Lemma 3.5. *The Deletion/Insertion algorithm yields a one-stack sortable permutation.*

Proof. In the case of deletion, the proof is straightforward given the one-to-one correspondence with trees and one-stack sortable permutations. Consider a tree labeled by a depth first traversal. Deleting the edge i from this tree changes all labels greater than i by subtracting 1.

For the insertion operation, let σ be a one-stack sortable permutation and f be a complete factor of $\sigma = ufv$. By Proposition 3.2, f corresponds to a subtree of $\mathcal{T}(\sigma)$.

(1) $(\Lambda \to f)$: Let $T = \mathcal{T}(\sigma)$ and $(e_1, e_2, \ldots, e_n)$ be the edges of T ordered by a prefix DFS of the tree. Note that $\sigma = \alpha_T(e_1)\alpha_T(e_2)\ldots\alpha_T(e_n)$, where $\alpha(i)$ is the label of the edge i in T.

Let T' be the tree obtained by the insertion of an internal vertex v ($a = (p(v)v)$) at the root-vertex of the subtree $\mathcal{T}(f)$, which is a subtree

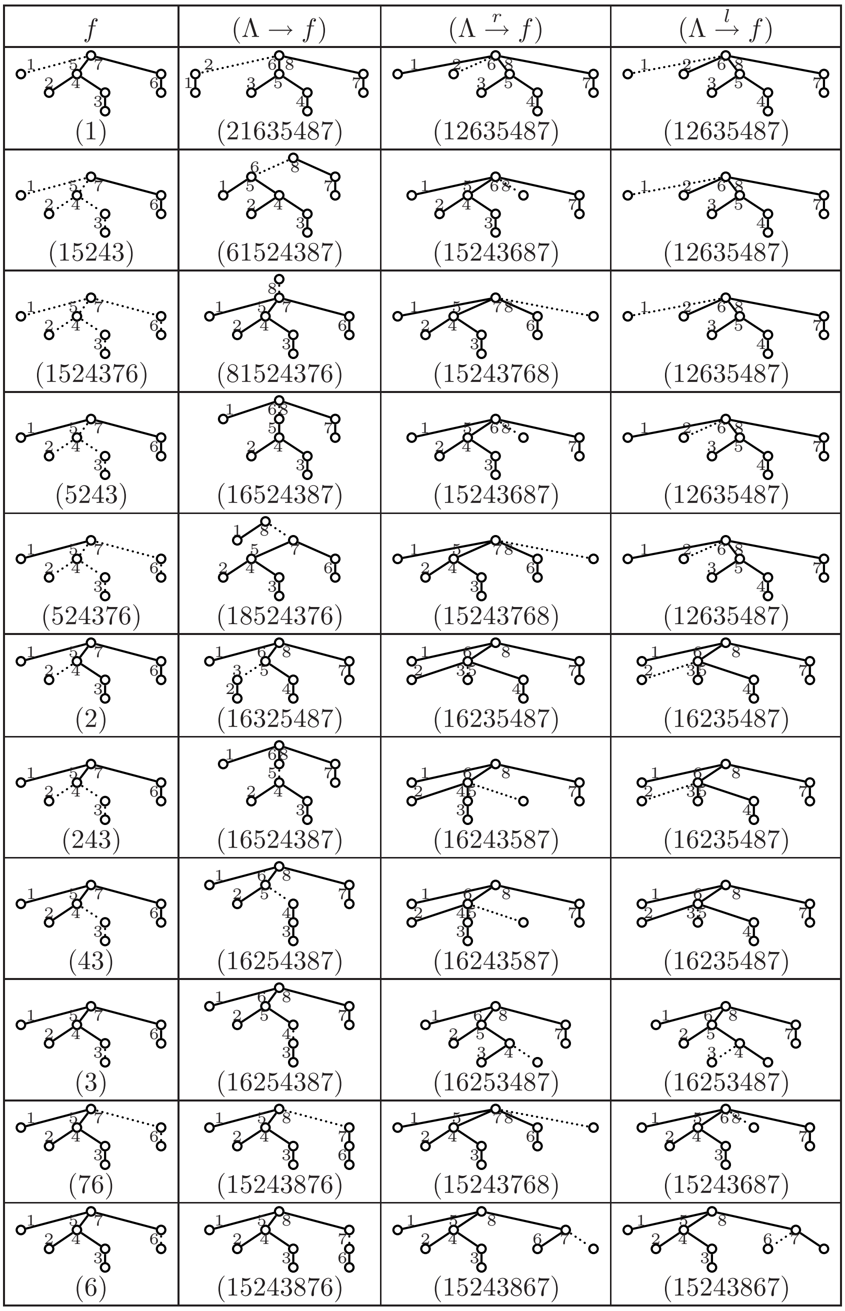


FIGURE 7. Insertion in permutation $\sigma = 1524376$.

hanging on v. Let $\sigma'' = \Theta(T')$. A prefix traversal of T' orders the edges of T' as follows: $(e_1, e_2, \ldots, e_l, a, e_{l+1}, \ldots, e_n)$.

Since σ'' is obtained by a prefix traversal, $\sigma'' = u'af'v'$. Since the edges of f appear before a in the postfix DFS, $f' = f$. The edge a in a postfix DFS appears just after f. Thus its label is $\max\{f\} + 1$. All the edges visited after f in T (and thus after a in T') by the postfix DFS have their labels increased by 1. Thus $\sigma'' = [u]_a af[v]_a = \sigma'$.

(2) $(\Lambda \xrightarrow{l} f), (\Lambda \xrightarrow{r} f)$: The same argument used as for $(\Lambda \to f)$ hold. □

Proposition 3.6. *Insertion and deletion are inverse operations.*

Proof. There are two different kinds of deletion in a tree T.

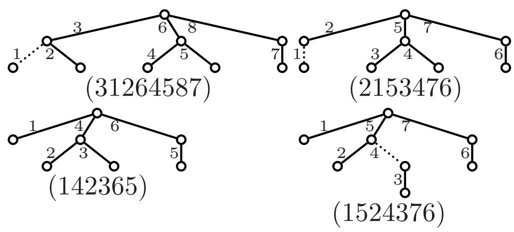
(1) Deletion of an inner vertex v. Consider the subtree T' of T hanging on v. It corresponds to a complete factor f in $\sigma = \Theta(T)$. This contraction corresponds to the inverse operation of $(\Lambda \to f)$.

(2) Deletion of a leaf. There are three different cases:
- Deletion of a vertex with no sibling. This is the same as deleting the parent of this vertex which is an inner vertex except if the tree is reduced to a single edge.
- Otherwise, this vertex has either:
  - A left sibling v'. Consider the subtree hanging at v' (including $(p(v')v')$). It corresponds to the factor f. The inverse operation is $(\Lambda \xrightarrow{r} f)$.
  - A right sibling v'. Consider the subtree hanging at v' (including $(p(v')v')$). It corresponds to the factor f. The inverse operation is $(\Lambda \xrightarrow{l} f)$. □

Definition 3.7. The distance between two one-stack sortable permutations σ_1 and σ_2 is the minimal number of operations – deletion or insertion – to transform σ_1 into σ_2.

For example, let $\sigma_1 = 31264587$ and $\sigma_2 = 1524376$. We want to transform σ_1 into σ_2.

- 31264587 $\xrightarrow{(1\to\Lambda)}$ 2153476
- 2153476 $\xrightarrow{(1\to\Lambda)}$ 142365
- 142365 $\xrightarrow{(\Lambda\to 3)}$ 1524376



Theorem 3.8. *The edit distance between ordered trees is the distance between the associated one-stack sortable permutations.*

Proof. This is a consequence of Proposition 3.6. □

Theorem 3.9. *The edit distance between one-stack sortable permutations σ_1 and σ_2 is equal to*

$$|\sigma_1| + |\sigma_2| - 2|u|$$

where u is a largest normalized subsequence (pattern) of σ_1 and σ_2.

Proof. The edit distance $d(\sigma_1, \sigma_2)$ between σ_1 and σ_2 is given by the minimal number of insertions and deletions. If t_1 is an insertion and t_2 is a deletion, then there exist a deletion t_1' and an insertion t_2' such that $t_1 t_2(\sigma) = t_1' t_2'(\sigma)$. Note that t_1' and t_2' depend on the one-stack sortable permutation σ.

Considering the sequence of edit operations, there exists a sequence composed of deletions and consequent insertions that transforms σ_1 into σ_2. We denote this sequence by $D_1 \dots D_l O_1 \dots O_k$, $l + k = d(\sigma_1, \sigma_2)$.

Consider the one-stack sortable permutation $\sigma' = D_1 \dots D_l(\sigma_1)$. Let $u = \sigma'$. Then u is a normalized subsequence of σ_1 because deleting an edge from a one-stack sortable permutation yields a normalized subsequence of the original one-stack sortable permutation. In addition u is also a normalized subsequence of σ_2 because inserting an edge in a one-stack sortable permutation s yields a one-stack sortable permutation s' and s is a normalized subsequence of s'.

Conversely, let u be a maximal normalized subsequence of σ_1 and σ_2. It is straightforward to find $|\sigma_1| - |u|$ operations of deletions such that those deletions transform σ_1 into u. The same goes for σ_2 and u. □

Corollary 3.10. *Finding the greatest common pattern between two one-stack sortable permutations is polynomial.*

In [1], the authors proved that finding the greatest common pattern between two permutations is NP-complete. We prove here that the problem becomes polynomial when restricted to one-stack sortable permutations, *i.e.* (132) or (231) pattern-avoiding permutations. In fact, the algorithm of Zhang and Shasha [14] on trees solves the problem on one-stack sortable permutations because the algorithm outputs not only the distance but also the greatest common subtree.

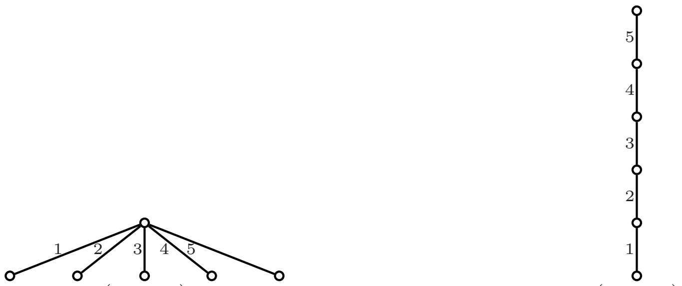
4. Lower bounds on average edit distance

In this section we study the average edit distance between a given rooted plane tree T with n vertices and all other rooted plane trees with n vertices. We show that this average distance is bounded below by $\frac{n}{ln(n)}$.

Lemma 4.1. *Let T be a rooted plane tree with n vertices. There are at most $n-1$ different deletions and $3n^3$ insertions allowed in T.*

Proof. The number of deletions is bounded above by the number of edges *i.e.* $(n-1)$.

The number of insertions is bounded by 3 times the number of subtrees (or complete factors of the corresponding permutation). The number of subtrees of T rooted at vertex v is bounded by $d(v)^2$ where $d(v)$ denotes the degree of vertex v. Thus the total number of subtrees is bounded by $\sum_v d(v)^2$. □



(12345) (54321)

FIGURE 8. Some canonical trees.

Theorem 4.2. *Let T_0 be a tree with n vertices. The proportion of rooted plane trees with n vertices at distance at most $\mathcal{O}(n/ln(n))$ tends to* 0.

The average distance between T_0 and the set of rooted plane trees is bounded below by $n/ln(n)$.

Proof. Let T_0 be a rooted plane tree. Let $A_k = \{T \in \mathcal{T}_n, \mathrm{dist}(T_0, T) \leq 2k\}$. Note that $A_0 = \{T_0\}$. A tree $T_k \in A_k$ is obtained from T_0 by $l \leq k$ deletions and l insertions. Thus $|A_k| < (n-1)^k (n^3)^k < n^{4k}$. But the number of rooted plane trees is $C_n \sim \frac{4^n}{n\sqrt{\pi n}}$, then the proportion of rooted plane trees at distance at most $\mathcal{O}(n/ln(n))$ tends to 0.

Hence the average distance is bounded below by $n/ln(n)$. □

5. Generating functions

By using the combinatorial interpretation of the distance, we compute the generating functions of the edit distance between rooted plane trees with n edges and some special ones as shown in Figure 8. Moreover, we deduce the average distances from the generating functions.

5.1. Generating function of the edit distance between one-stack sortable permutations and $Id = 1\ 2\ \ldots\ n$

We denote by $S_1(t, q)$ the generating function of one-stack sortable permutations, where t counts the size of the permutation and q the edit distance between one-stack sortable permutations and Id. This is the distance between a tree and the trivial one which is made of n edges and of height 1.

5.1.1. *Tree interpretation of the largest increasing subsequence*

Proposition 5.1. *The length of a largest increasing subsequence of a one-stack sortable permutation is the number of leaves of the associated tree.*

Proof. Let T be a rooted plane rooted tree and σ the associated one-stack sortable permutation. We call *leaf-edge* an edge incident to a leaf; any other edge will be called an internal edge.

(1) The subsequence of σ made of the leaf-edges is increasing because the order in which the leaf-edges are visited by a prefix traversal is the same as by a postfix traversal.
(2) Suppose that we take an increasing subsequence σ' of σ. This subsequence is in one-to-one correspondence with some of the edges in the tree. Suppose that there is an internal edge $\gamma = (p(\nu)\nu)$. Then, by the postordering of the edges, each edge $(p(v)v)$ such that $\nu = p(v)$ has a smaller label and appears in σ after the edge γ. Thus, none of these edges are in σ'. Moreover, there is at least one leaf-edge belonging to the subtree T_γ hanging on ν. Replace the edge γ by a leaf of T_γ. The prefix traversal ensures that the subsequence so obtained is an increasing one. □

Proposition 5.2. *The number of rooted plane trees with n edges and k leaves is equal to the number of rooted plane trees with n edges and $n+1-k$ leaves.*

Proof. This is a direct consequence of the symmetry of the Narayana numbers $\frac{1}{n}\binom{n}{k}\binom{n}{k-1}$ (see [8], Article 495, [9]) which count the number of rooted plane trees with n edges and k leaves [10]. □

5.1.2. *Generating function*

We now compute the generating function $I(t,p)$ of one-stack sortable permutations of size t and largest increasing subsequence of size p. The first terms of $I(t,p)$ are $[I(t,p)]_0 = 1$, $[I(t,p)]_1 = p$ and $[I(t,p)]_2 = (p+p^2)$. The general term follows the recursive formula:

$$[I(t,p)]_n = p[I(t,p)]_{n-1} + \sum_{i=0}^{n-2}[I(t,p)]_i[I(t,p)]_{n-1-i}. \tag{1}$$

This formula comes from the decomposition of a one-stack sortable permutation σ into InJ with $n \geq 1$. The largest increasing subsequence of σ is the union of the largest one of I and the largest one of J unless J is empty; in this case, the largest subsequence is the largest one for In.

From this formula we deduce:

$$I(t,p) = 1 + (p-1)tI(t,p) + tI^2(t,p) \tag{2}$$

- 1 comes from the case where $n = 0$ in the equation (1);
- $ptI(t,p)$ comes from $p[I(t,p)]_{n-1}$.

It follows from equation (2) that:

$$I(t,p) = \frac{1 + (1-p)t - \sqrt{(p-1)^2t^2 - 2(p+1)t + 1}}{2t}. \tag{3}$$

Let $\tilde{S}_1(t,q)$ be the generating function of the difference between the lengths of the one-stack sortable permutation and the largest increasing subsequence in it. The first terms are $[\tilde{S}_1(t,q)]_0 = -1$, $[\tilde{S}_1(t,q)]_1 = 1$ and $[\tilde{S}_1(t,q)]_2 = q+1$. Then

Lemma 5.3. *$I(t,p)$ and $\tilde{S}_1(t,p)$ are related by:*

$$I(t,p) = 1 + p + p\tilde{S}_1(t,p). \tag{4}$$

Proof.

$$\begin{aligned} I(t,p) &= \sum_{\tau\geq 1}\sum_{\alpha=1}^{\tau}[I(t,p)]_{\tau,\alpha}t^\tau p^\alpha + 1 = \sum_{\tau\geq 1}\sum_{\beta=1}^{\tau}[I(t,p)]_{\tau,\tau+1-\beta}t^\tau p^{\tau+1-\beta} + 1 \\ &= \sum_{\tau\geq 1}\sum_{\beta=0}^{\tau}[I(t,p)]_{\tau,\tau+1-\beta}t^\tau p^{\tau+1-\beta} + 1 = 1 + p(\tilde{S}_1(t,p)+1). \end{aligned}$$

The end of the proof is straightforward using Proposition 5.2. □

Lemma 5.3 and Formula 3 yield:

Theorem 5.4.

$$\begin{aligned} S_1(t,q) &= \tilde{S}_1(t,q^2) \\ &= \frac{1-(1+3q^2)t-\sqrt{(q^2-1)^2t^2-2(q^2+1)t+1}}{2tq^2}. \end{aligned}$$

5.1.3. *Average distance*

Theorem 5.5. *The average edit distance between rooted plane trees with n edges and Id is $n-1$.*

Proof. We provide two different proofs of this theorem. The first one is analytic and the second one combinatorial.
No. 1: the average distance δ can be obtained from the generating function $S_1(t,q)$ in the following way:

- $F(t) = \left.\frac{\partial S_1(t,q)}{\partial q}\right|_{q=1}$;
- $\delta = \frac{[F(t)]_n}{C(n)}$, where $C(n)$ is the n-th Catalan number.

This easy computation yields $\delta = n-1$ but a direct combinatorial interpretation proves this result in a more comprehensive way.
No. 2: this is a direct consequence of Propositions 5.1 and 5.2. Another proof can be found in [4, 11]. In [11] a more general result is presented. We provide here a simpler proof for this special case. □

5.2. Generating function of the edit distance between one-stack sortable permutations and $n(n-1)\dots 1$

This is the distance between a tree and the trivial one which is a chain with n edges. It is equivalent to finding the largest decreasing subsequence in the one-stack sortable permutation.

5.2.1. *Generating function*

We compute the generating function $D(x,y,z)$ of trees, where x counts the number of edges, y the height of the tree and z the number of leaves at maximal depth.

Proposition 5.6.

$$D(x,y,z) = yD\left(x,y,\frac{1}{1-xz}\right) - yD(x,y,1) + \frac{xyz}{1-xz}\cdot \tag{5}$$

Proof.

$$\begin{aligned}
[D(x,y,z)]_{i,j,k} &= \sum_{l=1}^{i-j+1} \binom{l+k-1}{k} [D(x,y,z)]_{i-k,j-1,l} \text{ if } j>1\\
[D(x,y,z)]_{i,1,k} &= \delta_{i,k}.
\end{aligned}$$

The coefficient $[D(x,y,z)]_{i,j,k}$ is equal to the number of ways to add k leaves at depth j to any tree with $i-k$ edges, depth $j-1$ and l leaves at depth $j-1$. $\binom{l+k-1}{k}$ is the number of ways to add k leaves to l leaves at depth j.

$$\begin{aligned}
D(x,y,z) &= \sum_{i\geq 1}\sum_{j\geq 1}\sum_{k\geq 1} d_{i,j,k} x^i y^j z^k\\
&= \sum_{i\geq 1}\sum_{j\geq 2}\sum_{k\geq 1}\sum_{l\geq 1} \binom{k+l-1}{k} d_{i-k,j-1,l} x^i y^j z^k + y\sum_{i\geq 1}(xz)^i\\
&= \sum_{i\geq 1}\sum_{j\geq 2}\sum_{k\geq 1}\sum_{l\geq 1} (-1)^k \binom{-l}{k} d_{i-k,j-1,l} x^i y^j z^k + y\sum_{i\geq 1}(xz)^i\\
&= \sum_{i\geq 1}\sum_{j\geq 2}\sum_{k\geq 1}\sum_{l\geq 1} (-1)^k \binom{-l}{k} d_{i,j-1,l} x^i y^j (xz)^k + y\sum_{i\geq 1}(xz)^i.
\end{aligned}$$

Using

$$(x+a)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k a^{-n-k},$$

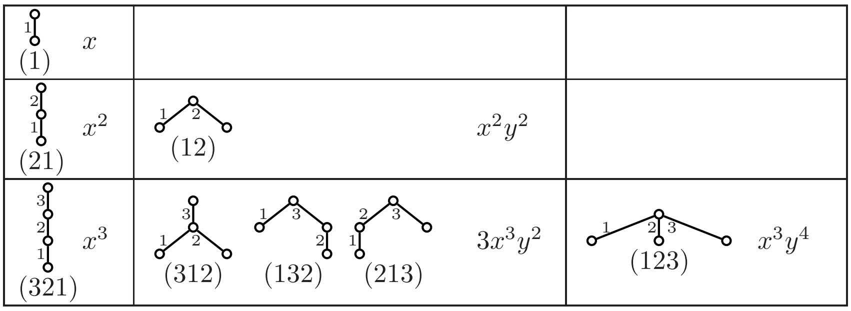


FIGURE 9. The first terms of S_2.

we have:

$$\begin{aligned} D(x,y,z) &= \sum_{i\geq 1}\sum_{j\geq 2}\sum_{l\geq 1}((1-zx)^{-l}-1)d_{i,j-1,l}x^i y^j + y\sum_{i\geq 1}(xz)^i \\ &= yD\left(x,y,\frac{1}{1-xz}\right) - yD(x,y,1) + \frac{xyz}{1-xz}\cdot \end{aligned}$$

□

Let $S_2(x,y)$ be the generating function with respect to the length n of the one-stack sortable permutation and the edit distance between this one-stack sortable permutation and $n(n-1)(n-2)\ldots 1$. Then, $S_2(x,y) = D(xy^2, \frac{1}{y^2}, 1)$.

In [3, 12], the authors give a solution for $D(x,y,1)$ in terms of a continued fraction:

$$D(x,y,1) = \sum_{k\geq 1} D_k(y)x^k, D_k(y) = \cfrac{1}{k\left\{1 - \cfrac{y}{1-\cfrac{y}{1-\ldots}}\right.}\cdot$$

This yields the following solution for S_2:

$$S_2(x,y) = \sum_{k\geq 1} y^{2k} D_k\left(\frac{1}{y^2}\right)x^k.$$

The first terms of S_2 (see Fig. 9) are given by:

$$S_2(x,y) = x + x^2y^2 + x^2 + x^3y^4 + 3\,x^3y^2 + x^3 + x^4y^6 + 7\,x^4y^4 + 5\,x^4y^2 + x^4.$$

5.2.2. *Average edit distance*

In [3], the average height of a rooted plane tree with n edges, which is $\sqrt{\pi n} - \frac{1}{2}$ is analytically determined. Thus, the average edit distance is $2(n - \sqrt{\pi n} + \frac{1}{2}) \sim 2n$.

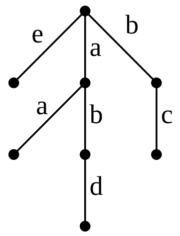


FIGURE 10. Example of labeled tree.

6. Conclusion

The general case where the trees are labeled (see Fig. 10) and the different edit operations have different costs can be obtained in a similar way. Define a decorated one-stack sortable permutation as a one-stack sortable permutation where each number is indexed by a letter; $1_e5_a2_a4_b3_d7_b6_c$ represents the tree in Figure 10. Note that the root is unlabeled.

The operations on decorated one-stack sortable permutations are almost the same as before and the relabeling operation consists of changing one letter. Let c_i, c_d, c_r be, respectively, the insert, delete and relabeling unitary costs. The only difference is for the insertion of a new free edge. In the unlabeled case, we did not take into account the insertion of a leaf with no sibling. Thus we define a fourth insertion operation as:

- $(\Lambda \xrightarrow{1} i)$, where i is a complete factor of size 1 of the permutation $\sigma = uiv$ and $\sigma' = [u]_a[i]_a a[v]_a$, where $a = i$.

Let σ_1 and σ_2 be two decorated one-stack sortable permutations with the same underlying permutation. The label distance $d(\sigma_1, \sigma_2)$ is equal to the string distance between the two labeled words.

Let T_1 and T_2 be two decorated one-stack sortable permutations. We denote by a subpermutation σ of T_1 and T_2 a normalized subpermutation without labels. Then Σ_{T_1} is the set of all sub-decorated one-stack sortable permutations of T_1 whose underlying permutation is σ.

The relabeling distance between T_1 and T_2 with respect to σ is:

$$d_\sigma(T_1, T_2) = \min\{c_r d(\alpha, \beta), \forall \alpha \in \Sigma_{T_1}, \beta \in \Sigma_{T_2}\}.$$

The distance between these two decorated one-stack sortable permutations T_1 and T_2 is given by $\min\{c_i(|T_1| - |\sigma|) + c_d(|T_2| - |\sigma|) + d_\sigma(T_1, T_2), \sigma$ normalized subpermutation of $T_1, T_2\}$.

References

[1] P. Bose, J.F. Buss and A. Lubiw, Pattern matching for permutations. *Inf. Proc. Lett.* **65** (1998) 277–283.

[2] M. Bousquet-Mélou, Sorted and/or sortable permutations. *Disc. Math.* **225** (2000) 25–50.

[3] N.G. De Bruijn, D.E. Knuth and S.O. Rice, *Graph theory and Computing.* Academic Press (1972) 15–22.

[4] E. Deutsch, A.J. Hildebrand and H.S. Wilf, Longest increasing subsequences in pattern-restricted permutations. *Elect. J. Combin.* **9** (2003) R12.

[5] M. Garofalakis and A. Kumar, Correlating XML data streams using tree-edit distance embeddings, in *Proc. PODS'03* (2003).

[6] P.N. Klein, Computing the edit-distance between unrooted ordered trees, in *ESA '98* (1998) 91–102.

[7] D.E. Knuth, *The Art of Computer Programming: Fundamental Algorithms.* Addison-Wesley (1973) 533.

[8] P.A. MacMahon, *Combinatorial Analysis* **1–2**. Cambridge University Press (reprinted by Chelsea in 1960) 1915–1916.

[9] T.V. Narayana, Sur les treillis formés par les partitions d'un entier et leurs applications à la théorie des probabilités. *C. R. Acad. Sci. Paris* **240** (1955) 1188–1189.

[10] T.V. Narayana, A partial order and its application to probability theory. *Sankhyā* **21** (1959) 91–98.

[11] A. Reifegerste, *On the diagram of 132-avoiding permutations.* Technical Report 0208006, Math. CO (2002).

[12] E. Roblet and X.G. Viennot, Théorie combinatoire des t-fractions et approximants de Padé en deux points. *Discrete Math.* **153** (1996) 271–288.

[13] J. West, *Permutations and restricted subsequences and Stack-sortable permutations.* Ph.D. thesis, M.I.T., 1990.

[14] K. Zhang and D. Shasha, Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.* **18** (1989) 1245–1262.

Communicated by C. Choffrut.
Received June 15, 2005. Accepted February 15, 2006.