Reconstructing words from a fixed palindromic length sequence*

Alexandre Blondin Massé[1], Srečko Brlek[1], Andrea Frosini[2], Sébastien Labbé[1], and Simone Rinaldi[2]

[1] Laboratoire de Combinatoire et d'Informatique Mathématique,
Université du Québec à Montréal,
C. P. 8888 Succursale "Centre-Ville", Montréal (QC), CANADA H3C 3P8,
Brlek.Srecko@uqam.ca, [blondin_masse.alexandre, Labbe.Sebastien]@courrier.uqam.ca

[2] Dip. di Scienze Matematiche ed Informatiche Roberto Magari,
Università degli Studi di Siena
Pian dei Mantellini 44, 53100 Siena, Italy
[frosini, rinaldi]@unisi.it

Abstract. To every word w is associated a sequence G_w built by computing at each position i the length of its longest palindromic suffix. This sequence is then used to compute the palindromic defect of a finite word w defined by $D(w) = |w| + 1 - |\mathrm{Pal}(w)|$ where $\mathrm{Pal}(w)$ is the set of its palindromic factors. In this paper we exhibit some properties of this sequence and introduce the problem of reconstructing a word from G_w. In particular we show that up to a relabelling the solution is unique for 2-letter alphabets.

Key words: Palindromic complexity, defect, lacunas, reconstruction.

1 Introduction

Among the many ways of measuring the information content of a finite word, counting the number of its distinct factors or subwords of given length has been widely used and known as its complexity. A refinement of this notion amounts to restrict the factors to palindromes. The motivations for the study of palindromic complexity comes from many areas ranging from the study of Schrödinger operators in physics [4, 7, 20] to number theory [6] and combinatorics on words where it appears as a powerful tool for understanding the local structure of words. It has been recently studied in various classes of infinite words, an account of which may be found in the survey provided by Allouche et al. [5].

In particular, the palindromic factors give an insight on the intrinsic structure, due to its connection with the usual complexity, of many classes of words. For instance, they completely characterize Sturmian words [23], and for the class of smooth words they provide a connection with the notion of recurrence [12, 13].

* with the support of NSERC (Canada)

The problem of reconstructing words from partial information arise naturally. We mention a few of them that have been solved for a fixed alphabet Σ:

Some set A of factors is fixed. Find the shortest words containing the set A of all factors of given length k. This leads to the De Bruijn sequences [15, 17, 19] whose construction uses a graph G_k where vertices are the given words of length k, and where edges model the scanning of the word by a window of size k. The solution is then obtained by computing all Eulerian cycles in the graph. It is worth noting that finding the lexicographically smallest such word is much easier: it is given by the lexicographic concatenation of Lyndon words on Σ whose lengths are divisible by k (See Fredericksen et al. [18]).

Some set A is fixed along with some suitable hypothesis. Construct all words w such that the set of its factors $A = F(w)$. The technique used for this problem is based on constructing a set of minimal forbidden words, that is the extensions of words in A that do not belong to A [9]. That technique was also used in [11] to construct words whose language of palindromes is a fixed set P. It turns out that it is a rational language. Concerning multisets of subsequences, instead of factors we mention a general result. If the set A contains sufficiently many subsequences of length k, then the solution is unique [26]: indeed, for a word w of lengh $n > 7$ and $k \geq [n/2]$ the subsequences uniquely determine w, and for $k < \log_2 n$ they do not. See also an interesting combinatorial approach depending on the Burrows-Wheeler transform (See Mantaci et al. [25]).

Fixed complexity. The most famous example is that of Sturmian words (see Lothaire [22] for a substantial review) which are characterized by the complexity $P(n) = n + 1$ established by M. Morse [28]. Sturmian words are the discretization of lines with irrational slopes, and they are easily constructed from the continued fraction expansion corresponding to the irrational slope. The complexity is therefore not enough to characterize completely a word. However, in the case of the Thue-Morse complexity [10, 24], there are essentially only two such words [1, 2].

In this paper we introduce the problem of reconstructing a word from sequences describing its palindromic complexity. Droubay, Justin and Pirillo [16] noted that the palindrome complexity $|\mathrm{Pal}(w)|$ of a word w is bounded by $|w| + 1$, and observed that it is computed by a sequential algorithm listing the first occurrences of longest palindromic suffixes, called *unioccurrent* in [16]. For our study we need the following two auxiliary functions on words. Given a word of length n, $w : [0..(n-1)] \longrightarrow \Sigma$, we define two functions $G_w, H_w : \mathbb{N} \longrightarrow \mathbb{N}$ by $G_w(i) = |\mathrm{LPS}(w[0..i])|$ and

$$H_w(i) = \begin{cases} G_w(i) & \text{if it is the first occurrence of } \mathrm{LPS}(w[0..i]) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We first exhibit some combinatorial properties of the palindromic factors in words (Section 3) and use them in order to obtain properties of the sequences

G and H (Section 4). Finally we study the problem of reconstructing words from given sequences, and establish conditions for unicity on 2-letter alphabets.

2 Preliminaries

In what follows, Σ is a finite *alphabet* whose elements are called *letters.* By *word* we mean a finite sequence of letters $w : [0..(n-1)] \longrightarrow \Sigma$, where $n \in \mathbb{N}$. The length of w is $|w| = n$ and $w[i]$ or w_i denote its i-th letter. The set of n-length words over Σ is denoted Σ^n. By convention, the *empty* word is denoted ε and its length is 0. The free monoid generated by Σ is defined by $\Sigma^* = \bigcup_{n\geq 0} \Sigma^n$.

The set of right infinite words is denoted by Σ^ω and we set $\Sigma^\infty = \Sigma^* \cup \Sigma^\omega$. Given a word $w \in \Sigma^\infty$, a *factor* f of w is a word $f \in \Sigma^*$ satisfying

$$\exists x \in \Sigma^*, \exists y \in \Sigma^\infty, w = xfy.$$

If $x = \varepsilon$ (resp. $y = \varepsilon$) then f is called *prefix* (resp. *suffix*). The set of all factors of w is denoted by $\text{Fact}(w)$, those of length n is $\text{Fact}_n(w) = \text{Fact}(w) \cap \Sigma^n$, and $\text{Pref}(w)$ is the set of all prefixes of w. The number of occurrences of a factor f in w is denoted $|w|_f$. A *period* of a word w is an integer $p < |w|$ such that $w[i] = w[i+p]$, for all $i < |w| - p$. If $w = pu$, with $|w| = n$ and $|p| = k$, then $p^{-1}w = w[k..(n-1)] = u$ is the word obtained by erasing p. A word is said to be *primitive* if it is not a power of another word. Two words u and v are *conjugate* when there are words x, y such that $u = xy$ and $v = yx$. The conjugacy class of a word w is denoted by $[w]$; note that the length is invariant under conjugacy. For a given word w of length n, any of its conjugates is obtained by cyclic permutation, that is $\sigma^i(w) = w[i..(n-1)]w[0..(i-1)]$.

The *reversal* of $u = u_0u_1 \cdots u_{n-1} \in \Sigma^n$ is the word $\widetilde{u} = u_{n-1}u_{n-2} \cdots u_0$, and a *palindrome* is a word p such that $p = \widetilde{p}$. Since every word contains palindromes, the letters and ε being necessarily part of them, the set of its palindromic factors is $\text{Pal}(w)$, and its *palindromic complexity* is denoted by $|\text{Pal}(w)|$. Conjugacy is an equivalence relation having numerous properties and for our purpose we need the following one easily obtained by induction: let p and q be two palindromes, then $\sigma^i(pq) = p'q'$, for some palindromes p' and q'. We start by quoting Lemma 1 of [8] in order to establish a useful combinatorial property.

Lemma 1 (Blondin Massé et al. [8]) *Assume that $w = xy = yz$. Then for some u, v, and some $i \geq 0$ we have from* [21]

$$x = uv, y = (uv)^i u, z = vu; \tag{2}$$

and the following conditions are equivalent :

(i) $x = \widetilde{z}$;
(ii) *u and v are palindromes;*

(iii) *w is a palindrome;*
(iv) *xyz is a palindrome.*

Moreover, if one of the equivalent conditions above holds then

(v) *y is a palindrome.*

As a consequence we have the following proposition.

Proposition 1 *Assume that $w = xp = qz$ where p and q are palindromes such that $|q| > |x|$. Then w has period $|x| + |z|$, and $x\widetilde{z}$ is a product of two palindromes.*

Proof. Since $|q| > |x|$, there exists a non-empty word y such that $q = xy$ and $p = yz$. It follows that

$$w\,\widetilde{x} = q\,z\,\widetilde{x} = x\,y\,z\,\widetilde{x} = x\,p\,\widetilde{x} = x\,\widetilde{p}\,\widetilde{x} = x\,\widetilde{z}\,\widetilde{y}\,\widetilde{x} = x\,\widetilde{z}\,\widetilde{q} = x\,\widetilde{z}\,q.$$

Considering $qz\widetilde{x} = x\widetilde{z}q$, we obtain from Equation (2) that $|x\widetilde{z}|$ is a period of $w\widetilde{x}$. From Lemma 1 (iii), there exist palindromes u, v such that $x\widetilde{z} = uv$. □

In order to compute the palindromic complexity we need the function LPS : $\Sigma^* \longrightarrow \Sigma^*$ which associates to any word w its longest palindromic suffix LPS(w).

Droubay, Justin and Pirillo [16] noted that the palindrome complexity $|\text{Pal}(w)|$ of a word w is bounded by $|w| + 1$, and that finite Sturmian (and even episturmian) words realize the upper bound. Moreover they implicitly show that the palindrome complexity is computed by an algorithm listing the longest palindromic suffixes which amounts to compute for a word w the functions $G_w, H_w : \mathbb{N} \longrightarrow \mathbb{N}$ defined by

$$G_w(i) = |\text{LPS}(w[0..i])|;$$

$$H_w(i) = \begin{cases} G_w(i) & \text{if it is the first occurrence of LPS}(w[0..i]); \\ 0 & \text{otherwise.} \end{cases}$$

We often omit the subscript w in G_w and H_w when the context is clear. As an example let $w = aababbaababaaabaab$. Then we have the following table :

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
w	a	a	b	a	b	b	a	a	b	a	b	a	a	a	b	a	a	b
G	1	2	1	3	3	2	4	2	4	3	3	5	7	3	5	7	5	4
H	1	2	1	3	3	2	4	0	4	0	0	5	7	3	5	7	5	0

A position in the word w where H vanishes is called a *lacuna* in [8]. For instance the set of lacunas for w in the example above is $\{7, 9, 10, 17\}$. Equivalent words, that is words obtained by relabelling of the alphabet, have obviously the same functions G and H. For instance, on the 2-letter alphabet $\{a, b\}$, we have $G_w = G_{\overline{w}}$ and $H_w = H_{\overline{w}}$, where $\overline{(\ )}$ is the morphism defined by $\overline{a} = b, \overline{b} = a$.

The palindromic defect of a finite word w is defined in Brlek et al. [11] by $D(w) = |w| + 1 - |\mathrm{Pal}(w)|$, and words for which $D(w) = 0$, that is, such that H does not vanish for any index are called *full.* In that paper it is also shown that there exist periodic full words, and an optimal algorithm is provided to check if an infinite periodic word is full or not. Moreover, a characterization by means of a rational language is given for the language L_P of words whose palindromic factors belong to a fixed and finite set P of palindromes.

3 Properties of the functions *G* and *H*

First observe that a word w is full if and only if $G_w = H_w$. Now we describe the shortest words having a fixed defect value d. For instance, on a 2-letter alphabet, the shortest words having one lacuna, i.e. when $d = 1$, are

$$w_1 = aababbaa, w_2 = aabbabaa, w_3 = bbabaabb \text{ and } w_4 = bbaababb.$$

Observe that this set is closed under reversal ($w_1 = \widetilde{w_2}; w_3 = \widetilde{w_4}$) and complementation ($w_1 = \overline{w_3}; w_2 = \overline{w_4}$). On the other hand, one of the shortest words having two lacunas is the following.

i	0	1	2	3	4	5	6	7	8	9
w	b	a	a	b	a	b	b	a	a	b
G	1	1	2	4	3	3	2	4	2	4
H	1	1	2	4	3	3	2	4	0	0

The example above extends to the infinite periodic word $W = (baab.ab)^\omega$

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...
W	b	a	a	b	a	b	b	a	a	b	a	b	b	a	a	b	a	b	...
G	1	1	2	4	3	3	2	4	2	4	3	3	2	4	2	4	3	3	...
H	1	1	2	4	3	3	2	4	0	0	0	0	0	0	0	0	0	0	...

where $baab.ab$ is not the product of two palindromes, so that $|\mathrm{Pal}(W)|$ is finite by virtue of a previous result (see Theorems 4 and 6 in [11]). More generally, we have the following result.

Proposition 2 *Let $M(k, d)$ be the length of a shortest word on a k-letter alphabet Σ having defect d, we have :*

$$M(k, d) = \begin{cases} 8 & \text{if } k = 2, d = 1, \\ d + 8 & \text{if } k = 2, d \geq 2, \\ d + k & \text{if } k \geq 3. \end{cases}$$

Proof. The first two cases follow from the observations above. For $k \geq 3$, let w be a word such that $|w| = M(k, d)$. Since every letter occurs in w and w

has defect value d, we have $M(k,d) \geq d+k$. Now, consider the infinite periodic word $w = (\alpha_1\alpha_2 \cdots \alpha_k)^\omega$, where α_i is a letter. Observe that $\mathrm{Pal}(w) = \Sigma$, so that each prefix of length $n \geq k+1$ has defect value $n-k$. Hence $M(k,d) = d+k$ for $k \geq 3$. □

Lemma 2 *Let w be a nonempty word, and let $W = w^\omega$. Then we have*

(i) $G_w(0) = 1$*, and if* $H_w(i) = 1$ *then* $w[i]$ *is the first occurrence of a letter;*
(ii) *if* $w = pq$ *is primitive with* $p, q \in \mathrm{Pal}(\Sigma^*)$ *then* $\lim_{n \longrightarrow \infty} G_W(n) = \infty$*;*
(iii) *if* w *is not the product of two palindromes then* G_W *is eventually periodic.*

Proof. (i) Obvious. (ii) In this case by Theorem 4 of [11] the palindromic language of W is infinite. Since for all $k \geq 0, (pq)^k p$ is a palindromic prefix of W, there are infinitely many palindromic prefixes of W. Moreover, we have $G_W(i) = G_W(i - |w|) + |w|$ for $i \geq 2|w|$.

(iii) Here again by Theorem 4 of [11], the palindromic language of W is finite. Therefore, let u be the shortest prefix of W containing all the palindromes, and let k be the smallest integer such that $u \in \mathrm{Pref}(w^k)$ then we have $G_W(i) = G_W(i + k|w|)$. □

Examples. Let $W = (abc)^\omega$, whose palindromic language is $P = \{a, b, c\}$ taken from [11] (Section 3). Then we have the following values for G and H:

i	0	1	2	3	4	5	6	7	8	
W	a	b	c	a	b	c	a	b	c	...
G	1	1	1	1	1	1	1	1	1	...
H	1	1	1	0	0	0	0	0	0	...

Here are some typical periodic words with their characteristic functions:

W	G_W
a^n	$[1, 2, 3, 4, 5, \cdots]$
$a.b^n$	$[1, 1, 2, 3, 4, 5, \cdots]$
$(ab)^n$	$[1, 1, 3, 3, 5, 5, 7, 7, 9, 9, \cdots, (2n+1), (2n+1), \cdots]$

Moreover they are all full, since G and H coincide.

Another periodic example illustrating Lemma 2 (ii) is $W = (aba.cbc)^\omega$. Its palindromic language is infinite and W has infinitely many palindromic prefixes, and we have

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
W	a	b	a	c	b	c	a	b	a	a	b	a	c	b	c	a	b	a	c	b	c	...
H	1	1	3	1	0	3	5	7	9	5	7	9	11	13	15	11	13	15	17	19	21	...

Observe also that there are non periodic words U such that G_U is periodic. Indeed, take any nonperiodic word, for instance the Fibonacci word defined as

$$F = \varphi^{\omega}(a) = abaababaabaabab\cdots, \quad \text{where} \quad \varphi(a) = ab; \varphi(b) = a.$$

Define the morphism $\theta : \{a, b\} \longrightarrow \{a, b, c, d\}^*$ by $a \mapsto abcd; b \mapsto acbd$. Then the word $W = \theta(F)$ is nonperiodic, but $G_W = (1111)^{\omega}$. Nevertheless we have the following result showing a local periodical behaviour.

Lemma 3 *Let $w \in \Sigma^*$. If there exists i such that $G(i) = G(i + k) = l$, with $l \geq k$, then the factor $f = w[(i - l + 1)..(i + k)]$ has period $2k$, and any factor of length $2k$ of f is the product of two palindromes.*

Proof. Assume that q and p are the longest palindromic suffixes of length l at positions respectively i and $i + k$. Then there exist x and z such that $f = qz = xp$. In the case $l = k$, we have $f = qp$ and the claim is true. If $l > k$ there exists a non-empty word y such that $q = xy$ and $p = yz$. It follows from Proposition 1 that $2|x|$ is a period of f, and $x\widetilde{z}$ is a product of two palindromes. Therefore, any factor of length $2k$ is the product of two palindromes since it is a conjugate of $x\widetilde{z}$. □

The function G satisfies the following properties

Proposition 3 *For any finite word $w \in \Sigma^*$, the following properties hold :*

(i) $G(i) \leq \max\{|p| : p \in \mathrm{Pal}(w[0..i])\} \leq i + 1$
(ii) $G(j) \leq G(i) + 2(j - i)$, *for all* $j \geq i$;
(iii) $G(i{+}1) = G(i) \implies G(i)$ *and* $G(i{+}1)$ *are odd, and* $G(i{+}2) \in \{G(i){+}2, 2\}$;
(iv) $G(i + 1) = G(i) + 1 \implies \mathrm{LPS}(w[0..i]) = \alpha^{G(i)+1}$, *for some* $\alpha \in \Sigma$;

Proof. (i) is obvious. (ii) First, note that $G(i + 1) \leq G(i) + 2$ since the longest palindromic suffix at position $i + 1$ contains a palindrome of length $G(i + 1) - 2$ ending at position i. The result follows by induction.

(iii) Follows from Lemma 3. (iv) Let p and q be the respective palindromes at positions i and $(i + 1)$. Then we have $q = p\alpha$ for some $\alpha \in \Sigma$, and we conclude by using Proposition 1. □

Lemma 4 *Let $i \leq k$. If $G(k) = G(i) + 2(k - i)$, then $G(j) = G(i) + 2(j - i)$ for all $i \leq j \leq k$.*

Proof. $G(k) - 2(k - j) \leq G(j) \leq G(i) + 2(j - i)$ and the left term is equal to

$$G(i) + 2(k - i) - 2(k - j) = G(i) + 2(j - i). \quad \square$$

The next proposition is obtained by adapting the proof of Proposition 3.

Proposition 4 *For any finite word $w \in \{a, b\}^*$, the function H satisfies*

(i) $H(i + 1) - H(i) \leq 2$;
(ii) $H(i + 1) = H(i) \implies H(i + 1)$ *and* $H(i)$ *are both odd;*
(iii) $H(i) \leq \max\{|p| : p \in \mathrm{Pal}(w)\}$;
(iv) *if* $H([i..(i + k + 2)]) = [n, 0, \cdots, 0, m]$ *for some* i, *then* $m < n + 2k$.

4 Reverse engineering the functions G and H

Here we tackle the following problems. Given a (finite or infinite) sequence s of integers, does there exist a word w such that $H_w = s$ or $G_w = s$? If such a word w exists, under which conditions is it unique up to permutation of the letters ?

We say that a finite/infinite sequence s is *G-consistent* (resp. *H-consistent*) on Σ if there is at least one nonempty word $w \in \Sigma^\infty$ such that for all i, $G_w(i) = s[i]$ (resp. $H_w(i) = s[i]$). If there is only one such word (up to permutation of the letters) then s is said to be *unambiguous*. A first simple result follows:

Proposition 5 *Let Σ be an alphabet of at most 3 letters. Then any G-consistent sequence on Σ is unambiguous.*

Proof. Let s be a G-consistent sequence. We proceed by induction on the length of s. Then $s[0] = 1$ so that the base of the induction is trivially satisfied by choosing one letter in Σ. Assume that $s[0..i]$ is unambiguous. Then there exist a word w, such that $G_w[0..i] = s[0..i]$. Two cases arise:

(a) $s[i+1] > 1$: we set $w[i+1] = w[i+2-s[i+1]]$.
(b) $s[i+1] = 1$: if $|s|_1 = 2$ then $|\Sigma| = 2$, so that $w[i+1] \in \Sigma \setminus \{w[0]\}$.
If $|s|_1 > 2$ then $|\Sigma| = 3$ and we have to consider two cases:
- if $|s[0..i]|_1 = 2$, then we set $w[i+1]$ to the remaining letter;
- if $|s[0..i]|_1 > 2$, then $w[0..i] = p\gamma\beta^k\alpha^l$ where $\Sigma = \{\alpha, \beta, \gamma\}$, $p \in \Sigma^*$, and $k, l \geq 1$, and we set $w[i+1] = \gamma$. □

Observe that for larger alphabets, that is when $|\Sigma| > 3$, G-consistent sequences are not necessarily unambiguous, as shown in the following examples.

Example. Let $\Sigma = \{a, b, c, d\}$ and consider the sequence $s = [1, 1, 1, 3, 2, 1, 3, 5]$. There is a unique word $w[0..4]$ which is G-consistent with $s[0..4]$:

i	0	1	2	3	4	5	6	7
s	1	1	1	3	2	1	3	5
w	a	b	c	b	b	**a**	b	b
w'	a	b	c	b	b	**d**	b	b

while two different words are consistent with $s[0..5]$, a fact that follows from Lemma 2(i).

One can easily see that the previous ambiguity is related with the presence of more than three 1's in the sequence s. However here is a word w having four occurrences of 1, but uniquely determined by G as well.

Example. Let $\Sigma = \{a, b, c, d\}$ and let $s = [1, 1, 1, 3, 1, 3, 5, 7, 9]$. There is a unique word which is G-consistent with s:

i	0	1	2	3	4	5	6	7	8
s	1	1	1	3	1	3	5	7	9
w	a	b	c	b	d	b	c	b	a

The situation is clearly explained by the following statement

Proposition 6 *Let s be a G-consistent sequence. If there exist two distinct words w, w' consistent with s, then there exists i such that $G_w(i) = G_{w'}(i) = 1$ and $H_w(i) = 0$ or $H_{w'}(i) = 0$.*

Proof. Indeed, if $s[i] = 1$, then $w[i]$ is either a new letter or, a previously encountered letter such that the longest palindromic LPS($w[0..i]$) is the letter itself. □

Consider now the same problem for the function H. Since the functions G and H coincide for full words, we have immediately the next result.

Corollary 1 *Any full word (thus any Sturmian word) is uniquely determined by the function H.*

So, the function G_w encodes all the information on w, but this is no longer true for the function H_w. Indeed, there exist H-consistent sequences that are not unambiguous as shown in the following example: consider the word $w = abbabbbabaabb$. Then, we have

$$H_{wa} = H_{wb} = (1, 1, 2, 4, 3, 5, 3, 5, 7, 3, 2, 4, 0, 0) \tag{3}$$

but $wa \neq wb$.

Observe that the counterpart of Proposition 6 does not hold for the function H. Indeed, every 1 in the sequence $s = H_w$ corresponds necessarily to a new letter in w. Consequently the presence of 1's does not cause ambiguity, and $|s|_1 = |\Sigma|$, as shown below for a 5-letter alphabet $\{a, b, c, d, e\}$.

i	0	1	2	3	4	5	6	7	8
s	1	1	2	1	3	5	1	3	5
w	a	b	b	**c**	b	b	**e**	b	b
w'	a	b	b	**d**	b	b	**c**	b	b

We point out that w and w' are the same word up to a relabelling of the letters.

For words which are not full, the study of the H function is more complex. However, there are some special conditions ensuring that an H-consistent sequence s on a given Σ is also unambiguous.

Proposition 7 *Let s be an H-consistent sequence such that $s = s_1 0^k m s_2$ with $m \neq 0$, and s_1 does not contain any 0. If $m > 2k+1$ then there is a unique word $w[0..(|s_1| + k)]$ such that $H_w = s[0..(|s_1| + k)]$.*

The proof is similar to that of proof of Proposition 5. As an example, consider the following H-consistent sequence on $\Sigma = \{a, b\}$

i	0	1	2	3	4	5	6	7	8	9	10	11
s	1	2	1	2	4	6	4	3	3	0	0	6
w	a	a	b	b	a	a	b	a	b	x	y	z

where $s_1 = [\,1, 2, 1, 2, 4, 6, 4, 3, 3]$, $m = 6$, and $k = 2$. The last three elements of w can be uniquely determined since the factor $b\,a\,b\,x\,y\,z$ has to be a palindrome, that is $x = z = b$, and $y = a$.

Note that the bound k on the length of a subsequence of 0's in s given in Proposition 5 does not depend on the cardinality of Σ. On the other hand, observe that if $|\,\Sigma\,| = 2$ and $k = 1$ the sequence s is still uniquely determined. For instance, consider the sequence $s[0..12] = [1, 2, 1, 3, 3, 2, 4, 6, 5, 3, 5, 0, 3]$, with $\Sigma = \{a, b\}$, and therefore $m = 3$:

i	0	1	2	3	4	5	6	7	8	9	10	11	12
s	1	2	1	3	3	2	4	6	5	3	5	0	3
w	a	a	b	a	b	b	a	b	b	b	a	a	a

Here the cardinality of the alphabet allows only one possible choice of the letter consistent with $H_w(11) = 0$, consequently $m \not> 2k + 1 = 3$, but the word w is uniquely determined as well.

4.1 Infinite words

In the case of infinite words, the situation is similar and ambiguous H can also occur. We start by recalling some facts. From [8, 11], we know that, when analyzing the defect and the lacunas of an infinite word, it can present

(a) an infinite palindromic complexity with a finite number of lacunas;
(b) a finite palindromic complexity with an infinite number of lacunas;
(c) both infinite palindromic complexity and number of lacunas.

In general, in none of the three cases the function H is unambiguous, as it is shown in the following examples.

Case (a): consider two words U and V having the same prefix of length 23

$$u_1 = a\,b\,b\,a\,a\,b\,b\,a\,b\,a\,b\,a\,a\,a\,a\,b\,a\,b\,b\,b\,b\,a\,a,$$

and such that $U = u_1\,a\,(\,a\,b\,)^{\omega}$ and $V = u_1\,b\,(\,b\,a\,)^{\omega}$. The two sequences $H_U([0..22])$ and $H_V([0..22])$ are equal since they share a common prefix. Now, since the suffix parts of U and V starting at position 23 satisfy $\overline{U[\geq 23]} = V[\geq 23]$, we have $H_U = H_V$. Since, both u and v are eventually periodic, and since their period is the product of two palindromes, the palindromic complexity of both u and v is infinite. Finally, an easy check reveals that the suffix sequence of the function H, for $n > 2$, is

$$H_U([\geq 23]) \;=\; (0,0,0,0,0,0,0,7,7,9,9,11,11,\ldots,(2n+3),(2n+3),\ldots).$$

Case (b): let $w = abbabbbabaabb$, already used in Equation (3), and consider the words U and V defined as follows, by means of the word w:

$$U = w \,\cdot\, ab \,\cdot\, bbaaa \,\cdot\, (baabba)^{\omega}\,,$$

$$V = w \,\cdot\, ba \,\cdot\, bbaaa \,\cdot\, (baabba)^{\omega}\,.$$

The sequences H_U and H_V coincide and are

$$\begin{aligned} H_U = H_V = (&1,1,2,4,3,5,3,5,7,3,2,4,0,0,0,0,0,0,0,3,5,0,5,0,0,0,\\ &0,0,0,0,0,0,0,0,\cdots) \end{aligned}$$

All the terms after position 22 are equal to 0 since the words U and V are eventually periodic, with a period which is not the product of two palindromes.

Case (c): finally, consider the words U and V defined as follows (using again $w = abbabbbabaabb$):

$$U = w \cdot ab \cdot bbaaa \cdot baab \cdot baabba \cdot (baab)^2 \cdot baabba \ldots (baab)^n \cdot baabba \ldots$$

$$V = w \cdot ba \cdot bbaaa \cdot baab \cdot baabba \cdot (baab)^2 \cdot baabba \ldots (baab)^n \cdot baabba \ldots.$$

The sequences H_U and H_V coincide and their first terms are

$$\begin{aligned} H_U = H_V = (&1,1,2,4,3,5,3,5,7,3,2,4,0,0,0,0,0,0,0,3,5,0,5,0,0,0,\\ &6,8,6,8,0,0,0,0,0,0,0,0,0,0,10,12,10,12,0,0,0,0,0,0,\\ &0,0,0,0,0,0,0,0,14,16,14,16,0,0,\ldots). \end{aligned}$$

The two sequences have an infinite number of new palindromes since the palindromic factor $baab$ is repeated an increasing number of times at each step. At the same time the set of lacunas is infinite since the factor $baabba$, which is not the product of two palindromes, occurs infinitely many times.

5 Further work

The problem of reconstructing words from the functions G and H leads to many interesting developments, some of them requiring a deeper analysis in order to produce efficient decision algorithms.

Consistency. Deciding if a given finite sequence s of numbers is G-consistent (resp. H-consistent) may be easily achieved. Indeed, let $k = |s|_1$. This implies that the smallest alphabet Σ we have to consider contains at most k letters (exactly k for H-consistency, by virtue of Lemma 2 (i)). Taking an order on the letters of Σ permits to restrict the study to classes of words equivalent under permutations of letters. A close look to the proof of Proposition 5 reveals all the information in order to construct sequentially all words consistent with s: indeed, it suffices to check at each position i, if $\mathrm{LPS}(w[0..i]) = s[i]$.

Random and exhaustive generation. The algorithms described above may be used for constructing trees of words. Indeed, at each step i one constructs a trie of words having height $i+1$ and satisfying $G[i] = s[i+1]$ (resp. $H[i] = s[i+1]$). The process stops if either it is impossible to construct the next step, or ends successfully if $i = |s|$. In case of a successful termination it is easy to check if every non leaf node has a unique son, solving the unambiguity problem of the sequence s. These are the basic tools for constructing randomly or exhaustively many classes of words, for instance all full words of length n.

Enumeration. Counting classes of G-consistent or H-consistent sequences follows naturally. For instance, given a fixed length n, it amounts to count for a fixed alphabet Σ the set $\{ H_w \,:\, w \in \Sigma^n \}$. Indeed, a greedy algorithm can be implemented to obtain the first values: it suffices to generate all words in Σ^n, and to compute H for each such word.

The enumeration formula of the finite Sturmian words is known [27]. Since they are full, a closely related counting problem is that of determining a formula for the number of non-Sturmian full words on the alphabet $\{a, b\}^*$. Determining the number of words having a fixed number of lacunas is also challenging.

Characterization of special classes of G or H functions. For instance, given a 2-letter alphabet $\Sigma = \{a, b\}$ one might look for a description of the following sets of functions:

$$\mathcal{G} = \{G_w \,:\, w \in \{a,b\}^*\} \qquad \mathcal{H} = \{H_w \,:\, w \in \{a,b\}^* \text{ such that } w \text{ is full}\}.$$

In another direction it would be interesting to describe infinite words on fixed alphabets whose G (or H) sequence is automatic.

Constrained reconstruction. Given a finite set of palindromes P, how can we determine the shortest full word containing all the palindromes of P and only those palindromes? The answer is based on Theorem 1 of [11]. Indeed, the language of words having exactly P as palindromic factors is rational. Therefore there exists a deterministic minimal automaton recognizing all these words. For each palindrome q in P, there is a unique path starting from the initial state whose trace is q. Collecting the target states $\mathrm{T}(P)$ of all paths computing P, it suffices then to compute the shortest path starting from the initial state and containing all states in $\mathrm{T}(P)$. It may or may not exist, and if it does not, one might relax the conditions by allowing some extra palindromes in order to find a solution.

Structure of full words. Let w be a finite full word on a 2-letter alphabet Σ. One can easily prove that H_w and $H_{\widetilde{w}}$ have the same elements, while $H_w = H_{\widetilde{w}}$ if and only if $\widetilde{w} = \overline{w}$ or $\widetilde{w} = w$. The two sets of longest unioccurrent palindromic suffixes of w and $\widetilde{w}$ naturally define a permutation on the set $\{1, 2, \ldots, |w|\}$. More precisely, let p_1, p_2, …, $p_{|w|}$ be the longest palindromic suffixes of w in order of their first occurrence in w and let x_i be the position of the last occurrence of p_i in w. We define the permutation π_w on $\{1, 2, \ldots, |w|\}$ by

$$\pi_w(i) = |w| + |p_i| - |x_i|.$$

Now, let q_1, q_2, …, $q_{|w|}$ be the longest palindromic suffixes of $\widetilde{w}$ in order of their first occurrence in $\widetilde{w}$. Then $p_i = q_{\pi_w(i)}$, for $i = 1, 2, \ldots, |w|$. We illustrate this fact by an example: let $w = ababbabab$, so that $\widetilde{w} = bababbaba$. Then we have the following table showing that $H_w \neq H_{\widetilde{w}}$,

i	1	2	3	4	5	6	7	8	9
LPSU(w)	*a*	*b*	*aba*	*bab*	*bb*	*abba*	*babbab*	*ababbaba*	*babab*
LPSU($\widetilde{w}$)	*b*	*a*	*bab*	*aba*	*babab*	*bb*	*abba*	*babbab*	*ababbaba*

and the permutation π_w is $(2, 1, 4, 3, 6, 7, 8, 9, 5)$.

We would like to study the combinatorial properties of π_w in relation with those of the word w. In particular we are interested in characterizing the permutations π_w associated with full words. A similar study can also be performed on arbitrary alphabets provide one replaces the $\overline{(\ )}$ operation by an arbitrary permutation of the alphabet Σ.

Acknowledgements

We are grateful to the anonymous referees for their careful reading and useful comments.

References

1. Aberkane, A. and Brlek, S. (2002) Suites de même complexité que celle de Thue-Morse, *Actes des Journées Montoises d'informatique théorique (9-11 septembre 2002, Montpellier, France)* 85–89.
2. Aberkane, A., Brlek, S. and Glen, A. (2007) Sequences having the Thue-Morse complexity, *Disc. Math.* 15p. (submitted)
3. Allouche, J.-P. (1994) Sur la complexité des suites infinies, *Bull. Belg. Math. Soc.* 1:133–143.
4. Allouche, J.P. (1997) Schrödinger operators with Rudin-Shapiro potentials are not palindromic, *J. Math. Phys.* 38:1843–1848.
5. Allouche, J.P., Baake, M., Cassaigne, J., and Damanik, D. (2003) Palindrome complexity, *Theoret. Comput. Sci.* 292:9–31.

6. Allouche, J.P., and Shallit, J. (2000) Sums of digits, overlaps, and palindromes, *Disc. Math. and Theoret. Comput. Sci.* 4:1–10.
7. Baake, M. (1999) A note on palindromicity, *Lett. Math. Phys.* 49:217–227.
8. Blondin Massé, A., Brlek, S., and Labbé, S. (2008) Palindromic lacunas of the Thue-Morse word, GASCOM 2008 (To appear)
9. Fici, G., Mignosi, F., Restivo, A. and Sciortino, M. (2006) Word assembly through minimal forbidden words,*Theoret. Comput. Sci.*, 359/1-3: 214–230.
10. Brlek, S. (1989) Enumeration of factors in the Thue-Morse word, *Disc. Appl. Math.* 24:83–96.
11. Brlek, S., Hamel, S., Nivat, M., and Reutenauer, C. (2004) On the Palindromic Complexity of Infinite Words, in J. Berstel, J. Karhumäki, D. Perrin, Eds, Combinatorics on Words with Applications, Int. J. of Found. of Comput. Sci., 15/2:293–306
12. Brlek, S., and Ladouceur, A. (2003) A note on differentiable palindromes, *Theoret. Comput. Sci.* 302:167–178.
13. Brlek, S., Jamet, D., and Paquin, G., (2008) Smooth Words on 2-letter alphabets having same parity, *Theoret. Comput. Sci.* 393/1-3:166181.
14. Brlek, S., Dulucq, S., Ladouceur, A. and Vuillon L. (2006) Combinatorial properties of smooth infinite words, *Theoret. Comput. Sci.* 352/1-3:306–317.
15. de Bruijn N. G. (1946) A Combinatorial Problem, *Koninklijke Nederlandse Akademie v. Wetenschappen* 49: 758764.
16. Droubay, X., Justin, J., and Pirillo, G. (2001) Episturmian words and some constructions of de Luca and Rauzy, *Theoret. Comput. Sci.* 255:539–553.
17. Flye Sainte-Marie, C. (1894). Question 48, *L'Intermdiaire Math.* 1: 107110.
18. Fredericksen, H. and Maiorana, J. (1978) Necklaces of beads in k colors and k-ary de Bruijn sequences *Disc. Math.* 23/3, 207–210
19. Good, I. J. (1946) Normal recurring decimals, *J. London Math. Soc.* 21 (3): 167–169.
20. Hof, A., Knill, O., and Simon, B. (1995) Singular continuous spectrum for palindromic Schrödinger operators, *Commun. Math. Phys.* 174:149–159.
21. Lothaire M. (1983) Combinatorics on words, Addison-Wesley.
22. Lothaire, M. (2002) Algebraic Combinatorics on words, Cambridge University Press.
23. de Luca, A. (1997) Sturmian words: structure, combinatorics, and their arithmetics, *Theoret. Comput. Sci.* 183:45–82.
24. de Luca, A. , and Varricchio, S. (1989) Some combinatorial properties of the Thue-Morse sequence, *Theoret. Comput. Sci.* 63:333–348.
25. Mantaci, S., Restivo, A., Rosone, G. and Sciortino, M. (2008) A New Combinatorial Approach to Sequence Comparison*Theory Comput. Syst.* 42/3:411–429.
26. Manvel, B., Meyerowitz, A., Schwenk*, A., Smith, K. and Stockmeyer, P. (1991) Reconstruction of sequences, *Disc. Math.* 94/3: 209–219.
27. Mignosi, F. (1991) On the number of factors of Sturmian words, *Theoret. Comput. Sci.* 82/1: 71–84.
28. Morse, M. and Hedlund, G. (1940) Symbolic Dynamics II. Sturmian trajectories, *Amer. J. Math.* 62:1–42.