

Grammar Compression by Induced Suffix Sorting*

DANIEL S. N. NUNES, Federal Institute of Brasília, and Department of Computer Science, University of Brasília, Brazil

FELIPE A. LOUZA, Faculty of Electrical Engineering, Federal University of Uberlândia, Brazil

SIMON GOG, eBay Inc., USA

MAURICIO AYALA-RINCÓN, Departments of Computer Science and Mathematics, University of Brasília, Brazil

GONZALO NAVARRO, Center for Biotechnology and Bioengineering (CeBiB) and Department of Computer Science, University of Chile, Chile

A grammar compression algorithm, called GCIS, is introduced in this work. GCIS is based on the induced suffix sorting algorithm SAIS, presented by Nong *et al.* in 2009. The proposed solution builds on the factorization performed by SAIS during suffix sorting. A context-free grammar is used to replace factors by non-terminals. The algorithm is then recursively applied on the shorter sequence of non-terminals. The resulting grammar is encoded by exploiting some redundancies, such as common prefixes between right-hands of rules, sorted according to SAIS. GCIS excels for its low space and time required for compression while obtaining competitive compression ratios. Our experiments on regular and repetitive, moderate and very large texts, show that GCIS stands as a very convenient choice compared to well-known compressors such as Gzip 7-Zip; and RePair the gold standard in grammar compression; and recent compressors such as SOLCA, LZRR and LZD. In exchange, GCIS is slow at decompressing. Yet, grammar compressors are more convenient than Lempel-Ziv compressors in that one can access text substrings directly in compressed form, without ever decompressing the text. We demonstrate that GCIS is an excellent candidate for this scenario because it shows to be competitive among its RePair based alternatives. We also show that the relation with SAIS makes GCIS a good intermediate structure to build the suffix array and the LCP array during decompression of the text.

CCS Concepts: • **Theory of computation** → **Data compression; Data structures and algorithms for data management; Grammars and context-free languages;**

Additional Key Words and Phrases: data compression, suffix sorting, extract, suffix-array, LCP-array.

ACM Reference Format:

Daniel S. N. Nunes, Felipe A. Louza, Simon Gog, Mauricio Ayala-Rincón, and Gonzalo Navarro. 2019. Grammar Compression by Induced Suffix Sorting. *ACM J. Exp. Algor.* 0, 0, Article 0 (2019), 35 pages. <https://doi.org/0000001.0000001>

*A preliminary version of this work appeared in DCC 2018 [Nunes *et al.* 2018].

Authors' addresses: Daniel S. N. Nunes, Federal Institute of Brasília, Taguatinga, Department of Computer Science, University of Brasília, Brasília, Brazil, daniel.nunes@ifb.edu.br; Felipe A. Louza, Faculty of Electrical Engineering, Federal University of Uberlândia, Uberlândia, Brazil, louza@ufu.br; Simon Gog, eBay Inc. San Jose, USA, sgog@ebay.com; Mauricio Ayala-Rincón, Departments of Computer Science and Mathematics, University of Brasília, Brasília, Brazil, ayala@unb.br; Gonzalo Navarro, Center for Biotechnology and Bioengineering (CeBiB), Department of Computer Science, University of Chile, Santiago, Chile, gnavarro@dcc.uchile.cl.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.

Manuscript submitted to ACM

1 INTRODUCTION

Text compression is a fundamental task in Computer Science that consists in transforming an input string into another string whose bit sequence representation is smaller. The suffix array [Gonnet et al. 1992; Manber and Myers 1993] is a key data structure used to compute lossless compression transforms [Goto and Bannai 2014; Kärkkäinen et al. 2013; Ohlebusch and Gog 2011], such as the Burrows-Wheeler transform (BWT) [Burrows and Wheeler 1994] and the Lempel-Ziv factorization (LZ77) [Liu et al. 2016; Ziv and Lempel 1977], which are at the heart of the popular data compression tools like BZIP2, 7-ZIP and GZIP.

The suffix array construction (or suffix sorting) may be performed in linear time (see Puglisi et al. [2007] and Dhaliwal et al. [2012] for good reviews). Nong et al. [2009] introduced a remarkable suffix array construction algorithm called SAIS, which runs in linear time and is fast in practice. Subsequently, SAIS was adapted to directly compute the BWT [Okanojara and Sadakane 2009], the Φ -array [Goto and Bannai 2014; Kärkkäinen et al. 2009], the LCP array [Fischer 2011], and the suffix array for string collections [Louza et al. 2017a].

Grammar compression [Kieffer and Yang 2000] is a compression technique based on finding a small context-free grammar that generates (only) the text. Finding the smallest such grammar is NP-hard [Charikar et al. 2005], but heuristics like REPAIR [Larsson and Moffat 1999] work very well in practice. Like Lempel-Ziv, grammar compression performs particularly well on repetitive text collections. An advantage of grammar compression is that text substrings can be extracted from the compressed representation without the need to decompress the text [Bille et al. 2015]. One can then aim at never decompressing the text but work on it directly in compressed form.

This article introduces GCIS, a new grammar-based compression algorithm that builds on SAIS framework. GCIS constructs a context-free grammar recursively, based on the string factorization performed by SAIS. The rules are encoded exploiting the common prefixes between the right-hands of consecutive rules, which are sorted lexicographically by SAIS. GCIS is the first grammar-compressor algorithm, as far as we know, based on induced suffix sorting.

Our experiments show that, on repetitive texts and compared to the best grammar-compressor (REPAIR [Larsson and Moffat 1999]) and the best Lempel-Ziv compressor (7-ZIP [Pavlov 2017]), GCIS is an interesting alternative in practice because it displays the fastest compression speed and low memory usage, while reaching a compression ratio close to that of REPAIR. We also show that GCIS is competitive with the relevant alternatives when processing regular (not highly repetitive) and very large texts.

In exchange, GCIS is slower than most alternatives for decompressing the text. As explained, however, we may aim at extracting any desired substring while never decompressing the whole text. When compared to variants of REPAIR that allow extraction, GCIS turns out to be very efficient while using less space. Further, it is possible to build the suffix and LCP arrays as a byproduct of text decompression by GCIS in competitive time.

This work differs from its early version [Nunes et al. 2018] in that we support efficient extraction of substrings and computation of suffix and LCP arrays directly from decompression. We also present more thorough descriptions and experimental results.

2 BACKGROUND

2.1 Suffix array

Let T be a string of length $|T| = n$, over an ordered alphabet Σ . We assume that our alphabet Σ has an integer size, but limited to n , that is, $1 \leq |\Sigma| \leq n$. The concatenation of strings or symbols is denoted by the dot operator (\cdot). The symbol $<$ is used for the lexicographic order relation between strings.

Let $T[i]$ be the i -th symbol of T . The substring (factor) from $T[i]$ to $T[j]$, both included, is denoted by $T[i, j]$, for $1 \leq i \leq j \leq n$. A prefix of T is a substring of the form $T[1, i]$ and a suffix is a substring of the form $T[i, n]$, also denoted by T_i . For convenience, it is assumed that T always ends with a special symbol $T[n] = \$$, which is not present elsewhere in T and lexicographically precedes every symbol in $T[1, n-1]$.

The suffix array (SA) [Gonnet et al. 1992; Manber and Myers 1993] of a string $T[1, n]$ is an array of integers in the range $[1, n]$ that gives the lexicographic order of all suffixes of T , such that $T_{SA[1]} < T_{SA[2]} < \dots < T_{SA[n]}$. The length of the longest common prefix (LCP) of two strings X and Y in Σ^* is denoted $\text{lcp}(X, Y)$. The LCP array of $T[1, n]$ is an array of integers that stores the lcp value between consecutive suffixes in SA, such that $\text{LCP}[i] = \text{lcp}(T_{SA[i-1]}, T_{SA[i]})$, for $1 < i \leq n$, and we define $\text{LCP}[1] = 0$. The suffixes starting with the same symbol $c \in \Sigma$ form a c -bucket in SA. The head and the tail of a c -bucket refer to the first and the last position of the c -bucket in SA.

2.2 Grammar compression

Let $G = (\Sigma, \Gamma, P, X_S)$ be a reduced context-free grammar (i.e, with no unreachable non-terminals) such that Σ is the terminal alphabet of G ; Γ is the set of non-terminal symbols (disjoint from Σ); $P \subseteq \Gamma \times (\Sigma \cup \Gamma)^*$ is the set of production rules; and $X_S \in \Gamma$ is the start symbol.

A production rule (X_i, α_i) is also denoted $X_i \rightarrow \alpha_i$. In this case, it is said that α_i is derived from X_i . For strings $S, R \in (\Sigma \cup \Gamma)^*$, if R is obtained from S by production rules in P , then R is derived from S . When R is obtained by a (possibly empty) sequence of derivations from S , then R is generated from S .

Given a string T , the grammar compression problem is to find a grammar G that generates only T , such that G can be represented in less space than the original T . Given that G grammar-compresses T , for $(X_i, \alpha_i) \in P$, $\mathcal{G}(X_i) = S$ is defined as the single string $S \in \Sigma^*$ that is generated from X_i . The language generated by G contains the unique string $\mathcal{G}(X_S) = T$. This notion can be extended further for a string of terminals and non-terminals S , such that:

$$\mathcal{G}(S) = \{W_1 \cdot W_2 \cdot \dots \cdot W_{|S|} \mid W_k \in \mathcal{G}(S[k]), 1 \leq k \leq |S|\}.$$

When each $S[i]$, $1 \leq i \leq |S|$, generates a single sequence, the previous definition can be replaced by the concatenation of the strings generated by $S[i]$, $1 \leq i \leq |S|$:

$$\mathcal{G}(S) = \mathcal{G}(S[1]) \cdot \mathcal{G}(S[2]) \cdot \dots \cdot \mathcal{G}(S[|S|])$$

2.3 Integer Encoding

We cover various techniques to encode sequences of integers, when most of them are expected to be small. Some of the techniques allow us to directly access any integer in the sequence.

Simple8b. The Simple8b scheme, proposed by Anh and Moffat [2010], encodes a sequence of small integers in a 64-bit word using the number of bits required by the largest integer. Basically, it identifies a word with a 4-bit tag called *selector*, which specifies the number of integers encoded in the rest of the word and the width of such integers. Simple8b also has specific selectors for a run consisting of zeroes. If a run of 240 or 120 zeros is encountered, it can be represented with a single 64-bit word.

Directly Addressable Codes. The Directly Addressable Codes (DAC) proposed by Brisaboa et al. [2013] allow efficient retrieval of any given value $A[i]$ from an array of integers $A[1, n]$ while encoding such integers compactly. Let l_i be

the length (number of bits) of $A[i]$, then this encoding splits each $A[i]$ into $\lceil l_i/b \rceil$ blocks $v_{i,1}, v_{i,2}, \dots, v_{i,k}$ of b bits each. A bit $b_{i,j} = 1$ is associated with a block $v_{i,j}$ if $j < k$, that is, $v_{i,j}$ is not the last block of $A[i]$. Otherwise, $b_{i,j} = 0$. Then a layered data structure is constructed in such a way that the k -th layer contains two bitmaps: the first bitmap is the concatenation of blocks $v_{i,k}$, for every $1 \leq i \leq n$, whereas the second bitmap is the concatenation of the bits $b_{i,k}$ associated with each block $v_{i,k}$.

To retrieve any given value $A[i]$, one must first recover $b_{i,1}$ and check if its value is zero. If so $A[i]$ equals $v_{i,1}$, otherwise, it is necessary to proceed recursively to the j -th entry of the next layer, where $j = \sum_{m=1}^i b_{m,1}$, and append the result of the recursive call to $v_{i,1}$. The prefix sum value j can be computed in constant time by using auxiliary data structures on top of the bitmaps.

Elias-Fano Encoding. This format permits the encoding of a monotonically increasing sequence of n integers over the interval $[0, m - 1]$ within $2n + n \lceil \lg \frac{m}{n} \rceil$ bits and allows the retrieval of any integer of such sequence in constant time [Vigna 2013]. Each integer a_i is divided into two parts: u_i , the $\lceil \lg n \rceil$ most significant bits of a_i and l_i , the $\lceil \lg \frac{m}{n} \rceil$ remaining bits of a_i . The l_i values are concatenated in a single array of $n \lceil \lg \frac{m}{n} \rceil$ bits and each value a_i is classified in one of the total of n possible buckets. Then, the number of elements of each bucket is represented in a negated unary representation and such representations are concatenated in a bitmap B of $2n$ bits, n bits for each possible bucket and further n bits for every element a_i .

To retrieve the i -th value of the sequence of integers, one just needs to search for the position k of the i -th one bit on B , and append l_i to the binary representation of $k - i$. The position k can be retrieved in constant time using auxiliary data structures on top of B .

3 SAIS: INDUCED SUFFIX SORTING

SAIS [Nong et al. 2009] builds on the induced suffix sorting technique introduced by previous algorithms [Itoh and Tanaka 1999; Ko and Aluru 2003]. Induced suffix sorting consists in deducing the order of unsorted suffixes from a (smaller) set of already ordered suffixes.

The next definition classifies suffixes and symbols of strings.

DEFINITION 1 (L-TYPE AND S-TYPE). *For any string T , $T_n = \$$ has type S . A suffix T_i is an S -suffix if $T_i < T_{i+1}$, otherwise T_i is an L -suffix. Each symbol $T[i]$ has the type of T_i .*

The suffixes can be classified in linear time by scanning T once from right to left, so that the type of each suffix is stored in a bitmap of size n .

Note that, within a c -bucket, the L -suffixes precede the S -suffixes.

Further, the classification of suffixes is refined as follows:

DEFINITION 2 (LMS-TYPE). *Let T be a string. Then T_i is an LMS -suffix if T_i is an S -suffix and T_{i-1} is an L -suffix.*

Nong et al. [2009] showed that the order of the LMS -suffixes is enough to induce the order of all suffixes. This is the basis of the SAIS algorithm.

3.1 SAIS framework

(1) Sort the LMS -suffixes. This step is explained later.

- (2) Insert the LMS-suffixes into the tail of their respective c -buckets in $SA[1, n]$, without changing their order. Now SA contains the LMS-suffixes positions, in sorted order, on the end of each c -bucket. The remaining values of SA are initialized with a sentinel \perp value.
- (3) Induce L-suffixes by scanning $SA[1, n]$ from left to right: for each suffix $SA[i] \neq \perp$, if $T[SA[i] - 1]$ is L-type, insert $SA[i] - 1$ into the head of its c -bucket.
- (4) Induce S-suffixes by scanning $SA[1, n]$ from right to left: for each suffix $SA[i] \neq \perp$, if $T[SA[i] - 1]$ is S-type, insert $SA[i] - 1$ into the tail of its c -bucket.

Whenever a value is inserted in the head (or tail) of a c -bucket, the pointer to the head (or tail) is increased (or decreased) by one.

In order to sort the LMS-suffixes in Step 1, $T[1, n]$ is divided (factorized) into LMS-substrings.

DEFINITION 3. $T[i, j]$ is an LMS-substring if both T_i and T_j are LMS-suffixes, but no suffix between i and j has LMS-type. The last suffix T_n is an LMS-substring.

Let $r_1^1, r_2^1, \dots, r_n^1$ be the n^1 LMS-substrings of T read from left to right. A modified version of SAIS is applied to sort the LMS-substrings. Starting from Step 2, $T[1, n]$ is scanned (right-to-left) and each new LMS-suffix starting with c is inserted (bucket-sorted) at the tail of its c -bucket. Steps 3 and 4 work exactly the same. At the end, the LMS-substrings are sorted and the beginning positions of each LMS-substring are stored in their corresponding c -buckets in SA .

3.2 Naming

A name v_i^1 is assigned to each LMS-substring r_i^1 according to its lexicographical rank in $[1, \sigma^1]$, such that $v_i^1 \leq v_j^1$ iff $r_i^1 \leq r_j^1$, where σ^1 is the number of different LMS-substrings in T . In order to compute the names, each consecutive LMS-substrings in SA , say r_i^1 and r_{i+1}^1 , are compared to determine if either $r_i^1 = r_{i+1}^1$ or $r_i^1 < r_{i+1}^1$. In the former case v_{i+1}^1 is set to v_i^1 , whereas in the latter case v_{i+1}^1 is set to $v_i^1 + 1$. This procedure may be sped up by comparing the LMS-substrings first by symbol and then by type, with L-type symbols being smaller than S-type symbols in case of ties [Nong et al. 2011].

3.3 Recursive call

A new (reduced) string $T^1 = v_1^1 \cdot v_1^1 \cdot \dots \cdot v_n^1$ is created, whose length n^1 is at most $n/2$, and the alphabet size σ^1 is integer. If every $v_i^1 \neq v_j^1$ then all LMS-suffixes are already sorted. Otherwise, SAIS is recursively applied to sort all the suffixes of T^1 . Nong et al. [Nong et al. 2009] showed that the relative order of the LMS-suffixes in T and the order of the respective suffixes in T^1 are the same. Therefore, the order of all LMS-suffixes can be determined by the result of the recursive algorithm.

4 GCIS: GRAMMAR COMPRESSION BY INDUCED SUFFIX SORTING

This section introduces the Grammar Compression algorithm by Induced Sorting (GCIS).

4.1 Compressing

First, a context-free grammar $G = (\Sigma, \Gamma, P, X_S)$ that generates only $T[1, n]$ is computed. To do this SAIS is modified as follows.

Consider the j -th recursion level. In Step 1, after the input string $T^{j-1}[1, n^{j-1}]$ is divided into the LMS-substrings $r_1^j, r_2^j, \dots, r_{n^{j+1}}^j$ and named $v_1^j, v_2^j, \dots, v_{n^{j+1}}^j$, a new rule $v_i^j \rightarrow r_i^j$ is created for each different LMS-substring r_i^j . Moreover,

an additional rule $v_0^j \rightarrow v_0^{j-1}T^j[1, j_1 - 1]$ if $j > 1$ or $v_0^j \rightarrow T[1, j_1 - 1]$ if $j = 1$, with j_1 standing for the index of the leftmost LMS-type suffix of T^j , is created for the prefix of T^j that is not included in any LMS-substring. In this context, $n^0 = n$ and $T^0 = T$.

The algorithm is then called recursively with the reduced string $T^{j+1} = v_1^j \cdot v_2^j \cdots v_{n^{j+1}}^j$ as input as long as $\sigma^j < n^j$, that is, the LMS-substrings are not pairwise distinct. At the end, when $\sigma^\ell = n^\ell$, the last recursion level $j = \ell$ is reached, and the start symbol of X_S of G is created so that the initial production $X_S \rightarrow v_0^\ell \cdot v_1^\ell \cdot v_2^\ell \cdots v_{n^\ell}^\ell$ generates the original string $T[1, n]$.

The algorithm stops after computing X_S , since we are not interested in constructing the suffix array; Steps 2, 3 and 4 of SAIS are not executed. The recursive calls return to the top level and a grammar G that generates only $T[1, n]$ has been computed.

Since for each LMS-substring there is a unique v_i^j , there are no cycles in any generation. Further, there is only one path of derivations that from a string S generates a string S' . The consequence of this deterministic choice, for every derivation, is that $\mathcal{G}(X_i)$, for $X_i \in \Gamma$, is a fixed string of terminals. Figure 1 shows the grammar construction on GCIS.

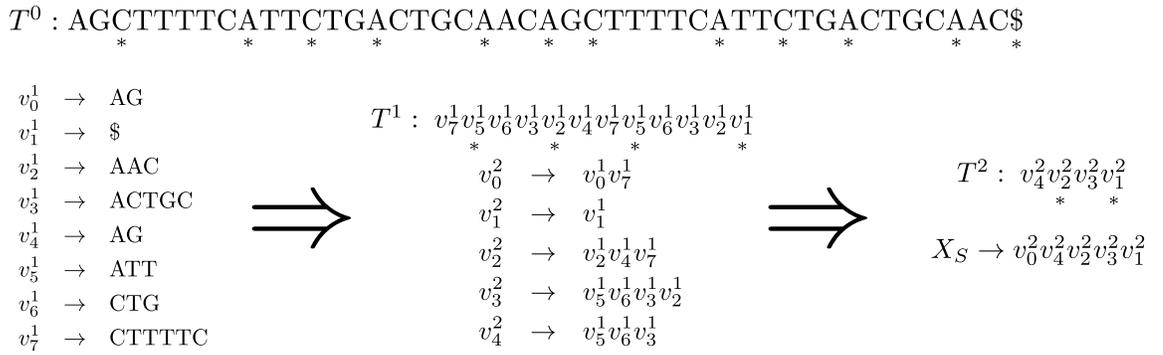


Fig. 1. Grammar construction during GCIS. All LMS-substrings (those starting with a ‘*’ symbol), are sorted according to SAIS framework, and then rules $v_0^1 \rightarrow \text{AG}$, $v_1^1 \rightarrow \$$, \dots , $v_7^1 \rightarrow \text{CTTTTC}$ are created. Next, T^1 is obtained by replacing every LMS-substring by the left-hand side of its rule. The procedure is applied recursively to T^1 . When T^2 is created, the alphabet size is equal to $|T^2| = n^2$, and thus the starting rule X_S that generates T^0 is obtained.

Grammar compression. Consecutive entries in the set of productions P are likely to share a common prefix, since the LMS-substrings are given lexicographically ordered by SAIS. Therefore, each rule $X_i \rightarrow \alpha_i \in P$ is encoded using two values $(l_i, s(\alpha_i))$, such that l_i encodes $\text{lcp}(\alpha_{i-1}, \alpha_i)$, and the remaining symbols of α_i are given by $s(\alpha_i) = \alpha_i[l_i + 1, |\alpha_i|]$. For each starting rule v_0^j , we define $l_i = 0$. This technique is known as Front-coding [Witten *et al.* 1999].

The computation of $(l_i, s(\alpha_i))$ is performed with no additional cost with a slight modification in the naming procedure of SAIS. Each consecutive LMS-substring in SA, say r_{i-1}^j and r_i^j , are compared first by symbol until a mismatch is found, and then compared by type, to check if either $r_{i-1}^j = r_i^j$ or $r_{i-1}^j < r_i^j$. The symbol-wise comparison reveals $\text{lcp}(r_{i-1}^j, r_i^j)$ as well, so the resulting complexity is the same with a small slowdown in the running time.

Time complexity. GCIS runs in $O(n)$ time, since each step of the modified SAIS runs in linear time and the length of the reduced string T^j is at most $|T^{j-1}|/2$.

Grammar size. The number of rules for a given string T^{k-1} is dependent of the number of distinct LMS-Substrings (σ^k) plus one: the additional rule for the prefix that does not belong to any LMS-substring of T^{k-1} . Let ℓ be the number of grammar levels, r_s^ℓ be the right-hand side of the starting symbol X_S and r_0^k the prefix of T^{k-1} that does not belong to any LMS-Substring, hence the grammar size can be computed as:

$$|r_s^\ell| + \sum_{i=1}^{\ell} \sum_{j=0}^{\sigma^i} |r_j^i|$$

In the worst case the grammar size might be $\Theta(n)$ for easily compressible texts such as $T = a^n\$$, but empirical results demonstrate that GCIS achieves low compression ratios regarding real data (cf. Section 5).

Implementation details. Each non-terminal X_i is represented by a pair $\alpha_i = (l_i, s(\alpha_i))$, as explained. The l_i values tend to be small and, considering the j -th recursion value, the sum of such values cannot be greater than n^j , since no two LMS-substrings overlap by more than one symbol.

One can encode all l_i values by using the Simple8b encoding in an integer array W . All strings $s(\alpha_i)$ are encoded in a single fixed-width integer array Y , of cell width $\lfloor \lg(\sigma^j) \rfloor + 1$ bits. The length of each $s(\alpha_i)$ is also encoded using Simple8b into a word array Z . The same observation of the lcp sum can be done here: the sum of all $|s(\alpha_i)|$ on the j -th recursion level is no larger than n^j .

4.2 Decompressing

The decoding process is done level-wise, starting from the last recursion level $j = \ell$, by decoding the right side of each rule. At the end, T is decoded from T^1 .

In the j -th recursion level, the values (w, y, z) from W, Y and Z , the data structures mentioned in the implementation details of Section 4.1, are decoded sequentially. In order to obtain the right-hand side of the production rules α_{k+1} from α_k , the first w symbols of α_k are copied to α_{k+1} and the z symbols from Y , which correspond to the string y , are appended to α_{k+1} . After this process the plain representation of each rule is stored, in a single array of cells with fixed width $\lfloor \lg(\sigma^j) \rfloor + 1$ bits. An additional array of pointers D is also created to find the starting position of each rule in this fixed-width array.

With the fixed-width array and the array of pointers D , T^{j-1} now can be decoded from T^j . First, the right side of v_0^j is copied into T^{j-1} . Then, T^j is scanned in a left-to-right fashion and for each $T^j[i]$ the algorithm appends to T^{j-1} the right-hand side of the non-terminal $T^j[i]$, which can be easily found with support of array D in constant time.

Time complexity. The whole decompression process takes $O(n)$ time.

4.3 Extracting substrings

In order to support extraction of substrings from the compressed text, it is necessary to augment the dictionary with two additional data structures: P_S , a partial-sum on the lengths of the symbols in the reduced string T^ℓ of the last recursion level, and L , a data structure that for each non-terminal X_i stores $|\mathcal{G}(X_i)|$. Formally, those data structures are defined as:

$$P_S(i) = \sum_{j=1}^{i-1} |\mathcal{G}(X_{S_j})|, \quad X_S \rightarrow X_{S_1}, \dots, X_{S_k} \text{ and } 1 \leq i \leq k+1$$

$$L(X) = |\mathcal{G}(X)|, \quad X \in \Sigma \cup \Gamma$$

The data structure L can also be defined recursively as:

$$L(X) = \begin{cases} 1, & X \in \Sigma \\ \sum_{i=1}^{|S|} L(S[i]), & X \rightarrow S \end{cases}$$

To obtain a substring $T[l, r]$, we then proceed as follows:

- (1) With a binary search, locate *indices* a and b from P_S such that:

$$a = \max\{1 \leq k \leq |T^\ell| \mid P_S(k) \leq l\}$$

$$b = \min\{1 \leq k \leq |T^\ell| + 1 \mid P_S(k) > r\} - 1$$

- (2) Let ℓ be the number of levels in GCIS grammar and S the string derived from X_S . Then define $E^\ell = S[a, b]$ and follow the next steps for $i = \ell$ to $i = 1$.
- (3) Apply a derivation step to each non-terminal $X \in E^i$ to obtain a new string E^{i-1} . Note that $\mathcal{G}(E^{i-1}) = T[l', r']$ is a superstring of $T[l, r]$.
- (4) Trim E^{i-1} from the left and right as much as possible as long as it generates a superstring of $T[l, r]$. This can be done efficiently because we know the length of $\mathcal{G}(X)$, for every $X \in \Sigma \cup \Gamma$.
- (a) If $i = 1$, then E^0 contains only terminal symbols and generates a superstring $T[l', r'] = E^0$ of $T[l, r]$. Thus, one simply extracts the symbols $E^0[l - l' + 1, r - l' + 1]$ to obtain $T[l, r]$.
- (b) If $i > 1$, then E^{i-1} contains only non-terminal symbols and generates a superstring $T[l', r'] = \mathcal{G}(E^{i-1})$ of $T[l, r]$. We then trim E^{i-1} by using L and finding, with a linear search, two indices a and b of E^{i-1} such that:

$$a = \max \left\{ 1 \leq k \leq |E^i| \mid l' + \sum_{j=1}^{k-1} L(E^i[j]) \leq l \right\}$$

$$b = \max \left\{ 1 \leq k \leq |E^i| \mid r' - \sum_{j=k+1}^{|E^i|} L(E^i[j]) \geq r \right\}$$

E^{i-1} is then trimmed to $E^{i-1}[a, b]$ before proceeding.

Figure 2 shows an example for extracting a text using the aforementioned procedure.

Implementation details. Since the length of the string T^ℓ is much shorter than the original text in practice, the verbatim representation of P_S as an array of integers is affordable.

The array L is represented using DACs. This representation allows efficient access while representing the data in a compact way.

To support fast extraction, we need to efficiently decompress a single rule. Simple8b encoding works very well when the objective is compressing or decompressing since all the rules are first expanded sequentially in the decompressing stage. However, when the aim is to extract symbols, we need to expand individual rules. Thus, instead of encoding all the lcp values with the simple8b scheme, Elias-Fano encoding is employed, allowing us to retrieve a random lcp value

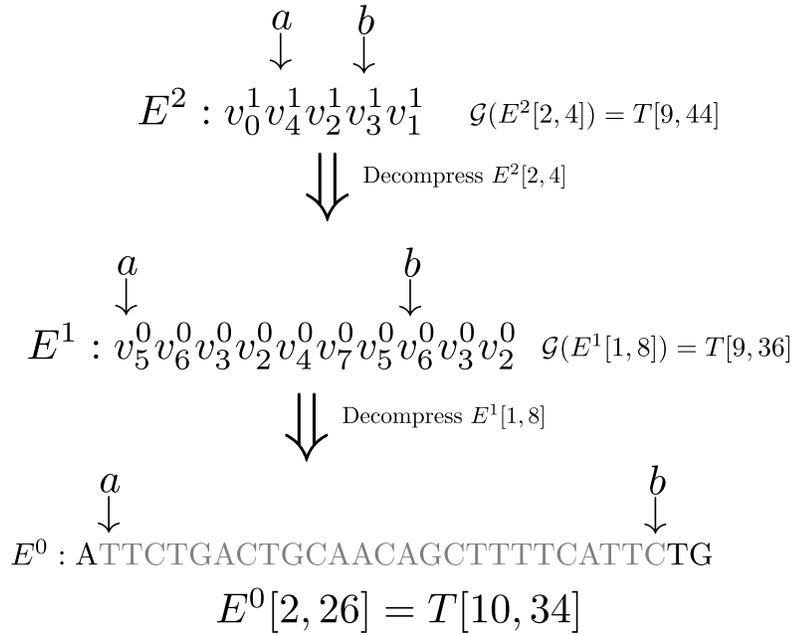


Fig. 2. Extraction of the substring $T[10, 34]$ of the text of Figure 1. Initially, a binary search is performed on P_S to identify the substring of E^2 that shall be decompressed: $E^2[2, 4]$, which generates $T[9, 43]$, is decompressed to obtain E^1 . A linear scan is performed in both ends considering the length of the terminals generated by each rule of E^1 to find the indexes $a = 2$ and $b = 8$. $E^1[2, 8]$ is then decompressed and $E^0 = T[9, 36]$ is obtained, which makes possible to extract $T[10, 34]$ by simply ignoring both ends.

of a rule efficiently and hence the decoding of a random rule. The length of each $s(\alpha_i)$ is also encoded using Elias-Fano and the $s(\alpha_i)$ values are encoded in a fixed-width integer array. Since the lcp values are front-encoded, we force that every k -th lcp value is set to 0, with $k \in O(1)$. This setting does not have a significant impact on compression and ensures that we have to backtrack a constant number of rules to extract an individual rule prefix.

4.4 Suffix array construction

The suffix array (SA) construction boils down to sorting all suffixes of T . Although GCIS compression does not sort suffixes, it executes Step 1 of SAIS and the production rules created correspond to the LMS-substrings already sorted, which is used by SAIS for sorting all suffixes. We show next how to modify our decompression algorithm for building SA as a byproduct with, asymptotically, no additional overhead.

First, when $j = \ell$, $X_S \rightarrow T^\ell$ is taken, and its suffix array SA^j is built directly as $SA^j[T^j[i]] = i$. Nong et al. [2009] observed that SA^j also gives the order of all LMS-suffixes of string T^{j-1} . Then, T^{j-1} is decoded (Section 4.2), and Steps 2, 3 and 4 of SAIS (Section 3) are executed to obtain SA^{j-1} , and so on. The algorithm proceeds for $j = \ell - 1, \dots, 1$, obtaining the reduced string T^{j-1} together with SA^{j-1} at each iteration. At the end, the original string T is decoded from T^1 and SA is induced from SA^1 .

Time Complexity. SA is built in $O(n)$ time, since each step of SAIS is linear and the length of all reduced strings is $O(n)$.

4.5 LCP array construction

When $|\Sigma| \in O(1)$, the longest common prefix (LCP) array can also be computed in linear time within the induced suffix sorting framework [Fischer 2011; Louza et al. 2017b]. We show below how to modify our decompression algorithm to compute SA and LCP together with, asymptotically, no additional cost.

When $j = 1$, the original string T is decoded from T^1 , and SA^1 stores the order of all LMS-suffixes of T . Then, in linear time, we compute the LCP array of the LMS-suffixes using a sparse variant of Φ -algorithm by Kärkkäinen et al. [2009], which avoids storing auxiliary arrays by reusing the space of $SA[n/2, n]$ and $LCP[n/2, n]$. The `lcp`-values between the LMS-suffixes are used to induce the `lcp`-values between the L-suffixes during Step 3 (Section 3), and these are used to induce the S-suffixes during Step 4 (see Louza et al. [2017b] for details). Given an additional stack of $O(\sigma \log n)$ bits [Gog and Ohlebusch 2011], each `lcp`-value induction is done in $O(\sigma)$ time, which we assume to be constant at the top recursion level.

Therefore, at the end we have computed SA and LCP as a byproduct of GCIS decompression.

Time complexity. SA and LCP are built in $O(n)$ time, since each step of SAIS is linear and the LCP-values induction can be done in constant time.

5 EXPERIMENTS

To confirm the practical value of GCIS, we conducted experiments using several corpora. We measured compression and decompression speed, compression ratio, and memory usage during compression and decompression of GCIS against classical and grammar-based compressors (Section 5.4), evaluated the extraction of symbols (Section 5.6) and showed the efficiency of suffix array construction of GCIS during the decompression (Section 5.7).

In the following subsections we describe and discuss the experimental setup and the results.

5.1 Compressors and Extractors

To evaluate GCIS in compression speed, decompression speed and compression ratio, we chose the well-known compressors: GZIP [Gailly and Adler 2017]; BZIP2 [Seward 1996]; 7-ZIP [Pavlov 2017]; the statistical compressor PPMJ [Shkarin 2006]; the grammar compressor REPAIR [Wan 2014]; the grammar compressor SOLCA [Takabatake et al. 2017]; the LZD compressor that relies on a clever LZ78 modification [Goto et al. 2015]; the bidirectional parsing scheme LZRR [Nishimoto and Tabei 2019]; the Lempel-Ziv approximation for very large texts ReLZ [Kosolobov et al. 2020]; the REPAIR approximation for very large texts BIGREPAIR [Gagie et al. 2019].

Regarding extraction of symbols, we compared GCIS with different encodings of REPAIR grammars that allow fast extraction. These encodings can be represented in a more straight-forward way, storing $\mathcal{G}(X)$, for $X \in \Sigma \cup \Gamma$, or in a more elaborated way, creating succinct tree data structures that replace the original grammar encoding while allowing one obtain the right-hand side of any rule, as described by Maruyama and Tabei [2013]. The implementation of such data structures was based on the work of Gagie et al. [2020] and can be found in I [2020]. We used the following encodings:

- `PlainSlp_32Fblc`: uses 32-bit integers for the array representations.
- `PlainSlp_FblcFblc`: employs the minimum bit length required to represent the maximum value of a given integer array.
- `PlainSlp_IblcFblc`: uses roughly $\lceil \lg i \rceil$ bits to represent the i -th rule exploiting that the i -th rule is less than i . For representing $\mathcal{G}(X)$, for $X \in \Sigma \cup \Gamma$, it uses the same strategy of `PlainSlp_FblcFblc`.

- PoSlp_Iblc: employs the approach POSLP of [Maruyama and Tabei \[2013\]](#) to represent the parse tree and encodes the leaves using roughly $\lceil \lg i \rceil$ bits for the i -th rule.
- PoSlp_Sd: applies the POSLP approach of [Maruyama and Tabei \[2013\]](#) to represent the parse tree and encodes the leaves with Elias-Fano.

In order to assess the computation of suffix and LCP arrays directly from decompression, GCIS was compared with efficient suffix and LCP construction algorithms implemented by `sais-lite` [[Kurpicz 2015; Mori 2010](#)] and `divsufsort` [[Kurpicz 2016; Mori 2008](#)].

GCIS source code and a detailed description of the processed data are available at <https://github.com/danielsaad/gcis>.

5.2 Texts

Regular texts were taken from the corpora `large-corpus` [[Trigell 1998](#)], `enwiki` [[Mahoney 2006](#)], `manzini` [[Manzini 2003](#)], `pizza-chili` [[Ferragina and Navarro 2005a](#)] and `silesia` [[Deorowicz 2003](#)]. Repetitive texts were chosen from `pizza-chili-repetitive corpus` [[Ferragina and Navarro 2005b](#)]. Very large inputs were built by repeating and mutating strings such as `chr19` [[Consortium 2009](#)], `sars-cov` [[NCBI 2020](#)] and `salmonella` [[NCBI 2007](#)] with a mutation rate of 0.1%, thus these texts are highly repetitive as well; each filename has an integer suffix that represents the number of repetitions. In addition, a 20GB prefix from November 2019 Wikipedia dump was taken [[Wikipedia 2019](#)]. Tables 1, 2 and 3 summarize the chosen texts and their size, grouping in boxes texts from the same corpus.

Along with GCIS, we considered a set of compressors for each *corpus* type. For regular texts, we considered the compressors: GZIP, BZIP2, PPMJ, 7-ZIP, REPAIR, LZD and SOLCA. For repetitive texts, the following compressors were adopted: 7-ZIP, REPAIR, RELZ, LZD, SOLCA and LZRR. Finally, for very large texts, the compressor set was defined by: 7-ZIP, RELZ, BIGREPAIR, LZD and SOLCA. These choices were made regarding the effectiveness of the compressors for each *corpus* type and the relevance in comparison with GCIS. For example: LZRR and REPAIR cannot deal with large volumes of data due to its high memory consumption, hence it were not considered for very large-texts; GCIS, BZIP2 and PPMJ does not capture repetitiveness very well, thus they were disconsidered for repetitive and very large texts.

5.3 Environment Setup

Due to memory capacity and availability, we conducted the experiments in two machines, one for the regular and repetitive corpora and another for the very large datasets. Their specifications follow:

Machine #1, used for regular and repetitive texts:

- CPU: 2x Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz CPUs;
- RAM Memory: 64GB;
- Operating System: Centos7, kernel version 3.10.

Machine #2, used for very large datasets:

- CPU: 2x Intel(R) Xeon(R) E5-2630 v3 @ 2.40GHz;
- RAM Memory: 386GB;
- Operating System: Debian GNU/Linux 8, kernel version 3.16.

Table 1. Regular texts.

Regular text	Size (MB)
archive	27.07
emacs	47.46
linux	47.60
samba	41.58
spamfile	84.22
enwiki8	100.00
enwiki9	1000.00
chr22	34.55
etext99	105.28
gcc-3.0.tar	86.83
howto	39.42
jdk13c	69.73
linux-2.4.5.tar	116.25
rctail96	114.71
rfc	116.42
sprot34.dat	109.62
w3c2	104.20
dblp.xml	296.14
dna	403.93
english	2210.40
pitches	55.83
sources	210.87
dickens	10.19
mozilla	51.22
mr	9.97
nci	33.55
ooffice	6.15
osdb	10.09
reymont	6.63
samba	21.61
sao	7.25
webster	41.46
xray	8.457
xml	5.35

Table 2. Repetitive texts.

Repetitive text	Size (MB)
cere	461.29
coreutils	205.28
dblp.xml.00001.1	104.86
dblp.xml.00001.2	104.86
dblp.xml.0001.1	104.86
dblp.xml.0001.2	104.86
dna.001.1	104.86
einstein.de.txt	92.76
einstein.en.txt	467.63
english.001.2	104.86
Escherichia_Coli	112.69
influenza	154.81
kernel	257.96
para	429.27
proteins.001.1	104.86
sources.001.2	104.86
world_leaders	46.97

Table 3. Very large texts.

Very large text	Size (MB)
c050	2956.45
c100	5912.90
c150	8869.35
c200	11 825.80
c250	14 782.25
c300	17 738.69
c350	20 695.14
sars-cov100000	2990.30
sars-cov200000	5980.60
sars-cov300000	8970.90
sars-cov400000	11 961.20
sars-cov500000	14 951.50
sars-cov600000	17 941.80
sars-cov700000	20 932.10
enwiki-20191120-20G	20 000.00
salmonella1000	4928.40
salmonella2000	9856.80
salmonella3000	14 785.20
salmonella4000	19 713.60

We compiled GCIS, REPAIR (and its extractors), BIGREPAIR, PPMJ and RELZ under gcc with `-O3 -NDEBUG` flags. The default command line parameters of GZIP, BZIP2, PPMJ, RELZ were used on the experiments. A dictionary size of 1 GB was used in 7-ZIP. BIGREPAIR RAM usage was limited to 10 GB.

GCIS was implemented in C++11 using the Succinct Data Structure Library (SDSL) version 2.0 [Gog et al. 2014].

5.4 Compression and decompression

We evaluated all compressors in terms of compression ratio, compression and decompression speed. We also considered their peak memory usage during compression and decompression. Some compressors could not compress some texts, so its corresponding data in the graphs are missing. Decompression in RELZ is not implemented, nonetheless, RELZ serves as a compression benchmark since it approximates the Lempel-Ziv parse.

It is important to remark that BIGREPAIR does not produce a compact representation of rules, since it represents the right-hand side of its rules with 2 integers (all the rules are of length two). However, we optimized it by representing each rule with at most $\lceil \log_2 r \rceil$ bits, r being the number of rules, and integrating the non-terminals that occur only once in their corresponding right-hand side. This saves $\lceil \log_2 r \rceil$ bits for each eliminated non-terminal.

Compression ratio. It stands for the ratio between the compressed and the original text size, and it is given as percentage.

Figure 3 shows that, for regular texts, LZD and REPAIR outperforms GCIS, and it is competitive with a basic Lempel-Ziv compressor such as GZIP. However, it is clearly outperformed by BZIP2, PPMJ and 7-ZIP. The latter displays the best compression ratio overall, being PPMJ a close competitor in some cases. GCIS presented better compression ratio than SOLCA.

Figure 4 shows the compression ratios for the repetitive texts. In particular, 7-ZIP obtains the best compression ratio in all cases, closely followed by REPAIR. The compression ratio of GCIS is about twice that of REPAIR in most cases, but it is still very good in absolute terms and outperforms LZD, SOLCA, LZRR and RELZ.

The results for the very large texts are depicted in Figure 5. 7-ZIP is better for the text `enwiki-20191120-20G`. The situation stays as in the smaller repetitive files, with LZD pursuing 7-ZIP closely. BIGREPAIR and LZD compresses more than GCIS, and GCIS compresses better than SOLCA and RELZ, except for the the chromosome 19 based texts.

Compression speed. Figures 6,7 and 8 show the compression speed, in MB/s, of the compressors for each text type.

GZIP is the fastest compressor in most regular texts. GCIS is the second-fastest compressor, followed by BZIP2, PPMJ and LZD. In particular, GCIS is typically an order of magnitude faster than the other grammar compressors, REPAIR and SOLCA, which are its direct competitors.

Considering repetitive texts, GCIS is faster than RELZ, SOLCA and 7-ZIP; it is a order of magnitude faster than REPAIR and LZRR, but is outperformed by LZD, the fastest compressor.

For very large texts, GCIS is much faster than 7-ZIP, which becomes the slowest of the considered compressors, and slightly faster than SOLCA. However, RELZ and BIGREPAIR become much faster than GCIS, as expected from being designed for this scenario. LZD is clearly the fastest compressor. A problem for GCIS on these very large files is that, once the text exceeds 2 GiB, it needs to use 64-bit integers, which doubles the memory requirements. BIGREPAIR and RELZ do not suffer from this problem and require a small amount of main memory during compression.

Decompression speed. Figure 9 depicts the results for regular texts. GZIP and REPAIR are the fastest at decompressing, followed by 7-ZIP LZD and GCIS. PPMJ and SOLCA are the slowest decompressors.

Figure 10 shows that the situation is similar on repetitive texts, except that 7-ZIP becomes way faster than the others in almost all cases. Despite the relative differences, in absolute terms GCIS is still fast, decompressing the files in around 50 MB/s. SOLCA was outperformed by LZRR and it is the slowest decompressor.

For very large texts, as shown in Figure 11, 7-ZIP is the fastest, followed by LZD, GCIS and then by BIGREPAIR. SOLCA is almost an order of magnitude slower than GCIS.

Peak memory. We evaluated the peak memory consumption (resident size) of all compressors during compression and decompression for each type of text; the results are shown in Figures 12 to 17.

For regular texts, GZIP, BZIP2 and PPMDJ require negligible space to compress or decompress. GCIS is outperformed by SOLCA and followed by LZD, 7-ZIP, REPAIR, the last being behind by a large margin. The situation has some changes during decompression: SOLCA is followed by REPAIR, LZD and 7-ZIP, and GCIS.

On repetitive texts, regarding compression, GCIS is outperformed by SOLCA and LZD and followed by RELZ, 7-ZIP and REPAIR. On decompression, GCIS stays behind every compressor.

On very large texts, during compression, SOLCA is the most space-efficient, followed by RELZ, BIGREPAIR, 7-ZIP LZD and lastly by GCIS. Considering decompression the order changes: LZD is the most efficient compressor, followed by BIGREPAIR and SOLCA. GCIS consumed much more memory than the alternatives in this scenario because for the internal arrays needed for the suffix sorting procedure, which requires 64-bits per entry.

As expected, SOLCA excels for its memory consumption during compression or decompression when compared to other compressors designed for repetitive data.

5.5 Overview

Figures 18 to 20 present conceptual radar charts that summarize, for each text type, the performance of all compressors in each rated aspect. The closer the values are to the pentagon borders, the better the compressor performed on the corresponding aspect.

5.6 Extract operation

Results depicted by Figures 21 and 22 show that GCIS is faster than the extractors on succinct encodings of REPAIR but slower than those running on the more straightforward representation using integer arrays. In turn, regarding space on regular and repetitive texts, GCIS is more space-efficient than the straightforward encodings but less space-efficient than the POSLP alternatives, as shown in Figures 23 and 24. GCIS is then a competitive alternative regarding the space-time trade-off.

5.7 Suffix Array and LCP Construction

Considering the computation of SA and LCP arrays during decompression, we measured the total time to decompress the files with GCIS without generating a plain-text file, but instead inducing the SA and the LCP arrays. We compared these results with SAIS [Kurpicz 2015] and `divsufsort` [Fischer and Kurpicz 2017; Kurpicz 2016] implementations based on those of Yuta Mori, which are known as the fastest suffix array construction algorithms in practice.

Figure 25 shows the SA and LCP construction on the 8 most repetitive real texts when only the GCIS compressed texts are available. Very large texts were not considered because the implementations of [Kurpicz 2015] and [Kurpicz 2016] only deal with 32-bit integers. GCIS builds the SA and LCP arrays faster than decompressing and then using the suffix array construction algorithms over the plain text.

The hatched part corresponds to the LCP computation and the black bar corresponds to the time spent in decompressing the text with GCIS to calculate SA and LCP values using the SAIS and `divsufsort` implementations.

6 CONCLUSIONS

We have introduced GCIS, a new grammar-based compression algorithm based on the induced suffix sorting framework of SAIS [Nong et al. 2009]. GCIS uses the meta-symbols introduced by SAIS to generate non-terminals of a balanced grammar that reproduces the original text. Our experiments on repetitive texts show that GCIS compresses 3–7 times faster than REPAIR and 7-ZIP. Compared to REPAIR, the grammar compressor that compresses the most, GCIS compresses using 3–5 times less memory, yet it obtains a compressed file twice as large (yet the absolute compression is still attractive, below 5% in most cases). GCIS decompresses 2–8 times slower than REPAIR and 7-ZIP, though.

Grammar-based compression is attractive because, unlike Lempel-Ziv, it can be enriched to support fast extraction of arbitrary text substrings. From this perspective and regarding the space-time relation, our experiments show that GCIS is a competitive option when compared to REPAIR-based extractors, being faster and less space-efficient than the succinct encoding of REPAIR extractors and slower, but more space-efficient, than the more straightforwardly encoded REPAIR extractors.

Finally, as a by-product of GCIS, the suffix array of the text can be obtained during the decompression algorithm, faster than decompressing and running on the original text. Díaz-Domínguez and Navarro [2021] recently exploited this capability by building the BWT of a collection of DNA reads using a grammar construction that optimizes GCIS.

All previously discussed features make GCIS especially attractive in scenarios where it is required to support random access on the compressed text. Grammar compression of very large files is challenging with REPAIR because of its large main memory footprint, for which GCIS offers an interesting alternative. Given its slowness at decompression, the GCIS grammar is best suited as a compressed data structure to be repeatedly accessed without decompressing it completely. Díaz-Domínguez et al. [2021] and Akagi et al. [2021] recently exploited this fact on close variants of the GCIS grammar, building self-indexes that provide efficient pattern searches on grammar-compressed texts. They prove that their GCIS-like grammars are *locally consistent*, which means that similar text substrings are parsed largely in the same way, and use this feature to speed up searches in comparison with indexes built on arbitrary grammars.

A possible further improvement on GCIS would be to replace the variant of Nong et al. [2009] used during compression by the more space-efficient SACA-K [Nong 2013] algorithm. This could decrease the working space used during compression. Other possible aspect of investigation is the encoding possibilities considering the right hand side of rules using a hybrid strategy such as employing a compressor like REPAIR or a LZ77 based compressor. Grammar-based compressed indexes [Bille et al. 2018; Claude and Navarro 2010, 2012; Nishimoto et al. 2020a,b; Takabatake et al. 2014, 2015] are of particular interest.

ACKNOWLEDGMENTS

Part of the research of DSNN was performed during visits at University of Chile and supported by FAP-DF, grants 13619.54.33621.1708/2016 and 0193.001380/2017. FAL was supported by the grant 2017/09105-0 from the São Paulo Research Foundation (FAPESP). MAR was supported by FAP-DF and CNPq, grants DE 193.001.369/2016 and PQ 307672/2017-4. GN was funded by Basal Funds FB0001 and Fondecyt Grant 1-200038, Conicyt, Chile.

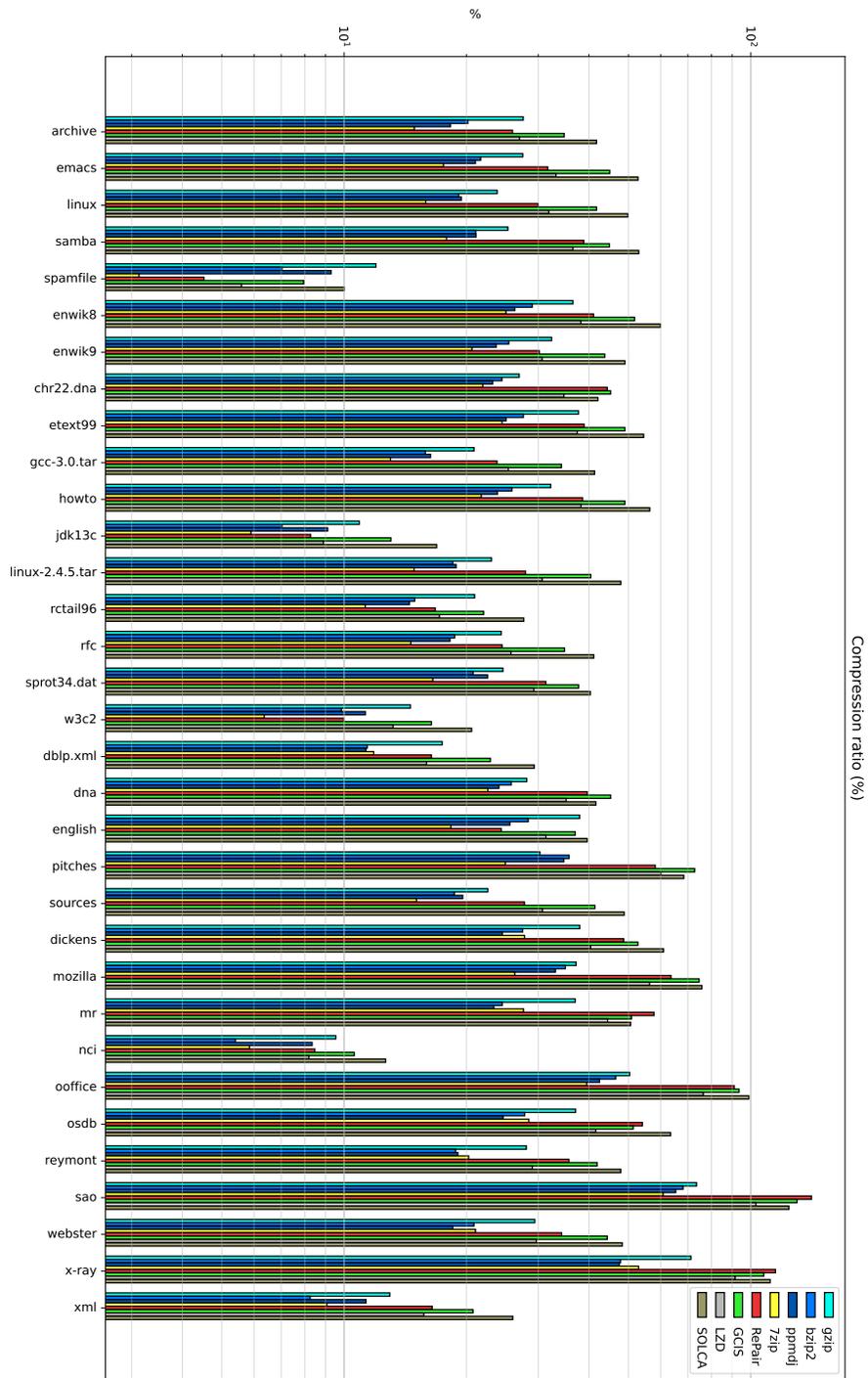


Fig. 3. Compression ratio on regular texts.

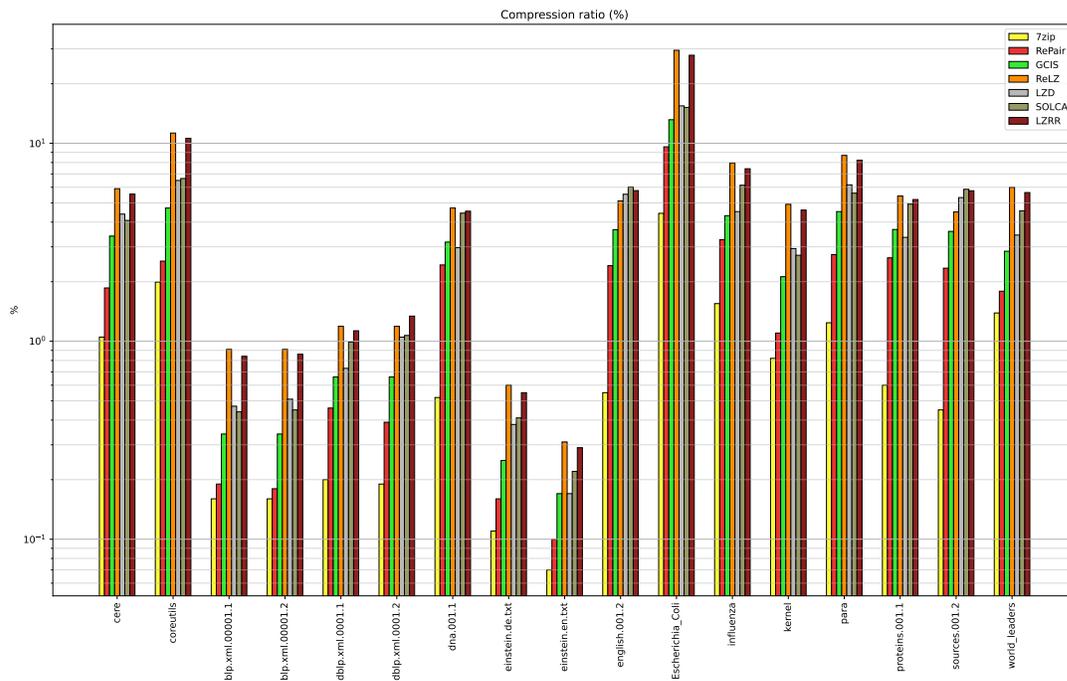


Fig. 4. Compression ratio on repetitive texts.

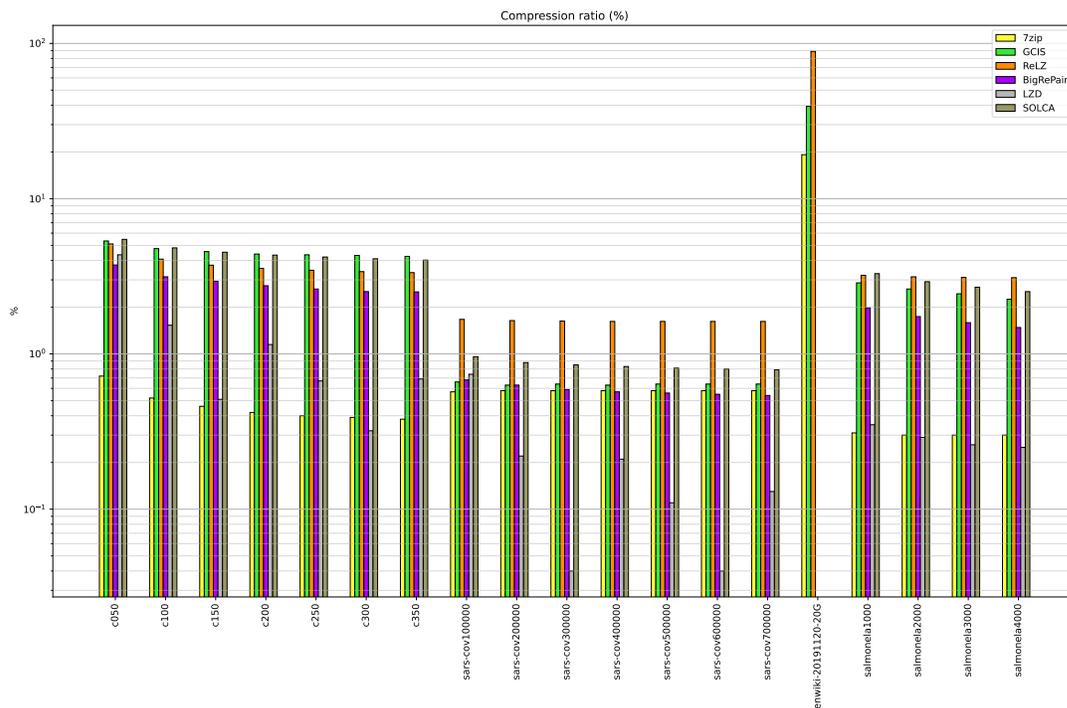


Fig. 5. Compression ratio on very large texts.

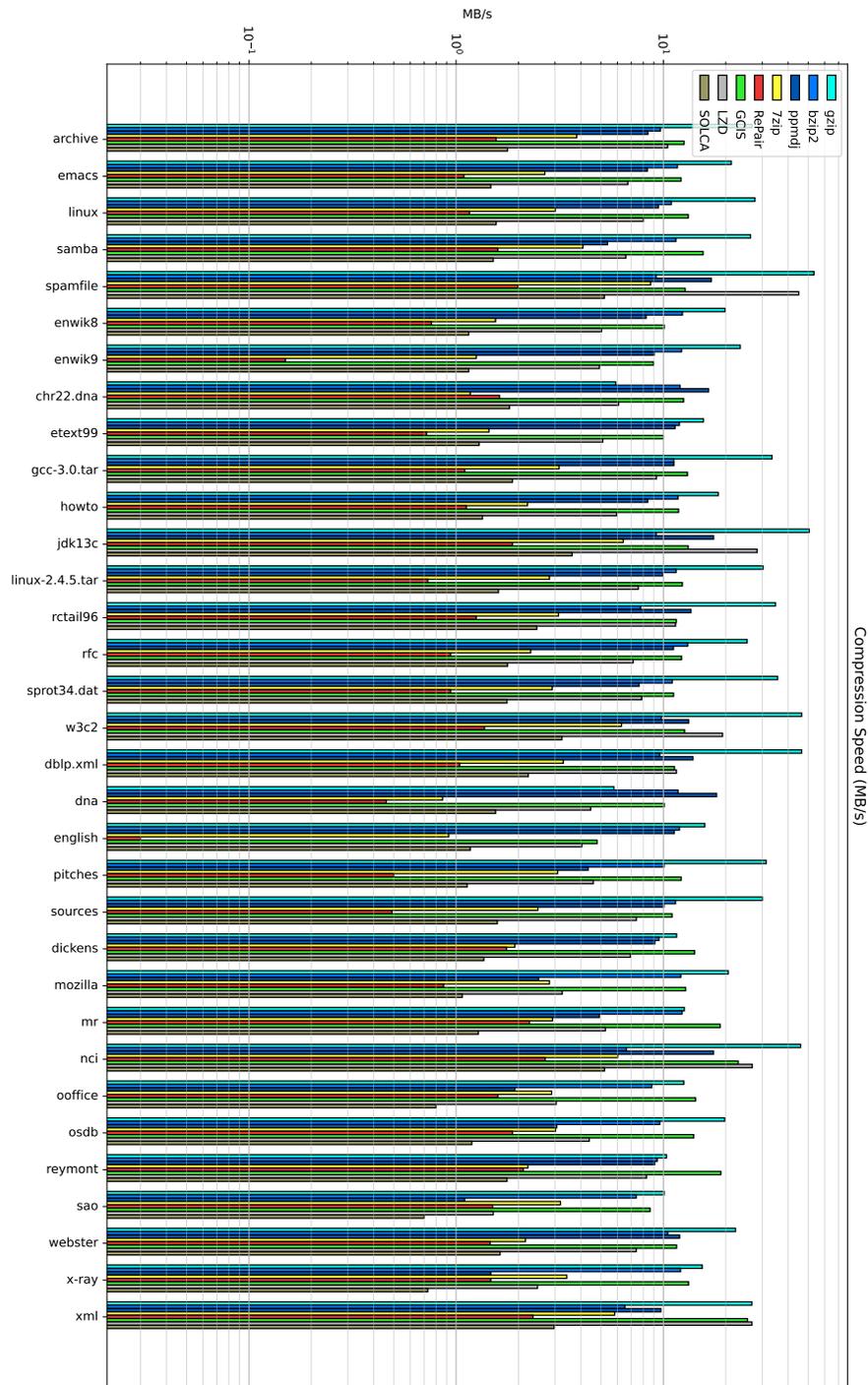


Fig. 6. Compression speed on regular texts.

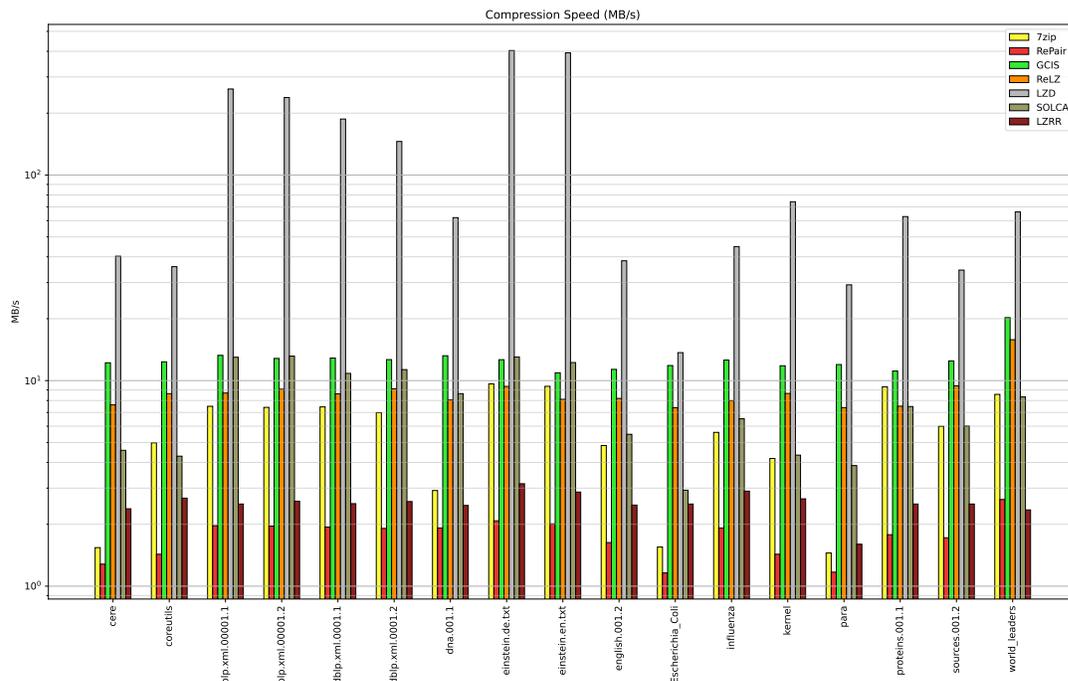


Fig. 7. Compression speed on repetitive texts.

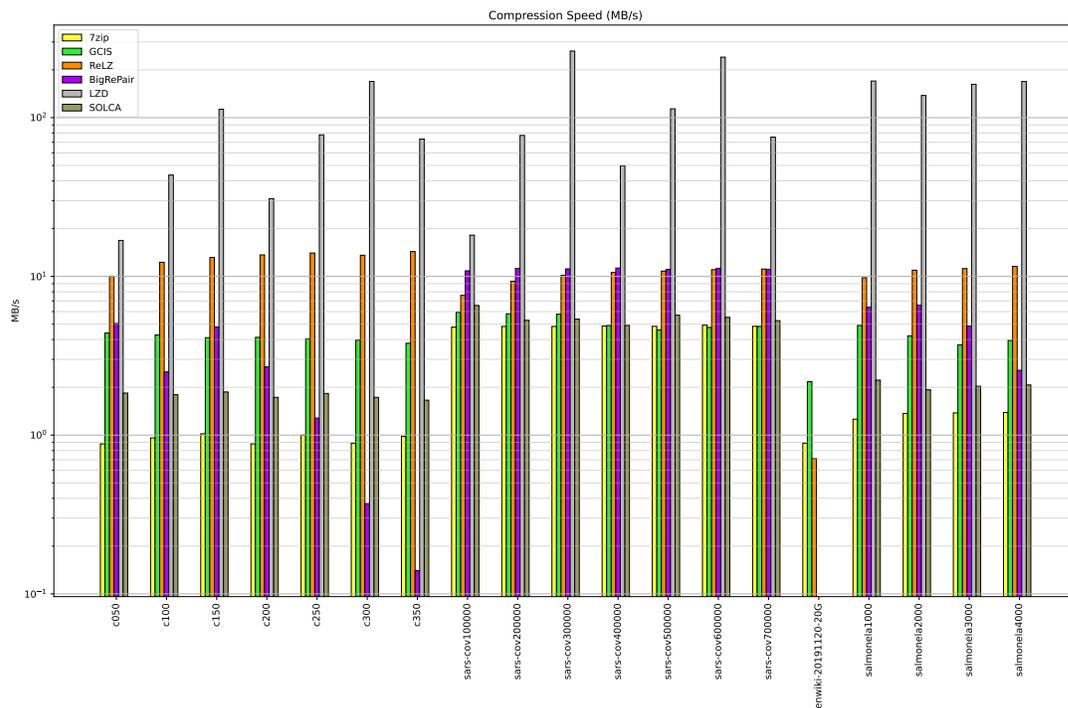


Fig. 8. Compression speed on very large texts.

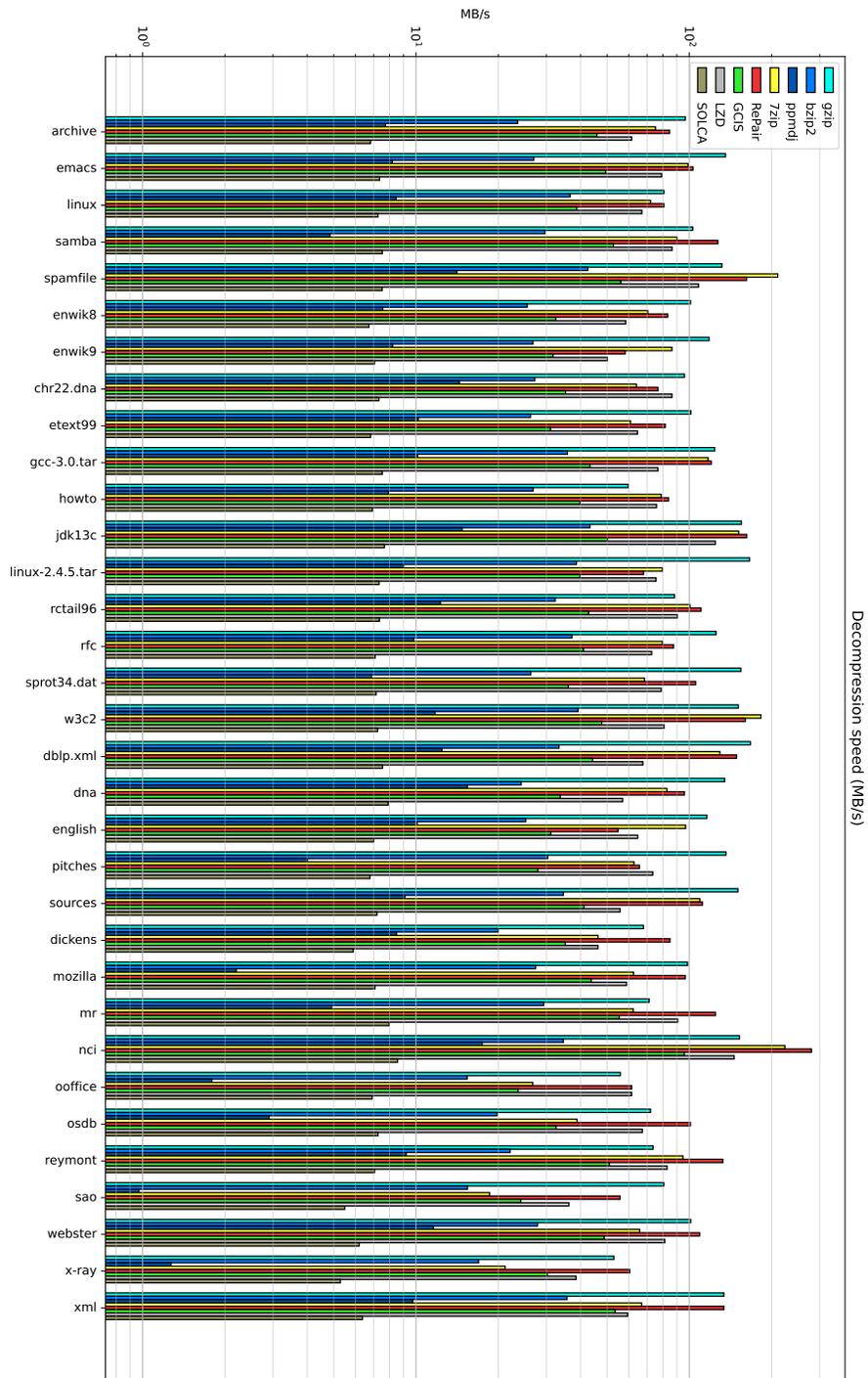


Fig. 9. Decompression speed on regular texts.

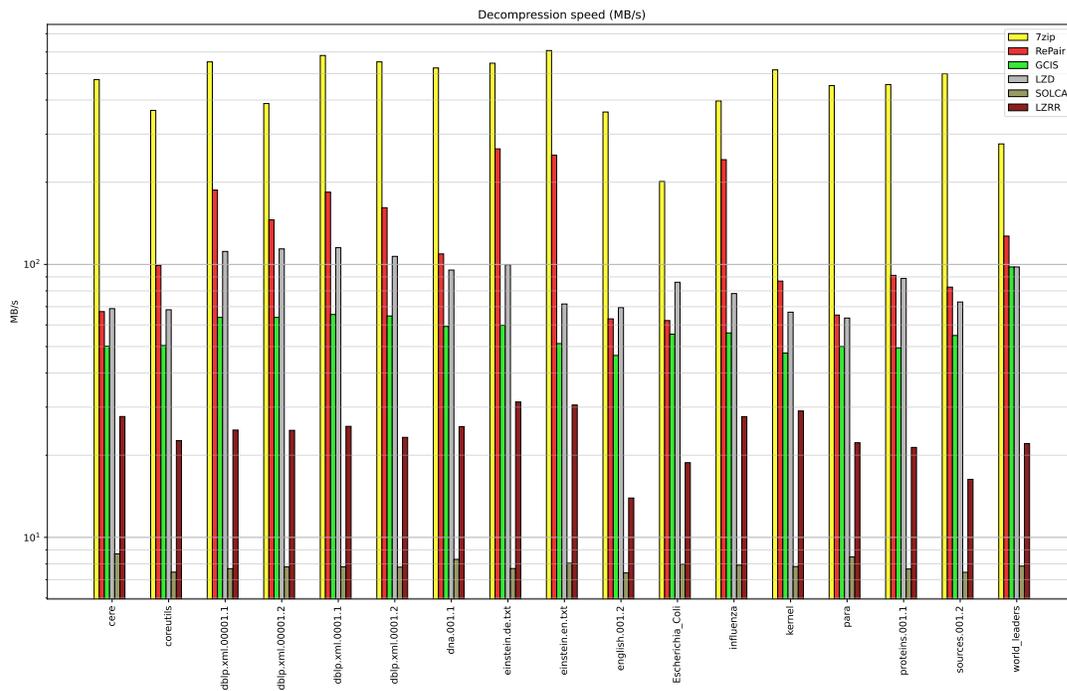


Fig. 10. Decompression speed on repetitive texts.

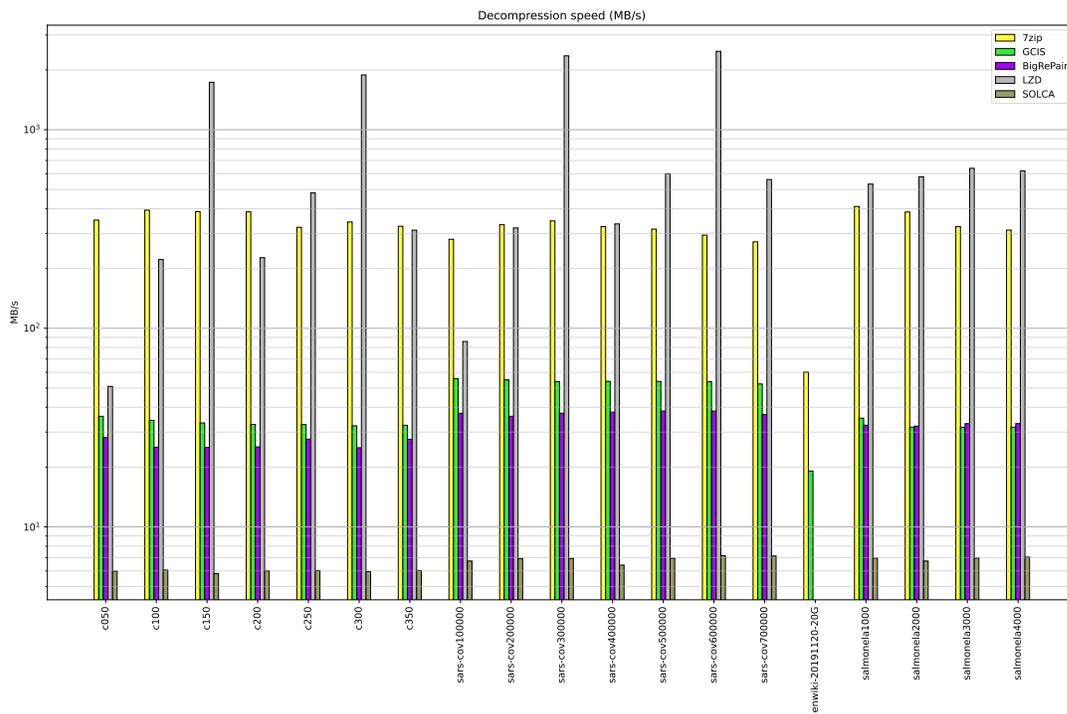


Fig. 11. Decompression speed on very large texts.

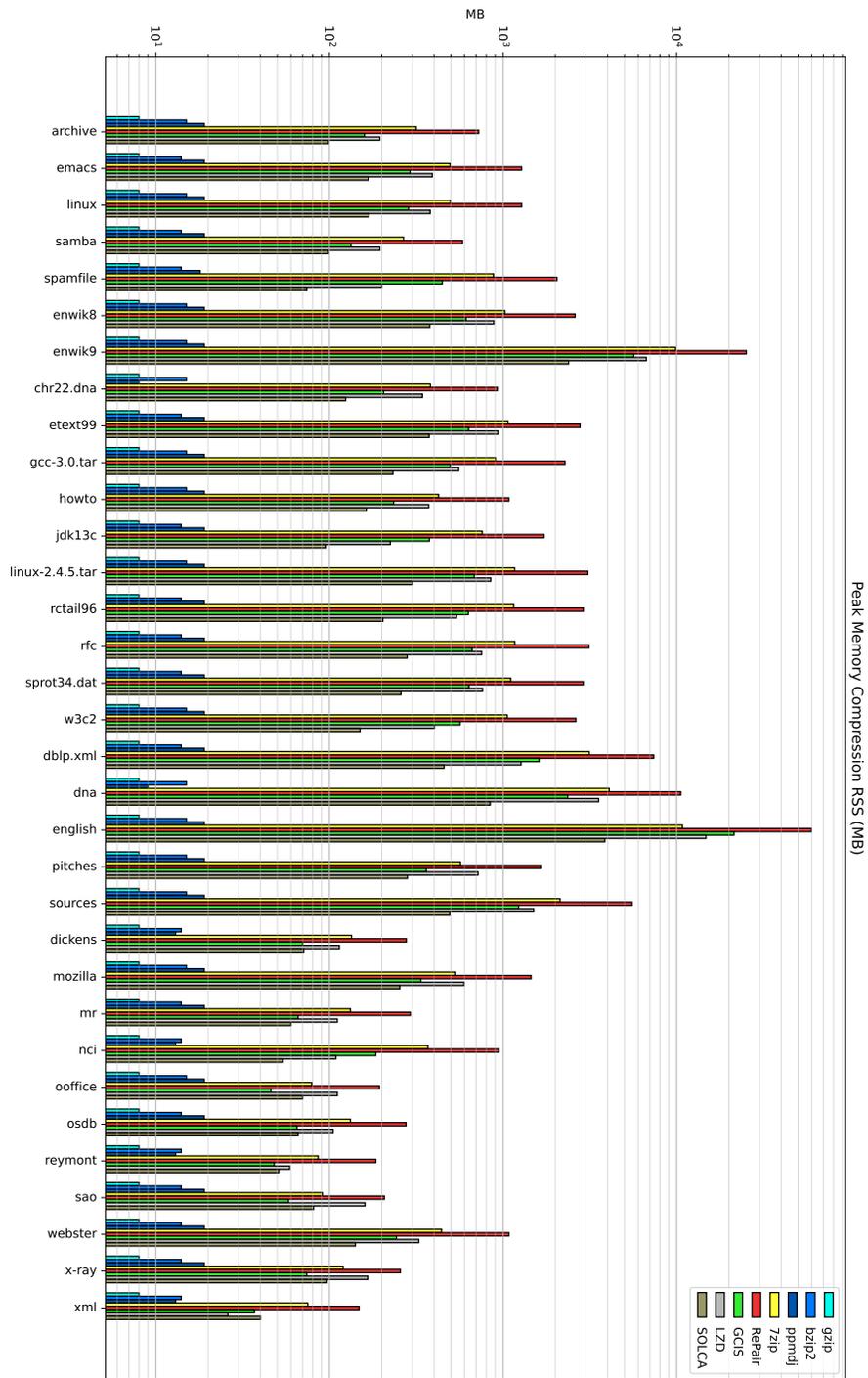


Fig. 12. Peak memory (in MB) used by the compressors during compression for regular texts.

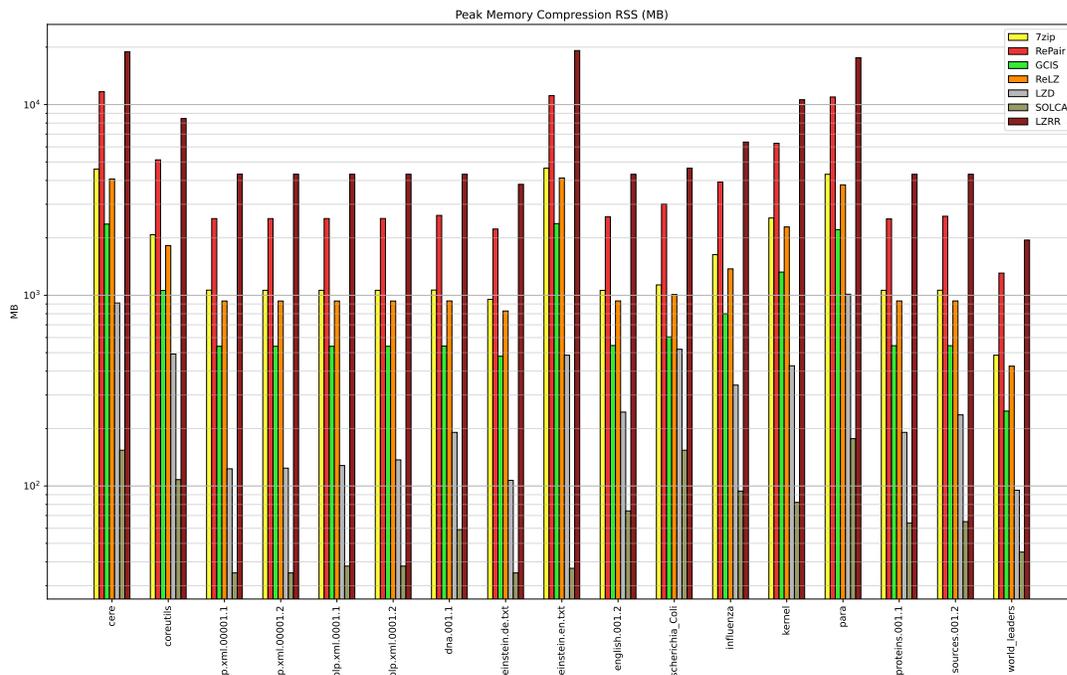


Fig. 13. Peak memory (in MB) used by the compressors during compression for repetitive texts.

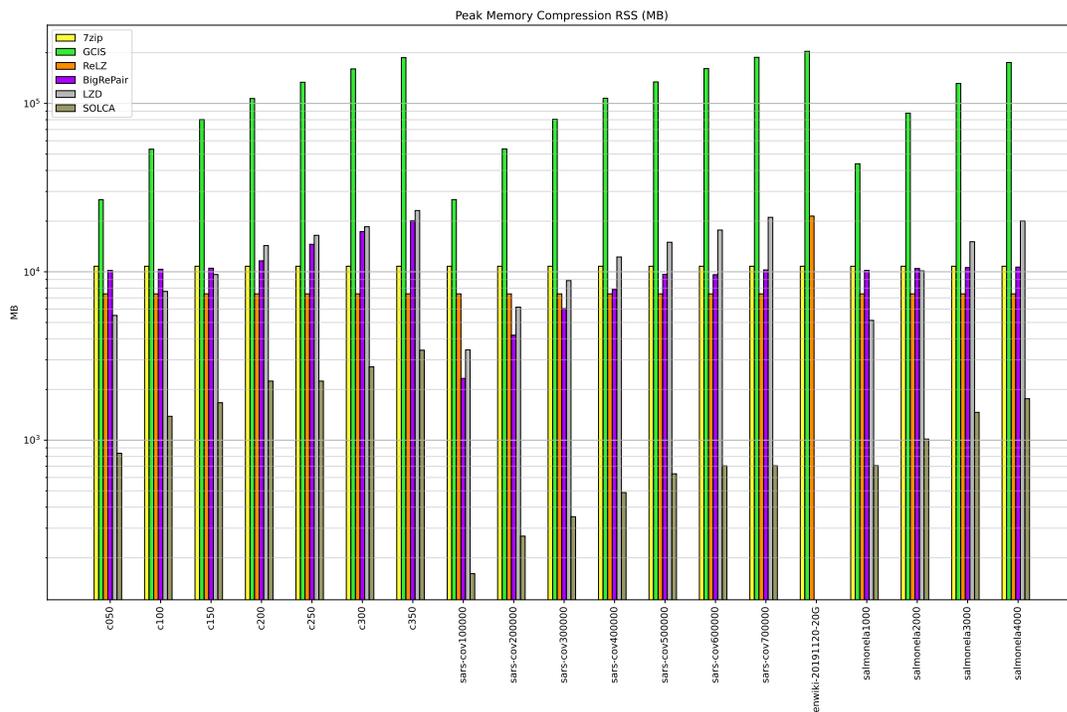


Fig. 14. Peak memory (in MB) used by the compressors during compression for very large texts.

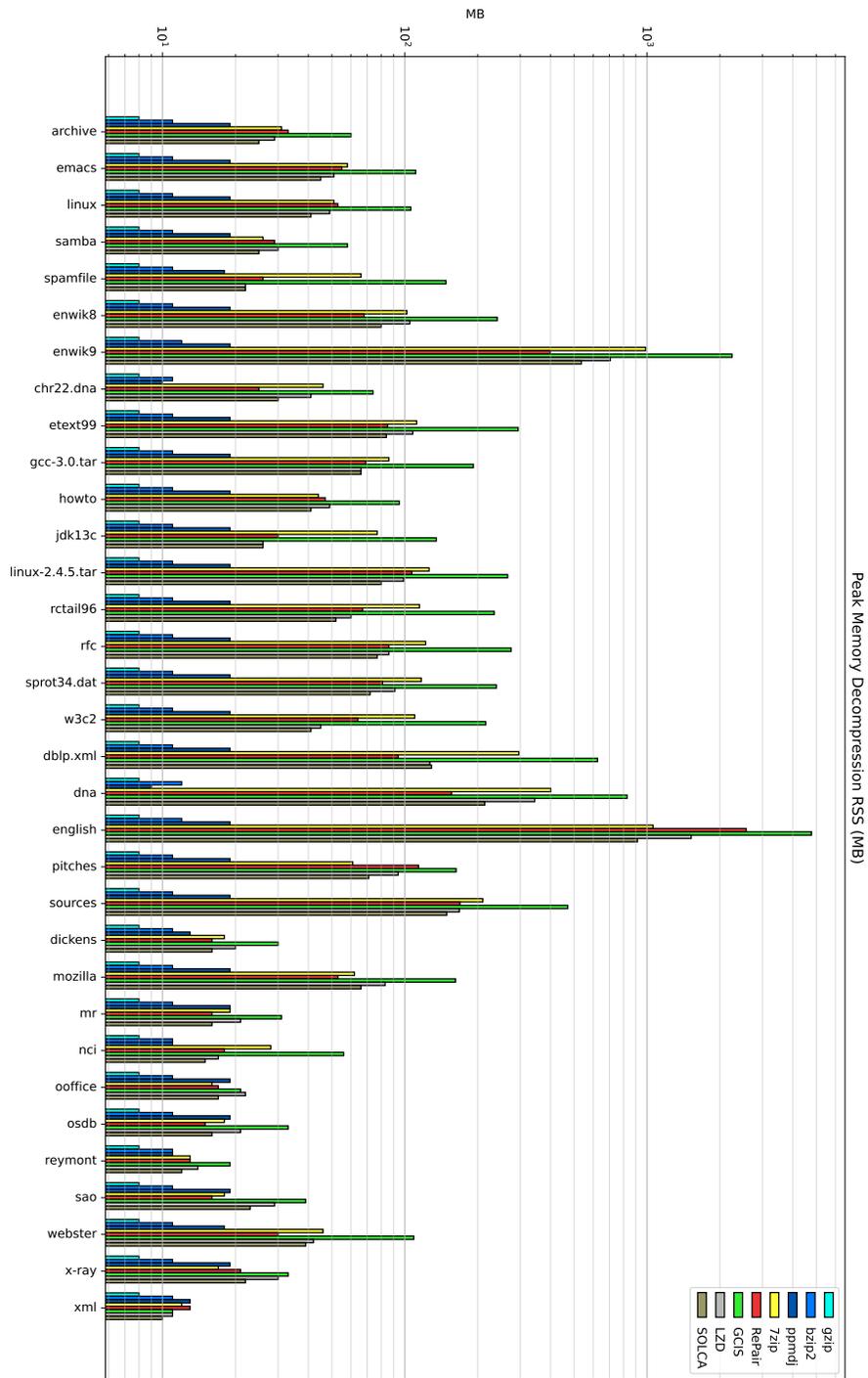


Fig. 15. Peak memory (in MB) used by the compressors during decompression for regular texts.

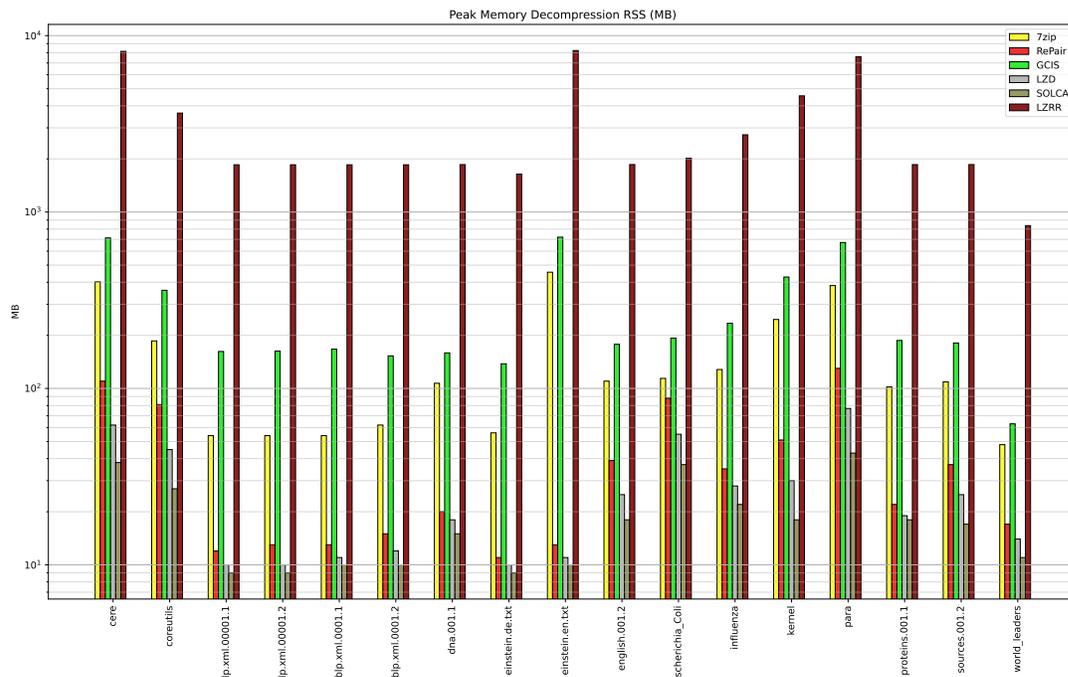


Fig. 16. Peak memory (in MB) used by the compressors during decompression for repetitive texts.

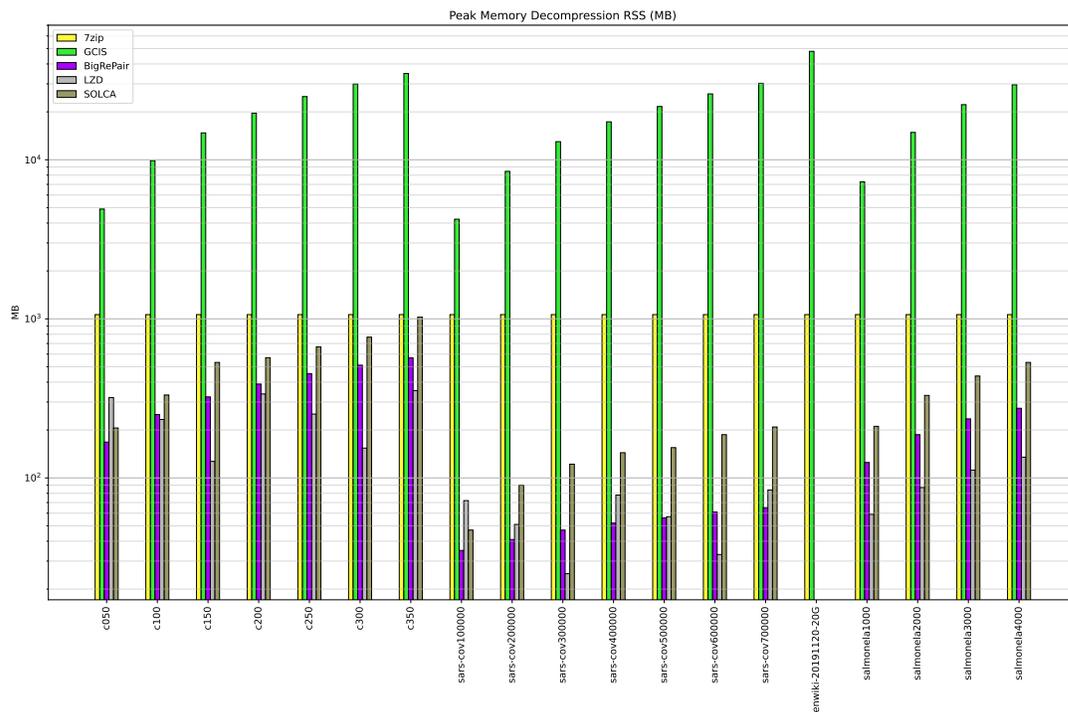


Fig. 17. Peak memory (in MB) used by the compressors during decompression for very large texts.

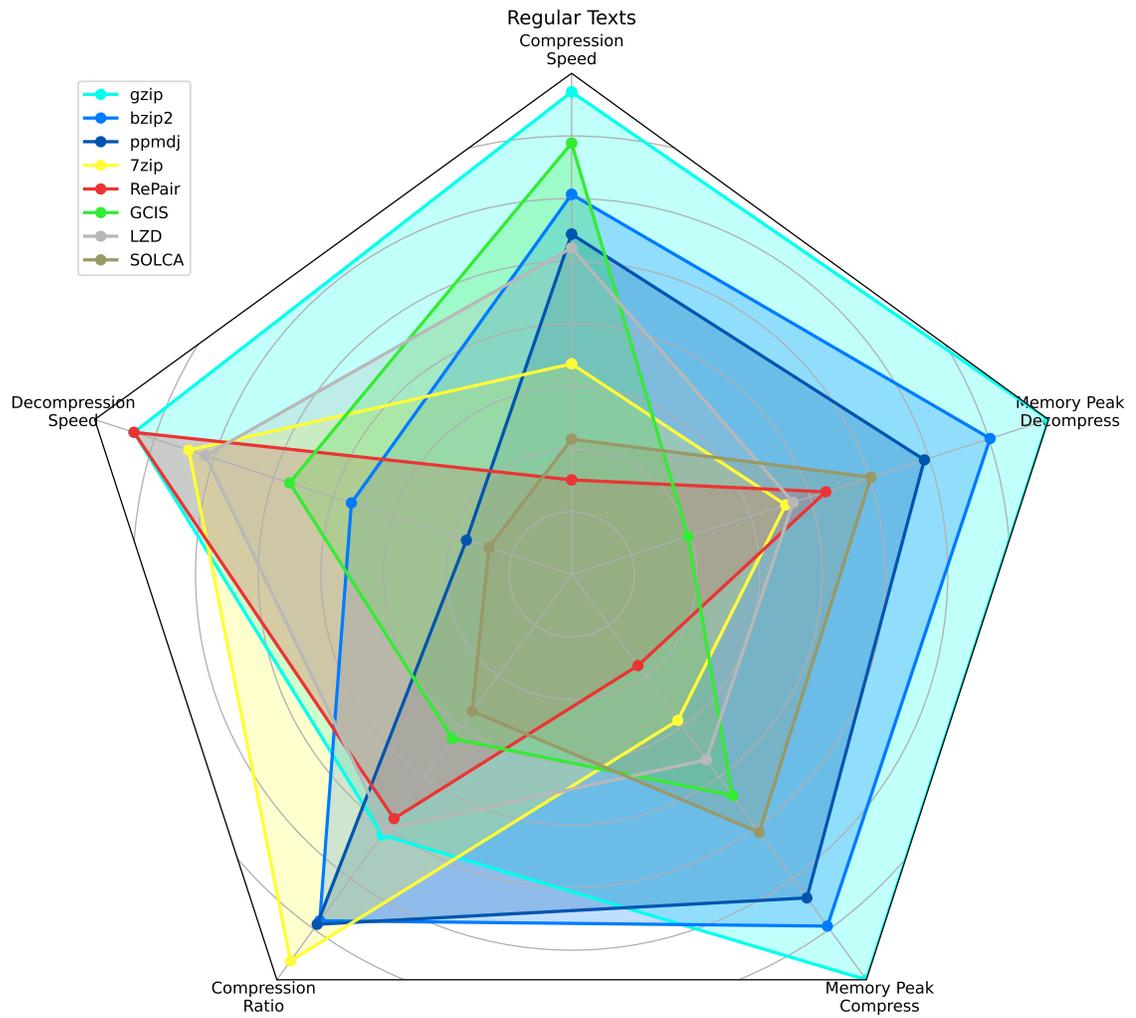


Fig. 18. Compressor comparison on regular texts.

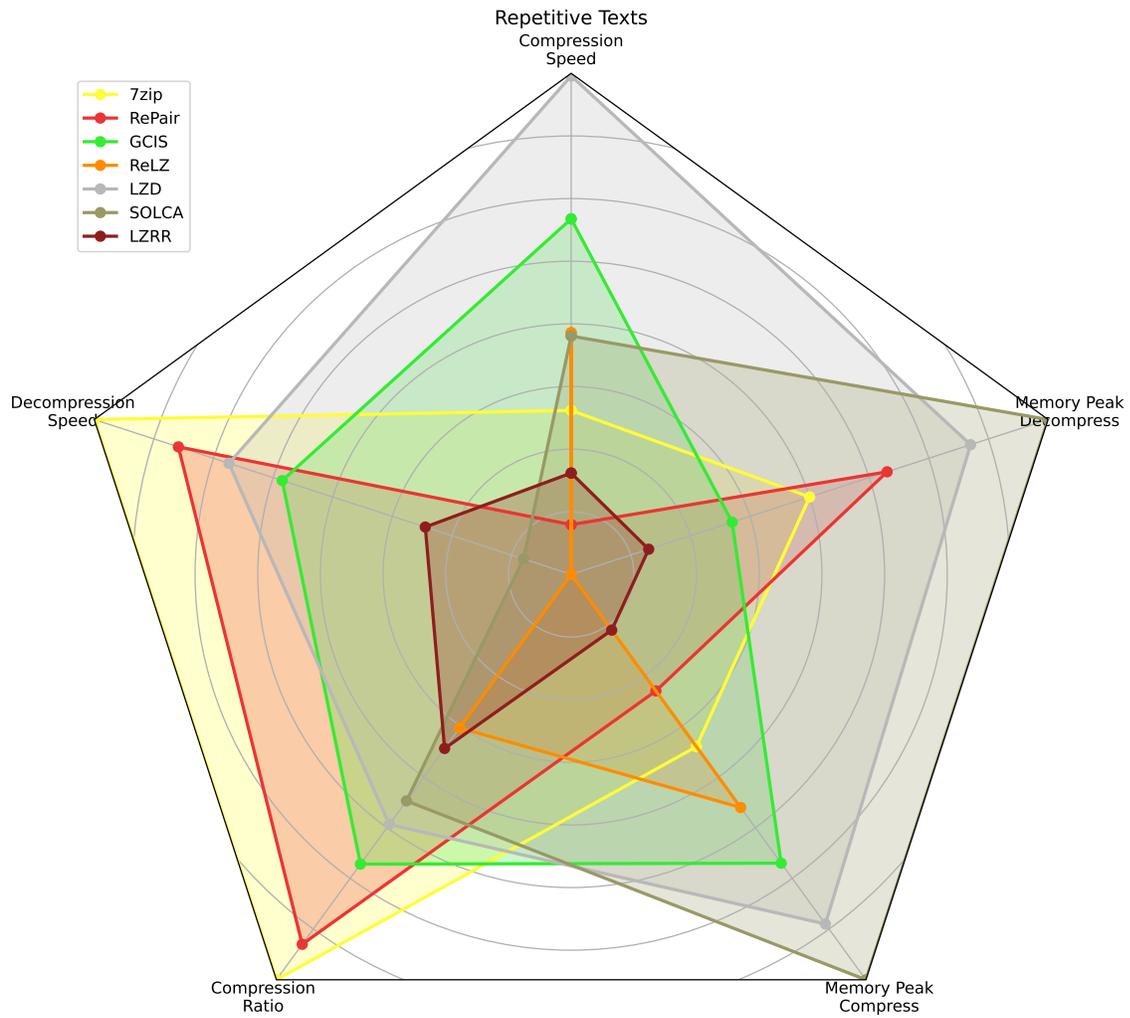


Fig. 19. Compressor comparison on repetitive texts.



Fig. 20. Compressor comparison on very large texts.

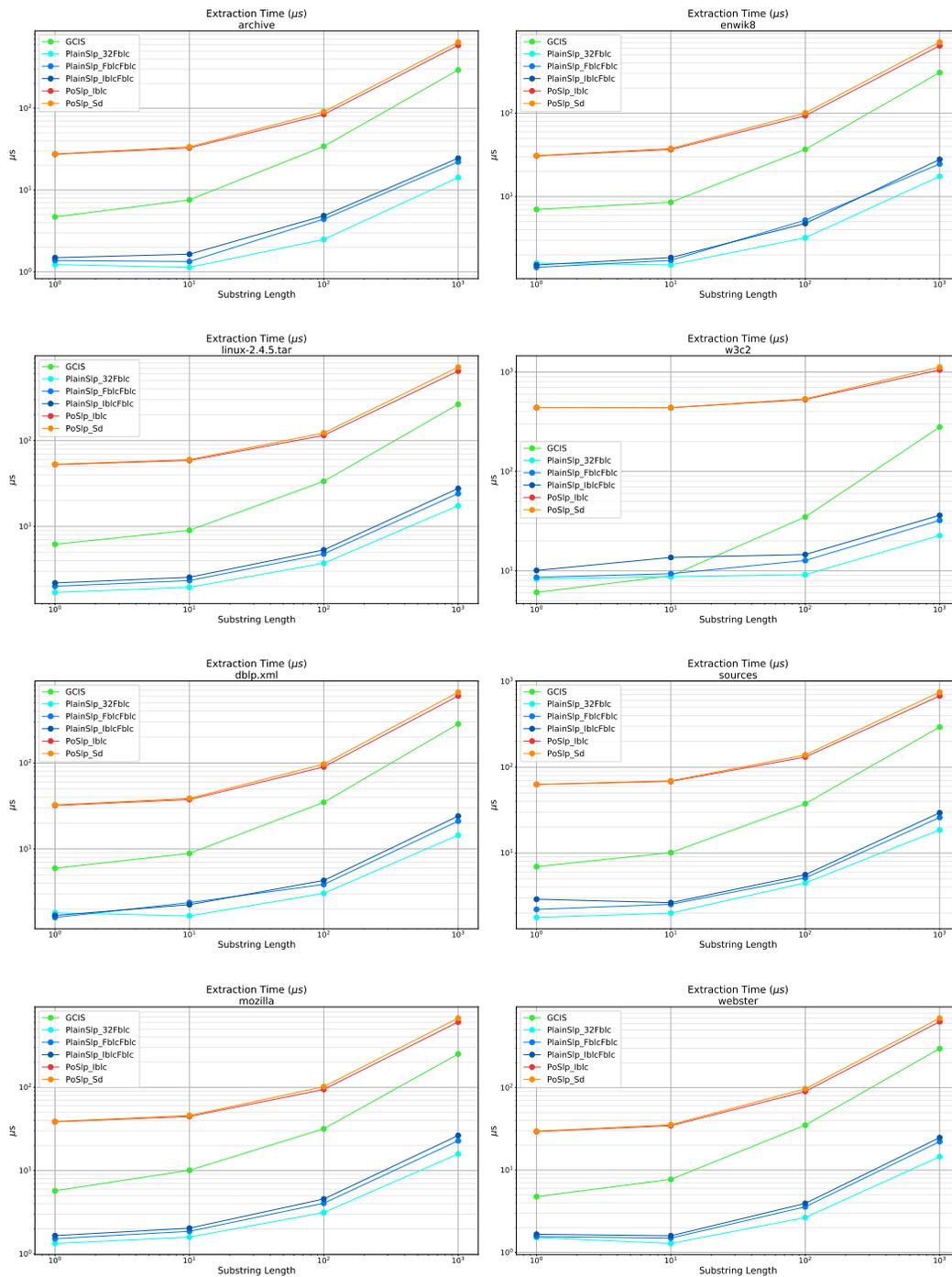


Fig. 21. Substring length vs. extraction time (microseconds) on regular texts.

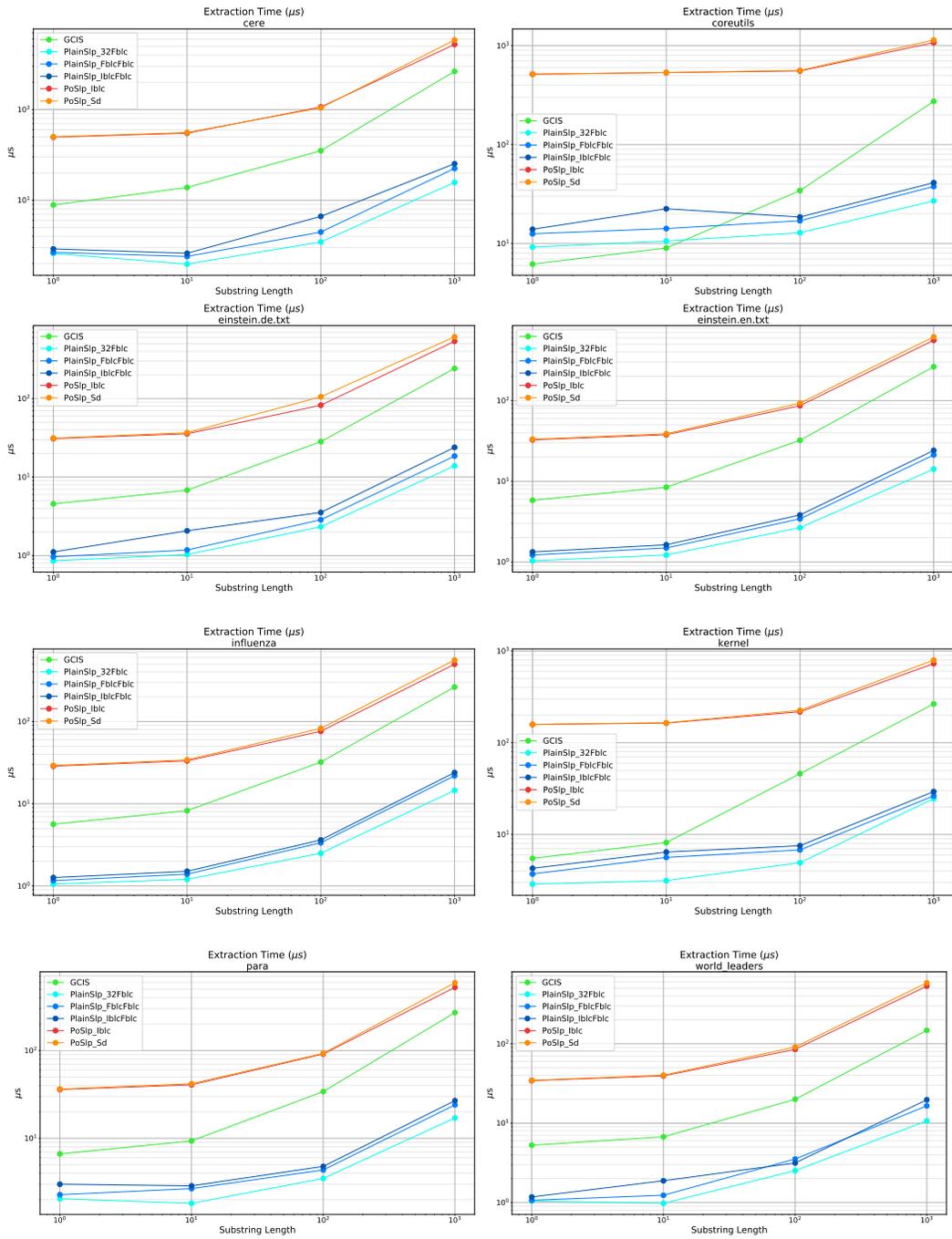


Fig. 22. Substring length vs. extraction time (microseconds) on repetitive texts.

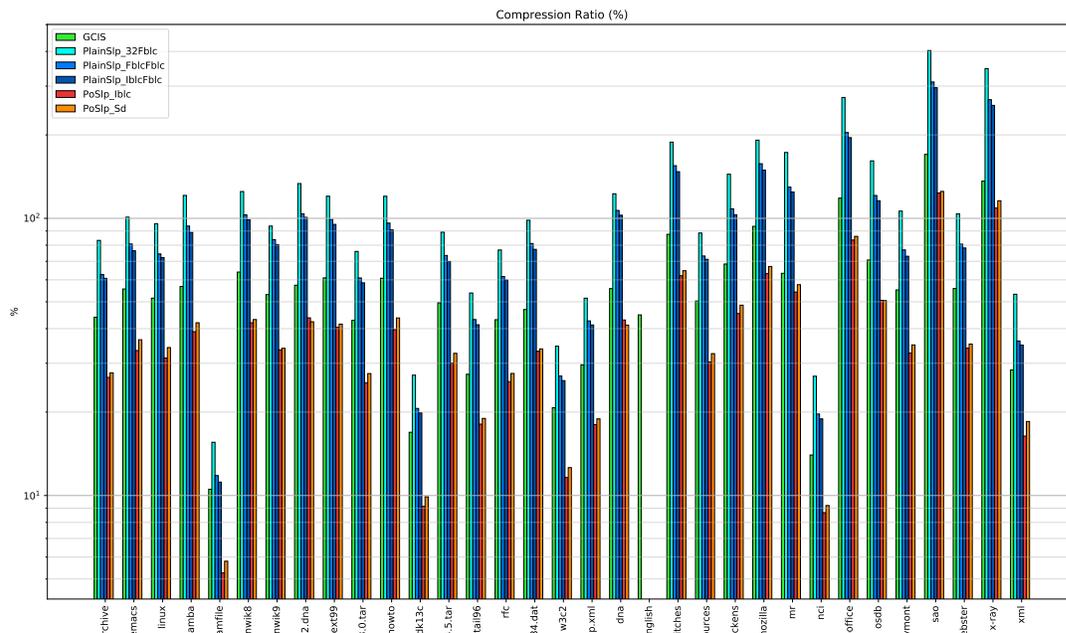


Fig. 23. Compression ratio of the extractors on regular texts.

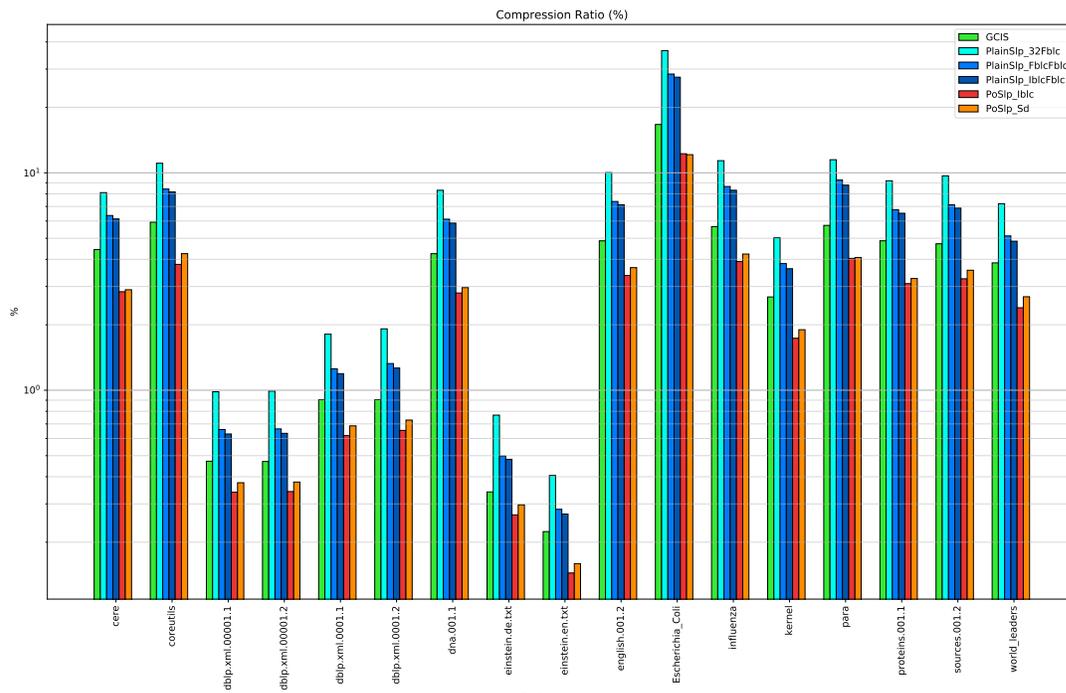


Fig. 24. Compression ratio of the extractors on repetitive texts.

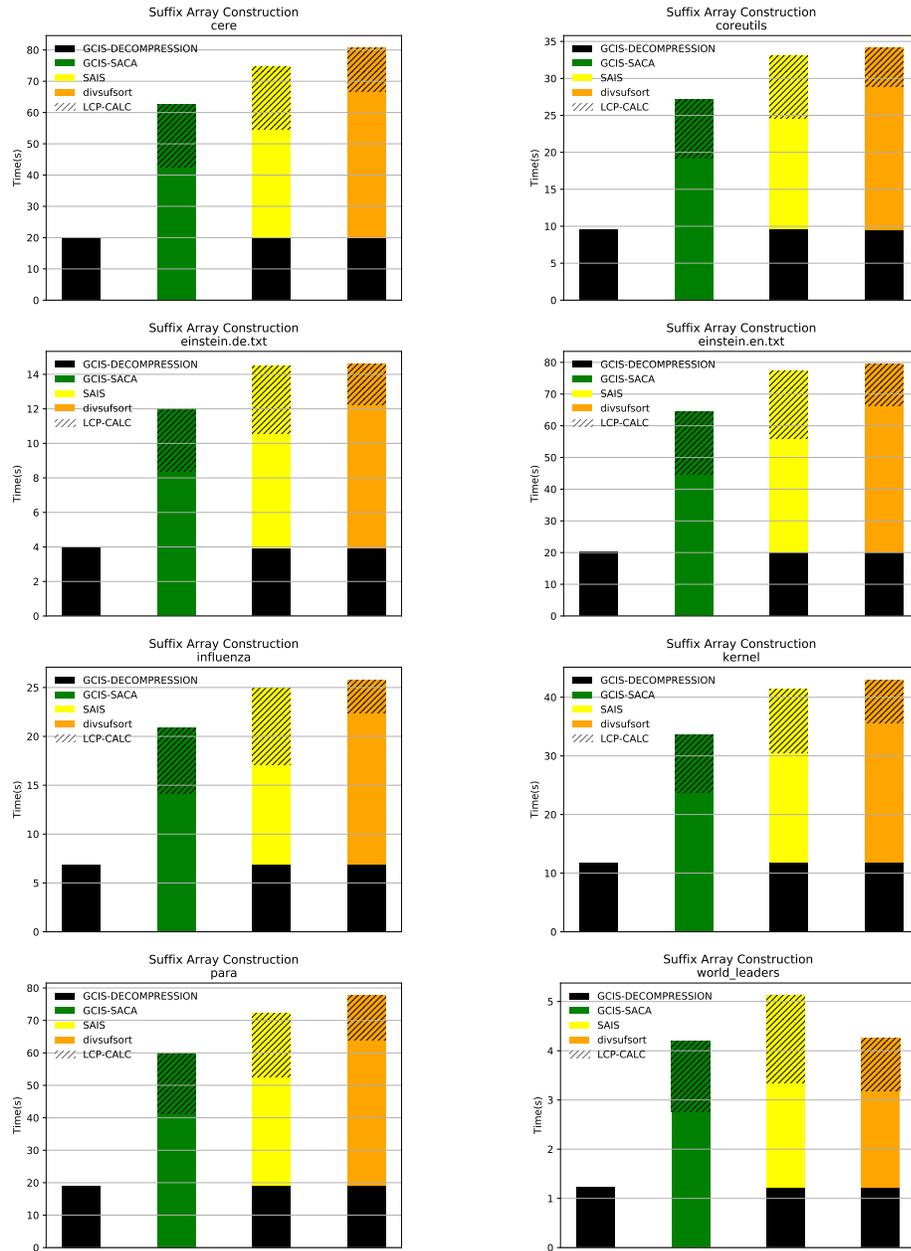


Fig. 25. Time consumed during Suffix Array and LCP construction.

REFERENCES

- Tooru Akagi, Dominik Köppl, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. 2021. Grammar Index By Induced Suffix Sorting. *CoRR* abs/2105.13744 (2021). arXiv:2105.13744 <https://arxiv.org/abs/2105.13744>
- Vo Ngoc Anh and Alistair Moffat. 2010. Index compression using 64-bit words. *Softw., Pract. Exper.* 40, 2 (2010), 131–147.
- Philip Bille, Mikko Berggren Ettienné, Inge Li Gørtz, and Hjalte Wedel Vildhøj. 2018. Time-space trade-offs for Lempel-Ziv compressed indexing. *Theor. Comput. Sci.* 713 (2018), 66–77. <https://doi.org/10.1016/j.tcs.2017.12.021>
- Philip Bille, Gad M. Landau, Rajeev Raman, Kunihiko Sadakane, Srinivasa Rao Satti, and Oren Weimann. 2015. Random Access to Grammar-Compressed Strings and Trees. *SIAM J. Comput.* 44, 3 (2015), 513–539.
- Nieves R. Brisaboa, Susana Ladra, and Gonzalo Navarro. 2013. DACs: Bringing direct access to variable-length codes. *Inf. Process. Lett.* 49, 1 (2013), 392–404.
- Michael Burrows and David J. Wheeler. 1994. *A block-sorting lossless data compression algorithm*. Technical Report. Digital SRC Research Report. 1–18 pages.
- Moses Charikar, Eric Lehman, Ding Liu, Rina Panigrahy, Manoj Prabhakaran, Amit Sahai, and abhi shelat. 2005. The smallest grammar problem. *IEEE Transactions on Information Theory* 51, 7 (2005), 2554–2576.
- Francisco Claude and Gonzalo Navarro. 2010. Self-Indexed Grammar-Based Compression. *Fundamenta Informaticae* 111, 3 (2010), 313–337.
- Francisco Claude and Gonzalo Navarro. 2012. Improved Grammar-Based Compressed Indexes. In *Proc. 19th International Symposium on String Processing and Information Retrieval (SPIRE) (LNCS 7608)*. Springer, Heidelberg, 180–192.
- Genome Reference Consortium. 2009. Genome Reference Consortium Human Reference 37. <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/>. Accessed: 06/2020.
- Sebastian Deorowicz. 2003. Silesia Corpus. <http://sun.aei.polsl.pl/~sdeor/index.php?page=silesia>. Accessed: 06/2020.
- Jasbir Dhaliwal, Simon J. Puglisi, and Andrew Turpin. 2012. Trends in Suffix Sorting: A Survey of Low Memory Algorithms. In *Proc. Australasian Computer Science Conference (ACSC)*. Australian Computer Society Inc., Sydney, 91–98.
- D. Díaz-Domínguez and G. Navarro. 2021. A Grammar Compressor for Collections of Reads with Applications to the Construction of the BWT. In *Proc. 31th Data Compression Conference (DCC)*. 93–102.
- D. Díaz-Domínguez, G. Navarro, and A. Pacheco. 2021. An LMS-based Grammar Self-index with Local Consistency Properties. In *Proc. 28th International Symposium on String Processing and Information Retrieval (SPIRE)*. <https://users.dcc.uchile.cl/~gnavarro/ps/spire21.2.pdf> To appear.
- Paolo Ferragina and Gonzalo Navarro. 2005a. Pizza-Chili Corpus. <http://pizzachili.dcc.uchile.cl/texts.html>. Accessed: 06/2020.
- Paolo Ferragina and Gonzalo Navarro. 2005b. Pizza-Chili Repetitive Corpus. <http://pizzachili.dcc.uchile.cl/repcorpus.html>. Accessed: 06/2020.
- Johannes Fischer. 2011. Inducing the LCP-Array. In *Proc. Workshop on Algorithms and Data Structures (WADS) (Lecture Notes in Computer Science, Vol. 6844)*. Springer, Berlin, Heidelberg, 374–385.
- Johannes Fischer and Florian Kurpicz. 2017. Dismantling DivSufSort. In *Proceedings of the Prague Stringology Conference*. Department of Theoretical Computer Science, Faculty of Information Technology, Czech Technical University in Prague, Prague, 62–76.
- Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Louisa Seelbach Benkner, and Yoshimasa Takabatake. 2020. Practical Random Access to SLP-Compressed Texts. (2020). Accepted short paper, *SPIRE*.
- Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, and Yoshimasa Takabatake. 2019. Rpair: Scaling up RePair with Rsync. In *Proc. 26th International Symposium on String Processing and Information Retrieval (SPIRE) (Lecture Notes in Computer Science, Vol. 11811)*. Springer-Verlag, Berlin, Heidelberg, 35–44.
- Jean-Loup Gailly and Mark Adler. Accessed: 3/2017. The gzip home page. <http://www.gzip.org/>.
- Simon Gog, Timo Beller, Alistair Moffat, and Matthias Petri. 2014. From Theory to Practice: Plug and Play with Succinct Data Structures. In *Proc. Symposium on Experimental and Efficient Algorithms (SEA) (Lecture Notes in Computer Science, Vol. 8504)*. Springer, Cham., 326–337.
- Simon Gog and Enno Ohlebusch. 2011. Fast and Lightweight LCP-Array Construction Algorithms. In *Proc. Workshop on Algorithm Engineering and Experimentation (ALENEX)*. ACM Digital Library, US, 25–34.
- Gaston H. Gonnet, Ricardo A. Baeza-Yates, and Tim Snider. 1992. New Indices for Text: PAT Trees and PAT Arrays. In *Information Retrieval*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 66–82.
- Keisuke Goto and Hideo Bannai. 2014. Space Efficient Linear Time Lempel-Ziv Factorization for Small Alphabets. In *Proc. IEEE Data Compression Conference (DCC)*. IEEE, NY, 163–172.
- Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda. 2015. LZD Factorization: Simple and Practical Online Grammar Compression with Variable-to-Fixed Encoding. In *Combinatorial Pattern Matching - 26th Annual Symposium, CPM 2015, Ischia Island, Italy, June 29 - July 1, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9133)*. Springer, 219–230. https://doi.org/10.1007/978-3-319-19929-0_19
- Tomohiro I. 2020. Shaped SLP implementation. <https://github.com/itomomoti/ShapedSlp>. Accessed: 08/2020.
- Hideo Itoh and Hozumi Tanaka. 1999. An Efficient Method for in Memory Construction of Suffix Arrays. In *Proc. International Symposium on String Processing and Information Retrieval (SPIRE)*. IEEE, NY, 81–88.
- Juha Kärkkäinen, Dominik Kempa, and Simon J. Puglisi. 2013. Linear Time Lempel-Ziv Factorization: Simple, Fast, Small. In *Proc. Annual Symposium on Combinatorial Pattern Matching (CPM) (Lecture Notes in Computer Science, Vol. 7922)*. Springer, Berlin, Heidelberg, 189–200.

- Juha Kärkkäinen, Giovanni Manzini, and Simon J. Puglisi. 2009. Permuted Longest-Common-Prefix Array. In *Proc. Annual Symposium on Combinatorial Pattern Matching (CPM) (Lecture Notes in Computer Science, Vol. 5577)*. Springer, Berlin, Heidelberg, 181–192.
- John C. Kieffer and En-Hui Yang. 2000. Grammar-based codes: A new class of universal lossless source codes. *IEEE Transactions on Information Theory* 46, 3 (2000), 737–754.
- Pang Ko and Srinivas Aluru. 2003. Space efficient linear time construction of suffix arrays. In *Proc. Annual Symposium on Combinatorial Pattern Matching (CPM) (Lecture Notes in Computer Science, Vol. 2676)*. Springer, Berlin, Heidelberg, 200–210.
- Dmitry Kosolobov, Daniel Valenzuela, Gonzalo Navarro, and Simon J. Puglisi. 2020. Lempel-Ziv like parsing in small space. (2020). <https://doi.org/10.1007/s00453-020-00722-6> Available On line and to appear in *Algorithmica*.
- Florian Kurpicz. 2015. Sais-lite suffix and LCP arrays construction algorithm. <https://github.com/kurpicz/sais-lite-lcp>. Accessed: 06/2020.
- Florian Kurpicz. 2016. Divsufsort suffix and LCP arrays construction algorithm. <https://github.com/kurpicz/libdivsufsort>. Accessed: 06/2020.
- [N.] Jesper Larsson and Alistair Moffat. 1999. Offline Dictionary-Based Compression. In *Proc. IEEE Data Compression Conference (DCC)*. IEEE, NY, 296–305.
- Weijun Liu, Ge Nong, Wai Hong Chan, and Yi Wu. 2016. Improving a lightweight LZ77 computation algorithm for running faster. *Softw. Pract. Exp.* 46, 9 (2016), 1201–1217.
- Felipe A. Louza, Simon Gog, and Guilherme P. Telles. 2017a. Inducing enhanced suffix arrays for string collections. *Theor. Comput. Sci.* 678 (2017), 22–39.
- Felipe A. Louza, Simon Gog, and Guilherme P. Telles. 2017b. Optimal suffix sorting and LCP array construction for constant alphabets. *Inf. Process. Lett.* 118 (2017), 30–34.
- Matt Mahoney. 2006. Large Text Compression Benchmark. <http://mattmahoney.net/dc/text.html>. Accessed: 06/2020.
- Udi Manber and Eugene W. Myers. 1993. Suffix Arrays: A New Method for On-Line String Searches. *SIAM J. Comput.* 22, 5 (1993), 935–948.
- Giovanni Manzini. 2003. Manzini’s Lightweight Corpus. <http://people.unipmn.it/~manzini/lightweight/>. Accessed: 06/2020.
- Shirou Maruyama and Yasuo Tabei. 2013. Fully Online Grammar Compression in Constant Space. In *Proc. IEEE Data Compression Conference (DCC)*. IEEE, NY, 173–182. <https://doi.org/10.1109/DCC.2014.69>
- Yuta Mori. 2008. Divsufsort suffix array construction algorithm. <https://github.com/y-256/libdivsufsort>. Accessed: 06/2020.
- Yuta Mori. 2010. Sais-lite suffix sorting algorithm. <https://sites.google.com/site/yuta256/sais>. Accessed: 06/2020.
- NCBI. 2007. Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7, complete sequence. https://www.ncbi.nlm.nih.gov/nucleotide/NC_010102. Accessed: 06/2020, NCBI Reference Sequence: NC 010102.1.
- NCBI. 2020. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512. Accessed: 06/2020, NCBI Reference Sequence: NC 045512.2.
- Takaaki Nishimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. 2020a. Dynamic index and LZ factorization in compressed space. *Discret. Appl. Math.* 274 (2020), 116–129. <https://doi.org/10.1016/j.dam.2019.01.014>
- Takaaki Nishimoto and Yasuo Tabei. 2019. LZRR: LZ77 Parsing with Right Reference. In *Proc. IEEE Data Compression Conference (DCC)*. IEEE, 211–220. <https://doi.org/10.1109/DCC.2019.00029>
- Takaaki Nishimoto, Yoshimasa Takabatake, and Yasuo Tabei. 2020b. A compressed dynamic self-index for highly repetitive text collections. *Inf. Comput.* 273 (2020), 104518. <https://doi.org/10.1016/j.ic.2020.104518>
- Ge Nong. 2013. Practical linear-time $O(1)$ -workspace suffix sorting for constant alphabets. *ACM Trans. Inform. Syst.* 31, 3 (2013), 1–15.
- Ge Nong, Sen Zhang, and Wai H. Chan. 2009. Linear Suffix Array Construction by Almost Pure Induced-Sorting. In *Proc. IEEE Data Compression Conference (DCC)*. IEEE, NY, 193–202.
- Ge Nong, Sen Zhang, and Wai H. Chan. 2011. Two Efficient Algorithms for Linear Time Suffix Array Construction. *IEEE Trans. Comput.* 60, 10 (2011), 1471–1484.
- Daniel Saad Nogueira Nunes, Felipe Alves da Louza, Simon Gog, Mauricio Ayala-Rincón, and Gonzalo Navarro. 2018. A Grammar Compression Algorithm Based on Induced Suffix Sorting. In *Proc. IEEE Data Compression Conference (DCC)*. IEEE, NY, 42–51.
- Enno Ohlebusch and Simon Gog. 2011. Lempel-Ziv Factorization Revisited. In *Proc. Annual Symposium on Combinatorial Pattern Matching (CPM) (Lecture Notes in Computer Science, Vol. 6661)*. Springer, Berlin, Heidelberg, 15–26.
- Daisuke Okanohara and Kunihiko Sadakane. 2009. A Linear-Time Burrows-Wheeler Transform Using Induced Sorting. In *Proc. International Symposium on String Processing and Information Retrieval (SPIRE) (Lecture Notes in Computer Science, Vol. 5721)*. Springer, Berlin, Heidelberg, 90–101.
- Igor Pavlov. Accessed: 10/2017. The 7zip home page. <http://www.7-zip.org/>.
- Simon J. Puglisi, William F. Smyth, and Andrew H. Turpin. 2007. A taxonomy of suffix array construction algorithms. *ACM Comp. Surv.* 39, 2 (2007), 1–31.
- Julian Seward. 1996. The bzip home page. <http://www.bzip.org/>. Accessed: 3/2017.
- Dmitry Shkarin. 2006. PPMd algorithm variant j revision 1. <http://www.compression.ru/ds/>. Accessed: 07/2020.
- Yoshimasa Takabatake, Tomohiro I, and Hiroshi Sakamoto. 2017. A Space-Optimal Grammar Compression. In *25th Annual European Symposium on Algorithms, ESA 2017, September 4-6, 2017, Vienna, Austria (LIPIcs, Vol. 87)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 67:1–67:15. <https://doi.org/10.4230/LIPIcs.ESA.2017.67>
- Yoshimasa Takabatake, Yasuo Tabei, and Hiroshi Sakamoto. 2014. Improved ESP-index: A Practical Self-index for Highly Repetitive Texts. In *Experimental Algorithms - 13th International Symposium, SEA 2014, Copenhagen, Denmark, June 29 - July 1, 2014. Proceedings (Lecture Notes in Computer Science, Vol. 8504)*, Joachim Gudmundsson and Jyrki Katajainen (Eds.). Springer, 338–350. https://doi.org/10.1007/978-3-319-07959-2_29
- Yoshimasa Takabatake, Yasuo Tabei, and Hiroshi Sakamoto. 2015. Online Self-Indexed Grammar Compression. In *String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9309)*.

- Costas S. Iliopoulos, Simon J. Puglisi, and Emine Yilmaz (Eds.). Springer, 258–269. https://doi.org/10.1007/978-3-319-23826-5_25
- Andrew Trigell. 1998. Andrew Trigell’s Large Corpus. <https://www.samba.org/ftp/tridge/large-corpus/>. Accessed: 06/2020.
- Sebastiano Vigna. 2013. Quasi-succinct indices. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*. ACM Digital Library, US, 83–92.
- Raymond Wan. 2014. Offline Dictionary-based Compression (RePair, Recursive Pairing) . <https://github.com/rwanwork/Re-Pair>. Accessed: 06/2020.
- Wikipedia. 2019. Wikipedia’s Pages and Articles XML Dump. <http://wikipedia.c3sl.ufpr.br/enwiki/20191120/>. Accessed: 06/2020.
- Ian H Witten, Alistair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes (2nd Ed.): Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco CA.
- Jacob Ziv and Abraham Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* 23, 3 (1977), 337–343.