Burrows-Wheeler Transform and Sturmian Words

Sabrina Mantaci* Antonio Restivo* Marinella Sciortino*

1 Introduction

Michael Burrows and David Wheeler introduced in 1994 (cf. [1]) a reversible transformation on strings (BWT from now on) that arouses considerable interest and curiosity in the field of Data Compression. Such a transformation produces a permutation $BWT(w)$ of an input string w that is easier to compress than the original one, in the sense that there exist very simple universal data compression algorithms providing surprisingly good performances when the original string is preprocessed by applying BWT transform. Actually they achieve compression rates that are very close to the best known compression rate. Moreover, due to their simplicity, such algorithms can be implemented with relatively slow complexity. The underlying idea consists of grouping together symbols that appear in similar contexts. The output of such a transformation gives a sequence where it is very likely to have many instances of a same character close one another. A theoretical analysis of some BWT-based algorithms is available in [6].

Sturmian words were introduced in the forties by Morse and Hedlund(cf. [7]). They are defined as the infinite sequences having exactly $n+1$ distinct factors of length n (cf. [5]). Their numerous properties have lead to a great development of many fields of research. Such a versatility explains also the existence of many equivalent definitions. In particular, a Sturmian word can be defined by considering the intersections with a squared-lattice of a semiline having a slope which is an irrational number. A vertical intersection is denoted by the letter a, an horizontal intersection by b and the intersection with a corner by ab or ba. If the semiline starts from the origin the corresponding Sturmian words is called *characteristic*. Since the language of factors of a Sturmian word depends only on the slope of the corresponding semiline, characteristic words capture most of the properties of Sturmian words.

Characteristic Sturmian words can be constructed by a family of finite words called *Standard words*, in the sense that every characteristic word is the limit of a sequence of standard words (cf. [5]). These words have several characterizations as, for instance, a special decomposition into palindrome words and an extremal property on the periods of

*Dipartimento di Matematica ed Applicazioni. Università di Palermo, via Archirafi 34 - 90123 Palermo - ITALY (e-mail: {sabrina, restivo, mari}@math.unipa.it)

the word that is closely related to Fine and Wilf's theorem (cf. [2, 3]). Moreover they also appear as extremal case in the pattern matching algorithm of Knuth-Morris-Pratt (see [4]).

In this paper we prove that Standard words characterize also the extremal case, for binary alphabets, of BWT in the sense that the transformation produces a total clustering of all the instances of any character. In particular, we show that, given a word u over the alphabet $\{a, b\}$, $BWT(u) = b^k a^h$ *if and only if* u *is the power of a conjugate of a standard word.*

From this result we derive a further characterization of Standard words in terms of the Burrows-Wheeler transform.

2 Burrows-Wheeler transform

The Burrows-Wheeler transform (BWT from now on) was introduced in 1994 by Burrows and Wheeler [1] and it represents a extremely useful tool for textual lossless data compression. The idea is to apply a reversible transformation in order to produce a permutation $BWT(w)$ of an input string w, defined over an alphabet A, so that the string becomes easier to compress. Actually the transformation leads to group characters together so that the probability of finding a character close to another instances of the same character is substantially increased. BWT transforms a word $w = w_0 w_1 \cdots w_n$ by constructing all n cyclic rotations of w, sorting them lexicographically and extracting the last character of each rotation. The word $BWT(w)$ consists of the sequence of these characters. Moreover the transformation computes the index I, that is the row containing the original word in the sorted list of rotations.

For instance, suppose we want to compute $BWT(w)$ where $w = abraca$. Consider the following matrix M, which consists of all cyclic shifts of w, lexicographically sorted.

$$
\begin{array}{ccccccccc}
 & & & F & & & & & L \\
 & & & \downarrow & & & & & \downarrow \\
 & & 0 & a & a & b & r & a & c \\
I & \rightarrow & 1 & a & b & r & a & c & a \\
 & & 2 & a & c & a & a & b & r \\
 & & 3 & b & r & a & c & a & a \\
 & & 4 & c & a & a & b & r & a \\
 & & 5 & r & a & c & a & a & b
\end{array}
$$

Figure 1: The matrix of all cyclic rotations of the word $w = abraca$

The last column L of the matrix represents $BWT(w) = caraab$ and $I = 1$ since the original word w appears in row 1. The first column F, instead, contains the sequence of the characters of w lexicographically sorted.

In [1] the following properties concerning BWT have been proved:

1. For all $i = 0, \ldots, n-1$, $i \neq I$, the character $L[i]$ is followed in the original string by $F[i]$;

2. for each character z, the i-th occurrence of z in F corresponds to the i-th occurrence of z in L.

From above properties it follows that the Burrows-Wheeler transform is reversible in the sense that, given $BWT(w)$ and the index I, it is possible to recover the original string w. Actually, according to Property 2, we can define a function $\tau\colon \{0, 1, \ldots, n-1\} \to \{0, 1, \ldots, n-1\}$ giving the correspondence between the positions of characters of the first and the last column of the matrix M. The function τ represent also the order in which we have to rearrange the elements of F to reconstruct the original word w. Hence, starting from the position I, we can recover the word w as follows:

$$w_i = F[\tau^i(I)] \ , \ \text{where } \tau^0(x) = x, \ \text{and } \tau^{i+1}(x) = \tau(\tau^i(x)).$$

We show, for instance, how the reconstruction works for the example in Figure 1:

$$\tau = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 0 & 2 \end{pmatrix},$$

$$\begin{aligned} w_1 &= F[1] = a \\ w_2 &= F[3] = b \\ w_3 &= F[5] = r \\ w_4 &= F[2] = a \\ w_5 &= F[4] = c \\ w_6 &= F[0] = a. \end{aligned}$$

Definition 1 Two words $x, y \in A^*$ are called *conjugate* if $x = uv$ and $y = vu$ for some $u, v \in A^*$.

From the definition of BWT, the following characterization of conjugate words holds.

Proposition 2 *Two words $x, y \in A^*$ are conjugate if and only if $BWT(x) = BWT(y)$.*

There exists a relation between the BWT of a word and the BWT of a power of the word itself:

Proposition 3 *For $u, v \in A^*$ one has:*

1. *If $u = v^d$ and $BWT(v) = a_0 a_1 \cdots a_{n-1}$ then $BWT(u) = a_0^d a_1^d \cdots a_{n-1}^d$;*

2. *If $BWT(v) = a_0 a_1 \cdots a_{n-1}$ and $BWT(u) = a_0^d a_1^d \cdots a_{n-1}^d$, then there exist a conjugate u' of u such that $u' = v^d$.*

Proof. The proof of Statement 1 is straightforward.

As to concern Statement 2, let $BWT(v) = a_0 a_1 \cdots a_{n-1}$. If we denote by F_v the word obtained by lexicographically sorting the symbols of $BWT(v)$, F_v induces a partition of $\{0, \ldots, n-1\}$ into a set of intervals

$$J_z = \{i \mid F_v(i) = z\}, \text{ with } z \in A.$$

From the invertibility of BWT, it follows that there exists a permutation σ on $\{0, 1, \ldots, n-1\}$ such that for a fixed index $I \in \{0, \ldots, n-1\}$

$$v_i = F_v(\sigma^i(I)). \tag{1}$$

Let π_d: $\{0, \ldots, nd-1\} \to \{0, \ldots, n-1\}$ the projection defined as

$$\pi(x) = \lfloor x/d \rfloor.$$

Let F_u be the word obtained by lexicographically sorting the letters of $BWT(u) = a_0^d a_1^d \cdots a_{n-1}^d$. The partition into intervals M_z induced by F_u is such that

$$\pi_d(M_z) = J_z.$$

Moreover we can notice that

$$F_u = F_v \circ \pi_d.$$

Let τ be the permutation on $\{0, \ldots, nd-1\}$ such that, for a fixed index $K \in \{0, \ldots, nd-1\}$, $u_i = F_u(\tau^i(K))$ for $i = 0, \ldots, nd-1$. If $x \in \{0, \ldots, nd-1\}$ and $\pi_d(x) = j$, that is $x = jd + r$ with $0 \leq r \leq d-1$, then

$$\tau(x) = \sigma(j)d + r.$$

Let $H \in \{0, \ldots, nd-1\}$ be an integer such that $\pi_d(H) = I$. Let u' be the word such that $u'_i = F_u(\tau^i(H)) = F_v(\sigma^i(I))$ for $i = 0, \ldots, nd-1$. Since $\sigma^n = id$, then $u'_i = u'_{i+n}$ for $i = 0, \ldots, (d-1)n-1$, and so from Equality 1, $u' = v^d$. From Proposition 2 it follows that u' and u are conjugates. □

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

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

The sequence $\{s_n\}_{n \geq 0}$ converges to a limit s that is a *characteristic Sturmian word.* Moreover any characteristic Sturmian word is obtained in this way. The sequence $\{s_n\}_{n \geq 0}$

is called the *approximating sequence* of s and $(d_1, d_2, \ldots, d_n, \ldots)$ is the *directive sequence* of s.

For instance, the Fibonacci word is the standard word having directive sequence equal to $(1, 1, \ldots, 1, \ldots)$.

Each finite word s_n in the sequence is called a *standard word.* It is univocally determined by the (finite) directive sequence $(d_0, d_1, \ldots, d_{n-1})$.

Standard words can also be characterized by the notion of *mechanical word* (cf. [5]). Let h, k be two relatively prime integers with $1 \leq h < k$. We define the words $X_{h,k} = x_0 x_1 \cdots x_{k-1}$ and $Y_{h,k} = y_0 y_1 \cdots y_{k-1}$ on the alphabet $\{a, b\}$ by:

$$x_i = \begin{cases} a & \text{if } \lfloor (i+2)\frac{h}{k} \rfloor - \lfloor (i+1)\frac{h}{k} \rfloor = 0 \\ b & \text{if } \lfloor (i+2)\frac{h}{k} \rfloor - \lfloor (i+1)\frac{h}{k} \rfloor = 1, \end{cases}$$

and

$$y_i = \begin{cases} a & \text{if } \lceil (i+2)\frac{h}{k} \rceil - \lceil (i+1)\frac{h}{k} \rceil = 0 \\ b & \text{if } \lceil (i+2)\frac{h}{k} \rceil - \lceil (i+1)\frac{h}{k} \rceil = 1. \end{cases}$$

Remark 4 Note that in the above definition $x_i = y_i$ for all $0 \leq i \leq k-3$ and that $x_{k-2}x_{k-1} = ba$ and $y_{k-2}y_{k-1} = ab$.

It is interesting to give a graphical interpretation of the notion of mechanical words. Consider the straight line with equation $y = (h/k)\, x$. The points with integer coordinates just below this line are $P_i = (i, \lfloor i\frac{h}{k} \rfloor)$. In a mechanical word, the segment joining two consecutive point P_i and P_{i+1} is encoded by the letter a if it is horizontal, by the letter b if it is diagonal. The same observation holds for the point $Q_i = (i, \lceil i\frac{h}{k} \rceil)$

Example 5 Figure 2 shows the two mechanical words $X_{3,8} = abaababa$ and $Y_{3,8} = abaabaab$.

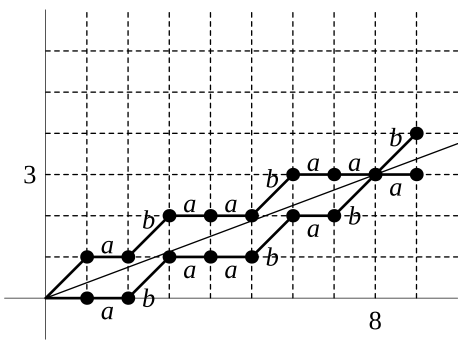

Figure 2: Mechanical words associated with the straight line $y = \frac{3}{8}x$

Consider the function

$$\eta \colon \{a, b\} \to \mathbb{Q} \cup \infty$$

defined as

$$\eta(\epsilon) = 1,\ \eta(w) = |w|_b / |w|_a, \text{ for } w \neq \epsilon.$$

We assume that $1/0 = \infty$. For any $w \in \{a,b\}^*$ we call $\eta(w)$ the *slope* of w.
The following proposition holds (cf. [5]):

Proposition 6 *Let $w \in \{a,b\}^*$ with $\eta(w) = p/q$ and length $n = p+q$. Then w is a finite standard word if and only if $GCD(p,q) = 1$ and $w = X_{p,n}$ or $w = Y_{p,n}$.*

Finite standard words of length n can also be defined by rotations on the cyclic group $\mathbb{Z}_n$ of residue classes modulo n. Let p, q be two integers such that $GCD(p,q) = 1$ and let $n = p+q$. Let us consider the rotation over $\mathbb{Z}_n$, $\varrho_p \colon \mathbb{Z}_n \to \mathbb{Z}_n$, defined as:

$$\varrho_p([z]) = [z+p]$$

where $[z]$ denotes the residue class modulo n represented by the integer z.

We can then define a partition of $\mathbb{Z}_n$ in two intervals: $I_a = \{[0], \ldots, [q-1]\}$ and $I_b = \{[q], \ldots, [n-1]\}$. We define a function $\varphi \colon \mathbb{Z}_n \to \{a,b\}$ given by

$$\varphi([z]) = \begin{cases} a & \text{if } [z] \in I_a \\ b & \text{if } [z] \in I_b \end{cases}$$

Given the functions ϱ_p and φ, for each $[x] \in \mathbb{Z}_n$ we can univocally define a word $R_{p,x} = r_0 r_1 \cdots r_{n-1} \in \{a,b\}^n$ where $r_i = R_{p,x}(i) = \varphi(\varrho_p^i([x]))$.

In the following theorem we show that the two standard words of length $n = p+q$ and slope p/q can be obtained by the rotation ϱ_p.

Theorem 7 *Let $w = w_1 w_2 \cdots w_n$ be a finite word with slope p/q and length $n = p+q$ with $GCD(p,q) = 1$. Then*

1. *w is a standard word with suffix ba if and only if $w = R_{p,p}$;*

2. *w is a standard word with suffix ab if and only if $w = R_{p,p-1}$.*

Proof. Let w be a standard word with suffix ba. Then, by Proposition 6, there exist two coprime integers p and q such that $p+q = n$ and $w = X_{p,n} = x_0 x_1 \cdots x_{n-1}$. It is possible to verify that

$$\lfloor (i+2)p/n \rfloor = 1 + \lfloor (i+1)p/n \rfloor \iff (i+1)p \bmod n \geq q \tag{2}$$

From this relation it follows that $x_i = b$ if and only if $(i+1)p \bmod n \geq q$. Then $w_i = x_i = b$ if and only if $[p+ip] = \varrho_p^i([p]) \in I_b$, that is $\varphi(\varrho_p^i([p])) = b$. This implies that $w_i = \varphi(\varrho_p^i([p]))$ for all $i = 0, \ldots, n-1$. Then, by definition, $w = R_{p,p}$.
The converse is proved by reversing the relations and using the Equality 2.

Let us suppose that w is a standard word with suffix ab. By Proposition 6 there exist two coprime integers p and q such that $p+q = n$ and $w = Y_{p,n} = y_0 y_1 \cdots y_{n-1}$. It is possible to prove that

$$\lceil (i+2)p/n \rceil = 1 + \lceil (i+1)p/n \rceil \iff 0 < ((i+1)p + p) \bmod n \leq p \tag{3}$$

From this inequality we can deduce that $w_i = y_i = b$ if and only if $(i+1)p \bmod n = 0$ or $(i+1)p \bmod n \geq q+1$. We can notice that $[(i+1)p-1] = [p-1+ip] = \varrho_p^i([p-1]) \in I_b$, that is $\varphi(\varrho_p^i([p-1])) = b$. So $w_i = \varphi(\varrho_p^i([p-1]))$ for all $i = 0, \ldots, n-1$. Then, by definition, $w = R_{p,p-1}$.
The converse is obtained by reversing the previous relations and using the Equality 3. □

We observe that all the words $R_{p,x}$ with $x = 0, \ldots, n-1$ are one another related by a cyclic rotation.

Definition 8 Let $w = uv$. The *j-th conjugate* or *j-th cyclic rotation* of w is the word $C_j(w) = vu$, with $|v| = j$.

The following lemma establishes that all the words $R_{p,x}$ are one another conjugate.

Lemma 9 *For each $[y] \in \mathbb{Z}_n$, there exists a non-negative integer $k \leq n-1$ such that $R_{p,y} = C_k(R_{p,0})$.*

Proof. Since $GCD(q,n) = 1$ then $\mathbb{Z}_n = \langle [q] \rangle$. If $[y] \in \mathbb{Z}_n$, then $[y] = [kq]$ for some k. We can deduce that $R_{p,y}(i) = \varphi(\varrho_p^i([y])) = \varphi([kq+ip]) = \varphi(\varrho_p^{i-k}([0])) = C_k(R_{p,0})(i)$, where $\varrho_p^j([z]) = \varrho_p^{n-j}([z])$ if $-n+1 \leq j \leq 0$. □

4 Standard words and Burrows-Wheeler Transform

In this section we prove the following theorem that provides a new characterization of standard words by means of the Burrows-Wheeler transform.

Theorem 10 *Let $u \in \{a,b\}^*$. Then $BWT(u) = b^p a^q$ with $GCD(p,q) = 1$ if and only if u is a conjugate of a standard word.*

Proof. Let us suppose that $BWT(u) = b^p a^q$. Let $n = p+q$ the length of u. We are going to show that there exists an index $I \in \{0, \ldots, n-1\}$ such that the antitransform of $b^p a^q$ is a standard word. The thesis will follow by Proposition 2.

Let us consider $I = p$. If we denote by F_u the word obtained by lexicographically sorting $BWT(u)$ and by τ the permutation defined in Section 2, then for $i = 0, \ldots, n-1$,

$$u_i = F_u(\tau^i(p)).$$

Since $F_u = a^q b^p$, we notice that $\tau(x) = (x+p) \bmod n$ for all $x \in \{0, \ldots, n-1\}$. Moreover F_u induces a partition of $\{0, \ldots, n-1\}$ into $J_a = \{0, \ldots, q-1\}$ and $J_b = \{q, \ldots, n-1\}$ and according with the definition of the rotation ϱ_p on $\mathbb{Z}_n$ and the function φ given in Section 3 we have that

$$u_i = \varphi(\varrho_p^i([p])) = R_{p,p}.$$

By Theorem 7 it follows that if $I = p$ then $u = X_{p,n}$ that is the standard word of length n and slope p/q terminating by ba.

Conversely, let u be a conjugate of a standard word.

All the cyclic rotations of u are the words $R_{p,x}$, with $0 \leq x \leq n-1$. We can observe that such words are lexicographically sorted as follows:

$$R_{p,0} < R_{p,1} < \ldots < R_{p,n-1}.$$

Moreover we can notice that the sequence of the first p characters of $BWT(u)$ is the word $\varphi([0-p])\varphi([1-p])\ldots\varphi([p-1-p])$, where $\varphi\colon \mathbb{Z}_n \rightarrow \{a,b\}$ is defined as:

$$\varphi([z]) = \begin{cases} a & \text{if } [z] \in I_a = \{[0], \ldots, [q-1]\} \\ b & \text{if } [z] \in I_b = \{[q], \ldots, [n-1]\} \end{cases}$$

It is easy to see that if $0 \leq x \leq p-1$ then $[q] \leq [x-p] \leq [n-1]$, that is $[x-p] \in I_b$. So

$$BWT(v) = b^p a^q.$$

□

From the proof of previous theorem it follows that:

Corollary 11 *Let $u \in \{a,b\}^*$ a word such that $BWT(u) = b^p a^q$ with $GCD(p,q) = 1$. Let $I \in \{0, 1, \ldots, n-1\}$ be the index used to reconstruct u.*

1. *If $I = p$ then $u = X_{p,n}$.*
2. *If $I = p-1$ then $u = Y_{p,n}$.*

The following corollary generalizes Theorem 10 in the case where the number of a's and the number of b's are not coprime.

Corollary 12 *Let $u \in \{a,b\}^*$. Then $BWT(u) = b^k a^h$ with $GCD(k,h) = d$ if and only if $u = v^d$ where v is a conjugate of a standard word.*

Proof. Let $BWT(u) = b^k a^h$. Since $GCD(k,h) = d$, then

$$BWT(u) = \underbrace{b^d b^d \cdots b^d}_{k/d \text{ times}} \underbrace{a^d a^d \cdots a^d}_{h/d \text{ times}}$$

where $GCD(k/d, h/d) = 1$. By Theorem 10 if $BWT(v) = b^{k/d} a^{h/d}$, then v is a conjugate of a standard word, hence by Proposition 3 $u = v^d$.

Conversely, let $u = v^d$ with v is a conjugate of a standard word. If k and h are respectively the number of b's and the number of a's in u, then k/d and h/d are respectively the number of b's and the number of a's in v, with $GCD(k/d, h/d) = 1$. By Theorem 10 we have $BWT(v) = b^{k/d} a^{h/d}$ and, by Proposition 3, $BWT(u) = b^k a^h$. □

From Corollary 11 and Proposition 3 it follows that:

Corollary 13 *Let $u \in \{a,b\}^*$ a word such that $BWT(u) = b^k a^h$ with $GCD(k,h) = d$. Let $I \in \{0, 1, \ldots, n-1\}$ be the index used to reconstruct u.*

1. *If $k \leq I \leq k+d-1$ then $u = (X_{k/d,n})^d$.*
2. *If $k-d \leq I \leq k-1$ then $u = (Y_{k/d,n})^d$.*

References

[1] M. Burrows and D.J. Wheeler, A block sorting data compression algorithm. Tech. rep., DIGITAL System Research Center, 1994.

[2] A. de Luca, Combinatorics of standard sturmian. In *Structures in Logic and Computer Science* (1997), J. Milcielski, G. Rozenberg, and A. Salomaa, Eds., vol. 1261 of *Lect. Notes Comp. Sci.*, Springer-Verlag, pp. 249–267.

[3] A. de Luca and F. Mignosi, Some combinatorial properties of sturmian words. *Theoret. Comput. Sci.*, 136 (1994), 361–385.

[4] D. Knuth, J. H. Morris, and V. R. Pratt, Fast pattern matching in strings. *Siam J. Comput. 6*, 2 (1977), 323–350.

[5] M. Lothaire, *Algebraic combinatorics on words.* Cambridge University Press, 2001, ch. 2, "Sturmian Words", pp. 40–97.

[6] G. Manzini, An analysis of the Burrows-Wheeler transform. *Journal of the ACM 48*, 3 (2001), 407–430.

[7] M. Morse and G. A. Hedlund, Symbolic dynamics ii: Sturmian trajectories. *Amer. J. Math.*, 62 (1940), 1–42.