

Minimizing the Minimizers via Alphabet Reordering

Hilde Verbeek  



CWI, Amsterdam, The Netherlands

Lorraine A.K. Ayad  

Brunel University London, London, UK

Grigorios Loukides  

King's College London, London, UK

Solon P. Pissis  

CWI, Amsterdam, The Netherlands

Vrije Universiteit, Amsterdam, The Netherlands

Abstract

Minimizers sampling is one of the most widely-used mechanisms for sampling strings [Roberts et al., Bioinformatics 2004]. Let $S = S[1] \dots S[n]$ be a string over a totally ordered alphabet Σ . Further let $w \geq 2$ and $k \geq 1$ be two integers. The minimizer of $S[i \dots i + w + k - 2]$ is the smallest position in $[i, i + w - 1]$ where the lexicographically smallest length- k substring of $S[i \dots i + w + k - 2]$ starts. The set of minimizers over all $i \in [1, n - w - k + 2]$ is the set $\mathcal{M}_{w,k}(S)$ of the minimizers of S .

We consider the following basic problem:

Given S , w , and k , can we efficiently compute a total order on Σ that minimizes $|\mathcal{M}_{w,k}(S)|$?

We show that this is unlikely by proving that the problem is NP-hard for any $w \geq 3$ and $k \geq 1$. Our result provides theoretical justification as to why *there exist no exact algorithms* for minimizing the minimizers samples, while *there exists a plethora of heuristics* for the same purpose.

2012 ACM Subject Classification Theory of computation \rightarrow Pattern matching

Keywords and phrases sequence analysis, minimizers, alphabet reordering, feedback arc set

Digital Object Identifier 10.4230/LIPIcs.CPM.2024.28

Related Version *Extended Version*: <https://arxiv.org/abs/2405.04052>

Funding *Hilde Verbeek*: Supported by a Constance van Eeden Fellowship.

Solon P. Pissis: Supported by the PANGAIA and ALPACA projects that have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No 872539 and 956229, respectively.

1 Introduction

The minimizers sampling mechanism has been introduced independently by Schleimer et al. [17] and by Roberts et al. [16]. Since its inception, it has been employed ubiquitously in modern sequence analysis methods underlying some of the most widely-used tools [11, 12, 19].

Let $S = S[1] \dots S[n]$ be a string over a totally ordered alphabet Σ . Further let $w \geq 2$ and $k \geq 1$ be two integers. The minimizer of the fragment $S[i \dots i + w + k - 2]$ of S is the smallest position in $[i, i + w - 1]$ where the lexicographically smallest length- k substring of $S[i \dots i + w + k - 2]$ starts. We then define the set $\mathcal{M}_{w,k}(S)$ of the minimizers of S as the set of the minimizers positions over all fragments $S[i \dots i + w + k - 2]$, for $i \in [1, n - w - k + 2]$. Every fragment $S[i \dots i + w + k - 2]$ containing w length- k fragments is called a *window* of S .

► **Example 1.** Let $S = \text{aacaaacgcta}$, $w = 3$, and $k = 3$. Assuming $\text{a} < \text{c} < \text{g} < \text{t}$, we have that $\mathcal{M}_{w,k}(S) = \{1, 4, 5, 6, 7\}$. The minimizers positions are colored red: $S = \text{aacaaacgcta}$.



© Hilde Verbeek, Lorraine A.K. Ayad, Grigorios Loukides, and Solon P. Pissis; licensed under Creative Commons License CC-BY 4.0

35th Annual Symposium on Combinatorial Pattern Matching (CPM 2024).

Editors: Shunsuke Inenaga and Simon J. Puglisi; Article No. 28; pp. 28:1–28:13

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

28:2 Minimizing the Minimizers via Alphabet Reordering

Note that by choosing the *smallest* position in $[i, i + w - 1]$ where the lexicographically smallest length- k substring starts, we resolve ties in case the latter substring has multiple occurrences in a window.

It is easy to prove that minimizers samples enjoy the following three useful properties [23]:

- **Property 1 (approximately uniform sampling):** Every fragment of length at least $w + k - 1$ of S has at least one representative position sampled by the mechanism.
- **Property 2 (local consistency):** Exact matches between fragments of length at least $\ell \geq w + k - 1$ of S are preserved by means of having the same (relative) representative positions sampled by the mechanism.
- **Property 3 (left-to-right parsing):** The minimizer selected by any fragment of length $w + k - 1$ comes at or after the minimizers positions selected by all previous windows.

Since Properties 1 to 3 hold *unconditionally*, and since the ordering of letters does not affect the correctness of algorithms using minimizers samples [6, 18, 14, 1], one would like to choose the ordering that minimizes the resulting sample as a means to improve the space occupied by the underlying data structures; contrast Example 1 to the following example.

► **Example 2.** Let $S = \text{aacaaacgcta}$, $w = 3$, and $k = 3$. Assuming $c < a < g < t$, we have that $\mathcal{M}_{w,k}(S) = \{3, 6, 7\}$. The minimizers positions are colored red: $S = \text{aac} \color{red}{\text{aaac}} \text{gcta}$. In fact, this ordering is a best solution in minimizing $|\mathcal{M}_{w,k}(S)|$, together with the orderings $c < g < t < a$ and $c < g < a < t$, which both, as well, result in $|\mathcal{M}_{w,k}(S)| = 3$.

Our Problem. We next formalize the problem of computing a best such total order on Σ :

MINIMIZING THE MINIMIZERS

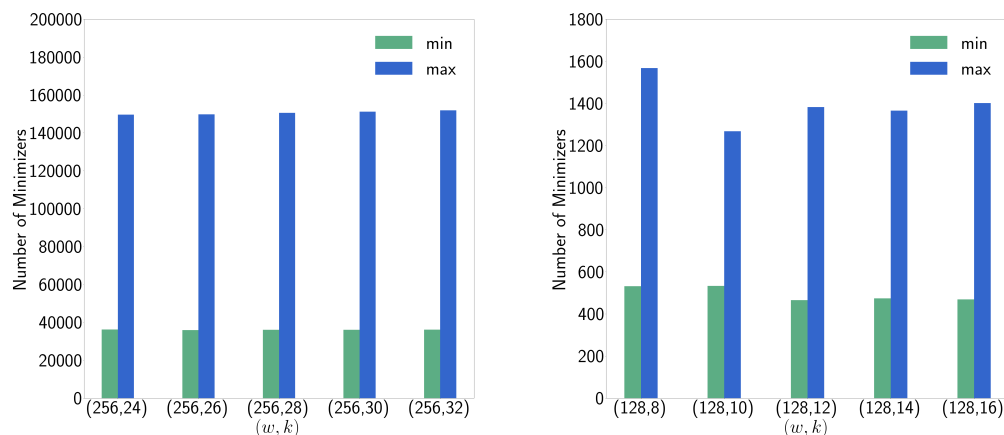
Input: A string $S \in \Sigma^n$ and two integers $w \geq 2$ and $k \geq 1$.

Output: A total order on Σ that minimizes $|\mathcal{M}_{w,k}(S)|$.

Motivation. A lot of effort has been devoted by the bioinformatics community to designing practical algorithms for minimizing the resulting minimizers sample [3, 4, 15, 21, 8, 22, 7]. Most of these approaches consider the space of all orderings on Σ^k (the set of all possible length- k strings on Σ) instead of the ones on Σ ; and employ *heuristics* to choose some ordering resulting in a small sample (see Section 3 for a discussion). To illustrate the impact of reordering on the number of minimizers, we considered two real-world datasets and measured the difference in the number of minimizers between the worst and best reordering, among those we could consider in a reasonable amount of time. The first dataset we considered is the complete genome of Escherichia coli str. K-12 substr. MG1655. For selecting minimizers, we considered different orderings on Σ^k . We thus mapped every length- k substring to its lexicographic rank in $\{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}^k$ (assuming $\mathbf{A} < \mathbf{C} < \mathbf{G} < \mathbf{T}$) constructing a new string S over $[1, |\Sigma|^k]$. We then computed $|\mathcal{M}_{w,1}(S)|$ for different values of (w, k) and orderings on $[1, |\Sigma|^k]$. It should be clear that this corresponds to computing the size of $\mathcal{M}_{w,k}$ for the original sequence over $\{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$. The second dataset is the complete genome of SARS-CoV-2 OL663976.1. Figure 1 shows the min and max values of the size of the obtained minimizers samples. The results in Figure 1 clearly show the impact of alphabet reordering on $|\mathcal{M}_{w,1}(S)|$: the gap between the min and max is quite significant as in all cases we have $2 \min < \max$. Note that we had to terminate the exploration of the whole space of orderings when $2 \min < \max$ was achieved; hence the presented gaps are not even the largest possible.

This begs the question:

Given S , w , and k , can we efficiently compute a total order on Σ that minimizes $|\mathcal{M}_{w,k}(S)|$?



(a) Complete genome of Escherichia coli.

(b) Complete genome of SARS-CoV-2.

■ **Figure 1** The min and max values of the size of the minimizers sample, among *some* of the possible orderings of $[1, |\Sigma|^k]$, on two real datasets using a range of (w, k) parameter values.

Our Contribution. We answer this basic question in the negative. Let us first define the decision version of MINIMIZING THE MINIMIZERS.

MINIMIZING THE MINIMIZERS (DECISION)

Input: A string $S \in \Sigma^n$ and three integers $w \geq 2$, $k \geq 1$, and $\ell > 0$.

Output: Is there a total order on Σ such that $|\mathcal{M}_{w,k}(S)| \leq \ell$?

Our main contribution in this paper is the following result.

► **Theorem 3.** *MINIMIZING THE MINIMIZERS (DECISION) is NP-complete if $w \geq 3$ and $k \geq 1$.*

Theorem 3 provides theoretical justification as to why *there exist no exact algorithms* for minimizing the minimizers samples, while *there exists a plethora of heuristics* for the same purpose. Notably, Theorem 3 almost completes the complexity landscape of the MINIMIZING THE MINIMIZERS problem – the only exception is the case $w = 2$ and $k \geq 1$. To cover *all practically interesting combinations* of input parameters w and k (i.e., for any $w \geq 3$ and $k \geq 1$), we design a non-trivial reduction from the feedback arc set problem [9].

The reduction we present is specifically for the case in which the size of the alphabet Σ is variable. If $|\Sigma|$ is bounded by a constant, the problem can be solved in polynomial time: one can simply iterate over the $|\Sigma|!$ permutations of the alphabet, compute the number of minimizers for each ordering in linear time [13], and output a globally best ordering.

Other Related Work. Choosing a best total order on Σ is generally not new; it has also been investigated in other contexts, e.g., for choosing a best total order for minimizing the number of runs in the Burrows-Wheeler transform [2]; for choosing a best total order for minimizing (or maximizing) the number of factors in a Lyndon factorization [5]; or for choosing a best total order for minimizing the number of bidirectional string anchors [14].

Paper Organization. Section 2 presents the proof of Theorem 3. Section 3 presents a discussion on orderings on Σ^k in light of Theorem 3. Final remarks are presented in Section 4.

2 Minimizing the Minimizers is NP-complete

We show that the MINIMIZING THE MINIMIZERS problem is NP-hard by a reduction from the well-known FEEDBACK ARC SET problem [9]. Let us first formally define the latter problem.

FEEDBACK ARC SET

Input: A directed graph $G = (V, A)$.

Output: A set $F \subseteq A$ of minimum size such that $(V, A \setminus F)$ contains no directed cycles.

We call any such $F \subseteq A$ a *feedback arc set*. The decision version of the FEEDBACK ARC SET problem is naturally defined as follows.

FEEDBACK ARC SET (DECISION)

Input: A directed graph $G = (V, A)$ and an integer $\ell' > 0$.

Output: Is there a set $F \subseteq A$ such that $(V, A \setminus F)$ contains no directed cycles and $|F| \leq \ell'$?

An equivalent way of phrasing this problem is to find an ordering on the set V of the graph's vertices, such that the number of arcs (u, v) with $u > v$ is minimal [20]. Then this is a topological ordering of the graph $(V, A \setminus F)$, and will be analogous to the alphabet ordering in the MINIMIZING THE MINIMIZERS problem; see [14] for a similar application of this idea.¹ If MINIMIZING THE MINIMIZERS is then solved on the instance constructed by our reduction, producing a total order on V , taking all arcs (u, v) with $u > v$ should produce a feedback arc set of minimum size, solving the original instance of the FEEDBACK ARC SET problem.

2.1 Overview of the Technique

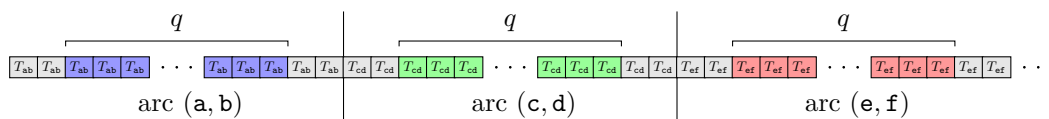
Given any instance $G = (V, A)$ of FEEDBACK ARC SET, we will construct a string S over alphabet V and of length polynomial in $|A|$. Specifically, we define string S as follows:

$$S = \prod_{(a,b) \in A} T_{ab}^{q+4},$$

where T_{ab} is a string consisting of the letters \mathbf{a} and \mathbf{b} , whose length depends only on w and k , and q is an integer polynomial in $|A|$, both of which will be defined later. The product \prod of some strings is defined as their concatenation, and X^q denotes q concatenations of string X starting with the empty string; e.g., if $X = \mathbf{ab}$ and $q = 4$, we have $X^q = (\mathbf{ab})^4 = \mathbf{abababab}$.

String T_{ab} will be designed such that each occurrence, referred to as a *block*, will contain few minimizers if $\mathbf{a} < \mathbf{b}$ in the alphabet ordering, and many minimizers if $\mathbf{b} < \mathbf{a}$, analogous to the “penalty” of removing the arc (\mathbf{a}, \mathbf{b}) as part of the feedback arc set. We denote by $M_{\mathbf{a} < \mathbf{b}}$ the number of minimizers starting within some occurrence of T_{ab} in S , provided that this T_{ab} is both preceded and followed by at least two other occurrences of T_{ab} (i.e., the middle q blocks), when $\mathbf{a} < \mathbf{b}$ in the alphabet ordering. We respectively denote by $M_{\mathbf{b} < \mathbf{a}}$ the number of minimizers starting in such a block when $\mathbf{b} < \mathbf{a}$ in the alphabet ordering. This will allow us (see Figure 2) to express the total number of minimizers in S in terms of $|F|$, the size of the feedback arc set, minus some discrepancy denoted by λ . This *discrepancy* is determined by the blocks T_{ab} that are not preceded or followed by two occurrences of T_{ab} itself; namely, those that occur near some T_{cd} , for another arc (c, d) , or those that occur near the start or the end of S .

¹ Our proof is more general and thus involved because it works for any values $w \geq 3$ and $k \geq 1$, whereas the reduction from [14] works only for some fixed parameter values.



■ **Figure 2** Illustration of the structure of string S , with the different gadgets for different arcs in G . The highlighted blocks are the ones for which the minimizers are counted in $M_{a < b}$ and $M_{b < a}$.

Let us start by showing an upper and a lower bound on the discrepancy λ .

► **Lemma 4.** $|A| - 1 \leq \lambda \leq 4 \cdot |A| \cdot |T_{ab}|$ if $|T_{ab}| \geq \frac{1}{4}(w + k - 1)$.

Proof. We are counting the number of minimizers in q blocks of T_{ab} , for each arc (a, b) . Note that we ignore four blocks for each arc, which is $4 \cdot |A|$ blocks of length $|T_{ab}|$ in total. This is $4 \cdot |A| \cdot |T_{ab}|$ positions in total, which gives the upper bound on the number of disregarded minimizers. For the lower bound, note that, by hypothesis, four consecutive blocks are at least as long as a single minimizer window, meaning at least one minimizer must be missed among the four blocks surrounding the border between each pair of consecutive arcs. The lower bound follows by the fact that for $|A|$ arcs we have $|A| - 1$ such borders. ◀

Given the values $M_{a < b}$, $M_{b < a}$, and λ , we can express the total number of minimizers as a function of some feedback arc set F : if an arc (a, b) is part of the feedback arc set, this corresponds to $b < a$ in the alphabet ordering, so the corresponding blocks T_{ab} will each have $M_{b < a}$ minimizers, whereas if (a, b) is not in F , we have $a < b$ and the blocks will each have $M_{a < b}$ minimizers. Using these values, we can define the number of minimizers in S given some feedback arc set F as

$$\begin{aligned} \mathcal{M}_{w,k}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda. \end{aligned} \quad (1)$$

With this in mind, we can prove the following relationship between $\mathcal{M}_{w,k}(S, F)$ and $|F|$:

► **Lemma 5.** Let ℓ' be some positive integer and let $\ell = q \cdot (M_{b < a} - M_{a < b}) \cdot (\ell' + 1) + q \cdot M_{a < b} \cdot |A|$. If $M_{b < a} > M_{a < b}$, $|T_{ab}| \geq \frac{1}{4}(w + k - 1)$, and q is chosen such that $\lambda < q \cdot (M_{b < a} - M_{a < b})$, then $\mathcal{M}_{w,k}(S, F) \leq \ell$ if and only if $|F| \leq \ell'$.

Proof. By hypothesis, $M_{b < a} - M_{a < b}$ is positive, thus, by Equation 1, $\mathcal{M}_{w,k}(S, F)$ grows linearly with $|F|$. Suppose we have a feedback arc set F with $|F| \leq \ell'$. Consider the alphabet ordering inducing F and let λ be the corresponding discrepancy for $\mathcal{M}_{w,k}(S, F)$. By hypothesis, we have $\lambda < q \cdot (M_{b < a} - M_{a < b})$. Substituting the bounds on $|F|$ and λ into Equation 1 gives

$$\begin{aligned} \mathcal{M}_{w,k}(S, F) &\leq q \cdot (M_{b < a} - M_{a < b}) \cdot \ell' + q \cdot M_{a < b} \cdot |A| + q \cdot (M_{b < a} - M_{a < b}) \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot (\ell' + 1) + q \cdot M_{a < b} \cdot |A| = \ell, \end{aligned}$$

completing the proof in one direction.

For the other direction, suppose we have picked F such that $\mathcal{M}_{w,k}(S, F) \leq \ell$ and assume that $|F| \geq \ell' + 1$ towards a contradiction. Then we have the following two inequalities:

$$\begin{aligned} \mathcal{M}_{w,k}(S, F) &\leq \ell = q \cdot (M_{b < a} - M_{a < b}) \cdot (\ell' + 1) + q \cdot M_{a < b} \cdot |A| \\ \mathcal{M}_{w,k}(S, F) &\geq q \cdot (M_{b < a} - M_{a < b}) \cdot (\ell' + 1) + q \cdot M_{a < b} \cdot |A| + \lambda. \end{aligned} \quad (\text{by Equation 1})$$

By Lemma 4, for any non-trivial instance with $|A| > 1$, λ is strictly positive, meaning these inequalities are contradictory. Therefore, if $\mathcal{M}_{w,k}(S, F) \leq \ell$, it must be that $|F| \leq \ell'$. ◀

28:6 Minimizing the Minimizers via Alphabet Reordering

Given w and k , we must determine a string T_{ab} such that $M_{b < a} > M_{a < b}$ and $|T_{ab}| \geq \frac{1}{4}(w + k - 1)$. We then simply have to choose some q , which is polynomial in $|A|$, satisfying $\lambda < q \cdot (M_{b < a} - M_{a < b})$. At that point we will have constructed a string S for which it holds that the feedback arc set induced by the minimum set of minimizers is also a minimum feedback arc set on G , thus completing the reduction.

The following three subsections address the T_{ab} construction:

- Section 2.2: $w \geq k + 2$ (Case A);
- Section 2.3: $w = 3$ and $k \geq 2$ (Case B);
- Section 2.4: $3 < w < k + 2$ (Case C).

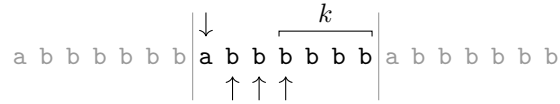
It should be clear that the above sections cover all the cases for $w \geq 3$ and $k \geq 1$. Section 2.5 puts everything together to complete the proof.

2.2 Case A: $w \geq k + 2$

► **Lemma 6.** *Let $T_{ab} = ab^{w-1}$, for $w \geq k + 2$. Then $M_{a < b} = 1$ and $M_{b < a} = w - k$.*

Proof. The block has length w ; inspect Figure 3. Recall that, for the window starting at position i , the candidates for its minimizer are the length- k fragments starting at positions $[i, i + w - 1]$. Therefore, for every window starting in a block T_{ab} (provided it is succeeded by another T_{ab}), a candidate minimizer is ab^{k-1} ; so if $a < b$, each T_{ab} will contain just one minimizer. Thus we have $M_{a < b} = 1$.

For $b < a$, consider that T_{ab} contains $w - k$ occurrences of b^k , and that for each window, at least one of the candidates for its minimizer is b^k . Since there is no length- k substring that is lexicographically smaller than b^k , each occurrence of b^k (and nothing else) is a minimizer, so it follows that $M_{b < a} = w - k$. Note that $M_{b < a} > M_{a < b}$ only if $w \geq k + 2$. ◀



■ **Figure 3** Illustration of 3 copies of T_{ab} in S for $w = 7$ and $k = 4$, along with its respective minimizers when $a < b$ (top) and when $b < a$ (bottom). It can be seen that $M_{a < b} = 1$ and $M_{b < a} = 3$.

2.3 Case B: $w = 3$ and $k \geq 2$

► **Lemma 7.** *Let $T_{ab} = (ab)^t bb$ with $t = \lceil \frac{w+k}{2} \rceil$, for $w = 3$ and $k \geq 2$. Then $M_{a < b} = \lfloor \frac{k}{2} \rfloor + 3$ and $M_{b < a} = \lfloor \frac{k}{2} \rfloor + 4$.*

Proof. Since $w = 3$, for every window, the minimizer is one out of three length- k fragments; inspect Figure 4. Every a in the block has a b before it. For any window starting at a position preceding an a , two of the candidates start with a b and the other starts with an a . As an example consider the window $babab$ preceding an a in Figure 4. We have that the first and the third candidates start with a b and the second starts with an a . Therefore, if $a < b$, the candidate starting with an a will be chosen and every a in T_{ab} is a minimizer. Only the window starting at the third-to-last position of the block will not consider any length- k substring starting with an a as its minimizer, as therein we have three b 's occurring in a row. Since $k \geq 2$, the last b of the block will be chosen if $a < b$. Thus, $M_{a < b}$ counts every a and one b , which gives:

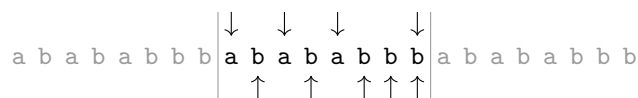
$$M_{a < b} = t + 1 = \left\lceil \frac{w+k}{2} \right\rceil + 1 = \left\lceil \frac{3+k}{2} \right\rceil + 1 = \left\lfloor \frac{k+2}{2} \right\rfloor + 1 + 1 = \left\lfloor \frac{k}{2} \right\rfloor + 3.$$

For $M_{b < a}$, we apply the same logic to conclude that every b surrounded by a 's is a minimizer, which accounts for all b 's except the final three, which occur at positions $[2t, 2t + 2]$:

- For the window starting at position $2t$, the three minimizer candidates start, respectively, with bb , bb and ba . Since $k \geq 2$, the first candidate ($2t$) will be the minimizer because it is lexicographically a smallest and the leftmost ($b < a$).
- For the window starting at position $2t + 1$, the first two candidates start, respectively, with bb and ba , and the third starts with an a . The first candidate ($2t + 1$) will be the minimizer, because it is lexicographically smaller ($b < a$).
- For the window starting at position $2t + 2$, the first and third candidates start with a b whereas the second starts with an a . The third candidate starts at the second position of the next T_{ab} -block. Since $2t > k + 1$, this candidate consists of only $baba\dots$ alternating for k letters. It is equal to the first candidate, so by tie-breaking the first candidate ($2t + 2$) is the minimizer as it is the leftmost.

Thus, every b in the block will be a minimizer if $b < a$, and we have:

$$M_{b < a} = t + 2 = \left\lceil \frac{3+k}{2} \right\rceil + 2 = \left\lfloor \frac{k}{2} \right\rfloor + 4. \quad \blacktriangleleft$$



■ **Figure 4** T_{ab} for $w = 3$ and $k = 3$, with its respective minimizers. The last b is a minimizer even when $a < b$, because $w = 3$. In this situation, $M_{a < b} = 4$ and $M_{b < a} = 5$.

2.4 Case C: $3 < w < k + 2$

► **Lemma 8.** Let $T_{ab} = (ab)^t bb$ with $t = \lceil \frac{w+k}{2} \rceil$, for $3 < w < k + 2$. Then

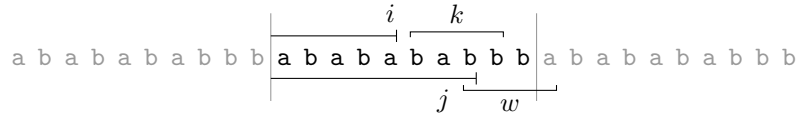
- if k is even, $M_{a < b} = \frac{k}{2} + 2 + p$ and $M_{b < a} = \frac{k}{2} + 3 + p$, where $p = (w + k) \bmod 2$;
- if k is odd, $M_{a < b} = \lfloor \frac{k}{2} \rfloor + 3$ and $M_{b < a} = \lfloor \frac{k}{2} \rfloor + 4$.

Proof. Every length- w fragment of the block contains at least one a and at least one b ; inspect Figure 5. Because of this, only a 's will be minimizers if $a < b$ and only b 's if $b < a$ (unlike when $w = 3$, as shown in Section 2.3). We start by counting $M_{a < b}$. Suppose we are determining the minimizer at position i . Every candidate we consider is a string of alternating a 's and b 's (starting with an a), in which potentially one a is substituted by a b (if the length- k fragment contains the bbb at the end of the block). A lexicographically smallest length- k fragment is one in which this extra b appears the latest, or not at all.

First, we will consider the number of length- k fragments in which the extra b does not occur. For these fragments, it is the case that no other fragment in the block is lexicographically smaller when $a < b$, so it is automatically picked as minimizer at the position corresponding to the start of the length- k fragment. The extra b appears at position $2t + 1$ in the block, so this applies to all length- k fragments starting with an a that end before position $2t + 1$. That is, all a 's up to (and including) position $i = 2t - (k - 1) = 2 \lceil \frac{w+k}{2} \rceil - k + 1 = w + k + p - k + 1 = w + p + 1$, where $p = (w + k) \bmod 2$.

28:8 Minimizing the Minimizers via Alphabet Reordering

Next, we consider the length- k fragments that do include the extra b . At any position past i , the smallest candidate will be the first one starting with an a , *unless* one of the candidates appears in the next T_{ab} -block, in which case the minimizer will be the first position of this next block (because this candidate does not include the extra b and is therefore smaller than any candidate before it). Specifically, this is the case if position $|T_{ab}| + 1$ is one of the w candidates. Therefore, all windows starting at positions up to and including $j = |T_{ab}| + 1 - w = (2\lceil \frac{w+k}{2} \rceil + 2) + 1 - w = w + k + 3 + p - w = k + p + 3$ will have as their minimizer the first position with an a , meaning that all a 's up to position $j + 1$ are minimizers.



■ **Figure 5** T_{ab} for $w = 4$ and $k = 4$, showing the positions i and j for counting $M_{a < b}$. Position i is the final position at which the length- k fragment does not contain bb , whereas j is the final position for which the starting position of the next T_{ab} -block is not a candidate. When $a < b$, the minimizers in the block are all a 's up to position $\max\{i, j + 1\}$.

We now have that all a 's up to position $i = w + p + 1$ and all a 's up to position $j + 1 = k + p + 4$ are minimizers. Thus we need to count the a 's up to position $\max\{w + p + 1, k + p + 4\}$. Because, by hypothesis, $w < k + 2$, this maximum is equal to $k + p + 4$. The first $k + p + 4$ letters of the block are alternating a 's and b 's, so we get

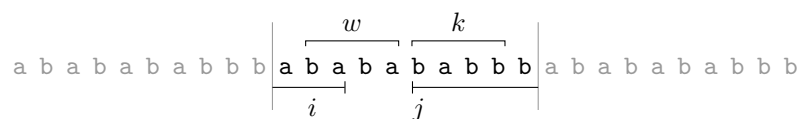
$$M_{a < b} = \left\lceil \frac{k + p + 4}{2} \right\rceil = \left\lceil \frac{k + p}{2} \right\rceil + 2 = \begin{cases} \frac{k}{2} + 2 + p & \text{if } k \text{ is even;} \\ \lfloor \frac{k}{2} \rfloor + 3 & \text{if } k \text{ is odd.} \end{cases}$$

Next, we compute $M_{b < a}$. We start by showing that the final three b 's in T_{ab} are all minimizers. There is only one length- k fragment that starts with bbb and one that starts with bba , so the first two of these final b 's will both be minimizers for the windows that start with bbb and bba . For the window that starts at the third b , which is position $|T_{ab}|$, note that the entire window does not contain bb at all; it consists of only alternating b 's and a 's as the window has length $w + k - 1$ whereas the next occurrence of bb is after $w + k + p$ positions. Because the window does not contain bb , none of its candidates are smaller than $baba \dots$ alternating, which first appears at the start of the window. Therefore, the third b is also a minimizer.

The rest of the minimizers consist of two sets. The first set corresponds to positions for which no candidate is smaller than $baba \dots$ (alternating for k letters). These are all positions with a b , up to a certain position i (to be computed later), after which there will also be a smaller minimizer candidate, i.e., one that contains bb ; inspect Figure 6. This is the second set of minimizers: ones that start with b and contain bb at some point. These are all positions with a b from some position j onwards.

We start by computing j . Position j is the first position such that the length- k fragment starting at j starts with a b and contains bb . If k is odd, the fragment ends at position $2t + 2$ with bbb as suffix; if k is even, the fragment ends at position $2t + 1$ with bb as suffix. We have

$$j = \begin{cases} 2t + 1 - k + 1 = w + p + 2 & \text{if } k \text{ is even;} \\ 2t + 2 - k + 1 = w + p + 3 & \text{if } k \text{ is odd.} \end{cases}$$



■ **Figure 6** T_{ab} for $w = 4$ and $k = 4$, showing the positions i and j when counting $M_{b < a}$: j is the position of the first **b** at which the corresponding length- k fragment contains **bb**; i is the last position at which j is not a candidate for its minimizer. When $b < a$, the minimizers in this block are all **b**'s up to position $i + 1$ and all **b**'s from position j onwards.

Note that $j = w + p + 2 + (k \bmod 2)$. Every **b** from position j onwards is a minimizer. This includes the three **b**'s at the end of the pattern (at positions $2t$ through $2t + 2$), as well as the ones between positions j and $2t - 1$ (both inclusive). Thus we have

$$\begin{aligned} 3 + \left\lfloor \frac{2t - j}{2} \right\rfloor &= 3 + \left\lfloor \frac{w + k + p - (w + p + 2 + (k \bmod 2))}{2} \right\rfloor \\ &= 3 + \left\lfloor \frac{k - 2 - (k \bmod 2)}{2} \right\rfloor = 2 + \left\lfloor \frac{k}{2} \right\rfloor \end{aligned}$$

b's from position j onwards.

Next, we compute i and count the number of **b**'s up to i . We take the last position for which the length- k fragment starting at j is not a candidate. This is $i = j - w$. The minimizer for the window starting at position $i + 1$ is the length- k fragment starting at j , since this is the only candidate that contains **bb**. However, if there is a **b** at position $i + 1$,² then $i + 1$ will still be a minimizer: when we take the minimizer for position i , the length- k fragment containing **bb** will not be a candidate so it will take the first length- k fragment starting with a **b**, which is at position $i + 1$. Therefore, we count all **b**'s that appear up to $i + 1$:

$$\begin{aligned} \left\lfloor \frac{i + 1}{2} \right\rfloor &= \left\lfloor \frac{j - w + 1}{2} \right\rfloor = \left\lfloor \frac{(w + p + 2 + (k \bmod 2)) - w + 1}{2} \right\rfloor = \left\lfloor \frac{p + 3 + (k \bmod 2)}{2} \right\rfloor \\ &= 1 + \left\lfloor \frac{1 + p + (k \bmod 2)}{2} \right\rfloor = \begin{cases} 1 + p & \text{if } k \text{ is even;} \\ 2 & \text{if } k \text{ is odd.} \end{cases} \end{aligned}$$

Adding the two numbers of **b**'s together gives (inspect Figure 7):

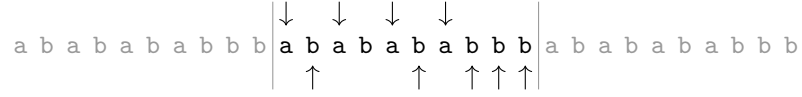
$$\begin{aligned} M_{b < a} &= 2 + \left\lfloor \frac{k}{2} \right\rfloor + \begin{cases} 1 + p & \text{if } k \text{ is even;} \\ 2 & \text{if } k \text{ is odd;} \end{cases} \\ &= \begin{cases} \frac{k}{2} + 3 + p & \text{if } k \text{ is even;} \\ \left\lfloor \frac{k}{2} \right\rfloor + 4 & \text{if } k \text{ is odd.} \end{cases} \end{aligned}$$

2.5 Wrapping up the Reduction

Proof of Theorem 3. MINIMIZING THE MINIMIZERS (DECISION) asks whether or not there exists some ordering on Σ such that a string $S \in \Sigma^n$ has at most ℓ minimizers for parameters w and k . Given w , k and an ordering on Σ , one can compute the number of minimizers

² Consider the case when $T_{ab} = \text{abababababb}$ with $w = 5$ and $k = 4$. For this block, we have $i = 3$ and $j = 8$. Indeed $i = j - w = 3$ and at position $i + 1 = 4$ of the block we have a **b**. Position 4 will be selected as the minimizer for the window starting at position 3.

28:10 Minimizing the Minimizers via Alphabet Reordering



■ **Figure 7** T_{ab} for $w = 4$ and $k = 4$, showing its minimizers for $a < b$ (top) and $b < a$ (bottom). In this situation, $M_{a < b} = 4$ and $M_{b < a} = 5$.

for those parameters in linear time [13, Theorem 3]. Therefore, one can use an alphabet ordering as a certificate to verify a YES instance of MINIMIZING THE MINIMIZERS (DECISION) simply by comparing the computed number of minimizers to ℓ . This proves that the MINIMIZING THE MINIMIZERS (DECISION) problem is in NP. To prove that MINIMIZING THE MINIMIZERS (DECISION) is NP-hard, we use a reduction from FEEDBACK ARC SET (DECISION) (see Section 2 for definition), which is a well-known NP-complete problem [9].

We are given an instance $G = (V, A)$ of FEEDBACK ARC SET and an integer ℓ' , and we are asked to check if G contains a feedback arc set with at most ℓ' arcs. We will construct an instance S of MINIMIZING THE MINIMIZERS (DECISION), for given parameters $w \geq 3$ and $k \geq 1$, such that: the minimum number of minimizers in S , over all alphabet orderings, is at most some value ℓ if and only if G contains a feedback arc set of size at most ℓ' .

By Lemma 4, we have $\lambda \leq 4 \cdot |A| \cdot |T_{ab}|$. Given w and k , we must determine a string T_{ab} such that $M_{b < a} > M_{a < b}$ and $|T_{ab}| \geq \frac{1}{4}(w + k - 1)$, and also choose some q satisfying $\lambda < q \cdot (M_{b < a} - M_{a < b})$ (see Lemma 5). Let $\Sigma = V$ and let $S = \prod_{(a,b) \in A} T_{ab}^{q+4}$, with T_{ab} and q to be determined depending on w and k .

Case A: $w \geq k + 2$. Let $T_{ab} = ab^{w-1}$, so $|T_{ab}| = w$. Since, by hypothesis, the maximal value of k is $w - 2$, and since $|T_{ab}| = w$, we have that $4|T_{ab}| \geq 2w - 3$. Thus, the condition on the length of T_{ab} always holds. By Lemma 6, $M_{b < a} - M_{a < b} = w - k - 1$. We choose $q = 4 \cdot w \cdot |A| + 1$, so that $\lambda \leq 4 \cdot |A| \cdot w < q \cdot (w - k - 1)$. Thus, $\lambda < q \cdot (M_{b < a} - M_{a < b})$.

Case B and Case C: $w < k + 2$. Let $T_{ab} = (ab)^t bb$ for $t = \lceil \frac{w+k}{2} \rceil$. We have $|T_{ab}| = 2t + 2 = 2(\lceil \frac{w+k}{2} \rceil) + 2 = w + k + p + 2$, where $p = (w + k) \bmod 2$. The condition on the length of T_{ab} always holds because $w + k + p + 2 > w + k - 1$.

- If $w = 3$, then by Lemma 7, $M_{b < a} - M_{a < b} = \lfloor \frac{k}{2} \rfloor + 4 - (\lfloor \frac{k}{2} \rfloor + 3) = 1$.
- If $w > 3$, then by Lemma 8:
 - if k is even, $M_{b < a} - M_{a < b} = \frac{k}{2} + 3 + p - (\frac{k}{2} + 2 + p) = 1$;
 - if k is odd, $M_{b < a} - M_{a < b} = \lfloor \frac{k}{2} \rfloor + 4 - (\lfloor \frac{k}{2} \rfloor + 3) = 1$.

In any case, $M_{b < a} - M_{a < b} = 1$. We choose $q = 4 \cdot |A| \cdot (w + k + 3) + 1$, so that $\lambda \leq 4 \cdot |A| \cdot (w + k + p + 2) < q$. Thus, $\lambda < q \cdot (M_{b < a} - M_{a < b})$.

Finally, we set $\ell = q \cdot (M_{b < a} - M_{a < b}) \cdot (\ell' + 1) + q \cdot M_{a < b} \cdot |A|$. By Lemma 5, we have that $\mathcal{M}_{w,k}(S, F) \leq \ell$ if and only if $|F| \leq \ell'$; in other words, G contains a feedback arc set of size at most ℓ' if and only if S has an alphabet ordering with at most ℓ minimizers.

Hence we have shown that (G, ℓ') is a YES instance of MINIMIZING THE MINIMIZERS (DECISION) if and only if (S, ℓ) is a YES instance of FEEDBACK ARC SET (DECISION). Moreover, the length of S is $(q + 4) \cdot |A| \cdot |T_{ab}|$, with T_{ab} being of polynomial length, so the reduction can be performed in polynomial time. The existence of a polynomial-time reduction from FEEDBACK ARC SET (DECISION) to MINIMIZING THE MINIMIZERS (DECISION) proves our claim: MINIMIZING THE MINIMIZERS (DECISION) is NP-complete if $w \geq 3$ and $k \geq 1$. ◀

3 Considering the Orderings on Σ^k

Most of the existing approaches for minimizing the minimizers samples consider the space of all orderings on Σ^k instead of the ones on Σ . Such an approach has the advantage of an easy and efficient construction of the sample by using a rolling hash function $h : \Sigma^k \rightarrow \mathbb{N}$, such as the popular Karp-Rabin fingerprints [10]; this results in a random ordering on Σ^k that usually performs well in practice [23]. Let us denote by MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) the version of MINIMIZING THE MINIMIZERS that seeks to minimize $|\mathcal{M}_{w,k}(S)|$ by choosing a best ordering on Σ^k (instead of a best ordering on Σ). It is easy to see that any algorithm solving MINIMIZING THE MINIMIZERS solves also MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) with a polynomial number of extra steps: We use an arbitrary ranking function rank from the set of length- k substrings of S to $[1, n - k + 1]$. We construct the string S' such that $S'[i] = \text{rank}(S[i..i+k-1])$, for each $i \in [1, n - k + 1]$. Let Σ' be the set of all letters in S' . It should be clear that $|\Sigma'| \leq n$ because S has no more than n substrings of length k . We then solve the MINIMIZING THE MINIMIZERS problem with input $\Sigma := \Sigma'$, $S := S'$, $w := w$, and $k := 1$. It is then easy to verify that an optimal solution to MINIMIZING THE MINIMIZERS for this instance implies an optimal solution to MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) for the original instance. We thus conclude that MINIMIZING THE MINIMIZERS is at least as hard as MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$); they are clearly equivalent for $k = 1$.

► **Example 9.** Let $S = \text{aacaaacgcta}$, $w = 3$, and $k = 3$. We construct the string $S' = 235124687$ over $\Sigma' = [1, 8]$ and solve MINIMIZING THE MINIMIZERS with $w = 3$, $k = 1$, and $\Sigma = \Sigma'$. Assuming $1 < 3 < 5 < 6 < 2 < 4 < 7 < 8$, $\mathcal{M}_{3,1}(S') = \mathcal{M}_{3,3}(S) = \{2, 4, 7\}$. The minimizers positions are colored red: $S' = 2\mathbf{3}5\mathbf{1}2\mathbf{4}6\mathbf{8}7$. This is one of many best orderings.

Another advantage of MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) is that a best ordering on Σ^k is at least as good as a best ordering on Σ at minimizing the resulting sample. Indeed this is because every ordering on Σ implies an ordering on Σ^k but not the reverse.

Unfortunately, MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) comes with a major disadvantage. Suppose we had an algorithm solving MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) (either exactly or with a good approximation ratio or heuristically) and applied it to a string S of length n , with parameters w and k . Now, in order to compare a query string Q to S , the first step would be to compute the minimizers of Q , but to ensure local consistency (Property 2), we would need access to the ordering output by the hypothetical algorithm. The size of the ordering is $\mathcal{O}(\min(|\Sigma|^k, n))$ and storing this defeats the purpose of creating a sketch for S . This is when it might be more appropriate to use MINIMIZING THE MINIMIZERS instead.

Since MINIMIZING THE MINIMIZERS is NP-hard for $w \geq 3$ and $k = 1$, MINIMIZING THE MINIMIZERS ($\leq \Sigma^1$) is NP-hard for $w \geq 3$; hence the following corollary of Theorem 3.

► **Corollary 10.** *MINIMIZING THE MINIMIZERS ($\leq \Sigma^1$) is NP-hard if $w \geq 3$.*

4 Final Remarks

The most immediate open questions are:

- Is MINIMIZING THE MINIMIZERS NP-hard for $w = 2$ and $k \geq 1$?
- Is MINIMIZING THE MINIMIZERS ($\leq \Sigma^k$) NP-hard for $k > 1$?

References

- 1 Lorraine A. K. Ayad, Grigorios Loukides, and Solon P. Pissis. Text indexing for long patterns: Anchors are all you need. *Proc. VLDB Endow.*, 16(9):2117–2131, 2023. doi:10.14778/3598581.3598586.
- 2 Jason W. Bentley, Daniel Gibney, and Sharma V. Thankachan. On the complexity of BWT-runs minimization via alphabet reordering. In Fabrizio Grandoni, Grzegorz Herman, and Peter Sanders, editors, *28th Annual European Symposium on Algorithms, ESA 2020, September 7-9, 2020, Pisa, Italy (Virtual Conference)*, volume 173 of *LIPICs*, pages 15:1–15:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.ESA.2020.15.
- 3 Rayan Chikhi, Antoine Limasset, Shaun Jackman, Jared T. Simpson, and Paul Medvedev. On the representation of de Bruijn graphs. *J. Comput. Biol.*, 22(5):336–352, 2015. doi:10.1089/CMB.2014.0160.
- 4 Sebastian Deorowicz, Marek Kokot, Szymon Grabowski, and Agnieszka Debudaj-Grabysz. KMC 2: fast and resource-frugal k -mer counting. *Bioinform.*, 31(10):1569–1576, 2015. doi:10.1093/BIOINFORMATICS/BTV022.
- 5 Daniel Gibney and Sharma V. Thankachan. Finding an optimal alphabet ordering for Lyndon factorization is hard. In Markus Bläser and Benjamin Monmege, editors, *38th International Symposium on Theoretical Aspects of Computer Science, STACS 2021, March 16-19, 2021, Saarbrücken, Germany (Virtual Conference)*, volume 187 of *LIPICs*, pages 35:1–35:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.STACS.2021.35.
- 6 Szymon Grabowski and Marcin Raniszewski. Sampled suffix array with minimizers. *Softw. Pract. Exp.*, 47(11):1755–1771, 2017. doi:10.1002/SPE.2481.
- 7 Minh Hoang, Hongyu Zheng, and Carl Kingsford. Differentiable learning of sequence-specific minimizer schemes with DeepMinimizer. *J. Comput. Biol.*, 29(12):1288–1304, 2022. doi:10.1089/CMB.2022.0275.
- 8 Chirag Jain, Arang Rhie, Haowen Zhang, Claudia Chu, Brian Walenz, Sergey Koren, and Adam M. Phillippy. Weighted minimizer sampling improves long read mapping. *Bioinform.*, 36(Supplement-1):i111–i118, 2020. doi:10.1093/BIOINFORMATICS/BTAA435.
- 9 Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Proceedings of a symposium on the Complexity of Computer Computations, held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA*, The IBM Research Symposia Series, pages 85–103. Plenum Press, New York, 1972. doi:10.1007/978-1-4684-2001-2_9.
- 10 Richard M. Karp and Michael O. Rabin. Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.*, 31(2):249–260, 1987. doi:10.1147/RD.312.0249.
- 11 Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinform.*, 32(14):2103–2110, 2016. doi:10.1093/BIOINFORMATICS/BTW152.
- 12 Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinform.*, 34(18):3094–3100, 2018. doi:10.1093/BIOINFORMATICS/BTY191.
- 13 Grigorios Loukides and Solon P. Pissis. Bidirectional string anchors: A new string sampling mechanism. In Petra Mutzel, Rasmus Pagh, and Grzegorz Herman, editors, *29th Annual European Symposium on Algorithms, ESA 2021, September 6-8, 2021, Lisbon, Portugal (Virtual Conference)*, volume 204 of *LIPICs*, pages 64:1–64:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.ESA.2021.64.
- 14 Grigorios Loukides, Solon P. Pissis, and Michelle Sweering. Bidirectional string anchors for improved text indexing and top- k similarity search. *IEEE Trans. Knowl. Data Eng.*, 35(11):11093–11111, 2023. doi:10.1109/TKDE.2022.3231780.
- 15 Yaron Orenstein, David Pellow, Guillaume Marçais, Ron Shamir, and Carl Kingsford. Compact universal k -mer hitting sets. In Martin C. Frith and Christian Nørgaard Storm Pedersen, editors, *Algorithms in Bioinformatics - 16th International Workshop, WABI 2016, Aarhus, Denmark, August 22-24, 2016. Proceedings*, volume 9838 of *Lecture Notes in Computer Science*, pages 257–268. Springer, 2016. doi:10.1007/978-3-319-43681-4_21.

- 16 Michael Roberts, Wayne B. Hayes, Brian R. Hunt, Stephen M. Mount, and James A. Yorke. Reducing storage requirements for biological sequence comparison. *Bioinform.*, 20(18):3363–3369, 2004. doi:10.1093/bioinformatics/bth408.
- 17 Saul Schleimer, Daniel Shawcross Wilkerson, and Alexander Aiken. Winnowing: Local algorithms for document fingerprinting. In Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, editors, *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, June 9-12, 2003*, pages 76–85. ACM, 2003. doi:10.1145/872757.872770.
- 18 Yoshihiro Shibuya, Djamal Belazzougui, and Gregory Kucherov. Space-efficient representation of genomic k-mer count tables. *Algorithms Mol. Biol.*, 17(1):5, 2022. doi:10.1186/S13015-022-00212-0.
- 19 Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.
- 20 Daniel H. Younger. Minimum feedback arc sets for a directed graph. *IEEE Transactions on Circuit Theory*, 10(2):238–245, 1963. doi:10.1109/TCT.1963.1082116.
- 21 Hongyu Zheng, Carl Kingsford, and Guillaume Marçais. Improved design and analysis of practical minimizers. *Bioinform.*, 36(Supplement-1):i119–i127, 2020. doi:10.1093/BIOINFORMATICS/BTAA472.
- 22 Hongyu Zheng, Carl Kingsford, and Guillaume Marçais. Sequence-specific minimizers via polar sets. *Bioinform.*, 37(Supplement):187–195, 2021. doi:10.1093/BIOINFORMATICS/BTAB313.
- 23 Hongyu Zheng, Guillaume Marçais, and Carl Kingsford. Creating and using minimizer sketches in computational genomics. *J. Comput. Biol.*, 30(12):1251–1276, 2023. doi:10.1089/CMB.2023.0094.