

Minimizing the Minimizers through Alphabet Reordering

Hilde Verbeek¹, Lorraine A. K. Ayad³,
Grigorios Loukides⁴, Solon P. Pissis^{1,2}

¹CWI, Amsterdam, Netherlands

²Vrije Universiteit, Amsterdam, Netherlands

³Brunel University London, UK

⁴King's College London, UK

CPM 2024

福岡市, 25 June 2024

String sampling

- Minimizers are a popular form of string sampling
- Important applications such as indexing and sequence alignment
- Minimizers satisfy some useful properties:
 - approximately uniform sampling
 - local consistency
 - left-to-right parsing

Minimizers

Definition (Minimizers)

Let $w \geq 2, k \geq 1$. The **minimizer** of a length- $(w + k - 1)$ window is the smallest k -mer within it.

For a string S , $M_{wk}(S)$ is the set of the minimizers of all windows in S .

- Overlapping windows can have the same minimizer
- w is the number of minimizer candidates for each window; k is the length of the minimizers
- In case of a tie, we choose the **left-most** smallest candidate

Example

Let $w = k = 3$ and $S = \text{AACAAACGCTA}$.

↓		↓	↓	↓	↓							$A < C < G < T$ (5)
A	A	C	A	A	C	G	C	T	A			$C < T < A < G$ (4)
		↑	↑	↑	↑							

The ordering of the k -mers matters!

Minimizing the Minimizers

- The right ordering on k -mers can make a significant difference in the number of minimizers
- Practitioners use some heuristics to improve this, but no theoretical results yet
- We focus specifically on **lexicographic** orders

Minimizing the Minimizers

Definition (Minimizing the Minimizers)

Given string $S \in \Sigma^n$ and $w \geq 2, k \geq 1$, find the ordering on Σ that minimizes $M_{wk}(S)$.

Theorem

Minimizing the Minimizers is NP-complete for all $w \geq 3, k \geq 1$.

Feedback Arc Set

Definition (Feedback Arc Set)

Given a directed graph $G = (V, A)$ and integer ℓ , find a set $F \subseteq A$ with $|F| \leq \ell$ such that $(V, A \setminus F)$ is acyclic.

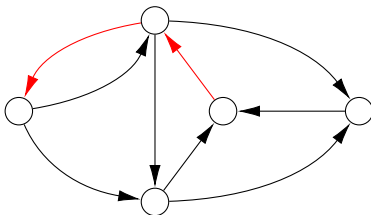


Figure: A directed graph with a feedback arc set of size 2.

Theorem (Karp 1972)

Feedback Arc Set is NP-complete.

Feedback Arc Set

- A feedback arc set F can be induced by an ordering on G 's vertices, such that $F = \{(u, v) \in A \mid v < u\}$.
 - This is a topological ordering of $(V, A \setminus F)$.

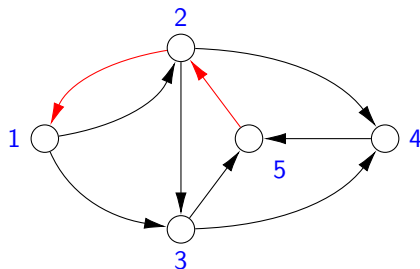


Figure: The feedback arc set consists of arcs (u, v) with $v < u$.

Summary

Given a FAS instance $G = (V, A)$, we create a string S on alphabet V .

- Alphabet orders with few minimizers on S will induce a small FAS on G .
- Create a gadget for every arc (u, v) with
 - few minimizers when $u < v$;
 - many minimizers when $v < u$, as “penalty” for being in the FAS.
- We use these gadgets to count the minimizers in terms of the FAS.

Construction of S

Let T_{ab} be some string consisting of letters a and b , and $q \in \mathbb{N}$ (to be determined at the end). We construct S as

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

We compute the number of minimizers **exactly** for the middle q blocks of every T_{ab}^{q+4} ; the rest is the **discrepancy** λ .

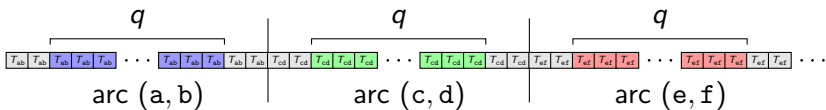


Figure: Structure of S . We count the minimizers for windows within the highlighted blocks, whereas the minimizers in grey blocks count as discrepancy.

Counting minimizers

- Only count the middle q blocks, so the $(w + k - 1)$ -windows do not overlap some other T_{cd} .
- Let $M_{a < b}$ and $M_{b < a}$ be the number of minimizers in T_{ab} if $a < b$ and $b < a$ respectively.
- Let $M_{wk}(S, F)$ denote the number of minimizers in S under the alphabet ordering inducing feedback arc set F .

Counting minimizers

We count the number of minimizers as

$$\begin{aligned}M_{wk}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda.\end{aligned}$$

Counting minimizers

We count the number of minimizers as

$$\begin{aligned}M_{wk}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda.\end{aligned}$$

$M_{wk}(S, F)$: number of minimizers in S for feedback arc set F

Counting minimizers

We count the number of minimizers as

$$\begin{aligned}M_{wk}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda.\end{aligned}$$

$q \cdot M_{b < a}$ minimizers for the $|F|$ arcs in the FAS

Counting minimizers

We count the number of minimizers as

$$\begin{aligned}M_{wk}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda.\end{aligned}$$

$q \cdot M_{a < b}$ minimizers for the $|A| - |F|$ arcs **not** in the FAS

Counting minimizers

We count the number of minimizers as

$$\begin{aligned}M_{wk}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda.\end{aligned}$$

λ : minimizers we don't count explicitly

Counting minimizers

We count the number of minimizers as

$$\begin{aligned}M_{wk}(S, F) &= q \cdot M_{b < a} \cdot |F| + q \cdot M_{a < b} \cdot (|A| - |F|) + \lambda \\ &= q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda.\end{aligned}$$

Lemma

If $M_{b < a} > M_{a < b}$ and $\lambda < q \cdot (M_{b < a} - M_{a < b})$, then $M_{wk}(S, F)$ is minimal if and only if $|F|$ is minimal.

Construction of T_{ab}

We distinguish three cases:

- Case A ($w \geq k + 2$):
 - $T_{ab} = ab^{w-1}$.
- Case B ($w = 3, k \geq 2$) and Case C ($3 < w < k + 2$):
 - $T_{ab} = (ab)^t b b$ with $t = \lceil \frac{w+k}{2} \rceil$.
 - T_{ab} is the same, but the proof is different.

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

We will determine the minimizer for every window starting in some T_{ab} , for both $a < b$ and $b < a$.

a b b b b b b | a b b b b b b | a b b b b b b

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

Find the minimum k -mer in the window.

a b b b b b b | a b b b b b | a b b | b b b b

$w + k + 1$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

Find the minimum k -mer in the window.

a b b b b b b | a b b b b b b | a b b | b b b b

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

The minimizer is in the next T_{ab} , so we ignore it.

a b b b b b b | a | b b b b b b | a b b b | b b b

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

The minimizer is in the next T_{ab} , so we ignore it.

a b b b b b b | a b | b b b b b | a b b b b | b b

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

The minimizer is in the next T_{ab} , so we ignore it.

a b b b b b b | a b b | b b b b | a b b b b b | b

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

Thus we have one minimizer if $a < b$.

a b b b b b b | \downarrow a b b b b b b | a b b b b b b

$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

We do the same for $b < a$.

a b b b b b b | a b b b b b | a b b | b b b b

The diagram shows the string 'abbbbbb' with a dashed box around the substring 'abbbb'. A vertical line is placed between the 'a' and the first 'b' of the boxed substring. A downward arrow points to the 'a' and an upward arrow points to the first 'b'.

$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

We do the same for $b < a$.

a b b b b b b b $\left| \begin{array}{c} \downarrow \\ a \ b \end{array} \right| \begin{array}{c} \uparrow \ \uparrow \\ b \ b \ b \ b \ b \end{array} \left| \begin{array}{c} \\ a \ b \ b \ b \ b \end{array} \right| b \ b$

$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

We do the same for $b < a$.

a b b b b b b b $\left| \begin{array}{c} \downarrow \\ a \ b \ b \end{array} \right| \begin{array}{c} \boxed{b \ b \ b \ b} \end{array} \left| \begin{array}{c} \text{---} \\ a \ b \ b \ b \ b \ b \end{array} \right| b$
 $\uparrow \ \uparrow \ \uparrow$

$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

Minimizer is in the next T_{ab} so we ignore it.

$a\ b\ b\ b\ b\ b\ b\ |$
 $\begin{array}{c} \downarrow \\ a\ b\ b\ b \\ \uparrow\ \uparrow\ \uparrow \end{array}$
 $|b\ b\ b|$
 $|a\ b\ b\ b\ b\ b\ b|$

$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

Minimizer is in the next T_{ab} so we ignore it.

$a\ b\ b\ b\ b\ b\ b\ |$
 $\begin{array}{c} \downarrow \\ a\ b\ b\ b\ b \\ \uparrow\ \uparrow\ \uparrow \end{array}$
 $b\ b\ |$
 $\boxed{a\ b\ b\ b\ b\ b\ b}$

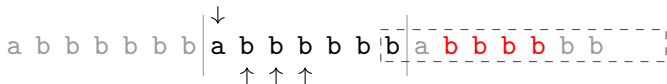
$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

Minimizer is in the next T_{ab} so we ignore it.



$$M_{a < b} = 1$$

Case A: $w \geq k + 2$

Let $w = 7$ and $k = 4$.

We have $T_{ab} = ab^{w-1} = abbbbbb$.

We have three minimizers if $b < a$.

$$a \ b \ b \ b \ b \ b \ b \ \left| \begin{array}{c} \downarrow \\ a \ b \ b \ b \ b \ b \ b \\ \uparrow \ \uparrow \ \uparrow \end{array} \right| a \ b \ b \ b \ b \ b \ b$$

$$M_{a < b} = 1$$

$$M_{b < a} = 3$$

Case A: $w \geq k + 2$

In general:

- every window contains ab^{k-1} , which is the minimizer if $a < b$, so $M_{a < b} = 1$;
- every T_{ab} contains $w - k$ occurrences of b^k , so $M_{b < a} = w - k$.

Given that $w - k \geq 2$, we have $M_{b < a} > M_{a < b}$.

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | a b a b a b b b | a b a b a b b b

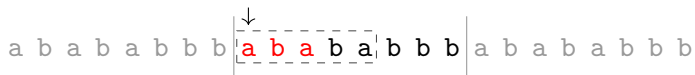
Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t b b = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | a b a b a | b b b | a b a b a b b b



Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | a b a b a b | b b | a b a b a b b b

\downarrow \downarrow
a b a b a b

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | a b [a b a b b] b | a b a b a b b b

↓ ↓

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | $\begin{array}{c} \downarrow \\ \text{a} \end{array}$ $\begin{array}{c} \downarrow \\ \text{b} \end{array}$ a $\begin{array}{c} \downarrow \\ \text{b} \end{array}$ a b b b | a b a b a b b b

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | $\begin{matrix} \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & b & b \end{matrix}$ | a b a b a b b b

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We count minimizers in the same way as before.

a b a b a b b b | a b a b a b b b | a b a b b b

↓
↓
↓
↓

b b b
a b

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

Minimizer is in the next T_{ab} , so we don't count it.

a b a b a b b b | $\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & b & b \end{matrix}$ | a b a b b b

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

Minimizer is in the next T_{ab} , so we don't count it.

a b a b a b b b | $\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & b \end{matrix}$ | $\begin{matrix} \downarrow \\ \boxed{b \ a \ b \ a \ b} \end{matrix}$ | a b b b

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

We have four minimizers if $a < b$.

a b a b a b b b | $\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & b & b \end{matrix}$ | a b a b a b b b

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

Now for $b < a$.

a b a b a b b b | a b a b a | b b b | a b a b a b b b

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

Now for $b < a$.

a b a b a b b b | a b a b a b | b b | a b a b a b b b

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

Now for $b < a$.

$a \ b \ a \ b \ a \ b \ b \ b \left| \begin{array}{c} \downarrow \\ a \ b \end{array} \right. \left[\begin{array}{c} \downarrow \\ a \ b \end{array} \right] \left[\begin{array}{c} \downarrow \\ a \ b \end{array} \right] \left| \begin{array}{c} \downarrow \\ b \end{array} \right. a \ b \ a \ b \ a \ b \ b \ b$

(Note: In the original image, the subwords ab are highlighted with dashed boxes, and the a in the second ab and the b in the third ab are highlighted in red. Arrows point to the a in the first ab and the b in the second ab .)

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t bb = abababbb$.

Now for $b < a$.

$a \ b \ a \ b \ a \ b \ b \ b \left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b \end{array} \right. a \ b \ a \ b \ a \ b \ b \ b$

 $\quad \quad \quad \uparrow \quad \uparrow \quad \uparrow$

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t b b = abababbb$.

Now for $b < a$.

a b a b a b b b | $\begin{array}{c} \downarrow \\ a \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ $\begin{array}{c} \downarrow \\ a \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ $\begin{array}{c} \downarrow \\ a \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ | $\begin{array}{c} \downarrow \\ a \end{array}$ b a b a b b b

$\begin{array}{c} \uparrow \\ a \end{array}$ $\begin{array}{c} \uparrow \\ b \end{array}$ $\begin{array}{c} \uparrow \\ a \end{array}$ $\begin{array}{c} \uparrow \\ b \end{array}$ $\begin{array}{c} \uparrow \\ a \end{array}$ $\begin{array}{c} \uparrow \\ b \end{array}$ $\begin{array}{c} \uparrow \\ b \end{array}$ $\begin{array}{c} \uparrow \\ b \end{array}$

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t b b = abababbb$.

Now for $b < a$.

$a \ b \ a \ b \ a \ b \ b \ b$

 $\left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b \\ \uparrow & \uparrow & \uparrow & \uparrow \end{array} \right|$

 $b \ b \ a \ b \ a$
 $b \ a \ b \ b \ b$

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t b b = abababbb$.

Now for $b < a$.

$a \ b \ a \ b \ a \ b \ b \ b \left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & b & b \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \end{array} \right. \boxed{b} \text{ } \boxed{a \ b \ a \ b} \text{ } a \ b \ b \ b$

$$M_{a < b} = 4$$

Case B: $w = 3, k \geq 2$

Let $w = 3$ and $k = 3$.

We have $t = \lceil \frac{3+3}{2} \rceil = 3$ so $T_{ab} = (ab)^t b b = abababbb$.

We have five minimizers if $b < a$.

$$a \ b \ a \ b \ a \ b \ b \ b \left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ a \ b & a \ b & a \ b & b \ b \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \end{array} \right. a \ b \ a \ b \ a \ b \ b \ b$$

$$M_{a < b} = 4$$

$$M_{b < a} = 5$$

Case B: $w = 3, k \geq 2$

In general:

- for $a < b$, we count every k -mer starting with a , plus the very last k -mer, and $M_{a < b} = \lfloor \frac{k}{2} \rfloor + 3$;
- for $b < a$, we count every k -mer starting with b , and $M_{b < a} = \lfloor \frac{k}{2} \rfloor + 4$.

Therefore $M_{b < a} - M_{a < b} = 1$.

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | a b a b a b a b b b | a b a b a b ...

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b a b a b a b b b | a b a b a b ...

↓

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b $\left| \begin{array}{c} \downarrow \\ \text{a} \end{array} \right. \left[\begin{array}{c} \downarrow \\ \text{b a b a} \end{array} \right] \left| \begin{array}{c} \text{b b b} \end{array} \right. \left| \text{a b a b a b} \dots \right.$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | $\begin{array}{c} \downarrow \\ \text{a b} \end{array}$ $\begin{array}{c} \downarrow \\ \text{a b a b a b} \end{array}$ | b b | a b a b a b ...

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b $\left| \begin{array}{c} \downarrow \\ \text{a} \end{array} \right. \text{b} \text{a} \left[\begin{array}{c} \downarrow \\ \text{b} \end{array} \right. \text{a} \text{b} \text{a} \text{b} \text{b} \left. \right] \text{b} \left| \text{a} \text{b} \text{a} \text{b} \text{a} \text{b} \dots$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b $\left| \begin{array}{c} \downarrow \quad \downarrow \quad \downarrow \\ \text{a b a b } \boxed{\text{a b a b b b}} \end{array} \right| \text{a b a b a b } \dots$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | a b a b a | b a b b | a | b a b a b ...

↓
↓
↓
↓

b a b b

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | $\begin{array}{c} \downarrow \\ a \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ $\begin{array}{c} \downarrow \\ a \end{array}$ $\begin{array}{c} \downarrow \\ b \end{array}$ a b b b | a b | a b a b ...

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | $\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & a \end{matrix}$ [b b b] a b a | b a b ...

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

\dots a b a b b b $\left| \begin{array}{c} \downarrow \\ \downarrow \\ \downarrow \\ \downarrow \end{array} \right.$ a b a b a b a b $\left[\begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \right]$ b b $\left[\begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \right]$ a b a b $\left[\begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \right]$ a b \dots

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | $\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & b & b \end{matrix}$ | a b a b a b ...

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b $\left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ \text{a} & \text{b} & \text{a} & \text{b} \\ \uparrow & & & \end{array} \right| \text{a b b b} \left| \text{a b a b a b} \dots \right.$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

... a b a b b b | a b a b a b a | b b b | a b a b a b ...

↓
↓
↓
↓

↑

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b \left| \begin{array}{c} \downarrow \\ a \end{array} \right. \begin{array}{c} \downarrow \\ b \end{array} \left[\begin{array}{c} \downarrow \\ a \end{array} \right. \begin{array}{c} \downarrow \\ b \end{array} \left. \begin{array}{c} \downarrow \\ a \end{array} \right. \begin{array}{c} \downarrow \\ b \end{array} \left. \begin{array}{c} \downarrow \\ b \end{array} \right. \begin{array}{c} \downarrow \\ b \end{array} \left. \begin{array}{c} \downarrow \\ b \end{array} \right| a b a b a b \dots$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b \left| \begin{array}{ccccccc} \downarrow & \downarrow & \downarrow & \downarrow & & & \\ a & b & a & \boxed{b a b a b b} & & & \\ \uparrow & & \uparrow & & & & \\ & & & & & & \end{array} \right| b a b a b a b \dots$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b \left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b \\ \uparrow & \uparrow & & \uparrow \end{array} \right. \left[a b a \color{red}{b b b} \right] a b a b a b \dots$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b \mid \begin{array}{c} \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ a \ b \ a \ b \ a \ [b \ a \ b \ b \ b] \ a \mid b \ a \ b \ a \ b \ \dots \\ \uparrow \quad \uparrow \quad \quad \quad \uparrow \end{array}$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b \mid \begin{array}{c} \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ a \ b \ a \ b \ a \ b \ a \ b \ b \ b \end{array} \mid \begin{array}{c} \text{---} \\ a \ b \ b \ b \end{array} \mid \begin{array}{c} \text{---} \\ a \ b \end{array} \mid a \ b \ a \ b \dots$

$\uparrow \quad \uparrow \quad \uparrow$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b$

 $\left| \begin{array}{cccc|cccc} \downarrow & \downarrow & \downarrow & \downarrow & & & & \\ a & b & a & b & a & b & b & b \\ \uparrow & \uparrow & & & \uparrow & \uparrow & & \end{array} \right.$

 $\left[\begin{array}{cccc} b & b & a & b & a & b \end{array} \right]$
 $a b \dots$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$\dots a b a b b b \left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ a & b & a & b & a & b & a & b & b \\ \uparrow & & \uparrow & & & \uparrow & \uparrow & \uparrow & \uparrow \end{array} \right| \begin{array}{c} \text{---} \\ b \text{---} \\ \text{---} \\ a \text{---} \\ \text{---} \\ b \text{---} \\ \text{---} \\ a \text{---} \\ \text{---} \\ b \end{array} \dots$

$$M_{a < b} = 4$$

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

Let $w = 4$ and $k = 3$.

We have $t = \lceil \frac{4+3}{2} \rceil = 4$ so $T_{ab} = (ab)^t bb = ababababbb$.

$$\dots ababbb \left| \begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ ababababbb \\ \uparrow & \uparrow & & \uparrow \uparrow \uparrow \end{array} \right| ababab \dots$$

$$M_{a < b} = 4$$

$$M_{b < a} = 5$$

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

Let $p = (w + k) \bmod 2$.

If k is even:

- $M_{a < b} = \frac{k}{2} + 2 + p$
- $M_{b < a} = \frac{k}{2} + 3 + p$

If k is odd:

- $M_{a < b} = \lfloor \frac{k}{2} \rfloor + 3$
- $M_{b < a} = \lfloor \frac{k}{2} \rfloor + 4$

Either way, $M_{b < a} - M_{a < b} = 1$.

Wrapping up the reduction

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

$$M_{wk}(S, F) = q \cdot (M_{b < a} - M_{a < b}) \cdot |F| + q \cdot M_{a < b} \cdot |A| + \lambda$$

Lemma

If $M_{b < a} > M_{a < b}$ and $\lambda < q \cdot (M_{b < a} - M_{a < b})$, then $M_{wk}(S, F)$ is minimal if and only if $|F|$ is minimal.

Wrapping up the reduction

Lemma

If $M_{b<a} > M_{a<b}$ and $\lambda < q \cdot (M_{b<a} - M_{a<b})$, then $M_{wk}(S, F)$ is minimal if and only if $|F|$ is minimal.

- We have seen $M_{b<a} > M_{a<b}$ for each case:
 - Case A: $M_{b<a} - M_{a<b} = w - k$;
 - Cases B and C: $M_{b<a} - M_{a<b} = 1$.
- We count minimizers for all but 4 T_{ab} -blocks for each arc, so $\lambda \leq 4 \cdot |A| \cdot |T_{ab}|$.
- Pick q such that $q \cdot (M_{b<a} - M_{a<b}) > 4 \cdot |A| \cdot |T_{ab}| \geq \lambda$.

Non-lexicographic orders

What if we can order k -mers in any way, instead of lexicographically?

- Now there are up to $(\min\{n - k + 1, |\Sigma|^k\})!$ permutations to consider.
- For $k = 1$, this is the same as the lexicographic version.
- For $k \geq 2$, it is likely harder due to it being more general.

Open problems

- Hardness for $w = 2, k \geq 1$ (done, on arXiv)
- Hardness for non-lexicographic orders
- Algorithms, approximability, etc.