On the Near-Tightness of $\chi \leqslant 2r$: A General σ-ary Construction and a Binary Case via LFSRs

Vinicius Tikara Venturi Date[1] Leandro Miranda Zatesko[2,1]

LATIN 2026

[1]Federal University of Paraná [2]Federal University of Technology – Paraná

tikaradate@gmail.com

Context

Repetitiveness measures capture how compressible a string is.

Bounds between them shape what compressed data structures can achieve.

Navarro et al. (2025): $\chi \leqslant 2r$

Cenzato et al. (2024): empirically, on genomes, $\chi/r \leqslant 1.33$

Can the ratio reach 2?

Conventions

$$w = \mathrm{aabddc}, \quad |w| = 6, \quad \Sigma = \{a, b, c, d\}, \quad \sigma = |\Sigma|$$

ϵ is the empty string.

i	0	1	2	3	4	5
$w[i]$	a	a	b	d	d	c

(We use 0-based indexing.)

Conventions

$w = \mathrm{aabddc}, \quad |w| = 6, \quad \Sigma = \{a, b, c, d\}, \quad \sigma = |\Sigma|$

ϵ is the empty string.

i	0	1	2	3	4	5
$w[i]$	a	a	b	d	d	c

(We use 0-based indexing.)

$w\$ = \mathrm{aabddc}\$, \quad \$ \notin \Sigma, \quad \$ < \text{all of } \Sigma$

Burrows–Wheeler Transform and r

Construction:

- Append a unique endmarker \$ to w.

Example: $w = \mathrm{aabddc}$

$w\$ = \mathrm{aabddc}\$$

Burrows–Wheeler Transform and r

Construction:

- Append a unique endmarker \$ to w.
- Take all cyclic rotations of $w\$$.
- Sort the rotations lexicographically.

Example: $w = \mathrm{aabddc}$

$w\$ = \mathrm{aabddc\$}$

$aabddc\$a$
$abddc\$aa$
$bddc\$aab$
$c\$aabdd$
$dc\$aabd$
$ddc\$aab$

Burrows–Wheeler Transform and r

Construction:

- Append a unique endmarker \$ to w.
- Take all cyclic rotations of $w\$$.
- Sort the rotations lexicographically.
- BWT($w\$$) = last column.

r = number of maximal runs of equal characters in BWT($w\$$).

Example: w = aabddc

$w\$$ = aabddc\$

$\$aabddc$
$aabddc\$$
$abddc\$a$
$bddc\$aa$
$c\$aabdd$
$dc\$aabd$
$ddc\$aab$

BWT($w\$$) = c\$aaddb, $r = 5$

The measure χ

Example: $w\$ = \mathrm{aabddc}\$$

The measure χ

Example: $w\$ = \mathrm{aabddc}\$$

Right-maximal substrings:

$$\epsilon : \{a, b, c, d, \$\}$$
$$a : \{aa, ab\}$$
$$d : \{dd, dc\}$$

Right-extensions:

$\{a, b, c, d, \$, aa, ab, dd, dc\}$

Right-maximal substring: x occurs at least twice and has two distinct **right-extensions** xc, xd with $c \neq d$.

The measure χ

Example: $w\$ = \mathrm{aabddc}\$$

Right-maximal substrings:

$$\epsilon : \{a, b, c, d, \$\}$$
$$a : \{aa, ab\}$$
$$d : \{dd, dc\}$$

Right-extensions:
$\{a, b, c, d, \$, aa, ab, dd, dc\}$

Super-maximal extensions:
$\{aa, ab, dc, dd, \$\}$

Right-maximal substring: x occurs at least twice and has two distinct **right-extensions** xc, xd with $c \neq d$.

Super-maximal extension: a right-extension that is not a suffix of any other.

The measure χ

Example: $w\$ = \mathrm{aabddc}\$$

Right-maximal substrings:

$$\epsilon : \{a, b, c, d, \$\}$$
$$a : \{aa, ab\}$$
$$d : \{dd, dc\}$$

Right-extensions:
$\{a, b, c, d, \$, aa, ab, dd, dc\}$

Super-maximal extensions:
$\{aa, ab, dc, dd, \$\}$

$\chi = 5$

Right-maximal substring: x occurs at least twice and has two distinct **right-extensions** xc, xd with $c \neq d$.

Super-maximal extension: a right-extension that is not a suffix of any other.

$\chi = |\text{super-maximals}|$

(Equivalently, the size of a smallest suffixient set.)

Our result

Theorem

The bound $\chi \leqslant 2r$ is asymptotically tight.

We construct string families where $\dfrac{\chi}{r} \to 2$.

Our result

Theorem

The bound $\chi \leqslant 2r$ is asymptotically tight.

We construct string families where $\frac{\chi}{r} \to 2$.

Two constructions:

1. **Clustered family** K — works for all σ, approaches 2 as $\sigma \to \infty$
2. **Binary family** M — achieves ratio $\to 2$ with $\sigma = 2$ via de Bruijn sequences

Clustered family K

Idea: concatenate runs of distinct characters in decreasing order.

$$K = s_{\sigma-1}^{k_{\sigma-1}} s_{\sigma-2}^{k_{\sigma-2}} \cdots s_1^{k_1} s_0^{k_0}$$

with $s_i < s_j$ when $i < j$ and all $k_i > 1$.

Example: $K = \texttt{eeeddccaaa} \quad (\sigma = 4)$

Clustered family K

Idea: concatenate runs of distinct characters in decreasing order.

$$K = s_{\sigma-1}^{k_{\sigma-1}} s_{\sigma-2}^{k_{\sigma-2}} \cdots s_1^{k_1} s_0^{k_0}$$

with $s_i < s_j$ when $i < j$ and all $k_i > 1$.

Example: $K = \texttt{eeeddccaaa} \quad (\sigma = 4)$

- Each character forms a single contiguous block.
- Boundaries between blocks create the super-maximal extensions.
- Descending order ensures the BWT has a clean structure.

Clustered family K

Idea: concatenate runs of distinct characters in decreasing order.

$$K = s_{\sigma-1}^{k_{\sigma-1}} s_{\sigma-2}^{k_{\sigma-2}} \cdots s_1^{k_1} s_0^{k_0}$$

with $s_i < s_j$ when $i < j$ and all $k_i > 1$.

Example: $K = \texttt{eeeddccaaa} \quad (\sigma = 4)$

- Each character forms a single contiguous block.
- Boundaries between blocks create the super-maximal extensions.
- Descending order ensures the BWT has a clean structure.

Goal: show that $r = \sigma + 1$ and $\chi = 2\sigma$, giving $\chi/r \to 2$.

Clustered family K — BWT analysis

$eeeddccaaa

$a$eeeddccaa

aaeeeddcca

aaaeeeddcc

$caaa$eeeddc

$ccaaa$eeedd

$dccaaa$eeed

$ddccaaa$eee

$eddccaaa$ee

$eeddccaaa$e

$eeeddccaaa$

Clustered family K — BWT analysis

$eeeddccaa\mathbf{a}$

$a\$eeeddcca\mathbf{a}$

$aa\$eeeddcc\mathbf{a}$

$aaa\$eeeddc\mathbf{c}$

$caaa\$eeedd\mathbf{c}$

$ccaaa\$eeed\mathbf{d}$

$dccaaa\$eee\mathbf{d}$

$ddccaaa\$ee\mathbf{e}$

$eddccaaa\$e\mathbf{e}$

$eeddccaaa\$\mathbf{e}$

$eeeddccaaa\mathbf{\$}$

- Suffixes group by first character.
- For almost every row, the last character equals the first.
- Generally,
  $\mathrm{BWT} = s_0^{k_0}\, s_1^{k_1} \cdots s_{\sigma-1}^{k_{\sigma-1}}\, \$$

Clustered family K — BWT analysis

$eeeddccaa\mathbf{a}$

$a\$eeeddcca\mathbf{a}$

$aa\$eeeddcc\mathbf{a}$

$aaa\$eeeddc\mathbf{c}$

$caaa\$eeedd\mathbf{c}$

$ccaaa\$eeed\mathbf{d}$

$dccaaa\$eee\mathbf{d}$

$ddccaaa\$ee\mathbf{e}$

$eddccaaa\$e\mathbf{e}$

$eeddccaaa\$\mathbf{e}$

$eeeddccaaa\mathbf{\$}$

- Suffixes group by first character.
- For almost every row, the last character equals the first.
- Generally, $\mathrm{BWT} = s_0^{k_0}\, s_1^{k_1} \cdots s_{\sigma-1}^{k_{\sigma-1}}\, \$$

$$\mathrm{BWT}(K\$) = \mathbf{aaaccddeee\$}$$

$$r = \sigma + 1 = 5$$

Clustered family K — χ

$K\$ = \text{eeeddccaaa}\$$

Right-extensions:

$\{ee, ed, eee, eed,$

$dd, dc, cc, ca,$

$aa, a\$, aaa, aa\$\}$

Clustered family K — χ

$K\$ = \mathtt{eeeddccaaa\$}$

Right-extensions:

$$\{ee, ed, eee, eed,$$
$$dd, dc, cc, ca,$$
$$aa, a\$, aaa, aa\$\}$$

Super-maximal extensions:

$\{eee, eed, dd, dc, cc, ca, aaa, aa\$\}$

- Two super-maximal extensions per cluster boundary.
- $\chi = 2\sigma = 8$

Clustered family K — χ

$K\$ = \mathtt{eeeddccaaa\$}$

Right-extensions:

$$\{ee, ed, eee, eed,$$
$$dd, dc, cc, ca,$$
$$aa, a\$, aaa, aa\$\}$$

Super-maximal extensions:

$$\{eee, eed, dd, dc, cc, ca, aaa, aa\$\}$$

- Two super-maximal extensions per cluster boundary.
- $\chi = 2\sigma = 8$

Ratio

$$\frac{\chi}{r} = \frac{2\sigma}{\sigma + 1} \xrightarrow{\sigma \to \infty} 2$$

The binary case is loose

Problem

For $\sigma = 2$, the clustered family K gives:

$$\frac{\chi}{r} = \frac{2 \cdot 2}{2+1} = \frac{4}{3}$$

This is still far from 2.

The binary case is loose

Problem

For $\sigma = 2$, the clustered family K gives:

$$\frac{\chi}{r} = \frac{2 \cdot 2}{2+1} = \frac{4}{3}$$

This is still far from 2.

Solution: Use de Bruijn sequences.

De Bruijn sequence of order k

A cyclic string where every k-mer over Σ appears exactly once.

Example ($k = 3, \sigma = 2$): $B = 00010111$

Windows: $\{000, 001, 010, 101, 011, 111, 110, 100\}$ ✓

De Bruijn family M — construction

Problem: Not all de Bruijn sequences have the same BWT.

$\chi = |w|$ for all de Bruijn, but r varies. We need run-minimal ones.

De Bruijn family M — construction

Problem: Not all de Bruijn sequences have the same BWT.

$\chi = |w|$ for all de Bruijn, but r varies. We need run-minimal ones.

Approach: Build de Bruijn from polynomial $x^k + x + 1$ over $\mathbb{F}_2$.

This gives a recurrence: each bit determined by the previous $k \rightarrow$ cBWT is predictable.

De Bruijn family M — construction

Problem: Not all de Bruijn sequences have the same BWT.

$\chi = |w|$ for all de Bruijn, but r varies. We need run-minimal ones.

Approach: Build de Bruijn from polynomial $x^k + x + 1$ over $\mathbb{F}_2$.

This gives a recurrence: each bit determined by the previous $k \rightarrow$ cBWT is predictable.

Construction:

1. $x^k + x + 1 \rightarrow$ recurrence $\rightarrow$ almost-de Bruijn (missing 0^k)
2. Insert $0^k \rightarrow$ true de Bruijn
3. Reverse + complement $\rightarrow$ run-minimal

The recurrence survives all transformations.

De Bruijn family M — cBWT

Cyclic BWT: BWT without \$. Rows sorted by k-prefix.

The recurrence fixes each cBWT entry from the last 2 bits of its prefix:

De Bruijn family M — cBWT

Cyclic BWT: BWT without \$. Rows sorted by k-prefix.

The recurrence fixes each cBWT entry from the last 2 bits of its prefix:

last 2 bits of k-prefix	cBWT
00	1
01	0
10	0
11	1

Base pattern: $(1001)^{2^{k-2}}$

De Bruijn family M — cBWT

Cyclic BWT: BWT without \$. Rows sorted by k-prefix.

The recurrence fixes each cBWT entry from the last 2 bits of its prefix:

last 2 bits of k-prefix	cBWT
00	1
01	0
10	0
11	1

Base pattern: $(1001)^{2^{k-2}}$

After adjusting for the missing 0^k:

$$1(0011)^{2^{k-2}-1}010$$

De Bruijn family M — cBWT

Cyclic BWT: BWT without \$. Rows sorted by k-prefix.

The recurrence fixes each cBWT entry from the last 2 bits of its prefix:

last 2 bits of k-prefix	cBWT
00	1
01	0
10	0
11	1

Base pattern: $(1001)^{2^{k-2}}$

After adjusting for the missing 0^k:

$$1(0011)^{2^{k-2}-1}010$$

This is run-minimal (Mantaci et al., 2017).

De Bruijn family M — cBWT

Cyclic BWT: BWT without \$. Rows sorted by k-prefix.

The recurrence fixes each cBWT entry from the last 2 bits of its prefix:

last 2 bits of k-prefix	cBWT
00	1
01	0
10	0
11	1

Base pattern: $(1001)^{2^{k-2}}$

After adjusting for the missing 0^k:

$$1(0011)^{2^{k-2}-1}010$$

This is run-minimal (Mantaci et al., 2017). $r = 2^{k-1} + 2$ runs

De Bruijn family M — linearization and termination

Linearization: The value of χ is known in linearized de Bruijn strings.
Termination: The bound $\chi \leqslant 2r$ applies to terminated strings.

Choose the cycle that starts at 0^k, append first $k-1$ characters, terminate with \$.

De Bruijn family M — linearization and termination

Linearization: The value of χ is known in linearized de Bruijn strings.
Termination: The bound $\chi \leqslant 2r$ applies to terminated strings.

Choose the cycle that starts at 0^k, append first $k-1$ characters, terminate with \$.

Effect on r:

cBWT: $1(0011)^{2^{k-2}-1}010$

BWT: $0^{k-1}1\$(0011)^{2^{k-2}-1}010$

$2^{k-1}+2 \to 2^{k-1}+4$

De Bruijn family M — linearization and termination

Linearization: The value of χ is known in linearized de Bruijn strings.
Termination: The bound $\chi \leqslant 2r$ applies to terminated strings.

Choose the cycle that starts at 0^k, append first $k-1$ characters, terminate with \$.

Effect on r:

cBWT: $1(0011)^{2^{k-2}-1}010$

BWT: $0^{k-1}1\$(0011)^{2^{k-2}-1}010$

$2^{k-1}+2 \to 2^{k-1}+4$

Effect on χ:

$\chi = |w| = 2^k$ for linearized de Bruijn (Navarro et al.)

$\chi = 2^k + 1$ after appending \$

De Bruijn family M — linearization and termination

Linearization: The value of χ is known in linearized de Bruijn strings.
Termination: The bound $\chi \leqslant 2r$ applies to terminated strings.

Choose the cycle that starts at 0^k, append first $k-1$ characters, terminate with \$.

Effect on r:

cBWT: $1(0011)^{2^{k-2}-1}010$

BWT: $0^{k-1}1\$(0011)^{2^{k-2}-1}010$

$2^{k-1}+2 \to 2^{k-1}+4$

Effect on χ:

$\chi = |w| = 2^k$ for linearized de Bruijn (Navarro et al.)

$\chi = 2^k + 1$ after appending \$

$$\frac{\chi}{r} = \frac{2^k+1}{2^{k-1}+4} \xrightarrow{k\to\infty} 2$$

Summary

Main result

$\chi \leqslant 2r$ is asymptotically tight for large alphabet sizes and for the binary alphabet.

Family K	$\sigma \geqslant 2$	$\chi/r = 2\sigma/(\sigma+1) \to 2$
Family M	$\sigma = 2$	$\chi/r = (2^k+1)/(2^{k-1}+4) \to 2^*$

*Requires $x^k + x + 1$ primitive over $\mathbb{F}_2$; whether infinitely many such k exist is open

Summary

Main result

$\chi \leqslant 2r$ is asymptotically tight for large alphabet sizes and for the binary alphabet.

Family K	$\sigma \geqslant 2$	$\chi/r = 2\sigma/(\sigma+1) \to 2$
Family M	$\sigma = 2$	$\chi/r = (2^k+1)/(2^{k-1}+4) \to 2^*$

*Requires $x^k + x + 1$ primitive over $\mathbb{F}_2$; whether infinitely many such k exist is open

Open problems:

- Intermediate alphabets $\sigma = 3, 4, 5, \ldots$
- Why does real repetitive data stay below 2?

On the Near-Tightness of $\chi \leqslant 2r$: A General σ-ary Construction and a Binary Case via LFSRs

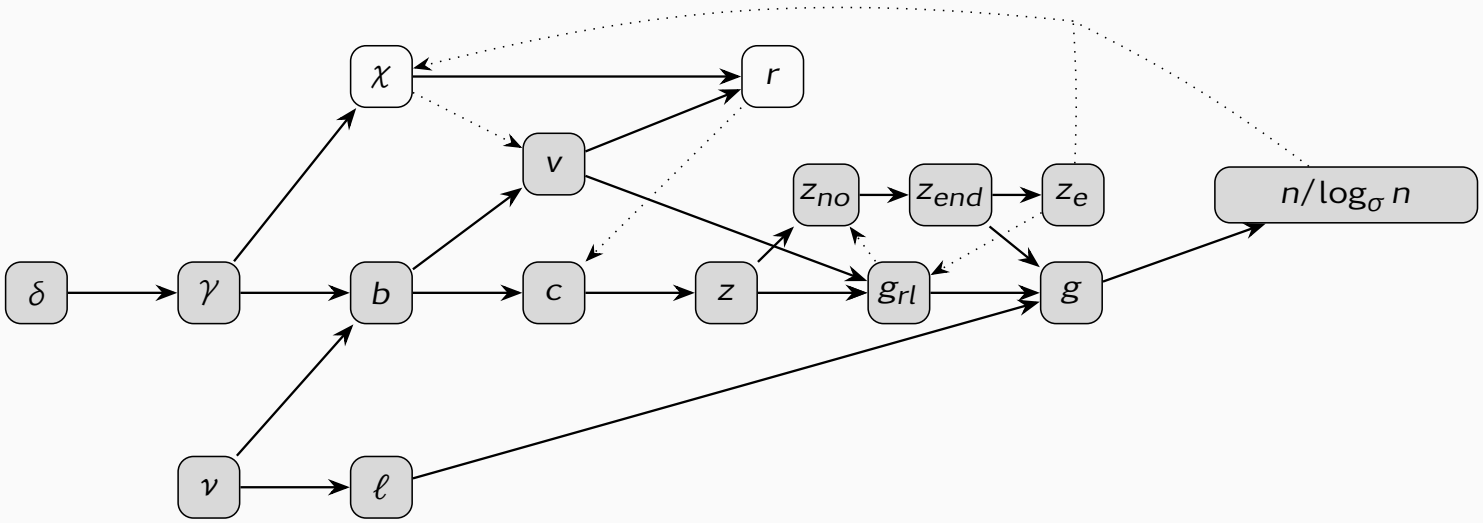
Vinicius Tikara Venturi Date[1] Leandro Miranda Zatesko[2,1]

LATIN 2026

[1]Federal University of Paraná [2]Federal University of Technology – Paraná

tikaradate@gmail.com

Backup: Measure relationships (full)



χ and r are highlighted — our work shows the $\chi \leqslant 2r$ edge is tight.

Backup: LFSR recurrence transformations

Reversal

For primitive $C(x) = 1 + c_1 x + \cdots + c_{k-1} x^{k-1} + x^k$, its reciprocal $C^*(x) = x^k C(1/x)$ is also primitive and gives a sequence with reversed taps.

Complement

Setting $p_t := s_t \oplus 1$:

$$p_{t+k} = \bigoplus_{i=0}^{k-1} c_i p_{t+i} \oplus \left(\bigoplus_{i=0}^{k-1} c_i \right) \oplus 1$$

Combined for $T(x) = x^k + x + 1$:

$$s'_t = s'_{t+k} \oplus s'_{t+k-1} \oplus 1 \oplus \omega$$

Backup: Linearization details

Suffixes containing \$:

$$
\begin{array}{lll}
\$\cdots & 0\cdots & 0 \\
0\$\cdots & 0\cdots & 0 \\
\vdots & & \\
0\cdots & 0\$\cdots & 0 \\
0\cdots & 00\$\cdots & 1 \\
\vdots & &
\end{array}
$$

Prefix of BWT: $x = 0^{k-1}1$.

Suffixes without \$:

$$
\begin{array}{c}
\vdots \\
0\cdots 0\cdots \$ \\
0\cdots 1\cdots 0 \\
\vdots \\
1\cdots 0\cdots 1 \\
1\cdots 1\cdots 0
\end{array}
$$

Suffix of BWT:
$y = \$(0011)^{2^{k-2}-1}010$.

Backup: De Bruijn family M — worked example ($k = 3$)

Example: $M = 00010111$, linearized: $0001011100\$$

$\$0001011100$

$0\$000101110$

$00\$00010111$

$0001011100\$$

$001011100\$0$

$01011100\$00$

$011100\$0001$

$100\$0001011$

$1011100\$000$

$1100\$00010\,1$

$11100\$00010$

BWT $= 001\$0011010$

$r = 8 = 2^{k-1} + 4$ ✓

$\chi = 2^k + 1 = 9$ ✓

$\chi/r = 9/8 = 1.125$

(Ratio improves as k grows)

Backup: Why de Bruijn fails for $\sigma \geqslant 3$

For σ-ary de Bruijn of order k:

- $\chi = \sigma^k + 1$
- $r \geqslant \sigma^{k-1}(\sigma - 1) + 1$

$$\frac{\chi}{r} < \frac{\sigma}{\sigma - 1} \xrightarrow{\sigma \to \infty} 1$$

$\sigma = 3$: ratio < 1.5 $\quad$ $\sigma = 4$: ratio < 1.33

Worse than family K for large σ!

Backup: $\sigma \geqslant 3$ constructions

2-branching property: A cyclic string is 2-branching at order k if every $(k-1)$-mer extends to exactly two k-mers.

Binary de Bruijn sequences are 2-branching (extend to 0 and 1).

Results:

- Closed-form ratio: $\chi/r = (2\sigma^{k-1}+1)/(\sigma^{k-1}+4)$
- Explicit order-3 construction for all $\sigma \geqslant 2$
- Gap to 2 narrows from $O(1/\sigma)$ to $O(1/\sigma^2)$
- For $\sigma \in \{3,4\}$: order-5 instances with ratio > 1.91

Available on arXiv.