In-place update of suffix array while recoding words

Matthias Gallé, Pierre Peterlongo, François Coste

International Journal of Foundations of Computer Science

IN-PLACE UPDATE OF SUFFIX ARRAY WHILE RECODING WORDS

MATTHIAS GALLÉ, PIERRE PETERLONGO and FRANÇOIS COSTE
Centre de Recherche INRIA Rennes - Bretagne Atlantique,
Campus de Beaulieu, 35042 Rennes cedex, France
matthias.galle@irisa.fr
pierre.peterlongo@irisa.fr
francois.coste@irisa.fr

Received (Day Month Year)
Accepted (Day Month Year)
Communicated by (xxxxxxxxxx)

Motivated by grammatical inference and data compression applications, we propose an algorithm to update a suffix array while in the indexed text some occurrences of a given word are substituted by a new character. Compared to other published index update methods, the problem addressed here may require the modification of a large number of distinct positions over the original text. The proposed algorithm uses the specific internal order of suffix arrays in order to update simultaneously groups of indices, and ensures that only indices to be modified are visited. Experiments confirm a significant execution time speed-up compared to the construction of suffix array from scratch at each step of the application.

Keywords: suffix array, in-place update, dynamic indexing, word-interval

68-04, 68P05

1. Motivation

In this paper, we propose an algorithm to efficiently update a suffix array, after substituting a word by a new character in the indexed text. This work is motivated by grammatical inference or grammar-based compression, along the lines initiated by SEQUITUR [24] in the framework formalized by Kieffer and Yang [11, 12]. The goal is to infer a grammar G which generates only a given (long) sequence s in order to discover the structure that underlies the sequence, or simply, to compress the sequence thanks to a code based on the grammar. Learning and compression being often subtly intertwined (as for instance in the Occam's razor principle), in both cases the grammar is expected to be as small as possible. Kieffer and Yang introduced the definition of irreducible grammars and proposed several reduction rules allowing to transform a reducible grammar into an irreducible one, giving rise to efficient universal compression algorithms [11]. The sketch of these algorithms is to begin with a unique $S \rightarrow s$ rule generating the whole given sequence and to

reduce iteratively the size of the grammar at each step by: 1) choosing a repeated substring, 2) replacing the occurrences of the repeated substring by a new symbol and 3) adding a new rewriting rule from this new symbol into the repeated substring. For instance, the sequence $uRvRw$, where u, v, w and R are substrings, and the length of R is strictly bigger than one, would be rewritten by a rule $S \rightarrow uRvRw$. At the first step, this rule can be reduced into two rules $S \rightarrow uAvAw$ and $A \rightarrow R$, where A is a new (non-terminal) symbol. At the following step, another repeated substring, including eventually the new inserted symbol A, is selected and factorized by the introduction of a third rule, and so forth for the next steps. As a result, the algorithm returns a compact grammar which can be used to get a hierarchical point of view on the structure of the sequence or which can be encoded in order to get a better compression than by encoding directly the sequence.

Algorithms of this kind are thus mainly based on the successive detection of repeats. They differ mostly in the order in which repeats are factorized. In SEQUITUR [24] and its variant [12], each repeat is replaced as soon as it is detected by a left to right scan of the sequence. More elaborate strategies for choosing the repeat to replace have been proposed. Kieffer and Yang [11], Nakamura and Bannai [23] and Lanctot, Li, and Yang [16] proposed to replace longest matching substring. Apostolico and Lonardi [3] proposed in their algorithm Off-Line to choose the substring yielding the best compression in a steepest-descent fashion. A comparison between the different strategies can be found in [25].

Efficient implementation of an elaborate choice of repeat often requires the use of data structures from the suffix tree family. These index structures are well suited for efficient computations on repeats but they have to be built at initialization, and then updated at each step of the algorithm with respect to sequence modifications. Yet, as pointed out by Apostolico and Lonardi [3], most of the published work on dynamic indexing problem [27], by updating a suffix tree [5–8,20] or suffix array [28] focuses on localized modifications of the string. They do not seem appropriate for efficiently replacing *more than one* occurrence of a given substring, as they would require one update operation for each occurrence.

Thus, index structures have usually to be built from scratch at each step of the algorithm. To our knowledge, only GTAC [16], an algorithm applied successfully on genomic sequences by Lanctot, Li and Yang, updates a suffix tree data structure after the deletion of all occurrences of a word. More recently, [23] solved the same problem also in linear time. However, their updating scheme are specific to the longest matching substrings and seems difficult to adapt to other strategies.

In this paper, we propose a solution to the problem of updating efficiently an index structure while replacing some non-overlapping occurrences of a word of the indexed text by a new symbol. The first originality of our approach relies on the use of enhanced suffix arrays instead of suffix trees. Enhanced suffix arrays are known to be equivalent to suffix trees while being more space efficient [1]. They can be

built in linear time [10, 13, 15] but non-linear algorithms [17, 19] are usually more efficient for practical applications. A simple way of updating suffix array (instead of enhanced suffix array, thus without the same efficiency objective) by lazy bubble sort has been used in [25]. We propose here, to take advantage of the internal order offered by enhanced suffix arrays, to simultaneously handle groups of indices. This enables us to efficiently implement an update procedure for grammatical inference or grammar-based compression algorithm, choosing at each step a repeated substring, and replacing some or all of its occurrences by a new symbol.

2. Definitions and notations

A *sequence* is a concatenation of zero or more characters from an alphabet Σ. The number of characters in Σ is denoted by $|\Sigma|$. A sequence s of length n on Σ is represented by $s[0]s[1]\dots s[n-1]$, where $s[i] \in \Sigma \ \forall 0 \leq i < n$. We denote by $s[i,j](j \geq i)$ the sequence $s[i]s[i+1]\dots s[j]$ of s (if $j < i$ then $s[i,j] = \epsilon$, the empty string). In this case, we say that the sequence $s[i,j]$ occurs at position i in s. Its length, denoted by $|s[i,j]|$, is equal to $j - i + 1$. Furthermore, the sequence $s[0,j]$ $(0 \leq j < n)$, also denoted by $s[..j]$, is called a prefix of s, and symmetrically, $s[i,n-1]$ $(0 \leq i < n)$, also denoted by $s[i..]$, is called a suffix of s.

Definition 1 (Suffix Array) *Consider a sequence s of length n over an alphabet Σ with an order $\prec$ extensible to Σ^*. This lexicographically extension will be denoted also by $\prec$. Let $\tilde{s} = s\$$, with a special character $\$$ not contained in Σ, smaller than every element of Σ.*

The suffix array, denoted by sa, is a permutation of $[0..n]$ such that:

$$\forall\ i,\ 0 < i \leq n\ :\ \tilde{s}[sa[i-1]..] \prec \tilde{s}[sa[i]..]$$

Usually, the suffix array is used conjointly with an array called lcp, that gives the longest common prefix length between two suffixes whose starting positions are adjacent in sa. Formally,

$$lcp[0] = 0,$$
$$\text{and } \forall\ i \in [1,n]\ :\ lcp[i] = k \text{ such that}$$
$$\tilde{s}[sa[i-1]..][0,k-1] = \tilde{s}[sa[i]..][0,k-1] \text{ and } \tilde{s}[sa[i-1]..][k] \neq \tilde{s}[sa[i]..][k].$$

Eventually, a third array called isa (for inverse suffix array) may be used conjointly with sa and lcp. This array gives, for a position p in s, the index i in sa such that $sa[i] = p$. Thus $sa[isa[p]] = p$.

The union of *sa*, *lcp* and *isa* arrays is called an *Enhanced Suffix Array* (*ESA*). An *ESA* enables $O(n)$ computation of occurrences of different kinds of repeats (repeats, maximal repeats [9, 14] or super maximal repeats [1, 9]).

To avoid confusion, we will use the term *position* when referring to the index over a sequence and *index* when referring to any of the arrays of an ESA.

In this paper, we propose to update an *ESA*, deleting and moving some of its indices and keeping *lcp* consistent. In order to avoid shifting set of indices, we link consecutive indices using two additional arrays called *next* and *prev*. Thus, $next[i]$ (resp. $prev[i]$) gives the index of the next (resp. previous) valid entry in the *ESA*. Initially, $next[i] = i+1$ and $prev[i+1] = i$. We call the set *ESA* plus *next* and *prev* arrays the ESA_{DL} for *Double Linked Enhanced Suffix Array*.

It is worth noticing that an ESA_{DL} does not have the exact same properties as an *ESA*. Indeed, going from an index i to index $i + j$ may be done in constant time on an *ESA*, while this operation in an ESA_{DL} requires $O(j)$ time, as the *next* array has to be used j times.

Anyway, an ESA_{DL} still allows the detection of repeats (general repeats, maximal repeats or super maximal repeats) in linear time, because the algorithms used advance one by one over the arrays like most of the algorithm over *ESA* (a notable exception is the algorithm searching for a substrings proposed in [30]).

We propose an *in-place* solution, where we always work with the same arrays and only update the values of their fields. Moreover, during the whole process, we modify only the *prev*, *next* and *lcp* arrays. Arrays *sa* and *isa* remain unchanged. This approach forces to extend the in-place behavior to the sequence: we also add two arrays to imitate a double linked list over the sequence.

The j^{th} position after position i, is denoted by $i \oplus j$. We compute $i \oplus j$ using links between sequence positions, indicating for each position its successor. Similarly $i \ominus j$ points to the j^{th} position before i. We define that, if $i \oplus j$ (respectively $i \ominus j$) is out of range, then $i \oplus j = n + 1$ (respectively $i \ominus j = -1$).

We consider that the grammatical inference or grammar based compression algorithm proceeds by steps. At each step, the alphabet grows because of the introduction of a new character: Σ_k will denote the alphabet in step k. At each of these steps the algorithm **i)** finds a repeat $\mathcal{R}_k$ in a sequence $\tilde{s}^{(k)}$ defined on the alphabet Σ_k and returns a list $\mathcal{O}_k$ of non-overlapping occurrences of $\mathcal{R}_k$ **ii)** updates the sequence $\tilde{s}^{(k)}$ and its associated ESA_{DL} replacing the given occurrences of $\mathcal{R}_k$ by a single new character $\mathcal{C}_k$, thus defining a new alphabet $\Sigma_{k+1} = \Sigma_k \cup \{\mathcal{C}_k\}$. The modified sequence is then called $\tilde{s}^{(k+1)}$. The whole iterative process stops either if no more repeat is found in the sequence or after a fixed number of iterations.

Our contribution focuses on updating the ESA_{DL}, at each step k of this algorithm (part **ii**).

In the next sections, we describe how to perform the three tasks needed for updating an ESA_{DL} at each step k: **1)** delete indices of suffixes starting inside an $\mathcal{R}_k$ occurrence; **2)** move indices with respect to the new alphabetic order; and **3)** update *lcp* array with respect to recoding occurrences of $\mathcal{R}_k$ by one single character. Note that a few values of the *lcp* array are also modified during step 1 and 2, but only as a consequence of deletions and moves.

3. Algorithm

Before describing the different steps of the algorithms, we will define the *left context tree* which result useful to better understand the algorithm and the modifications it made over the suffix array.

3.1. *The left context tree.*

One of the most useful characteristics of a suffix array is that all indices corresponding to suffixes starting with the same word correspond to an adjacent block. We define here the corresponding concept of word interval. Based on this, we will define the *left context tree* of a word ω where the nodes correspond to a left context of ω.

An ω-interval is the set $\{k : \exists \ell,\ k = isa[\ell] \wedge \tilde{s}[\ell..\ell + |\omega| - 1] = \omega\}$. This can also be denoted as an $[i..j]$-interval, where i and j are respectively the lowest and highest indices of an ω-interval. Let us note that different words can share the same interval. More precisely, any pair of words ω and $\omega\alpha$ share the same interval if each occurrence of ω is followed by α.

This definition is thus slightly more general than the definition of ω-interval given by Abouelhoda, Kurtz and Ohlebusch [1], since we also define ω-interval for words leading to implicit nodes of a compact suffix tree, and not only to internal nodes.

The *left context tree of* ω ($\omega \in \Sigma^*$) for a sequence $\tilde{s}$ is an implicit tree whose nodes are v-intervals ($v \in \Sigma^*$) such that:

- the root is the ω-interval
- for each v-interval node corresponding to a non-empty interval, its children are all the av-intervals, for all $a \in \Sigma$
- the leaves are empty intervals

Given the isa array, it is easy to obtain the parent of a node. Let $[i..j]$ be an av-interval node. Given $k \in [i..j]$, $isa[sa[k] + 1]$ is an index belonging to the v-interval. Inversely, $isa[sa[k] - 1]$ belongs to one of the child interval. The exact child depends on the character at $\tilde{s}[sa[k] - 1]$. We introduce the *successor* and *predecessor* notations:

$$successor(i) = \begin{cases} isa[sa[i] \oplus 1] & \text{if } sa[i] \oplus 1 \neq n + 1 \\ n + 1 & \text{otherwise,} \end{cases}$$

and

$$predecessor(i) = \begin{cases} isa[sa[i] \ominus 1] & \text{if } sa[i] \neq 0 \\ -1 & \text{otherwise.} \end{cases}$$

One may remark that $predecessor(i)$ is the equivalent of the "*suffix link*" in a suffix tree [31].

The problem that an *ESA* update algorithm must face is that the changes over the occurrences of a word ω not only affect the ω-interval, but also some of the $v\omega$-intervals ($v \in \Sigma^*$). The core of our algorithm is based on moving $v\omega$-interval in

index	*lcp*	suffix
$prev[j]$	4	$ATAC\ldots$
~~j~~	~~2~~	~~$ATGA\ldots$~~
$next[j]$	~~3~~ **2**	$ATGT\ldots$

Fig. 1. Deletion of index j.

constant time, using the two following properties implied by the internal order of suffix arrays:

Proposition 2. *Let $[i..j]$ be an v-interval ($v \in \Sigma^*$), and $k_1, k_2 \in [i..j]$ with $k_1 > k_2$ and such that $predecessor(k_1)$ and $predecessor(k_2)$ belong to the same αv-interval ($\alpha \in \Sigma$). Then $predecessor(k_1) > predecessor(k_2)$.*

Proposition 3. *With $i < j$, the longest common prefix between $\tilde{s}[sa[i]..]$ and $\tilde{s}[sa[j]..]$ is $min_{k \in [next[i],j]}(lcp[k])$.*

3.2. *Delete indices of suffixes occurring inside $\mathcal{R}_k$ substituted occurrences*

By replacing the word $\mathcal{R}_k$ by a single letter, the sequence is compressed and so is its ESA_{DL}: consequently, any suffix of sequence $\tilde{s}^{(k)}$ appearing inside an $\mathcal{R}_k$ substituted occurrence must be deleted. Thus for i in $\mathcal{O}_k$ and for ℓ in $[1, |\mathcal{R}_k| - 1]$, suffix $\tilde{s}^{(k)}[i \oplus \ell..]$ and the associated index in the suffix array $j = isa[i \oplus \ell]$ have to be removed. We simulated this deletion by *jumping over it* by setting $next$ and $prev$ arrays to their previous and next index: $next[prev[j]] \leftarrow next[j]$ and $prev[next[j]] \leftarrow prev[j]$. Furthermore, the lcp value of the index following j ($lcp[next[j]]$) has to be modified according to the deletion of index j. As a consequence of proposition 3, after the deletion of index j, the longest common prefix of index $next[j]$ is equal to the minimal longest common prefix value of indices j and $next[j]$.

An example is shown in Figure 1 where the deletion of index j affects the $lcp[next[j]]$ that now should contain the length of longest common prefix between $ATGT$ and $ATAC$ which is 2, equal to the longest common prefix of $ATGT$, $ATGA$ and $ATAC$.

Algorithm 1 presents the procedure for deleting indices. The notation END refers to the last index of the suffix array ($prev[n+1]$).

3.3. *Move indices, with respect to new alphabetic order*

After replacing the word $\mathcal{R}_k$ by the new character $\mathcal{C}_k$, some ESA_{DL} lines may be misplaced with respect to the chosen order of $\mathcal{C}_k$ in Σ_{k+1}.

Indices in the $\mathcal{R}_k$-interval are potentially misplaced. Moreover, for $v \in \Sigma_k^*$, index inside an $v\mathcal{R}_k$-interval are misplaced if the substitution of $\mathcal{R}_k$ into $\mathcal{C}_k$ affects their lexicographical order with respect to the previous and next index over the suffix

Algorithm 1 Delete indices at step k, replacing $\mathcal{R}_k$ by $\mathcal{C}_k$

$delete\_indices\{ESA_{DL}^{(k)}, \mathcal{R}_k, \mathcal{O}_k\}$

```
1: for i ∈ O_k do
2:   for ℓ ∈ [1, |R_k| − 1] do
3:     j ← isa[i ⊕ ℓ]
4:     if next[j] ≠ END then
5:       lcp[next[j]] ← min(lcp[j], lcp[next[j]])
6:     end if
7:     next[prev[j]] = next[j]
8:     prev[next[j]] = prev[j]
9:   end for
10: end for
```

array. Thus, lines belonging to node-intervals of the left-context tree of $\mathcal{R}_k$ may have to be moved.

In our approach, we decided to give to $\mathcal{C}_k$ the largest rank in the lexicographic order of the alphabet Σ_k, *i.e.* $\forall\, \alpha \in \Sigma_k \,:\, \alpha \prec \mathcal{C}_k$.

With respect to this arbitrary choice, the $\mathcal{R}_k$-interval is moved to the end of the suffix array. Furthermore, for any $v \in \Sigma_k^*$, the $v\mathcal{R}_k$-interval is moved after the last index of the v-interval.

If an $v\mathcal{R}_k$-interval is already at the end of the v-interval (it is naturally well ordered), for any $v' \in \Sigma_k^*$, the $v'v\mathcal{R}_k$-interval is also at the end of the $v'v\mathcal{R}_k$-interval and does not have to be moved.

Based on this property, our algorithm uses a recursive approach in order to move groups. The recursion starts on the initial $\mathcal{R}_k$-interval. During recursion, if the group of an $v\mathcal{R}_k$-interval is moved, the recursion continues on groups of $\alpha v\mathcal{R}_k - intervals$, with $\alpha \in \Sigma_k$.

From a theoretical point of view, the algorithm starts on the root of the left-context tree of $\mathcal{R}_k$ and if the group corresponding to the interval of the node is moved, it recursively treats its children in a breadth first traversal (a FIFO is used).

In practice, the recursion on a $v\mathcal{R}_k$-interval works as follows:

(1) detects the end position of the $v\mathcal{R}_k$-interval,
(2) detects the end position of the v-interval,
(3) if necessary:

3.a. moves the group to the end position of the v-interval,
3.b. call the recursion on predecessors of indices of the group.

During a call on predecessor of an index of the group, either this is the first time the matched group is called and by construction the call is done on its first element, or the group was already treated, and the recursion stops.

The algorithm for this step is shown in algorithm 2. This recursive function

receives three parameters besides the data structures: the starting position of the group, the current depth over the left-context and a boolean flag (see below).

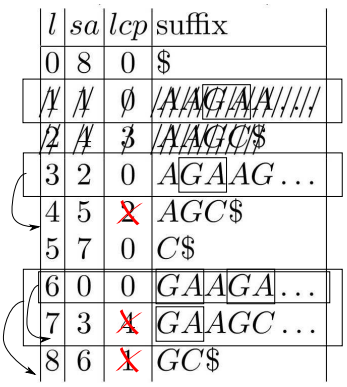
Algorithm 2 Restore consistency of suffix array order

$update\_order\{ESA_{DL}^{(k)}, \mathcal{R}_k, \mathcal{O}_k, i_{start}, depth, move\}$

1: **if** Couple $(i_{start}, depth)$ already treated during another recursion call **then**
2: End procedure
3: **end if**
4: $i \leftarrow i_{start}$
5: **while** $i \neq END \wedge lcp[next[i]] \geq depth + |\mathcal{R}_k|$ **do**
6: $i \leftarrow next[i]$
7: **end while**
8: $i_{end} \leftarrow i$
9: $minLCP \leftarrow min_{j \in [i_{start}, i_{end}]} lcp[j]$
10: **if** *move* **then**
11: **while** $i \neq END \wedge lcp[next[i]] \geq depth$ **do**
12: $i \leftarrow next[i]$
13: **end while**
14: **end if**
15: $i_{dest} \leftarrow i$
16: **if** $i_{end} \neq i_{dest}$ **then**
17: $lcp[next[i_{end}]] \leftarrow min(lcp[next[i_{end}]], minLCP)$
18: $lcp[i_{start}] \leftarrow depth$
19: **if** $i_{start} = i_{first} \wedge depth \neq 0$ **then**
20: $i_{first} \leftarrow next[i_{end}]$
21: **end if**
22: move_group$(i_{start}, i_{end}, i_{dest})$
23: **else**
24: $lcp[i_{start}] \leftarrow min(lcp[i_{start}, depth)$
25: $move \leftarrow false$
26: **end if**
27: $i \leftarrow i_{start}$
28: **while** $i \neq next[i_{end}]$ **do**
29: $newdepth \leftarrow depth+$ (**if** $predecessor(i) \in \mathcal{O}_k$ **then** len **else** 1)
30: **if** $move \vee (sa[prev[predecessor(i)]] > newdepth \wedge sa[prev[predecessor(i)]] \oplus newdepth \in \mathcal{O}_k)$ **then**
31: $update\_order(ESA_{DL}^{(k)}, \mathcal{R}_k, \mathcal{O}_k, predecessor(i), newdepth, i_{dest} \neq i_{end})$
32: **end if**
33: $i \leftarrow next[i]$
34: **end while**

In first place, the end of the $v\mathcal{R}_k$-interval is found (lines 5, 6 and 8).

Fig. 2. Moves induced by substituting GA by $\mathcal{C}_1$.



This is done from the first element of the interval, following the *next* array while the visited index corresponds to a suffix starting with $v\mathcal{R}_k$ ($lcp \geq |v| + |\mathcal{R}_k|$). After finding the extremes of the group, the destination index of this group according to the chosen order for the new character is found (lines 11, 12 and 15). This is done by finding the end of the v-interval in the same way ($lcp \geq |v|$).

Moving the group to its new position is now simple and is done in constant time. Thanks to the well-ordered property of the suffix array, the whole interval is moved by changing only the delimiting positions. Let $i_{start}, i_{end}, i_{dest}$ be respectively the starting and ending positions of the $v\mathcal{R}_k$-interval, and the last position of the v-interval. Moving the group $[i_{start}, i_{end}]$ to the position after i_{dest} is simply done by jumping over the group and *inserting* it into i_{dest} and $next[i_{dest}]$. See the algorithm 3 for implementation details.

Two longest common prefix values are modified as a consequence of the deletion of the group and its insertion:

(1) $lcp[next[i_{end}]]$: contains the value of the length of the longest common prefix between $prev[i_{start}]$ and $next[i_{end}]$, which according to proposition 3, is the minimum of the lcp values of the group and itself
(2) $lcp[i_{start}]$: we assign to it the value of $depth$, that is the correct value over $\tilde{s}_{k+1}$. This serves also to set a stop-point for future recursions calls (see below).

As i_{first} points to the first line over the suffix array that contains a selected repetition, we also update i_{first} (line 19) if this line is moved.

Figure 2 shows the ESA_{DL} of sequence $GAAGAAGC$, where $\mathcal{R}_1 = GA$ is substituted by $\mathcal{C}_1$. One remarks that the initial interval of suffixes starting with GA (indices 6 and 7) is moved as well as suffix starting with AGA (index 3). Note also that suffix starting with $GAAGA$ has to be moved with respect to suffix $GAAGC$.

Algorithm 3 Move the group $[i_{start}, i_{end}]$ after the position i_{dest}

$move\_group\{ESA_{DL}^{(k)}, \mathcal{R}_k, \mathcal{O}_k, i_{start}, i_{end}, i_{dest}\}$

1: $next[prev[i_{start}]] = next[i_{end}]$
2: $prev[next[i_{end}]] = prev[i_{start}]$
3: $next[i_{end}] = next[i_{dest}]$
4: $prev[next[i_{dest}]] = i_{end}$
5: $next[i_{dest}] = start$
6: $prev[i_{start}] = i_{dest}$

3.3.1. *A special case*

Once an interval is treated, the recursion continues either if the current group was moved, or in the special case described in what follows.

Consider for instance the following case, where the substituted repeat is TA.

i C**TA**TT**TA**C...
i+1 C**TA**TT**TA**G...
i+2 C**TA**<u>**TTA**...</u>,

and suppose that the TTA-interval containing the index $isa[sa[i+2] \oplus 3]$ (the underlined suffix in the figure) was already at its right position and therefore does not have to be moved. So, its children in the left-context tree are not considered for future moves, and as a consequence, neither is index $i+2$. Supposing that we cut the recursion here, that means that when treating the $CTATT$-interval, $lcp[i+2] = 5$. This interval ends at the index $i+1$, but because we use the *lcp* array to detect it, we also consider index $i+2$ as part of the $CTATT$-interval.

To resolve this special case, the recursion continues even when the current interval was not moved. In this case, it will never be necessary to move an interval, but maybe update some *lcp* values to set *stop-points* for future recursion calls.

This is the reason for introducing the last parameter in algorithm 2 (the boolean flag *move*). It differentiates the normal case (when it is necessary to detect the destination index and move the interval) from the case in which the current interval is considered only to set a *stop-point* at the first index of the interval. The recursion continues in both cases.

3.3.2. *Filtering non substituted $\mathcal{R}_k$ occurrences*

Among each $v\mathcal{R}_k$-interval, suffixes starting with $v\mathcal{R}_k$ where $\mathcal{R}_k$ is not substituted (whose position does not belong to $\mathcal{O}_k$) may occur. The associated indices in the ESA_{DL} should not be moved with the $v\mathcal{R}_k$-interval. Thus, before applying the recursive procedure previously exposed, a straightforward *filtering step* is applied. During the recursion, each line i of each group is first checked in order to detect if it corresponds to an index of a selected occurrence ($sa[i] \oplus depth \in \mathcal{O}_k$). Once a non-selected occurrence is detected, we move it to the beginning of the group (before i_{start}). As previously mentioned, this also involves modifications of the *lcp*

array for maintaining its consistency.

3.4. *Update lcp values after the substitution of $\mathcal{R}_k$ occurrences to a single character*

The substitution of any occurrence of $\mathcal{R}_k$ of length $|\mathcal{R}_k| \geq 2$ by $\mathcal{C}_k$ of length 1 involves the modification of the length of all common prefixes involving such an occurrence.

In the previous step, it was easy to update the *lcp* values of the limits of the intervals while they were moved. In this step, we update the *lcp* values of the internal position of the intervals.

For this, we traverse the left-context tree of $\mathcal{R}_k$. Contrary to the moving step, where it was possible to move one line several times, in this step we update each *lcp* index only once. To do this, we recalculate all the *lcp* values for the root ($\mathcal{R}_k$-interval) and use this information to update the *lcp* of the other intervals.

As a consequence of propositions 2 and 3, the *lcp* between two indices of the same interval-node is simply one plus the *lcp* between their successor indices belonging to the parent interval-node:

Let i, j belong to the same aw-interval and let us assume that $i > j$.

Then $lcp(\tilde{s}[sa[i]..], \tilde{s}[sa[j]..]) = min_{\ell \in [next[successor(i)], successor(j)]} lcp[\ell]$

With this inductive approach, it is sufficient to re-calculate the *lcp* of only the first interval (the root of the left-context tree). This is straightforward (see algorithm 4).

Algorithm 4 Calculate the value of the *lcp* for index i

$recalculate\_lcp\{ESA_{DL}, i\}$

```
1: lcp[i] ← 0
2: if prev[i] ≥ 0 then
3:    i ← sa[i]
4:    j ← sa[prev[i]]
5:    while i < n ∧ j < n ∧ s[i] = s[j] do
6:       i ← i ⊕ 1
7:       j ← j ⊕ 1
8:       lcp[i] ← lcp[i] + 1
9:    end while
10: end if
```

During the iterative call, if an index that is already treated appears, it is skipped. Indeed, its *lcp* value is then up-to-date. The pseudo-code for this step is exposed in algorithm 5.

Because in each step we use the value of all the lines of the previous group, we traverse once again the left context tree in a breadth-first order.

Algorithm 5 Update *lcp* of step k

update_lcp$\{ESA_{DL}^{(k)}, \mathcal{R}_k, \mathcal{O}_k\}$

1: $q \leftarrow queue()$
2: **for** $i \in \mathcal{O}_k$ **do**
3: $\quad recalculate\_lcp(ESA_{DL}^{(k)}, isa[i])$
4: $\quad q.push((predecessor(isa[i]), 1))$
5: **end for**
6: **while** not *q.empty()* **do**
7: $\quad (i, depth) \leftarrow q.top$
8: $\quad q.pop$
9: $\quad$ **if** $i \geq 0 \wedge lcp[i]$ not already updated $\wedge\, lcp[i] \geq depth$ **then**
10: $\quad\quad lcp[i] \leftarrow (min_{j \in [next[successor(prev[i])], successor(i)]} lcp[j]) + 1$
11: $\quad\quad q.push((predecessor(i), depth + 1))$
12: $\quad$ **end if**
13: **end while**

4. Efficiency

The space complexity is $O(n)$. The ESA_{DL} structure needs to complete the *ESA* with two arrays of length n. During the execution, a queue of length $O(n)$, and an array of length n are used to check in constant time whether a couple $(i, depth)$ was already used.

The worst case time complexity of the update algorithm is bounded by $O(n^2)$. This case is reached while replacing for instance AA occurrences in an ESA_{DL} indexing the text A^nT. A better bound on time complexity could be obtained by considering amortized complexity, but it will still be unlikely to be better than the $O(n)$ complexity required for building the suffix array from scratch. Nevertheless, the algorithms building suffix arrays that currently perform best in practical cases, are not the linear ones (see [29] for a description of the different suffix array construction algorithms and their strengths). We propose in this section to evaluate the practical efficiency of our update algorithm, comparing it to the standard approach that recreate the suffix array.

A prototype implementing the proposed algorithm has been developed using the C++ language. It is available at *http://www.irisa.fr/symbiose/mgalle/suffix_array_update*

. It has been tested on different types of text. For the sake of brevity, in this paper we only report the results on the following classical corpora from the literature:

- the standard and large Canterbury corpus (*http://corpus.canterbury.ac.nz/* [4]),
- the Purdue corpus (*http://www.cs.ucr.edu/ stelo/Offline/* [2])

Results on other corpora can be found on our internet site.

To compare the execution time with a recreating from scratch approach, we took three different suffix array creation algorithms: the linear time one proposed by Kärkkäinen and Sanders [10], the non-linear algorithm of Larsson and Sadakane [17] and the Induced Sorting algorithm of Zhang, Nong and Chan [32] (again a linear one). The source code of the first two were retrieved from the Internet sites specified in the associated articles. Note that Kärkkäinen and Sanders' code "strives for conciseness rather than for speed" [10]. For the Induced Sorting algorithm, we used the optimized implementation of Mori [21].

In the last years, suffix array creation algorithms has proven to be a rich field of research. New strategies and improvements are proposed each year, and for a complete taxonomy of the state of art we refer to [26]. But some of them do assumptions over the alphabet that could no be fulfilled by our grammar based application:

(1) the size of the alphabet. Manzini and Ferragina's algorithm [19] and Yuta Mori's *libdivsufsort* [22] suppose a size of alphabet less than 256. In our approach, in each iteration we introduce a new non-terminal, so this bound is too tight.
(2) it is possible that, after a replacement, a letter does not occur any more in the sequence because all its occurrence where inside the selected repeat. That is because we discarded algorithms that suppose a contiguous alphabet (like [18]).

The tests were executed on 1GHz AMD Opteron processors with 4Gb of memory.

First, to have an idea of the complexity of the algorithm, we studied how the length of the sequence influences the execution time of the algorithm. From the large Calgary corpus, we extracted sequences of different lengths by considering successively bigger (by steps of 100 characters) prefixes of the sequences. On each extracted sequence, we performed 250 iterations of selecting a random repeat, replacing it over the sequence by a new character and updating the associated suffix array. Time (user + system time) required for updating the suffix array was reported, averaged over 5 different runs corresponding to 5 different random seeds. The same experiments, replacing the update algorithm by the "from scratch" construction algorithms of the suffix array by Kärkkäinen and Sanders (*K & S*), Larsson and Sadakane (*L & S*) and Zhang, Nong and Chan (*ISA*) have been performed. The plots, shown in figure 3, confirm that the execution time of our updating algorithm is not directly correlated to the length of the sequence, and is significantly smaller than the execution time required by reconstruction "from scratch" algorithms, especially when the length of the sequence increases.

We present a more exhaustive evaluation and comparison on all the corpora using different strategies for the selection of the repeated word. In each test we performed 500 iterations of selecting a repeat, replacing it over the sequence and updating (or building from scratch) the associated suffix array. The different strategies for the selection of the repeat were:

- take a random one (using the same seed for the random number generator),
- take the longest,
- take the one that covers the maximal number of positions[a].

Information over these files are summarized in figure 4. Results are given in figure 5 (page 18). For each selection strategy, we measured time (user + system time) spent in updating ESA_{DL} with our algorithm (column *update*), and time spent in building *ESA* from scratch at each iteration with the three creation algorithms. For easier comparison, we only report the times given by the update algorithm and the ratios of the time spent by each of the three "from scratch" algorithms over the update algorithm. A ratio lower than 1 means that the from scratch algorithm was faster than the update. Time spent by a from scratch algorithm can be obtained by multiplying the time reported in the "update" column by the respective ratio.

Some of the files (notably *fields.c, grammar.lsp* and *xargs.1*) are too small to draw significant conclusions, but results are shown here for the sake of completeness. On the other files, results show that a significant speedup is usually achieved by using our algorithm. The main exceptions are the $Spor\_All\_2x.fasta$ file (an artificial file obtained by concatenating $Spor\_All.fasta$ with itself) from the Purdue corpus, and the *ptt*5 file from the Canterbury corpus (a fax image with very long zones of the same byte). One can also remark that the ratio is less favorable when the repeat to replace is chosen according to the maximal compression strategy. On the one hand, in each iteration the resulting sequence is smaller and the suffix array creation from scratch for this sequence faster. On the other hand, there are more positions affected by the substitution and this affects the update algorithm.

These cases allow us to illustrate an intrinsic limit of the update approach when the length of the sequence is highly reduced by recoding: when the number of positions to update is larger than the number of positions in the resulting sequence, it may be worth adopting the "from scratch" construction algorithm (let us remark that the best algorithm to use can vary along the iterations). A solution to handle these extreme cases, would be to design a criterion on the repeat and its coverage to automatically choose the best algorithm to use (eventually at each iteration).

5. Conclusion and future work

We introduced in this paper an approach allowing to update an enhanced suffix array while substituting some of the occurrences of a word in the indexed text We did not consider singular insertions or deletions, but simultaneuos substitution.This is of particular interest for grammatical inference or grammar based compression methods which use these data structures and are iteratively performing a large number of such substitutions.

[a]choose at each step the repeat that maximize $(|\mathcal{O}_k|-1)*(|w|-1)-2$, which actually corresponds to a maximal compression approach [25]

Our approach uses the specific internal order of suffix arrays to simultaneously update groups of adjacent indices and ensures that only indices to be modified are visited. This specific property of the suffix arrays allows to design an efficient update procedure which has been implemented and tested on classical corpora. The experimentation confirms that, in regard to the direct method reconstructing the suffix array, our approach enables significant speed-up of the execution time of a factor up to 70 when choosing randomly a repeat to replace.

The time complexity of the new algorithm depends mainly of the size of the left context tree. This grows with both the average *lcp* of the sequence, and the number of positions the chosen repeat covers. In some cases - specially when these two factors are big -, the update method is less efficient than building the enhanced suffix array from scratch. Intuitively, when the number of lines to change is larger than the number of lines in the new suffix array, a reconstruction algorithm is likely to be more efficient than an update approach. In order to be even more efficient, a criterion allowing to decide automatically which algorithm to use could be designed. This would require a finer complexity analysis of the update algorithm, but also of the chosen building algorithm, in order to identify easy-to-compute key parameters involved in the execution time complexity.

Of course, the question of the existence of a practical efficient $O(n)$ algorithm remains open. But the results on the construction of suffix arrays suggest that a better way of improvement could be the design of other practical update algorithms. Finally, these results have been obtained by using a suffix array. It would be interesting to study how easily this approach can be adapted to suffix trees and how much it depends on the properties specific to suffix arrays.

References

[1] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch. "Replacing suffix trees with enhanced suffix arrays," *Journal of Discrete Algorithms* **2** (2004) pp 53–86.

[2] A. Apostolico and S. Lonardi. "Compression of biological sequences by greedy off-line textual substitution,". *Proc. Data Compression Conference*. 28-30 March 2000. pp 143–152.

[3] A. Apostolico and S. Lonardi. "Off-line compression by greedy textual substitution,". *Proc. IEEE* volume 88. November 2000. pp 1733–1744.

[4] R. Arnold and T. Bell. "A corpus for the evaluation of lossless compression algorithms,". *Proc. Conference on Data Compression*. Washington, DC, USA. 1997. pp 201.

[5] H.-L. Chan, W.-K. Hon, T.-W. Lam, and K. Sadakane. "Compressed indexes for dynamic text collections," *ACM Transactions on Algorithms* **3** (May 2007).

[6] P. Ferragina, R. Grossi, and M. Montangero. "On updating suffix tree labels," *Theoretical Computer Science* **201** (1998) pp 249–262.

[7] E. Fiala and D. H. Greene. "Data compression with finite windows," *Communications ACM* **32** (1989) pp 490–505.

[8] M. Gu, M. Farach, and R. Beigel. "An efficient algorithm for dynamic text indexing,". *Proc. ACM-SIAM symposium on Discrete Algorithms*. 1994. pp 697–704.

[9] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and*

Computational Biology. (Cambridge University Press January 1997).

[10] J. Kärkkäinen and P. Sanders. "Simple linear work suffix array construction,". *Proc. International Conference on Automata, Languages and Programming*. 2003.

[11] J. Kieffer and E.-H. Yang. "Grammar-based codes: A new class of universal lossless source codes," *IEEE Transactions on Information Theory* **46** (2000).

[12] J. Kieffer and E.-H. Yang. "Grammar-based codes: a new class of universal lossless source codes," *IEEE Transactions on Information Theory* **46** (2000).

[13] P. Ko and S. Aluru. "Space efficient linear time construction of suffix arrays,". *Proc. Combinatorial Pattern Matching* volume 2676. 2003. pp 200–210.

[14] R. Kolpakov and G. Kucherov. "Finding maximal repetitions in a word in linear time,". *Proc. Annual Symposium on Foundations of Computer Science*. New York, USA. 1999. pp 596–604.

[15] K. D. Kyue, S. J. Seop, P. Heejin, and P. Kunsoo. "Linear time construction of suffix arrays,". *Proc. Combinatorial Pattern Matching* volume 2676. 2003. pp 186–2003.

[16] J. K. Lanctot, M. Li, and E.-H. Yang. "Estimating dna sequence entropy,". *Proc. ACM-SIAM symposium on Discrete Algorithms*. 2000. pp 409–418.

[17] N. J. Larsson and K. Sadakane. "Faster suffix sorting,". Technical report Department of Computer Science, Lund University. Sweden. May 1999.

[18] M. A. Maniscalco and S. J. Puglisi. "An efficient, versatile approach to suffix sorting," *J. Exp. Algorithmics* **12** (2008) pp 1–23.

[19] G. Manzini and P. Ferragina. "Engineering a lightweight suffix array construction algorithm," *Algorithmica* **40** (2004) pp 33–50.

[20] E. M. McCreight. "A space-economical suffix tree construction algorithm," *Journal ACM* **23** (1976) pp 262–272.

[21] Y. Mori. "An implementation of the induced sorting algorithm," 2008. `http://yuta.256.googlepages.com/sais`.

[22] Y. Mori. "libdivsufsort," August 2008. `http://code.google.com/p/libdivsufsort/`.

[23] R. Nakamura, H. Bannai, S. Inenaga, and M. Takeda. "Simple linear-time off-line text compression by longest-first substitution,". *Proc. Data Compression Conference*. 2007.

[24] C. Nevill-Manning and I. Witten. "Identifying hierarchical structure in sequences: A linear-time algorithm," *Journal of Artificial Intelligence Research* **7** (1997) pp 67–82.

[25] C. Nevill-Manning and I. Witten. "On-line and off-line heuristics for inferring hierarchies of repetitions in sequences,". *Proc. Data Compression Conference*. Nov 2000.

[26] S. Puglisi, W. Smyth, and A. Turpin. "A taxonomy of suffix array construction algorithms," *ACM Computing Surveys* **39** (2007).

[27] S. C. Sahinalp and U. Vishkin. "Efficient approximate and dynamic matching of patterns using a labeling paradigm,". *Proc. Annual Symposium on Foundations of Computer Science*. 1996.

[28] M. Salson, T. Lecroq, M. Léonard, and L. Mouchard. "Dynamic burrows-wheeler transform,". *Proc. Prague Stringology Club*. 2008.

[29] K.-B. Schürmann and J. Stoye. "An incomplex algorithm for fast suffix array construction," *Software - Practice and Experience* **37** (2007) pp 309–329.

[30] J. S. Sim. "Time and space efficient search for small alphabets with suffix arrays,". *Proc. Conference on Fuzzy Systems and Knowledge Discovery*. 2005.

[31] E. Ukkonen. "On-line construction of suffix trees," *Algorithmica* **14** (1995) pp 249–260.

[32] S. Zhang, G. Nong, and W. H. Chan. "Fast and space efficient linear suffix array construction,". *Proc. Data Compression Conference*. 2008.

Fig. 3. Large corpus: *bible.txt*, *world192.txt* and *E.coli*. Times are given in hundredth of seconds and the size in thousands of characters

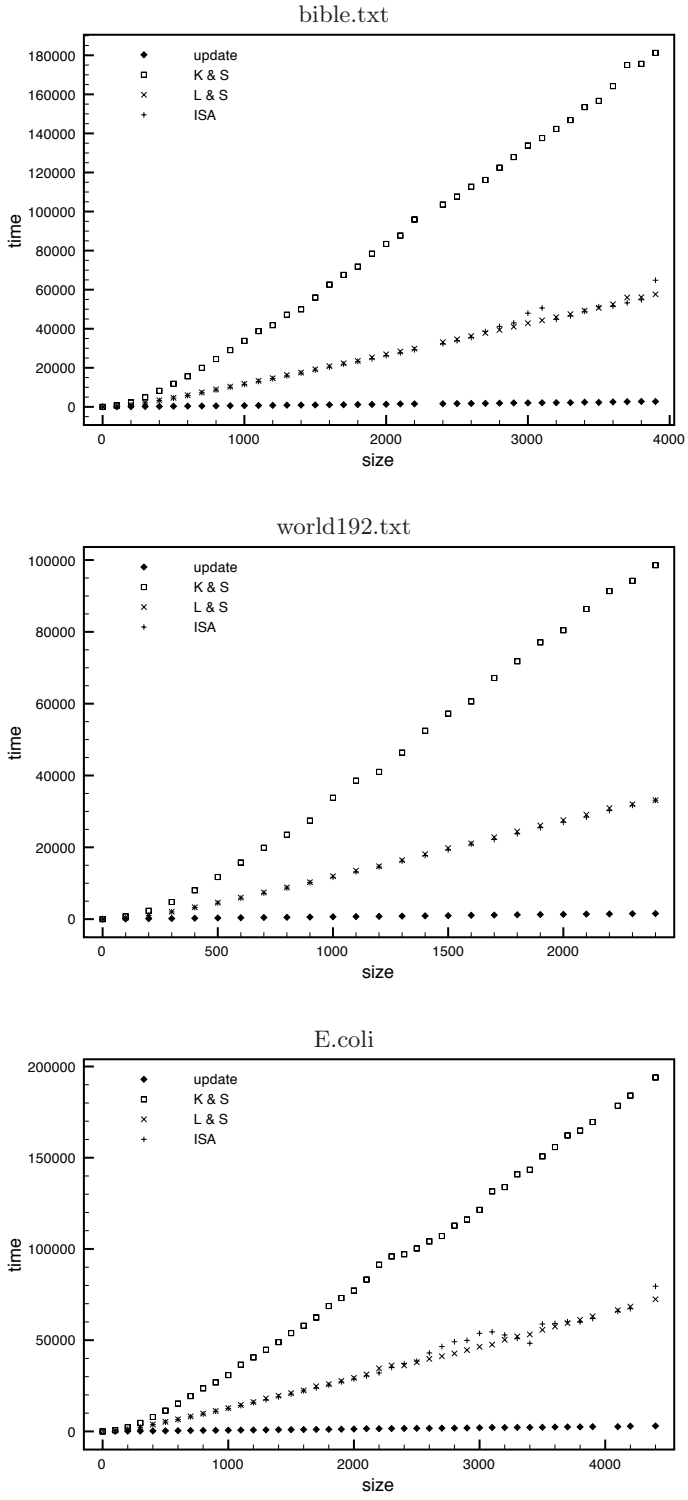


Fig. 4. Information over the tested files.

file	**size (number of symbols)**	**average lcp**	**alphabet size**
CANTERBURY CORPUS			
alice29.txt	152089	7.76	74
asyoulik.txt	125179	6.61	68
cp.html	24603	12.47	86
fields.c	11150	12.67	90
grammar.lsp	3721	8.63	76
kennedy.xls	1029744	7.56	256
lcet10.txt	426754	10.32	84
plrabn12.txt	481861	7.12	81
ptt5	513216	2353.31	159
sum	38240	51.31	255
xargs.1	4227	5.35	74
LARGE CORPUS			
bible.txt	4047392	13.97	63
E.coli	4638690	17.38	4
world192.txt	2473400	23.0	94
PURDUE CORPUS			
All_Up_1M.fasta	1001002	18.75	46
All_Up_400k.fasta	399615	15.32	46
Helden_All.fasta	112507	20.84	61
Helden_CGN.fasta	32871	7.2	51
Spor_All_2x.fasta	444906	56022.6	54
Spor_All.fasta	222453	818.061	54
Spor_EarlyI.fasta	31039	7.12	49
Spor_EarlyII.fasta	25008	6.94	46
Spor_Middle.fasta	54325	7.57	51

Note: The files from the purdue corpus contains comments in fasta notation, but most of the sequences are composed only of 4 symbols

Fig. 5. Comparison between update and reconstruction from scratch of the suffix array. Times are given in hundredth of seconds. A ratio lower than 1 means that the from scratch algorithm was faster than the update algorithm

	random				maximal length				maximal compression			
	update time	K & S ratio	L & S ratio	ISA ratio	update time	K & S ratio	L & S ratio	ISA ratio	update time	K & S ratio	L & S ratio	ISA ratio
CANTERBURY												
alice29.txt	163	17.25	9.18	9.47	192	12.28	7.14	7.45	269	4.06	1.9	2.61
asyoulik.txt	131	16.11	8.47	8.78	127	13.6	8.34	8.69	182	4.76	2.23	2.88
cp.html	15	8.8	6.33	6.6	15	6.4	4.27	5	18	3.06	2.22	2.83
fields.c	6	6.33	5.17	6.17	8	2.38	2.63	3.88	3	6	2	5
grammar.lsp	3	1.67	1.67	3	0	div 0	div 0	div 0	0	div 0	div 0	div 0
kennedy.xls	1323	26.38	9.7	9.73	1230	29.24	11.22	10.87	1541	3.16	1.08	1.5
lcet10.txt	1248	5.92	3.7	6.07	522	31.51	12.35	14.01	749	7.76	3.02	3.97
plrabn12.txt	516	33.24	13.32	19.42	606	31.84	15.35	16.26	887	8.84	3.28	4.49
ptt5	588	38.53	15.06	4.81	696	7.65	5.32	3.41	1900	0.44	0.19	0.28
sum	42	5.57	3.6	3.9	34	5.5	2.91	4	28	2.93	1.71	2.18
xargs.1	6	4.17	1.5	4.83	2	3	1	4	2	2	1	7
LARGE												
bible.txt	5055	66.81	22.84	22.8	5168	64.39	22.5	21.96	10285	15.37	3.7	5.41
E.coli	5534	69.14	27.36	26.59	6307	53.46	24.03	21.8	14808	9.51	2.11	3.4
world192.txt	3084	65.06	21.75	22.12	3089	60.7	21.11	21.2	5573	16.28	4.54	5.8
PURDUE												
All_Up_1M.fasta	1238	49.8	19.87	19.03	1200	46.16	19.99	19.34	2350	6	1.77	2.66
All_Up_400k.fasta	501	27.86	13.53	13.88	481	27.64	13.93	13.88	884	3.12	1.28	1.74
Helden_All.fasta	119	12.7	8.09	7.51	122	11.17	7.65	7.11	165	2.32	1.16	1.61
Helden_CGN.fasta	31	7.87	5.55	4.87	34	6.82	5.24	5.65	19	2.89	2.63	2.95
Spor_All_2x.fasta	112	0.73	0.84	0.5	57	0.6	0.77	0.26	61	0.57	1.16	0.25
Spor_All.fasta	246	14.87	8.57	8.29	250	13.26	8.56	8.25	413	1.88	0.97	1.22
Spor_EarlyI.fasta	34	5.5	4.47	4.76	26	8.46	7.31	8.27	25	2.24	1.88	4.92
Spor_EarlyII.fasta	20	7.25	7.55	7.9	15	11.07	8.07	10.2	33	1.82	1.18	1.27
Spor_Middle.fasta	51	10.31	6.88	6.78	62	8.16	6.39	5.85	73	1.6	0.9	1.62