



HAL
open science

Substring Complexity in Sublinear Space

Giulia Bernardini, Gabriele Fici, Pawel Gawrychowski, Solon P Pissis

► **To cite this version:**

Giulia Bernardini, Gabriele Fici, Pawel Gawrychowski, Solon P Pissis. Substring Complexity in Sublinear Space. ISAAC 2023 - 34th International Symposium on Algorithms and Computation, Dec 2023, Kyoto, Japan. 10.4230/LIPIcs.ISAAC.2023.12 . hal-04385532

HAL Id: hal-04385532

<https://inria.hal.science/hal-04385532>

Submitted on 10 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License


Substring Complexity in Sublinear Space

Giulia Bernardini ✉ 

University of Trieste, Italy

Gabriele Fici ✉ 

Dipartimento di Matematica e Informatica, University of Palermo, Italy

Paweł Gawrychowski ✉ 

Institute of Computer Science, University of Wrocław, Poland

Solon P. Pissis ✉ 

CWI, Amsterdam, The Netherlands

Vrije Universiteit, Amsterdam, The Netherlands

Abstract

Shannon’s entropy is a definitive lower bound for statistical compression. Unfortunately, no such clear measure exists for the compressibility of repetitive strings. Thus, ad hoc measures are employed to estimate the repetitiveness of strings, e.g., the size z of the Lempel–Ziv parse or the number r of equal-letter runs of the Burrows–Wheeler transform. A more recent one is the size γ of a smallest string attractor. Let T be a string of length n . A string attractor of T is a set of positions of T capturing the occurrences of all the substrings of T . Unfortunately, Kempa and Prezza [STOC 2018] showed that computing γ is NP-hard. Kociumaka et al. [LATIN 2020] considered a new measure of compressibility that is based on the function $S_T(k)$ counting the number of distinct substrings of length k of T , also known as the *substring complexity* of T . This new measure is defined as $\delta = \sup\{S_T(k)/k, k \geq 1\}$ and lower bounds all the relevant ad hoc measures previously considered. In particular, $\delta \leq \gamma$ always holds and δ can be computed in $\mathcal{O}(n)$ time using $\Theta(n)$ working space. Kociumaka et al. showed that one can construct an $\mathcal{O}(\delta \log \frac{n}{\delta})$ -sized representation of T supporting efficient direct access and efficient pattern matching queries on T . Given that for highly compressible strings, δ is significantly smaller than n , it is natural to pose the following question:

Can we compute δ efficiently using sublinear working space?

It is straightforward to show that in the comparison model, any algorithm computing δ using $\mathcal{O}(b)$ space requires $\Omega(n^{2-o(1)}/b)$ time through a reduction from the element distinctness problem [Yao, SIAM J. Comput. 1994]. We thus wanted to investigate whether we can indeed match this lower bound. We address this algorithmic challenge by showing the following bounds to compute δ :

- $\mathcal{O}(\frac{n^3 \log b}{b^2})$ time using $\mathcal{O}(b)$ space, for any $b \in [1, n]$, in the comparison model.
- $\tilde{\mathcal{O}}(n^2/b)^1$ time using $\tilde{\mathcal{O}}(b)$ space, for any $b \in [\sqrt{n}, n]$, in the word RAM model. This gives an $\tilde{\mathcal{O}}(n^{1+\epsilon})$ -time and $\tilde{\mathcal{O}}(n^{1-\epsilon})$ -space algorithm to compute δ , for any $0 < \epsilon \leq 1/2$.

Let us remark that our algorithms compute $S_T(k)$, for all k , within the same complexities.

2012 ACM Subject Classification Theory of computation → Pattern matching

Keywords and phrases sublinear-space algorithm, string algorithm, substring complexity

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2023.12

Related Version *Full Version:* <https://arxiv.org/abs/2007.08357>

Funding *Giulia Bernardini:* MUR – FSE REACT EU – PON R&I 2014-2020.

Gabriele Fici: Projects MUR PRIN 2017 ADASCOML – 2017K7XPAN and MUR PRIN 2022 APML – 20229BCXNW.

Solon P. Pissis: Supported by the PANGAIA (No 872539) and ALPACA (No 956229) projects.

¹ The $\tilde{\mathcal{O}}(f)$ notation denotes $\mathcal{O}(f \cdot \text{polylog}(n))$.



1 Introduction

We are currently witnessing our world drowning in data. These datasets are generated by a large gamut of applications: databases, web applications, genome sequencing projects, scientific computations, sensors, e-mail, entertainment, and others. The biggest challenge is thus to develop theoretical and practical methods for processing datasets efficiently.

Compressed data representations that can be *directly* used in compressed form have a central role in this challenge [61]. Indeed, much of the currently fastest-growing data is highly repetitive; this, in turn, enables space reductions of orders of magnitude [35]. Prominent examples of such data include genome, versioned text, and software repositories collections. A common characteristic is that each element in a collection is very similar to every other.

Since a significant amount of this data is sequential, a considerable amount of algorithmic research has been devoted to text indexes over the past decades [68, 59, 29, 45, 31, 42, 44, 6, 23, 60, 46, 35, 47]. String processing applications (see [43, 2] for reviews) require fast access to the substrings of the input string. These applications rely on such text indexes, which arrange the string suffixes lexicographically in an ordered tree [68] or an ordered array [59].

This significant amount of research has resulted in compressed text indexes that support fast pattern searching in space close to the statistical entropy of the text collection. The problem, however, is that this kind of entropy is unable to capture repetitiveness [57, 58]. To achieve orders-of-magnitude space reductions, one thus needs to resort to other compression methods, such as Lempel-Ziv (LZ) [71], grammar compression [50] or run-length compressed Burrows-Wheeler transform (BWT) [35], to name a few; see [35] for a review.

Unlike Shannon's entropy, which is a definitive lower bound for statistical compression, no such clear measure exists for the compressibility of repetitive texts. Other than Kolmogorov's complexity [55], which is not computable, repetitiveness is measured in ad hoc terms, based on what the compressors may achieve. Such measures on a string T include: the number z of phrases produced by the LZ parsing of T ; the size g of the smallest grammar generating T ; and the number r of maximal equal-letter runs in the BWT of T . See [62] for a survey.

An improvement is the recent introduction of the *string attractor* [49] notion. Let T be a string of length n . An attractor Γ is a set of positions over $[1, n]$ such that any substring of T has an occurrence covering a position in Γ . The size γ of a smallest attractor asymptotically lower bounds all the repetitiveness measures listed above (and others; see [52]). Unfortunately, using indexes based on γ comes also with some challenges. Other than computing γ is NP-hard [49], it is unclear if γ is the definitive measure of repetitiveness: we do not know whether one can always represent T in $\mathcal{O}(\gamma)$ space (machine words). This motivated Christiansen et al. [18] to consider a new measure δ of compressibility, initially introduced in the area of string compression by Raskhodnikova et al. [66], and for which $\delta \leq \gamma$ *always holds* [18].

► **Definition 1** ([18]). *Let T be a string and $S_T(k)$ its substring complexity: the function counting the number of distinct substrings of length k of T . The normalized substring complexity of T is the function $S_T(k)/k$ and we set $\delta = \sup\{S_T(k)/k, k \geq 1\}$ its supremum.*

Christiansen et al. also showed that δ can be computed in $\mathcal{O}(n)$ time using $\Theta(n)$ working space. Kociumaka et al. [52, 53] showed that δ can also be strictly smaller than γ by up to a logarithmic factor: for any n and any δ , there are strings with $\gamma = \Omega(\delta \log \frac{n}{\delta})$. Moreover, Kociumaka et al. developed a representation of T of size $\mathcal{O}(\delta \log \frac{n}{\delta})$, which is worst-case optimal in terms of δ and allows for accessing any $T[i]$ in time $\mathcal{O}(\log \frac{n}{\delta})$ and for finding all *occ* occurrences of any pattern $P[1..m]$ in T in near-optimal time $\mathcal{O}(m \log n + \text{occ} \log^\epsilon n)$, for any constant $\epsilon > 0$ (see also [51] and [48] for further improvements). Since for highly compressible strings, δ is significantly smaller than n , we pose the following basic question:

Can we compute δ efficiently using sublinear working space?

The question on computing δ in *bounded space* arises naturally: it extends a large body of work on problems on strings, which admit a straightforward solution if we have the space to construct and store the suffix tree [68]; but as this is often not the case, one needs to overcome the space challenge by investigating space-time trade-offs for these problems.

Related Work. The standard approach for showing space-time trade-off lower bounds for problems answered in polynomial time has been to analyze their complexity on (*multi-way branching programs*). In this model, the input is stored in read-only memory, the output in write-only memory, and neither is counted towards the space used by any algorithm. This model is powerful enough to simulate both Turing machines and standard RAM models that are unit-cost with respect to time and log-cost with respect to space. It was introduced by Borodin and Cook, who used it to prove that any multi-way branching program requires a time-space product of $\Omega(n^2/\log n)$ to sort n integers in the range $[1, n^2]$ [11, 4]. Unfortunately, the techniques in [11] yield only trivial bounds for problems with single outputs.

String algorithms that use sublinear space have been extensively studied over the past decades [36, 26, 64, 13, 67, 12, 54, 19, 34, 20, 22, 41, 40, 39, 38, 21, 37, 3, 63, 15, 65, 56]. The perhaps most relevant problem to our work is the classic longest common substring of two strings. Formally, given two strings X and Y of total length n , the *longest common substring* (LCS) problem consists in computing a longest string occurring as a substring of both X and Y . The LCS problem was conjectured by Knuth to require $\Omega(n \log n)$ time. This conjecture was disproved by Weiner who, in his seminal paper on suffix tree construction [68], showed how to solve the LCS problem in $\mathcal{O}(n)$ time for constant-sized alphabets. Farach showed that the same problem can be solved in the optimal $\mathcal{O}(n)$ time for polynomially-sized integer alphabets [29]. A straightforward space-time trade-off lower bound of $\Omega(b)$ space and $\Omega(n^2/b)$ time for the LCS problem can be derived from the problem of checking whether the length of an LCS is 0; i.e., deciding if X and Y have a common letter or not. Thus, in some sense, the LCS problem can be seen as a generalization of the *element distinctness* problem: given n elements over a domain D , decide whether all n elements are distinct.

On the upper bound side, Starikovskaya and Vildhøj showed that for any $b \in [n^{2/3}, n]$, the LCS problem can be solved in $\tilde{\mathcal{O}}(n^2/b)$ time and $\mathcal{O}(b)$ space [67]. In [54], Kociumaka et al. gave an $\mathcal{O}(n^2/b)$ -time algorithm to find an LCS for any $b \in [1, n]$, and also provided a lower bound, which states that any deterministic multi-way branching program that uses $b \leq \frac{n}{\log n}$ space must take $\Omega(n\sqrt{\log(n/(b \log n))}/\log \log(n/(b \log n)))$ time. This lower bound implies that the classic $\mathcal{O}(n)$ -time solution for the LCS problem [68, 29] is optimal in the sense that we cannot hope for an $\mathcal{O}(n)$ -time algorithm using $o(n/\log n)$ space. Unfortunately, we do not know if the $\mathcal{O}(b)$ -space and $\mathcal{O}(n^2/b)$ -time trade-off is generally the best possible for the LCS problem. For the easier element distinctness problem, Beame et al. [5] showed a randomized multiway branching program using $\tilde{\mathcal{O}}(n^{3/2}/\sqrt{b})$ -time and $\mathcal{O}(b)$ space.

It is thus a big open question to answer whether the LCS problem can be solved asymptotically faster than $\mathcal{O}(n^2/b)$ using $\mathcal{O}(b)$ space. Towards this direction, Ben-Nun et al. exploited the intuition suggesting that an LCS of X and Y can be computed more efficiently when its length L is large [63] (see also [16]). The authors showed an algorithm which runs in $\tilde{\mathcal{O}}(\frac{n^2}{L \cdot b} + n)$ time, for any $b \in [1, n]$, using $\mathcal{O}(b)$ space. Still, a straightforward lower bound for the aforementioned problem is in $\Omega(\frac{n^2}{L \cdot b} + n)$ time when $\mathcal{O}(b)$ space is used; it seems that further insight is required to match this space-time trade-off lower bound.

Our Results and Techniques. Our goal is to efficiently compute δ using $\mathcal{O}(b)$ space. As a preliminary step towards this algorithmic challenge, we show the following theorem.

► **Theorem 2.** *Given a string T of length n , we can compute $\delta = \sup\{S_T(k)/k, k \geq 1\}$ in $\mathcal{O}(\frac{n^3 \log b}{b^2})$ time using $\mathcal{O}(b)$ space, for any $b \in [1, n]$, in the comparison model.*

It is straightforward to show that any comparison-based branching program to compute δ using $\mathcal{O}(b)$ space requires $\Omega(n^{2-o(1)}/b)$ time through a reduction from the element distinctness problem [70]. By Yao's lemma, this lower bound also applies to randomized branching programs [70]. This suggests that a natural intermediate step towards fully understanding the computation complexity of computing δ in small space should be designing an $\tilde{\mathcal{O}}(n^2/b)$ -time algorithm using $\mathcal{O}(b)$ space (not necessarily in the comparison model).

The natural approach for computing δ is through computing all values of $S_T(k)$. In particular, this is the idea behind the straightforward $\mathcal{O}(n)$ -time computation of δ using $\mathcal{O}(n)$ space [18]. It is unclear to us if a more direct approach exists (see also Section 6 for a combinatorial analysis on the behaviour of δ). Under this plausible assumption, we stress that computing $S_T(k)$, for all k one-by-one, is a more general problem than computing the length L of an LCS of X and Y , as an algorithm computing $S_T(k)$ can be used to compute L within the same complexities. This follows by the following argument: we compute $S_X(k)$, $S_Y(k)$, and $S_{X\#Y}(k)$ (where $\#$ is a special letter that does not occur in X or in Y) in parallel, and set L equal to the largest k such that $S_X(k) + S_Y(k) > S_{X\#Y}(k) - k$. As the best-known time upper bound for the very basic question of computing LCS in $\mathcal{O}(b)$ space remains to be $\mathcal{O}(n^2/b)$, this further motivates the algorithmic challenge of designing an algorithm with such bounds for computing δ . We address it by proving the following theorem.

► **Theorem 3.** *Given a string T of length n , we can compute $\delta = \sup\{S_T(k)/k, k \geq 1\}$ in $\tilde{\mathcal{O}}(n^2/b)$ time using $\tilde{\mathcal{O}}(b)$ space, for any $b \in [\sqrt{n}, n]$, in the word RAM model.*

Our algorithms compute $S_T(k)$, for all k , within the same complexities. To arrive at the $\mathcal{O}(\frac{n^3 \log b}{b^2})$ -time bound, we split the computation of the values $S_T(k)$ in n/b phases: in each phase, we restrict to substrings whose length is in a range of size b . In turn, in each phase, we process the substrings that start within a range of b positions of T at a time, from left to right. With this scheme, we process in $\mathcal{O}(n \log b)$ time each block of n/b positions of T in each of the n/b phases, resulting in $\mathcal{O}(\frac{n^3 \log b}{b^2})$ time using $\mathcal{O}(b)$ space. For large enough b , we can process all the substrings of a single phase at once, saving a factor of n/b . We show in fact that a representation of all the occurrences of all the substrings of a phase can be packed in $\tilde{\mathcal{O}}(b)$ space if b is large enough, and process them in different ways depending on their period, following a scheme similar to [8]; we also adapt a method used in [9] to select a small set of anchors (length- b substrings), so that each fragment of T contains at least one anchor but their total number of occurrences in T is bounded. Note that Theorem 3 implies an $\tilde{\mathcal{O}}(n^{1+\epsilon})$ -time and $\tilde{\mathcal{O}}(n^{1-\epsilon})$ -space algorithm to compute δ , for any $0 < \epsilon \leq 1/2$.

Paper Organization. Section 2 introduces the basic definitions and notation we use and the space-time trade-off lower bound for computing δ . In Section 3, we present a simple $\mathcal{O}(n^3/b)$ -time and $\mathcal{O}(b)$ -space algorithm, for any $b \in [1, n]$. This algorithm is refined to run in $\mathcal{O}(\frac{n^3 \log b}{b^2})$ time using $\mathcal{O}(b)$ space, for any $b \in [1, n]$, in Section 4. Our main result, the $\tilde{\mathcal{O}}(n^2/b)$ -time and $\tilde{\mathcal{O}}(b)$ -space algorithm, for any $b \in [\sqrt{n}, n]$, is presented in Section 5. In Section 6, we consider the notion of substring complexity from the combinatorial point of view; and in Section 7, we conclude this paper with a final remark on approximating δ .

2 Preliminaries

An *alphabet* Σ is a finite nonempty set of elements called *letters*. We fix throughout a *string* $T = T[1] \cdots T[n]$ of *length* $|T| = n$ over an ordered alphabet Σ . By ε we denote the *empty string* of length 0. For two indices $1 \leq i \leq j \leq n$, the (i, j) -*fragment* of T is an *occurrence* of the underlying *substring* $T[i..j] = T[i] \cdots T[j]$. A *prefix* of T is a fragment of T of the form $T[1..j]$ and a *suffix* of T is a fragment of T of the form $T[i..n]$. A prefix (resp. suffix) of T is *proper* if it is not equal to T . We let $T^r = T[n]T[n-1] \cdots T[1]$ denote the *reversal* of T .

A positive integer p is a *period* of a string T if $T[i] = T[j]$ whenever $i = j \pmod{p}$; we call *the period* of T , denoted by $\text{per}(T)$, the smallest such p . A string T is said to be *strongly periodic* if $\text{per}(T) \leq |T|/4$ and *periodic* if $\text{per}(T) \leq |T|/2$. We call the lexicographically smallest cyclic shift of $T[1.. \text{per}(T)]$ the (*Lyndon*) *root* of T . Notice that if T is periodic, then the root of T is always a fragment of T (that is, it has an occurrence in T).

For every string t and every natural number ℓ , we define the ℓ th *power* of t , denoted by t^ℓ , by $t^0 = \varepsilon$ and $t^k = t^{k-1}t$, for integer $k = [1, \ell]$. A *run* with (Lyndon) root t in a string T is a periodic fragment $T[i..j] = t[q..|t|]t^\beta t[1..\gamma]$, with $q, \gamma \in [1, |t|]$ and β a positive integer, such that both $T[i-1..j]$ and $T[i..j+1]$, if defined, have their smallest period larger than $|t|$; we say that $q \in [1, |t|]$ is the *offset* of the run $t[q..|t|]t^\beta t[1..\gamma]$ and that two runs with the same root are *synchronized* if they have the same offset. We represent a run $t[q..|t|]t^\beta t[1..\gamma]$ by its starting and ending positions (i, j) in T , its root t , and its offset q .

The *element distinctness* problem asks to determine if all the elements of an array A of size n are pairwise distinct. Yao showed that, in the comparison-based branching program model, the time required to solve the element distinctness problem using $\mathcal{O}(b)$ space is in $\Omega(n^{2-o(1)}/b)$ [70]. We show the following lower bound for computing δ in the same model.

► **Theorem 4.** *The time required to compute δ for a string T of length n using $\mathcal{O}(b)$ space in the comparison model is in $\Omega(n^{2-o(1)}/b)$.*

Proof. We reduce the element distinctness problem to computing δ in $\mathcal{O}(n)$ time as follows. Let A be the input array for the element distinctness problem. Further let $\#_1, \#_2, \dots, \#_n$ be pairwise distinct elements not occurring in A . We set $T = A \cdot \#_1 \#_2 \dots \#_n$, with $|T| = 2n$, $\#_i \neq A[j]$, for all $i, j \in [1, n]$. Observe that $S_T(k)/k < n$, for all $k \geq 2$, and thus $\delta = S_T(1) = n + |\{A\}|$. Then A has a repeating element if and only if $\delta < 2n$. ◀

3 $\mathcal{O}(n^3/b)$ Time Using $\mathcal{O}(b)$ Space in the Comparison Model

We start with a warm-up lemma to guide the reader smoothly to the $\mathcal{O}(n^3/b)$ -time algorithm.

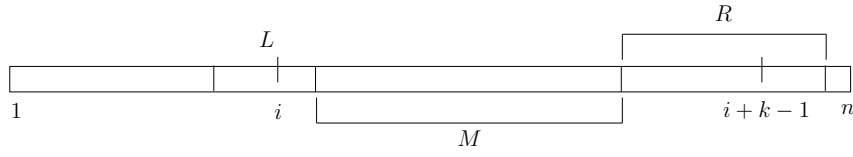
► **Lemma 5.** *Given a string T of length n , we can compute $\delta = \sup\{S_T(k)/k, k \geq 1\}$ in $\mathcal{O}(n^3)$ time using $\mathcal{O}(1)$ space in the comparison model.*

Proof. Let us consider each $S_T(k)$ separately, for all $k \in [1, n]$.

Set $S_T(k) = 0$. For all $i \in [1, n]$, we increase $S_T(k)$ if $T[i..i+k-1]$ is the first occurrence in $T[1..i+k-1]$. To perform this we check whether $T[j..j+k-1] = T[i..i+k-1]$, for all $j \in [1, i-1]$. We employ any linear-time constant-space pattern matching algorithm [36, 26, 13] to do this check in $\mathcal{O}(n)$ time using $\mathcal{O}(1)$ space for a single i . The statement follows. ◀

We next generalize Lemma 5 by employing the following straightforward observation.

► **Observation 6.** *Let S be a substring of T . If S occurs at least twice in T , then every substring of S occurs at least twice in T ; if S occurs only once in T , then any substring of T containing S as a substring occurs only once in T .*



■ **Figure 1** The main setting of the algorithm underlying Proposition 7.

Main Idea. Recall that we have $\mathcal{O}(b)$ budget for space. At any phase of the algorithm, we maintain $S_T(k)$ for b values of k , and iterate on consecutive non-overlapping substrings of T of length b , which we call *blocks*. This gives n/b phases and n/b iterations per phase, respectively. For each iteration, we define a substring M of T , which we call *anchor*. We search for occurrences of this anchor in T and extend each of the (at most) n occurrences of M in $\mathcal{O}(b)$ time per occurrence. This gives $\mathcal{O}(n^3/b)$ time and $\mathcal{O}(b)$ space.

► **Proposition 7.** *Given a string T of length n , we can compute $\delta = \sup\{S_T(k)/k, k \geq 1\}$ in $\mathcal{O}(n^3/b)$ time using $\mathcal{O}(b)$ space, for any $b \in [1, n]$, in the comparison model.*

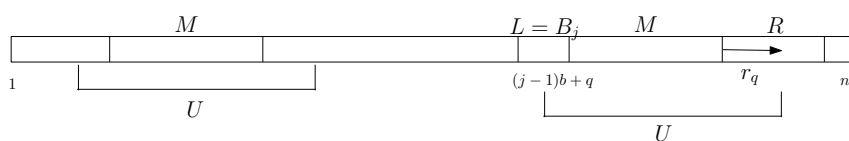
Proof. Our algorithm consists of $n/b - 2$ phases. In phase α , for all $\alpha \in [2, 3, \dots, n/b - 1]$,² we compute altogether the b values of $S_T(k)$, for all $k \in [\alpha b + 1, (\alpha + 1)b]$. Let $\mathcal{S} = \mathcal{S}[1..b]$ be an array of size b where we store the values of $S_T(k)$ corresponding to phase α : $\mathcal{S}[h] = S_T(\alpha b + h)$, for $h \in [1, b]$. At the end of phase α we maintain the maximum of $\mathcal{S}[h]/(\alpha b + h)$. Clearly, at the end of the whole procedure, we can output $\delta = \sup\{S_T(k)/k \mid k \geq 1\}$.

We start by decomposing T into n/b blocks $B_1, B_2, \dots, B_{n/b}$, each of length b . We next describe our algorithm for a fixed phase $\alpha > 1$. First we set $\mathcal{S}[h] = 0$, for all $h \in [1, b]$. Let i be a position on T . For each k in the range of α , we want to know if $T[i..i+k-1]$ has its first occurrence in T at position i or if it occurs also at some position to the left of i . We process together all positions i in the same block $B_j = T[(j-1)b + 1..jb]$, for every $j \in [1, n/b]$. Let $L = B_j$ be the block we are currently processing (inspect also Figure 1). To compute $S_T(k)$ we consider, for all $k \in [\alpha b + 1, (\alpha + 1)b]$, the length- k fragments with starting position i in L . All such fragments share the same *anchor* $M = T[jb + 1..(j + \alpha - 1)b]$. The fragment of length k ends at position $i + k - 1$, which belongs to one of the *two blocks* succeeding M for all $k \in [\alpha b + 1, (\alpha + 1)b]$; we denote the concatenation of these two succeeding blocks as fragment R . In particular, we have $|M| = (\alpha - 1)b$ and $R = B_{j+\alpha}B_{j+\alpha+1}$.

We will use the occurrences of M in T that start before its *starting position* $jb + 1$ as anchors for finding possible occurrences of the length- k fragments starting within L . We search for such occurrences of M with any linear-time constant-space pattern matching algorithm [36, 26, 13]. For each such occurrence of M , we then need to check the b letters preceding it and the $2b - 1$ letters following it in order to determine whether it generates a previous occurrence of some $(i, i + k - 1)$ -fragment, where i is a position within the block L . In particular, we check the b letters preceding it because L is the block of b positions preceding M ; we check the $2b - 1$ letters following it because $k \leq (\alpha + 1)b$.

While processing $L = B_j = T[(j-1)b + 1..jb]$, we also maintain an array $\text{END}_L[1..b]$ of size b . After we have finished processing L , $\text{END}_L[q]$ will store the length r_q of the longest prefix of R such that $T[(j-1)b + q..(j-1)b + q + |M| + r_q - 1]$ occurs in T before

² We process the substrings of length $k \in [1, 2b]$ separately: for each block B_j , we compute an array LS_j of size b such that $\text{LS}_j[q]$ is the length the *longest substring of length up to $2b$* starting at position q in B_j that occurs in T before position $(j-1)b + q$. This is done as described in the proof of Lemma 8 and requires $\mathcal{O}(\frac{n^2 \log b}{b})$ total time. At the end of this procedure, we just maintain $\max\{S_T(k) \mid k \in [1, 2b]\}$.



■ **Figure 2** Largest r_q such that $U = T[(j-1)b+q..(j-1)b+q+|M|+r_q-1]$.

position $(j-1)b+q$ (inspect Figure 2). We compute END_L as follows. We search for all the occurrences of $M = T[jb+1..(j+\alpha-1)b]$ in $T[1..(j+\alpha-1)b-1]$, from left to right. Let $M = T[i'..i'+|M|-1]$ be one such occurrence. Let ℓ be the length of the longest common suffix of L and $T[1..i'-1]$; let r be the length of the longest common prefix of R and $T[i'+|M|..n]$. For each $q \geq b-\ell+1$, we update $\text{END}_L[q]$ with the maximum between its previous value $\text{END}_L[q]$ and r (note that we do not update any values if $\ell = 0$). After we have processed all the occurrences of M , for each q we increase by 1 all $S_T(k)$ such that $k > b-q+1+|M|+\text{END}_L[q] = b-q+1+(\alpha-1)b+\text{END}_L[q] = \alpha b-q+\text{END}_L[q]+1$. This is an application of Observation 6: all these occurrences correspond to a substring that is longer than a substring that occurs for the first time in T at position $(j-1)b+q$.

The whole algorithm takes time $\mathcal{O}(\frac{n^3}{b})$: there are $\frac{n}{b}$ phases; in each phase, we consider $\frac{n}{b}$ blocks and for each block we spend $\mathcal{O}(n)$ time for pattern matching anchor M ; for each occurrence of the anchor, we spend $\mathcal{O}(b)$ time for finding and updating the possible extensions, thus $\mathcal{O}(nb)$ time overall. We finally need $\mathcal{O}(b^2)$ time for updating the values of $S_T(k)$ for all k 's in the range and all positions i in L . Overall this is $\mathcal{O}(\frac{n}{b} \cdot \frac{n}{b} \cdot (nb+b^2)) = \mathcal{O}(\frac{n^3}{b})$ time. ◀

4 $\mathcal{O}(\frac{n^3 \log b}{b^2})$ Time Using $\mathcal{O}(b)$ Space in the Comparison Model

Recall that in Proposition 7, we spend $\mathcal{O}(nb+b^2)$ time to process the at most n occurrences of a single anchor M in T . We show here that all these occurrences can be processed in $\mathcal{O}(n \log b)$ time. This is made possible by processing together batches of occurrences of M that are *close enough* in T . This is done by means of answering longest common extension queries on suffix trees constructed for certain length- $\mathcal{O}(b)$ fragments of T .

The first trick is based on the following remark: The pattern matching algorithm for reporting the occurrences of M (e.g., [13]) reports the occurrences of M in real-time from left to right. Every such occurrence m of M is preceded by a block L' of length b on the left of m starting at position $m-b$ and ending at position $m-1$, and it is succeeded by a fragment R' of length $2b$ starting at position $m+|M|$ and ending at position $m+|M|+2b-1$. We thus need to find the longest common prefix of R and R' and the longest common suffix of L and L' . Let us describe the process for the longest common prefix of R and R' . (The procedure for the longest common suffix of L and L' is analogous and is executed simultaneously.)

We use the so-called *standard trick* to construct a sequence of $n/(4b)$ suffix trees for fragments of T of length $4b$ overlapping by $2b$ positions. We first concatenate each such fragment of length $4b$ with R . Constructing one such suffix tree takes $\mathcal{O}(b \log b)$ time using $\mathcal{O}(b)$ space [68]. Recall that an occurrence m of M implies an occurrence of R' at position $m+|M|$ and thus this position is part of some fragment of length $4b$. We preprocess this suffix tree in $\mathcal{O}(b)$ time and space to answer longest common prefix queries in $\mathcal{O}(1)$ time [7]. The whole preprocessing thus takes $n/(4b)\mathcal{O}(b \log b) = \mathcal{O}(n \log b)$ time. Thus, for any occurrence m of M we can find the longest right extension (and the longest left extension with a similar procedure) in $\mathcal{O}(1)$ time; recall that each extension cannot be of length greater than $2b$ so we do not miss any of them. To memorize the extensions we use an array END_L of size b .

For each occurrence of M , if we have a left extension of length $\ell > 0$ and a right extension of length r , we set $\text{END}_L[b - \ell + 1] = \max\{\text{END}_L[b - \ell + 1], r\}$ in $\mathcal{O}(1)$ time. At the end of this process we sweep through END_L and set $\text{END}_L[q] = \max\{\text{END}_L[q - 1], \text{END}_L[q]\}$, for all $i \in [2, b]$ by Observation 6: if we can extend a position q in L r positions to the right of M , then we must be able to extend position $q + 1$ in L at least r positions to the right of M .

The second trick updates all values of $S_T(k)$ using array END_L in $\mathcal{O}(b)$ time instead of $\mathcal{O}(b^2)$ time. We use an array I of size b with all its entries initialized to 0; $I[h]$ will store the number of positions q in L such that the shortest unique substring starting at q is of length $\alpha b + h$. We fill in I scanning END_L : the shortest unique substring starting at q is by definition of length $\alpha b - q + \text{END}_L[q] + 2$, which equals $\alpha b + h$ when $h = \text{END}_L[q] - q + 2$. We thus increment $I[h]$ by one. We finally increase $S_T(\alpha b + h)$ by $\sum_{j=1}^h I[j]$ for all $h = 1, \dots, b$. Thus, updating all values of $S_T(k)$ is implemented in $\mathcal{O}(b)$ time. We have arrived at Theorem 2.

5 $\tilde{\mathcal{O}}(\frac{n^2}{b})$ Time Using $\tilde{\mathcal{O}}(b)$ Space for $b \geq \sqrt{n}$ in the word RAM model

The algorithm underlying Theorem 2 is organized in n/b phases. In phase α we process b values of $S_T(k)$ making use of evenly-spaced fragments of T , each of length $(\alpha - 1)b$, as *anchors* for finding possible multiple occurrences of the length- k fragments of T . Considering $\mathcal{O}(n/b)$ anchors in each phase and processing them one by one is the bottleneck of this algorithm. Our approach here is thus to avoid the burden of considering new anchors at every phase by carefully selecting a set of anchors that will remain *unchanged* in each phase of the algorithm. Let $c > 1$ be any integer constant. We will process the values of $S_T(k)$ for $k \leq cb$ (Section 5.1) and for $k > cb$ (Sections 5.2 to 5.5) in two different ways.

We work in the word RAM model and our goal is a deterministic algorithm. Recall that a suffix tree of any string of length d can be constructed in $\tilde{\mathcal{O}}(d)$ time using $\mathcal{O}(d)$ space [68, 29].

5.1 Computing $S_T(k)$ for Small k

We process together all values $k \in [1, cb]$. Like in Sections 3-4, for such values of k we split T into n/b blocks of b positions and work with each such block separately; we compute all values $S_T(k)$ and keep track of $\max_{k \leq cb} S_T(k)/k$ before computing $S_T(k)$ for all $k > cb$.

Consider block $L = B_j = T[(j - 1)b + 1 .. jb]$. We compute an array LS_L of size b such that $\text{LS}_L[q]$ is the length the *longest substring* starting at position q in L that occurs in T before position $(j - 1)b + q$, if this length does not exceed cb , otherwise we set it to ∞ . This is done by constructing multiple generalized suffix trees of windows of length $2cb$ and L .

► **Lemma 8.** $\max_{k \leq cb} \frac{S_T(k)}{k}$ can be computed in $\tilde{\mathcal{O}}(n^2/b)$ time and $\mathcal{O}(b)$ space, for any $b \in [1, n]$.

Proof. We consider a block $L = T[(j - 1)b + 1 .. jb]$ of b positions of T at a time; for each position i of T within L , we must compute the length of the longest fragment $T[i .. \ell]$ that occurs to the left of position i , if this length does not exceed cb . We consider windows of length $2cb$ over the prefix $T[1 .. (j + c)b]$, overlapping by cb positions. Clearly, if a fragment $T[i .. \ell]$ occurs earlier in T , then it must be a substring of at least one such window. For a fixed L we initialize all the b positions of an array LS_L to 0; we then consider one window W of $2cb$ positions at a time, from left to right. At the end of the computation for a window W , $\text{LS}_L[q]$ will store the length of the longest fragment starting at position $(j - 1)b + q$ which occurs earlier in T . We proceed as follows to achieve this computation.

For the current window W of length $2cb$, we concatenate W and $T[(j - 1)b + 1 .. (j + c)b] = L \cdot T[jb + 1 .. (j + c)b]$ (that is, block L and the following cb positions) constructing a new string S ; we use a separator letter that does not occur in either of the two strings. We then

construct the suffix tree of S ; and from there on the Longest Previous Factor (LPF) array of S in $\mathcal{O}(|S|) = \mathcal{O}(b)$ time [25]. The LPF array is an array of length $|S|$; for each position i of S , it gives the length of the longest substring of S that occurs both at i and to the left of i in S . Finally, we use this information to update the values of LS_L : $\text{LS}_L[q]$ maintains the maximum between its previous value and the new value computed for the current W . We proceed to the next window. Once we have processed all the windows, we use LS_L to update the corresponding values of S_T in $\mathcal{O}(b)$ time the same way as we used END_L in Section 4.

The time and space complexity is as follows. There are n/b blocks in T , each of length b . For each such block, we consider $\mathcal{O}(n/b)$ windows of $2cb$ positions each, and for each window, we construct the suffix tree and the LPF array of the two underlying fragments of length $\mathcal{O}(b)$ in $\tilde{\mathcal{O}}(b)$ time using $\mathcal{O}(b)$ words of space. The whole procedure, for all n/b blocks and all $\mathcal{O}(n/b)$ windows, thus requires $\tilde{\mathcal{O}}(\frac{n}{b} \frac{n}{b} b) = \tilde{\mathcal{O}}(n^2/b)$ time using $\mathcal{O}(b)$ words of space. ◀

5.2 b -Runs and b -Gaps

When $k > cb$, we process b values of $S_T(k)$ at each phase, just like we did in Section 4. Different from Section 4, though, we aim at selecting a *global* set of anchors, carefully chosen among the length- b substrings of T . At each phase, we will distinguish three types of substrings, depending on the period of their length- b substrings. A *b -run* is a maximal fragment of length at least b such that each of its length- b substrings is strongly periodic; a standard reasoning based on the periodicity lemma [33] shows that the period of each b -run is at most $b/4$, and so a b -run is indeed a run. A *b -gap* is a maximal fragment such that none of its length- b substrings is strongly periodic. Any fragment of T of length at least b and period at most $b/4$ is fully contained in a unique b -run; and every fragment of T of length at least b and such that none of its length- b substrings is strongly periodic is fully contained in a unique b -gap. At each phase, the substrings to be processed are thus of three types: (i) either they are fully contained in a b -gap, or (ii) they are fully contained in a b -run, or (iii) neither of the two. We will process the substrings differently depending on their type. A standard reasoning using the periodicity lemma [33] shows that two b -runs cannot overlap by more than $b/2$ letters, so there are only $\mathcal{O}(n/b)$ of them. Lemma 10 states that we can identify and store the b -runs of T in such space complexity. For proving it we rely on the space-efficient construction of sparse suffix trees. The term “sparse” refers to constructing the compacted trie of an arbitrary subset of the set of the suffixes of the input string.

► **Theorem 9** ([10]). *Given a set $B \subseteq [n]$ of size $\Omega(\log n) \leq |B| \leq n$, there exists a deterministic algorithm which constructs the (sparse) suffix tree of B in $\mathcal{O}(n \log \frac{n}{|B|})$ time using $\mathcal{O}(|B|)$ words of space.*

► **Lemma 10.** *A representation of the b -runs of T can be computed in $\tilde{\mathcal{O}}(n)$ time using $\mathcal{O}(n/b)$ space, which is $\mathcal{O}(b)$ space when $b \geq \sqrt{n}$.*

Proof. We process windows of b positions of T at a time, with any two consecutive windows overlapping by $b/2$ positions. At each step, we compute the longest suffix, which has period at most $b/4$, of the window in $\mathcal{O}(b)$ time [24]. If such a suffix has nonzero length, we keep track of its starting position in T and extend it naïvely to the right as much as possible. If this extension results in a run of length at least b , we store its starting and ending position in a list ordered by starting position and resume the process using the window starting $b - 1$ positions before the end of the run. Otherwise, if the extension results in a run shorter than b , we ignore it. Whenever we identify a b -run, we compute its root t in $\mathcal{O}(b)$ time [28], and store in a list the starting and ending position (s_r, e_r) of its root and the starting and

12:10 Substring Complexity in Sublinear Space

ending position (s, e) of the b -run (as mentioned above). After computing all b -runs in T , we construct the sparse suffix tree over the set of all s_r positions in the list. Each internal node of the sparse suffix tree, corresponding to a root of a b -run of T , is associated with the list of the starting and ending positions (s, e) of the b -runs corresponding to this root.

This procedure identifies all the b -runs of T . Indeed, consider a window $T[i..i+b-1]$. If a b -run Y with period $p \leq b/4$ begins between position i and position $i+b-1-p$, a prefix of it of length greater than p is a suffix of the window with period p . If it is the longest such suffix, it will be extended to the right allowing the identification of the whole Y . Otherwise, suppose there is a longer suffix of $T[i..i+b-1]$ with period $b/4 \geq p' > p$ (it cannot be $p' < p$, because otherwise, p' would have been the period of the whole suffix) that includes the whole prefix of Y in $T[i..i+b-1]$. In this case, we only extend the longer suffix and do not find Y at this stage. However, the longer suffix with period $p' \leq b/4$ is part of a run that overlaps with Y , and therefore such overlap must be shorter than $b/2$ because of the periodicity lemma [33]. This means: (a) this situation can only happen when the prefix of Y in $T[i..i+b-1]$ is shorter than $b/2$, thus a longer prefix of Y will be a suffix of the next window $T[i+b/2..i+3b/2-1]$; and (b) the period p' must break before the end of $T[i+b/2..i+3b/2-1]$, thus the prefix of Y in $T[i+b/2..i+3b/2-1]$ must be the longest suffix with period at most $b/4$ and will therefore be extended, allowing to identify the whole Y . Finally, if Y begins between position $i+b-p$ and position $i+b-1$ of $T[i..i+b-1]$, its prefix included in the window does not have a period p , and will therefore not be extended. However, the next window is $T[i+b/2..i+3b/2-1]$: since the length of any b -run is at least b , a prefix of the b -run of length greater than $b/2$ is now a suffix of the window, and since $p \leq b/4$, it will be extended to the right allowing the identification of the whole b -run.

The time and space complexity is as follows. We consider $\mathcal{O}(n/b)$ windows of length b . At each step, we spend $\mathcal{O}(b)$ time to compute the longest suffix of the current window with period at most $b/4$. Whenever we identify a suffix of a run Y with period at most $b/4$, we extend it naively to the right in $\mathcal{O}(|Y|)$ time, and the next window we consider only covers the last $b-1$ positions of Y . Since consecutive b -runs can only overlap by less than $b/2$ positions because of the periodicity lemma [33], they are at most $\mathcal{O}(n/b)$ and their total length is $\mathcal{O}(n)$, so it takes $\mathcal{O}(n)$ time to perform all extensions. For each b -run, we spend $\mathcal{O}(b)$ time to compute its root. For the sparse suffix tree, we employ Theorem 9. Hence the overall time complexity is $\tilde{\mathcal{O}}(n)$. As for the space, we process blocks of $\mathcal{O}(b)$ positions in $\mathcal{O}(b)$ space. We also store a pair of positions for each b -run, therefore the space required to store them is $\mathcal{O}(n/b)$, which is $\mathcal{O}(b)$ when $b \geq \sqrt{n}$. ◀

The output of Lemma 10 is a list representing all the b -runs of T in the natural left-to-right order. The b -gaps can be deduced from this list as follows: if $T[i..j]$ and $T[i'..j']$ are two consecutive b -runs in the list, then $T[j-b+2..i'+b-2]$ is a b -gap (if $T[i..j]$ is the first run, then so is $T[1..i+b-2]$, and similarly for the last run).

A subset of the length- b substrings of T is a *valid set of anchors* if two properties hold: (i) at least one anchor occurs in each fragment of T of length cb ; and (ii) the total number of occurrences of all anchors in T is in $\mathcal{O}(n/b \cdot \log n)$. Lemma 11 shown next will be useful to prove that there always exists a set of valid anchors included in the b -gaps of T .

► **Lemma 11.** *Let Z be a string with all length- d substrings not strongly periodic, and $c > 1$ be any integer constant. Then we can compute in $\tilde{\mathcal{O}}(|Z|^2/d)$ time and $\tilde{\mathcal{O}}(|Z|/d + d)$ space a subset A of the length- d substrings of Z such that: (i) at least one $h \in A$ occurs in each fragment of Z of length cd ; and (ii) the total number of occurrences of all $h \in A$ in Z is $\mathcal{O}(|Z|/d \cdot \log |Z|)$.*

Sketch of Proof. The high-level idea of the proof is to first reduce the problem to the following: we have $\mathcal{O}(|Z|/d)$ strings Z_i , each of length $5d/4$ and with all length- d substrings not strongly periodic, and a set of $\mathcal{O}(|Z|)$ possible anchors consisting of all length- d substrings of the Z_i s. We want to choose a subset A of the anchors such that (i) at least one $h \in A$ occurs in each Z_i , (ii) the total number of occurrences of all $h \in A$ in the Z_i s is $\mathcal{O}(|Z|/d \cdot \log |Z|)$. This is a special case of the Node Selection problem, considered in [9] as a strengthening of the well-known Hitting Set problem.³ Indeed, we can take U to be the set of strings Z_i , V to be the set of possible anchors, and add an edge (u, v) in $G(U, V, E)$ when the possible anchor corresponding to v occurs in the string Z_i corresponding to u . Because every possible anchor is not strongly periodic and every Z_i is of the same length $5d/4$, the degree of every node $u \in U$ is $5d/4$. Then, by Lemma 5.4 of [9] (the weights are irrelevant) we can choose a set $V' \subseteq V$ such that (i) $N[u] \cap V' \neq \emptyset$ for every $u \in U$, (ii) $\sum_{u \in U} |N[u] \cap V'| = \mathcal{O}(|U| \log |U|) = \mathcal{O}(|Z|/d \cdot \log |Z|)$, so V' corresponds to a set of anchors A' with the sought properties. Furthermore, V' can be found in linear time and space in the size of G , which is $\mathcal{O}(|Z|)$. This is however not enough for our purposes, as we cannot store the whole G . Analysing the algorithm used inside the proof of Lemma 5.4 of [9] we see that it considers the nodes $v \in V$ one-by-one while maintaining some information of size $\mathcal{O}(|U|) = \mathcal{O}(|Z|/d)$ and a precomputed table of a size that can be bounded by the maximum degree of any $u \in U$, which is $\mathcal{O}(d)$. Furthermore, the algorithm accesses G only by iterating a constant number of times over the neighbours of the current node $v \in V$. In the full version, we show how to implement this efficiently in our model to achieve the claimed bounds. ◀

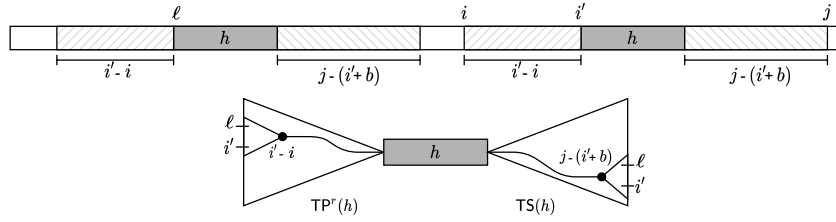
5.3 Processing the b -Gaps

For ease of presentation, in this section, we will assume that all length- b substrings of T are not strongly periodic, but no major changes are required to apply the same reasoning on the set of all b -gaps. Assume we have already computed a set A of valid anchors over T . For each $h \in A$, we compute a list of its occurrences in T . The overall size of these lists is $\mathcal{O}(n/b \cdot \log n)$ because of property (ii), and the occurrences of each $h \in A$ can be generated in $\mathcal{O}(n)$ time and $\mathcal{O}(1)$ space (plus the space to store the list) with any linear-time constant-space pattern matching algorithm, so $\tilde{\mathcal{O}}(n^2/b)$ time overall. We divide the computation of $S_T(k)$ in n/b phases. Consider phase α , in which we consider substrings of length $k \in [\alpha b + 1, (\alpha + 1)b]$. Because of property (i), at least one anchor occurs in the first cb positions of each such substring. We conceptually associate such a substring with the leftmost anchor $h \in A$ occurring therein, and we say that a fragment of T is *anchored* at an occurrence i of some anchor h if the leftmost occurrence of any anchor in the fragment is i . We then process the substrings according to the anchor with which they are associated.

All substrings associated with an anchor $h \in A$ have a (possibly empty) prefix of length $\mathcal{O}(b)$ where no anchors occur, followed by h and then by a suffix where any anchor can occur. This implies that any occurrence of such substrings can only start in a range of $\mathcal{O}(b)$ positions preceding some occurrence of h in T . In particular, if h occurs at position i in T and the closest anchor to its left is at position $i' < i$, the *starting range* of substrings of T associated with h is $[i' + 1, i]$, or $[1, i]$ if i is the first occurrence of any anchors in T . All starting ranges for all anchors can be computed in $\mathcal{O}(n)$ time by scanning the list of occurrences of the anchors. To update the values of $S_T(k)$ with the substrings associated with h we need to

³ Let us remark that this problem has already been considered in the conference version [8], with a slightly different definition but essentially the same proof. However, our goal is a deterministic algorithm and to this end we need [9], the extended version of [8].

12:12 Substring Complexity in Sublinear Space



■ **Figure 3** A previous occurrence of $T[i..j]$ anchored at i' can be detected using $D(h)$.

know, for each occurrence i of h in T and each of its previous occurrences $i' < i$, the longest left extension within the starting ranges of i and i' , and the longest right extension of the fragments of T following the occurrences of h at i and i' . We cannot afford to store all these pairs of values explicitly as this would require $\tilde{O}(n^2/b^2)$ space. We thus construct a separate data structure, denoted by $D(h)$, for each anchor $h \in A$. This data structure encode the same information in a compact form. We next describe the data structure and its construction.

$D(h)$ consists of two *compacted* tries $\text{TP}^r(h)$ and $\text{TS}(h)$. For every occurrence i of h in T , $\text{TS}(h)$ contains a leaf corresponding to $T[i+b..n]$, and $\text{TP}^r(h)$ a leaf corresponding to $(T[1..i-1])^r$, both labelled with position i . We only store the list of children and the length of the path label of each node, which we call its *depth*. Because of property (ii), the overall size of these data structures for all anchors is thus in $\mathcal{O}(n/b \cdot \log n)$. For any two occurrences i, i' of h , the depth of the lowest common ancestor of leaves i and i' in $\text{TP}^r(h)$ gives the length of their longest left extension, and the depth of their lowest common ancestor in $\text{TS}(h)$ gives the length of their longest right extension: see Figure 3 for an example. $D(h)$ can be efficiently constructed for all $h \in A$, as shown by Lemma 12.

► **Lemma 12.** *Data structures $D(h)$, for all $h \in A$, can be constructed in $\tilde{O}(n^2/b)$ total time using $\tilde{O}(n/b)$ space, which is $\tilde{O}(b)$ when $b \geq \sqrt{n}$.*

Proof. Let $\text{occ}(h)$ be the list of occurrences of anchor h in T , let $B = \bigcup_{h \in A} \text{occ}(h)$ and $B' = \bigcup_{h \in A} \text{occ}(h) + b - 1$. Recall that $D(h)$ consists of two *compacted* tries $\text{TP}^r(h)$ and $\text{TS}(h)$. We will first construct two global compacted tries $\text{TP}^r(A)$ and $\text{TS}(A)$ for all anchors in A , and then extract from them subtrees $\text{TP}^r(h)$ and $\text{TS}(h)$ for each $h \in A$.

$\text{TP}^r(A)$ and $\text{TS}(A)$ are constructed in the same way, except that for $\text{TP}^r(A)$ we consider the reversal of strings. To construct $\text{TS}(A)$ we employ Theorem 9 on set B , as it is essentially the sparse suffix tree for the suffixes starting at positions in B ; and to construct $\text{TP}^r(A)$ we employ Theorem 9 on the reverse of T and set B' . Once we have constructed $\text{TS}(A)$ and $\text{TP}^r(A)$, to extract subtrees $\text{TS}(h)$ and $\text{TP}^r(h)$ for $h \in A$ it suffices to spell h from the root of $\text{TS}(A)$ (resp. h^r from the root of $\text{TP}^r(A)$) and take the subtree below.

The time and space complexity of computing $\text{TS}(A)$ and $\text{TP}^r(A)$ is as follows. The size of sets B and B' is $\mathcal{O}(n/b \cdot \log n) = \mathcal{O}(b \log n)$ when $b \geq \sqrt{n}$, thus by Theorem 9 we make use of $\mathcal{O}(b \log n)$ words of space. Again by Theorem 9, the overall time complexity to construct them is $\tilde{O}(n)$. To find the right subtree for each $h \in A$ we then spend $\tilde{O}(b)$ time for each of the $\mathcal{O}(n/b \log n)$ anchors of A , thus again $\tilde{O}(n)$ time overall. ◀

Computing $S_T(k)$ Using $D(h)$. Similar to Section 4, in each phase α we fill in an auxiliary array $I = I[1..b]$ such that, at the end of the phase, $I[q]$ contains the number of positions i in T such that the shortest substring that does not occur in T before position i is of length $\alpha b + q$. We proceed as follows. We consider one position of T at the time, from left to right. When we are at position i , let h be the leftmost anchor occurring at some position $i' \geq i$.

We binary search for the smallest position j such that $T[i..j]$ does not occur to the left of i using $D(h)$. We first identify in $\text{TP}^r(h)$ the highest ancestor u of leaf i' with string depth at least $i' - i$. This corresponds to answering a *weighted level ancestor* query [30] on $\text{TP}^r(h)$, where the weight of each node is its depth. After linear-time preprocessing, weighted ancestor queries for nodes of a weighted tree with integer weights from a universe $[1..U]$ can be answered in $\mathcal{O}(\log \log U)$ time [1]. In our case, the queries thus cost $\mathcal{O}(\log \log n)$ time.

We then start binary searching for the leftmost position j such that $T[i..j]$ does not occur to the left of position i and such that $|T[i..j]| \in [\alpha b + 1, (\alpha + 1)b]$: we thus look for j in the range $[i + \alpha b, i + (\alpha + 1)b - 1]$. For each value j considered in the binary search, we find in $\text{TS}(h)$ the highest ancestor v of leaf i' with string depth at least $j - (i' + b)$, by answering a weighted level ancestor query. We then need to check whether $T[i..j]$ occurs somewhere to the left of i , in correspondence of a previous occurrence of anchor h , in which case we increase j in the next step; or it does not occur before, in which case we decrease j . We do so by looking at the leaves (occurrences of h) in the subtree below u in $\text{TP}^r(h)$, denoted by $\text{TP}^r(h)|u$, and in the subtree below v in $\text{TS}(h)$, $\text{TS}(h)|v$. Every leaf in the intersection of the two subsets of leaves corresponds to an occurrence of $T[i..j]$ in T . The information we need is whether i' is the smallest leaf in the intersection, meaning that $T[i..j]$ does not occur anywhere before. This reduces to a 2D range searching problem.

We assume that each leaf of each tree has a unique identifier, independent from their label and such that the identifiers of the leaves of any subtree form a contiguous range. For each leaf ℓ , its identifiers in $\text{TP}^r(h)$ and $\text{TS}(h)$ give the coordinates of a point on a plane, to which we assign ℓ as weight. By construction, the points corresponding to leaves in the intersection of $\text{TP}^r(h)|u$ and $\text{TS}(h)|v$ are contained in a rectangle: we need to find the point with the smallest weight there and check whether it is i' or not. Such queries can be answered in time $\mathcal{O}(\log s)$ with a data structure that is constructed in time and space $\mathcal{O}(s \log s)$, where s is the total number of points [17]. At the end of the binary search, if $j = i + \alpha b + q$ we increase the counter at $I[q]$ by one, unless $j = i + (\alpha + 1)b - 1$ and $T[i..j]$ occurs before i , in which case we do not increase any counters. We finally move to the next position of T .

► **Lemma 13.** *Assume that all length- b substrings of T are not strongly periodic. Then δ can be computed in $\tilde{\mathcal{O}}(n^2/b)$ time using $\tilde{\mathcal{O}}(n/b + b)$ space, which is $\tilde{\mathcal{O}}(b)$ when $b \geq \sqrt{n}$.*

Proof. Set A is selected in $\tilde{\mathcal{O}}(n^2/b)$ time and $\tilde{\mathcal{O}}(n/b + b)$ space as per Lemma 11, and $D(h)$ can be computed in the same time and space for all $h \in A$ and all phases, as per Lemma 12. In each phase α , we go over the n positions of T one at a time. At each position i we binary search for the shortest substring not occurring before i in $\mathcal{O}(\log \alpha b)$ steps, each requiring $\mathcal{O}(\log n)$ time. Over all $\mathcal{O}(n/b)$ phases, this requires $\tilde{\mathcal{O}}(n^2/b)$ time and $\tilde{\mathcal{O}}(n/b)$ space. ◀

5.4 Processing the b -Runs

Recall that we have computed, as per Lemma 10, a representation of all the b -runs of T . In this section, we only focus on the substrings of length at least b and periods at most $b/4$. Every occurrence of such a substring is fully contained in some b -run, and for ease of presentation we will assume that in phase α , in which we process substrings of length $k \in [\alpha b + 1, (\alpha + 1)b]$, every b -run is longer than αb . Observe that each substring of a b -run $T[s..e]$ with root t occurs also as a prefix of some fragment starting within the first $|t|$ positions of the run, which we call its *relevant* range. Since we aim to identify the leftmost occurrence of each substring of T , we can ignore all positions of a b -run after its relevant range. By slightly abusing notation, we select as anchors some fragments of the b -runs of T , instead of selecting substrings together with the whole set of their occurrences. However,

this set of anchors must have the following property, that for the anchors of Section 5.3 held naturally: for any two occurrences of the same substring in the relevant ranges, the leftmost occurrence of any anchor therein is at the same offset from the beginning of the substring. In phase α we use as anchors the first two occurrences of the root in each b -run: let H be this set of fragments of T . Clearly, H is of size $\mathcal{O}(n/b)$ because the representation of all the b -runs is of such size.

► **Lemma 14.** *For any two occurrences of the same substring of length at least b and period at most $b/4$, both starting in the relevant ranges of the b -runs of T , the leftmost occurrence of any $h \in H$ in each of them is at the same offset from the beginning of the substring.*

Proof. Let $Y = t[d..|t|]t^\beta t[1..f]$ be a fragment occurring at the first t positions in some b -run $T[s..e] = t[q..|t|]t^\gamma t[1..g]$. The anchors within $T[s..e]$ are, by definition, the occurrences of t at position $p_1 = s + (|t| - q + 1) \bmod |t|$ and $p_2 = p_1 + |t|$. If $d \geq q$, the leftmost occurrence of any anchors in Y is at p_1 , which is at offset $|t| - d + 2$ in Y . Otherwise, if $d < q$, the leftmost occurrence of any anchors in Y is at p_2 , which is in any case at offset $|t| - d + 2$ in Y .

Consider another occurrence of Y in the first $|t|$ positions of some other b -run $T[s'..e'] = t[q'..|t|]t^{\gamma'} t[1..g']$. The anchors are the occurrences of t at position $p'_1 = s' + (|t| - q' + 1) \bmod |t|$ and $p'_2 = p'_1 + |t|$; depending on whether $d \geq q'$ or not, the leftmost occurrence of any anchor in this occurrence of Y is either p'_1 or p'_2 , in either case at offset $|t| - d + 2$ in Y . ◀

Let P be the set of roots of the b -runs of T . We construct a data structure $D(P)$ for all roots $t \in P$ similar to what we do in Section 5.3, but we use only the occurrences of t corresponding to fragments in H . We then proceed as described in Section 5.3 to fill in array I , except that, in each b -run with root t , we disregard any position after the first $|t|$.

We have arrived at the following lemma.

► **Lemma 15.** *The substrings of T that are fully contained within a b -run can be processed in $\tilde{\mathcal{O}}(n^2/b)$ time using $\mathcal{O}(n/b)$ space, which is $\mathcal{O}(b)$ when $b \geq \sqrt{n}$.*

5.5 Computing $S_T(k)$ for Large k

The occurrences of anchors $h \in A$ selected for the b -gaps anchor all the fragments fully contained in a b -gap and possibly some other fragments. However, we are not guaranteed that this holds for any fragment not fully contained in a b -run. Consider a fragment $T[i..j]$ of length at least b with period larger than $b/4$ (thus, not contained in any b -run) but containing a strongly periodic length- b fragment $T[i'..j']$ inside (so, not contained in any b -gap). Then, $T[i'..j']$ is fully contained in some b -run $T[s..e]$. Because $T[i..j]$ is not fully contained in $T[s..e]$, either $T[s-1..s+b-2]$ or $T[e-b+2..e+1]$ (that is, a length- b substring with exactly one letter before or after the b -run) is fully within $T[i..j]$. This suggest that we should augment A with the following length- b substrings: for each b -run $T[s..e]$, $T[s-1..s+b-2] \in A$ and $T[e-b+2..e+1] \in A$, and we consider all their occurrences in T . By the above reasoning, this guarantees that $T[i..j]$ contains an occurrence of some anchor inside. We are defining only $\mathcal{O}(n/b)$ new anchors, but then we need to consider all of their occurrences. Therefore, we need to argue that the total number of occurrences of the new anchors is $\mathcal{O}(n/b)$. It is enough to show this for the occurrences of the anchors $T[s-1..s+b-2]$, where the period of $T[s..s+b-2]$ is at most $b/4$. We claim that for any two such occurrences $T[s-1..s+b-2]$ and $T[s'-1..s'+b-2]$ with $s < s'$ we have $s + b/2 < s'$: otherwise $T[s..s+b-2]$ and $T[s'..s'+b-2]$ overlap by at least $b/2$ positions, but two b -runs cannot overlap by $b/2$ positions, a contradiction. We generate all

these occurrences and then process all the anchors as in Section 5.3. The only difference is the starting range associated with the anchors obtained from the suffix of some b -run: when they are not preceded by another anchor within cb positions, we take as starting range the cb positions preceding the anchor.

Let us put everything together. Before computing $S_T(k)$ in phases, we identify the b -runs and the b -gaps of T as per Lemma 10. We then extract a set of anchors from the b -gaps as described in Lemma 11, and we complement it with the length- b substrings that start one position before each b -run, and with the length- b substrings that end one position after the end of each b -run, to complete the set A of anchors. We then compute the list of occurrences of each $h \in A$; we also identify the relevant ranges within each b -run. We then proceed in phases. In each phase, we scan T from left to right and process all positions in b -gaps as per Section 5.3. All positions within a b -run are processed as per Section 5.4, and additionally as per Section 5.3, when they are within the starting range of an occurrence of some $h \in A$. At the end of a phase α , we have computed an auxiliary array I such that $I[h]$ gives the number of positions i of T such that the shortest substring that does not occur in T before position i is of length $ab + h$. We use I to compute $S_T(k)$ for each $k \in [ab + 1, (\alpha + 1)b]$ as in Section 4.

By combining Lemmas 13 and 15 we arrive at Theorem 3, the main result of this paper.

6 Substring Complexity from the Combinatorial Point of View

Knowing the substring complexity of a string can also be used to find other regularities. To mention a few, we have the following straightforward implications in sublinear working space:

- T has a substring of length k repeating in T if and only if $S_T(k) < n - k + 1$. This yields the length r of the longest repeated substring of T (also known as the repetition index of T) [68]. It is worth noticing that $S_T(k + 1) = S_T(k) - 1$ for every $k > r$ [27] and that r approximates $\mathcal{O}(\log_{|\Sigma|} n)$ when T is randomly generated by a memoryless source [32].
- A string S is called a *minimal absent word* of T if S does not occur in T but all proper substrings of S occur in T . The length ℓ of a longest minimal absent word of T is equal to $2 + r$ [32]. This quantity is important because if two strings X and Y have the same set of distinct substrings up to length ℓ , then $X = Y$ [32, 14]. The length of a shortest absent word [69] of T over alphabet Σ is equal to the smallest k such that $S_T(k) < |\Sigma|^k$.
- The longest common substring of strings X and Y is equal to the largest k such that $S_X(k) + S_Y(k) > S_{X\#Y}(k) - k$, where $\#$ does not occur in X nor in Y , since there are precisely k distinct substrings of length k containing the letter $\#$ in $X\#Y$.

The substring complexity function is well studied in the area of combinatorics on words, both for finite and infinite strings. However, the normalization $S(k)/k$ and its supremum δ have not been considered until very recently. In [27] it is proved that the substring complexity $S_T(k)$ of a string T takes its maximum precisely for $k = R$, where R is the minimum length for which no substring of T has occurrences followed by different letters, and one has $S_T(R) = n + 1 - \max\{R, K\}$, where K is the length of the shortest unrepeated suffix of T . But this seems to be of little help in understanding the behaviour of the *normalized* substring complexity $S_T(k)/k$.

7 Approximating δ in Sublinear Space

Our algorithms compute the *exact value* of δ . If one is interested in a constant-factor approximation of δ (e.g., an algorithm's complexity has a polynomial dependency on δ [53]), then there is a simple algorithm in our model based on the following combinatorial observation, which follows directly by the number of fragments of length ℓ of a string of length n being $n - \ell + 1$, and by the fact that each fragment of length $\ell' > \ell$ has a prefix of length ℓ .

► **Observation 16.** For any string T , let $S_T(k)$ be the number of distinct substrings of length k . The number $S_T(k')$ of distinct substrings of any length $k' > k$ is at least $S_T(k) - (k' - k)$.

► **Lemma 17.** Let $\delta' = \sup\{\frac{S_T(2^d)}{2^d} \mid d = 0, \dots, \log n\}$. Then $\delta \leq 2\delta' + 1$.

Proof. Let $\delta = \frac{S_T(k)}{k}$ for some $k \geq 1$, and let d be the integer such that

$$2^d \leq k < 2^{d+1}. \quad (1)$$

By the definition of δ' , we have that $\delta' \geq \frac{S_T(2^{d+1})}{2^{d+1}}$. By applying Observation 16, we obtain:

$$\frac{S_T(2^{d+1})}{2^{d+1}} \geq \frac{S_T(k) - (2^{d+1} - k)}{2^{d+1}} \geq \frac{S_T(k) - (2^{d+1} - 2^d)}{2^{d+1}} \geq \frac{S_T(k)}{2k} - \frac{2^d}{2^{d+1}} = \frac{1}{2}\delta - \frac{1}{2}. \blacktriangleleft$$

Recall that the algorithm underlying Theorem 2 works in $\frac{n}{b}$ phases, where each phase handles a range of b lengths k . By plugging in Lemma 17, the number of phases become $\Theta(\log n)$ – instead of $\Theta(n/b)$ – and so we obtain a simple $\tilde{O}(n^2/b)$ -time and $\mathcal{O}(b)$ -space algorithm to approximate δ , within a constant factor, in the comparison model.

References

- 1 Amihood Amir, Gad M. Landau, Moshe Lewenstein, and Dina Sokol. Dynamic text and static pattern matching. *ACM Trans. Algorithms*, 3(2):19, 2007. doi:10.1145/1240233.1240242.
- 2 Alberto Apostolico, Maxime Crochemore, Martin Farach-Colton, Zvi Galil, and S. Muthukrishnan. 40 years of suffix trees. *Commun. ACM*, 59(4):66–73, 2016. doi:10.1145/2810036.
- 3 Lorraine A. K. Ayad, Golnaz Badkobeh, Gabriele Fici, Alice Héliou, and Solon P. Pissis. Constructing antidictionaries in output-sensitive space. In *29th Data Compression Conference (DCC)*, pages 538–547, 2019. doi:10.1109/DCC.2019.00062.
- 4 Paul Beame. A general sequential time-space tradeoff for finding unique elements. *SIAM J. Comput.*, 20(2):270–277, 1991. doi:10.1137/0220017.
- 5 Paul Beame, Raphaël Clifford, and Widad Machmouchi. Element distinctness, frequency moments, and sliding windows. In *54th Symposium on Foundations of Computer Science (FOCS)*, pages 290–299, 2013. doi:10.1109/FOCS.2013.39.
- 6 Djamel Belazzougui. Linear time construction of compressed text indices in compact space. In *46th Symposium on Theory of Computing, (STOC)*, pages 148–193, 2014. doi:10.1145/2591796.2591885.
- 7 Michael A. Bender and Martin Farach-Colton. The LCA problem revisited. In *4th Latin American Symposium (LATIN)*, pages 88–94, 2000. doi:10.1007/10719839_9.
- 8 Giulia Bernardini, Pawel Gawrychowski, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Even faster elastic-degenerate string matching via fast matrix multiplication. In *46th International Colloquium on Automata, Languages, and Programming, (ICALP)*, volume 132 of *LIPICs*, pages 21:1–21:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPICs.ICALP.2019.21.
- 9 Giulia Bernardini, Pawel Gawrychowski, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Elastic-degenerate string matching via fast matrix multiplication. *SIAM J. Comput.*, 51(3):549–576, 2022. doi:10.1137/20m1368033.
- 10 Or Birenzweige, Shay Golan, and Ely Porat. Locally consistent parsing for text indexing in small space. In *31st Symposium on Discrete Algorithms, (SODA)*, pages 607–626. SIAM, 2020. doi:10.1137/1.9781611975994.37.
- 11 Allan Borodin and Stephen A. Cook. A time-space tradeoff for sorting on a general sequential model of computation. *SIAM J. Comput.*, 11(2):287–297, 1982. doi:10.1137/0211022.
- 12 Dany Breslauer and Zvi Galil. Real-time streaming string-matching. *ACM Trans. Algorithms*, 10(4):22:1–22:12, 2014. doi:10.1145/2635814.
- 13 Dany Breslauer, Roberto Grossi, and Filippo Mignosi. Simple real-time constant-space string matching. *Theoret. Comput. Sci.*, 483:2–9, 2013. doi:10.1016/j.tcs.2012.11.040.

- 14 Arturo Carpi and Aldo de Luca. Words and special factors. *Theoret. Comput. Sci.*, 259(1-2):145–182, 2001. doi:10.1016/S0304-3975(99)00334-5.
- 15 Timothy M. Chan, Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, and Ely Porat. Approximating text-to-pattern Hamming distances. In *52nd Symposium on Theory of Computing (STOC)*, pages 643–656, 2020. doi:10.1145/3357713.3384266.
- 16 Panagiotis Charalampopoulos, Maxime Crochemore, Costas S. Iliopoulos, Tomasz Kociumaka, Solon P. Pissis, Jakub Radoszewski, Wojciech Rytter, and Tomasz Walen. Linear-time algorithm for long LCF with k mismatches. In *29th Symposium on Combinatorial Pattern Matching (CPM)*, pages 23:1–23:16, 2018. doi:10.4230/LIPIcs.CPM.2018.23.
- 17 Bernard Chazelle. A functional approach to data structures and its use in multidimensional searching. *SIAM J. Comput.*, 17(3):427–462, 1988. doi:10.1137/0217026.
- 18 Anders Roy Christiansen, Mikko Berggren Ettienne, Tomasz Kociumaka, Gonzalo Navarro, and Nicola Prezza. Optimal-time dictionary-compressed indexes. *ACM Trans. Algorithms*, 17(1):8:1–8:39, 2021. doi:10.1145/3426473.
- 19 Raphaël Clifford, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana Starikovskaya. Dictionary matching in a stream. In *23rd Annual European Symposium on Algorithms (ESA)*, pages 361–372, 2015. doi:10.1007/978-3-662-48350-3_31.
- 20 Raphaël Clifford, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana Starikovskaya. The k -mismatch problem revisited. In *37th Symposium on Discrete Algorithms (SODA)*, pages 2039–2052, 2016. doi:10.1137/1.9781611974331.ch142.
- 21 Raphaël Clifford, Tomasz Kociumaka, and Ely Porat. The streaming k -mismatch problem. In *30th Symposium on Discrete Algorithms (SODA)*, pages 1106–1125, 2019. doi:10.1137/1.9781611975482.68.
- 22 Raphaël Clifford and Tatiana Starikovskaya. Approximate Hamming distance in a stream. In *43rd International Colloquium on Automata, Languages, and Programming, (ICALP)*, pages 20:1–20:14, 2016. doi:10.4230/LIPIcs.ICALP.2016.20.
- 23 Richard Cole, Tsvi Kopelowitz, and Moshe Lewenstein. Suffix trays and suffix trists: Structures for faster text indexing. *Algorithmica*, 72(2):450–466, 2015. doi:10.1007/s00453-013-9860-6.
- 24 Maxime Crochemore, Christophe Hancart, and Thierry Lecroq. *Algorithms on strings*. Cambridge University Press, 2007.
- 25 Maxime Crochemore, Lucian Ilie, Costas S. Iliopoulos, Marcin Kubica, Wojciech Rytter, and Tomasz Walen. Computing the longest previous factor. *Eur. J. Comb.*, 34(1):15–26, 2013. doi:10.1016/j.ejc.2012.07.011.
- 26 Maxime Crochemore and Dominique Perrin. Two-way string matching. *J. ACM*, 38(3):651–675, 1991. doi:10.1145/116825.116845.
- 27 Aldo de Luca. On the combinatorics of finite words. *Theoret. Comput. Sci.*, 218(1):13–39, 1999. doi:10.1016/S0304-3975(98)00248-5.
- 28 Jean Pierre Duval. Factorizing words over an ordered alphabet. *Journal of Algorithms*, 4(4):363–381, 1983.
- 29 Martin Farach. Optimal suffix tree construction with large alphabets. In *38th Symposium on Foundations of Computer Science (FOCS)*, pages 137–143, 1997. doi:10.1109/SFCS.1997.646102.
- 30 Martin Farach and S. Muthukrishnan. Perfect hashing for strings: Formalization and algorithms. In *7th Symposium on Combinatorial Pattern Matching (CPM)*, volume 1075 of *Lecture Notes in Computer Science*, pages 130–140. Springer, 1996. doi:10.1007/3-540-61258-0_11.
- 31 Paolo Ferragina and Giovanni Manzini. Indexing compressed text. *J. ACM*, 52(4):552–581, 2005. doi:10.1145/1082036.1082039.
- 32 Gabriele Fici, Filippo Mignosi, Antonio Restivo, and Marinella Sciortino. Word assembly through minimal forbidden words. *Theoret. Comput. Sci.*, 359(1-3):214–230, 2006. doi:10.1016/j.tcs.2006.03.006.
- 33 Nathan J. Fine and Herbert S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965. doi:10.2307/2034009.

- 34 Johannes Fischer, Travis Gagie, Pawel Gawrychowski, and Tomasz Kociumaka. Approximating LZ77 via small-space multiple-pattern matching. In *23rd European Symposium on Algorithms (ESA)*, pages 533–544, 2015. doi:10.1007/978-3-662-48350-3_45.
- 35 Travis Gagie, Gonzalo Navarro, and Nicola Prezza. Fully functional suffix trees and optimal text searching in BWT-runs bounded space. *J. ACM*, 67(1):2:1–2:54, 2020. doi:10.1145/3375890.
- 36 Zvi Galil and Joel I. Seiferas. Time-space-optimal string matching. *J. Comput. Syst. Sci.*, 26(3):280–294, 1983. doi:10.1016/0022-0000(83)90002-8.
- 37 Pawel Gawrychowski and Tatiana Starikovskaya. Streaming dictionary matching with mismatches. In *30th Symposium on Combinatorial Pattern Matching (CPM)*, pages 21:1–21:15, 2019. doi:10.4230/LIPIcs.CPM.2019.21.
- 38 Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, and Ely Porat. The streaming k-mismatch problem: Tradeoffs between space and total time. In *31st Symposium on Combinatorial Pattern Matching (CPM)*, pages 15:1–15:15, 2020. doi:10.4230/LIPIcs.CPM.2020.15.
- 39 Shay Golan, Tsvi Kopelowitz, and Ely Porat. Towards optimal approximate streaming pattern matching by matching multiple patterns in multiple streams. In *45th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 65:1–65:16, 2018. doi:10.4230/LIPIcs.ICALP.2018.65.
- 40 Shay Golan, Tsvi Kopelowitz, and Ely Porat. Streaming pattern matching with d wildcards. *Algorithmica*, 81(5):1988–2015, 2019. doi:10.1007/s00453-018-0521-7.
- 41 Shay Golan and Ely Porat. Real-time streaming multi-pattern search for constant alphabet. In *25th Annual European Symposium on Algorithms (ESA)*, pages 41:1–41:15, 2017. doi:10.4230/LIPIcs.ESA.2017.41.
- 42 Roberto Grossi and Jeffrey Scott Vitter. Compressed suffix arrays and suffix trees with applications to text indexing and string matching. *SIAM J. Comput.*, 35(2):378–407, 2005. doi:10.1137/S0097539702402354.
- 43 Dan Gusfield. *Algorithms on Strings, Trees, and Sequences – Computer Science and Computational Biology*. Cambridge University Press, 1997. doi:10.1017/cbo9780511574931.
- 44 Wing-Kai Hon, Kunihiko Sadakane, and Wing-Kin Sung. Breaking a time-and-space barrier in constructing full-text indices. *SIAM J. Comput.*, 38(6):2162–2178, 2009. doi:10.1137/070685373.
- 45 Juha Kärkkäinen, Peter Sanders, and Stefan Burkhardt. Linear work suffix array construction. *J. ACM*, 53(6):918–936, 2006. doi:10.1145/1217856.1217858.
- 46 Dominik Kempa and Tomasz Kociumaka. String synchronizing sets: sublinear-time BWT construction and optimal LCE data structure. In *51st Symposium on Theory of Computing (STOC)*, pages 756–767, 2019. doi:10.1145/3313276.3316368.
- 47 Dominik Kempa and Tomasz Kociumaka. Breaking the $O(n)$ -barrier in the construction of compressed suffix arrays and suffix trees. In *34th Symposium on Discrete Algorithms, SODA*, pages 5122–5202. SIAM, 2023. doi:10.1137/1.9781611977554.ch187.
- 48 Dominik Kempa and Tomasz Kociumaka. Collapsing the hierarchy of compressed data structures: Suffix arrays in optimal compressed space. *CoRR*, abs/2308.03635, 2023. doi:10.48550/arXiv.2308.03635.
- 49 Dominik Kempa and Nicola Prezza. At the roots of dictionary compression: string attractors. In *50th Symposium on Theory of Computing (STOC)*, pages 827–840, 2018. doi:10.1145/3188745.3188814.
- 50 John C. Kieffer and En-Hui Yang. Grammar-based codes: a new class of universal lossless source codes. *IEEE Trans. Inf. Theory*, 46(3):737–754, 2000. doi:10.1109/18.841160.
- 51 Tomasz Kociumaka, Gonzalo Navarro, and Francisco Olivares. Near-optimal search time in δ -optimal space. In *15th Latin American Symposium (LATIN)*, volume 13568 of *Lecture Notes in Computer Science*, pages 88–103. Springer, 2022. doi:10.1007/978-3-031-20624-5_6.
- 52 Tomasz Kociumaka, Gonzalo Navarro, and Nicola Prezza. Towards a definitive measure of repetitiveness. In *14th Latin American Symposium (LATIN)*, volume 12118 of *Lecture Notes in Computer Science*, pages 207–219. Springer, 2020. doi:10.1007/978-3-030-61792-9_17.

- 53 Tomasz Kociumaka, Gonzalo Navarro, and Nicola Prezza. Toward a definitive compressibility measure for repetitive sequences. *IEEE Trans. Inf. Theory*, 69(4):2074–2092, 2023. doi:10.1109/TIT.2022.3224382.
- 54 Tomasz Kociumaka, Tatiana Starikovskaya, and Hjalte Wedel Vildhøj. Sublinear space algorithms for the longest common substring problem. In *22th European Symposium on Algorithms (ESA)*, pages 605–617, 2014. doi:10.1007/978-3-662-44777-2_50.
- 55 Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1–4):157–168, 1968. doi:10.1080/00207166808803030.
- 56 Dmitry Kosolobov, Daniel Valenzuela, Gonzalo Navarro, and Simon J. Puglisi. Lempel-ziv-like parsing in small space. *Algorithmica*, 82(11):3195–3215, 2020. doi:10.1007/s00453-020-00722-6.
- 57 Sebastian Kreft and Gonzalo Navarro. On compressing and indexing repetitive sequences. *Theoret. Comput. Sci.*, 483:115–133, 2013. doi:10.1016/j.tcs.2012.02.006.
- 58 Veli Mäkinen, Gonzalo Navarro, Jouni Sirén, and Niko Välimäki. Storage and retrieval of highly repetitive sequence collections. *J. Comput. Biol.*, 17(3):281–308, 2010. doi:10.1089/cmb.2009.0169.
- 59 Udi Manber and Eugene W. Myers. Suffix arrays: A new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993. doi:10.1137/0222058.
- 60 J. Ian Munro, Gonzalo Navarro, and Yakov Nekrich. Space-efficient construction of compressed indexes in deterministic linear time. In *28th Symposium on Discrete Algorithms (SODA)*, pages 408–424, 2017. doi:10.1137/1.9781611974782.26.
- 61 Gonzalo Navarro. *Compact Data Structures – A Practical Approach*. Cambridge University Press, 2016. URL: <http://www.cambridge.org/de/academic/subjects/computer-science/algorithmics-complexity-computer-algebra-and-computational-g/compact-data-structures-practical-approach?format=HB>.
- 62 Gonzalo Navarro. Indexing highly repetitive string collections, part I: repetitiveness measures. *ACM Comput. Surv.*, 54(2):29:1–29:31, 2022. doi:10.1145/3434399.
- 63 Stav Ben Nun, Shay Golan, Tomasz Kociumaka, and Matan Kraus. Time-space tradeoffs for finding a long common substring. In *31st Symposium on Combinatorial Pattern Matching (CPM)*, pages 5:1–5:14, 2020. doi:10.4230/LIPIcs.CPM.2020.5.
- 64 Benny Porat and Ely Porat. Exact and approximate pattern matching in the streaming model. In *50th Symposium on Foundations of Computer Science (FOCS)*, pages 315–323, 2009. doi:10.1109/FOCS.2009.11.
- 65 Jakub Radoszewski and Tatiana Starikovskaya. Streaming k -mismatch with error correcting and applications. *Inf. Comput.*, 271:104513, 2020. doi:10.1016/j.ic.2019.104513.
- 66 Sofya Raskhodnikova, Dana Ron, Ronitt Rubinfeld, and Adam D. Smith. Sublinear algorithms for approximating string compressibility. *Algorithmica*, 65(3):685–709, 2013. doi:10.1007/s00453-012-9618-6.
- 67 Tatiana Starikovskaya and Hjalte Wedel Vildhøj. Time-space trade-offs for the longest common substring problem. In *24th Symposium on Combinatorial Pattern Matching (CPM)*, pages 223–234, 2013. doi:10.1007/978-3-642-38905-4_22.
- 68 Peter Weiner. Linear pattern matching algorithms. In *14th Symposium on Switching and Automata Theory*, pages 1–11, 1973. doi:10.1109/SWAT.1973.13.
- 69 Zong-Da Wu, Tao Jiang, and Wu-Jie Su. Efficient computation of shortest absent words in a genomic sequence. *Inf. Process. Lett.*, 110(14-15):596–601, 2010. doi:10.1016/j.ipl.2010.05.008.
- 70 Andrew Chi-Chih Yao. Near-optimal time-space tradeoff for element distinctness. *SIAM J. Comput.*, 23(5):966–975, 1994. doi:10.1137/S0097539788148959.
- 71 Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343, 1977. doi:10.1109/TIT.1977.1055714.