Burrows-Wheeler Transform on Purely Morphic Words

Andrea Frosini\textsuperscript{1}, Ilaria Mancini\textsuperscript{2}, Simone Rinaldi\textsuperscript{2}, Giuseppe Romana\textsuperscript{3}, and Marinella Sciortino\textsuperscript{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 \( w \) (\( \text{BWT}(w) \)) is the concatenation of the last characters of the lexicographically sorted rotations of \( w \)

\[
\text{w} = a \ b \ b \ a \ b \ a \ a \ b
\]

\[
\text{BWT} = \\
a \ a \ b \ a \ b \ a \ b \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b \ a \ b
\]

- **Run-length encoding**

\[
\text{BWT}(w) = (b,2) (a,1) (b,1) (a,1) (b,1) (a,2)
\]

- Number of phrases of the rle on the BWT

\[
\rho(w) = \frac{r_{\text{BWT}}(w)}{r(w)}: \text{BWT-clustering ratio} \quad [\text{Mantaci et al., Theoret. Comput. Sci. 2017}]
\]
Morphisms and Purely Morphic Words

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

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

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

$$\varphi(a) = ab, \quad \varphi^2(a) = abbc, \quad \varphi^3(a) = abbcacbcacacabac, \quad \varphi^4(a) = abbcacbcacbcacacabacacabacabacabacabacabacabacabac...$$

**Purely morphic finite words**

**Purely morphic word**

**Fixed-point** $\varphi^\infty(a) : abbcacbcacabacabacacabacacabac...$
Purely morphic words: Thue-Morse & Fibonacci

\[ T = \text{Thue-Morse word} \]

\[ F = \text{Fibonacci word} \]
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 $\gamma$ [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 \( \varphi \) such that \( \varphi^\infty(a) \) is a purely morphic word, can we bound \( r_{BWT}(\varphi^i(a)) \)?

- **Question 2**
  - Can we evaluate the BWT-clustering ratio \( \rho(\varphi^i(a)) \)?
So far: \( r_{\text{BWT}} \) on finite Thue-Morse & Fibonacci words

- \([\text{Brlek et al., IWOCA 2019}]

  \( r_{\text{BWT}}(T_i) = 2i \) for any \( i > 0 \)

- \([\text{Mantaci et al., Inf. Process. Lett. 2003}]

  \( r_{\text{BWT}}(F_i) = 2 \) for any \( i > 0 \)
Factor complexity of purely morphic words

- **Periodic fixed-points**
  \[ x = \varphi^\infty(a) = v^\omega = vvvv \ldots vvv \ldots \]
  \[ uv^\omega = uvvvv \ldots vvv \ldots \]
  \[ f_x(n) = \Theta(1) \]

- **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:
    - \( f_x(n) = \Theta(n) \)
    - \( f_x(n) = \Theta(n \log \log n) \)
    - \( f_x(n) = \Theta(n \log n) \)
    - \( f_x(n) = \Theta(n^2) \)
    - \( r_{BWT}(\varphi^i(a)) \in \Theta(1) \)
    - \( r_{BWT}(\varphi^i(a)) \in ? \)

- **factor complexity** \( f_x(n) \): number of distinct factors of length \( n \) that occur in \( x \).

- \( x \): infinite or finite word
Upper bounds for $r_{BWT}$

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

  - [Kempa & Kociumaka, FOCS 2020]
  - [Raskhodnikova et al., Algorithmica 2013]

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

- Such a result does not provide a significative upper-bound when $f_x(n) = \Theta(n^2)$
\[ f_x(n) = \Theta(n^2) : \text{binary alphabet } A=\{a, b\} \]

\[ \varphi: \begin{cases} a \mapsto auab^k \\
    b \mapsto b \end{cases} \quad \text{with} \quad k > 0 \quad \Rightarrow \quad f_{\varphi^\infty(a)}(n) = \Theta(n^2) \]

- There exists \( i_0 \) such that at each step \( i \geq i_0 \), we add a constant number of runs
  - \( r_{BWT}(\varphi^i(a)) \in O(i) \), 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:

<table>
<thead>
<tr>
<th>$f_x(n)$</th>
<th>$r_{BWT}(\varphi^i(a))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta(1)$</td>
<td>$\Theta(1)$</td>
</tr>
<tr>
<td>$\Theta(n)$</td>
<td>$O(i)$</td>
</tr>
<tr>
<td>$\Theta(n \log \log n)$</td>
<td>$0(i \log i \log \log i)$</td>
</tr>
<tr>
<td>$\Theta(n \log n)$</td>
<td>$0(i^2 \log i)$</td>
</tr>
<tr>
<td>$\Theta(n^2)$</td>
<td>$O(i)$</td>
</tr>
</tbody>
</table>

$x = \varphi^\infty(a)$

- On the other hand, we proved that:

$r(\varphi^i(a)) \in$

- $\Theta(1)$ if $\varphi: \begin{cases} a \mapsto ab^k, & \text{with } k, \ell \geq 1 \\ b \mapsto b^\ell \end{cases}$

- $\Omega(2^i)$ otherwise

Purely morphic word $x = abbbbbb ...$

for all $k, \ell \geq 1$

Exists $i_0$ such that, for all $i \geq i_0$,

$\rho(\varphi^i(a)) = \frac{r_{BWT}(\varphi^i(a))}{r(\varphi^i(a))} < 1$
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)]
  - $r_{BWT}(\varphi^i(a)) \in O(i)$ for any binary prolongable morphism
  - If $f_{\varphi}(n)$ is $\Theta(n \log \log n)$ or $\Theta(n \log n)$ or $\Theta(n^2)$, then $r_{BWT}(\varphi^i(a)) \in \Theta(i)$

- 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