[1]University of Florence, IT [2]University of Siena, IT [3]University of Palermo, IT

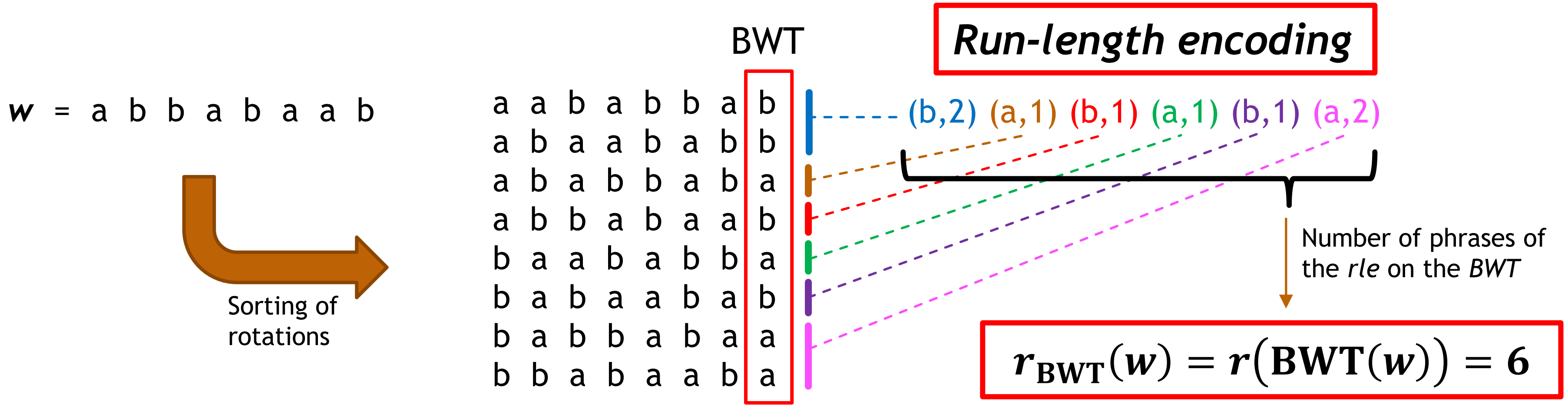
Burrows-Wheeler Transform on Purely Morphic Words

Andrea Frosini[1], *Ilaria Mancini*[2], *Simone Rinaldi*[2],
Giuseppe Romana[3], and *Marinella Sciortino*[3]

Data Compression Conference 2022
Snowbird, Utah, March 24

Burrows-Wheeler Transform and Run-Length Encoding

- Given a word w, the ***Burrows-Wheeler Transform*** of w $(\mathrm{BWT}(w))$ is the concatenation of the last characters of the lexicographically sorted rotations of w



- $r(w)$: number phrases of the run-length encoding applied to w
- $\rho(w) = \frac{r_{BWT}(w)}{r(w)}$: BWT-***clustering ratio*** [Mantaci et al., Theoret. Comput. Sci. 2017]

Morphisms and Purely Morphic Words

- Given two alphabets A and B, a ***morphism*** is a map $\varphi: A^* \mapsto B^*$ such that $\varphi(uv) = \varphi(u)\varphi(v)$ for all $u, v \in A^*$.

$$\varphi: \begin{cases} a \mapsto ab \\ b \mapsto bc \\ c \mapsto ac \end{cases}$$

a

$\varphi(a) = ab$

$\varphi^2(a) = abbc$

$\varphi^3(a) = abbcbcac$

$\varphi^4(a) = abbcbcacbcacabac$

…

Purely morphic finite words

If $\varphi(a) = au$, $u \in A^*$, then φ is called *prolongable* on a

Purely morphic word

Fixed-point $\longrightarrow$ $\boldsymbol{\varphi^\infty(a)}$: $abbcbcacbcacabacbcacabacabbcabac$…

Purely morphic words: Thue-Morse & Fibonacci

$$\tau: \begin{cases} a \mapsto ab \\ b \mapsto ba \end{cases}$$

a

ab

$abba$

$abbabaab$

$\dots$

$T = \quad abbabaabbaababbabaababbaabbabaab\dots$

Thue-Morse word

$$\theta: \begin{cases} a \mapsto ab \\ b \mapsto a \end{cases}$$

a

ab

aba

$abaab$

$\dots$

$F = \ abaababaabaababaababaabaababaabaab\dots$

Fibonacci word

Morphisms & Data Compression

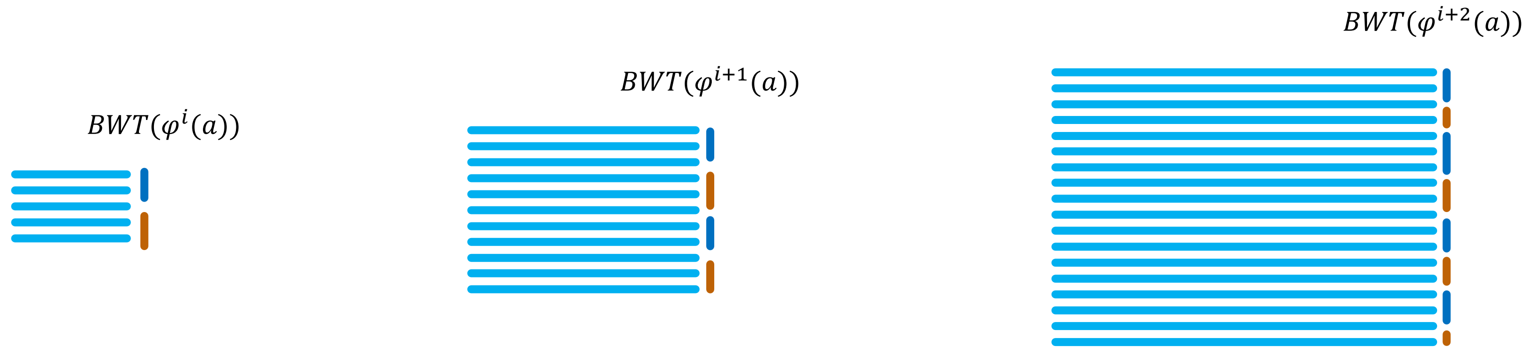
- Some repetitiveness measures have been studied for families of words generated by morphisms
  - LZ77 complexity z **[Constantinescu & Ilie, SIAM J. Discret. Math., 2007]**
  - Smallest string attractor γ **[Schaeffer & Shallit, arXiv, 2020]**

[Kempa & Prezza, STOC 2018]

- NU-systems **[Navarro & Urbina, SPIRE 2021]** are based on morphisms

r_{BWT} on purely morphic finite words

- **Question 1**
  - Given a morphism φ such that $\varphi^{\infty}(a)$ is a purely morphic word, can we bound $r_{BWT}\left(\varphi^{i}(a)\right)$?



- **Question 2**
  - Can we evaluate the BWT-clustering ratio $\rho(\varphi^{i}(a))$?

So far: r_{BWT} on finite Thue-Morse & Fibonacci words

$$\tau: \begin{cases} a \mapsto ab \\ b \mapsto ba \end{cases}$$

$$T_i = \tau^i(a)$$

$$T_3 \mapsto \begin{array}{ccccccc|c} & & & & & & & \text{BWT} \\ a & a & b & a & b & b & a & b \\ a & b & a & a & b & a & b & b \\ a & b & a & b & b & a & b & a \\ a & b & b & a & b & a & a & b \\ b & a & a & b & a & b & b & a \\ b & a & b & a & a & b & a & b \\ b & a & b & b & a & b & a & a \\ b & b & a & b & a & a & b & a \end{array}$$

$$\theta: \begin{cases} a \mapsto ab \\ b \mapsto a \end{cases}$$

$$F_i = \theta^i(a)$$

$$F_4 \mapsto \begin{array}{ccccccc|c} & & & & & & & \text{BWT} \\ a & a & b & a & a & b & a & b \\ a & a & b & a & b & a & a & b \\ a & b & a & a & b & a & a & b \\ a & b & a & a & b & a & b & a \\ a & b & a & b & a & a & b & a \\ b & a & a & b & a & a & b & a \\ b & a & a & b & a & b & a & a \\ b & a & b & a & a & b & a & a \end{array}$$

- [Brlek et al., IWOCA 2019]
  - $r_{BWT}(T_i) = 2i$ for any $i > 0$
- [Mantaci et al., Inf. Process. Lett. 2003]
  - $r_{BWT}(F_i) = 2$ for any $i > 0$

Factor complexity of purely morphic words

- x: infinite or finite word
- ***factor complexity*** $f_x(n)$: number of distinct factors of length n that occur in x.

Periodic fixed-points

$$x = \varphi^{\infty}(a) = \begin{cases} v^{\omega} = vvvvv \dots vvv \dots \\ uv^{\omega} = uvvvv \dots vvv \dots \end{cases} \Rightarrow r_{\mathrm{BWT}}\left(\varphi^i(a)\right) \in \Theta(1)$$

- $f_x(n) = \Theta(1)$

Aperiodic fixed-points classification [Pansiot, ICALP 1984]

- Let x=$\varphi^{\infty}(a)$ be an *aperiodic* purely morphic word. Then, only one of the following is true:

Thue-Morse

Fibonacci

- $f_x(n) = \Theta(n)$
- $f_x(n) = \Theta(n \log \log n)$
- $f_x(n) = \Theta(n \log n)$
- $f_x(n) = \Theta(n^2)$

$$\Rightarrow r_{\mathrm{BWT}}\left(\varphi^i(a)\right) \in ?$$

Upper bounds for r_{BWT}

- **Proposition**

- Let $x = \varphi^{\infty}(a)$ be an infinite aperiodic word. Then the following upper bounds for $r_{\mathrm{BWT}}\left(\varphi^{i}(a)\right)$ hold:
  - if $f_x(n) \in \Theta(n)$ then $r_{\mathrm{BWT}}\left(\varphi^{i}(a)\right) \in O(i)$;
  - if $f_x(n) \in \Theta(n \log\log n)$ then $r_{\mathrm{BWT}}\left(\varphi^{i}(a)\right) \in O(i \log i \log\log i)$;
  - if $f_x(n) \in \Theta(n \log n)$ then $r_{\mathrm{BWT}}\left(\varphi^{i}(a)\right) \in O(i^2 \log i)$.

[Kempa & Kociumaka, FOCS 2020]

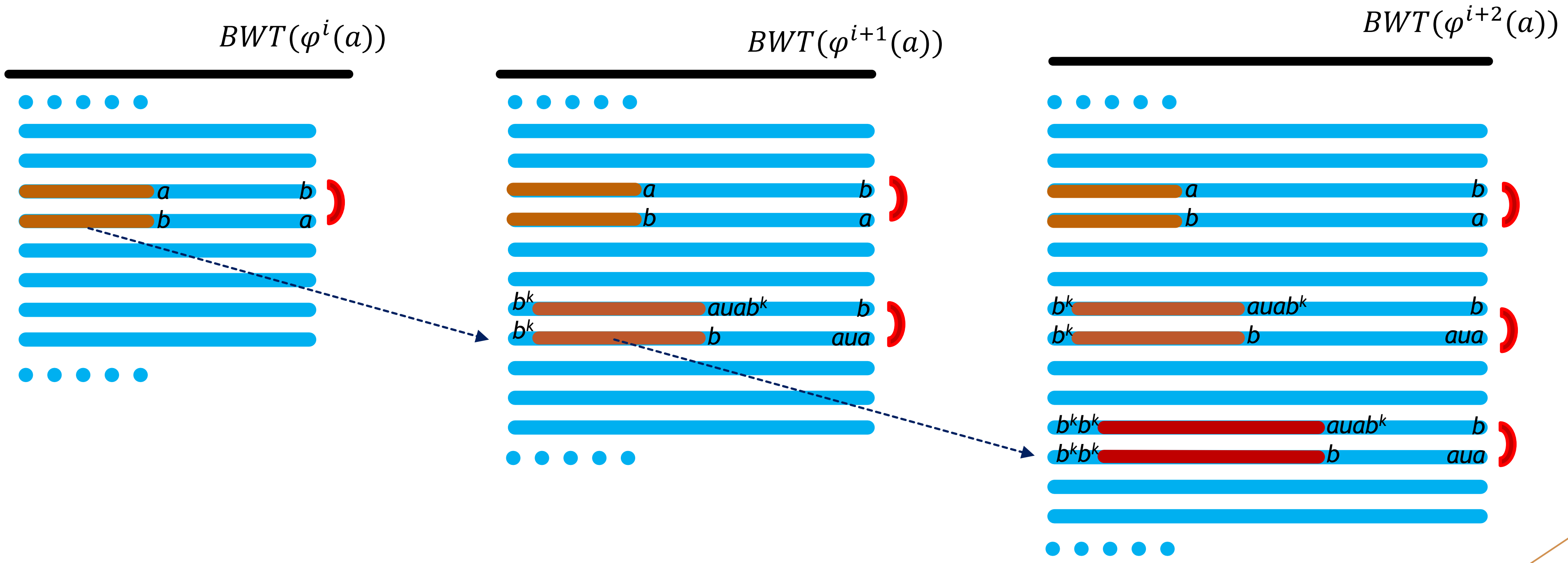
[Raskhodnikova et al., Algorithmica 2013]

- In the proof a relationship between r_{BWT} and the measure δ (related to the factor complexity) is also used

- Such a result does not provide a significative upper-bound when $f_x(n) = \Theta(n^2)$

$f_x(n) = \Theta(n^2)$: binary alphabet A=$\{a, b\}$

$$\varphi: \begin{cases} a \mapsto auab^k \\ b \mapsto b \end{cases} \text{ with } \begin{matrix} k > 0 \\ u \in A^* \end{matrix} \quad \Leftrightarrow \quad f_{\varphi^\infty(a)}(n) = \Theta(n^2)$$



- There exists i_0 such that at each step $i \geq i_0$, we add a constant number of runs
  - $r_{\text{BWT}}\left(\varphi^i(a)\right) \in O(i)$, for any $i > 0$

Binary morphisms

- Summing up, for binary morphisms we have the following bounds for r_{BWT} on binary purely morphic finite words

	$\boldsymbol{f_x(n)}$	$\boldsymbol{r_{BWT}\left(\varphi^i(a)\right)}$
	$\Theta(1)$	$\Theta(1)$
$x = \varphi^\infty(a)$	$\Theta(n)$	$O(i)$
	$\Theta(n \log\log n)$	$O(i \log i \log\log i)$
	$\Theta(n \log n)$	$O(i^2 \log i)$
	$\Theta(n^2)$	$O(i)$

$$\boxed{\begin{array}{l} i \in \Theta\left(\log|\varphi^i(a)|\right) \\ \text{if } \varphi \neq \begin{cases} a \mapsto ab^k \\ b \mapsto b \end{cases} \end{array}}$$

- On the other hand, we proved that

$$r\left(\varphi^i(a)\right) \in \begin{cases} \Theta(1) \text{ if } \varphi\colon \begin{cases} a \mapsto ab^k \\ b \mapsto b^\ell \end{cases}, \text{ with } k, \ell \geq 1 \longrightarrow \text{Purely morphic word } x = abbbbb\ldots \text{ for all } k, \ell \geq 1 \\ \Omega(2^i) \text{ otherwise} \longrightarrow \end{cases}$$

$$\boxed{\text{Exists } i_0 \text{ such that, for all } i \geq i_0,\quad \boldsymbol{\rho\left(\varphi^i(a)\right) = \frac{r_{BWT}\left(\varphi^i(a)\right)}{r\left(\varphi^i(a)\right)} \ll 1}}$$

Further works and open problems

- Results on binary morphisms have been improved in [Frosini, Mancini, Rinaldi, R. and Sciortino, *Logarithmic equal-letter runs for BWT of purely morphic words*, Developments in Language Theory (DLT-2022)]
  - $r_{BWT}\left(\varphi^{i}(a)\right) \in O(i)$ for any binary prolongable morphism
  - If $f_x(n)$ is $\Theta(n \log\log n)$ or $\Theta(n \log n)$ or $\Theta(n^2)$, then $r_{BWT}\left(\varphi^{i}(a)\right) \in \Theta(i)$

- Open problems
  - Can we extend the bounds on r_{BWT} for all prefixes of the fixed point?
  - Can we extend the tighter upper-bounds for larger alphabet?

Thanks for your attention