Burrows-Wheeler transform of words defined by morphisms

S. Brlek[1] , A. Frosini[2], I. Mancini[3⋆], E. Pergola[2], S. Rinaldi[3]

[1] Laboratoire de combinatoire et d'informatique mathématique, UQAM, Canada
brlek.srecko@uqam.ca
[2] Dipartimento di Matematica e Informatica "U. Dini", Università di Firenze, Italy
andrea.frosini@unifi.it, elisa.pergola@unifi.it
[3] Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche, Università di Siena, Italy
ilaria.mancini@student.unisi.it, rinaldi@unisi.it

Abstract. The *Burrows-Wheeler transform* (BWT) is a popular method used for text compression. It was proved that BWT has optimal performance on standard words, i.e. the building blocks of Sturmian words. In this paper, we study the application of BWT on more general morphic words: the Thue-Morse word and to generalizations of the Fibonacci word to alphabets with more than two letters; then, we study morphisms obtained as composition of the Thue-Morse morphism with a Sturmian one. In all these cases, the BWT efficiently clusters the iterates of the morphisms generating prefixes of these infinite words, for which we determine the compression clustering ratio.
Keywords: Burrows-Wheeler transform, morphisms, Thue-Morse word, Generalized Fibonacci words.

1 Introduction

The *Burrows-Wheeler transform* (BWT) is a powerful technique used at the preprocessing stage in text compression algorithms [1,4]. Actually, it produces a permutation of the characters of an input word and tends to group characters in runs, so that the output word is easier to compress because it has a lower number of runs (clusters). More precisely, the Burrows-Wheeler transform of a word w of length n is built by lexicographically sorting all its n conjugates and extracting the last character of each conjugate. The BWT is interesting in many aspects, and in particular because it is linked to a remarkable bijection due to Gessel and Reutenauer [8] on permutations, of which it is a special case [5].

In general it is difficult to determine a priori the compression ratio of the BWT on a given word, so recent studies use combinatorics on words tools in order to give a measure of efficiency of the BWT applied to some classes of words [7,10,15,19]. In order to investigate the clustering effect of the BWT from a combinatorial viewpoint, it is interesting to consider the structural properties

⋆ Corresponding author

of the words for which the BWT produces the maximal or the minimal compression ratio; a perfect clustering produced by the BWT corresponds to optimal performances of the run-length encoding.

The clustering effect of BWT on balanced words is studied in [18]. In particular, standard words - building blocks of infinite Sturmian words - have a BWT of the form $b^h\, a^j$ [14]. On a k-letter alphabet, words having a BWT with minimal number of clusters have been characterized in [17] in the case of balanced words, and in [6] in the general case.

The study on Sturmian words suggests that more general morphic words are expected to show a similar behaviour. In this paper, we support this thesis by studying the BWT performance on the building blocks of the Thue-Morse word and k-bonacci words, and, we determine the compression clustering ratio.

Finally, we study morphisms obtained by composition of the Thue-Morse morphism with a Sturmian one. In all these cases the BWT proves to efficiently cluster the building blocks of the considered infinite words.

2 Burrows-Wheeler transform

We assume the reader familiar to the basic terminology on words defined on a finite alphabet (ref A [11]). The *Burrows-Wheeler transform* is defined as a map $\mathtt{bwt}(w) : A^* \to A^* \times \mathbb{N}$ such that $\mathtt{bwt}(w) = (L, I)$ where

- L is the last column of a matrix M whose lines are all the conjugates of w sorted lexicographically;
- I is the index of the line of M containing the original word w.

Example 1. The construction of BWT for $w =$ *filosofia* is

$$\begin{array}{rcccccccc} & F & & & & & & & L \\ & \downarrow & & & & & & & \downarrow \\ 1 & a & f & i & l & o & s & o & f & i \\ 2 & f & i & a & f & i & l & o & s & o \\ I \to 3 & f & i & l & o & s & o & f & i & a \\ 4 & i & a & f & i & l & o & s & o & f \\ 5 & i & l & o & s & o & f & i & a & f \\ 6 & l & o & s & o & f & i & a & f & i \\ 7 & o & f & i & a & f & i & l & o & s \\ 8 & o & s & o & f & i & a & f & i & l \\ 9 & s & o & f & i & a & f & i & l & o \end{array}$$

so $\mathtt{bwt}(w) = (\mathit{ioaffislo}, 3)$. In what follows we shall denote $\mathtt{bwt}_1(w)$, the first component, that is the column L.

Note that the first column F of the matrix M is the sequence of lexicographically sorted letters of w (see Example 1). The Burrows-Wheeler transform is reversible by using the properties described in the following proposition [4].

Proposition 1. *Let $w \in A^*$ be a word such that $\mathtt{bwt}(w) = (L, I)$ and let F be as above. The following properties hold:*

1. *$\forall i, 1 \leq i \leq n, i \neq I$, the letter $F[i]$ follows the letter $L[i]$ in w;*
2. *$\forall \alpha \in A$, the i-th occurrence of α in F matches the i-th occurrence of α in L;*
3. *the first letter of w is $F[I]$.*

According to property 2 of Proposition 1, there is a permutation τ_w giving the correspondence between the positions of letters in F and L. Hence, starting from the position I, the word w is obtained as follows:

$$w[i] = F[\tau_w^{i-1}(I)],$$

where $\tau_w^0(x) = x$ and $\tau_w^i(x) = \tau_w(\tau_w^{i-1}(x))$, with $1 \leq i \leq n$. For the word of Example 1 , we have $\tau_w = (345168297)$.

Observation 1 For any two words u, v such that $|u| = |v|$ and any letter α, we have $\alpha u < \alpha v$ if and only if $u\alpha < v\alpha$ (if and only if $u < v$). Thus, given a word w, for all indices i, j if $i < j$ and $F[i] = F[j]$, then $\tau_w(i) < \tau_w(j)$.

The conjugation relation between words is denoted by $u \equiv v$. It is easy to see that in the BWT the column L is stable by conjugation.

Proposition 2. *$u \equiv v$ if and only if $\mathtt{bwt}_1(u) = \mathtt{bwt}_1(v)$.*

We denote by $\rho(u)$ the number of equal-letter runs or *clusters* of a word $u \in A^*$.

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

$$\gamma(w) = \frac{\rho(\mathtt{bwt}_1(w))}{\rho(w)}.$$

For $w =$ *filosofia* we have $\rho(w) = 9$, $\rho(\mathtt{bwt}_1(w)) = \rho(ioaffislo) = 8$, $\gamma(w) = 8/9$. More details about bounds for the compression ratio can be found in [12,13].

Burrows-Wheeler transform of Sturmian words. The family of Sturmian words has been extensively studied in several contexts, and recently in particular under the BWT transformation. For the unfamiliar reader we simply recall that Sturmian words are infinite words that approximate lines of irrational slopes on a square grid, so that they are conveniently encoded on the two-letter alphabet $A = \{a, b\}$. In [14], Mantaci et al. established that, given a word $w \in A^*$ then $\mathtt{bwt}_1(w) = b^k\, a^h$ if and only if u is the power of a conjugate of a standard word. In [20], Simpson and Puglisi provided an alternative proof of the previous statement and a characterization of words on the alphabet $\{a, b, c\}$ whose transforms have the form $c^i\, b^j\, a^k$.

Definition 2. A morphism f on A is *Sturmian* if $f(x)$ is a Sturmian word whenever x is a Sturmian word.

In particular, it is a well-known fact that each Sturmian morphism is obtained by composition from the following three morphisms

$$E : \begin{matrix} a \mapsto b \\ b \mapsto a \end{matrix} \qquad \varphi : \begin{matrix} a \mapsto ab \\ b \mapsto a \end{matrix} \qquad \widetilde{\varphi} : \begin{matrix} a \mapsto ba \\ b \mapsto a. \end{matrix}$$

3 Burrows-Wheeler transform of non-Sturmian words

The study on Sturmian words suggests that the compression ratio of words obtained by morphism iteration is likely to be high. Indeed, as these words are defined recursively, it is expected that their BWT can be expressed recursively as well, resulting in a lower number of clusters. In what follows we consider (right) infinite fixed points of morphisms on finite alphabets.

An alphabet A comes equipped with an order $\preceq$ which extends to A^* in the usual lexicographic way. Let $\Phi : A^* \to A^*$ be an order-preserving morphism on A, that is $\Phi(\alpha) \preceq \Phi(\beta)$ whenever $\alpha \preceq \beta$ for all $\alpha, \beta \in A$. Then we have

Lemma 1. *If Φ is an order-preseving uniform morphism then*

$$\forall x, y \in A^*,\ x \preceq y \implies \Phi(x) \preceq \Phi(y).$$

3.1 Burrows-Wheeler transform of standard Thue-Morse words

The Thue-Morse word m is a recurrent cube-free infinite word having many combinatorial properties, and appearing in many contexts (see [3] for a primer).

The word m is the limit of the sequence $(u_n)_{n\geq 0}$, with $u_n = \mu^n(a)$, obtained by iterating the morphism μ, defined by $\mu(a) = ab$, $\mu(b) = ba$, which is order preserving. So the first letters are the following

$$m = abbabaabbaababbabaab\cdots .$$

The word $u_n = \mu^n(a)$ is the n-standard Thue-Morse word. The BWT applied to the first n-standard Thue-Morse words shows some interesting regularities:

$n = 0$ $\mathtt{bwt}_1(a) = a$
$n = 1$ $\mathtt{bwt}_1(ab) = ba$
$n = 2$ $\mathtt{bwt}_1(abba) = baba$
$n = 3$ $\mathtt{bwt}_1(abbabaab) = b^2ababa^2$
$n = 4$ $\mathtt{bwt}_1(abbabaabbaababba) = b^4a^2babab^2a^4$
$n = 5$ $\mathtt{bwt}_1(abbabaabbaababbabaababbaabbabaab) = b^8a^4b^2ababa^2b^4a^8$

To study the performance of the BWT on standard Thue-Morse words we need the following properties.

Proposition 3 (Thue, 1912). *For $n \geq 0$ the n-standard word u_n satisfies:*

(i) *u_n does not contain the factors aaa and bbb for any $n \geq 0$.*
(ii) *$|u_n|_a = |u_n|_b = 2^{n-1}$.*

Theorem 2. *For $n \geq 2$, it holds that:*

$$\mathtt{bwt}_1(u_n) = \begin{cases} b^{2^{n-2}}\, a^{2^{n-3}} \cdots b\,(ab)\,a \cdots b^{2^{n-3}}\, a^{2^{n-2}} & \text{if } n \text{ is even} \\ b^{2^{n-2}}\, a^{2^{n-3}} \cdots a\,(ba)\,b \cdots b^{2^{n-3}}\, a^{2^{n-2}} & \text{if } n \text{ is odd.} \end{cases}$$

Proof. The proof is obtained by induction on n. The statement holds for u_2 since $\mathtt{bwt}_1(u_2) = baba$ and $2^0 = 1$. Suppose the thesis holds for u_{n-1} and that n in even. So $\mathtt{bwt}_1(u_{n-1}) = b^{2^{n-3}}\, a^{2^{n-4}} \cdots a\,(ba)\, b \cdots b^{2^{n-4}}\, a^{2^{n-3}}$. The conjugates of u_{n-1} are words of the form

$$aaX_1b,\ abX_2b,\ abX_3a,\ baX_4a,\ baX_5b,\ bbX_6a \tag{1}$$

and, by inductive hypothesis, the final letters of the lexicographically ordered conjugates form the above $\mathtt{bwt}_1(u_{n-1})$. Applying μ to (1) we get some conjugates of u_n:

$$ababμ(X_1)ba,\ abbaμ(X_2)ba,\ abbaμ(X_3)ab,\ baabμ(X_4)ab,\ baabμ(X_5)ba,\ babaμ(X_6)ab.$$

Let $\mathcal{C}$ be the set of these conjugates obtained by μ. Lemma 1 ensures that the order of the conjugates in $\mathcal{C}$ is preserved by μ, and moreover, μ swaps the final letter of each conjugate from a to b and vice versa. The set of conjugates of u_n also includes

$$aababμ(X_1)b \prec aabbaμ(X_2)b \prec abaabμ(X_5)b$$

and

$$babbaμ(X_3)a \prec bbaabμ(X_4)a \prec bbabaμ(X_6)a.$$

So all conjugates ending with b form a set $\mathcal{P}$ and precede those in $\mathcal{C}$, and all conjugates ending with a form a set $\mathcal{F}$ and follow those in $\mathcal{C}$ in lexicographic order. It follows that the central part of $\mathtt{bwt}_1(u_n)$ is $a^{2^{n-3}}\, b^{2^{n-4}} \cdots b\,(ab)\, a \cdots a^{2^{n-4}}\, b^{2^{n-3}}$. Finally, by Proposition 3(ii) $|\mathcal{P}| = 2^{n-2} = |\,\mathtt{bwt}_1(u_{n-1})|_a$, the number of occurrences of a in $\mathtt{bwt}_1(u_{n-1})$. Similarly, $|\mathcal{F}| = 2^{n-2} = |\,\mathtt{bwt}_1(u_{n-1})|_b$. The thesis follows:

$$\mathtt{bwt}_1(u_n) = b^{2^{n-2}}\, a^{2^{n-3}}\, b^{2^{n-4}} \cdots b\,(ab)\, a \cdots a^{2^{n-4}}\, b^{2^{n-3}}\, a^{2^{n-2}}.$$

The case n odd is similar. □

Corollary 1. *For any $n > 0$ it holds that $I_{u_n} = 2^{n-1}$, i.e. the word u_n is the last word (in the lexicographical order) of all its conjugates starting with a.*

Proof. The proof is obtained by induction on n. For $n = 1$, $u_1 = ab$ so we have that $I_{u_1} = 1$. Suppose $I_{u_{n-1}} = 2^{n-2}$. By Lemma 1, $u_n = \mu(u_{n-1})$ keeps the same position with respect to the images of the conjugates of u_{n-1}. Moreover there are 2^{n-2} conjugates of u_n that precede u_n. So the index of u_n in the lexicographic order shifts to 2^{n-1}. □

Denoting $\widetilde{w}$ the reversal of w, and $\overline{w} = E(w)$, observe that $\mathtt{bwt}_1(u_n)$ is also a pseudo-palindrome, that is

Corollary 2. *For any $n > 0$, we have $\widetilde{\mathtt{bwt}(u_n)} = \overline{\mathtt{bwt}(u_n)}$.*

Example 2. The conjugates of u_2 and u_3 are:

$$I_{u_2} = \begin{matrix} 1\ a\ a\ b\ b \\ 2\ a\ b\ b\ a \\ 3\ b\ a\ a\ b \\ 4\ b\ b\ a\ a \end{matrix} \xrightarrow{\mu} I_{u_3} = \begin{matrix} 1\ a\ a\ b\ a\ b\ b\ a\ b \\ 2\ a\ b\ a\ a\ b\ a\ b\ b \\ 3\ a\ b\ a\ b\ b\ a\ b\ a \\ 4\ a\ b\ b\ a\ b\ a\ a\ b \\ 5\ b\ a\ a\ b\ a\ b\ b\ a \\ 6\ b\ a\ b\ a\ a\ b\ a\ b \\ 7\ b\ a\ b\ b\ a\ b\ a\ a \\ 8\ b\ b\ a\ b\ a\ a\ b\ a \end{matrix} \begin{matrix} \left.\vphantom{x}\right\} \mathcal{P} \\ \left.\vphantom{x}\right\} \\ \left.\vphantom{x}\right\} \\ \left.\vphantom{x}\right\} \mathcal{C} \\ \left.\vphantom{x}\right\} \\ \left.\vphantom{x}\right\} \\ \left.\vphantom{x}\right\} \mathcal{F} \\ \left.\vphantom{x}\right\} \end{matrix}$$

An easy consequence of Theorem 2 shows that the number of clusters satisfies $\rho(\mathtt{bwt}_1(u_n)) = 2\,n$ and $\rho(u_n) = f_n$, where $f_0 = 1$, $f_1 = 2$, and:

$$f_n = \begin{cases} 2\,f_{n-1} & \text{if } n \text{ is odd} \\ 2\,f_{n-1} - 1 & \text{if } n \text{ is even.} \end{cases}$$

So $f_n \sim 2^n$, and therefore, for n large enough, the *clustering ratio* tends to zero:

$$\gamma(u_n) = \frac{\rho(\mathtt{bwt}_1(u_n))}{\rho(u_n)} \ll 1,$$

confirming the fact that the BWT is efficient on standard Thue-Morse words.

3.2 Generalizations of the Fibonacci word

We already know that the BWT of the standard Fibonacci words (being Sturmian) have the minimum number of clusters. We show now that it does not hold for generalizations to a three-letters alphabet. Actually, in view of the characterization of words such that $\mathtt{bwt}(w) = c^i\,b^j\,a^k$ established in [20], the BWT of the n-standard Tribonacci word have more than three clusters, for n sufficiently large. Consider the morphism $\tau : A^* \to A^*$defined by

$$\tau(a) = ab,\ \tau(b) = ac,\ \tau(c) = a.$$

The *Tribonacci word* is the limit of the sequence $(t_n)_{n>0}$, where $t_n = \tau^n(a)$ is the *n-standard Tribonacci word*. Some of its properties can be found in [2,21]. The sequence of Tribonacci numbers is defined as $(T_n)_{n\geq 0}$ with

$$T_0 = 1,\ T_1 = 1,\ T_2 = 2,\ T_{n+3} = T_n + T_{n+1} + T_{n+2}.$$

The first few terms are: $1, 1, 2, 4, 7, 13, 24, 44, 81, 149, \ldots$ (A000073 in [16]). The change of the initial conditions to $S_0 = 1$, $S_1 = 2$ and $S_2 = 3$ gives the integer sequence $S_n = T_n + T_{n-1}$.

Let us compute the BWT of the first standard Tribonacci words:

$n = 1\ \mathtt{bwt}_1(a) = a$
$n = 2\ \mathtt{bwt}_1(ab) = ba$
$n = 3\ \mathtt{bwt}_1(abac) = cba^2$
$n = 4\ \mathtt{bwt}_1(abacaba) = bcaba^3$
$n = 5\ \mathtt{bwt}_1(abacabaabacab) = bc^2bab^2a^6$
$n = 6\ \mathtt{bwt}_1(abacabaabacababacabaabac) = b^2c^4ba^2b^4a^{11}$
$n = 7\ \mathtt{bwt}_1(abacabaabacababacabaabacabacabaabacababacaba) = b^4c^6bcaba^3b^7a^{20}$

The morphism τ is not order-preserving as $\tau(c) \prec \tau(a) \prec \tau(b)$ neither uniform, so that we cannot use Lemma 1 in order to compute the BWT for Tribonacci words. Instead, let us consider the morphism

$$\tau^3 : a \mapsto abacaba, b \mapsto abacab, c \mapsto abac,$$

which is order-reversing since $\tau^3(c) \prec \tau^3(b) \prec \tau^3(a)$. However, we can use the following result

Lemma 2. *Let $\alpha \prec \beta \in A$. Then $\tau^3(\alpha s) \preceq \tau^3(\beta t)$ for all $s, t \in A^+$.*

Proof. Since s and t are nonempty, $\tau^3(s)$ and $\tau^3(t)$ have $abac$ for prefix. It follows that $\tau^3(as) = abacaba \cdot abac \cdots \preceq abacab \cdot abac \cdots = \tau^3(bt)$ and so $\tau^3(as) \preceq \tau^3(bt)$. The cases $(\alpha, \beta) = (a, c)$ and $(\alpha, \beta) = (b, c)$ are similar. □

Theorem 3. *For every standard Tribonacci word t_n, with $n > 0$, we have*

$$\mathtt{bwt}_1(t_{n+3}) = b^{T_{n-1}}\, c^{S_{n-1}}\, \mathtt{bwt}_1(t_n)\, b^{T_n}\, a^{S_{n+1}}.$$

Proof. The proof follows a scheme similar to the one of Theorem 2 by applying Lemma 2 and the fact that the word t_n does not contain the factors bb, cc, bc, cb and so neither aaa, aac, caa, cac. □

The number of clusters of $\mathtt{bwt}_1(t_n)$ satisfies the recurrence

$$\rho(\mathtt{bwt}_1(t_{n+3})) = \rho(\mathtt{bwt}_1(t_n)) + 4, \text{ for } n > 3.$$

The first terms are $(h_n)_{n>0} = 1, 2, 3, 5, 6, 6, 9, 10, 10, 13, \ldots$, and $h_n \sim n$.

Lemma 3. *The number of clusters in the n-standard Tribonacci word is $\rho(t_n) = f_n$, where $f_1 = 1$, and:*

$$f_n = \begin{cases} 2\,S_{n-2} + 1 & \text{if } n = 3\,m + 1 \text{ for some } m \in \mathbb{N} \\ 2\,S_{n-2} & \text{otherwise.} \end{cases}$$

Proof. By definition of t_n and τ^3 we have that $\rho(t_n) = |t_n| - |t_n|_{aa}$ and every factor aa in t_n derives from an a in t_{n-3}, except the case where a is the final letter of t_{n-3}, which happens if $n = 3\,m + 1$. So if $n \neq 3\,m + 1$ we have that

$$\rho(t_n) = |t_n| - |t_{n-3}|_a = T_n - T_{n-4} = T_{n-1} + T_{n-2} + T_{n-3} - T_{n-4} = 2\,S_{n-2}.$$

Otherwise, if $n = 3\,m + 1$ we have one additional cluster:

$$\rho(t_n) = |t_n| - |t_{n-3}|_a + 1 = 2\,S_{n-2} + 1. \quad \square$$

Since $f_n \sim 2^n$, we have that, for n large enough, the *clustering ratio* tends to 0:

$$\gamma(t_n) = \frac{\rho(\mathtt{bwt}_1(t_n))}{\rho(t_n)} \ll 1,$$

confirming that the BWT is remarkably efficient on standard Tribonacci words.

Further generalizations. The results holding for Tribonacci words can be naturally extended to the generalization of Fibonacci words to an alphabet of cardinality $k \in \mathbb{N}$. So, with $A_k = \{a_1, \ldots, a_k\}$, let φ_k be defined as

$$\varphi_k(a_1) = a_1a_2,\ \varphi_k(a_2) = a_1a_3,\ \ldots\ \varphi_k(a_{k-1}) = a_1a_k,\ \varphi_k(a_k) = a_1.$$

The *k-bonacci word* is the limit of the sequence $(g_{k,n})_{n>0}$, where $g_{k,n} = \varphi_k^n(a)$ is the *n-standard k-bonacci word* and the length of $g_{k,n}$ is the n-th k-bonacci number G^1_n:

$$G^1_0 = 1, \ldots, G^1_{k-1} = \sum_{i=0}^{k-2} G^1_i \text{ and } G^1_{n+k} = \sum_{i=n}^{n+k-1} G^1_i.$$

From the sequence $(G^1_n)_{n\geq 0}$ we can define $k-2$ other sequences:

$$\begin{aligned} G^2_n &= G^1_n + G^1_{n-1} \\ G^3_n &= G^1_n + G^1_{n-1} + G^1_{n-2} \\ &\vdots \\ G^{k-1}_n &= G^1_n + \ldots + G^1_{n-(k-2)} = \sum_{i=n-k+2}^{n} G^1_i\,. \end{aligned}$$

Theorem 4. *For every n-standard k-bonacci word $g_{k,n}$, it holds that*

$$\mathtt{bwt}_1(g_{k,n+k}) \;=\; a_2^{G^1_{n-1}}\, a_3^{G^2_{n-1}} \ldots a_k^{G^{k-1}_{n-1}}\, \mathtt{bwt}_1(g_{k,n})\, a_{k-1}^{G^1_n} \ldots a_2^{G^{k-2}_{n+k-3}}\, a_1^{G^{k-1}_{n+k-2}}.$$

The proof is a simple generalization of the one of Theorem 3 to an alphabet with k letters. Finally, the number of clusters of $\mathtt{bwt}_1(g_{k,n})$ is $\rho(\mathtt{bwt}_1(g_{k,n+k})) = \rho(\mathtt{bwt}_1(g_{k,n})) + 2\,k - 2$ for each $n > k$. On the other hand the number of clusters in the standard k-bonacci word is $\rho(g_{k,n}) = f_n$, where $f_1 = 1$, and:

$$f_n = \begin{cases} 2\,G^{k-1}_{n-2} + 1 & \text{if } n = k\,m + 1 \text{ for some } m \in \mathbb{N} \\ 2\,G^{k-1}_{n-2} & \text{otherwise.} \end{cases}$$

Since $f_n \sim 2^n$, againt, for n large enough, the *clustering ratio* tends to 0:

$$\gamma(g_{k,n}) = \frac{\rho(\mathtt{bwt}_1(g_{k,n}))}{\rho(g_{k,n})} \ll 1,$$

and so, the BWT reduces the number of clusters in the general case as well.

Example 3. For $k = 4$, we have the *Tetranacci word* and the BWT of the standard Tetranacci words q_n, $n > 0$, has the following form:

$$\mathtt{bwt}_1(q_{n+4}) \;=\; b^{Q_{n-1}}\, c^{R_{n-1}}\, d^{V_{n-1}}\, \mathtt{bwt}_1(q_n)\, c^{Q_n}\, b^{R_{n+1}}\, a^{V_{n+2}}$$

where $(Q_n)_{n\geq 0}$ is the sequence of Tetranacci numbers (sequence A000078 in [16]) and $(R_n)_{n\geq 0}$, $(V_n)_{n\geq 0}$ are such that $R_n = Q_n + Q_{n-1}$, $V_n = Q_n + Q_{n-1} + Q_{n-2}$.

3.3 Composition of Thue-Morse with Sturmian morphisms

A final result concerns the inspection of the BWT behaviour when applied to words obtained by composition of the Thue-Morse and Sturmian morphisms. We first consider the two different compositions of μ and φ.

The composition $\mu \circ \varphi$. It is defined on $A = \{a, b\}$ by

$$\mu \circ \varphi : a \mapsto abba; b \mapsto ab.$$

Let us consider the standard words $p_n = (\mu \circ \varphi)^n(a)$, whose first terms are: a, $abba$, $abbaababababba$, ... In general, we have the recurrence formula:

Lemma 4. *For every $n > 1$, it holds that*

$$p_n = p_{n-1}\ p_{n-2} \cdots p_1\ p_0\ b\ p_{n-2}\ p_{n-3} \cdots p_1\ p_0\ b\ p_{n-1}.$$

Proof. By induction on n. If $n = 2$ we have $p_2 = abbaababababba = p_1\, p_0\, b\, p_0\, b\, p_1$. Suppose that the statement holds for $n - 1$, i.e.:

$$p_{n-1} = p_{n-2}\ p_{n-3} \cdots p_1\ p_0\ b\ p_{n-3}\ p_{n-4} \cdots p_1\ p_0\ b\ p_{n-2}.$$

Then, by applying $\mu \circ \varphi$, the thesis follows, because $(\mu \circ \varphi)(b) = p_0\, b$. □

Corollary 3. *For every word p_n with $n > 1$, we have that $|p_n| = 3\,|p_{n-1}|$.*

The first terms of the sequence $(L_n)_{n\geq 0}$ of the lengths of the words p_n are: $1, 4, 12, 36, 108, 324, \ldots$ (sequence A003946 in [16]). The BWT applied to the first few terms of $(p_n)_{n\geq 0}$ yields:

$$\begin{aligned}
B_0 &= a \\
B_1 &= baba \\
B_2 &= b^2abab^3a^4 \\
B_3 &= b^6a^4b^3abab^8a^{12} \\
B_4 &= b^{18}a^{12}b^8abab^3a^4b^{24}a^{36} \\
B_5 &= b^{54}a^{36}b^{24}a^4b^3abab^8a^{12}b^{72}a^{108}
\end{aligned}$$

The Thue-Morse morphism μ being order-preserving, while the Fibonacci morphism φ is order reversing we immediately have

Lemma 5. *Let $\alpha \prec \beta \in A$. Then $(\mu \circ \varphi)(\alpha s) \succeq (\mu \circ \varphi)(\beta t)$ for all $s, t \in A^+$.*

Theorem 5. *For every word p_n, and $n > 0$, we have*

$$B_{n+1} = b^{\frac{L_n}{2}}\, \widetilde{B}_n\, b^{\frac{L_n}{2}}\, a^{L_n}.$$

Proof. The proof is similar to the proof of Theorem 2, and relies on: the application of Lemma 5, the fact that $|p_n|_a = |p_n|_b = L_n/2$ for $n > 0$, and the fact that p_1 has the same number of a's and b's, a property preserved by the morphism $\mu \circ \varphi$. □

For $n > 0$ the number of clusters of the BWT of the n-standard word is $\rho(B_{n+1}) = \rho(B_n) + 2$, as every reversal of B_n ends with the letter b. So the first terms of the sequence $(h_n)_{n\geq 0}$ are $1, 4, 6, 8, 10, 12, 14, \ldots$ and $h_n \sim n$.

Lemma 6. *The number of clusters in p_n for each $n \geq 0$ is $\rho(p_n) = f_n$ where $f_0 = 1$ and $f_n = L_n - L_{n-1} + 1$.*

Proof. By definition of $\mu \circ \varphi$, we have that in each p_n there are at most two consecutive a or b. It follows that

$$\rho(p_n) = |p_n| - |p_n|_{aa} - |p_n|_{bb}.$$

Every factor aa in p_n derives from an a in p_{n-1} except if a is the last letter, and every factor bb derives from an a. We have seen that $|p_{n-1}|_a = L_{n-1}/2$, so

$$\rho(p_n) = L_n - \left(\frac{L_{n-1}}{2} - 1\right) - \frac{L_{n-1}}{2}$$

and the thesis follows. □

The first terms of the sequence $(f_n)_{n\geq 0}$ are $1, 2, 9, 25, 73, 217, \ldots$ and $f_n \sim 3^n$. It follows that, for n large enough, the *clustering ratio* tends to zero:

$$\gamma(p_n) = \frac{\rho(B_n)}{\rho(p_n)} \ll 1$$

which states that the BWT is very effective for the n-standard words p_n defined by the morphism $\mu \circ \varphi$.

Observe that if we replace φ with $\widetilde{\varphi}$, and we define $p'_n = (\mu \circ \widetilde{\varphi})^n(a)$, we obtain words that are conjugates of p_n for each $n \geq 0$ and so, by Proposition 2, they have the same BWT.

The composition $\varphi \circ \mu$. Let us consider now the composition:

$$\varphi \circ \mu : \begin{array}{l} a \mapsto aba \\ b \mapsto aab \end{array}$$

Let q_n denote the n-standard words associated with $\varphi \circ \mu$, i.e. $q_n = (\varphi \circ \mu)^n(a)$, $n \geq 0$. The first terms are: a, aba, $abaababa$, ... In general we have

Lemma 7. *For each $n > 1$ it holds that*

$$q_n = q_{n-1}\ q_{n-2}\ q_{n-2}\ q_{n-3}\ q_{n-3} \cdots q_1\ q_1\ q_0\ q_0\ b\ q_{n-1}.$$

The proof is similar to the one of Lemma 4.

As a consequence of the previous lemma, we have that $|q_n| = 3^n$ and, by definition of $\varphi \circ \mu$, $|q_n|_a = 2\,|q_{n-1}| = 2 \cdot 3^{n-1}$ and $|q_n|_b = |q_{n-1}| = 3^{n-1}$. By applying the BWT to q_n, we get

$$B_{n+1} = b^{2\cdot 3^{n-1}}\, \widetilde{B}_n\, a^{4\cdot 3^{n-1}}$$

where $4 \cdot 3^{n-1}$ is obtained from $3^n + 3^{n-1}$.

The number of clusters of the BWT for each $n > 0$ is $\rho(B_{n+1}) = \rho(B_n) + 2$. So the first terms of the sequence $(h_n)_{n\geq 0}$ are $1, 4, 6, 8, 10, 12, 14, \ldots$ and $h_n \sim n$.

Lemma 8. *The number of clusters in q_n for each $n \geq 0$ is $\rho(q_n) = f_n$ where $f_0 = 1$, $f_1 = 3$ and $f_n = 3\,f_{n-1} - 2$.*

Proof. As q_n begins and ends with an a for each $n \in \mathbb{N}$, by Lemma 7, we have that for each $n > 1$

$$f_n = 2\,(f_{n-1} + f_{n-2} + \ldots + f_1 + f_0) + 1 - 2\,(n-1)$$

so $f_n + 2 = 2\,f_{n-1} + 2\,(f_{n-2} + \ldots + f_1 + f_0) - 2\,(n-1) + 3 = 3\,f_{n-1}$. □

The first terms of the sequence $(f_n)_{n\geq 0}$ are $1, 3, 7, 19, 55, 163, 487, \ldots$ and $f_n \sim 3^n$. It follows that, for n large enough, the *clustering ratio* tends to zero:

$$\gamma(q_n) = \frac{\rho(B_n)}{\rho(q_n)} \ll 1$$

which states that the BWT is again efficient for this word defined by the morphism $\varphi \circ \mu$.

3.4 Conclusions and further developments.

The study done in this paper supports our initial idea that the BWT is particularly efficient on standard words associated with morphisms. Experimental evidence leads us to consider that this property also holds more generally. Further studies could investigate a formal way to express standard words defined by a generic composition of non-Sturmian morphisms, similarly to the case of Sturmian ones. For instance, consider the morphism $\mu \circ \varphi \circ \widetilde{\varphi}$:

$$\mu \circ \varphi \circ \widetilde{\varphi} : \begin{array}{l} a \mapsto ababba \\ b \mapsto abba \end{array}$$

with $w_n = (\mu \circ \varphi \circ \widetilde{\varphi})^n(a)$. The first iterates are: a, $ababba$, … and

$$\begin{array}{l} B_0 = a \\ B_1 = bab^2a^2 \\ B_2 = b^6a^4ba^2b^8a^9 \\ B_3 = b^{30}a^{21}b^4ab^2a^8b^{39}a^{45} \\ B_4 = b^{150}a^{105}b^{21}a^4ba^2b^8a^{39}b^{195}a^{225} \end{array}$$

In this case, though experimentally BWT shows a positive behaviour on these words, a formal proof of this fact is complicated to achieve since a generic expression for the terms w_n and B_n is not easy to find. The same problem shows up when we study several other morphisms such as: $\varphi \circ \mu \circ \widetilde{\varphi}$ and $\varphi \circ \widetilde{\varphi} \circ \mu$.

References

1. Adjeroh D., Bell T., Mukherjee A. (2008). *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*, Springer Science+Business Media, New York.
2. Barcucci E., Bélanger L., Brlek S. (2004). *On Tribonacci sequences*. Fibonacci Quart. 42, 314–319.
3. Berstel J. (1995). *Axel Thue's papers on repetitions in words: a translation*, Publications du LaCIM 20, Montreal.
4. Burrows M., J. Wheeler D. (1995). *A Block-Sorting Lossless Data Compression Algorithm*. Digital Systems Research Center Research Reports.
5. Crochemore M., Désarménien J., and Perrin D. (2005) *A note on the Burrows-Wheeler transformation*. Theoretical Computer Science 332 (1-3) 567–572.
6. Ferenczi, S., Zamboni, L.Q. (2013). *Clustering Words and Interval Exchanges*. Journal of Integer Sequences 16(2), Article 13.2.1
7. Ferragina P., Giancarlo R., Manzini G., Sciortino M. (2005). *Boosting textual compression in optimal linear time*. Journal of the ACM 52(4) 688–713.
8. Gessel I., Reutenauer C. (1993). *Counting permutations with given cycle structure and descent set*. Journal of Combinatorial Theory, Series A 64(2) 189–215.
9. Hedlund G.A., Morse M. (1938). *Symbolic dynamics*. Amer. J. Math 60, 815–866.
10. Kaplan H., Landau S., Verbin E. (2007). *A Simpler Analysis of Burrows-Wheeler Based Compression*. Theoretical Computer Science 387(3) 220–235,
11. Lothaire M. (2002). *Algebraic Combinatorics on Words* (*Encyclopedia of Mathematics and its Applications*). Cambridge: Cambridge University Press.
12. Mantaci S., Restivo A., Rosone G., Sciortino M. (2017). *Burrows-Wheeler Transform and Run-Length Enconding*. Proceedings WORDS 2017, 11th Int. Conf. on Words, Montréal, Canada, Springer LNCS 10432, 228–239.
13. Mantaci S., Restivo A., Rosone G., Sciortino M., Versari L. (2017). *Measuring the clustering effect of BWT via RLE*. Theoretical Computer Science 698, 79–87.
14. Mantaci S., Restivo A., Sciortino M. (2003). *Burrows-Wheeler transform and Sturmian words*. Inf. Process. Lett. 86, 241–246.
15. Manzini, G. (1999). *An analysis of the Burrows-Wheeler transform*. Journal of the ACM (JACM), 48, 669–677.
16. OEIS Foundation Inc., *The On-line Encyclopedia of Integer Sequences*, `http://oeis.org`, 2011
17. Restivo A., Rosone G. (2009). *Balanced Words Having Simple Burrows-Wheeler Transform*. In: Diekert V., Nowotka D. (eds) Developments in Language Theory. DLT 2009. Springer LNCS 5583, 431–442
18. Restivo A., Rosone G. (2011). *Balancing and clustering of words in the Burrows-Wheeler transform*. Theoretical Computer Science 412 (27) 3019–3032,
19. Restivo A., Rosone G. (2009). *Burrows-Wheeler transform and palindromic richness*. Theoretical Computer Science 410 (30-32) 3018–3026.
20. Simpson J., Puglisi S. J. (2008). *Words with Simple Burrows-Wheeler Transforms*. Electr. J. Comb. 15.
21. Tan B., Wen Z.-Y. (2007). *Some properties of the Tribonacci sequence*. European Journal of Combinatorics 28 (6) 1703–1719.