Theoretical Computer Science 410 (2009) 3018–3026

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Burrows–Wheeler transform and palindromic richness

Antonio Restivo*, Giovanna Rosone

University of Palermo, Dipartimento di Matematica ed Applicazioni, Via Archirafi 34, 90123 Palermo, Italy

ARTICLE INFO

Keywords:
Combinatorics on words
Burrows–Wheeler transform
Palindromes
Rich words

ABSTRACT

The investigation of the extremal case of the Burrows–Wheeler transform leads to study the words w over an ordered alphabet $A = \{a_1, a_2, \ldots, a_k\}$, with $a_1 < a_2 < \cdots < a_k$, such that $bwt(w)$ is of the form $a_k^{n_k} a_{k-1}^{n_{k-1}} \cdots a_2^{n_2} a_1^{n_1}$, for some non-negative integers $n_1, n_2, \ldots, n_k$. A characterization of these words in the case $|A| = 2$ has been given in [Sabrina Mantaci, Antonio Restivo, Marinella Sciortino, Burrows-Wheeler transform and Sturmian words, Information Processing Letters 86 (2003) 241–246], where it is proved that they correspond to the powers of conjugates of standard words. The case $|A| = 3$ has been settled in [Jamie Simpson, Simon J. Puglisi, Words with simple Burrows-Wheeler transforms, Electronic Journal of Combinatorics 15, (2008) article R83], which also contains some partial results for an arbitrary alphabet. In the present paper we show that such words can be described in terms of the notion of "palindromic richness", recently introduced in [Amy Glen, Jacques Justin, Steve Widmer, Luca Q. Zamboni, Palindromic richness, European Journal of Combinatorics 30 (2) (2009) 510–531]. Our main result indeed states that a word w such that $bwt(w)$ has the form $a_k^{n_k} a_{k-1}^{n_{k-1}} \cdots a_2^{n_2} a_1^{n_1}$ is strongly rich, i.e. the word w^2 contains the maximum number of different palindromic factors.

1. Introduction

Michael Burrows and David Wheeler introduced in 1994 (cf. [3]) a reversible transformation on words that turns out to be an extremely useful tool for textual data compression.

Compression algorithms based on the Burrows–Wheeler Transform (BWT) take advantage of the fact that the word output of BWT shows a local similarity (occurrences of a given symbol tend to occur in clusters) and then turns out to be highly compressible.

In order to investigate such a "clustering effect" of BWT it is interesting to consider the extremal case when all occurrences of each letter make up a factor of the transform, i.e. the transform produces a perfect clustering. Perfect clustering corresponds indeed to optimal performances of compression algorithms.

So we consider the set E of the words w over a totally ordered alphabet $A = \{a_1, a_2, \ldots, a_k\}$, with $a_1 < a_2 < \cdots < a_k$, for which

$$bwt(w) = a_k^{n_k} a_{k-1}^{n_{k-1}} \cdots a_2^{n_2} a_1^{n_1}$$

for some non-negative integers $n_1, n_2, \ldots, n_k$.

The aim of this paper is to describe such words. A complete description of the set E in the case of a binary alphabet has been given in [9], where it is proved that a word is in E if and only if it is a power of a conjugate of a standard word (cf. [8]). In the case of a three letter alphabet a constructive characterization of the elements of E has been recently given by Simpson

* Corresponding author. Tel.: +39 091 6040307; fax: +39 091 6040311.
E-mail addresses: restivo@math.unipa.it (A. Restivo), giovanna@math.unipa.it (G. Rosone).

0304-3975/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2009.03.008

and Puglisi in [10]. In the same paper [10] Simpson and Puglisi approach the problem for an arbitrary alphabet and obtain some partial results (see Theorem 4.2 and Corollary 4.3).

In the present paper we deepen the investigation of the general case and show that the elements of E are "rich" in palindromes, in the sense that they contain the maximum number of different palindromic factors.

The notion of palindromic richness has been introduced very recently and it appears to play a relevant role in combinatorics on words. In [5], Droubay, Justin and Pirillo proved that any word w of length $|w|$ contains at most $|w| + 1$ distinct palindromic factors (including the empty word). Inspired by this result, Glen, Justin, Widmer and Zamboni in [6] initiated a unified study of both finite and infinite words characterized by this palindromic richness. Accordingly, we say that a finite word w is *rich* if and only if it has exactly $|w| + 1$ distinct palindromic factors, and an infinite word is rich if all its factors are rich. Rich words appear in many different contexts: in particular, all episturmian words are rich (cf. [5]). Several characterizations and nice properties of rich words are given in [6,2].

We say that a finite word w is *strongly rich* if the infinite word w^ω is rich. The main result of the present paper states that all words in E are strongly rich. The proof makes use of some special properties of the Burrows–Wheeler matrix M and it is obtained by a detailed analysis of several cases.

Note however that our result does not provide a complete characterization of the set E, since we show that there exist words which are strongly rich and do not belong to the set E.

2. Preliminaries

Let $A = \{a_1, a_2, \ldots, a_k\}$ be a finite ordered alphabet (with $a_1 < a_2 < \cdots < a_k$). We denote by A^* the set of words over A. Given a finite word $w = b_1b_2 \cdots b_n \in A^*$ with each $b_i \in A$, the length of w, denoted $|w|$, is equal to n. By convention, the empty word ε is the unique word of length 0. We denote by $\tilde{w}$ the reversal of w, given by $\tilde{w} = b_n \cdots b_2b_1$. If w is a word that has the property of reading the same in either direction, i.e. if $w = \tilde{w}$, then w is called a *palindrome*. A word has the *two palindrome property* if it can be written as uv where u and v are palindromes or empty.

We say that two words $x, y \in A^*$ are *conjugate*, if $x = uv$ and $y = vu$ for some $u, v \in A^*$. Conjugacy between words is an equivalence relation over A^*. We denote by $[x]$ the *conjugacy classes* containing x. A conjugacy class can also be represented as a circular word. Hence in what follows we will use "circular word" and "conjugacy class" as synonyms.

A word $v \in A^*$ is said to be a *factor* (resp. a *prefix*, resp. a *suffix*) of a word $w \in A^*$ if there exist words $x, y \in A^*$ such that $w = xvy$ (resp. $w = vy$, resp. $w = xv$). A factor (resp. the prefix, resp. the suffix) is *proper* if $xy \neq \varepsilon$ (resp. $y \neq \varepsilon$, resp. $x \neq \varepsilon$). If $L \subseteq A^*$, we denote by $F(L)$ the set of factors of the words in L and by $F_h(L)$ the elements of $F(L)$ of length h. In particular, $F([u])$ (resp. $F_h([u])$) denotes the set of factors (resp. factors of length h) of the conjugates of u.
A factor u of a word w is said to be *unioccurrent* in w if u has exactly one occurrence in w. Otherwise, u has at least two distinct occurrences in w, in which case there exists a factor r of w containing exactly two distinct occurrences of u, one as a prefix and one as a suffix. Such a factor r is called a *complete return* to u in w.

A word $w \in A^*$ is *primitive* if $w = u^h$ implies $w = u$ and $h = 1$. Notice that if a word is primitive, then all of its conjugates are primitive. A circular word, i.e. a conjugacy class, is primitive if any element of the class is primitive. Recall that (cf. [7]) every word $u \in A^*$ can be written in a unique way as a power of a primitive word, i.e. there exists a unique primitive word w and a unique integer k such that $u = w^k$.

If u is a word in A^*, we denote by u^ω the infinite word obtained by infinitely iterating u, i.e. $u^\omega = uuuuu\ldots$.

For all notions and results not explicitly reported here we refer to [8] and [4].

3. The Burrows–Wheeler transform

The Burrows–Wheeler transform was introduced in 1994 by Burrows and Wheeler [3] and represents an 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 sequence w, defined over an ordered alphabet A, so that the sequence becomes easier to compress. Actually the transformation tends to group characters together so that the probability of finding a character close to another instance of the same character is substantially increased. BWT transforms a sequence $w = b_1b_2 \cdots b_n$ by lexicographically sorting all the n conjugates of w and extracting the last character of each conjugate. The sequence $bwt(w)$ consists of the concatenation of these characters. We denote by M the matrix which consists of all conjugates $w_1, w_2, \ldots, w_n$ of w lexicographically sorted. In what follows we will refer to M as the "Burrows–Wheeler matrix" of w. Moreover the transformation computes the index I, that is the row containing the original sequence in the sorted list of the conjugates.

For instance, suppose we want to compute $bwt(w)$ where $w = abraca$. Consider the Burrows–Wheeler matrix M in Fig. 1.

The last column L of the matrix M represents $bwt(w) = caraab$ and $I = 2$ since the original sequence w appears in row 2. The first column F, instead, contains the sequence of the characters of w lexicographically sorted.
Next proposition is an easy consequence of the definition of BWT (cf. [3]).

Proposition 3.1. *The following properties hold:*

1. *For all $i = 1, \ldots, n$, $i \neq I$, the character $L[i]$ is followed in the original string by $F[i]$;*
2. *For each character α, the ith occurrence of α in F corresponds to the ith occurrence of α in L.*

$$\begin{array}{ccccccccc} & & F & & & & & L \\ & & \downarrow & & & & & \downarrow \\ & 1 & a & a & b & r & a & c \\ I \longrightarrow & 2 & a & b & r & a & c & a \\ & 3 & a & c & a & a & b & r \\ & 4 & b & r & a & c & a & a \\ & 5 & c & a & a & b & r & a \\ & 6 & r & a & c & a & a & b \end{array}$$

Fig. 1. The matrix M of the sequence $w = abraca$.

From the above properties of the *BWT*, it follows that the 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 of Proposition 3.1, we can define a permutation

$$\tau : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\} \tag{1}$$

giving the correspondence between the positions of characters of the first and the last column of the matrix M. For instance, the permutation τ of the word w in Fig. 1 is

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 6 & 1 & 3 \end{pmatrix}.$$

Starting from the position I, we can recover the sequence w as follows:

$$a_i = F[\tau^{i-1}(I)], \quad \text{where } \tau^0(x) = x, \quad \text{and} \quad \tau^{i+1}(x) = \tau(\tau^i(x)). \tag{2}$$

Notice that the reconstruction algorithm corresponds decomposing the permutation τ into a product of cycles. In our case there is only one cycle. For instance, the permutation τ of the word $w = abraca$ can be decomposed in this way:

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 6 & 1 & 3 \end{pmatrix} = (2\,4\,6\,3\,5\,1).$$

The permutation τ also represents the order in which we have to rearrange the elements of F to reconstruct the original sequence w. We show, for instance, how the reconstruction works for the example in Fig. 1:

$$\begin{aligned} a_1 &= F[2] = a \\ a_2 &= F[4] = b \\ a_3 &= F[6] = r \\ a_4 &= F[3] = a \\ a_5 &= F[5] = c \\ a_6 &= F[1] = a. \end{aligned}$$

Notice that if we except the index, all the mutual conjugate words have the same Burrows–Wheeler Transform. Actually the index has the only aim of denoting one representative in the conjugacy class. However this index is not necessary for the construction of the matrix M from L.

Notice also that BWT is not surjective on the set A^*, that is, there exist some words in A^* that are not the image of any word by the BWT. Consider for instance the word $u = bccaaab$. It is easy to see that there exists no word w such that $bwt(w) = u$.

4. Extremal case of BWT

In this section, we consider the set E of the words w over a totally ordered alphabet $A = \{a_1, a_2, \ldots, a_k\}$, with $a_1 < a_2 < \cdots < a_k$, for which

$$bwt(w) = a_k^{n_k} a_{k-1}^{n_{k-1}} \cdots a_2^{n_2} a_1^{n_1}$$

for some non-negative integers $n_1, n_2, \ldots, n_k$.

We recall that in the case $|A| = 2$, the set E has been characterized in [9] where it is proved the remarkable result that the set E coincides with the set of power of conjugates of standard words. In the case $|A| = 3$ a constructive characterization of the set E has been given by Simpson and Puglisi in [10]. An approach to the general case has been proposed in the same paper [10] and some partial results are derived (see below).

The next theorem provides a characterization of the words belonging to E in terms of the Burrows–Wheeler matrix M. We denote by R the matrix obtained from M by a rotation of 180°. We denote by F_M, L_M the first and the last column of M and by F_R, L_R the first and the last column of R.

For instance, given the word $w = abraca$, M and R are the following:

M							R					
F_M					L_M		F_R					L_R
a	a	b	r	a	c		b	a	a	c	a	r
a	b	r	a	c	a		a	r	b	a	a	c
a	c	a	a	b	r		a	a	c	a	r	b
b	r	a	c	a	a		r	b	a	a	c	a
c	a	a	b	r	a		a	c	a	r	b	a
r	a	c	a	a	b		c	a	r	b	a	a

Notice that the rows of R correspond to the conjugates of $\tilde{w}$.

Remark 4.1. By construction, the properties 1 and 2 stated in Proposition 3.1 for the matrix M hold true also for the matrix R:

1. For all $i, j = 1, \ldots, n$, $i \neq j$, the character $L_R[i]$ is followed by $F_R[i]$ in the jth row of R.
2. For each character α, the ith occurrence of α in F_R corresponds to the ith occurrence of α in L_R.

As a consequence, given F_R and L_R, one can uniquely reconstruct the matrix R by the same procedure used for reversing BWT.

Theorem 4.2. *A word $w \in E$ if and only if $M = R$.*

Proof. Let w be a word in E and let M be the corresponding Burrows–Wheeler matrix. Since $bwt(w) = L_M = a_k^{n_k} a_{k-1}^{n_{k-1}} \cdots a_2^{n_2} a_1^{n_1}$, one has $L_M = \widetilde{F_M}$. Since, by definition of R, $L_R = \widetilde{F_M}$ and $F_R = \widetilde{L_M}$, it follows that

$$L_R = L_M \quad \text{and} \quad F_R = F_M. \tag{3}$$

By Remark 4.1, $M = R$.

Conversely, if $M = R$, it follows trivially that $bwt(w) = a_k^{n_k} a_{k-1}^{n_{k-1}} \cdots a_2^{n_2} a_1^{n_1}$, i.e. $w \in E$. □

We mention that a result equivalent of Theorem 4.2 has been obtained, with a different proof, by Simpson and Puglisi [10, Theorem 4.3]. They also derive the following corollary (cf. [10, Corollary 4.4]).

Corollary 4.3. *Each conjugate of $w \in E$ has the two palindrome property.*

Proof. From Theorem 4.2 one easily derives that for any $w \in E$, $\tilde{w}$ is conjugate of w. Then $w = uv$ and $\tilde{w} = vu$ for some u and v. It follows that $uv = (\widetilde{vu}) = \tilde{u}\tilde{v}$ so that u and v are palindromes, and w has the two palindrome property. □

5. Rich words

Recall that (cf. [5]) any word w of length $|w|$ contains at most $|w| + 1$ distinct palindromic factors (including the empty word). Glen et al. in [6,2] introduced and studied rich words, that constitute a new class of finite and infinite words characterized by containing the maximal number of distinct palindromes. We denote by $P(x)$ the set of distinct palindromic factors of x (including ε).

More precisely, a finite word w is *rich* if it has exactly $|w| + 1$ distinct palindromic factors.

We also mention an explicit description of finite and periodic infinite rich words that are established in [6] and [5] (see also [1]).

Proposition 5.1. *For any finite or infinite word w, the following conditions are equivalent:*

1. *w is rich;*
2. *every factor u of w contains $|u| + 1$ distinct palindromes;*
3. *every prefix (resp. suffix) of w has a unioccurrent palindromic suffix (*`ups` *for short) (resp. prefix (*`upp` *for short));*
4. *for each palindromic factor p of w, every complete return to p in w is a palindrome.*

Proposition 5.2. *For a finite word w, the following properties are equivalent:*

1. *w^ω is rich;*
2. *w^2 is rich;*
3. *w is a product of two palindromes and all of the conjugates of w (including itself) are rich.*

We say that a finite word w is *strongly rich* if the infinite word w^ω is rich.

Remark 5.3. The hypothesis that all of the conjugates of w are rich is not sufficient in order to have a strongly rich word: abc is rich, but it is not strongly rich. The hypothesis that w is rich and a product of two palindromes is not sufficient either: $w = ba^2bab^2aba^2b$ is a rich palindrome, but the conjugate $w' = a^2bab^2aba^2b^2$ is not rich.

The following propositions (cf. [2]) will be useful in what follows.

Proposition 5.4. *A finite or infinite word w is rich if and only if, for each factor $v \in F(w)$, any factor of w beginning with v and ending with $\tilde{v}$ and not containing v or $\tilde{v}$ as an interior factor is a palindrome.*

Proposition 5.5. *Suppose w is a rich word. Then, for any non-palindromic factor v of w, $\tilde{v}$ is a unioccurrent factor of any complete return to v in w.*

Remark 5.6. The above proposition tells us that for any factor v of a rich word w, occurrences of v and $\tilde{v}$ alternate in w.

Clearly, if w has a upp, say p, then p is the unique upp and the longest palindromic prefix of w.

The following lemmas, which are fundamental for the proof of our main result (Theorem 6.2), take into account two words of the form bw and wa, where $w \in A^*$ and $a, b \in A$. We suppose that bw and wa are rich and we denote by p the upp of wa and q the upp of bw. Remark that $|p|, |q| \leq |w| + 1$.

Lemma 5.7. $|q| \leq |p| + 2$.

Proof. By contradiction, assume $|q| > |p| + 2$, hence bp is a prefix of q, it follows that pb is a suffix of q, so p has two occurrences in wa, which is a contradiction. □

Lemma 5.8. *If $|q| > 1$ and $|p| \leq |q|$ then bwa is rich.*

Proof. We have to prove that bwa has a upp. We set $|wa| = |bw| = h$, hence $|p|, |q| \leq h$. Since $|q| \leq |p| + 2$, the following cases are allowed.
Case 1: $|p| < |q| \leq |p| + 2$.
Suppose, by contradiction, that bwa is not rich, so there are two different occurrences of q in bwa; since p is a factor of q, there are also two occurrences of p in wa, and then p is not a upp of wa, a contradiction. So $P(bwa) = P(wa) \cup \{q\}$ and bwa is rich.
Case 2: $1 < |q| = |p|$.
In this case one has that $q = bz$ and $p = zc$, for some $z \in A^*$ and $b, c \in A$. Since p, q are palindromes, it follows, from the property of the palindrome word, that $z = (cb)^j$, with $j > 0$, so we can write $q = b(cb)^j = (bc)^j b$, $p = c(bc)^j = (cb)^j c$ and $q = bz = \tilde{z}b$.

Consider first the case $b \neq c$. We suppose, by contradiction, that bwa is not rich. So q is not a upp of bwa and there are at least two occurrences of q in bwa. It follows that $a = b$. If the two occurrences of q overlap or are separated by one letter, then $p = w$ and bpb is the upp of bwb, so $P(bwb) = P(wb) \cup \{bpb\}$ and bwb is rich. Otherwise bwb is of the form $bwb = qxq$ for some word $x \in A^*$ and $|x| \geq 2$:

$$bwb = \overbrace{bz}^{q}\,\underbrace{c\cdots}_{x}\,\overbrace{\tilde{z}b}^{q}.$$

Since p is the only upp of wb, it follows that the final letter of x is not c, so x is not a palindrome. As wb is rich, by Remark 5.6, the factors z and $\tilde{z}$ alternate in wb. Hence $bwb = b\underbrace{zr\tilde{z}tzs\tilde{z}}_{w}b$, where the factors z and $\tilde{z}$ do not appear in $zr\tilde{z}$ except as prefix and suffix, and t, s can contain the factors z and $\tilde{z}$ alternating. So, by Proposition 5.4, the factor $zr\tilde{z}$ is a palindrome and is a palindromic prefix of wa of length greater than $|p|$ and $|q|$, which leads contradiction. Hence $q = b(cb)^j$ occurs once in bwb, $P(bwb) = P(wb) \cup \{b(cb)^j\}$, so bwb is rich.
In the case $b = c$, since q is the longest palindromic prefix of bw, one has that $|p| = |q| = h$ and $v = b^{h+1}$. Indeed, if $|p| = |q| = j < h$, then b^{j+1} is the longest palindromic prefix of bw, so $|q| > |p|$, which is a contradiction. So $p = b^h$, $q = b^h$, it follows that $bwb = b^{h+1}$ and thus bwb is a palindrome and the upp of itself, so $P(bwb) = P(wb) \cup \{b^{h+1}\}$ and bwb is rich. □

Lemma 5.9. *If $|p| \geq |w|$ and $|q| < |p|$ then bwa is rich.*

Proof. We proceed by contradiction and suppose bwa is not rich. Thus, there is a second occurrence of q in bwa and it must contain the final letter of wa, which, since q is a palindrome, necessarily equals b. By setting $|wa| = |bw| = h$, one has that $|p|, |q| \leq h$. Since $|p| \geq |w|$, then $|p| = h$ or $|p| = h - 1$.

Consider first the case where $|p| = h - 1$. In this case, $p = w$, so $bwb = bpb$. Therefore bpb is a upp of itself, hence $P(bwb) = P(wb) \cup \{bpb\}$ and bwb is rich. Now, we consider the case where $|p| = h$ and divide the case in two subcases depending on the length of q.
First we prove the case where $|q| = 1$. In this case, we can write $q = b$, so b is a upp of bw and, by hypothesis, the letter b does not appear in w. We observe that a second occurrence of q in bwa must contain the final letter of wa, i.e. one has $a = b$. Since $|p| = h$ then $p = wb$. Moreover p is a palindrome, so we can write $w = bx$ and one has $bwb = b\underbrace{bxb}_{p}$, against the hypothesis that $q = b$ is a upp of bw. Hence q is unioccurrent in bwb and bwb is rich.
Now we prove the case where $|q| > 1$. We can write $q = bzb$, where z is a palindrome. So if q is not unioccurrent in bwb, then it follows that $bwb = b\underbrace{zbrbz}_{w}b$, for some $r \in A^*$.

Since $|p| = h$, one has $bwb = bp$. We observe that the prefix of p is zb and the suffix of p is bz. As q is not unioccurrent in bwb, the suffix of bwb is bzb, so the suffix of bwb of length $|z| + 1$ is equal both to bz and zb. Then $bz = zb$, hence z is a power of b: $z = b^j$, with $j \geq 0$. So $bwb = bp = b\underbrace{b^j brbb^j}_{p}$, where r is a palindrome and the first and the last letter of r are not b. So $q = b^{j+2}$ and $p = b^{j+1}rb^{j+1}$. Since, by contradiction, we supposed that the second occurrence of q is a suffix of bwb, then the last letter of r is b. Since r is a palindrome, the first letter of r is b and then $q = b^{j+3}$ and $p = b^{j+2}rb^{j+2}$. We repeat again the argument, until, from the property of the palindrome word, we reach that $p = q = b^h$ and this contradicts the fact $|q| < |p|$. Since $|q| < |p|$ one has that $r \neq \varepsilon$, the first and the last letter of r are not b, hence q occurs only once. So $q = b^{j+2}$ and $p = b^{j+1}rb^{j+1}$, but q does not appear in r, because it is a upp of bw. So $P(bwb) = P(wb) \cup \{b^{j+2}\}$ and bwb is rich. □

6. Main result

This section is devoted to the proof of our main result. In order to prove it, we first prove the following lemma.

Lemma 6.1. *If $v = bu'b$ is a prefix of a word $w \in E$, where bu' and $u'b$ are rich and b does not appear in u', then the first and last letters of u' are equal.*

Proof. Suppose, on the contrary, that the first and the last letters of u' are distinct.

As u' is rich, we can write $u' = p_1p_2 \cdots p_k$, where $k \leq |u'|$ and every p_i, for $i = 1, \ldots, k$, is recursively defined as follows:

- p_1 is the upp of u'.
- p_i is the upp of suffix of u' that is obtained by deleting $p_1, \ldots, p_{i-1}$.

By construction, $p_i \neq p_j$ for each $i \neq j$, with $i, j = 1, \ldots, k$.

Since $v = bp_1 \cdots p_kb$ is prefix of w, that is $w = vt$ for some word t, then, by Theorem 4.2, in $[w]$, there exist the conjugates of the form $\underbrace{p_1 \cdots p_k}_{u'} btb$ and $b\tilde{t}b \underbrace{p_k \cdots p_1}_{\tilde{u'}}$. Since the first and the last letters of u' are distinct, we suppose, without loss of generality, that $p_k \cdots p_1 b\tilde{t}b$ is lexicographically less than $p_1 \cdots p_kbtb$, hence the last letter of u' is less than the first letter of u'. So the two conjugates of w appear in the following order in the Burrows–Wheeler matrix M of w:

$$\begin{array}{lcr} F & & L \\ p_k & \cdots & p_1b\tilde{t}b \\ \vdots & & \vdots \\ p_1 & \cdots & p_kbtb \end{array}$$

Now we prove, by induction, that each conjugate that begins with p_i, for $i \geq 3$ odd, is greater than the conjugate $p_1 \cdots p_kbtb$. We first prove the statement for $i = 3$. Since the b's in the last column of M are consecutive and p_2 does not contain b, in M we have:

$$\begin{array}{lcr} F & & L \\ p_k & \cdots & p_1b\tilde{t}b \\ \vdots & & \vdots \\ p_1 & \cdots & p_kbtb \\ \vdots & & \vdots \\ p_1b\tilde{t}bp_k & \cdots & p_2 \end{array}$$

Since the letters of w of last column of M are non-increasing, also the other conjugates which end with p_2 must be greater than the conjugate which ends with b. Hence $p_3 \cdots p_kbtbp_1p_2 > p_1 \cdots p_kbtb$. The same argument shows that $p_3 \cdots p_1b\tilde{t}bp_k \cdots p_4 > p_1 \cdots p_kbtb$.

Now suppose the statement is true for all integers up to $2i - 1$, i.e. each conjugate that begins with p_{2i-1} is greater than the conjugate $p_1 \cdots p_kbtb$ and we prove that the conjugate that begins with p_{2i+1} is greater than the conjugate $p_1 \cdots p_kbtb$. Hence the conjugates in M are ordered so:

$$\begin{array}{lcr} F & & L \\ p_1 & \cdots & p_{2i-1}p_{2i} \cdots p_kbtb \\ \vdots & & \vdots \\ p_{2i-1} & \cdots & p_1b\tilde{t}bp_k \cdots p_{2i} \end{array}$$

Since p_{2i} does not contain b, we have that the last letter of p_{2i} is greater than b, so the conjugate $p_{2i+1}\cdots p_k btbp_1\cdots p_{2i}$ is greater than the conjugate $p_1\cdots p_k btb$. Hence in M we have the following order:

$$\begin{array}{lcr} F & & L \\ p_1 & \cdots & p_{2i}p_{2i+1}\cdots p_k btb \\ \vdots & & \vdots \\ p_{2i+1} & \cdots & p_k btbp_1\cdots p_{2i} \end{array}$$

We proved that the last letter of each p_i, where i is even, is less than b. Hence if k is odd, the last letter of p_{k-1} is greater than b, so in M we have:

$$\begin{array}{lcr} F & & L \\ p_k & \cdots & p_1 b\tilde{t}b \\ \vdots & & \vdots \\ p_1 & \cdots & p_k btb \\ \vdots & & \vdots \\ p_k btbp_1 & \cdots & p_{k-1} \end{array}$$

Since the first letter of p_k is less and greater than the first letter of p_1, it follows that they are equal, a contradiction.

If k is even then by similar arguments we can prove that the last letter of each p_i, with i odd, is greater than b. Hence, it follows that the last letter of each p_i, with i even, is less than b and the last letter of each p_i, with i odd, is greater than b. So the situation in the matrix M is the following:

$$\begin{array}{lcr} F & & L \\ b\tilde{t}bp_k & \cdots & p_1 \\ \vdots & & \vdots \\ p_k & \cdots & p_1 b\tilde{t}b \\ \vdots & & \vdots \\ p_1 & \cdots & p_k btb \\ \vdots & & \vdots \\ btbp_1 & \cdots & p_k \end{array}$$

This is a contradiction, because the b's in the first column of M are not consecutive. So u' begins and ends with the same letter. This concludes the proof of the lemma. □

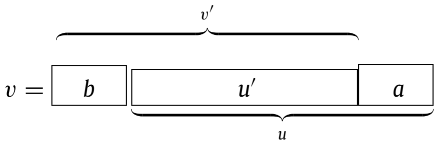
Theorem 6.2. *If the word w belongs to E then w is strongly rich.*

Proof. By Corollary 4.3 each $w \in E$ has the two palindrome property. Hence, by Proposition 5.2 it suffices to prove that all the conjugates of $w \in E$ (including itself) are rich. So we prove, by induction on h ($1 \le h \le n$), that each factor of length h of words in $[w]$, or equivalently each prefix of length h of a conjugate of w is rich.

The result is clearly true if $h \le 3$, in fact it is easy to verify that all words of length 3 or less are rich.

Now suppose the statement is true for all factors of length less than or equal to h, i.e. each factor $u \in F_h([w])$ is rich and we prove that each factor $v \in F_{h+1}([w])$ is rich.

If $v \in F_{h+1}([w])$ then v is of the form $v = bu$, with $b \in A$ and $u \in F_h([w])$. If a is the last letter of u, we can write $v = bu = bu'a$, with $a \in A$ and $u' \in F_{h-1}([w])$. Set $v' = bu'$. Clearly $v' \in F_h([w])$. The situation is depicted in the figure below.



By the induction hypothesis $u \in F_h([w])$ is rich, so u has a unioccurrent palindromic prefix (upp) p. Clearly $|p| \le h$.

By using again the induction hypothesis $v' \in F_h([w])$ is rich, so v' has a upp q. Clearly $|q| \le h$.

By Lemma 5.7, we have that $|q| \le |p| + 2$. We have to prove that v is rich. The proof can be divided in several cases depending on the relative lengths of p and q. We observe that if $|q| > 1$ and $|p| \le |q| \le |p| + 2$, then from Lemma 5.8, v is rich. Moreover if $|q| < |p|$ and $h - 1 \le |p| \le h$, then from Lemma 5.9, v is rich. Therefore it suffices to consider the case where $|q| < |p| < h - 1$ and $|q| > 1$ (case 1) and the case where $|q| \le |p|$ and $|q| = 1$ (case 2).

Suppose, on the contrary, v is not rich. Then v contains two occurrences of q and, in particular, q appears as suffix of v. So $v = bu'b$ and u' is not a palindrome (otherwise v is a palindrome too and the upp of itself). We will show that the condition that u' is not a palindrome leads to a contradiction.

Let $u' = \gamma_1 \cdots \gamma_{i_0} \cdots \gamma_{h-i_0} \cdots \gamma_{h-1}$. Since u' is not a palindrome, there exists the smallest integer i_0 such that $\gamma_{i_0} \neq \gamma_{h-i_0}$. We set

$$z = \gamma_1 \cdots \gamma_{i_0-1} \quad \text{and} \quad \tilde{z} = \gamma_{h-i_0+1} \cdots \gamma_{h-1}.$$

So we have that $v = bu'b = bz\gamma_{i_0} \cdots \gamma_{h-i_0}\tilde{z}b$, where $\gamma_{i_0} \neq \gamma_{h-i_0}$.
Now, we examine the two cases and prove that in both cases $z \neq \varepsilon$.

Case 1. If $1 < |q| < |p| < h - 1$ then we can write $q = bz'b$ where z' is a palindrome. As $bz'b$ is prefix and suffix of v, it follows that $z'b$ is a prefix of u' and bz' is a suffix of u'. In this case, $z \neq \varepsilon$, in fact $q = bz'b$, with z' palindrome, so one has $v = b\underbrace{z'b \cdots bz'}_{u'}b$ and hence $z'b$ is a prefix of z.

Case 2. If $1 = |q| \leq |p| < h - 1$, then we can write $q = b$. In this case, z could be an empty word. From Lemma 6.1 such a case cannot occur, so $z \neq \varepsilon$.

In both cases one has that $z \neq \varepsilon$. Since u' is rich, if z occurs only as a prefix and $\tilde{z}$ occurs only as a suffix of u', according to Property 4 of Proposition 5.1 (if z is a palindrome) or to Proposition 5.4 (if z is not a palindrome), it follows that between z and $\tilde{z}$ there is a palindrome factor. Hence u' is a palindrome, which contradicts the condition $\gamma_{i_0} \neq \gamma_{h-i_0}$.

Thus the factors z and $\tilde{z}$ occur several times in the word v. By Proposition 5.4 and Remark 5.6 we can write $u' = zp_1\tilde{z}p_2z \cdots \tilde{z}y_2zy_1\tilde{z}$, where every $(p_i)_{i\geq 1}$ (resp. $(y_i)_{i\geq 1}$), is the sequence of palindromic factors between z and $\tilde{z}$ constructed from left to right (resp. from right to left).

We denote by α_i the first and last letter of p_i and by β_i the first and last letter of y_i. By hypothesis $\alpha_1 \neq \beta_1$. Since q does not appear in u', if $|q| = 1$ then $\alpha_i, \beta_i \neq b$, for any i. If $|q| > 1$ then bz' is a suffix of $\tilde{z}$ and so $\alpha_{2i}, \beta_{2i} \neq b$, for any i. We will prove that $\alpha_i \neq \beta_i$ for any i.

As $\alpha_1 \neq \beta_1$, we can suppose, without loss of generality, that $\alpha_1 < \beta_1$. Since this inequality is often used in the sequel of the proof we refer to it as the Property *P1*.

Since, by Theorem 4.2, $[w]$ and its factors are closed under reverse, then, for any factor v of $[w]$, there exists in $[w]$ also the factor $\tilde{v}$.

Recall that $v = bzp_1\tilde{z} \cdots zp_{2i-1}\tilde{z}p_{2i}zp_{2i+1} \cdots y_{2i+1}\tilde{z}y_{2i}zy_{2i-1}\tilde{z} \cdots zy_1\tilde{z}b$ is a prefix of w, that is $w = vt$, for some word t, so there exist the two conjugates

$$w' = zp_1\tilde{z} \cdots zp_{2i-1}\tilde{z}p_{2i}zp_{2i+1} \cdots y_{2i+1}\tilde{z}y_{2i}zy_{2i-1}\tilde{z} \cdots zy_1\tilde{z}btb$$

and

$$w'' = zy_1\tilde{z} \cdots zy_{2i-1}\tilde{z}y_{2i}zy_{2i+1} \cdots p_{2i+1}\tilde{z}p_{2i}zp_{2i-1}\tilde{z} \cdots zp_1\tilde{z}b\tilde{t}b.$$

We now show the following properties (*P2* and *P3*) concerning the pairs of letters (α_i, β_i):
P2: For all i, if $\alpha_{2i-1} \leq \alpha_1 < \beta_1 \leq \beta_{2i-1}$ then $\alpha_{2i} > b > \beta_{2i}$.

As the last column in the Burrows–Wheeler matrix M of w is anti-lexicographically ordered, if follows that the conjugates of w that end with b are consecutive rows in M. Moreover, as the conjugate that begins with $z\alpha_{2i-1}$ (resp. $z\beta_{2i-1}$) is less (greater) than or equal to the conjugate w' (resp. w''), in M the conjugates appear in the following order:

$$\begin{array}{lr}
F & L \\
zp_{2i-1}\tilde{z} \cdots zp_1\tilde{z}b\tilde{t}bzy_1\tilde{z} \cdots zy_{2i-1}\tilde{z}y_{2i} \cdots & \cdots p_{2i} \\
\vdots & \vdots \\
zp_1\tilde{z} \cdots zp_{2i-1}\tilde{z}p_{2i} \cdots & \cdots y_{2i}zy_{2i-1}\tilde{z} \cdots zy_1\tilde{z}btb \\
\vdots & \vdots \\
zy_1\tilde{z} \cdots zy_{2i-1}\tilde{z}y_{2i} \cdots & \cdots p_{2i}zp_{2i-1}\tilde{z} \cdots zp_1\tilde{z}b\tilde{t}b \\
\vdots & \vdots \\
zy_{2i-1}\tilde{z} \cdots zy_1\tilde{z}btbzp_1\tilde{z} \cdots zp_{2i-1}\tilde{z}p_{2i} \cdots & \cdots y_{2i}
\end{array}$$

Hence $\alpha_{2i} > b > \beta_{2i}$.

P3: For all i, if $\alpha_{2i} > b > \beta_{2i}$ then $\alpha_{2i+1} \leq \alpha_1 < \beta_1 \leq \beta_{2i+1}$.

Since α_{2i} (resp. β_{2i}) is greater (resp. less) than b then the other conjugates that end with α_{2i} (resp. β_{2i}) are less than w' (resp. greater than w''), hence the conjugate that begins with zp_{2i+1} (resp. zy_{2i+1}) and ends with p_{2i} (resp. y_{2i}) is less (resp. greater) than the conjugate w' (resp. w''). So in M the conjugates appear in the following order:

$$\begin{array}{lr}
F & L \\
zp_{2i+1}\cdots & \cdots y_{2i+1}\tilde{z}y_{2i}z\cdots zy_1\tilde{z}btbzp_1\tilde{z}\cdots\tilde{z}p_{2i} \\
\vdots & \vdots \\
zp_1\tilde{z}\cdots\tilde{z}p_{2i}zp_{2i+1}\cdots & \cdots y_{2i+1}\tilde{z}y_{2i}z\cdots zy_1\tilde{z}btb \\
\vdots & \vdots \\
zy_1\tilde{z}\cdots\tilde{z}y_{2i}zy_{2i+1}\cdots & \cdots p_{2i+1}\tilde{z}p_{2i}z\cdots zp_1\tilde{z}b\tilde{t}b \\
\vdots & \vdots \\
zy_{2i+1}\cdots & \cdots p_{2i+1}\tilde{z}p_{2i}z\cdots zp_1\tilde{z}b\tilde{t}bzy_1\tilde{z}\cdots\tilde{z}y_{2i} \\
\vdots & \vdots
\end{array}$$

Hence $\alpha_{2i+1} \leq \alpha_1 < \beta_1 \leq \beta_{2i+1}$.
Now we prove, by induction, that for all integers j one has:

$$\alpha_{2j-1} \leq \alpha_1 < \beta_1 \leq \beta_{2j-1} \quad \text{and} \quad \alpha_{2j} > b > \beta_{2j}.$$

From the Property *P1*, the result is clearly true for $j = 1$. From the Property *P2*, since $\alpha_1 < \beta_1$, it follows that $\alpha_2 > b > \beta_2$.

Now suppose the statement is true for all integers up to $2j - 1$. From the Property *P2*, it follows that $\alpha_{2j} > b > \beta_{2j}$ and from the Property *P3*, it follows that if $\alpha_{2j} > b > \beta_{2j}$ then $\alpha_{2j+1} \leq \alpha_1 < \beta_1 \leq \beta_{2j+1}$.

We can then conclude that, for all integers j, $\alpha_j \neq \beta_j$. Denote by k the number of occurrences of z in u' (which coincides with the number of occurrences of $\tilde{z}$). By the definition of the sequences of words p_i and y_i, one has $p_k = y_k$. It follows that $\alpha_k = \beta_k$, a contradiction.

So, assuming that u' is not a palindrome, we have obtained a contradiction. We conclude that u' is a palindrome and then $v = bu'b$ is rich.

This concludes the proof of the theorem. □

Example 6.3. The word $w = cacbcac$ is in E, in fact $bwt(w) = ccccbaa$, and one can easily verify that w is strongly rich.

The following example shows that the converse of Theorem 6.2 is false.

Example 6.4. The word $w = ccaaccb$ is strongly rich, but $bwt(w) = cacccba$, hence $w \notin E$.

Acknowledgments

The authors are very grateful to the anonymous referees: their remarks were helpful in substantially improving this paper by pointing out some mistakes and inaccuracies in the first version. Moreover their suggestions have enhanced the readability of the paper.

References

[1] Michelangelo Bucci, Alessandro De Luca, Amy Glen, Luca Q. Zamboni, A new characteristic property of rich words, Theoretical Computer Science 410 (30–32) (2009) 2860–2863.
[2] Michelangelo Bucci, Alessandro De Luca, Amy Glen, Luca Q. Zamboni, A connection between palindromic and factor complexity using return words, Advances in Applied Mathematics 42 (1) (2009) 60–74.
[3] Michael Burrows, David J. Wheeler, A block sorting data compression algorithm. Technical Report, DIGITAL System Research Center, 1994.
[4] Christian Choffrut, Juhani Karhumaki, Combinatorics of words, in: G. Rozenberg, A. Salomaa (Eds.), in: Handbook of Formal Language Theory, vol. 1, Springer-Verlag, Berlin, 1997.
[5] Xavier Droubay, Jacques Justin, Giuseppe Pirillo, Episturmian words and some constructions of de Luca and Rauzy, Theoretical Computer Science 255 (1–2) (2001) 539–553.
[6] Amy Glen, Jacques Justin, Steve Widmer, Luca Q. Zamboni, Palindromic richness, European Journal of Combinatorics 3 (2) (2009) 510–531.
[7] M. Lothaire, Combinatorics on Words, in: Encyclopedia of Mathematics, vol. 17, Addison-Wesley, Reading, Mass, 1983, Reprinted in the Cambridge Mathematical Library, Cambridge University Press, 1997.
[8] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, 2002.
[9] Sabrina Mantaci, Antonio Restivo, Marinella Sciortino, Burrows-Wheeler transform and Sturmian words, Information Processing Letters 86 (2003) 241–246.
[10] Jamie Simpson, Simon J. Puglisi, Words with simple Burrows-Wheeler transforms, Electronic Journal of Combinatorics 15 (2008) article R83.