

A Faster Implementation of Online Run-Length Burrows-Wheeler Transform

Tatsuya Ohno Yoshimasa Takabatake Tomohiro I
 Hiroshi Sakamoto
 Kyushu Institute of Technology, Japan
 {t_ohno, takabatake}@donald.ai.kyutech.ac.jp,
 {tomohiro, hiroshi}@ai.kyutech.ac.jp

Abstract

Run-length encoding Burrows-Wheeler Transformed strings, resulting in *Run-Length BWT (RLBWT)*, is a powerful tool for processing highly repetitive strings. We propose a new algorithm for online RLBWT working in run-compressed space, which runs in $O(n \lg r)$ time and $O(r \lg n)$ bits of space, where n is the length of input string S received so far and r is the number of runs in the BWT of the reversed S . We improve the state-of-the-art algorithm for online RLBWT in terms of empirical construction time. Adopting the dynamic list for maintaining a total order, we can replace rank queries in a dynamic wavelet tree on a run-length compressed string by the direct comparison of labels in a dynamic list. The empirical result for various benchmarks show the efficiency of our algorithm, especially for highly repetitive strings.

1 Introduction

1.1 Motivation

The *Burrows-Wheeler Transform (BWT)* [8] is one of the most successful and elegant technique for lossless compression. When a string contains several frequent substrings, the transformed string would have several *runs*, i.e., maximal repeat of a symbol. Then, such a BWT string is easily compressed by run-length compression. We refer to the run-length compressed string as the *Run-Length BWT (RLBWT)* of the original string. Because of the definition of BWT, the number r of runs in the RLBWT is closely related to the easiness of compression of the original string. In fact, r can be (up to) exponentially smaller than the text length, and several studies [4, 12, 18, 19] showed that r is available for a measure of repetitiveness.

After the invention of BWT, various applications have been proposed for string processing [7, 9, 10]. The most notable one would be the BWT based self-index, called FM index [10], which allows us to search patterns efficiently while storing text in the entropy-based compressed space. However, the traditional entropy-based compression is not enough to process highly repetitive strings because it does not capture the compressibility in terms of repetitiveness. Therefore several authors have studied “repetitive-aware” self-indexes based on RLBWT [4, 12, 18, 19]. In particular, a self-index in [4] works in space proportional to the sizes of the RLBWT and LZ77 [20], another powerful compressor that can capture repetitiveness.

When it comes to constructing the RLBWT, a major concern is to reduce the working space depending on the repetitiveness of a given text. Namely, the problem is to construct the RLBWT *online in run-length compressed space*. It has been suggested in [12] that we can solve the problem using a dynamic data structure supporting rank queries on run-length encoded strings. An implementation appears very recently in [1, 17], proving its merit in space reduction. However the throughput is considerably sacrificed probably due to its use of dynamic succinct data structure. To ameliorate the time consumption, we present a novel algorithm for online RLBWT and show experimentally that our implementation runs faster with reasonable increase of memory consumption. Since Policriti and Prezza [16] recently proposed algorithms to compute LZ77 factorization in compressed space via RLBWT, online RLBWT becomes more and more important, and therefore, practical time-space tradeoffs are worth exploring.

1.2 Our Contribution

Given an input string $S = S[1]S[2] \cdots S[n]$ of length n in online manner, the algorithm described in [16] constructs the RLBWT of the reversed string $S^R = S[n] \cdots S[2]S[1]$ in $O(r \lg n)$ bits of space and $O(n \lg r)$ time, where r is the number of runs appearing in the BWT of S^R . When a new input symbol c is appended, whereas the BWT of Sc requires (in the worst case) sorting all the suffixes again, the BWT of $(Sc)^R$ requires just inserting c into the BWT of S^R , and the insert position can be efficiently computed by rank operations on the BWT of S^R . Hence a dynamic data structure on a run-length compressed string supporting rank operations allows to construct the RLBWT online. However, the algorithm of [16] internally uses rank operations on dynamic wavelet trees, which is considerably slow in practice.

In order to get a faster implementation, we replace the work carried out on dynamic wavelet trees by a comparison of integers using the dynamic maintenance of a total order. Here, the *Order-Maintenance Problem* is to maintain a total order of elements subject to *insert*(X, Y): insert a new element Y immediately after X in the total order, *delete*(X): remove X from the total order, and *order*(X, Y): determine whether $X > Y$ in the total order. Bender et al. [5] proposed a simple algorithm for this problem to allow $O(1)$ amortized insertion and deletion time and $O(1)$ worst-case query time. Adopting this technique, we develop a novel data structure for computing the insert position of c in the current BWT by a comparison of integers, instead of heavy rank operations on dynamic wavelet trees.

Compared to the baseline [16], we significantly improve the throughput of RLBWT with reasonable increase of memory consumption. Although there is a tradeoff between memory consumption and throughput performance, as shown in the experimental results, the working space of our algorithm is still sufficiently smaller than the input size, especially for highly repetitive strings.

2 Preliminaries

Let Σ be an ordered *alphabet*. An element of Σ^* is called a *string*. The length of a string S is denoted by $|S|$. The empty string ε is the string of length 0, namely, $|\varepsilon| = 0$. For a string $S = XYZ$, strings X , Y , and Z are called a *prefix*, *substring*, and *suffix* of S , respectively. For $1 \leq i \leq |S|$, the i th character of a string S is denoted by $S[i]$. For $1 \leq i \leq j \leq |S|$, let $S[i..j] = S[i] \cdots S[j]$, i.e., $S[i..j]$ is the substring of S starting at position i and ending at position j in S . For convenience, let $S[i..j] = \varepsilon$ if $j < i$.

In the *run-length encoding (RLE)* of a string S , a maximal run c^e (for some $c \in \Sigma$ and $e \in \mathcal{N}$) of a single character in S is encoded by a pair (c, e) , where we refer to c and respectively e as the *head* and *exponent* of the run. Since each run is encoded in $O(1)$ words (under Word RAM model with word size $\Omega(\lg |S|)$), we refer to the number of runs as the size of the RLE. For example, $S = \text{aaaabbcccacc} = \mathbf{a}^4\mathbf{b}^2\mathbf{c}^3\mathbf{a}^1\mathbf{c}^2$ is encoded as $(\mathbf{a}, 4), (\mathbf{b}, 2), (\mathbf{c}, 3), (\mathbf{a}, 1), (\mathbf{c}, 2)$, and the size of the RLE is five.

For any string S and any $c \in \Sigma$, let $\text{occ}_c(S)$ denote the number of occurrences of c in S . Also, let $\text{occ}_{<c}(S)$ denote the number of occurrences of any character smaller than c in S , i.e., $\text{occ}_{<c}(S) = \sum_{c' < c} \text{occ}_{c'}(S)$. For any $c \in \Sigma$ and position i ($1 \leq i \leq |S|$), $\text{rank}_c(S, i)$ denotes the number of occurrences of c in $S[1..i]$, i.e., $\text{rank}_c(S, i) = \text{occ}_c(S[1..i])$. For any $c \in \Sigma$ and i ($1 \leq i \leq \text{occ}_c(S)$), $\text{select}_c(S, i)$ denotes the position of the i th c in S , i.e., $\text{select}_c(S, i) = \min\{j \mid \text{rank}_c(S, j) = i\}$. Also we let $\text{access}(S, i)$ denote the query to ask for $S[i]$. We will consider data structures to answer $\text{occ}_{<c}$, rank , select , and access without having S explicitly.

2.1 BWT

Here we define the BWT of a string $S \in \Sigma^+$, denoted by BWT_S . For convenience, we assume that S ends with a terminator $\$ \in \Sigma$ whose lexicographic order is smaller than any character in $S[1..|S| - 1]$. BWT_S is obtained by sorting all non-empty suffixes of S lexicographically and putting the immediately preceding character of each suffix (or $\$$ if there is no preceding character) in the order.

For the online construction of BWT, it is convenient to consider “prepending” (rather than appending) a character to S because it does not change the lexicographic order among existing suffixes.¹ Namely, for some $c \in \Sigma$, we consider updating BWT_S to BWT_{cS} efficiently. The task is to replace the unique occurrence of $\$$ in BWT_S with c , and insert $\$$ into appropriate position. Since replacing can be easily done if we keep track of the current position of $\$$, the main task is to find the new position of $\$$ to insert, which can be done with a standard operation on BWT as follows: Let i be the position of $\$$ in BWT_S , then the new position is computed by $\text{rank}_c(\text{BWT}_S, i) + \text{occ}_{<c}(S) + 1$ because the new suffix cS is the $(\text{rank}_c(\text{BWT}_S, i) + 1)$ th lexicographically smallest suffix among those starting with c , and there

¹Or appending a character but constructing BWT for reversed string.

are $\text{occ}_{<c}(S)$ suffixes starting with some c' ($< c$). Thus, BWT can be constructed online using a data structure that supports rank , $\text{occ}_{<c}$, and insert queries.

Let RLBWT_S denote the run-length encoding of BWT_S . In Section 3, we study data structures that supports rank_c , $\text{occ}_{<c}$ and insert queries on run-length encoded strings, which can be directly used to construct RLBWT_S online in $O(|S| \lg r)$ time and $O(r \lg |S|)$ bits of space, where r is the size of RLE of BWT_S .

2.2 Searchable partial sums with indels

We use a data structure for the *searchable partial sums with indels (SPSI)* problem as a tool. The SPSI data structure T ought to maintain a dynamic sequence $Z[1..m]$ of non-negative integers (called *weights*) to support the following queries as well as insertion/deletion of weights:

- $\mathsf{T.sum}(k)$: Return the partial sum $\sum_{j=1}^k Z[j]$.
- $\mathsf{T.search}(i)$: For an integer i ($1 \leq i \leq \mathsf{T.sum}(m)$), return the minimum index k such that $\mathsf{T.sum}(k) \geq i$.
- $\mathsf{T.update}(k, \delta)$: For a (possibly negative) integer δ with $Z[k] + \delta \geq 0$, update $Z[k]$ to $Z[k] + \delta$.

We employ a simple implementation of T based on a B+-tree whose k th leaf corresponds to $Z[k]$.² Let B (≥ 3) be the parameter of B+-trees that represents the arity of an internal node. Namely the number of children of each internal node ranges from $B/2$ to B (unless m is too small), and thus, the height of the tree is $O(\log_B m)$. An internal node has two integer arrays LA and WA of length B such that $LA[j]$ (resp. $WA[j]$) stores the sum of #leaves (resp. weights) under the subtrees of up to the j th child of the node.

Using these arrays, we can easily answer $\mathsf{T.sum}$ and $\mathsf{T.search}$ queries in $O(\log_B m)$ time while traversing the tree from the root to a leaf: For example, $\mathsf{T.sum}(k)$ can be computed by traversing to the k th leaf (navigated by LA) while summing up the weights of the subtrees existing to the left of the traversed path by WA . It is the same for $\mathsf{T.search}(i)$ (except switching the roles of LA and WA). For $\mathsf{T.update}(k, \delta)$ query, we only have to update LA and WA of the nodes in the path from the root to the k th leaf, which takes $O(B \log_B m)$ time. Also, indels can be done in $O(B \log_B m)$ time with standard split/merge operations of B+-trees.

Naively the space usage is $O(m \lg M)$ bits, where M is the sum of all weights. Here we consider improving this to $O(m \lg(M/m))$ bits. Let us call an internal node whose children are leaves a *bottom node*, for which we introduce new arity parameter B_L , differentiated from B for the other internal nodes. For a bottom node, we discard LA , WA and the pointers to the leaves. Instead we let it store the weights of its children in a space efficient way. For example, using gamma encoding, the total space usage for the bottom nodes becomes $O(\sum_{j=1}^m \lg Z[j]) = O(m \lg(M/m))$ bits. The other (upper) part of T uses $O(m \lg M/B_L)$ bits, which can be controlled by B_L . The queries can be supported in $O(B_L + B \log_B m/B_L)$ time. Hence, setting $B = O(1)$ and $B_L = \Theta(\lg m)$, we get the next lemma.

Lemma 1 *For a dynamic sequence of weights, there is a SPSI data structure of $O(m \lg(M/m))$ bits supporting queries in $O(\lg m)$ time, where m is the current length of the sequence and M is the sum of weights.*

3 Dynamic Rank/Select Data Structures on Run-length Encoded Strings

In this section, we study dynamic rank/select data structures working on run-length encoded strings. Note that *select* and *delete* queries are not needed for online RLBWT algorithms, but we also provide them as they may find other applications. Throughout this section, we let X denote the current string with $n = |X|$, RLE size r , and containing σ distinct characters. We consider the following update queries as well as rank_c , select_c , access and $\text{occ}_{<c}$ queries on X :

- $\text{insert}(X, i, c^e)$: For a position i ($1 \leq i \leq n + 1$), $c \in \Sigma$ and $e \in \mathcal{N}$, insert c^e between $X[i - 1]$ and $X[i]$, i.e., $X \leftarrow X[1..i - 1]c^e X[i..n]$.

²More sophisticated solutions can be found in [6, 11, 15], but none of them has been implemented to the best of our knowledge.

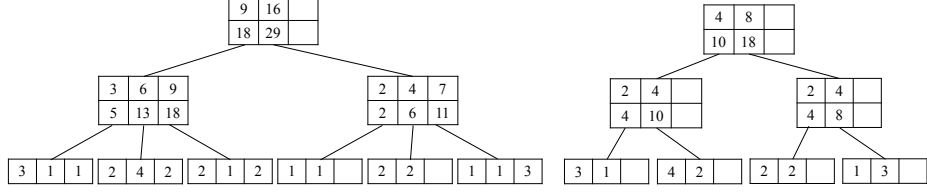


Figure 1: For $X = a^3b^1a^1c^2a^4b^2a^2c^1a^2b^1c^1a^2c^2a^1b^1a^3$, examples of T_{all} (left) and T_a (right) with $B = B_L = 3$ are shown. Note that the other components of the data structure (T_b , T_c and H) are omitted here. T_{all} holds the sequence $[3, 1, 1, 2, 4, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, 3]$ of the exponents in its leaves, and T_a holds the sequence $[3, 1, 4, 2, 2, 2, 1, 3]$ of the exponents of a 's runs in its leaves. For a node having two rows, the first row represents LA and the second WA .

- $\text{delete}(X, i, e)$: For a position i ($1 \leq i \leq n - e + 1$) such that $X[i..i + e - 1] \in c^e$ for some $c \in \Sigma$, delete $X[i..i + e - 1]$, i.e., $X \leftarrow X[1..i - 1]X[i + e..n]$.

Theorem 2 *There is a data structure that occupies $O(r \lg n)$ bits of space and supports rank_c , select_c , access , $\text{occ}_{<c}$, insert and delete in $O(\lg r)$ time.*

We will describe two data structures holding the complexities of Theorem 2 in theory but likely exhibiting different time-space tradeoffs in practice. In Subsection 3.1, we show an existing data structure. On the basis of this data structure, in Subsection 3.2, we present our new data structure to get a faster implementation.

We note that the problem to support $\text{occ}_{<c}$ in $O(\sigma \lg n)$ bits of space and $O(\lg \sigma)$ time is somewhat standard. For instance, we can think about the SPSI data structure of Lemma 1 storing $\text{occ}_c(X)$'s in increasing order of c . It is easy to modify the data structure so that, for a given c , we can traverse from the root to the leaf corresponding to the predecessor of c , where we mean by the predecessor of c the largest character c' that is smaller than c and appears in X . Then $\text{occ}_{<c}$ queries can be supported in a similar way to sum queries using WA . Thus in the following subsections, we focus on the other queries.

3.1 Existing data structure

Here we review the data structure described in [16] with implementation available in [1, 17].³ In theory it satisfies Theorem 2 though its actual implementation has the time complexity of $O(\lg \sigma \lg r)$ slower than $O(\lg r)$.

Let $(c_1, e_1), (c_2, e_2), \dots, (c_r, e_r)$ be the RLE of X . The data structure consists of three components (see also Fig. 1 for the first two):

1. T_{all} : SPSI data structure for the sequence $e_1e_2 \dots e_r$ of all exponents.
2. T_c (for every $c \in \Sigma$): SPSI data structure for the sequence of the exponents of c 's run.
3. H : Dynamic rank/select data structure for the head string $H = c_1c_2 \dots c_r$. There is a data structure (e.g., see [13, 14]), with which H can be implemented in $r \lg \sigma + o(r \lg \sigma) + O(\sigma \lg r)$ bits while supporting queries in $O(\lg r)$ time. (However, the actual implementation of [1, 17] employs a simpler algorithm based on wavelet trees that has $O(\lg \sigma \lg r)$ query time.)

Note that for every run c^e there are two copies of its exponent, one in T_{all} and the other in T_c . Since $\sigma \leq r \leq n$ holds, the data structure (excluding $\text{occ}_{<c}$ data structure) uses $r \lg \sigma + o(r \lg \sigma) + O(r \lg(n/r) + \sigma \lg r) = O(r \lg n)$ bits.

Let us demonstrate how to support $\text{rank}_c(X, i)$. Firstly by computing $k \leftarrow T_{all}.\text{search}(i)$ we can find that $X[i]$ is in the k th run. Next by computing $k_c \leftarrow H.\text{rank}_c(H, k)$ we notice that, up to the k th run, there are k_c runs with head c . Here we can check if the head of the k th run is c , and compute the number e of c 's in the k th run appearing after $X[i]$. Finally, $T_c.\text{sum}(k_c) - e$ tells the answer of $\text{rank}_c(X, i)$. It is easy to see that each step can be done in $O(\lg r)$ time.

Note that H plays an important role to bridge two trees T_{all} and T_c by converting the indexes k and k_c . The update queries also use this mechanism: We first locate the update position in T_{all} , then find

³The basic idea of the algorithm originates from the work of RLFM+ index in [12].

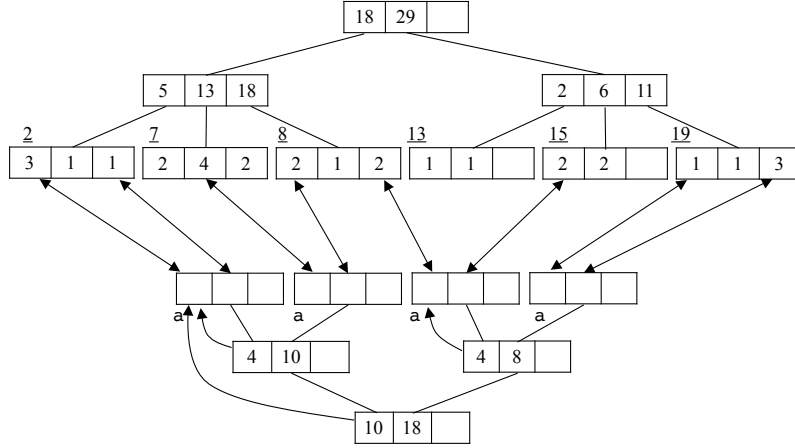


Figure 2: For $X = a^3b^1a^1c^2a^4b^2a^2c^1a^2b^1c^1a^2c^2a^1b^1a^3$ (same as the one in Fig. 1), examples of modified T_{all} (up) and T_a (down) with $B = B_L = 3$ are shown, where T_a is illustrated upside down. Note that the data structure related to T_b and T_c (e.g., pointers of Change1 to them) are omitted here. Each pair of leaves corresponding to the same run is connected by bidirectional pointers (Change1). Each internal node of T_a has pointer to its leftmost leaf (Change2). The character a is stored in each bottom node of T_a (Change3). Each bottom node of T_{all} stores a label (underlined number) that is monotonically increasing from left to right (Change4). LAs and weights in the leaves of T_a are discarded (Change5).

the update position in T_c by bridging two trees with H . After locating the positions, the updates can be done in each dynamic data structure. $\text{select}_c(X, i)$ can be answered by first locating i th c in T_c , finding the corresponding position in T_{all} with $H.\text{select}_c$, then computing the partial sum up to the position in T_{all} . Finally, $\text{access}(X, i)$ is answered by $H.\text{access}(H, T_{all}.\text{search}(i))$.

3.2 New data structure

Now we present our new data structure satisfying Theorem 2. We share some of the basic concepts with the data structure described in Section 3.1. For example, our data structure also uses the idea of answering queries by going back and forth between T_{all} and T_c . However we do not use H to bridge the two trees. Succinct data structures (like H) are attractive if the space limitation is critical, but otherwise the suffered slow-down might be intolerable. Therefore we design a fast algorithm that bridges the two trees in a more direct manner, while taking care not to blow up the space too much.

In order to do without H , we make some changes to T_{all} and T_c (see also Fig. 2):

1. We maintain bidirectional pointers connecting every pair of leaves representing the same run in T_{all} and T_c (recall that every run with head c has exactly one corresponding leaf in each of T_{all} and T_c).
2. For every internal node of T_c , we store a pointer to the leftmost leaf in the subtree rooted at the node.
3. For every bottom node of T_c , we store the character c .
4. For every bottom node of T_{all} , we store a label (a positive integer) such that the labels of bottom nodes are monotonically increasing from left to right. Since bottom nodes are inserted/deleted dynamically, we utilize the algorithm [5] for the order-maintenance problem to maintain the labels.
5. Minor implementation notes: Every LA can be discarded as our data structure does not use the navigation of indexes. Also, we can quit storing the leaf-level weights in T_c as it can be retrieved using the pointer to the corresponding leaf in T_{all} .

3.2.1 Space analysis.

In the change list, Changes1-4 increase the space usage while Change5 reduces. It is easy to see that the increase fits in $O(r \lg n)$ bits. More precisely, since Changes2-4 are made to internal nodes, the increase by these changes is $O(r \lg n / B_L)$ bits, which is more or less controllable by B_L (recall that B_L is

arity parameter for bottom nodes, and we have $O(r \lg n / B_L) = O(r \lg(n/r))$ by setting $B_L = \Theta(\lg r)$ for Lemma 1). On the other hand, Change1 is made to leaves and takes $2r \lg r$ bits of space. Thus, the total space usage of the data structure (excluding $\text{occ}_{<c}$ data structure) is $2r \lg r + O(r \lg(n/r)) = O(r \lg n)$ bits.

By this analysis, it is expected that $2r \lg r$ becomes a leading term when the ratio n/r is small, i.e., compressibility in terms of RLE is not high. It should be compared to $r \lg \sigma + o(r \lg \sigma) + O(r \lg(n/r) + \sigma \lg r)$ bits used by the data structure of Section 3.1, in which $r \lg r$ term does not exist. Hence, the smaller the ratio n/r , the larger the gap between the two data structures in space usage will be. On the other hand, when the $r \lg(n/r)$ term is leading, i.e., r is sufficiently smaller than n , the increase by the $r \lg r$ term would be relatively moderate.

3.2.2 Answering queries.

We show how to answer queries on our data structure. All queries are supported in $O(\lg r)$ time.

access(X, i): We first traverse from the root of T_{all} to the run containing $X[i]$ (navigated by *WA*), jump to the corresponding leaf of T_c by pointer of Change1, then read the character stored in the bottom node of T_c due to Change3.

select _{c} (X, i): We first traverse from the root of T_c to the run containing i th c (navigated by *WA*). At the same time, we can compute the rank i' of i th c within the run. Next we jump to the corresponding leaf in T_{all} by pointer of Change1, then compute the sum of characters appearing strictly before the leaf while going up the tree. The answer to **select** _{c} (X, i) is the sum plus i' .

rank _{c} (X, i): Recalling the essence of the algorithm described in Section 3.1, we can answer **rank** _{c} (X, i) if we locate the leaf of T_c representing the rightmost c 's run that starts at or before position i . In order to locate such leaf v , we first traverse from the root of T_{all} to the run containing $X[i]$ (navigated by *WA*). If we are lucky, we may find a c 's run in the bottom node containing $X[i]$, in which case we can easily get v or the successor of v by using the pointer of Change1 outgoing from the c 's run. Otherwise, we search for v traversing T_c from the root navigated by labels of Change4. Let t be the label of the bottom node containing $X[i]$. Then, it holds that v is the rightmost leaf pointing to a node of T_{all} with label smaller than t . Since the order of labels is maintained, we can use t as a key for binary search, i.e., we notice that an internal node u (and its succeeding siblings) cannot contain v if the leftmost leaf in the subtree rooted at u points to a node of T_{all} with label greater than t . Using the pointer of Change2 to jump to the leftmost leaf, we can conduct each comparison in $O(1)$ time, and thus, we can find v in $O(\lg r)$ time.

Update queries: The main task is to locate the update positions both in T_{all} and T_c , and this is exactly what we did in **rank** _{c} query—locating the run containing $X[i]$ and v . After locating the update positions, the update can be done in $O(\lg r)$ time in each tree. When the update operation invokes insertion/deletion of a bottom node of T_{all} , we maintain labels of Change4 using the algorithm of [5]. We note that the algorithm of [5] takes $O(\lg r)$ amortized time per “indel of bottom node”, and hence, takes $O(1)$ amortized time per “indel of leaf” (recall that $B_L = \Theta(\lg r)$, and one indel of bottom node needs $\Theta(\lg r)$ indels of leaves). In addition, the algorithm is quite simple and efficiently implementable without any data structure than labels themselves.

4 Experiments

We implemented in C++ the online RLBWT construction algorithm based on our new rank/select data structure described in Section 3.2 (the source code is available at [2]). We evaluate the performance of our method comparing with the state-of-the-art implementation [1] (we refer to it as PP taking the authors' initials of [16]) of the algorithm based on the data structure described in Section 3.1. We tested on highly repetitive datasets in *repcorpus*⁴, well-known corpus in this field, and some larger datasets created from git repositories. For the latter, we use the script [3] to create 1024MB texts (obtained by concatenating source files from the latest revisions of a given repository, and truncated to be 1024MB) from the repositories for *boost*⁵, *samtools*⁶ and *sdsl-lite*⁷ (all accessed at 2017-03-27). The programs were compiled using g++6.3.0 with `-Ofast -march=native` option. The experiments were conducted on a 6core Xeon E5-1650V3 (3.5GHz) machine with 32GB memory running Linux CentOS7.

⁴See <http://pizzachili.dcc.uchile.cl/repcorpus/statistics.pdf> for statistics of the datasets.

⁵<https://github.com/boostorg/boost>

⁶<https://github.com/samtools/samtools>

⁷<https://github.com/simongog/sdsl-lite>

Table 1: Computation time in seconds and working space in mega bytes to construct the RLBWT of r runs from each dataset of size $|S|$ using the proposed method (ours) and the previous method (PP).

dataset	$ S $ (MB)	r	computation time (sec)		working space (MB)	
			ours	PP	ours	PP
fib41	255.503	42	27	552	0.004	0.067
rs.13	206.706	76	16	623	0.005	0.068
tm29	256.000	82	24	802	0.005	0.068
dblp.xml.00001.1	100.000	172,195	37	2,060	2.428	1.307
dblp.xml.00001.2	100.000	175,278	37	2,070	2.446	1.322
dblp.xml.0001.1	100.000	240,376	40	2,100	4.381	1.586
dblp.xml.0001.2	100.000	269,690	40	2,105	4.565	1.730
dna.001.1	100.000	1,717,162	58	1,667	35.966	5.729
english.001.2	100.000	1,436,696	58	2,153	20.680	6.166
proteins.001.1	100.000	1,278,264	58	1,839	19.790	5.133
sources.001.2	100.000	1,211,104	49	2,141	19.673	5.721
cere	439.917	11,575,582	534	7,597	186.073	43.341
coreutils	195.772	4,732,794	128	4,479	81.642	22.301
einstein.de.txt	88.461	99,833	30	1,807	2.083	1.106
einstein.en.txt	445.963	286,697	182	9,293	4.836	2.296
Escherichia.Coli	107.469	15,045,277	154	2,047	316.184	36.655
influenza	147.637	3,018,824	91	2,501	72.730	12.386
kernel	246.011	2,780,095	146	5,333	41.758	12.510
para	409.380	15,635,177	547	7,364	329.901	52.005
world_leaders	44.792	583,396	17	857	9.335	2.891
boost	1024.000	63,710	320	20,327	1.161	0.904
samtools	1024.000	562,326	440	21,375	9.734	3.595
sdsl	1024.000	758,657	419	21,014	17.760	4.803

Table 1 shows the comparison of the two methods on construction time and working space. The result shows that our method significantly improves the construction time of PP as we intended. Especially for dumpfiles of Wikipedia articles (einstein.de.txt and einstein.en.txt), our method ran 60 times faster than PP. Our method also shows good performance for the 1024MB texts from git repositories. On the other hand, the working space is increased (except the artificial datasets, which are extremely compressible) by 1.3 to 8.7 times. Especially for less compressible datasets in terms of RLBWT like Escherichia.Coli, the space usage tends to be worse as predicted by space analysis in Section 3.2. Still for most of the other datasets the working space of our method keeps way below the input size.

5 Conclusion

We have proposed an improvement of online construction of RLBWT [1, 17], intended to speed up the construction time. We significantly improved the throughput of original RLBWT with reasonable increase of memory consumption for the benchmarks from various domain. By applying our new algorithm to the algorithm of computing LZ77 factorization in compressed space using RLBWT [16], we would immediately improve the throughput of [16]. As LZ77 plays a central role in many problems on string processing, engineering/optimizing implementation for compressed LZ77 computation is important future work.

6 Acknowledgments

This work was supported by JST CREST (Grant Number JPMJCR1402), and KAKENHI (Grant Numbers 17H01791 and 16K16009).

References

- [1] DYNAMIC: dynamic succinct/compressed data structures library. <https://github.com/xxsds/DYNAMIC>.
- [2] Online RLBWT. <https://github.com/itomomoti/OnlineRLBWT>.
- [3] get-git-revisions: Get all revisions of a git repository. <https://github.com/nicolaprezza/get-git-revisions>.
- [4] Djamel Belazzougui, Fabio Cunial, Travis Gagie, Nicola Prezza, and Mathieu Raffinot. Composite repetition-aware data structures. In *CPM*, pages 26–39, 2015.
- [5] Michael A. Bender, Richard Cole, Erik D. Demaine, Martin Farach-Colton, and Jack Zito. Two simplified algorithms for maintaining order in a list. In *ESA*, pages 152–164, 2002.
- [6] Philip Bille, Patrick Hagge Cording, Inge Li Gørtz, Frederik Rye Skjoldjensen, Hjalte Wedel Vildhøj, and Søren Vind. Dynamic relative compression, dynamic partial sums, and substring concatenation. In *ISAAC*, pages 18:1–18:13, 2016.
- [7] Alexander Bowe, Taku Onodera, Kunihiro Sadakane, and Tetsuo Shibuya. Succinct de bruijn graphs. In *WABI*, pages 225–235, 2012.
- [8] Michael Burrows and David J Wheeler. A block-sorting lossless data compression algorithm. Technical report, HP Labs, 1994.
- [9] Paolo Ferragina, Fabrizio Luccio, Giovanni Manzini, and S. Muthukrishnan. Structuring labeled trees for optimal succinctness, and beyond. In *FOCS*, pages 184–196, 2005.
- [10] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *FOCS*, pages 390–398, 2000.
- [11] Wing-Kai Hon, Kunihiro Sadakane, and Wing-Kin Sung. Succinct data structures for searchable partial sums with optimal worst-case performance. *Theor. Comput. Sci.*, 412(39):5176–5186, 2011.
- [12] Veli Mäkinen, Gonzalo Navarro, Jouni Sirén, and Niko Välimäki. Storage and retrieval of highly repetitive sequence collections. *J. Computational Biology*, 17(3):281–308, 2010.
- [13] J. Ian Munro and Yakov Nekrich. Compressed data structures for dynamic sequences. In *ESA*, pages 891–902, 2015.
- [14] Gonzalo Navarro and Yakov Nekrich. Optimal dynamic sequence representations. *SIAM J. Comput.*, 43(5):1781–1806, 2014.
- [15] Gonzalo Navarro and Kunihiro Sadakane. Fully functional static and dynamic succinct trees. *ACM Transactions on Algorithms*, 10(3):16, 2014.
- [16] Alberto Policriti and Nicola Prezza. Computing LZ77 in run-compressed space. In *DCC*, pages 23–32, 2016.
- [17] Nicola Prezza. A framework of dynamic data structures for string processing. In *SEA*, 2017. to appear.
- [18] Jouni Sirén. *Compressed Full-Text Indexes for Highly Repetitive Collections*. PhD thesis, University of Helsinki, 2012.
- [19] Jouni Sirén, Niko Välimäki, Veli Mäkinen, and Gonzalo Navarro. Run-length compressed indexes are superior for highly repetitive sequence collections. In *SPIRE*, pages 164–175, 2008.
- [20] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3):337–349, 1977.