Optimal Internal Shortest Absent Word Queries

Golnaz Badkobeh, Panagiotis Charalampopoulos, Dmitry Kosolobov, Solon P. Pissis

Abstract
Given a string $T$ of length $n$ over an alphabet $\Sigma \subset \{1, 2, \ldots, nO(1)\}$ of size $\sigma$, we are to preprocess $T$ so that given a range $[i, j]$, we can return a representation of a shortest string over $\Sigma$ that is absent in the fragment $T[i] \cdots T[j]$ of $T$. We present an $O(n)$-space data structure that answers such queries in constant time and can be constructed in $O(n \log n)$ time.

Keywords: string algorithms, internal queries, shortest absent word, bit parallelism

1. Introduction

Range queries are a classic data structure topic [59, 12, 11]. In 1d, a range query $q = f(A, i, j)$ on an array of $n$ elements over some set $U$, denoted by $A[1..n]$, takes two indices $1 \leq i \leq j \leq n$, a function $f$ defined over arrays of elements of $U$, and outputs $f(A[i..j]) = f(A[i], \ldots, A[j])$. Range query data structures in 1d can thus be viewed as data structures answering queries on a string in the internal setting, where $U$ is the considered alphabet.

Asking internal queries on a string has received much attention in recent years. In the internal setting, we are asked to preprocess a string $T$ of length $n$ over an alphabet $\Sigma$ of size $\sigma$, so that queries about substrings of $T$ can be answered efficiently. Note that an arbitrary substring of $T$ can be encoded in $O(1)$ words of space by the indices $i, j$ of its occurrence as a fragment $T[i] \cdots T[j]$ of $T$. Data structures for answering internal queries are interesting in their own sake, but also have numerous applications in the design of algorithms and (more sophisticated) data structures in stringology. Because of these numerous applications, we usually place particular emphasis on the construction time—other than on space/query-time tradeoffs, which is the main focus in the classic data structure literature.

In data structures on strings it is typically assumed that the input alphabet is integer and polynomially bounded, i.e., it is a subset of $\{1, 2, \ldots, nO(1)\}$ where $n$ is the length of the input string $T$. The most widely-used internal query is that of asking for the longest common prefix of two suffixes $T[i..n]$ and $T[j..n]$ of $T$. The classic data structure for this problem [45] consists of the suffix tree of $T$ [25] and a lowest common ancestor data structure [87] over the suffix tree. It occupies $O(n)$ space, it can be constructed in $O(n)$ time, and it answers queries in $O(1)$ time. In the word RAM model of computation with word size $\Theta(\log n)$ bits the construction time is not necessarily optimal when the input alphabet is $\{1, 2, \ldots, \sigma\}$ and the string is packed into $O(n/\log \sigma, n)$ machine words. A sequence of works [57, 49, 13] has culminated in the recent
optimal data structure of Kempa and Kociumaka [40]: it occupies $O(n/\log_\sigma n)$ space, it can be constructed in $O(n/\log_\sigma n)$ time, and it answers queries in $O(1)$ time.

Another fundamental problem in this setting is the internal pattern matching (IPM) problem. It consists in preprocessing $T$ so that we can efficiently compute the occurrences of a substring $U$ of $T$ in another substring $V$ of $T$. For the decision version of the IPM problem, Keller et al. [39] presented a data structure of nearly-linear size supporting sublogarithmic-time queries. Kociumaka et al. [44] presented a data structure of linear size supporting constant-time queries when the ratio between the lengths of $V$ and $U$ is bounded by a constant. The $O(n)$-time construction algorithm of the latter data structure was derandomised in [42]. In fact, Kociumaka et al. [44], using their efficient IPM queries as a subroutine, managed to show efficient solutions for other internal problems, such as for computing the periods of a substring (period queries, introduced in [43]), and for checking whether two substrings are rotations of one another (cyclic equivalence queries). Other problems that have been studied in the internal setting include string alignment [58] [13], approximate pattern matching [21], dictionary matching [20] [19], longest common substring [4], counting palindromes [55], range longest common prefix [3, 1, 46, 34], the computation of the lexicographically minimal approximate pattern matching [21], dictionary matching [20, 19], longest common substring [4], counting cyclic equivalence introduced in [43], and for checking whether two substrings are rotations of one another ([43]).

In this work, we extend this line of research by investigating the following basic internal query, which, to the best of our knowledge, has not been studied previously. Given a string $T$ of length $n$ over an alphabet $\Sigma \subset \{1, 2, \ldots, n^{O(1)}\}$, preprocess $T$ so that given a range $[i, j]$, we can return a shortest string over $\Sigma$ that does not occur in $T[i..j]$. The latter shortest string is also known as a shortest absent word in the literature. We work on the standard unit-cost word RAM model with machine word-size $w = \Theta(\log n)$ bits. We measure the space used by our algorithms and data structures in machine words, unless stated otherwise. We assume that we have random access to $T$ and so our algorithms return a constant-space representation of a shortest string (a witness) consisting of a substring of $T$ and a letter. A na"ive solution for this problem precomputes a table of size $O(n^2)$ that stores the answer for every possible query $[i, j]$. Our main result is the following theorem.

**Theorem 1.** Given a string $T$ of length $n$ over an alphabet $\Sigma \subset \{1, 2, \ldots, n^{O(1)}\}$ of size $\sigma$, we can construct in $O(n \log_\sigma n)$ time a data structure of size $O(n)$ that, for any given query $[a, b]$, can compute in $O(1)$ time a shortest string over $\Sigma$ that does not occur in $T[a..b]$.

In an earlier conference version of the present paper [8], we have obtained a weaker result: a data structure of size $O((n/k) \cdot \log \log_\sigma n)$ that can answer queries in $O(\log \log_\sigma k)$ time, where $k$ is a user-defined parameter from $[1, \log \log_\sigma n]$. Here, we remove the parameter, thus improving the result.

In the related range shortest unique substring problem, defined by Abedin et al. [2], the task is to construct a data structure over $T$ to be able to answer the following type of online queries efficiently. Given a range $[i, j]$, return a shortest string with exactly one occurrence (starting position) in $[i, j]$. Abedin et al. presented a data structure of size $O(n \log n)$ supporting $O(\log w)$-time queries, where $w = \Theta(\log n)$ is the word size. Additionally, Abedin et al. [2] presented a data structure of size $O(n)$ supporting $O(\sqrt{n} \log^\epsilon n)$-time queries, where $\epsilon$ is an arbitrarily small positive constant.

**Our Techniques.** For clarity of exposition, in this overview, we skip the time-efficient construction algorithms of our data structures and only describe how to compute the length of a shortest absent word (without a witness) in $T[a..b]$; note that this length is at most $\log_\sigma n$. Let us also note that the length of a shortest absent word of $T$ can be computed in $O(n)$ time using the suffix tree of $T$ [25]. It suffices to traverse the suffix tree of $T$ recording the shortest string-depth $\ell$, where an implicit or explicit node has less than $\sigma$ outgoing edges.

**First approach:** We precompute, for each position $i$ and for each length $j \in [1, \log_\sigma n]$, the starting position of the shortest suffix of $T[1..i]$ that contains an occurrence of each of the $\sigma^j$ distinct words of length $j$. Then, a query for the length of a shortest absent word of $T[a..b]$ reduces to a predecessor query among the starting positions we have precomputed for position $b$. By maintaining these $O(\log_\sigma n)$ starting positions in a fusion tree [32], we obtain a data structure of size $O(n \log_\sigma n)$ supporting queries in $O(\log_\sigma \log n) = O(1)$ time.
Second approach: We precompute, for each length \( j \in [1, \log_\sigma n] \), all minimal fragments of \( T \) that contain an occurrence of each of the distinct \( \sigma^j \) words of length \( j \). As these fragments are inclusion-free, we can encode them using two \( n \)-bit arrays storing their starting and ending positions in \( T \), respectively. We thus require \( O(n) \) words of space in total over all \( js \). Observe that \( T[a..b] \) does not have an absent word of length \( j \) if and only if it contains a minimal fragment for length \( j \); we can check this condition in \( O(1) \) time after augmenting the computed bit arrays with succinct rank and select data structures \( \mathcal{AS} \). Finally, due to monotonicity (if \( T[a..b] \) contains all strings of length \( j + 1 \) then \( T \) contains all strings of length \( j \)), we can binary search for the answer in \( O(\log \log_\sigma n) \) time.

Third approach: We optimize the first approach by utilizing succinct fusion trees to store the sets of size \( O(\log_\sigma n) \) associated with positions of \( T \), thus reducing the space on top of the sets to \( O(n \log_\sigma \log n) \).

Instead of storing the \( O(\log_\sigma n) \)-size sets explicitly, we compute their elements on demand using \( O(\log_\sigma n) \) select data structures, each occupying \( \mathcal{O}(n) \) bits. This leads to an \( O(n \log_\sigma \log n) \)-space solution. In order to optimize it further, we rely on the following combinatorial observation: if the length of a shortest absent word of a string \( X \) over \( \Sigma \) is \( \lambda \), we need to append \( \Omega(\sigma^{d-1} \cdot \lambda) \) letters to \( X \) in order to obtain a string with a shortest absent word of length \( \lambda + d \). (For intuition, think of \(|X|\) as a constant; then, we essentially need to append the de Bruijn sequence of order \( d \) over \( \Sigma \) to \( X \) in order to achieve the desired result.) This observation allows us to lower the memory consumption by truncating all succinct fusion trees at positions \( i \) that are not multiples of \( \log \log n \), by storing only their first \( O(\log n/\log \log n) \) entries. The total space thus reduces to \( O(n) \) words. A query for the length of a shortest absent word of \( T[a..b] \) is performed by first checking whether the answer is at most \( \log n/\log \log n \), which is done using the (truncated) fusion tree stored at \( b \), and, if not, a query on \( T[a..b'] \) is performed, where \( b' \) is the closest multiple of \( \log \log n \) after \( b \).

It can be shown using the combinatorial observation that the answer for \( T[a..b] \) is within an \( O(1) \)-length range of the answer for \( T[a..b'] \) and it is computed by the data structure from the second approach.

Other Related Work. Let us recall that a string \( S \) that does not occur in \( T \) is called absent from \( T \), and if all its proper substrings appear in \( T \) it is called a minimal absent word of \( T \). It should be clear that every shortest absent word is also a minimal absent word. Minimal absent words (MAWs) are used in many applications \([16, 22, 23, 35, 14, 11, 24]\) and their theory is well developed \([18, 27, 29]\), also from an algorithmic and data structure point of view \([17, 22, 9, 17, 16, 5, 33, 10, 23]\). For example, it is well known that, given two strings \( X \) and \( Y \), one has \( X = Y \) if and only if \( X \) and \( Y \) have the same set of MAWs \([48]\).

Paper Organization. Section 2 provides some preliminaries. The first approach is detailed in Section 3 and the second one in Section 4. Section 5 provides the combinatorial foundations for the third approach, which is detailed in Section 6. The material of Sections 3-6 essentially repeats the conference version of our paper \([3]\); the main difference and novelty is in Section 6.

2. Preliminaries

An alphabet \( \Sigma \) is a finite nonempty set whose elements are called letters. A string (or word) \( S = S[1..n] \) is a sequence of length \( |S| = n \) over \( \Sigma \). The empty string \( \varepsilon \) is the string of length 0. The concatenation of two strings \( S \) and \( T \) is the string composed of the letters of \( S \) followed by the letters of \( T \). It is denoted by \( S \cdot T \) or simply by \( ST \). The set of all strings (including \( \varepsilon \)) over \( \Sigma \) is denoted by \( \Sigma^* \). The set of all strings of length \( k > 0 \) over \( \Sigma \) is denoted by \( \Sigma^k \). For \( 1 \leq i \leq j \leq n \), \( S[i] \) denotes the \( i \)th letter of \( S \), and the fragment \( S[i..j] \) denotes an occurrence of the underlying substring \( P = S[i] \cdots S[j] \). We say that \( P \) occurs at (starting) position \( i \) in \( S \). A string \( P \) is called absent from \( S \) if it does not occur in \( S \). A substring \( S[i..j] \) is a suffix of \( S \) if \( j = n \) and it is a prefix of \( S \) if \( i = 1 \).

The following proposition is straightforward (as explained in Section 1).

**Proposition 1.** Let \( T \) be a string of length \( n \) over an alphabet \( \Sigma \subset \{1, 2, \ldots, n^{O(1)}\} \). A shortest absent word of \( T \) can be computed in \( O(n) \) time.

Given an array \( A \) of \( n \) items taken from a totally ordered set, the range minimum query \( \text{RMQ}_A(\ell, r) = \arg \min A[k] \) (with \( 1 \leq \ell \leq k \leq r \leq n \)) returns the position of the minimal element in \( A[\ell..r] \). The following result is known.
Theorem 2 ([31]). Let $A$ be an array of $n$ integers. A data structure of size $2n + o(n)$ bits can be constructed in $O(n)$ time supporting RMQs on $A$ in $O(1)$ time without the need to store and access $A$ itself.

We make use of rank and select data structures constructed over bit vectors. For a bit vector $H$ we define $\text{rank}_q(i, H) = \{k \in [1, i] : H[k] = q\}$ and $\text{select}_q(i, H) = \min\{k \in [1, n] : \text{rank}_q(k, H) = i\}$, for $q \in \{0, 1\}$. The following result is known.

Theorem 3 ([50]). Let $H$ be a bit vector of $n$ bits. A data structure of $o(n)$ additional bits can be constructed in $O(n)$ time supporting rank and select queries on $H$ in $O(1)$ time.

The static predecessor problem consists in preprocessing a set $Y$ of integers, over an ordered universe $U$, so that, for any integer $x \in U$ one can efficiently return the predecessor $\text{pred}(x) := \max\{y \in Y : y \leq x\}$ of $x$ in $Y$. The successor problem is defined analogously: upon a queried integer $x \in U$, the successor $\min\{y \in Y : y \geq x\}$ of $x$ in $Y$ is to be returned. Willard and Fredman designed the fusion tree data structure for this problem [52]. In the dynamic variant of the problem, updates to predecessor and successor queries. Pătraşcu and Thorup [53] presented a dynamic version of fusion trees, which, in particular, yields an efficient construction of this data structure.

Theorem 4 ([53]). Let $Y$ be a set of at most $n$ $w$-bit integers. A data structure of size $O(n)$ can be constructed in $O(n \log_w n)$ time supporting insertions, deletions, and predecessor queries on $Y$ in $O(\log_w n)$ time.

We also use a succinct version of the (static) fusion tree that utilizes only $O(n \log w)$ bits on top of a read-only array $Y$ of length $n$ (in contrast, the fusion tree from Theorem 4 uses $O(nw)$ bits). In this data structure there is no need to store the array $Y$ explicitly, $Y$ can be “emulated” by computing its elements on demand in $O(1)$ time.

Theorem 5 ([53]). Let $Y$ be a read-only array of at most $n$ $w$-bit integers. A data structure of size $O(\log w)$ bits can be constructed in $O(n \log_w n)$ time supporting predecessor queries on the elements of $Y$ in $O(\log_w n)$ time, provided a table computable in $o(2^w)$ time and independent of the array has been precomputed.

If $|U| = O(n)$, then, after an $O(n)$-time preprocessing, we can answer predecessor queries over the integer universe $U$ in $O(1)$ time as follows. For each $y \in U$, we set the $y$th bit of an initially all-zeros $|U|$-size bit vector. We then preprocess this bit vector as in theorem 3. Then, a predecessor query for any integer $x$ can be answered in $O(1)$ time due to the following readily verifiable formula: $\text{pred}(x) = \text{select}_1(\text{rank}_1(x))$.

The main problem considered in this paper is formally defined as follows.

**Internal Shortest Absent Word (ISAW)**

**Input:** A string $T$ of length $n$ over an alphabet $\Sigma \subseteq \{1, 2, \ldots, n^{O(1)}\}$ of size $\sigma > 1$.

**Output:** Given integers $a$ and $b$, with $1 \leq a \leq b \leq n$, output a shortest string in $\Sigma^*$ with no occurrence in $T[a \ldots b]$.

If $a = b$ then the answer is trivial. So, in what follows we assume that $a < b$. Let us also remark that the output (shortest absent word) can be represented in $O(1)$ space using: either a range $[i, j] \subseteq [1, n]$ and a letter $\alpha$ of $\Sigma$, such that the shortest string in $\Sigma^*$ with no occurrence in $T[a \ldots b]$ is $T[i \ldots j] \alpha$; or simply a range $[i, j] \subseteq [1, n]$ such that the shortest string in $\Sigma^*$ with no occurrence in $T[a \ldots b]$ is $T[i \ldots j]$.

**Example 1.** Given the string $T =abaabaaabbabbbbaab$ and the range $[a, b] = [8, 14]$ (shown in red), the only shortest absent word of $T[8 \ldots 14]$ is $T[i \ldots j] = T[7 \ldots 8] = aa$.

3. **$O(n \log_{\sigma} n)$ Space and $O(1)$ Query Time**

Let $T$ be a string of length $n$. We define $S_T(j)$ as the function counting the cardinality of the set of length-$j$ substrings of $T$. This is known as the *substring complexity* function [26] [54]. Note that $S_T(j) \leq n$, for all $j$. We have the following simple fact.
The length $\ell$ of a shortest absent word of a string $T$ of length $n$ over an alphabet of size $\sigma$ is equal to the smallest $j$ for which $S_T(j) < \sigma^j$ and hence $\ell \in [1, \lfloor \log_\sigma n \rfloor]$. We denote the set of shortest absent words of $T$ by $SAW_T$. Recall that, by Proposition 1, a shortest absent word of $T$ can be computed in $O(n)$ time. We denote the length of the shortest absent words of $T$ by $\ell$. By Fact 2, $\ell \leq \lfloor \log_\sigma n \rfloor$. Since $\ell$ is an upper bound on the length of the answer for any ISAW query on $T$, in what follows, we consider only lengths in $[1, \ell - 1]$. Let one such length be denoted by $j$. By constructing and traversing the suffix tree of $T$, we can assign to each $T[i \ldots i + j - 1]$ its lexicographic rank in $\Sigma^j$. The time required for each length $j$ is $O(n)$, since the suffix tree of $T$ can be constructed within this time. Thus, the total time for all lengths $j \in [1, \ell - 1]$ is $O(n \log_\sigma n)$ by Fact 6.

We design the following warm-up solution to the ISAW problem. For all $j \in [1, \ell - 1]$ we store an array $RNK_j$ of $n$ integers such that $RNK_j[i]$ is equal to the lexicographic rank of $T[i \ldots i + j - 1]$ in $\Sigma^j$. Then, given a range $[a, b]$, in order to check if there is an absent word of length $j$ in $T[a \ldots b]$ we only need to compute the number of distinct elements in $RNK_j[a \ldots b - j + 1]$. It is folklore that using a persistent segment tree, we can preprocess an array $A$ of $n$ integers in $O(n \log n)$ time so that upon a range query $[a, b]$ we can return the number of distinct elements in $A[a \ldots b]$ in $O(\log n)$ time. Thus, we could use this tool as a black box for every array $RNK_j$ resulting, however, in $O(\log n)$-time queries. We improve upon this solution as follows.

We employ a range minimum query (RMQ) data structure over a slight modification of $RNK_j$. For each $j$, we have an auxiliary procedure checking whether all strings from $\Sigma^j$ occur in $T[a \ldots b]$ or not (i.e., it suffices to check whether any lexicographic rank is absent from the corresponding range). Similar to the previous solution, we rank the elements of $\Sigma^j$ by their lexicographic order. We append $RNK_j$ with all integers in $[1, \sigma^j]$. Let this array be $APP_j$. By Fact 6, we have that $|APP_j| \leq 2n$. Then, we construct an array $PRE_j$ of size $|APP_j|$: $PRE_j[i]$ stores the position of the rightmost occurrence of $APP_j[i]$ in $APP_j[i \ldots i - 1]$ (or $0$ if such an occurrence does not exist). This can be done in $O(n)$ time per $j$ by sorting the list of pairs $(i \ldots i + j - 1, i)$, for all $i$, using the suffix tree of $T$ to assign ranks for $T[i \ldots i + j - 1]$ and then radix sort to sort the list of pairs.

We now rely on the following fact.

Fact 7. $S_T[a \ldots b](j) = \sigma^j$ if and only if $\min\{PRE_j[i] : i \in [b - j + 2, |PRE_j|]\} \geq a$.

Proof. If the smallest element in $P[\ldots b - 2 \ldots |P|]$, say $P[k]$, is such that $P[k] \geq a$, then all ranks of elements in $\Sigma^j$ occur in $P[a \ldots b - j + 1]$. This is because all elements (ranks) in $\Sigma^j$ occur at least once after $b - j + 2$ (due to appending all integers in $[1, \sigma^j]$ to $RNK_j$), thus all must have a representative occurrence after $b - j + 2$. Inspect Figure 1 for an illustration. (The opposite direction is analogous.)

The following two examples illustrate the construction of arrays $RNK_j$, $APP_j$, and $PRE_j$ as well as fact 7.

Example 2 (Construction). Let $T = \texttt{ababababbbababaa}$ and $\Sigma = \{a, b\}$. The set $SAW_T$ of shortest absent words of $T$ over $\Sigma$, each of length $\ell = 4$, is $\{\texttt{aaaa}, \texttt{abab}, \texttt{baba}, \texttt{bbbb}\}$. Arrays $RNK_j$, $APP_j$, and $PRE_j$, for all $j \in [1, \ell - 1]$, are as depicted in Table 1. For instance, $RNK_2[15] = APP_2[15] = 1$ denotes that the lexicographic rank of $\texttt{aa}$ in $\Sigma^2$ is 1; and $PRE_2[15] = 7$ denotes that the previous rightmost occurrence of $\texttt{aa}$ is at position 7.

5
Table 1: Arrays \( \text{RNK}_j \), \( \text{APP}_j \), and \( \text{PRE}_j \) in Example 3

| \( i \) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| \( T \) | a | b | a | b | a | a | b | b | b | b | a | a | a | b | b | b | b | b | a | a | b | b | b | b |
| \( \text{RNK}_1 \) | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| \( \text{APP}_1 \) | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 |
| \( \text{PRE}_1 \) | 0 | 0 | 1 | 3 | 2 | 4 | 6 | 7 | 5 | 9 | 8 | 10 | 12 | 13 | 11 | 5 | 6 | 11 | 17 | 16 | 14 | 17 | 18 |
| \( \text{RNK}_2 \) | 2 | 3 | 1 | 2 | 3 | 1 | 1 | 2 | 4 | 3 | 2 | 4 | 4 | 3 | 1 | 2 | 1 | 2 | 3 | 4 |
| \( \text{APP}_2 \) | 2 | 3 | 1 | 2 | 3 | 1 | 1 | 2 | 4 | 3 | 2 | 4 | 4 | 3 | 1 | 2 | 1 | 2 | 3 | 4 |
| \( \text{PRE}_2 \) | 0 | 0 | 0 | 1 | 2 | 3 | 6 | 4 | 0 | 5 | 8 | 9 | 12 | 10 | 7 | 15 | 11 | 16 | 17 | 14 | 13 |
| \( \text{RNK}_3 \) | 3 | 5 | 2 | 3 | 5 | 1 | 2 | 4 | 7 | 6 | 4 | 8 | 7 | 5 | 1 | 2 |
| \( \text{APP}_3 \) | 3 | 5 | 2 | 3 | 5 | 1 | 2 | 4 | 7 | 6 | 4 | 8 | 7 | 5 | 1 | 2 |
| \( \text{PRE}_3 \) | 0 | 0 | 0 | 1 | 2 | 3 | 6 | 4 | 0 | 5 | 8 | 9 | 12 | 10 | 7 | 15 | 16 | 14 | 11 | 14 | 10 | 13 | 12 |

Example 3 (Fact 7). Let \([a, b] = [7, 11]\) and \( j = 2\) (see Example 2). The smallest element in \( \{\text{PRE}_2[11], \ldots, \text{PRE}_2[21]\}\) is \( \text{PRE}_2[15] = 7 \geq a = 7\), which corresponds to rank \( \text{APP}_2[15] = 1\). Indeed all other ranks 2, 3, 4 have at least one occurrence within \( \text{APP}_2[7..11] = 1, 2, 4, 3, 2\).

To apply Fact 7, we construct an RMQ data structure over \( \text{PRE}_j \). By Theorem 2 it takes \( O(n) \) time and space and answers RMQs in \( O(1) \) time. This results in \( O(n\ell) = O(n \log_n \sigma) \) preprocessing time and space for all \( j \).

For querying, let us observe that \( \sigma^j = S_{T[a..b]}(j) \), for any \( T, a, b \) and increasing \( j \), is non-decreasing. We can thus apply binary search on \( j \) to find the smallest length \( j \) such that \( S_{T[a..b]}(j) < \sigma^j \). This results in \( O(\log \ell) = O(\log \log_n \sigma) \) query time. We obtain the following proposition (retrieving a witness shortest absent word is detailed later).

Proposition 2. Given a string \( T \) of length \( n \) over an alphabet \( \Sigma \subset \{1, 2, \ldots, n^{O(1)}\} \) of size \( \sigma \), we can construct a data structure of size \( O(n \log_n \sigma) \) in \( O(n \log_n \sigma) \) time, so that if query \([a, b] \) is given, we can compute a shortest string over \( \Sigma \) that does not occur in \( T[a..b] \) in \( O(\log \log_n \sigma) \) time.

We further improve the query time via employing fusion trees as follows. We create a 2d array \( \text{FTR}[1..\ell-1][1..n] \) of integers, where

\[
\text{FTR}[j][i] = \min\{\text{PRE}_j[i-j+2], \ldots, \text{PRE}_j[|\text{PRE}_j|]\},
\]

for all \( j \in [1, \ell-1] \) and \( i \in [1, n] \). Intuitively, \( \text{FTR}[j][i] \) is the rightmost index of \( T \) such that \( T[\text{FTR}[j][i]..i] \) contains all strings of length \( j \) over \( \Sigma \).

Array \( \text{FTR} \) can be constructed in \( O(n\ell) = O(n \log_n \sigma) \) time by scanning each array \( \text{PRE}_j \) from right to left maintaining the minimum. Within the same complexities we also maintain satellite information specifying the index \( k \in [i-j+2, |\text{PRE}_j|] \) where the range minimum \( \text{FTR}[j][i] \) came from in the sub-array \( \text{PRE}_j[i-j+2..|\text{PRE}_j|] \). We then construct \( n \) fusion trees, one for every collection of \( \ell-1 \) integers in \( \text{FTR}[1..\ell-1][i] \). This takes total preprocessing time and space \( O(n\ell) = O(n \log_n \sigma) \) by Theorem 4. Given the range query \([a, b] \), we need to find the smallest \( j \in [1, \ell-1] \) such that \( \text{FTR}[j][b] < a \). By Theorem 4 we find where the predecessor of \( a \) lies in \( \text{FTR}[1..\ell-1][b] \) in \( O(\log \ell \; \text{time}) \), where \( w \) is the word size; this time cost is \( O(1) \) since \( w = \Theta(\log n) \).

We finally retrieve a witness shortest absent word as follows. If there is no \( j \leq \ell \) such that \( \text{FTR}[j][b] < a \), then we output any shortest absent word of length \( \ell \) of \( T \) arbitrarily. If such a \( j < \ell \) exists, by the definition of \( \text{FTR}[j] \), we output \( T[\text{FTR}[j][b]..\text{FTR}[j][b] + j - 1] \) if \( \text{FTR}[j][b] > 0 \) or \( T[k..k+j-1] \) if \( \text{FTR}[j][b] = 0 \), where \( k \) is the index of \( \text{PRE}_j \), where the minimum came from. Inspect the following illustrative example.

Example 4 (Querying). We construct array FTR for \( T \) from Example 2. For a given \([a, b] \) we look up column \( b \), and find the topmost entry whose value is less than \( a \). If all entries have values greater than or equal to \( a \), we output any element from \( \text{SAW}_T \) arbitrarily.

If \([a, b] = [3, 14]\) then no entry in column \( b = 14 \) is less than \( a = 3 \), which means the length of the shortest absent word is \( 4 \); we output one from \( \{aaaa, abab, baba, bbbb\} \) arbitrarily. If \([a, b] = [5, 14] \) then
Theorem 8. Given a string $T$ of length $n$ over an alphabet $\Sigma \subset \{1, 2, \ldots, n^{O(1)}\}$ of size $\sigma$, we can construct a data structure of size $O(n \log \sigma \log n)$ in $O(n \log \sigma \log n)$ time, so that if query $[a, b]$ is given, we can compute a shortest absent word of length $n \log \sigma$ that does not occur in $T[a..b]$ in $O(1)$ time.

### 4. $O(n)$ Space and $O(\log \log n)$ Query Time

**Definition 1 (Order-$j$ Fragment).** Given a string $T$ over an alphabet of size $\sigma$ and an integer $j$, $V$ is called an order-$j$ fragment of $T$ if and only if $T$ is a fragment of $T$ and $S_T(j) = \sigma^j$. $V$ is further called a minimal order-$j$ fragment of $T$ if $S_T(j) < \sigma^j$ and $S_T(j) < \sigma^j$ for $U = V[1..|V| - 1]$ and $Z = V[2..|V|]$.

In particular, minimal order-$j$ fragments are pairwise not included in each other. The following fact follows directly.

**Fact 9.** Given a string $T$ of length $n$ over an alphabet of size $\sigma$ and an integer $j$, for all $j \in [1, \log \sigma, n]$, we have $O(n)$ minimal order-$j$ fragments. Moreover, an arbitrary fragment $F$ of $T$ has $|F| = \sigma^j$ if and only if it contains at least one of these minimal fragments.

For each $j \in [1, \log \sigma, n]$, we consider all minimal order-$j$ fragments $T$, separately. We encode the minimal order-$j$ fragments of $T$ using two bit vectors $SP_j$ and $EP_j$, standing for starting positions and ending positions, respectively. Inspect the following example:

**Example 5.** We consider $T$ from Example 2 and $j = 2$.

<table>
<thead>
<tr>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>APP$_2$</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>PRE$_2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>10</td>
<td>7</td>
<td>15</td>
<td>11</td>
<td>16</td>
<td>17</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>SP$_2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EP$_2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>


We construct a rank and select data structure on $SP_j$ and $EP_j$, for all $j \in [1, \ell - 1]$ supporting $O(1)$-time queries. The overall space is $O(n)$ by Theorem 3 and Fact 6.

Let us now explain how this data structure enables fast computation of absent words of length $j$. Given a range $[a, b]$, by Fact 3 we only need to find whether $T[a..b]$ contains a minimal order-$j$ fragment. We can do this in $O(1)$ time using one rank and two select queries: $t = \text{rank}_j(a - 1, SP_j) + 1$; and $\text{select}_j(t, SP_j)$ and $\text{select}_j(t, EP_j)$. The select queries return the starting and ending positions, respectively, of the leftmost minimal order-$j$ fragment that is located after the position $a - 1$; it remains to check whether this minimal order-$j$ fragment is inside $[a, b]$. 

7
Example 6. We consider \( T, SP_2 \) and \( EP_2 \) from Example 5. Let \([a, b] = [5, 14]\). We have \( t = \text{rank}_1(a - 1, SP_2) + 1 = \text{rank}_1(4, SP_2) + 1 = 1 \), \( \text{select}_1(t, SP_2) = \text{select}_1(1, SP_2) = 5 < b = 14 \) and \( \text{select}_1(t, EP_2) = \text{select}_1(1, EP_2) = 10 < b = 14 \), which means \( T[5, 14] \) contains a minimal order-2 fragment.

Let us now describe a time-efficient construction of \( SP_j \) and \( EP_j \). We use arrays \( PRE_j \) and \( APP_j \) of \( T \), which are constructible in \( O(n) \) time (see Section 3). Recall that \( PRE_j[i] \) stores the starting position of the rightmost occurrence of rank \( APP_j[i] \) in \( APP_j[1 \ldots i - 1] \) (or 0 if such an occurrence does not exist). We apply Fact 7 as follows. We start with all bits of \( SP \) unset. Then, for each \( b \in [1, n] \) for which \( PRE_j[b - j + 1] < \min\{PRE_j[i] : i \in [b - j + 2, |PRE_j|]\} = a \), we set the \( b \)th bit of \( SP_j \) and the \( a \)th bit of \( EP_j \).

Example 7. We consider \( T, SP_2 \) and \( EP_2 \) from Example 5. We start by setting \( b = n = 18 \) and scan \( PRE_2 \) from right to left: we have \( a = 13 \) because \( \min\{PRE_2[i] : i \in [18, 21]\} = 13 \). This gives fragment \( T[13 \ldots 18] \), which is minimal since \( \text{PRE}_2[b - 1] = \text{PRE}_2[17] < 13 \). Then we set \( b = n - 1 = 17 \) and have \( a = 11 \) because \( \min\{\text{PRE}_2[i] : i \in [17, 21]\} = 11 \). This gives fragment \( T[11 \ldots 17] \), which is not minimal since \( \text{PRE}_2[b - 1] = \text{PRE}_2[16] \geq 11 \). Then we set \( b = n - 2 = 16 \) and have \( a = 11 \) because \( \min\{\text{PRE}_2[i] : i \in [16, 21]\} = 11 \). This gives fragment \( T[11 \ldots 16] \), which is minimal since \( \text{PRE}_2[b - 1] = \text{PRE}_2[15] < 11 \).

Lemma 1. \( SP_j \) and \( EP_j \) can be constructed in \( O(n) \) time.

For all \( j \), the construction time is \( O(n \ell) = O(n \log_\sigma n) \) by Theorem 3, Lemma 4 and Fact 5. All the arrays \( SP_j \) and \( EP_j \) in total occupy \( O(n \ell) = O(n \log_\sigma n) \) bits of space, which is \( O(n) \) space when measured in \( \Theta(\log \ell) \)-bit machine words. We obtain the following lemma.

Lemma 2. Given a string \( T \) of length \( n \) over an alphabet \( \Sigma \subset \{1, 2, \ldots, n^{O(1)}\} \) of size \( \sigma \), we can construct a data structure of size \( O(n) \) in \( O(n \log_\sigma n) \) time, so that if query \( (j, [a, b]) \) is given, we can check in \( O(1) \) time whether there is any string in \( \Sigma^j \) that does not occur in \( T[a \ldots b] \), and if so return such a string.

We can now apply Lemma 2 using binary search on \( j \) to find the smallest length \( j \) such that \( S_{T[a \ldots b]}(j) < \sigma^j \). This results in \( O(\log \ell) = O(\log \log_\sigma n) \) query time by Fact 5. It should now be clear that when we find the \( j \) corresponding to the length of a shortest absent word, we can output the length-\( j \) suffix of the leftmost minimal order-\( j \) fragment starting after \( a \). Note that outputting this suffix is correct by the definition of minimal order-\( j \) fragments.

Example 8. We consider \( T, SP_2 \) and \( EP_2 \) from Example 5. Let \([a, b] = [2, 7]\). The length of a shortest absent word of \( T[2 \ldots 7] \) is 2. We output \( b \), which is the length-2 suffix of the leftmost minimal order-2 fragment \( T[5 \ldots 10] = \text{baaabb} \) starting after \( a = 2 \).

We obtain the following result.

Theorem 10. Given a string \( T \) of length \( n \) over an alphabet \( \Sigma \subset \{1, 2, \ldots, n^{O(1)}\} \) of size \( \sigma \), we can construct a data structure of size \( O(n) \) in \( O(n \log_\sigma n) \) time, so that if query \( [a, b] \) is given, we can compute a shortest string over \( \Sigma \) that does not occur in \( T[a \ldots b] \) in \( O(\log \log_\sigma n) \) time.

5. Combinatorial Insights

A positive integer \( p \) is a period of a string \( S \) if \( S[i] = S[i + p] \) for all \( i \in [1, |S| - p] \). We refer to the smallest period as the period of the string. Let us state the periodicity lemma, one of the most elegant combinatorial results on strings.

Lemma 3 (Periodicity Lemma (weak version) [30]). If a string \( S \) has periods \( p \) and \( q \) such that \( p + q \leq |S| \), then \( \gcd(p, q) \) is also a period of \( S \).

Lemma 4. If all strings in \( \{UW : U \in \Sigma^k\} \) for \( W \neq \varepsilon \) occur in some string \( S \), then \( |S| \geq |W| \cdot \sigma^k/4 \).
Proof. Let \( p \) be the period of \( W \), and let \( a \in \Sigma \) be such that the period of \( aW \) is also \( p \). All strings \( ZbW \) for a letter \( b \neq a \) and \( Z \in \Sigma^{k-1} \) must occur in \( S \). Let \( A = \{uW : U \in \Sigma^k\} \setminus \{ZaW : Z \in \Sigma^{k-1}\} \), and note that it is of size \( \sigma^k - \sigma^{k-1} \geq \sigma^k/2 \). The following claim immediately implies the statement of the lemma.

Claim. Let \( i \) and \( j \) be starting positions of occurrences of different strings \( UW, VW \in A \) in \( S \), respectively. Then, we have \( j - i \geq |W|/2 \).

Proof. Let us assume, without loss of generality, that \( j > i \). Further, let us assume towards a contradiction that \( j - i < |W|/2 \). Then, \( j - i \) is a period of \( W \) and \( p + j - i \leq |W| \) since \( p \leq j - i \). Therefore, due to the periodicity lemma (Lemma 3), \( j - i \) must be divisible by the period \( p \) of \( W \). Hence, \( V \) ends with the letter \( a \) and \( VW \notin A \), a contradiction.

This concludes the proof of this lemma.

Lemma 5. If a shortest absent word of a string \( X \) is of length \( \lambda \), then the length of a shortest absent word of \( XY \) is in \([\lambda, \lambda + \max\{10, 4 + \log_\sigma(|Y|/\lambda)\}]\).

Proof. Let \( W \) and \( W' \) be shortest absent words of \( X \) and \( XY \), respectively. Further, let \( d = |W'| - |W| \). In order to have \( d > 0 \), all strings \( UW \) for \( U \in \Sigma^{d-1} \) must occur in \( XY \), and hence in \( X[|X| - |W| + 2 \cdots |X|]Y \), since none of them occurs in \( X \). Lemma 4 implies that \( |Y| + \lambda + d > \lambda \cdot \sigma^{d-1}/4 \). Then, since \( \lambda + d \leq 2md \) for any positive integers \( \lambda, d \), we have \( |Y| > \lambda \cdot (\sigma^{d-1}/4 - 2d) \). Assuming that \( d \geq 10 \), and since \( \sigma \geq 2 \), we conclude that \( |Y| > \lambda \cdot \sigma^{d-1}/8 \). Consequently, \( \log_\sigma(8)|Y|/\lambda) + 1 > d \). Since \( \log_\sigma 8 \leq 3 \) we get the claimed bound.

Lemma 6. If a shortest absent word of \( XY \) is of length \( m \), a shortest absent word of \( X \) is of length \( \lambda \), and \( |Y| \leq m \cdot \tau \), for a positive integer \( \tau \geq 16 \), then \( m - \lambda \leq 10 + 2 \log_\sigma \tau \).

Proof. From lemma 5 we have \( \lambda \in [m - \max\{10, 4 + \log_\sigma(|Y|/\lambda)\}, m] \). If \( \max\{10, 4 + \log_\sigma(|Y|/\lambda)\} = 10 \), then \( m - \lambda \leq 10 \) and we are done.

In the complementary case, since \( |Y| \leq m \cdot \tau \), we get the following:

\[
\lambda \geq m - \log_\sigma(m \cdot \tau/\lambda) - 4 \iff \lambda \geq m + \log_\sigma \lambda - \log_\sigma m - \log_\sigma \tau - 4.
\]

In particular, \( \lambda \geq m - \log_\sigma m - \log_\sigma \tau - 4 \).

From the above, if \( m \leq \tau \), then \( m - \lambda \leq 4 + 2 \log_\sigma \tau \).

In what follows we assume that \( m > \tau \geq 16 \). Rearranging the original equation, and since \( \log_\sigma(\cdot) \) is an increasing function and \( \lambda \geq m - \log_\sigma m - \log_\sigma \tau - 4 \), we have

\[
m - \lambda \leq 4 + \log_\sigma(m \cdot \tau/\lambda) \leq 4 + \log_\sigma(m - \log_\sigma m - \log_\sigma \tau - 4) + \log_\sigma \tau \leq 4 + \log_\sigma \left(\frac{m}{m - 2 \log_\sigma m - 4}\right) + \log_\sigma \tau.
\]

Then, we have \( m - 2 \log_\sigma m - 4 \geq m/5 \) since, for any \( \sigma \geq 2 \), \( 4x/5 - 2 \log_\sigma x - 4 \) is an increasing function on \([16, \infty)\) and positive for \( x = 16 \). Hence, \( m - \lambda \leq 4 + \log_\sigma 5 + \log_\sigma \tau \leq 7 + \log_\sigma \tau \).

By combining the bounds on \( m - \lambda \) we get the claimed bound.

6. \( O(n) \) Space and \( O(1) \) Query Time

Our linear-space solution of the ISAW problem with constant query time is an optimization of the \( O(n \log_n n) \)-space solution from Section 3 with some “boundary” cases processed using the data structure of Section 4. Let us first describe a simpler \( O(n \log_\sigma \log n) \)-space data structure, which will be then optimized using combinatorial insights from Section 5.

Recall that we denote by \( \ell \) the length of a shortest absent word of \( T \). The issue with the solution of Section 3 is that the 2d array \( \text{FTR}[1 \ldots \ell - 1][1 \ldots n] \), equipped with fusion trees, occupies \( O(n \log_\sigma n) \) space.
In order to reduce the memory consumption, we store the array \( FTR \) implicitly, computing its entries on demand, and utilize succinct fusion trees from Theorem 5 instead of usual fusion trees.

Recall that \( FTR[j][i] \) is the rightmost index of \( T \) such that \( T^{FTR}[j][i] \ldots i \) contains all strings of length \( j \) over \( \Sigma \) and it is equal to \( \min \{ \text{PRE}_{[i − j + 2], \ldots, \text{PRE}_{[i]}} \{ \text{PRE}_{[i]} \} \} \). Therefore, the content of the 2d array \( FTR \) can “emulated” without storing it explicitly if one can calculate in \( O(1) \) time the minima \( \min \{ \text{PRE}_{[a], \ldots, \text{PRE}_{[b]}} \{ \text{PRE}_{[a]} \} \} \), for any \( a \in [1, n] \). For \( j \in [1, \ell − 1] \) and \( a \in [1, n] \), denote \( M_{j,a} = \min \{ \text{PRE}_{[a], \ldots, \text{PRE}_{[b]}} \{ \text{PRE}_{[a]} \} \} \). Fix \( j \). Since the sequence \( M_{j,1}, M_{j,2}, \ldots, M_{j,n} \) is non-decreasing, we can encode it in a 2n-bit array \( B_j \) using a select data structure from Theorem 3 as follows: we construct \( B_j \) (initially empty) by considering \( a = 1, 2, \ldots, n \) in the increasing order and, for each \( a \), we append to the end of \( B_j \) exactly \( M_{j,a} − M_{j,a−1} \) zeroes followed by 1, assuming \( M_{j,0} = 0 \) (i.e., we append the number \( M_{j,a} − M_{j,a−1} \) written in unary); then, we have \( M_{j,a} = \text{select}_1(a, B_j) − a \).

**Example 9.** We consider \( T \) from Example 2 and \( j = 2 \).

| \( i \) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| \( T \) | a | b | a | b | a | a | b | b | a | b | b | b | a | a | b | b | b | a | a | b |
| \( \text{APP}_2 \) | 2 | 3 | 1 | 2 | 3 | 1 | 1 | 2 | 4 | 3 | 2 | 4 | 3 | 4 | 3 | 1 | 1 | 2 | 1 | 2 | 3 | 4 |
| \( \text{PRE}_2 \) | 0 | 0 | 0 | 1 | 2 | 3 | 6 | 4 | 0 | 5 | 8 | 9 | 12 | 10 | 7 | 15 | 11 | 16 | 17 | 14 | 13 |
| \( M_{2,i} \) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 7 | 7 | 7 | 7 | 7 | 11 | 11 | 13 |

In this case, we have \( B_2 = 1111111110000010011111000011001 \).

Besides the 2d array \( FTR \), the algorithm of Section 3 required to access the values \( \arg \min \{ \text{PRE}_{[a], \ldots, \text{PRE}_{[b]}} \{ \text{PRE}_{[a]} \} \} \) in order to retrieve a witness shortest absent word. To this end, we build a 2n-bit RMQ data structure from Theorem 2 on each array \( \text{PRE}_j \); the data structure does not need to store the array \( \text{PRE}_j \) itself to calculate \( \arg \min \). The arrays \( B_j \), for \( j \in [1, \ell − 1] \), equipped with select data structures, and the RMQ data structures on arrays \( \text{PRE}_j \), for \( j \in [1, \ell − 1] \), can be constructed in total \( O(n) = O(n \log \sigma_n) \) time and they altogether occupy \( O(n \log \sigma_n) \) bits of space, which is \( O(n) \) space when measured in machine words.

To answer a query \( [a, b] \), it suffices to find the smallest \( j \) such that \( FTR[j][b] < a \). We do this by finding where the predecessor of a lies in \( FTR[1, \ell − 1][b] \). To this end, we constructed \( n \) fusion trees: one per \( FTR[1, \ell − 1][i] \), resulting in a data structure of size \( \Theta(n) = O(n \log \sigma_n) \) with \( O(1) \) query time. But now we do not store the arrays \( FTR[1, \ell − 1][i] \) explicitly, while still having \( O(1) \)-time “oracle” access to their entries on demand. Hence, we can construct a succinct fusion tree of Theorem 2 for each array \( FTR[1, \ell − 1][i] \), which takes \( O(\ell \log \sigma_n) \) bits of space since the size of machine words is \( w = \log n \) bits (a shared table mentioned in Theorem 5 is also precalculated for all the trees in \( o(2^n) = o(n) \) time).

Thus, all the succinct fusion trees can be constructed in \( O(n \log \sigma_n) \) time and they occupy \( O(n \log \sigma_n \log \log n) \) bits, which is \( O(n \log \sigma_n \log n) \) space when measured in \( O(\log \sigma_n) \)-bit machine words. The \( \text{ISAW} \) queries are answered in \( O(1) \) time by the same algorithm as in Section 3.

Now we are to reduce further the memory usage of the data structure. We truncate all the arrays \( FTR[0, \ell − 1][i] \) except those where \( i \) is a multiple of \( \lfloor \log \log n \rfloor \): namely, if \( i \) is a multiple of \( \lfloor \log \log n \rfloor \), then the succinct fusion tree for the whole array \( FTR[1, \ell − 1][i] \) is stored, occupying \( O(\ell \log \log n) \) bits, by Theorem 5; otherwise \( i \) is not a multiple of \( \lfloor \log \log n \rfloor \), we store the succinct fusion tree only for the subarray \( FTR[1, \ell − 1][i] \), thus taking \( O(\log \log n) \) bits, by Theorem 5. In total, the space used is \( O((\log \log n) \ell \log \log n + n \log n) = O(n \log \sigma_n) \) in bits or \( O(n) \) in words.

In order to answer an \( \text{ISAW} \) query for \( T[a..b] \), we first check whether the length \( \lambda \) of a shortest absent word in \( T[a..b] \) is smaller than \( \log n \log \log n \) by querying the fusion tree of \( FTR[1, \ell − 1][b] \). If it is the case, then we have computed the length \( \lambda \) and we find the absent word itself using RMQs exactly as in the \( O(n \log \log n) \)-space solution described above.

Suppose that \( \lambda \geq \log n \log \log n \). We compute \( b' \), the closest multiple of \( \lfloor \log \log n \rfloor \) after \( b: b' = \lfloor b / \log \log n \rfloor \cdot \lfloor \log \log n \rfloor \). Observe that \( [a, b] \subseteq [a, b'] \). Then, using the fusion tree of \( FTR[1, \ell − 1][b'] \), we compute the smallest \( m \) such that \( FTR[m][b'] < a \) so that \( m = \text{length} \) of a shortest absent word in \( T[a, b'] \). Denote \( X = T[a, b] \) and \( T[a, b'] = XY \) where \( Y \) is a suffix of \( T[a, b'] \) of length \( b' − b \). We obviously
have $\lambda \leq m$. Since $|Y| = b' - b < \log \log n$ and $m \geq \log \log n / \log \log n$ so that $|Y| < m$, it follows from Lemma 5 that the answer $\lambda$ is within a range of length 18 from $m$. Therefore, $\lambda$ belongs to the range $[m - 18, m]$ and we can find it in $O(1)$ time using $O(1)$ queries of the $O(n)$-space data structure from Section 4. We thus arrive at the main result of the paper.

**Theorem 1.** Given a string $T$ of length $n$ over an alphabet $\Sigma \subset \{1, 2, \ldots, n^{O(1)}\}$ of size $\sigma$, we can construct in $O(n \log n)$ time a data structure of size $O(n)$ that, for any given query $[a, b]$, can compute in $O(1)$ time a shortest string over $\Sigma$ that does not occur in $T[a..b]$.

**References**


11