Crochemore Factorization of Sturmian and Other Infinite Words

Jean Berstel[1] and Alessandra Savelli[1,2]

[1] Institut Gaspard Monge (IGM)
Université de Marne-la-Vallée
[2] Dipartimento di Elettronica e Informazione
Politecnico di Milano
{berstel, savelli}@univ-mlv.fr

Abstract. The Crochemore factorization was introduced by Crochemore for the design of a linear time algorithm to detect squares in a word. We give here the explicit description of the Crochemore factorization for some classes of infinite words, namely characteristic Sturmian words, (generalized) Thue-Morse words, and the period doubling sequence.

1 Introduction

In a seminal paper, Ziv and Lempel [7] defined several factorizations of finite words related to information theory and text processing. Several years later, Crochemore ([2,4,3]) introduced a similar factorization of words as a key tool in the design of a linear algorithm checking words for square freeness.

We study here both Ziv-Lempel and Crochemore factorization of special classes of infinite words, such as Sturmian words and some automatic words. It appears that these factorizations can be expressed by a closed formula in many significant examples. The proof of these formulas require some insight in the combinatorial structure of the infinite words considered.

Some factorizations are quite surprising. As an example, the Ziv-Lempel factorization of the Fibonacci word will be shown to be

$$f = a|b|aa|bab|aabaa|\cdots$$

This is precisely the so called singular factorization introduced by Wen and Wen ([8], see also [1]) in a completely different context.

Ziv-Lempel and Crochemore factorizations have similar properties. Both can be computed in linear time by preprocessing the suffix tree of the word. Furthermore, the number of factors in both factorizations are closely related: the number of factors of the Crochemore factorization is at most twice the number of factors of the Ziv-Lempel factorization. However, there are examples of factorizations which differ significantly infinitely many times.

In this paper we study the behavior of the Crochemore factorization in the case of some of the most known classes of words, i.e., characteristic Sturmian words, the Thue-Morse word, and the period doubling sequence.

R. Královič and P. Urzyczyn (Eds.): MFCS 2006, LNCS 4162, pp. 157–166, 2006.

As we shall see, the Crochemore factorization (or c-factorization for short) of special infinite words can be described explicitly, and it reflects the structure of these words.

The c-factorization $c(x)$ of a word x is defined as follows. Each factor of $c(x)$ is either a fresh letter, or it is a maximal factor of x already occurring in the prefix of the word; more formally, the *c-factorization* $c(x)$ of a word x is

$$c(x) = (x_1, x_2, \ldots, x_m, x_{m+1}, \ldots)$$

where x_m is the longest prefix of $x_m x_{m+1} \cdots$ occurring twice in $x_1 x_2 \cdots x_m$, or x_m is a letter a if a does not occur in $x_1 \cdots x_{m-1}$. For example, the c-factorization of $x = ababaab$ is (a, b, aba, ab), since aba occurs twice in $ababa$.

Note that the the c-factorization of a word differs slightly from the well known Ziv-Lempel factorization [7] (or z-factorization), so that these two factorizations are in general not comparable. The *z-factorization* $z(x)$ of a word x is

$$z(x) = (y_1, y_2, \ldots, y_m, y_{m+1}, \ldots)$$

where y_m is the shortest prefix of $y_m y_{m+1} \cdots$ which occurs only once in the word $y_1 y_2 \cdots y_m$.

For example, let x be the word $x = aabaaccbaabaabaa$. The c-factorization and the z-factorization of x are:

$$c(x) = (a, a, b, aa, c, c, baa, baabaa)$$
$$z(x) = (a, ab, aac, cb, aabaab, aa).$$

We shall discuss the relation between these factorizations in more detail in the final section.

The c-factorization has an interesting behavior in all of the well known infinite words we have considered. For example, take the Fibonacci word

$$\mathbf{f} = abaababaabaab \cdots$$

defined inductively by $f_{-1} = b$, $f_0 = a$, and $f_{n+2} = f_{n+1} f_n$. The c-factorization of f is

$$c(\mathbf{f}) = (a, b, a, aba, baaba, \ldots) = (a, b, a, \widetilde{f_2}, \widetilde{f_3}, \ldots)$$

Observe that each of the factors (except the first three) is the reverse the finite Fibonacci word f_n. We will see that a similar result holds for characteristic Sturmian words (Theorem 1 below).

The c-factorization is closely related to two other factorizations of the Fibonacci word. The first is the factorization into factors which are exactly the prefixes f_n, that is

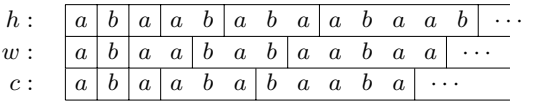
$$h(\mathbf{f}) = (a, b, a, ab, aba, abaab, \ldots) = (a, f_{-1}, f_0, f_1, f_2, \ldots)$$

The other is the Wen and Wen factorization (also called the singular factorization), in which the ith factor has the same length as f_{i-1}, thus resulting

$$w(\mathbf{f}) = (a, b, aa, bab, aabaa, \ldots) = (a, w_0, w_1, w_2, w_3, \ldots)$$

Note that in the Wen and Wen factorization the ith factor is a palindrome, and is the only factor of f of length $|f_{i-1}|$ that is not a conjugate of f_{i-1}.

The three factorizations can be visualized through the following scheme:



The relation between these factorizations is the following. Factors of h and w satisfy

$$bf_{2i} = w_{2i}a \text{ and } af_{2i+1} = w_{2i+1}b,$$

while factors of w and c satisfy

$$aw_{2i} = \widetilde{f}_{2i}b \text{ and } bw_{2i+1} = \widetilde{f}_{2i+1}a.$$

The c-factorization on Fibonacci word is a particular case of a more general result we obtained for the c-factorization on standard Sturmian words.

Theorem 1. *Let* $\mathbf{s}$ *be the standard Sturmian word defined as the limit of*

$$s_{-1} = b, s_0 = a, \text{ and } s_n = s_{n-1}^{d_n} s_{n-2},$$

where $d_i > 0$ *for each* i. *Then*

$$c(\mathbf{s}) = (a, a^{d_1-1}, b, a^{d_1}\widetilde{s}_1^{\,d_2-1}, \widetilde{s}_2^{\,d_3}, \widetilde{s}_3^{\,d_4}, \ldots, \widetilde{s}_n^{\,d_{n+1}}, \ldots)$$

Similar results hold for other familiar infinite words, such as the Thue-Morse and the period doubling sequence. However we do not yet have a full characterization of the c-factorization of automatic words.

The paper is organized as follows. Section 2 contains definitions and statements of theorems. Section 3 sketches the proof for Sturmian words, Section 4 sketches the proof for a family of Thue-Morse sequences and for the period doubling sequence, and Section 5 makes some comparison of Crochemore and Ziv-Lempel factorizations.

2 Basic Definitions and Main Results

Let A be an alphabet and A^* the the set of finite words on A. For any finite word $x = a_1a_2\cdots a_n$, $|x|$ denotes the length n of x and $\widetilde{x}$ denotes the reverse word $a_n\cdots a_2a_1$ of x. If A is a two-letter alphabet $A = \{a, b\}$, then $\overline{x}$ is the image of x of the morphism defined by $\overline{a} = b$ and $\overline{b} = a$. If $x = y\alpha$, with $x, y \in A^*$ and $\alpha \in A$, we denote by x' the word $x' = y\overline{\alpha}$.

A *factorization* of a finite word x is a sequence $(x_1, x_2, \ldots, x_n)$ such that $x = x_1x_2\cdots x_n$. Analogously, a factorization of an infinite word x is a sequence $(x_1, x_2, \ldots)$ such that $x = x_1x_2\cdots$. A recent introduction to factorizations of

words can be found in [6]. The c-factorization $c(x)$ of a word x can be constructively defined by induction on the length of x as follows. If x is a letter, his c-factorization is $c(x) = (x)$. Otherwise, let $x = y\alpha$ with $\alpha \in A$, and $c(y) = (u_1, \ldots, u_k)$. The c-factorization $c(x)$ of x is then

$$c(x) = \begin{cases} (u_1, \ldots, u_k, a) & \text{if } u_k a \text{ is not a factor of } y \\ (u_1, \ldots, u_k a) & \text{otherwise .} \end{cases}$$

The Ziv-Lempel factorization of a word x is

$$z(x) = (y_1, y_2, \ldots, y_m, y_{m+1}, \ldots)$$

where y_m is the shortest prefix of $y_m y_{m+1} \cdots$ which occurs only once in the word $y_1 y_2 \cdots y_m$.

Let $(d_1, d_2, \ldots)$ be an infinite sequence of integers (called a *directive sequence* such that $d_1 \geq 0$ and $d_i > 0$ for $i > 1$, and let $\{s_n\}_{n\geq 0}$ be the infinite sequence of words defined by

$$s_{-1} = b,\ s_0 = a, \text{ and } s_n = s_{n-1}^{d_n} s_{n-2}.$$

It is easy to see that this sequence converges to the infinite word s that is called a *standard Sturmian word*.

Note that the Sturmian word defined by a directive sequence $(0, d_2, d_3, \ldots)$ is obtained from the Sturmian word defined by $(d_2, d_3, \ldots)$ by changing each letter a with a letter b and viceversa, so that in the rest of this paper we will only refer to directive sequences with $d_1 > 0$.

With a result that is very similar to that obtained by de Luca in [5], we have that any standard Sturmian word has a particular decomposition in reverse finite words s_n:

$$\mathbf{s} = \widetilde{s_0}^{\,d_1} \widetilde{s_1}^{\,d_2} \cdots$$

The c-factorization of standard Sturmian words stated in Theorem 1

$$c(\mathbf{s}) = (a, a^{d_1-1}, b, a^{d_1} \widetilde{s}_1^{\,d_2-1}, \widetilde{s}_2^{\,d_3}, \widetilde{s}_3^{\,d_4}, \ldots, \widetilde{s}_n^{\,d_{n+1}}, \ldots)$$

is then clearly closely related to that decomposition.

Let τ be the Thue-Morse morphism on a two-letter alphabet defined by

$$\tau(a) = ab,\ \tau(b) = ba,$$

and let $\{t_n\}_{n\geq 0}$ be the infinite sequence of words such that $t_0 = a$ and $t_n = \tau(t_{n-1})$. This sequence converges to the well known *Thue-Morse infinite word*

$$\mathbf{t} = abbabaabbaababba \cdots$$

Each factor in the c-factorization of $\mathbf{t}$ can be obtained from the previous ones by applying the morphism τ as stated in the following Theorem.

Theorem 2. *The c-factorization $c(\mathbf{t}) = (c_1, c_2, \ldots)$ of the Thue-Morse sequence is $(a, b, b, ab, a, abba, c_7, c_8, \ldots)$ where $c_{n+2} = \tau(c_n)$ for every $n \geq 7$.*

The Thue-Morse word can be generalized in several different ways. We consider here the infinite word $\mathbf{t^{(m)}}$ on a m-letter alphabet $A = \{a_1, a_2, \ldots, a_m\}$ obtained as the limit of the morphism τ_m defined by

$$\tau_m(a_i) = a_i a_{i+1} \cdots a_m a_1 \cdots a_{i-1} \quad (i = 1, \ldots, m).$$

A result that is very similar and even better than that obtained for the Thue-Morse word is the following.

Theorem 3. *Let* $c(\mathbf{t^{(m)}}) = (c_1^{(m)}, c_2^{(m)}, \ldots)$ *be the c-factorization of the generalized Thue-Morse word* $\mathbf{t^{(m)}}$ *with* $m \geq 3$. *Then* $c^{(m)}_{n+2(m-1)} = \tau_m(c_n)$ *for every* $n > m$.

It shows that the c-factorization of the infinite words $\mathbf{t^{(m)}}$ for $m > 2$ are even more regular than in the binary case.

Finally, let δ be the morphism on a two-letter alphabet defined by

$$\delta(a) = ab, \quad \delta(b) = aa,$$

and let $\{q_n\}_{n\geq 0}$ be the infinite sequence of words such that

$$q_0 = a \text{ and } q_{n+1} = \delta(q_n).$$

The limit $\mathbf{q}$ of this sequence is the *period doubling sequence*.
We will denote the reverse of q_i by ${q_i}^R$ and the reverse of q_i' by ${q_i}^S$ (we recall that q_n' is q_n with just the last letter changed to its opposite).

Similarly to the case of standard Sturmian words, the period doubling sequence is the composition of the reverse of the finite period doubling sequence words q_n:

$$\mathbf{q} = {q_0}^R {q_1}^R {q_2}^R \cdots.$$

The c-factorization of $\mathbf{q}$ reflects indeed this property, as stated in the following Theorem, observing that the equality $q^R_{n+1} = q^S_n q^R_n$ holds for each n.

Theorem 4. *Let* $\mathbf{q}$ *be the doubling period sequence. The c-factorization of* $\mathbf{q}$ *is*

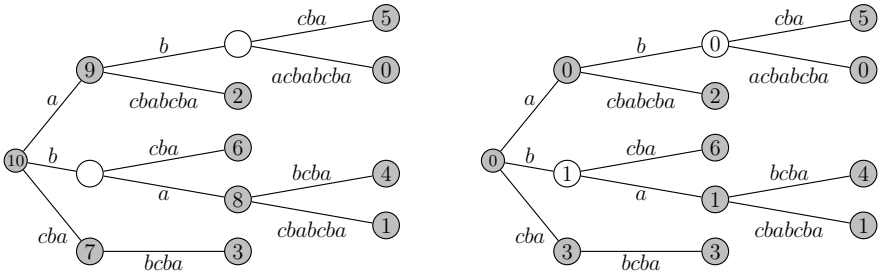
$$c(\mathbf{q}) = (q_0^R, q_0^S, q_0^R, q_1^S, q_1^R, q_2^S, q_2^R, \ldots).$$

We end this section by mentioning the following well-known result (see [3])

Proposition 1. *The Ziv-Lempel an the Crochemore factorizations of a finite word* x *can be computed in linear time.*

Indeed, one first computes the suffix tree of the word x, where each final state is labelled with the position of its suffix in x (see Figure 1(a)). Then, one computes, fore each vertex, the smallest of all positions of the factor corresponding to this vertex. This is done in linear time by a bottom up tree traversal to compute the minimum of all positions of its descendants (see Figure 1(b)).

To compute the Crochemore factorization $c(x) = (x_1, \ldots, x_n)$ of x, assume it is computed up to x_j. One enters the suffix $y = x_j \cdots x_n$ into the suffix tree as far as possible, provided the position red in the suffix tree remains strictly smaller than $|x_1 \cdots x_{j-1}|$. The maximal prefix of y obtained is x_j. This algorithm is clearly linear.



(a) The suffix tree of *abacbabcba* (b) The augmented suffix tree

Fig. 1. The suffix tree and the extended suffix tree of the word *abacbabcba*

3 Crochemore Factorization of Standard Sturmian Words

We recall that a standard Sturmian word is the limit **s** of the sequence

$$s_{-1} = b, s_0 = a, \text{ and } s_n = s_{n-1}^{d_n} s_{n-2},$$

with $\{d_n\}_{n>0}$ a sequence of positive integers. We want now to prove Theorem 1. It is a well known fact that $s_n = p_n \varepsilon_n$ for each n, where p_n is the palindrome word obtained by deleting the last two letters ε_n of s_n, and

$$\varepsilon_n = \begin{cases} ab & \text{if } n \text{ is odd} \\ ba & \text{otherwise.} \end{cases}$$

With an easy induction argument, one can obtain the following result.

Proposition 2. *$s_n = s_{n-1}^{d_n - 1} s_{n-2}^{d_{n-1}} \cdots s_0^{d_1} \varepsilon_n$ for each $n > 0$.*

Since $s_n = p_n \varepsilon_n$ and p_n is a palindrome word, one immediately has the following decomposition of **s** in reverse words.

Proposition 3. $\mathbf{s} = \widetilde{s_0}^{\,d_1} \widetilde{s_1}^{\,d_2} \widetilde{s_2}^{\,d_3} \cdots$.

Lemma 1. *Let w_n be the word $s_n^{d_{n+1}} s_{n-1}^{d_n} s_{n-2}^{d_{n-1}} \cdots s_0^{d_1}$. Then the only occurrences of s_n in w_n are the first $d_{n+1} + 1$ consecutive ones.*

Proof. Since s_n is primitive for every n, s_n is not a proper factor of s_n^2 and we only have to prove that the $(d_{n+1} + 1)$-th occurrence is the last one. We prove it by induction on n. If $n = 2$, $s_2 = (a^{d_1} b)^{d_2} a$ and $w_2 = s_2^{d_3} w_1 = ((a^{d_1} b)^{d_2} a)^{d_3} (a^{d_1} b)^{d_2} a^{d_1}$. Since $(a^{d_1} b)^{d_2}$ occurs in w_1 only as a prefix, we have the assertion. If $n = 3$, $s_3 = ((a^{d_1} b)^{d_2} a)^{d_3} a^{d_1} b$ and $w_3 = s_3^{d_4} w_2$. Since the last occurrence of $aa^{d_1} b$, which is a suffix of s_3, in w_3 is exactly the suffix of the initial $s_2^{d_3} s_1$ in w_2, we have the assertion. Let now suppose the assertion be true for every $2 \leq k < n$. $w_n = s_n^{d_n + 1} w_{n-1}$. By induction hypothesis, there are only two occurrences of $s_{n-1}^{d_n}$ in w_{n-1} and only the first of them is followed by s_{n-2}, so that s_n occurs in w_{n-1} only as a prefix. □

It is now easy to prove Theorem 1.

Proof. **(Theorem 1)** Consider **s** as a composition of reverse words as obtained in Proposition 3. The first 4 factors of the c-factorization can be easily obtained by hand. Notice that the composition of these 4 factors in the c-factorization is exactly $\widetilde{s_0}^{d_1}\widetilde{s_1}^{d_2}$:

$$c(\mathbf{s}) = (\underbrace{a, a^{d_1-1}, b, \widetilde{s}_0^{d_1}\widetilde{s}_1^{d_2-1}}_{\widetilde{s_0}^{d_1}\widetilde{s_1}^{d_2}}, \widetilde{s}_2^{d_3}, \widetilde{s}_3^{d_4}, \ldots).$$

We obtain the result as the limit of the c-factorization on the palindrome prefixes p_n of **s**. Suppose the c-factorization of $p_n = \widetilde{s}_0^{d_1} \cdots \widetilde{s}_{n-2}^{d_{n-1}}\widetilde{s}_{n-1}^{d_n-1}$ to be

$$c(p_n) = (\ldots, \widetilde{s}_{n-2}^{d_{n-1}}, \widetilde{s}_{n-1}^{d_n-1}).$$

By Lemma 1, the only occurrences of $\widetilde{s}_n$ in $\widetilde{w}_n$ are the last $d_{n+1}+1$ consecutive ones and since the first letter of $\widetilde{s}_{n-1}$ is different from the first letter of $\widetilde{s}_n$, we obtain the c-factorization of $p_{n+1} = \widetilde{w}_{n-1}\widetilde{s}_{n-1}^{d_n}\widetilde{s}_n^{d_{n+1}-1}$

$$c(p_{n+1}) = (\ldots, \widetilde{s}_{n-2}^{d_{n-1}}, \widetilde{s}_{n-1}^{d_n}, \widetilde{s}_n^{d_{n+1}-1}).$$ □

4 Crochemore Factorization of Thue-Morse and Period Doubling Sequences

Let **t** be the Thue-Morse word defined as the limit of the morphism τ such that $\tau(a) = ab$ and $\tau(b) = ba$.
We recall the following well known fact.

Lemma 2. *Let w be a factor of* **t** *such that $|w| \geq 4$. Then the occurrences of w in* **t** *begin all in pair positions or all in odd positions.*

Proof. **(Theorem 2)** Let $c(\mathbf{t}) = (c_1, c_2, \ldots)$ be the c-factorization of **t**. We will prove that each Crochemore factor c_{n+1} begins in a pair position, double than that of c_{n-1}, and that $c_{n+1} = \tau(c_{n-1})$ for every $n \geq 8$. One can verify by hands the first step, that is, c_9 begins in the pair position double than that of c_7. Let now c_{n+1} begin in the double position than that of c_{n-1}. Then $\tau(c_{n-1}) \in \text{Prefix}(c_{n+1}c_{n+2}\cdots)$. By definition of Crochemore factor and by Lemma 2, since $|c_{n-1}| > 4$ the factor c_{n-1} occurs earlier in a pair position, so that also $\tau(c_{n-1})$ has an earlier occurrence. Thus, $\tau(c_{n-1}) \in \text{Prefix}(c_{n+1})$.

By contradiction, suppose $\tau(c_{n-1})a \in \text{Prefix}(c_{n+1})$, where by definition of τ a is forced to be the first letter of c_n. Then there is an earlier occurrence of $\tau(c_{n-1})a$ in a pair position, so that also c_{n-1} occurs followed by a letter a earlier than as a Crochemore factor, that is absurd. □

Let now $\mathbf{t^{(m)}}$ be the generalized Thue-Morse word on the m-alphabet $A = \{a_1, \ldots, a_m\}$ defined as the limit of the morphism

$$\tau_m(a_i) = a_i a_{i+1} \cdots a_m a_1 \cdots a_{i-1} \quad (i = 1, \ldots, m).$$

Set $w_{n,i} = \tau_m^n(a_i)$ and let $t_n^{(m)} = \tau_m^n(a_1) = w_{n,1}$. The following result can be easily obtained by induction.

Lemma 3. *$w_{n,i} \notin Fact(w_{n,h}w_{n,k})$ for every i and $h \neq k$.*

The relationship between the structure of the generalized Thue-Morse words and the c-factorization is simpler than what we have obtained for the Thue-Morse word on a 2-letter alphabet. For example, let $m = 3$. Then

$$c(t_3^{(3)}) = (a, b, c, bc, a, ca, b, bcacab, abc, cababc, bca).$$

Using Lemma 3 it is not difficult to prove the following result, which is even stronger than what we stated in Theorem 3.

Theorem 5. *The c-factorization of $t_n^{(m)}$ is* $(w_{1,1}, \ldots, w_{1,m}, w_{1,2} \cdots w_{1,m}, w_{1,1}, \ldots, w_{1,m} \cdots w_{1,1}, w_{1,m-1}, \ldots, w_{n,2} \cdots w_{n,m}, w_{n,1}, \ldots, w_{n,m} \cdots w_{n,1}, w_{n,m-1})$.

In the case $m = 3$ we have

$$t_3^{(3)} = \overbrace{\underbrace{abc}_{w_{2,1}}\underbrace{bca}_{w_{2,2}}\underbrace{cab}_{w_{2,3}}}^{t_2^{(m)}}\underbrace{bcacababc}_{w_{2,2}w_{2,3}w_{2,1}}\underbrace{cababcbca}_{w_{2,3}w_{2,1}w_{2,2}},$$

in accordance with the c-factorization given above.

In the general case also, this factorization reflects exactly the decomposition of each prefix $t_n^{(m)}$ in $w_{h,i}$ words:

$$\underbrace{(\underbrace{w_{1,1}, \ldots, w_{1,m}}_{t_1^{(m)}=w_{2,1}}, \underbrace{w_{1,2} \cdots w_{1,m}, w_{1,1}}_{w_{2,2}}, \ldots, \underbrace{w_{1,m} \cdots w_{1,1}, w_{1,m-1}}_{w_{2,m}}, \ldots)}_{t_2^{(m)}}$$

Recall that the period doubling sequence is defined as the limit $\mathbf{q}$ of the morphism δ such that $\delta(a) = ab$ and $\delta(b) = aa$. In order to prove Theorem 4, we begin by providing some easy to prove results on the structure of $\mathbf{q}$.

Lemma 4. *The following facts hold:*

(i) $q_{n+1} = q_n q_n'$, *where if* $q_n = va$, $q_n' = v\bar{a}$.

(ii) $q_n = p_n u_n$, *where p_n is a palindrome word such that* $p_0 = \varepsilon$, $p_{n+1} = p_n u_n p_n$, *and*

$$u_n = \begin{cases} a & n \text{ pair} \\ b & n \text{ odd} \end{cases}$$

(iii) $p_n = q_0^R q_1^R \cdots q_{n-1}^R$.

Lemma 5. *For every $n \geq 1$, p_n occurs only as a prefix and as a suffix in p_{n+1}.*

Proof. True for $n = 1, 2$. Suppose the lemma true for $n \leq k$.

$$p_{k+2} = p_k u_k p_k u_{k+1} p_k u_k p_k$$

and $p_{k+1} = p_k u_k p_k$. Since p_{k+1} is different from the central factor of p_{k+2} $p_k u_{k+1} p_k$ and by the induction hypothesis on p_k, we have the assert. □

Proof. **(Theorem 4)** We prove that

$$c(q_{n+2}) = (q_0^R, q_0^S, q_0^R, q_1^S, q_1^R, \ldots, q_n^S, q_n^R, u(n)).$$

The assertion holds for $n = 2$. Suppose it is true for q_{n+1}.
By Lemma 4,

$$q_{n+2} = q_{n+1} q'_{n+1} = p_{n+1} u_{n+1} p_n u_n p_n u_n. \tag{1}$$

Note that $u_{n+1} p_n = q_n^S$ and $u_n p_n = q_n^R$.
By induction hypothesis, the last Crochemore factor of q_{n+1} is its last letter, so that the u_{n+1} of Equation (1) is the first letter of a Crochemore factor of q_{n+2}.

With the further expansion of the expression of q_{n+2}

$$q_{n+2} = p_{n-1} u_{n-1} p_{n-1} u_n p_{n-1} u_{n-1} p_{n-1} \underbrace{u_{n-1} p_{n-1} u_{n-1} p_{n-1}}_{q_n^S} \underbrace{u_n p_{n-1} u_{n-1} p_{n-1}}_{q_n^R} u_n$$

we can verify that the factors $u_{n+1} p_n$ and $u_n p_n$ of Equation (1) (here underlined in the same occurrences) occurred already before in q_{n+2}. Moreover, by using Lemma 5 it can be shown that they are not contained into larger already occurred factors, that is, $u_{n+1} p_n u_n$ and $u_n p_n u_n$ did not occur before, so that q_n^S, q_n^R, u_n are exactly the c-factors we need to add to those of q_{n+1} to complete the c-factorization of q_{n+2}. □

5 Crochemore Factorization Versus Ziv-Lempel Factorization

Ziv and Lempel have considered several variations of factorizations of words (see [7]; these are also discussed in [6]). We illustrate the relation between Crochemore and Ziv-Lempel factorizations by stating some simple facts, and by giving some examples.

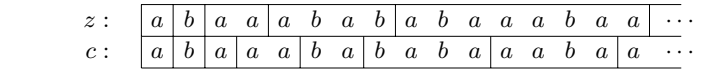
Lemma 6. *Let $(c_1, c_2, \ldots)$ and $(z_1, z_2, \ldots)$ be the Crochemore and the Ziv-Lempel factorizations of a word w. The following hold.*

- *For each i, j such that $|c_1 \cdots c_{i-1}| \geq |z_1 \cdots z_{j-1}|$ and $|c_1 \cdots c_i| < |z_1 \cdots z_j|$, then $|z_1 \cdots z_j| = |c_1 \cdots c_i| + 1$.*
- *For each i, j such that $|z_1 \cdots z_{j-1}| < |c_1 \cdots c_i| \leq |z_1 \cdots z_j|$ then $|c_1 \cdots c_{i+1}| \leq |z_1 \cdots z_{j+1}|$.*

Lemma 6 reflects the fact that by their definitions if a Ziv-Lempel factor includes a Crochemore factor, then it ends at most a letter after, and that a Crochemore factor cannot include a Ziv-Lempel factor. Accordingly, we have the following result.

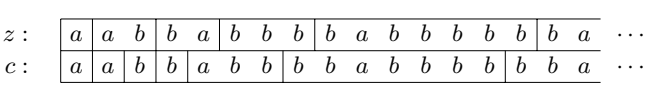
Proposition 4. *The number of factors of the Crochemore factorization is at most twice the number of factors of the Ziv-Lempel factorization.*

Consider for example the period doubling sequence. It is simple to show that each Ziv-Lempel factor of **q** properly includes a Crochemore factor by ending just a letter before, as illustrated in this figure:



In this example each Ziv-Lempel can therefore be associated to a couple of Crochemore factors. On the contrary, in the next example each Ziv-Lempel factor is associated to a sole factor.

Consider the word $\mathbf{v} = aabbabbbbabbbbbba\cdots$ in which for each $i \geq 0$ the letter in position i is defined as a if i is a perfect square and b otherwise. The Ziv-Lempel factors of **v** are shifted one letter ahead with respect to the Crochemore factors, as illustrated in the figure:



6 Conclusion

In the examples given here, the detailed knowledge of the structure of the infinite words yields enough information in order to compute the Crochemore factorization. Similar results hold for episturmian words. On the contrary, it is not yet clear whether a satisfactory description can be obtained for automatic sequences other than those which are uniform purely morphic sequences.

References

1. J.-P. Allouche and J. Shallit, *Automatic Sequences*, Cambridge University Press, 2003.
2. M. Crochemore, Recherche linéaire d'un carré dans un mot, *Comptes Rendus Sci. Paris Sér. I Math.*, 1983, **296**, 781–784.
3. M. Crochemore, C. Hancart, and T. Lecroq, *Algorithmique du texte*, Vuibert, 2001.
4. M. Crochemore and W. Rytter, *Text Algorithms*, The Clarendon Press Oxford University Press, 1994.
5. A. de Luca, A division property of the Fibonacci word, *Information Processing Letters*, 1995, **54**, 307–312.
6. R. Kolpakov and G. Kucherov, Periodic structures on words, in Lothaire, *Applied Combinatorics on Words*, Cambridge University Press, 2005.
7. A. Lempel and J. Ziv, On the complexity of finite sequences, *IEEE Transactions in Information Theory*, 1976, **IT-22**, 75–81.
8. Z.-X. Wen and Z.-Y. Wen, Some properties of the singular words of the Fibonacci word. *European Journal of Combinatorics* 1994, **15**, 587–598.