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=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.

$$
\begin{array}{ccccccc|l}
F & & & M & & & L & \\
\downarrow & & & & & & \downarrow & \\
d & i & e & k & e & r & \boxed{t} & \leftarrow I \\
e & k & e & r & t & d & i & \\
e & r & t & d & i & e & k & \\
i & e & k & e & r & t & d & \\
k & e & r & t & d & i & e & \\
r & t & d & i & e & k & e & \\
t & d & i & e & k & e & r & \\
\end{array}
$$

bwt(v)

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

$$bwt(v) = a_k^{i_k} a_{k-1}^{i_{k-1}} \cdots a_2^{i_2} a_1^{i_1}$$

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=*diekert*
$n=|v|=7$

$$\begin{array}{ccccccc} F_M & & & M & & & L_M \\ d & i & e & k & e & r & t \\ e & k & e & r & t & d & i \\ e & r & t & d & i & e & k \\ i & e & k & e & r & t & d \\ k & e & r & t & d & i & e \\ r & t & d & i & e & k & e \\ t & d & i & e & k & e & r \end{array} \quad \xrightarrow{180^\circ} \quad \begin{array}{ccccccc} F_R & & & R & & & L_R \\ r & e & k & e & i & d & t \\ e & k & e & i & d & t & r \\ e & i & d & t & r & e & k \\ d & t & r & e & k & e & i \\ k & e & i & d & t & r & e \\ i & d & t & r & e & k & e \\ t & r & e & k & e & i & d \end{array}$$

Since the matrix R is obtained from M by a rotation of 180° 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 $v \in S$ iff $M{=}R$

$F_R = F_M$ and $L_R = L_M$

a	c	b	c	b	c	a	d	a	d	
a	d	a	c	b	c	b	c	a	d	
a	d	a	d	a	c	b	c	b	c	v_i
b	c	a	d	a	d	a	c	b	c	
b	c	b	c	a	d	a	d	a	c	
c	a	d	a	d	a	c	b	c	b	
c	b	c	a	d	a	d	a	c	b	
c	b	c	b	c	a	d	a	d	a	v_{n-i+1}
d	a	c	b	c	b	c	a	d	a	
d	a	d	a	c	b	c	b	c	a	

$$v_i = \widetilde{v}_{n-i+1}$$

- So *[v]* and its factors are closed under reverse.
- Under these conditions each conjugate of *v* 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

Standard words:

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

$$s_0 = b \quad s_1 = a$$

$$s_{n+1} = s_n^{d_n} s_{n-1} \qquad n \geq 1$$

Standard words are special prefixes of Sturmian sequences.

Fibonacci words:

$f_0 = b$

$f_1 = a$

$f_2 = ab$

$f_3 = aba$

$f_4 = abaab$

$f_0 = b \qquad f_1 = a$

$f_{n+1} = f_n f_{n-1} \quad (n \geq 1)$

Balancing

A word v is **balanced** if for all letters a of the alphabet A we have for all factors $\boldsymbol{u}$ and $\boldsymbol{u}'$ of $\boldsymbol{v}$ s.t. $|\boldsymbol{u}|=|\boldsymbol{u}'|$ then $||\boldsymbol{u}|_{\boldsymbol{a}}-|\boldsymbol{u}'|_{\boldsymbol{a}}|\leq \boldsymbol{1}$.

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

For instance:

$w=cacbcac$ is circularly balanced word.

$v=acacbbc$ 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 <u>not always</u> produce a "perfect clustering".

For instance:

v=*cacbcac* is circularly balanced and *bwt(v)=ccccbaa*

w=*ababc* is circularly balanced and *bwt(w)=cbaa*<u>*b*</u>

Moreover there exist unbalanced words that produce perfect clustering.

For instance:

u=<u>*ac*</u>*ac*<u>*bb*</u>*c* is unbalanced and *bwt(u)= cccbbaa*

A generalization of Sturmian: Episturmian

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

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

t is *standard episturmian* if all of its left special factors are prefixes of it.

An infinite word on the finite alphabet A is standard episturmian if and only if it can be obtained by the ***Rauzy rules*** for A.

Let s be an infinite word, then a factor u of s is right (resp. left) special if there exist $x,y \in A$, $x \neq y$, such that $ux, uy \in F(s)$ (resp. $xu, yu \in F(s)$).

Finite epistandard and Rauzy rules

	Rules:	1	2	3	4
R_0		*a*	*b*	*c*	*d*
R_1	1	***a***	***a**b*	***a**c*	***a**d*
R_2	1	*a*	***a**ab*	***a**ac*	***a**ad*
R_3	4	***aad**a*	***aad**aab*	***aad**aac*	*aad*
R_4	3	***aadaac**aada*	***aadaac**aadaab*	*aadaac*	***aadaac**aad*

- 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 Episturmian

Theorem (Paquin and Vuillon, 2006):

Any balanced standard episturmian 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 $p = Pal(a_1{}^m\, a_k a_{(k-1)} \ldots a_3)$, where $k \geq 3$ and m>0;
2. $t = (pa_2)$ and $p = Pal(a_1\, a_k \ldots a_{(k-l)}\, a_1\, a_{(k-l-1)} \ldots a_3)$, where $0 \leq l \leq k-4$ and $k \geq 4$;
3. $t = Pal(a_1\, a_k \ldots a_3\, a_2)^\omega$, 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.

Rich words

- A finite word v *is rich if it has exactly* $|v| + 1$ *distinct* palindromic factors, including ε (Droubay, Justin, Pirillo, 2001).

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

- Example:

  v= *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 v, the following properties are equivalent (Glen, Justin, Widmer and Zamboni, 2009):

- v^{ω} is rich;
- v^2 is rich;
- v is a <u>product of two palindromes</u> and <u>all</u> of the conjugates of v (including itself) are rich.

- We say that a finite word v is circularly rich if the infinite word v^{ω} 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 $v \in A^*$ be a circularly balanced over A, 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.

Proof: 3→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⟷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→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 → 2

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

- The word v=*acbcbcadad* $\in S$, $|v|=10$, in fact $bwt(acbcbcadad)= ddcccbbaaa$

  $|P(v^2)| = 21$, so v is circularly rich.

We note that the converse of this result is false.

- The word u=*ccaaccb* is circularly rich, but $bwt(ccaaccb)=cacccba$ ($u \notin S$).

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 $w \notin EP$.
- $u = (adac)^2 adab (adac)^2 ada (adac)^2 adab (adac) \notin S$ and $u \in EP$.

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

- $v = \underline{aad}\underline{aaca}a\underline{d}$ 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 *clustering effect* of BWT).

Thank you
for your attention!