Burrows-Wheeler Transform and Run-Length Enconding*

Sabrina Mantaci[1], Antonio Restivo[1], Giovanna Rosone[2], and Marinella Sciortino[1]

[1]University of Palermo, Palermo, Italy
[2]University of Pisa, Pisa, Italy

Abstract

In this paper we study the clustering effect of the Burrows-Wheeler Transform (BWT) from a combinatorial viewpoint. In particular, given a word w we define the BWT-clustering ratio of w as the ratio between the number of clusters produced by BWT and the number of the clusters of w. The number of clusters of a word is measured by its Run-Length Encoding. We show that the BWT-clustering ratio ranges in $]0, 2]$. Moreover, given a rational number $r \in]0, 2]$, it is possible to find infinitely many words having BWT-clustering ratio equal to r. Finally, we show how the words can be classified according to their BWT-clustering ratio. The behavior of such a parameter is studied for very well-known families of binary words.

1 Introduction

Burrows-Wheeler Transform is a popular method used for text compression (cf. [1, 3]). It produces a permutation of the characters of an input word w in order to obtain a word easier to compress. Actually compression algorithms based on BWT take advantage of the fact that the word output of BWT

*The final authenticated publication is available online at https://doi.org/10.1007/978-3-319-66396-8_21. Please, cite the publisher version: Mantaci S., Restivo A., Rosone G., Sciortino M. (2017) Burrows-Wheeler Transform and Run-Length Enconding. In: Brlek S., Dolce F., Reutenauer C., Vandomme É. (eds) Combinatorics on Words. WORDS 2017. Lecture Notes in Computer Science, vol 10432. Springer, Cham DOI: https://doi.org/10.1007/978-3-319-66396-8_21

shows a local similarity (occurrences of a given symbol tend to occur in clusters) and then turns out to be highly compressible. Several authors refer to such a property as the "clustering effect" of BWT. The aim of this paper is to study such a clustering effect of BWT from the point of view of combinatorics on words.

In order to measure the amount of local similarity, or clustering, in a word we consider its Run-Length Encoding (RLE). RLE is another fundamental string compression technique: it replaces in a word occurrences of repeated equal symbols with a single symbol and a non-negative integer (run length) counting the number of times the symbol is repeated. RLE can be considered an efficient compression scheme when the input data is highly repetitive. In a more formal way, every word w over the alphabet Σ has a unique expression of the form $w = w_1^{l_1} w_2^{l_2} \cdots w_k^{l_k}$ with $l_i \in \mathbb{N}$ and $w_i \in \Sigma$ and $w_i \neq w_{i+1}$ for $i = 1, 2, \ldots, k$. The run-length encoding of w is the sequence $\mathtt{rle}(w) = (w_1, l_1)(w_2, l_2) \cdots (w_k, l_k)$. For instance if $w = aaabbbbbccbbbb$ the run-length encoding is $\mathtt{rle} = (a, 3)(b, 5)(c, 2)(b, 4)$. We set $\rho(w) = |\mathtt{rle}(w)|$, i.e., $\rho(w)$ is the number of maximal runs of equal letters in w. For instance, $\rho(aaabbbbbccbbbb) = 4$. It is straightforward that $1 \leq \rho(w) \leq |w|$. The quantity $|w|/\rho(w)$ provides a measure of the amount of local similarity of the word w, in the sense that the lower is the value $\rho(w)$ with respect to $|w|$, the greater is the length of the runs of individual symbols in w.

In this paper we are interested to investigate the "clustering effect" of BWT, extending some results presented in [8]. For this aim we introduce for any word its BWT-clustering ratio

$$\gamma(w) = \frac{\rho(\mathtt{bwt}(w))}{\rho(w)}$$

where $\mathtt{bwt}(w)$ denotes the output of BWT on the input word w. Our first result (Theorem 1) states that, for any word w, $0 < \gamma(w) \leq 2$. This means that, if the number of runs increases after the application of the BWT ("un-clustering effect"), in the worst case the number of runs in the output is at most twice the number of runs in the original word. In other words, whereas the "clustering effect" for some words w could be very high ($\gamma(w)$ close to 0), the "un-clustering effect" is in any case moderate. The fact that the worst case is not too bad provides an additional formal motivation of usefulness of BWT in Data Compression.

We further prove (Theorem 2) that, for any rational number r, with $0 < r \leq 2$, there exists a word w such that $\gamma(w) = r$.

Previous results suggest that the parameter $\gamma(w)$ could be an interesting tool for the study (or classification) of finite words. In particular, we derive

a characterization of Christoffel words w in terms of $\gamma(w)$ and we determine the possible values of $\gamma(w)$ for a de Bruijn word w.

Finally in Section 5 we show the results of some statistical experiments that classify words in terms of their BWT-clustering ratio γ.

2 Burrows-Wheeler Transform

Let $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ be a finite ordered alphabet with $a_1 < a_2 < \ldots < a_\sigma$, where $<$ denotes the standard lexicographic order. We denote by Σ^* the set of words over Σ. Given a finite word $w = w_1 w_2 \cdots w_n \in \Sigma^*$ with each $w_i \in \Sigma$, the length of w, denoted $|w|$, is equal to n. We denote by $alph(w)$ the subset of Σ containing all the letters that appear in w. Given a finite word $w = w_1 w_2 \cdots w_n$ with each $w_i \in \Sigma$, a *factor* of a word w is written as $w[i, j] = w_i \cdots w_j$ with $1 \leq i \leq j \leq n$. A factor of type $w[1, j]$ is called a *prefix*, while a factor of type $w[i, n]$ is called a *suffix*. We also denote by $w[i]$ the i-th letter in w for any $1 \leq i \leq n$.

We say that two words $x, y \in \Sigma^*$ are *conjugate*, if $x = uv$ and $y = vu$, where $u, v \in \Sigma^*$. Conjugacy between words is an equivalence relation over Σ^*. The *conjugacy class* (w) of $w \in \Sigma^n$ (or *necklace*) is the set of all words $w_i w_{i+1} \cdots w_n w_1 \cdots w_{i-1}$, for any $1 \leq i \leq n$. A necklace can be also thought as a cyclic word.

A nonempty word $w \in \Sigma^*$ is *primitive* if $w = u^h$ implies $w = u$ and $h = 1$.

A *Lyndon word* is a primitive word which is the minimum in its conjugacy class, with respect to the lexicographic order relation.

The *Burrows-Wheeler Transform* (BWT) can be described as follows: given a word $w \in \Sigma^*$, the output of BWT is the pair $(\texttt{bwt}(w), I)$, where:

- $\texttt{bwt}(w)$ is the permutation of the letters in the input word w obtained by considering the matrix M containing the lexicographically sorted list of the conjugates of w, and by concatenating the letters of the last column L of matrix M.
- I is the index of the row of M containing the original word w.

Note that if two words v and w are conjugate then $\texttt{bwt}(v) = \texttt{bwt}(w)$, i.e. the output of BWT is the same up to the second component of the pair. Note also that the first column F of the matrix M is the sequence of lexicographically sorted symbols of w.

The Burrows-Wheeler transform is reversible by using the properties (cf. [3]) described in the following proposition.

F															L
$\downarrow$															$\downarrow$
a	a	a	b	a	a	b	a	a	b	a	a	b	a	a	b
a	a	b	a	a	a	b	a	a	b	a	a	b	a	a	b
a	a	b	a	a	b	a	a	a	b	a	a	b	a	a	b
a	a	b	a	a	b	a	a	b	a	a	a	b	a	a	b
a	a	b	a	a	b	a	a	b	a	a	b	a	a	a	b
a	a	b	a	a	b	a	a	b	a	a	b	a	a	b	a
a	b	a	a	a	b	a	a	b	a	a	b	a	a	b	a
a	b	a	a	b	a	a	a	b	a	a	b	a	a	b	a
a	b	a	a	b	a	a	b	a	a	a	b	a	a	b	a
a	b	a	a	b	a	a	b	a	a	b	a	a	a	b	a
a	b	a	a	b	a	a	b	a	a	b	a	a	b	a	a
b	a	a	a	b	a	a	b	a	a	b	a	a	b	a	a
b	a	a	b	a	a	a	b	a	a	b	a	a	b	a	a
b	a	a	b	a	a	b	a	a	a	b	a	a	b	a	a
b	a	a	b	a	a	b	a	a	b	a	a	a	b	a	a
b	a	a	b	a	a	b	a	a	b	a	a	b	a	a	a

(a)

F															L
$\downarrow$															$\downarrow$
a	a	a	a	b	a	a	b	b	a	b	a	b	b	b	b
a	a	a	b	a	a	b	b	a	b	a	b	b	b	b	a
a	a	b	a	a	b	b	a	b	a	b	b	b	b	a	a
a	a	b	b	a	b	a	b	b	b	b	a	a	a	a	b
a	b	a	a	b	b	a	b	a	b	b	b	b	a	a	a
a	b	a	b	b	b	b	a	a	a	a	b	a	a	b	b
a	b	b	a	b	a	b	b	b	b	a	a	a	a	b	a
a	b	b	b	b	a	a	a	a	b	a	a	b	b	a	b
b	a	a	a	a	b	a	a	b	b	a	b	a	b	b	b
b	a	a	b	b	a	b	a	b	b	b	b	a	a	a	a
b	a	b	a	b	b	b	b	a	a	a	a	b	a	a	b
b	a	b	b	b	b	a	a	a	a	b	a	a	b	b	a
b	b	a	a	a	a	b	a	a	b	b	a	b	a	b	b
b	b	a	b	a	b	b	b	b	a	a	a	a	b	a	a
b	b	b	a	a	a	a	b	a	a	b	b	a	b	a	b
b	b	b	b	a	a	a	a	b	a	a	b	b	a	b	a

(b)

Figure 1: On the left (a) the matrix of all lexicographic sorted conjugates of the Lyndon word *aaabaabaabaabaab*. In this case the output of BWT is the pair $(bbbbbaaaaaaaaaaa, 1)$. On the right (b) the matrix M of the word *aaaabaabbbababbbb*. For such a word BWT outputs the pair $(baababababbababa, 1)$

Proposition 1. *Let (L, I) be a pair produced by the BWT applied to a word w. Let F be the sequence of the sorted letters of $L = \texttt{bwt}(w)$. The following properties hold:*

1. *for all $i = 1, \ldots, n$, $i \neq I$, the letter $F[i]$ follows $L[i]$ in the original string w;*
2. *for each letter c, the r-th occurrence of c in F corresponds to the r-th occurrence of c in L;*
3. *the first letter of w is $F[I]$.*

From the above properties it follows that the BWT is reversible in the sense that, given L and I, it is possible to reconstruct the original string w. Note that when $I = 1$, one can build the Lyndon conjugate of the original word.

Actually, according to Property 2 of Proposition 1, we can define a permutation $\tau\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$ where τ gives the correspondence between the positions of letters of F and L. The permutation τ is also called *FL-mapping*.

The permutation τ also represents the order in which we have to rearrange the elements of F to reconstruct the original word w. Hence, starting from I, we can recover the word w as follows:

$$w[i] = F[\tau^{i-1}(I)]\ , \text{ where } \tau^0(x) = x, \text{ and } \tau^i(x) = \tau(\tau^{i-1}(x)), \text{with } 1 \leq i \leq n.$$

Example 1. *Let us consider the words examined in Figure 1.*

Given the pair $(bbbbbaaaaaaaaaaa, 1)$ the permutation τ between the positions of $F = aaaaaaaaaaabbbbb$ and $L = bbbbbaaaaaaaaaaa$ is the following:

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

So, we can reconstruct the word $w = aaabaabaabaabaab$.

If we consider the pair $(baababababbababa, 1)$ the permutation τ between $F = aaaaaaaaaaabbbbb$ and $L = baababababbababa$ is :

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 2 & 3 & 5 & 7 & 10 & 12 & 14 & 16 & 1 & 4 & 6 & 8 & 9 & 11 & 13 & 15 \end{pmatrix}$$

So, the recovered word is $w = aaaabaabbababbbb$.

3 BWT-Clustering Ratio of a Word

The Run-Length Encoding is a fundamental string compression technique that replaces in a word occurrences of repeated equal symbols with a single symbol and a non negative integer (run length) counting the number of times the symbol is repeated. Formally, every word w over the alphabet Σ has a unique expression of the form $w = w_1^{l_1} w_2^{l_2} \cdots w_k^{l_k}$ with $l_i \in \mathbb{N}$ and $w_i \in \Sigma$ and $w_i \neq w_{i+1}$ for $i = 1, 2, \ldots, k$. The *run-length encoding* of a word w, denoted by $\texttt{rle}(w)$, is a sequence of pairs (w_i, l_i) such that $w_i w_{i+1} \cdots w_{i+l_i-1}$ is a maximal run of a letter w_i (i.e., $w_i = w_{i+1} = \cdots = w_{i+l_i-1}$, $w_{i-1} \neq w_i$ and $w_{i+l_i} \neq w_i$), and all such maximal runs are listed in $\texttt{rle}(w)$ in the order they appear in w. We denote by $\rho(w) = |\texttt{rle}(w)|$ i.e., is the number of pairs in w, or equivalently the number of equal-letter runs (also called *clusters*) in w.

Moreover we denote by $\rho(w)_{a_i}$ the number of pairs (w_j, l_j) in $\texttt{rle}(w)$ where $w_j = a_i$.

It is clear that for all $w \in \Sigma^*$ one has that $|alph(w)| \leq \rho(w) \leq |w|$. Notice also that if $w = uv$ then $\rho(w) \leq \rho(u) + \rho(v)$, that is, ρ is sub-additive.

In this section we introduce a parameter that gives a measure on how much the application of the BWT to a given word modifies the number of its clusters.

Definition 1. *The BWT-clustering ratio of a word w is*

$$\gamma(w) = \frac{\rho(\texttt{bwt}(w))}{\rho(w)}$$

Example 2. *Let us compute the BWT-clustering ratio for the words considered in Figure 1. If $w = aaabaabaabaabaab$ we have that $\rho(w) = 10$ and $\rho(\texttt{bwt}(w)) = \rho(bbbbbaaaaaaaaaaa) = 2$. So, $\gamma(w) = 1/5$.*

Let us consider $w = aaaabaabbababbbb$. In this case we have that $\rho(w) = 8$. Since $\texttt{bwt}(w) = baababababbababa$ then $\rho(\texttt{bwt}(w)) = 14$. So, $\gamma(w) = 7/4$.

Remark 1. *We note that if w is not a primitive word (i.e., $w = v^k$ for some $k > 1$) one can prove that $\rho(v^k) \leq k\rho(v)$. Moreover, in [9] it has been proved that if $\texttt{bwt}(v) = v_1 v_2 \cdots v_n$, where $v_i \in \Sigma$, then $\texttt{bwt}(v^k) = v_1^k v_2^k \cdots v_n^k$. So, $\rho(\texttt{bwt}(v^k)) = \rho(\texttt{bwt}(v))$. This implies that $\gamma(v^k) \geq \frac{1}{k}\gamma(v)$. In particular, it was proved (cf. [8]) that if v is a Lyndon word (different from a single letter), then $\gamma(v^k) = \frac{1}{k}\gamma(v)$.*

Remark 2. *We recall that if u and v are conjugate words, then $\texttt{bwt}(u) = \texttt{bwt}(v)$. On the other hand, one has that $|\rho(u) - \rho(v)| \leq 1$ and, within*

the conjugacy class, a power of a Lyndon word is one of the conjugates having least number of clusters. Since we are interested in evaluating how the number of clusters produced by BWT can grow compared to the number of clusters in the input necklace, we can consider words that are power of a Lyndon word as input of the parameter γ. Moreover, due to the property described in Remark 1 we can limit our attention to Lyndon words.

The following theorem, also reported in [8], shows that the number of clusters can at most be doubled by the BWT.

Theorem 1. *Given a Lyndon word w, we have that $0 < \gamma(w) \leq 2$.*

Proof. Let $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ with $a_1 < a_2 < \cdots < a_\sigma$ and let $\texttt{rle}(w) = (b_1, l_1), (b_2, l_2), \ldots, (b_k, l_k)$, where $b_1, b_2, \ldots b_k \in \Sigma$.

Recall that when computing $\texttt{bwt}(w)$, the column F of the matrix of sorted conjugates of w has the form $a_1^{|w|_{a_1}} a_2^{|w|_{a_2}} \cdots a_\sigma^{|w|_{a_\sigma}}$. It is then naturally defined a parsing of the column L according to the runs $(a_1, |w|_{a_1})(a_2, |w|_{a_2}) \cdots (a_\sigma, |w|_{a_\sigma})$ of F. We denote by u_{a_i} the factor in $L = \texttt{bwt}(w)$ associated to the run $(a_i, |w|_{a_i})$ of F, i.e. all the letters that in the original words precede an occurrence of the letter a_i. Then we can write $\texttt{bwt}(w) = u_{a_1} u_{a_2} \cdots u_{a_\sigma}$.

Consider any block u_{a_j}. In this block there are at most as many letters different from a_j as the number of different runs of a_j in w. In fact, in w, a_j is preceded by a letter different from a_j itself only in the beginning of each of its runs. So the greatest possible number of runs contained in u_{a_j} is achieved when all the letters different from a_j are spread in the block, never one next to another, producing on u_{a_j} a number of runs $\texttt{rle}(u_{a_j})$ at most equal to $2\rho(w)_{a_j}$. This happens for each block, then

$$\rho(\texttt{bwt}(w)) \leq \sum_{i=1}^{\sigma} \rho(u_{a_i}) \leq \sum_{i=1}^{\sigma} 2 \cdot \rho(w)_{a_i} = 2 \sum_{i=1}^{\sigma} \rho(w)_{a_i} = 2\ \rho(w).$$

□ □

In the following theorem we show that for any positive rational number r smaller than or equal to 2, it is possible to construct a binary word such that its BWT-clustering ratio is equal to r.

Theorem 2. *For any $r \in Q \cap (0, 2]$, there exists a Lyndon word $w \in \{a, b\}^*$ having $\gamma(w) = r$.*

Proof. Let p and q two coprime positive integers such that $r = \frac{p}{q}$. Let k be an integer such that $k \geq 2$.

Let us define $f_i = a^{2i-1}b^{2i-1}$ for $i = 2, 3, \ldots, k$ and $f_1 = abb$. Let h be an integer such that $h \geq 1$.

We can define a family of words

$$v_{h,k} = (f_k)^h f_{k-1} \cdots f_1 = (a^{2k-1}b^{2k-1})^h a^{2k-3}b^{2k-3} \cdots a^3b^3ab^2.$$

Since each f_i has two clusters, the first an a-cluster and the second a b-cluster, $\rho(v_{h,k}) = 2h + 2k - 2$.

We now compute $\rho(\mathtt{bwt}(v_{h,k}))$.

First of all we consider the case $h = 1$. In fact, $\mathtt{bwt}(v)$ can be factored in two parts: the first one corresponding to all the conjugates starting with a, and the second one corresponding to all the conjugates starting with b.

The first part starts with the only conjugate that has $a^{2k-1}b$ as prefix, then the one with $a^{2k-2}b$, then the two conjugates that start with $a^{2k-3}b$ (from the rightmost to the leftmost), and so on. From this we can see that the first part of $\mathtt{bwt}(v_{1,k})$ is $baba^3 \cdots ba^{2(k-1)-1}ba^{k-1}$ that has $2k$ clusters.

For the second part, we have exactly all the conjugates starting in the second part of each f_i. In particular, there are k conjugates starting with ba. All these conjugates are cyclicly preceded by b. Then we have all the conjugates starting with bba. The lexicographically smallest in this group is the one corresponding to the block f_1, then we have the conjugates corresponding to the other f_i from the leftmost to the rightmost. Such conjugates are lexicographically followed by the conjugates starting with $bbba$ that correspond to the blocks from f_k to f_2 and so on. This means that the second part of $\mathtt{bwt}(v_{1,k})$ is $b^kab^{2k-3}ab^{2k-5}a \cdots ba$ that has $2k$ clusters. It follows that

$$\mathtt{bwt}(v_{1,k}) = baba^3 \cdots ba^{2(k-1)-1}ba^{k-1}b^kab^{2k-3}ab^{2k-5}a \cdots ba,$$

that has $2k + 2k$ clusters. So $\rho(\mathtt{bwt}(v_{1,k})) = 4k$.

Finally, one can prove that for any $h \geq 1$, $\rho(\mathtt{bwt}(v_{h,k}) = \rho(\mathtt{bwt}(v_{1,k})$.

In fact, $\mathtt{bwt}(v_{h,k})$ is obtained by concatenating

$$b^ha^hba^{2h+1}ba^{2h+3} \cdots ba^{2h+2k-5}ba^{k-2+h}$$

and

$$b^{k-1+h}ab^{2k-5+2h}ab^{2k-7+2h}a \cdots b^{1+2h}ab^ha^h.$$

So, the thesis follows since

$$\gamma(v_{h,k}) = \frac{4k}{2h + 2k - 2} = \frac{2k}{h + k - 1} = \frac{p}{q}.$$

It is then sufficient to find suitable integer solutions to unknown h and k to the above equation. □ □

Example 3. *Let us consider the rational number* $6/5$ *(*> 1*). In this case a solution to the equation*

$$\frac{2k}{h+k-1} = \frac{6}{5}$$

is $k = 3$ *and* $h = 3$*. In fact one can verify that if* $w = (a^5b^5)^3a^3b^3abb$ *then* $\texttt{bwt}(w) = b^3a^3ba^7ba^4b^5ab^7ab^3a^3$*, so* $\rho(w) = 10$ *and* $\rho(\texttt{bwt}(w)) = 12$*.*

On the other hand if we consider the rational number $4/5$ *(*< 1*) a solution to*

$$\frac{2k}{h+k-1} = \frac{4}{5}$$

is $k = 2$ *and* $h = 4$*. One can verify that* $w = (a^3b^3)^4abb$ *and* $\texttt{bwt}(w) = b^4a^4ba^4b^5ab^4a^4$*, so* $\rho(w) = 10$ *and* $\rho(\texttt{bwt}(w)) = 8$*.*

Corollary 1. *For any rational number* $0 < r \leq 2$*, there are infinitely many words* w *with* $\gamma(w) = r$*.*

Proof. The solutions of the equation

$$\frac{2k}{h+k-1} = \frac{p}{q}$$

corresponds to the integer solutions to all of the following systems:

$$\begin{cases} 2k = lp \\ h+k-1 = lq \end{cases}$$

for any choice of l that gives integer solutions to h and k. In particular, if p is even, any integer value of l is allowed, if p is odd, only even values of l are allowed. □

4 Special Cases on Binary Alphabet

In this section we give some characterization and properties of families of words over two letters alphabets well known in combinatorics on words, according to their BWT-clustering ratio γ.

4.1 Clusters in Christoffel Words

In this subsection we take into account the BWT-clustering ratio of a class of words over a binary alphabet known in literature as Christoffel words (cf. [6, 2]). We start by giving the definition of a class of words strictly related to them, i.e. the Standard words. There exist many equivalent definitions of Standard words. Here we use the one that makes evident their relationships with the notion of characteristic Sturmian word.

Let $d_1, d_2, \ldots, d_n, \ldots$, $n \geq 1$ be a sequence of natural integers, with $d_1 \geq 0$ and $d_i > 0$ for $i = 2, \ldots, n, \ldots$. Consider the sequence of words $\{s_n\}_{n\geq 0}$ recursively defined by:

$$s_0 = b, \;\; s_1 = a, \quad \text{and } s_{n+1} = {s_n}^{d_n} s_{n-1} \quad \text{for } n \geq 1.$$

Each finite word s_n is called a *standard word*. It is univocally determined by the (finite) directive sequence $(d_1, d_2, \ldots, d_n)$. Such sequences are very important, since their limit, for $n \to \infty$ converges to infinite words called characteristic Sturmian words, well known in literature for its numerous and interesting combinatorial properties.

For any standard word w, the Lyndon word in its class is also called *Christoffel word*. We are now considering Christoffel words since, as usual, we take the Lyndon word for each class. For instance the word *aaabaabaabaabaab* considered in Figure 1(a) is a Christoffel word.

The following proposition gives a new characterization of Christoffel words in terms of the γ ratio.

Proposition 2. *A word w is a Christoffel word* $\iff \gamma(w) = \frac{1}{\min\{|w|_a, |w|_b\}}$.

Proof. Let w be a Christoffel word and suppose that $|w|_b = h$ and $|w|_a = k$ with $h < k$ (the other case has an analogous proof). Then no b in w appears next to another b, therefore w has $2h$ clusters, i.e. $\rho(w) = 2|w|_b$. On the other side in [9] it has been proved that any conjugate of a standard word (in particular any Christoffel word) has a totally clustered `bwt`; in particular $\texttt{bwt}(w) = b^h a^k$, i.e. $\rho(\texttt{bwt}(w)) = 2$. Therefore

$$\gamma(w) = \frac{2}{2h} = \frac{1}{|w|_b}.$$

Suppose now that $\gamma(w) = 1/|w|_b$, that is

$$\frac{\rho(\texttt{bwt}(w))}{\rho(w)} = \frac{1}{|w|_b}.$$

Then $\rho(\texttt{bwt}(w)) \cdot |w|_b = \rho(w) \leq 2|w|_b$, i.e. $\rho(\texttt{bwt}(w)) \leq 2$.

But on binary words $\rho(\texttt{bwt}(w)) \geq 2$, then $\rho(\texttt{bwt}(w)) = 2$ and this is true if and only if w is a Christoffel word (cf [9]). □ □

Remark 3. *For any $\epsilon > 0$ there exists a Christoffel word w such that $\gamma(w) < \epsilon$. In fact let us consider a Christoffel word where $|w| = 2n + 1$, $|w|_b = n$ and $|w|_a = n + 1$. By Proposition 2 $\gamma(w) = \frac{1}{n}$. For n sufficiently large, $1/n < \epsilon$.*

4.2 Clusters in Binary de Bruijn Words

In this section we consider another famous class of words called de Bruijn words. In particular here we consider de Bruijn words over a binary alphabet.

A de Bruijn word of order n on an alphabet Σ of size k is a cyclic word in which every word of length n on Σ occurs exactly once as a factor. By Remark 2, in the following when we refer to a de Bruijn word we mean the corresponding Lyndon word in its necklace. Such a word is denoted by $B(k, n)$ and has length k^n, which is also the number of distinct factors of length n on Σ. There are $\frac{(k!)^{k^{n-1}}}{k^n}$ many distinct de Bruijn words $B(k, n)$. In particular for two letters alphabets, all de Bruijn words $B(2, n)$ have length 2^n, and there are $\frac{2^{2^{n-1}}}{2^n}$ many distinct de Bruijn words $B(2, n)$.

One can verify that the word *aaaabaabbababbbb* considered in Figure 1(b) is a de Bruijn word of order 4 over the alphabet $\{a, b\}$, since every word in $\{a, b\}^4$ appears once in the corresponding cyclic word.

In the following proposition, also reported in [8], we find the number of runs of a de Bruijn word of order n on a binary alphabet. Note that this result can be inferred by using some combinatorial properties analyzed in [4].

Proposition 3. *Let $B(2, n)$ be any de Bruijn word of order n over a binary alphabet. Then $\rho(B(2, n)) = 2^{n-1}$.*

Proof. We first consider the runs of a's. First of all there is no run a^i with $i > n$ otherwise a^n would be a word of length n that appears more than once in $B(2, n)$. The run a^n is a particular word of length n, then, by definition, it appears exactly once as a factor in $B(2, n)$ (in particular as factor of ba^nb).

The words ba^{n-1} and $a^{n-1}b$ also appear once, but since they are factors of ba^nb, we have no runs of a's of length $n - 1$.

For any $1 \leq i \leq n - 2$ consider the runs of the form a^i. They appear as factors of all the words of the form ba^ibw where w is any word of length

$n-i-2$. Each of the words ba^ibw appear exactly once. There are 2^{n-i-2} of such words, therefore there are 2^{n-i-2} runs a^i. We have overall:

$$1+\sum_{i=1}^{n-2} 2^{n-i-2} = 1+\sum_{i=0}^{n-3} 2^i = 1+2^{n-2}-1 = 2^{n-2}$$

So there are 2^{n-2} runs of a's. For the same reason there are 2^{n-2} runs of b's, then overall $2 \cdot 2^{n-2} = 2^{n-1}$ runs. □ □

Remark 4. *As a byproduct of the theorem proved by Higgins in [5] (cf. also [10]) we have that if $B(n,k)$ is a de Bruijn word of order n, then $\mathtt{bwt}(B(k,n)) \in G^{k^{n-1}}$, where G is the set of all sequences of Σ of length $|\Sigma|$ obtained by permuting all the letters in Σ. In particular, if $k=2$, $\mathtt{bwt}(B(2,n)) \in \{ab,ba\}^{2^{n-1}}$.*

The following theorem is a consequence of Proposition 3 and of the above remark.

Theorem 3. *If w is a binary de Bruijn word then:*

$$1+\frac{4}{|w|} \leq \gamma(w) \leq 2-\frac{4}{|w|}.$$

Proof. Recall that any binary de Bruijn word of order n has length 2^n, with $n \geq 2$.

As remarked above, by Higgins's Theorem, $\mathtt{bwt}(w) \in \{ab,ba\}^{2^{n-1}}$. Moreover, one can note that ba must be the prefix and the suffix of $\mathtt{bwt}(B(2,n))$, since for any word its $\mathtt{bwt}$ cannot start with the smallest symbol and cannot end with the biggest symbol. Since, as proved in [8], $\rho(\mathtt{bwt}(B(2,n))) < |B(2,n)| = 2^n$, then $\mathtt{bwt}(w) \neq (ba)^{2^{n-1}}$ then both aa and bb must be factors of $\mathtt{bwt}(B(2,n))$. So, the upper bound follows because $\rho(\mathtt{bwt}(w)) \leq 2^n-2$. The lower bound on the number of runs is reached when $\mathtt{bwt}(w) = b(aabb)^{2^{n-2}-1}aba$. In this case this value is $2^{n-1}+2$. Then, the thesis follows. □

5 Experimental Results

It is commonly said that the application of BWT as a preprocessing to the application of a statistical compressor is useful since BWT tends to cluster together equal letters that appear in equal contexts, generating a so called "clustering effect". In this paper we highlight that this is not always the case, that is, there are words that are "un-clustered" by the BWT, that is,

the application of BWT generate on such words a greater number of shorter clusters.

The BWT-clustering ratio γ allows to classify words into *BWT-good words*, if $0 < \gamma(w) < 1$, and *BWT-bad words*, if $1 < \gamma(w) \leq 2$. The qualities *good* and *bad* reflects a good or bad behavior of BWT with respect to clustering, that is a good requirement for compression. For instance, since for any Christoffel word w, $\gamma(w) < 1$, then Christoffel words are BWT-good. On the other hand, any binary Bruijn word is BWT-bad.

Of course, the other special case is when this ratio is 1, i.e. the words, called *BWT-neutral*, where the BWT has no effect in terms of clustering. Among these words we can find fixed points (i.e. words w such that $\texttt{bwt}(w) = w$), that are studied in [7].

In this section, we show some experiments that highlight the distribution of the BWT-neutral, BWT-good and BWT-bad binary words when the length is fixed. In particular, table in Figure 2 shows such a distribution for all Lyndon words w of length 16 and 24.

length	number of words	$\gamma(w) = 1$	$\gamma(w) < 1$	$\gamma(w) > 1$	$\gamma(w) = 2$
16	4.080	1.160	1.247	1.673	142
24	698.870	156.652	237.636	304.582	4.362

Figure 2: Distribution of Lyndon words of length 16 and 24

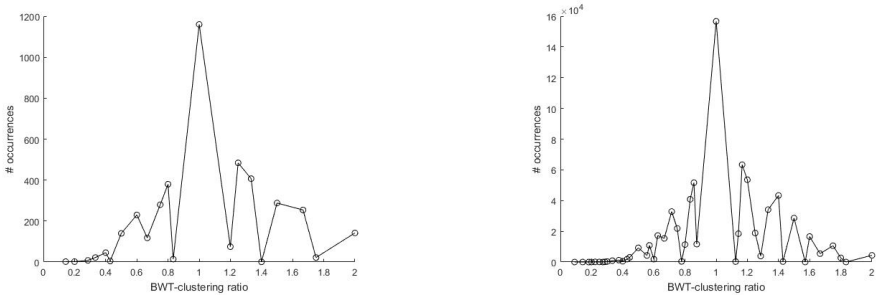

Figure 3: Lyndon words of length 16 and 24

On the other hand, table in Figure 4 shows such a distribution for all Lyndon words w of length $16, 20, 24$ and 28. with the same number of occurrences of letter a and letter b.

For completeness, the graphs in Figure 3 and Figure 5 show the number of Lyndon words of length 16 and 24 as a function of the BWT-clustering

length	number of words	$\gamma(w) = 1$	$\gamma(w) < 1$	$\gamma(w) > 1$	$\gamma(w) = 2$
16	800	224	239	337	26
20	9.225	2.183	3.042	4.000	129
24	112.632	23.866	38.884	49.882	666
28	1.432.613	288.485	504.505	639.623	3.556

Figure 4: Distribution of Lyndon words of length $16, 20, 24$ and 28. with the same number of letters a and b

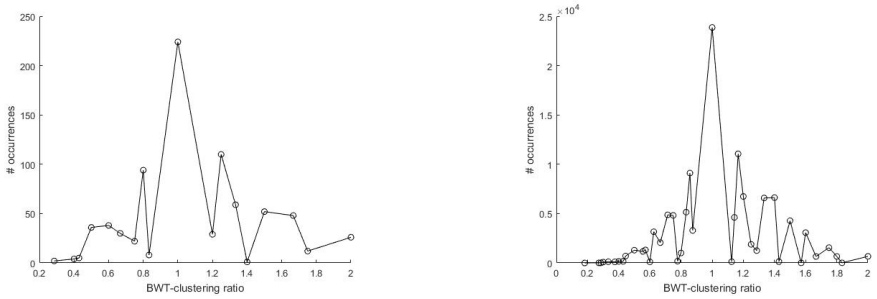

Figure 5: Lyndon words of length 16 and 24 with the same number of letters a and b

ratio. It is interesting to point out that the graphs show that the trend does not change substantially when words having the same number of a and b are considered. A possible further work could be to develop an analytic study of this behavior.

Acknowledgements

Thanks to Published source. S. Mantaci, G. Rosone and M. Sciortino are partially supported by the project MIUR-SIR CMACBioSeq ("Combinatorial methods for analysis and compression of biological sequences") grant n. RBSI146R5L and by the Gruppo Nazionale per il Calcolo Scientifico (GNCS-INDAM).

References

[1] D. Adjeroh, T. Bell, and A. Mukherjee. *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching.* Springer, 2008.

[2] J. Berstel, A. Lauve, C. Reutenauer, and F.V. Saliola. *Combinatorics on Words: Christoffel Words and Repetitions in Words*, volume 27 of *CRM monograph series*. American Mathematical Soc., 2008.

[3] M. Burrows and D. J. Wheeler. A block sorting data compression algorithm. Technical report, DIGITAL System Research Center, 1994.

[4] H. Fredricksen. A survey of full length nonlinear shift register cycle algorithms. *SIAM Review*, 24(2):195–221, 1982.

[5] P. M. Higgins. Burrows-Wheeler transformations and de Bruijn words. *Theoretical Computer Science*, 457:128–136, 2012.

[6] M. Lothaire. *Applied Combinatorics on Words (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press, New York, NY, USA, 2005.

[7] S. Mantaci, A. Restivo, G. Rosone, F. Russo, and M. Sciortino. On Fixed Points of the Burrows-Wheeler Transform. *Fundamenta Informaticae.* to appear.

[8] S. Mantaci, A. Restivo, G. Rosone, M. Sciortino, and L. Versari. Measuring the clustering effect of BWT via RLE. *Theoretical Computer Science.* to appear.

[9] S. Mantaci, A. Restivo, and M. Sciortino. Burrows-Wheeler transform and Sturmian words. *Information Processing Letters*, 86:241–246, 2003.

[10] D. Perrin and A. Restivo. Words. In Miklos Bona, editor, *Handbook of Enumerative Combinatorics*. CRC Press, 2015.