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 speedup 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, grammatical inference

1991 Mathematics Subject Classification: 68P05, 68P20

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 [23] in the framework formalized by Kieffer and Yang [11]. 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, can be rewritten by the 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 mainly based on the successive detection of repeats. They differ mostly in the order in which repeats are factorized. In SEQUITUR [23] and its variant [11], 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 et al [22] and Lanctot, Li, and Yang [15] 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 [24].

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 [26], by updating a suffix tree [5–8, 19] or suffix array [27] 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 [15], 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, [22] 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, 12, 13] but non-linear algorithms [16, 18] 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 [24]. 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 \( \Sigma \). The number of characters in \( \Sigma \) is denoted by \( |\Sigma| \). A sequence \( s \) of length \( n \) on \( \Sigma \) is represented by \( s[0]s[1]...s[n-1] \), where \( s[i] \in \Sigma \ \forall \ 0 \leq i < n \). We denote by \( s[i,j] \) (\( j \geq i \)) the sequence \( s[i]s[i+1]...s[j] \) 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 \( \Sigma \) with an order \( \prec \) extensible to \( \Sigma^* \). This lexicographically extension will be denoted also by \( \prec \). Let \( \tilde{s} = s\$ \), with a special character \$ not contained in \( \Sigma \), smaller than every element of \( \Sigma \).

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,

\[ lcp[0] = 0, \]

and \( \forall \ i \in [1, n] : \ lcp[i] = k \) 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 \( sa \), 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 computation in \( O(n) \) 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 \textit{position} when referring to the index over a sequence and \textit{index} when referring to any of the arrays of an ESA.
The algorithm presented below consist mostly of moving and deleting lines of the ESA 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. So, if for example the index i must be deleted, that can be easily done by setting next[prev[i]] to next[i] and prev[next[i]] to prev[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. Moreover, because of ESA_DL lines moving, the result of indices comparison may not coincide with the order of the associated suffixes. For instance, index i may correspond to a suffix with a lexicographically order greater than a suffix corresponding to index j, even if i < j. Anyway, an ESA_DL still allows the detection of repeats (general repeats, maximal repeats or super maximal repeats) in linear time, because the involved algorithms advance one by one over the arrays like most of the algorithm over ESA (a notable exception is the algorithm searching for a substring proposed in [29]). Finally, we remark that the standard ESA can directly be recovered in one simple pass from ESA_DL.

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 jth position after position i, is denoted by i ⊕ j. We compute i ⊕ j using links between sequence positions, indicating for each position its successor. Similarly i ⊖ j points to the jth position before i. We define that, if i ⊕ j (respectively i ⊖ j) is out of range, then i ⊕ j = n + 1 (respectively i ⊖ 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 at step k. At each step k the algorithm i) finds a repeat R_k in a sequence ˜s(k) defined on the alphabet Σ_k and returns a list O_k of non-overlapping occurrences of R_k ii) updates the sequence ˜s(k) and its associated ESA_DL replacing the given occurrences of R_k by a single new character C_k, thus defining a new alphabet Σ_{k+1} = Σ_k ∪ {C_k}. The modified sequence is then called ˜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 a $R_k$ occurrence; 2) move indices with respect to the alphabetic order of $C_k$; and 3) update $lcp$ array with respect to recoded occurrences of $R_k$ 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.

To better understand the different steps of the algorithms and the modifications they perform over the suffix array, we will define the concept of left context tree. It is worth noticing that we present this structure in order to help the understanding of our approach and that it is not actually implemented.

2.1. 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 (substring) 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 $\omega$ where the nodes correspond to a left context of $\omega$.

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

This definition is thus slightly more general than the definition of $\omega$-interval given by Abouelhoda, Kurtz and Ohlebusch [1], since in our approach $\omega$-interval are defined also for words leading to implicit nodes of a compact suffix tree, and not only to internal nodes.

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

- the root is the $\omega$-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{align*}
\text{isa}\[\text{sa}[i] \oplus 1]\text{ if } \text{sa}[i] \oplus 1 \neq n + 1 \\
n + 1 \text{ otherwise},
\end{align*}
\}

and \[\text{predecessor}(i) = \{
\begin{align*}
\text{isa}\[\text{sa}[i] \ominus 1]\text{ if } \text{sa}[i] \neq 0 \\
-1 \text{ otherwise}.
\end{align*}
\] \]

One may remark that \text{predecessor} is the equivalent of the “suffix link” in a suffix tree [30].

The problem that an ESA update algorithm must face is that the changes over the occurrences of a word \(\omega\) not only affect the \(\omega\)-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 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 \text{predecessor}(k_1) and \text{predecessor}(k_2) belong to the same \(\alpha v\)-interval \((\alpha \in \Sigma)\). Then \text{predecessor}(k_1) > \text{predecessor}(k_2).

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

### 3. Algorithm

We now detail the three tasks for updating an ESA\(_{DL}\) while replacing a set of occurrences \(O_k\) of a work \(R_k\) by a simple character \(C_k\).

#### 3.1. Delete indices of suffixes occurring inside \(R_k\) substituted occurrences

By replacing the word \(R_k\) by a single letter, the sequence is compressed and so is its ESA\(_{DL}\); consequently, any suffix of sequence \(\tilde{s}^{(k)}\) starting inside an \(R_k\) substituted occurrence must be deleted. Thus for \(i \in O_k\) and for \(\ell \in [1, |R_k| - 1]\), suffix \(\tilde{s}^{(k)}[i \oplus \ell..]\) and the associated index in the suffix array \(j = \text{isa}[i \oplus \ell]\) have to be removed. We simulated this deletion by jumping over it by setting next and prev arrays to their previous and next index: \(\text{next}[\text{prev}[j]] \leftarrow \text{next}[j]\) and \(\text{prev}[\text{next}[j]] \leftarrow \text{prev}[j]\). Furthermore, the \(lcp\) value of the index following \(j\) (\(lcp[\text{next}[j]]\)) 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 \(\text{next}[j]\) is equal to the minimal longest common prefix value of indices \(j\) and \(\text{next}[j]\).

An example is shown in Figure 1 where the deletion of index \(j\) affects the \(lcp[\text{next}[j]]\) 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 \(\text{prev}[n + 1]\).
In-Place Update of Suffix Array While Recoding Words

Algorithm 1: Delete indices at step $k$, replacing $R_k$ by $C_k$

```pseudo
delete_indices\{ESA^{(k)}_DL, R_k, \mathcal{O}_k\}

1: for $i \in \mathcal{O}_k$ do
2:    for $\ell \in [1, |R_k| - 1]$ do
3:       $j \leftarrow$ isa[$i \oplus \ell$]
4:       if $\text{next}[j] \neq \text{END}$ then
5:          $lcp[\text{next}[j]] \leftarrow \min(lcp[j], lcp[\text{next}[j]])$
6:       end if
7:       $\text{next}[\text{prev}[j]] = \text{next}[j]$
8:       $\text{prev}[\text{next}[j]] = \text{prev}[j]$
9:    end for
10: end for
```

### 3.2. Move indices, with respect to the alphabetic order of $C_k$

After replacing the word $R_k$ by the new character $C_k$, some $ESA_{DL}$ lines may be misplaced with respect to the chosen order of $C_k$ in $\Sigma_{k+1}$.

Indices in the $R_k$-interval are potentially misplaced. In fact, for $v \in \Sigma_k^*$, indices inside an $vR_k$-interval are misplaced if the substitution of $R_k$ into $C_k$ affects their lexicographical order with respect to the previous and next index over the suffix array. Thus, lines belonging to node-intervals of the left-context tree of $R_k$ may have to be moved.

In our approach, we decided to give to $C_k$ the largest rank in the lexicographical order of the alphabet $\Sigma_k$, i.e. $\forall a \in \Sigma_k : a \prec C_k$.

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

If a $vR_k$-interval is already at the end of the $v$-interval (it is naturally well ordered), for any $v' \in \Sigma_k^*$, the $v'vR_k$-interval is also at the end of the $v'vR_k$-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 $R_k$-interval. During recursion, if the group of a $vR_k$-interval is moved, the recursion continues on groups of $avR_k$ - intervals, with $a \in \Sigma_k$.

From a theoretical point of view, the algorithm starts on the root of the left-

<table>
<thead>
<tr>
<th>index</th>
<th>prev</th>
<th>next</th>
<th>lcp</th>
<th>suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$j-1$</td>
<td>$j-2$</td>
<td>$#j+1$</td>
<td>4</td>
<td>ATAC...</td>
</tr>
<tr>
<td>$j$</td>
<td>$#j+1$</td>
<td>$#j+1$</td>
<td>2</td>
<td>ATGT...</td>
</tr>
<tr>
<td>$j+1$</td>
<td>$#j-1$</td>
<td>$j+2$</td>
<td>2</td>
<td>ATGT...</td>
</tr>
</tbody>
</table>

Fig. 1. Deletion of index $j$. 

- **index**, **prev**, **next**, **lcp**, **suffix**
Gallé, Peterlongo and Coste

context tree of $R_k$ 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 $vR_k$-interval works as follows:

1. detects the end position of the $vR_k$-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. calls 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).

In first place, the end of the $vR_k$-interval is found (lines 5, 6 and 8).

This is done from the first element of the interval, following the \textit{next} array while the visited index corresponds to a suffix starting with $vR_k$ ($\text{lcp}[i] \geq |v| + |R_k|$). 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 ($\text{lcp}[i] \geq |v|$). 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 $i_{\text{start}}, i_{\text{end}}, i_{\text{dest}}$ be respectively the starting and ending positions of the $vR_k$-interval, and the last position of the $v$-interval. Moving the group $[i_{\text{start}}, i_{\text{end}}]$ to the position after $i_{\text{dest}}$ is easily done by
Algorithm 2 Restore consistency of suffix array order

\[\text{update-order}\{ ESA^{(k)}_{DL}, R_k, O_k, i_{start}, \text{depth}, \text{move}\} \]

1: if Couple \((i_{\text{start}}, \text{depth})\) already treated during another recursion call then
2:   End procedure
3: end if
4: \(i \leftarrow i_{\text{start}}\)
5: \(\text{while } i \neq \text{END} \land \text{lcp}[\text{next}[i]] \geq \text{depth} + |R_k| \text{ do}\)
6:   \(i \leftarrow \text{next}[i]\)
7: end while
8: \(i_{\text{end}} \leftarrow i\)
9: \(\text{minLCP} \leftarrow \min_{j \in [i_{\text{start}}, i_{\text{end}}]} \text{lcp}[j]\)
10: if \(\text{move}\) then
11:   \(\text{while } i \neq \text{END} \land \text{lcp}[\text{next}[i]] \geq \text{depth} \text{ do}\)
12:     \(i \leftarrow \text{next}[i]\)
13: end while
14: end if
15: \(i_{\text{dest}} \leftarrow i\)
16: if \(i_{\text{dest}} \neq i_{\text{end}}\) then
17:   \(\text{lcp}[\text{next}[i_{\text{end}}]] \leftarrow \min(\text{lcp}[\text{next}[i_{\text{end}}]], \text{minLCP})\)
18:   \(\text{lcp}[i_{\text{start}}] \leftarrow \text{depth}\)
19: if \(i_{\text{start}} = i_{\text{first}} \land \text{depth} \neq 0\) then
20:   \(i_{\text{first}} \leftarrow \text{next}[i_{\text{end}}]\)
21: end if
22: move group\((i_{\text{start}}, i_{\text{end}}, i_{\text{dest}})\)
23: else
24:   \(\text{lcp}[i_{\text{start}}] \leftarrow \min(\text{lcp}[i_{\text{start}}], \text{depth})\)
25:   \(\text{move} \leftarrow \text{false}\)
26: end if
27: \(i \leftarrow i_{\text{start}}\)
28: \(\text{while } i \neq \text{next}[i_{\text{end}}] \text{ do}\)
29:   \(\text{newdepth} \leftarrow \text{depth} + (\text{if } \text{predecessor}(i) \in O_k \text{ then } \text{ten } \text{else } 1)\)
30: if \(\text{move} \lor (\text{sa}[\text{prev}[\text{predecessor}(i)]) > \text{newdepth} \land \text{sa}[\text{prev}[\text{predecessor}(i)]) \oplus \text{newdepth} \in O_k) \text{ then}\)
31:   \text{update-order}\{ ESA^{(k)}_{DL}, R_k, O_k, \text{preddecessor}(i), \text{newdepth}, i_{\text{dest}} \neq i_{\text{end}}\)
32: end if
33: \(i \leftarrow \text{next}[i]\)
34: end while

jumping over the group and inserting it into \(i_{\text{dest}}\) and \(\text{next}[i_{\text{dest}}]\). 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:
10 Gallé, Peterlongo and Coste

(1) $lcp[\text{next}[i_{end}]]$: contains the value of the length of the longest common prefix between $\text{prev}[i_{start}]$ and $\text{next}[i_{end}]$, 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 $\text{depth}$, that is the correct value over $\tilde{s}_{k+1}$.

This serves also to set a stop-point for future recursions calls (see below).

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

```
move_group(ESADL_{(k)}, R_k, \mathcal{O}_k, i_{start}, i_{end}, i_{dest})
1: next[\text{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}
```

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 $ESADL$ of sequence $GAAGAAGC$, where $R_1 = GA$ is substituted by $C_1$. 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 $GAAGC$.

3.2.1. 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 situation, where the substituted repeat is $TA$.

```
i   CTTTAC...
i+1 CTTTAG...
i+2 CTATA....
```

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.

3.2.2. Filtering non substituted \( R_k \) occurrences

Among each \( vR_k \)-interval, suffixes starting with \( vR_k \) where \( R_k \) is not substituted (whose position does not belong to \( O_k \)) may occur. The associated indices in the \( ESA_{DL} \) should not be moved with the \( vR_k \)-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 (\( sa[i] \oplus \text{depth} \in O_k \)). Once a non-selected occurrence is detected, we move it to the beginning of the group (before \( i_{\text{start}} \)). As previously mentioned, this also involves modifications of the \( lcp \) array for maintaining its consistency. This step is basically a simplified version of algorithm 2. It adds an extra auxiliary array of size \( n \) to keep track, for each index, of the last depth with which it was analyzed.

3.3. Update lcp values after the substitution of \( R_k \) occurrences to a single character

The substitution of any occurrence of \( R_k \) of length \( |R_k| \geq 2 \) by \( C_k \) of length 1 involves the modification of the length of all common prefixes involving such an occurrence.

In the previous step, it was trivial to update the \( lcp \) values of the border lines. However, in this step, we update the \( lcp \) values of the internal position of the intervals. Straightforwardly subtracting \( |R_k| - 1 \) from each internal \( lcp \) value misses the cases where the common prefix between two successive suffixes include more than one occurrence of \( R_k \), or even worse, a part of a occurrence (consider for instance the example shown in section 3.2.1).

So we traverse again the left-context tree of \( R_k \). 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 (\( R_k \)-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 \( lcp(s[sa[i]..], s[sa[j]..]) = \min_{f \in \text{next}[\text{successor}(i)], \text{successor}(j)} \{ lcp[f] \} \)

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 $i$

\begin{align*}
\text{recalculate } & \text{lcp} \{ESA_{DL}, i\} \\
1: & \text{lcp}[i] \leftarrow 0 \\
2: & \text{if } \text{prev}[i] \geq 0 \text{ then} \\
3: & \quad i \leftarrow \text{sa}[i] \\
4: & \quad j \leftarrow \text{sa}\{\text{prev}[i]\} \\
5: & \quad \text{while } i < n \land j < n \land \text{s}[i] = \text{s}[j] \text{ do} \\
6: & \quad \quad i \leftarrow i \oplus 1 \\
7: & \quad \quad j \leftarrow j \oplus 1 \\
8: & \quad \quad \text{lcp}[i] \leftarrow \text{lcp}[i] + 1 \\
9: & \quad \text{end while} \\
10: & \text{end if}
\end{align*}

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.

Algorithm 5 Update lcp of step $k$

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

Because in each iteration 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.

4. Efficiency

4.1. Time efficiency

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 [28] 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 builds the suffix array from scratch.

A prototype implementing the proposed algorithm has been developed using the C++ language. 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:

- standard and large Canterbury corpus (http://corpus.canterbury.ac.nz/[4])
- 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 building from scratch approach, we used three different suffix array creation algorithms: the linear time one proposed by Kärkkäinen and Sanders [10], the non-linear algorithm of Larsson and Sadakane [16] and the Induced Sorting algorithm of Zhang, Nong and Chan [31] (again a linear one). The source code of the first two were retrieved from the web 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 [20].

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 [25]. But some of them do assumptions over the alphabet that could not be fulfilled by our grammar based application and that is because we could not compare them with our algorithm. The two assumption that excluded some of them were:

1. the size of the alphabet. Manzini and Ferragina’s algorithm [18] and Yuta Mori’s libdivsufsort [21] 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 why we discarded algorithms that suppose a contiguous alphabet (like [17]).

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

*available at http://www.irisa.fr/symbiose/projects/suffix_array_update
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 kilobytes) 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 (ZNC) 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 construction 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 pseudo-random number generator),
- take the longest,
- take the one that covers the maximal number of positions\(^b\).

Information about these files are summarized in figure 4. Results are given in figure 6 (page 22). 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

\(^b\) choose at each step the repeat that maximizes $|O_k| \times (|w| - 1) - 2$, which actually corresponds to a maximal compression approach [24]
the *ptt5* 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 (even at each iteration).

### 4.2. Space efficiency

The overall space complexity is $O(n)$. In this section we analyze this bound more precisely.

Storing the *ESA* requires 3 arrays of integer, and the sequence is also stored in an integer array (recall that our approach is supposed to work with integer alphabets). On a 32-bit architecture, this equals $10n$ bytes (where $n$ is the amount of symbols of the input). The *ESA* structure needs to extend the *ESA* with two arrays of length $n$ (*next* and *prev*). To implement the $\oplus$ and $\ominus$ operators, two extra arrays of size $n$ were used. The recursion in algorithm 2 was implement with a queue. Like the queue of algorithm 5, it is bounded by $n$. An array of length $n$ is used to check in constant time whether a couple $(i, depth)$ was already used and a last auxiliary array of size $n$ is used as specified in section 3.2.2. To sum up, the memory needed by the algorithm is $40n$, plus at most $4n$ for the queues.

To see the practical memory usage, we measured it during the execution of the first type of test (section 4.1) on E. Coli. In figure 5 we plotted the results for the update and the three from scratch algorithms. Note that in this case we measured the memory used by the whole process, while in section 4.1 the time spent in searching the repeats was not taken into account.

It is worth noticing that in the from scratch approach, the memory usage has a peak in the first iteration and then decreases, while in the update approach the memory occupied by the *ESA* remains the same. This cannot be observed in figure 5 because we only measured the maximal memory usage.

Each of the four curves shows a linear behavior. In general, both *L&S* and *ZN C* algorithm uses $26n$ memory. This is consistent with the 3 arrays used for the *ESA* and the sequence, plus one to store the new sequence. The other $6n$ can be attributed to the algorithm that recovers the repeats. The fact that there is no apparent difference between both algorithms can be explained again by the fact that we measured only the maximal memory used. *K&S* uses much more memory,
what is consistent with other reported results [21]. We can also observe that the
memory usage of our update algorithm reaches in this test the predicted 44$n$
upper bound.

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. Considering simultaneous substitutions 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.

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 speedup 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 specific cases - 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 interest-
ing to study how easily this approach can be adapted to suffix trees and how
much it depends on the specific properties of suffix arrays.

References

[2] A. Apostolico and S. Lonardi. “Compression of biological sequences by greedy off-
pp 143–152.


Fig. 3. Large corpus: bible.txt, world192.txt and E.coli. Times are given in hundredth of seconds and the size in kilobytes (1 byte = 1 character)
Fig. 4. Information about the tested files.

<table>
<thead>
<tr>
<th>file</th>
<th>size (number of symbols)</th>
<th>average lcp</th>
<th>alphabet size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CANTERBURY CORPUS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alice29.txt</td>
<td>152089</td>
<td>7.76</td>
<td>74</td>
</tr>
<tr>
<td>asyoulik.txt</td>
<td>125179</td>
<td>6.61</td>
<td>68</td>
</tr>
<tr>
<td>cp.html</td>
<td>24603</td>
<td>12.47</td>
<td>86</td>
</tr>
<tr>
<td>fields.c</td>
<td>11150</td>
<td>12.67</td>
<td>90</td>
</tr>
<tr>
<td>grammar.lsp</td>
<td>3721</td>
<td>8.63</td>
<td>76</td>
</tr>
<tr>
<td>kennedy.xls</td>
<td>1029744</td>
<td>7.56</td>
<td>256</td>
</tr>
<tr>
<td>lcet10.txt</td>
<td>426754</td>
<td>10.32</td>
<td>84</td>
</tr>
<tr>
<td>plrabn12.txt</td>
<td>481861</td>
<td>7.12</td>
<td>81</td>
</tr>
<tr>
<td>pt5</td>
<td>513216</td>
<td>2353.31</td>
<td>159</td>
</tr>
<tr>
<td>sum</td>
<td>38240</td>
<td>51.31</td>
<td>255</td>
</tr>
<tr>
<td>xargs.1</td>
<td>4227</td>
<td>5.35</td>
<td>74</td>
</tr>
<tr>
<td><strong>LARGE CORPUS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bible.txt</td>
<td>4047392</td>
<td>13.97</td>
<td>63</td>
</tr>
<tr>
<td>E.coli</td>
<td>4638690</td>
<td>17.38</td>
<td>4</td>
</tr>
<tr>
<td>world192.txt</td>
<td>2473400</td>
<td>23.0</td>
<td>94</td>
</tr>
<tr>
<td><strong>PURDUE CORPUS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All_Up_1M.fasta</td>
<td>1001002</td>
<td>18.75</td>
<td>46</td>
</tr>
<tr>
<td>All_Up_400k.fasta</td>
<td>399615</td>
<td>15.32</td>
<td>46</td>
</tr>
<tr>
<td>Helden_ALL.fasta</td>
<td>112507</td>
<td>20.84</td>
<td>61</td>
</tr>
<tr>
<td>Helden_CGN.fasta</td>
<td>32871</td>
<td>7.2</td>
<td>51</td>
</tr>
<tr>
<td>Spor_ALL_2x.fasta</td>
<td>444906</td>
<td>56022.6</td>
<td>54</td>
</tr>
<tr>
<td>Spor_ALL.fasta</td>
<td>222453</td>
<td>818.061</td>
<td>54</td>
</tr>
<tr>
<td>Spor_EarlyI.fasta</td>
<td>31039</td>
<td>7.12</td>
<td>49</td>
</tr>
<tr>
<td>Spor_EarlyII.fasta</td>
<td>25008</td>
<td>6.94</td>
<td>46</td>
</tr>
<tr>
<td>Spor_Middle.fasta</td>
<td>54325</td>
<td>7.57</td>
<td>51</td>
</tr>
</tbody>
</table>

*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. Memory consumption for E. Coli. Memory and size are given in kilobytes.
Fig. 6. Update time and speedup factor with respect to each of the from scratch algorithm. Times are given in hundredth of seconds.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>update time</th>
<th>random speedup factor</th>
<th>max. length update time</th>
<th>max. length speedup factor</th>
<th>max. compression update time</th>
<th>max. compression speedup factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CANTERBURY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alice29.txt</td>
<td>163</td>
<td>17.25</td>
<td>9.18</td>
<td>9.47</td>
<td>192</td>
<td>12.28</td>
</tr>
<tr>
<td>asyoulik.txt</td>
<td>131</td>
<td>16.11</td>
<td>8.47</td>
<td>8.78</td>
<td>127</td>
<td>13.6</td>
</tr>
<tr>
<td>cp.html</td>
<td>15</td>
<td>8.8</td>
<td>6.33</td>
<td>6.6</td>
<td>15</td>
<td>6.4</td>
</tr>
<tr>
<td>fields.c</td>
<td>6</td>
<td>6.33</td>
<td>5.17</td>
<td>6.17</td>
<td>8</td>
<td>2.38</td>
</tr>
<tr>
<td>grammar.lsp</td>
<td>3</td>
<td>1.67</td>
<td>1.67</td>
<td>3</td>
<td>0</td>
<td>div 0</td>
</tr>
<tr>
<td>kennedy.xls</td>
<td>1323</td>
<td>26.38</td>
<td>9.7</td>
<td>9.73</td>
<td>1230</td>
<td>29.24</td>
</tr>
<tr>
<td>lct10.txt</td>
<td>1248</td>
<td>5.92</td>
<td>3.7</td>
<td>6.07</td>
<td>522</td>
<td>31.51</td>
</tr>
<tr>
<td>plrabn12.txt</td>
<td>516</td>
<td>33.24</td>
<td>13.32</td>
<td>19.42</td>
<td>606</td>
<td>31.84</td>
</tr>
<tr>
<td>ptt5</td>
<td>588</td>
<td>38.53</td>
<td>15.06</td>
<td>4.8</td>
<td>696</td>
<td>7.65</td>
</tr>
<tr>
<td>sum</td>
<td>42</td>
<td>5.57</td>
<td>3.6</td>
<td>3.9</td>
<td>34</td>
<td>5.5</td>
</tr>
<tr>
<td>xargs.1</td>
<td>6</td>
<td>4.17</td>
<td>1.5</td>
<td>4.83</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>LARGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bible.txt</td>
<td>5055</td>
<td>66.81</td>
<td>22.84</td>
<td>22.8</td>
<td>5168</td>
<td>64.39</td>
</tr>
<tr>
<td>E.coli</td>
<td>5534</td>
<td>69.14</td>
<td>27.36</td>
<td>26.59</td>
<td>6307</td>
<td>53.46</td>
</tr>
<tr>
<td>world192.txt</td>
<td>3084</td>
<td>65.06</td>
<td>21.75</td>
<td>22.12</td>
<td>3089</td>
<td>60.7</td>
</tr>
<tr>
<td>PURDUE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AllUpM.fasta</td>
<td>1238</td>
<td>49.8</td>
<td>19.87</td>
<td>19.03</td>
<td>1200</td>
<td>46.16</td>
</tr>
<tr>
<td>AllUp_400k.fasta</td>
<td>501</td>
<td>27.86</td>
<td>13.53</td>
<td>13.88</td>
<td>481</td>
<td>27.64</td>
</tr>
<tr>
<td>Helden_All.fasta</td>
<td>119</td>
<td>12.7</td>
<td>8.09</td>
<td>7.51</td>
<td>122</td>
<td>11.17</td>
</tr>
<tr>
<td>Helden_CGN.fasta</td>
<td>31</td>
<td>7.87</td>
<td>5.55</td>
<td>4.87</td>
<td>34</td>
<td>6.82</td>
</tr>
<tr>
<td>Spor_All2x.fasta</td>
<td>112</td>
<td>4.73</td>
<td>0.84</td>
<td>0.5</td>
<td>57</td>
<td>0.6</td>
</tr>
<tr>
<td>Spor_All.fasta</td>
<td>246</td>
<td>14.87</td>
<td>8.57</td>
<td>8.29</td>
<td>250</td>
<td>13.26</td>
</tr>
<tr>
<td>Spor_EarlyI.fasta</td>
<td>34</td>
<td>5.5</td>
<td>4.47</td>
<td>4.76</td>
<td>26</td>
<td>8.46</td>
</tr>
<tr>
<td>Spor_EarlyII.fasta</td>
<td>20</td>
<td>7.25</td>
<td>7.55</td>
<td>7.9</td>
<td>15</td>
<td>11.07</td>
</tr>
<tr>
<td>Spor_Middle.fasta</td>
<td>51</td>
<td>10.31</td>
<td>6.88</td>
<td>6.78</td>
<td>62</td>
<td>8.16</td>
</tr>
</tbody>
</table>

Note: A speedup factor lower than 1 means that the from scratch algorithm was faster than the update algorithm (all these cases are underlined).