Balanced words having simple Burrows-Wheeler Transform

Antonio Restivo and Giovanna Rosone

Università degli Studi di Palermo
Italy
Motivations

• In 1994 M. Burrows and D. Wheeler introduced a new data compression method based on a preprocessing on the input string. Such a preprocessing is called the Burrows-Wheeler Transform (BWT).

• The application of the BWT produces a clustering effect (occurrences of a given symbol tend to occur in clusters).

• Perfect clustering corresponds to optimal performances of some BWT-based compression algorithms.

• We study the words where the BWT produces a perfect clustering.
How does BWT work?

BWT takes as input a text $v$ and produces:

- a permutation $bwt(v)$ of the letters of $v$.
- the index $I$, that is useful to recover the original word $v$.

Example: $v = \text{diekert}$

- Each row of $M$ is a conjugate of $v$ in lexicographic order.
- $bwt(v)$ coincides with the last column $L$ of the BW-matrix $M$.
- The index $I$ is the row of $M$ containing the original sequence.

Notice that if we except the index, all the mutual conjugate words have the same Burrows-Wheeler Transform. Hence, the BWT can be thought as a transformation acting on circular words.
Perfect clustering: Simple BWT words

Let $v$ be a word over a finite ordered alphabet $A=\{a_1,a_2,\ldots,a_k\}$ (with $a_1 < a_2 < \ldots < a_k$):

The word $v$ is a simple BWT words if

$$bwt(v) = a_{i_k}^i a_{i_{k-1}}^i \cdots a_2^i a_1^i$$

for some non-negative integers $i_k, i_{k-1}, \ldots, i_1 > 0$.

We denote by $S$ the set of the simple BWT words.

Example: $v=acbcbcadad \in S$, $bwt(v)=ddcccbbaaa$
Matrix $M$ and $R$

Example: $v=\text{diekert}$
$n=|v|=7$

Since the matrix $R$ is obtained from $M$ by a rotation of $180^\circ$ it follows that the $i$-th conjugate of $M$ is the reverse of the $(n-i+1)$-th conjugate of $R$.

Theorem.

A word $v \in S$ if and only if $M=R$. 
A word \( \nu \in S \) iff \( M=R \)

\[
F_R = F_M \text{ and } L_R = L_M
\]

\[
\nu_i = \tilde{\nu}_{n-i+1}
\]

• So \([\nu]\) and its factors are closed under reverse.
• Under these conditions each conjugate of \( \nu \) has the two palindrome property (cf. Simpson and Puglisi, 2008).
Simple BWT words

- In the case of binary alphabet, the elements of $S$ have been characterized by Mantaci, Restivo and Sciortino: they are related to Standard words and balanced words.

- In the case of a three letters alphabet a constructive characterization of the elements of $S$ has been given by Simpson and Puglisi, 2008.

- The case of larger alphabets is more complex.
Standard words:

**Directive sequence:** \(d_1, d_2, \ldots, d_n, \ldots\) \(d_i \geq 0, \ d_i > 0 \) for \(i > 1\)

\[
\begin{align*}
    s_0 &= b \\
    s_1 &= a \\
    s_{n+1} &= s_n^{d_n} s_{n-1} \quad n \geq 1
\end{align*}
\]

Standard words are special prefixes of Sturmian sequences.

**Fibonacci words:**

\[
\begin{align*}
    f_0 &= b \\
    f_1 &= a \\
    f_2 &= ab \\
    f_3 &= aba \\
    f_4 &= abaab
\end{align*}
\]

\[
\begin{align*}
    f_0 &= b \\
    f_1 &= a \\
    f_{n+1} &= f_n f_{n-1} \quad (n \geq 1)
\end{align*}
\]
A word \( \nu \) is **balanced** if for all letters \( a \) of the alphabet \( A \) we have for all factors \( u \) and \( u' \) of \( \nu \) s.t. \( |u| = |u'| \) then \( ||u|_a - |u'|_a| \leq 1 \).

A finite word is **circularly-balanced** if all its conjugates are balanced.

For instance:

- \( w = cacbcac \) is circularly balanced word.
- \( \nu = acacbbc \) is unbalanced word.
Binary alphabets

Theorem (Mantaci, Restivo and Sciortino, 2003)

In the binary case, the following sets of words coincide:

- *simple BWT words*;
- circularly balanced words;
- conjugates of a power of a Standard words.
Generalization to alphabets with more than two letters

In alphabets with more than two letters, the following sets do not coincide:

- circularly balanced words;
- simple BWT words;
- finite epistandard words (a generalization of the Standard words).

Remark

The problem of characterizing balanced words over large alphabets is still open and it is related to a conjecture of Fraenkel.
Balancing and BWT

The BWT of *circularly balanced* words over more than two letters alphabets does *not always* produce a “perfect clustering”.

For instance:

\[ \nu = \text{cacbcac} \text{ is circularly balanced and } \text{bwt}(\nu) = \text{ccccbaa} \]
\[ \omega = \text{ababc} \text{ is circularly balanced and } \text{bwt}(\omega) = \text{cbbaab} \]

Moreover, there exist *unbalanced* words that produce perfect clustering.

For instance:

\[ \upsilon = \text{acacbbc} \text{ is unbalanced and } \text{bwt}(\upsilon) = \text{cccbbaa} \]
A generalization of Sturmian: Episturmian

An infinite word \( t \) on \( A \) is \textit{episturmian} (Droubay, J. Justin, G. Pirillo, 2001) if:

- \( F(t) \) (its set of factors) is \textit{closed under reversal},
- \( t \) has \textit{at most one right special factor of each length}.

\( t \) is \textit{standard episturmian} if all of its left special factors are prefixes of it.

An infinite word on the finite alphabet \( A \) is \textit{standard episturmian} if and only if it can be obtained by the \textit{Rauzy rules} for \( A \).

Let \( s \) be an infinite word, then a factor \( u \) of \( s \) is \textit{right} (resp. \textit{left}) \textit{special} if there exist \( x,y \in A, x \neq y \), such that \( ux, uy \in F(s) \) (resp. \( xu, yu \in F(s) \)).
Let $|A|=k$ be. A word $v \in A^*$ is called **finite epistandard** if $v$ is an element of a $k$-tuples $R_n$, for some $n \geq 1$.

We denote by $EP$ the set of words that are a power of a conjugate of a finite epistandard word.
Balancing and Episturmiand Theorem (Paquin and Vuillon, 2006):

Any balanced standard episturmiand sequence $s$ over an alphabet with 3 or more letters is of the form $s = t^\omega$, where $t$ is a finite epistandard word that belongs to one of the following three families (up to letter permutation):

1. $t = (pa_2)$ and $p = Pal(a_1^m a_k a_{(k-1)} ... a_3)$, where $k \geq 3$ and $m > 0$;
2. $t = (pa_2)$ and $p = Pal(a_1 a_k ... a_{(k-l)} a_1 a_{(k-l-1)} ... a_3)$, where $0 \leq l \leq k-4$ and $k \geq 4$;
3. $t = Pal(a_1 a_k ... a_3 a_2)^\omega$, where $k \geq 3$ (Fraenkel’s sequence), where the operator $Pal$ is the iterated palindromic closure function.

Since $s$ is balanced, then the finite word $t$ is circularly balanced.
A finite word $v$ is *rich* if it has exactly $|v| + 1$ distinct palindromic factors, including $\varepsilon$ (Droubay, Justin, Pirillo, 2001).

A finite or infinite word is *rich* if all of its factors are rich.

**Example:**

$v = \text{diekert}$ is rich, $|v| = 7$, in fact:

$P(v) = \{\varepsilon, d, e, i, k, r, t, eke\}$, $|P(v)| = 8$. 

Circularly rich words

For a finite word $\nu$, the following properties are equivalent (Glen, Justin, Widmer and Zamboni, 2009):

- $\nu^\omega$ is rich;
- $\nu^2$ is rich;
- $\nu$ is a product of two palindromes and all of the conjugates of $\nu$ (including itself) are rich.

- We say that a finite word $\nu$ is circularly rich if the infinite word $\nu^\omega$ is rich.
- We say that $R$ is the set of the circularly rich words.
Our theorem

Let $A=\{a_1, a_2, \ldots, a_k\}$ be a totally ordered alphabet.

Let $\nu \in A^*$ be a **circularly balanced** over $A$, the following statements are equivalent:

1) $\nu \in S$ (simple BWT words);
2) $\nu$ is circularly rich;
3) $\nu$ is a conjugate of a power of a **finite epistandard**.
Proof: $3 \rightarrow 1$

The finite balanced epistandard words belong to $S$.

From the result by Paquin and Vuillon, we have to prove that each finite balanced epistandard word $t$ of the form:

1. $t = pa_3pa_2$ and $p = Pal(a_1^m a_k a_{(k-1)} \ldots a_4)$, where $k \geq 3$ and $m > 0$;

2. $t = pa_3pa_2$ and $p = Pal(a_1 a_k \ldots a_{(k-l)} a_1 a_{(k-l-1)} \ldots a_4)$, where $l \geq 1$ and $k \geq 4$;

3. $t = Pal(a_1 a_k \ldots a_3 a_2)$, where $k \geq 3$ (Fraenkel's word).

belongs to $S$.

The proof follows from the structure of $t$ and from the construction of BW-matrix.
Proof: $2 \iff 3$:

$v$ is circularly rich if and only if $v$ is a conjugate of a power of a finite epistandard

The proof is an immediate consequence of the fact that

- The set of the **episturmian sequences** is a **subset** of the set of the **rich words**. (Glen, Justin, Widmer and Zamboni, 2009).

- Recurrent **balanced rich** infinite words are precisely the **balanced episturmian** words (Glen, Justin, Widmer and Zamboni, 2009).
Proof: $1 \rightarrow 2$

If the word $w$ belongs to $S$ then $w$ is circularly rich.

We know that

- $w$ is circularly rich if and only if $w$ is a product of two palindromes and all the conjugates of $w$ (including itself) are rich.
- each word $w \in S$ has the two palindrome property.

We prove that

If $w \in S$ then all the conjugates of $w$ (including itself) are rich.
Example: 1 \rightarrow 2

If the word \( w \in S \) then \( w \) is circularly rich.

- The word \( v = acbcbcadad \in S \), \( |v| = 10 \), in fact
  \( \text{bwt}(acbcbcadad) = ddcccbbaaa \)
  \( |P(v^2)| = 21 \), so \( v \) is circularly rich.

We note that the converse of this result is false.
- The word \( u = ccaaccb \) is circularly rich, but
  \( \text{bwt}(ccaaccb) = caccccbba \) (\( u \notin S \)).
Conclusions and examples

Only under the condition of circularly balanced, the following statements are equivalent:

1) $v \in S$ (simple BWT words);
2) $v$ is circularly rich,
3) $v$ is a conjugate of a power of a finite epistandard.

In fact the circularly unbalanced word:

- $w=bbbbbacaca \in S$ (clearly it is circularly rich), but $w \notin EP$.
- $u=(adac)^2adab(adac)^2ada(adac)^2adab(adac) \notin S$ and $u \in EP$.

The following example shows that there exist words unbalanced which belong to $EP \cap S$:

- $v= aadaaca ad$ is a circularly unbalanced word: $v \in EP$ and $v \in S$. 
Further works

- Characterize the words in $S$
  (we have only characterized the balanced words in $S$).

- Introduce measures of balancing on words and study
  the effect of BWT on such measures
  (this corresponds to study the \textit{clustering effect of BWT}).
Thank you for your attention!