Characterization of some binary words with few squares

Golnaz Badkobeh[a], Pascal Ochem[b]

[a] *Department of Computer Science, University of Sheffield, UK*
[b] *CNRS - LIRMM, Montpellier, France*

Abstract

Thue proved that the factors occurring infinitely many times in square-free words over {0,1,2} avoiding the factors in {010,212} are the factors of the fixed point of the morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$. He similarly characterized square-free words avoiding {010,020} and {121,212} as the factors of two morphic words. In this paper, we exhibit smaller morphisms to define these two square-free morphic words and we give such characterizations for six types of binary words containing few distinct squares.

1. Introduction

Let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k\text{-}1\}$. Let ε denote the empty word. A finite word is *recurrent* in an infinite word w if it appears as a factor of w infinitely many times. An infinite word w is *recurrent* if all its finite factors are recurrent in w. If a morphism f is such that $f(0)$ starts with 0, then the *fixed point* of f is the unique word $w = f^\omega(0)$ starting with 0 and satisfying $w = f(w)$. An infinite word is *pure morphic* if it is the fixed point of a morphism. An infinite word is *morphic* if it is the image $g(f^\omega(0))$ by a morphism g of a pure morphic word $f^\omega(0)$. The *factor complexity* of an infinite word or a language is the number of factors of length n of the infinite word or the language. A pattern P is a finite word of variables over the alphabet $\{A, B, \ldots\}$. A word w (finite or infinite) *avoids* a pattern P if for every substitution ϕ of the variables of P with non-empty words, $\phi(P)$ is not a factor of w. Given a finite alphabet Σ_k, a finite set $\mathcal{P}$ of patterns, and a finite set $\mathcal{F}$ of factors over Σ_k, we say that $\mathcal{P} \cup \mathcal{F}$ *characterizes* a morphic word w over Σ_k if w avoids $\mathcal{P} \cup \mathcal{F}$ and every recurrent factor of an infinite word avoiding $\mathcal{P} \cup \mathcal{F}$ is a factor of w. In other words, $\mathcal{P} \cup \mathcal{F}$ characterizes w if and only if every recurrent word over Σ_k avoiding $\mathcal{P} \cup \mathcal{F}$ has the same set of factors as w. In our results, we do not specify the alphabet size k since Σ_k corresponds to the set of letters appearing in $\mathcal{F}$. A *repetition* is a factor of the form $r = u^n v$ where u is non-empty and v is a prefix of u. Then $|u|$ is the *period* of the repetition r and its *exponent* is $|r|/|u|$. A *square* is a repetition of exponent 2. Equivalently, it is an occurrence of the pattern AA. An overlap is a repetition with exponent strictly greater than 2.

Thue [3, 10, 11] gave the following characterization of overlap-free binary words: $\{ABABA\} \cup \{000, 111\}$ characterizes the fixed point of the morphism

$0 \mapsto 01$, $1 \mapsto 10$. Concerning ternary square-free words, he proved that

- $\{AA\} \cup \{010, 212\}$ characterizes the fixed point of $f_3 : 0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$,
- $\{AA\} \cup \{010, 020\}$ characterizes the morphic word $T_1(f_T^\infty(0))$,
- $\{AA\} \cup \{121, 212\}$ characterizes the morphic word $T_2(f_T^\infty(0))$,

where the morphisms f_T, T_1, and T_2 are given below.

$f_T(0) = 012$,	$T_1(0) = 01210212$,	$T_2(0) = 021012$,
$f_T(1) = 0432$,	$T_1(1) = 01210120212$,	$T_2(1) = 02102012$,
$f_T(2) = 0134$,	$T_1(2) = 01210212021$,	$T_2(2) = 02101201$,
$f_T(3) = 013432$,	$T_1(3) = 012102120210120212$,	$T_2(3) = 0210120102012$,
$f_T(4) = 0434$.	$T_1(4) = 0121012021$.	$T_2(4) = 0210201$.

To obtain the last two results, Thue first proved that $f_T^\infty(0)$ is characterized by $\{AA\} \cup \{02, 03, 10, 14, 21, 23, 24, 30, 31, 41, 42, 040, 132, 404, 1201, 2012\}$.

In this paper, we prove such characterizations mostly for the binary words considered by the first author [1]. We also obtain smaller morphisms for Thue's words avoiding $\{AA\} \cup \{010, 020\}$ and $\{AA\} \cup \{121, 212\}$ as well as a characterization for words avoiding the patterns $AABBCC$ (i.e., three consecutive squares), $ABCABC$ and a finite set of factors. The results are summarized in Table 1. The first column shows the description of the considered language given in the literature. It is either given by forbidden sets of patterns and factors, or by the notation (e, n, m), which means that we consider the binary words avoiding repetitions with exponent strictly greater than e, containing exactly n distinct repetitions with exponent e as a factor, and containing the minimum number m of distinct squares. We use the notation SQ_t for the pattern corresponding to squares with period at least t, that is, $SQ_1 = AA$, $SQ_2 = ABAB$, $SQ_3 = ABCABC$, and so on. These languages actually have an equivalent definition with one forbidden pattern SQ_t and a finite set of forbidden factors. This standardized definition, given in the second column, is more suited for proving the characterization. The third column gives the corresponding morphic word. The fourth column indicates the section containing the corresponding set F_{xx} and morphism g_{xx}.

To define a morphic word $g(f^\infty(0))$, we allow that g is an *erasing* morphism, i.e., that the g-image of a letter is empty. Notice that replacing g by $h_c = g \circ f^c$ defines the same morphic word, and that h_c is non-erasing for some small constant c.

The proofs are obtained by computer using the technique described in the next section. An example of proof by hand is given for Theorem 3. The morphic words in Table 1 are gathered according the pure morphic word they are built on. We introduce in Section 3 a pure morphic word $f_5^\infty(0)$ similar to Thue's word $f_T^\infty(0)$ and we characterize some of its morphic images. Section 4 is devoted to characterizations of some morphic images of Thue's ternary pure morphic word $f_3^\infty(0)$.

Original form	Standardized form	Morphic word	Section
$\{AA\} \cup \{\texttt{010,020}\}$	$\{AA\} \cup \{\texttt{010,020}\}$	$M_1(f_5^\infty(\texttt{0}))$	3.1
$\{AA\} \cup \{\texttt{121,212}\}$	$\{AA\} \cup \{\texttt{121,212}\}$	$M_2(f_5^\infty(\texttt{0}))$	3.1
$(5/2, 2, 8)$	$\{SQ_7\} \cup F_8$	$g_8(f_5^\infty(\texttt{0}))$	3.2
$(7/3, 2, 12)$	$\{SQ_9\} \cup F_{12}$	$g_{12}(f_5^\infty(\texttt{0}))$	3.3
$(7/3, 1, 14)$	$\{SQ_9\} \cup F_{14}$	$g_{14}(f_5^\infty(\texttt{0}))$	3.4
$\{AABBCC, SQ_3\} \cup F'_{cs}$	$\{SQ_3\} \cup F_{cs}$	$g_{cs}(f_5^\infty(\texttt{0}))$	3.5
$(5/2, 1, 11)$	$\{SQ_5\} \cup F_{11}$	$g_{11}(f_3^\infty(\texttt{0}))$	4.1
$(3, 2, 3) \cup F'_3$	$\{SQ_3\} \cup F_3$	$g_3(f_3^\infty(\texttt{0}))$	4.2
$\{AABBCABBA\} \cup \{\texttt{0011,1100}\}$	$\{SQ_5\} \cup F_q$	$g_q(f_3^\infty(\texttt{0}))$	4.3

Figure 1: Table of results

2. Characterizing a morphic word

A morphism $f : \Sigma_k^* \to \Sigma_k^*$ is *primitive* if there exists $n \in \mathbb{N}$ such that $f^n(a)$ contains b for every $a, b \in \Sigma_k$. We are given a primitive morphism $f : \Sigma_k^* \to \Sigma_k^*$, a morphism $g : \Sigma_k^* \to \Sigma_{k'}^*$, and a finite set of factors $\mathcal{F}_m \subset \Sigma_{k'}^*$. We want to prove that $g(f^\infty(\texttt{0}))$ is characterized by $\{SQ_t\} \cup \mathcal{F}_m$.

We assume that $g(f^\infty(\texttt{0}))$ avoids $\{SQ_t\} \cup \mathcal{F}_m$. This can be checked using Cassaigne's algorithm [5] that determines if a morphic word defined by circular morphisms avoids a given pattern with constants. We refer to Cassaigne [5] for the definitions of circular morphisms, synchronization point, and synchronization delay. We can use an online implementation [4] of this algorithm. We also assume that the pure morphic word $f^\infty(\texttt{0})$ is characterized by $\{AA\} \cup \mathcal{F}_p$ for some finite set of factors $\mathcal{F}_p \subset \Sigma_k^*$.

We compute the smallest integer c such that $\min\{|g(f^c(a))|,\ a \in \Sigma_k\} \geqslant t$. This c exists because f is primitive. We can consider the morphism $g' = g \circ f^c$ instead of g since we have $g'(f^\infty(\texttt{0})) = g(f^\infty(\texttt{0}))$.

First, we check that g' is circular. Then, we compute the set S_l of words v such that there exists a word $pvs \in \Sigma_{k'}^*$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$, where $l = \max\{|u|, u \in \mathcal{F}_p\} \times \max\{|g'(a)|, a \in \Sigma_k\}$, $|v| = l$, and $|p| = |s| = 4l$. To do this, we simply perform a depth-first exploration of the words of length $9l$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ and for each of them, we put the central factor of length l in S_l. The running time of this brute-force approach is not so prohibitive precisely because the characterization implies a polynomial factor complexity. Finally, we check that every word in S_l is a factor of $g'(f^\infty(\texttt{0}))$.

This implies that an infinite word over $\Sigma_{k'}$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ is the g'-image of an infinite word $w \in \Sigma_k^*$. Now w is square-free, since otherwise $g'(w)$ would contain a square of period at least t. Also w does not contain a word $y \in \mathcal{F}_p$, because $g'(y)$ is a word of length at most l that is not a factor of any word in S_l. So w avoids $\{AA\} \cup \mathcal{F}_p$, and thus has the same set of factors as $f^\infty(\texttt{0})$. Thus, every infinite recurrent word over $\Sigma_{k'}$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ has the same set of factors as $g'(f^\infty(\texttt{0}))$.

The programs we used are available at
`http://www.lirmm.fr/~ochem/morphisms/characterization.htm` .

3. A pure morphic word over Σ_5

We define the morphism f_5 from Σ_5^* to Σ_5^* as follows:

$$\begin{aligned} f_5(\mathtt{0}) &= \mathtt{01},\\ f_5(\mathtt{1}) &= \mathtt{23},\\ f_5(\mathtt{2}) &= \mathtt{4},\\ f_5(\mathtt{3}) &= \mathtt{21},\\ f_5(\mathtt{4}) &= \mathtt{0}. \end{aligned}$$

We also define the set

$$F_5 = \{\mathtt{02}, \mathtt{03}, \mathtt{13}, \mathtt{14}, \mathtt{20}, \mathtt{24}, \mathtt{31}, \mathtt{32}, \mathtt{40}, \mathtt{41}, \mathtt{43}, \mathtt{121}, \mathtt{212}, \mathtt{304}, \mathtt{3423}, \mathtt{4234}\}.$$

Theorem 1. *$\{AA\} \cup F_5$ characterizes $f_5^\infty(\mathtt{0})$.*

Proof. We adapt the method of the previous section for morphic words to the pure morphic word $f_5^\infty(\mathtt{0})$ by setting $g = g' = f_5$ and $\mathcal{F}_m = \mathcal{F}_p = F_5$. We set $l = \max\{|u|, u \in F_5\} \times \max\{|f_5(a)|, a \in \Sigma_k\} = 8$. We compute the set S_l of words v such that there exists a word $pvs \in \Sigma_5^*$ avoiding squares and F_5 with $|v| = l$ and $|p| = |s| = 4l$. Then we check that every word in S_l is a factor of $f_5^\infty(\mathtt{0})$.

The morphism f_5 is circular with synchronization delay 1. Indeed, for every factor of length 1 of the f_5-image of some word, we can insert at least one synchronization point | between letter images:

$$\begin{aligned} &\mathtt{0} \text{ implies } |\mathtt{0},\\ &\mathtt{1} \text{ implies } \mathtt{1}|,\\ &\mathtt{2} \text{ implies } |\mathtt{2},\\ &\mathtt{3} \text{ implies } \mathtt{3}|,\\ &\mathtt{4} \text{ implies } |\mathtt{4}|. \end{aligned}$$

This implies that every infinite recurrent word over Σ_5 avoiding $\{AA\} \cup F_5$ is the f_5-image of some infinite recurrent word w over Σ_5. Notice that w must be square-free, since otherwise $f_5(w)$ would not avoid squares. Now suppose that w contains a factor $y \in F_5$. Then $f_5(y)$ must appear as a factor in S_l since $|f_5(y)| \leqslant 8 = l$. Every word in S_l is a factor of $f_5^\infty(\mathtt{0})$, so $f_5(y)$ should also be a factor of $f_5^\infty(\mathtt{0})$, which is a contradiction. So w avoids squares and F_5, which implies by induction that it has the same set of factors as $f_5^\infty(\mathtt{0})$. Finally, we have that every infinite recurrent word over Σ_5 avoiding $\{AA\} \cup F_5$ is of the form $f_5(w)$ where w has the same set of factors as $f_5^\infty(\mathtt{0})$, so that $f_5(w)$ also has the same set of factors as $f_5^\infty(\mathtt{0})$.

□

Since many morphic words in this paper are obtained as the image of $f_5^\infty(\mathtt{0})$, let us state some of its properties. In $f_5^\infty(\mathtt{0})$, the letters $\mathtt{0}$, $\mathtt{1}$, and $\mathtt{2}$ have frequency $\sqrt{5} - 2$ and the letters $\mathtt{3}$ and $\mathtt{4}$ have frequency $\left(7 - 3\sqrt{5}\right)/2$. Notice that $\{AA\} \cup F_5$, and thus the set of factors of $f_5^\infty(\mathtt{0})$, is invariant by the operation

consisting in reversing the word and exchanging 3 and 4. This is trivially true for squares. For a word in F_5, say 40, we obtain 04 by reversing the word and we obtain 03 by exchanging 3 and 4, then we have that F_5 contains indeed 03. The factor complexity of $f_5^\infty(0)$ seems to be $4n+1$ for every factor length $n \geqslant 0$.

3.1. Smaller morphisms for Thue's words

Let M_1 and M_2 be the morphisms from Σ_5^* to Σ_3^* defined by

$$\begin{array}{ll} M_1(0) = 012, & M_2(0) = 02, \\ M_1(1) = 1, & M_2(1) = 1, \\ M_1(2) = 02, & M_2(2) = 0, \\ M_1(3) = 12, & M_2(3) = 12, \\ M_1(4) = \varepsilon. & M_2(4) = \varepsilon. \end{array}$$

Theorem 2.

- $\{AA\} \cup \{010, 020\}$ *characterizes the morphic word* $M_1(f_5^\infty(0))$,
- $\{AA\} \cup \{121, 212\}$ *characterizes the morphic word* $M_2(f_5^\infty(0))$.

Thue noticed that every word avoiding $\{AA\} \cup \{121,212\}$ can be obtained from a word avoiding $\{AA\} \cup \{010,020\}$ by deleting the letter immediately after each occurrence of the letter 0. This property is easy to check by comparing M_2 to M_1 and it explains why the same pure morphic word is used for both types of words. The morphisms M_1 and M_2 are the smallest possible. However, the morphisms $M_1' = M_1 \circ f_5$ and $M_2' = M_2 \circ f_5$ given below provide additional insight.

$$\begin{array}{ll} M_1'(0) = 0121, & M_2'(0) = 021, \\ M_1'(1) = 0212, & M_2'(1) = 012, \\ M_1'(2) = \varepsilon, & M_2'(2) = \varepsilon, \\ M_1'(3) = 021, & M_2'(3) = 01, \\ M_1'(4) = 012. & M_2'(4) = 02. \end{array}$$

The morphism M_1' exhibits natural properties of words avoiding $\{AA\} \cup \{010,020\}$ and of $M_1(f_5^\infty(0))$:

- The set $\{0121,0212,012,021\}$ is a code for words avoiding $\{AA\} \cup \{010,020\}$.
- The asymptotic frequencies of the factors 121 and 212 are equal since the letters 1 and 2 are symmetrical for words avoiding $\{AA\} \cup \{010,020\}$.
- Similarly, the asymptotic frequencies of 0120 and 0210 are equal.
- By applying the symmetry of the factors of $f_5^\infty(0)$ to M_1', that is, reversing the M_1'-images of every letter and exchanging 3 and 4, we obtain the conjugate morphism of M_1' such that the common prefix 0 becomes the common suffix.

Except for the last, similar remarks hold for M_2'. The factor complexity of $M_1(f_5^\infty(0))$ and $M_2(f_5^\infty(0))$ seems to be $4n-2$ for every factor length $n \geqslant 2$.

3.2. Words containing two 5/2-repetitions and 8 squares

If an infinite binary word contains the repetitions 01010 and 10101 of exponent 5/2 and no other overlap, then it contains at least 8 distinct squares. Moreover, if it contains exactly 8 distinct squares, then these 8 squares are 0^2, 1^2, $(01)^2$, $(10)^2$, $(0110)^2$, $(1001)^2$, $(011001)^2$, $(100110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_7 and the set

$$F_8 = \{000,111,00100,11011,010010,010101,101010,101101,00110011,\\ 11001100,1011001011,0100110100\}.$$

Let g_8 be the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned} g_8(0) &= 011,\\ g_8(1) &= 0,\\ g_8(2) &= 01,\\ g_8(3) &= \varepsilon,\\ g_8(4) &= \varepsilon. \end{aligned}$$

Theorem 3. $\{SQ_7\} \cup F_8$ *characterizes* $g_8(f_5^\infty(0))$.

Proof. We assume that $g_8(f_5^\infty(0))$ avoids SQ_7 and F_8 and we prove the other direction of Theorem 3. That is, we suppose that G_8 is an infinite recurrent word avoiding $\{SQ_7\} \cup F_8$ and we show that every factor of G_8 is a factor of $g_8(f_5^\infty(0))$. We consider the morphism $g_8' = g_8 \circ f_5^5$ given below instead of g_8 because we have $\min\{|g_8'(a)|,\ a \in \Sigma_5\} = 9 \geqslant 7 = t$, as specified in the method.

$$\begin{aligned} g_8'(0) &= 011001010011010110011010,\\ g_8'(1) &= 011001011001101,\\ g_8'(2) &= 011001010,\\ g_8'(3) &= 0110010110011010,\\ g_8'(4) &= 01100101001101. \end{aligned}$$

Let $p = 01100101$ be the common prefix of the factors $g_8'(a)$ for $a \in \Sigma_5$. It is easy to check that every occurrence of p in the g_8'-image of a word is the prefix of g_8'-image of a letter. So g_8' has bounded synchronization delay. Moreover, a computer check shows that the factors of G_8 are factors of the g_8'-image of a word. Let $L \subset \Sigma_5^*$ denote the language of words whose g_8'-image is a factor of G_8. We show that L is the set of factors of $f_5^\infty(0)$. Suppose that L contains a square uu for some $u \in \Sigma_5^+$. Then G_8 contains the square $g_8'(uu)$ with period $|g_8'(u)| \geqslant 9$. This is a contradiction since G_8 avoids SQ_7, so L is square-free.

Now, for every $w \in F_5$, we suppose that $w \in L$ and obtain a contradiction:

- $w \in \{02,32\}$: $g_8'(02)p$ and $g_8'(32)p$ both contain the square $1g_8'(2)p = (001100101)^2$ with period 9 as a suffix.
- $w = 03$: $g_8'(03)p$ contains the square $(1001101001100101)^2$ with period 16 as a suffix.

- $w \in \{\texttt{13},\texttt{41},\texttt{43}\}$: A common suffix of $g'_8(\texttt{1})$ and $g'_8(\texttt{4})$ is $\texttt{1}$. A common prefix of $g'_8(\texttt{1})$ and $g'_8(\texttt{3})$ is $\texttt{011001011}$. So, in every case, $g'_8(w)$ contains the factor $\texttt{1011001011} \in F_8$.

- $w = \texttt{14}$: $g'_8(\texttt{14})p$ contains the square $(\texttt{00110101100101})^2$ with period 14 as a suffix.

- $w \in \{\texttt{20},\texttt{24}\}$: $g'_8(\texttt{20})$ and $g'_8(\texttt{24})$ both contain the square $g'_8(\texttt{22})$ with period 9 as a prefix.

- $w = \texttt{31}$: $g'_8(\texttt{31})p$ contains the square $g'_8(\texttt{33})$ with period 16 as a prefix.

- $w = \texttt{40}$: $g'_8(\texttt{40})$ contains the square $g'_8(\texttt{44})$ with period 14 as a prefix.

- $w = \texttt{304}$: $g'_8(\texttt{304}) = \texttt{0110}(\texttt{0101100110100110010100 11})^2\texttt{01}$ contains a square with period 24.

- $w = \texttt{121}$: Since L is square-free and avoids $\{\texttt{13},\texttt{14}\}$, L must contain $\texttt{1210}$. However, $g'_8(\texttt{1210})$ contains the square $g'_8(\texttt{1212})$ with period 24 as a prefix.

- $w = \texttt{212}$: Since L is square-free and avoids $\{\texttt{20},\texttt{24}\}$, L must contain $\texttt{2123}$. However, $g'_8(\texttt{2123})$ contains the square $g'_8(\texttt{2121})$ with period 24 as a prefix.

- $w = \texttt{3423}$: Since L is square-free and avoids $\{\texttt{03},\texttt{13},\texttt{43}\}$, L must contain $\texttt{23423}$. Since L is square-free and avoids $\{\texttt{31},\texttt{32}\}$, L must contain $\texttt{234230}$. However, $g'_8(\texttt{234230})$ contains the square $g'_8(\texttt{234234})$ with period 39 as a prefix.

- $w = \texttt{4234}$: Since L is square-free and avoids $\{\texttt{40},\texttt{41},\texttt{43}\}$, L must contain $\texttt{42342}$. Since L is square-free and avoids $\{\texttt{20},\texttt{24}\}$, L must contain $\texttt{423421}$. However, $g'_8(\texttt{423421})p$ contains the square $g'_8(\texttt{423423})$ with period 39 as a prefix.

Therefore L is square-free and does not contain a factor in F_5, thus L is the set of factors as $f_5^\infty(\texttt{0})$ by Theorem 1. □

Notice that the last part of the proof above (when we prove that every word in F_5 is a forbidden factor in L) differs from the computer check described in Section 2. The proof by hand exhibits witness forbidden factors in $\{SQ_t\} \cup F_m$. The algorithm does the contrapositive: It lists all words avoiding SQ_t and F_m of some sufficient length and checks that they are g'-images of some The proof by hand exhibits witness forbidden factors in $\{SQ_t\} \cup F_m$. The algorithm does the contrapositive: It lists all words avoiding $\{SQ_t\} \cup F_m$ of some sufficient length and checks that they are images of some word avoiding $\{AA\} \cup F_p$.

The factor complexity of $g_8(f_5^\infty(\texttt{0}))$ seems to be $4n-6$ for every factor length $n \geqslant 3$.

3.3. Words containing two 7/3-repetitions and 12 squares

If an infinite binary word contains the repetitions 0110110 and 1001001 of exponent 7/3 and no other overlap, then it contains at least 12 distinct squares. Moreover, if it contains exactly 12 distinct squares, then these 12 squares are 0^2, 1^2, $(01)^2$, $(10)^2$, $(001)^2$, $(010)^2$, $(011)^2$, $(100)^2$, $(101)^2$, $(110)^2$, $(01101001)^2$, $(10010110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_9 and the set

$$F_{12} = \{000,111,01010,10101,001100,110011,0010010,0100100,1011011,\\ 1101101,0011010011,0101100101,1010011010,1100101100,\\ 01001011010010\}.$$

Let g_{12} be the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned} g_{12}(0) &= 01,\\ g_{12}(1) &= 0,\\ g_{12}(2) &= 011,\\ g_{12}(3) &= \varepsilon,\\ g_{12}(4) &= \varepsilon. \end{aligned}$$

Theorem 4. *$\{SQ_9\} \cup F_{12}$ characterizes $g_{12}(f_5^\infty(0))$.*

The factor complexity of $g_{12}(f_5^\infty(0))$ seems to be $4n - 6$ for every factor length $n \geqslant 3$.

3.4. Words containing one 7/3-repetition and 14 squares

If an infinite binary word contains the repetition 1001001 of exponent 7/3 and no other overlap, then it contains at least 14 distinct squares. Moreover, if it contains exactly 14 distinct squares, then these 14 squares are 0^2, 1^2, $(01)^2$, $(10)^2$, $(001)^2$, $(010)^2$, $(100)^2$, $(101)^2$, $(0110)^2$, $(1001)^2$, $(100110)^2$, $(0100110)^2$, $(0110010)^2$, and $(10010110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_9 and the set

$$F_{14} = \{000,111,11011,010101,101010,0010010,0100100,00110011,\\ 11001100,101001101,101100101,0100101101,1100101100,\\ 001001100100,010011010011,0011001001100,1011010010110011\}.$$

Let g_{14} be the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned} g_{14}(0) &= 01,\\ g_{14}(1) &= 00110,\\ g_{14}(2) &= 1,\\ g_{14}(3) &= 0010110,\\ g_{14}(4) &= 0110. \end{aligned}$$

Theorem 5. *$\{SQ_9\} \cup F_{14}$ characterizes $g_{14}(f_5^\infty(0))$.*

The factor complexity of $g_{14}(f_5^\infty(0))$ seems to be $4n - 1$ for every factor length $n \geqslant 11$.

3.5. Words avoiding AABBCC

The second author proved that the pattern $AABBCC$, i.e., three consecutive squares, can be avoided over the binary alphabet [8]. More precisely, there exist exponentially many binary words avoiding both $AABBCC$ and SQ_3. However, if we forbid also the factors in

$$F'_{cs} = \{\texttt{0001110010110}, \texttt{0110100111000}, \texttt{1001011000111}, \texttt{1110001101001}\},$$

we obtain a characterization of the morphic word $g_{cs}(f_5^\infty(\texttt{0}))$, where g_{cs} is the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned} g_{cs}(\texttt{0}) &= \texttt{00101100011010},\\ g_{cs}(\texttt{1}) &= \texttt{0111},\\ g_{cs}(\texttt{2}) &= \texttt{0010111010},\\ g_{cs}(\texttt{3}) &= \texttt{011100011010},\\ g_{cs}(\texttt{4}) &= \texttt{001011000111}. \end{aligned}$$

The word $g_{cs}(f_5^\infty(\texttt{0}))$ avoids SQ_3 and the set

$$\begin{aligned} F_{cs} = \{&\texttt{0000}, \texttt{1111}, \texttt{01010}, \texttt{10101}, \texttt{011001}, \texttt{100110}, \texttt{0011101}, \texttt{1011100},\\ &\texttt{1100010}, \texttt{00010111}, \texttt{11101000}, \texttt{0001110010110}, \texttt{0110100111000},\\ &\texttt{1001011000111}, \texttt{1110001101001}\} \end{aligned}$$

Theorem 6. *$\{AABBCC, SQ_3\} \cup F'_{cs}$ and $\{SQ_3\} \cup F_{cs}$ both characterize $g_{cs}(f_5^\infty(\texttt{0}))$.*

The factor complexity of $g_{cs}(f_5^\infty(\texttt{0}))$ seems to be $4n + 4$ for every factor length $n \geqslant 6$.

4. Thue's ternary pure morphic word

Thue [3, 10, 11] proved that $\{AA\} \cup \{\texttt{010}, \texttt{212}\}$ characterizes the fixed point of f_3. In this section, we give characterizations of three words that are morphic images of $f_3^\infty(\texttt{0})$. It is not surprising that $f_3^\infty(\texttt{0})$ appears in the context of characterizations: as soon as a morphism m is such that $m(\texttt{0}) = axb$ and $m(1) = ab$, the m-image of words of the form $0u1u0$, $u \in \Sigma_3^*$, contains a large square: $m(0u1u0) = axbm(u)abm(u)axb$ contains $(bm(u)a)^2$. Moreover, a ternary square-free word avoids factors of the form $0u1u0$ with $u \in \Sigma_3^*$ if and only if it avoids $\{010, 212\}$ [9]. So, the set of factors of a factorial langage containing only square-free factors in $\{m(0), m(1), m(2)\}^*$ such that $m(\texttt{0}) = axb$ and $m(1) = ab$ is the set of factors of $m(f_3^\infty(\texttt{0}))$. It is also easy to check that $\{AA\} \cup \{\texttt{010}, \texttt{212}\}$ characterizes the same ternary word as $\{AA\} \cup \{\texttt{1021}, \texttt{1201}\}$.

4.1. Words containing one 5/2-repetition and 11 squares

If an infinite binary word contains the repetition $\texttt{10101}$ of exponent $5/2$ and no other overlap, then it contains at least 11 distinct squares. Moreover, if it contains exactly 11 distinct squares, then these 11 squares are $\texttt{0}^2$, $\texttt{1}^2$, $(\texttt{01})^2$,

$(10)^2$, $(001)^2$, $(010)^2$, $(011)^2$, $(100)^2$, $(101)^2$, $(110)^2$, $(01100110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_7 and the set

$$F_{11} = \{000,111,01010,001100,0010010,0100100,1011011,1101101\}.$$

Let g_{11} be the morphism from Σ_3^* to Σ_2^* defined by

$$\begin{aligned} g_{11}(0) &= 1001001101011001101001011001001101100 \\ &\quad\ 1011010011011001001101001011001101 01, \\ g_{11}(1) &= 100100110100101, \\ g_{11}(2) &= 10010011011001011010011 01. \end{aligned}$$

Theorem 7. *$\{SQ_5\} \cup F_{11}$ characterizes $g_{11}(f_3^\infty(0))$.*

4.2. Words containing 3 squares

It is known that there exist exponentially many binary words containing only 3 distinct squares [7, 8]. Without loss of generality, we assume that these 3 squares are 00, 11, and 1010. To obtain a characterization, we forbid also the factors in $F_3' = \{01000110,10011101,1001101000,1110100110\}$. If w is a recurrent binary word avoiding F_3' and squares distinct from 00, 11, and 1010, then w avoids SQ_3 and the set

$$F_3 = \{0000,0101,1111,01000110,10011101,1001101000,1110100110\}.$$

Let g_3 be the morphism from Σ_3^* to Σ_2^* defined by

$$\begin{aligned} g_3(0) &= 000111, \\ g_3(1) &= 0011, \\ g_3(2) &= 01001110001101. \end{aligned}$$

Theorem 8. *$\{SQ_3\} \cup F_3$ characterizes $g_3(f_3^\infty(0))$.*

4.3. Words avoiding AABBCABBA

Another characterization has been obtained by the second author [9]: $\{AABBCABBA\} \cup \{0011,1100\}$ characterizes $g_q(f_3^\infty(0))$, where g_q is given below.

$$\begin{aligned} g_q(0) &= 0010110111011101001, \\ g_q(1) &= 00101101101001, \\ g_q(2) &= 00010. \end{aligned}$$

Equivalently, $g_q(f_3^\infty(0))$ is characterized by $\{SQ_5\} \cup F_q$ where

$$\begin{aligned} F_q = \{&0000,0011,1100,1111,01010,10101,010111,101000,0001001, \\ &1110110,00100100,01011010,10100101,11011011,0110111010,1001000101\} \end{aligned}$$

5. Concluding remarks

We have seen in Section 4 why $f_3^\infty(\mathtt{0})$ appears often in the context of characterization. Also, we have seen in Section 3.1 why Thue's words avoiding $\{AA\} \cup \{\mathtt{010,020}\}$ and $\{AA\} \cup \{\mathtt{121,212}\}$ use the same pure morphic word $f_5^\infty(\mathtt{0})$. However, we do not see why $f_5^\infty(\mathtt{0})$ is used in other "natural" languages. It would be interesting to investigate its properties, in particular to prove that its factor complexity is $4n+1$ and that its critical exponent is $(5+\sqrt{5})/4$.

The fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{0}$, known as the Fibonacci word, seems to have the same set of factors as $g_{\mathtt{fib}}(f_5^\infty(\mathtt{0}))$, where $g_{\mathtt{fib}}$ is given below. Moreover, the Rote-Fibonacci word studied in [6] seems to have the same set of factors as $g_{\mathtt{rf}}(f_5^\infty(\mathtt{0}))$, where $g_{\mathtt{rf}}$ is given below.

$$
\begin{array}{ll}
g_{\mathtt{fib}}(\mathtt{0}) = \mathtt{01}, & g_{\mathtt{rf}}(\mathtt{0}) = \mathtt{01}, \\
g_{\mathtt{fib}}(\mathtt{1}) = \mathtt{0}, & g_{\mathtt{rf}}(\mathtt{1}) = \mathtt{10}, \\
g_{\mathtt{fib}}(\mathtt{2}) = \mathtt{1}, & g_{\mathtt{rf}}(\mathtt{2}) = \varepsilon, \\
g_{\mathtt{fib}}(\mathtt{3}) = \mathtt{0}, & g_{\mathtt{rf}}(\mathtt{3}) = \mathtt{11}, \\
g_{\mathtt{fib}}(\mathtt{4}) = \mathtt{0}. & g_{\mathtt{rf}}(\mathtt{4}) = \mathtt{00}.
\end{array}
$$

The method discussed in this paper is not able to prove such equivalences because the languages are not defined by avoiding large squares and a finite set of factors. Maybe it can be proven by the method used in [6] to recover many known results about the Fibonacci word.

Baker, McNulty, and Taylor [2] obtained that $ABXBAYACZCAWBC \cup \{\mathtt{02}\}$ characterizes the fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{21}$, $\mathtt{2} \mapsto \mathtt{03}$, $\mathtt{3} \mapsto \mathtt{23}$ over Σ_4. Notice that the forbidden factor 02 is not crucial here, its only role is to distinguish one out of three symmetric versions obtained by permutation of the alphabet letters. So, characterizations are known for the patterns AA, $ABABA$, $ABCABC$, $AABBCC$, $AABBCABBA$, and $ABXBAYACZCAWBC$. An interesting open question is the following: Suppose that P is an avoidable pattern with avoidability index $\lambda(P) = k$. Is it possible to find a finite set $\mathcal{P}$ of patterns and a finite set $\mathcal{F}$ of factors such that $P \in \mathcal{P}$ and $\mathcal{P} \cup \mathcal{F}$ characterizes a morphic word over Σ_k ? This would be a strengthening of Cassaigne's conjecture stating that there exists a morphic word avoiding P over Σ_k.

References

[1] G. Badkobeh. Fewest repetitions vs maximal-exponent powers in infinite binary words, *Theoret. Comput. Sci.* **412** (2011), 6625–6633.

[2] K.A. Baker, G.F. McNulty, and W. Taylor. Growth problems for avoidable words, *Theoret. Comput. Sci* **69** (1989), 319–345.

[3] J. Berstel. Axel Thue's Papers on Repetitions in Words: a Translation. *Publications du Laboratoire de Combinatoire et d'Informatique Mathématique. Université du Québec à Montréal*, Number 20, February 1995.

[4] F. Blanchet-Sadri, K. Black, and A. Zemke. Avoidable patterns in partial words. `http://www.uncg.edu/cmp/research/patterns/implementation.html`

[5] J. Cassaigne. An algorithm to test if a given circular HD0L-language avoids a pattern. Information processing '94, Vol. I (Hamburg, 1994), 459–464, IFIP Trans. A Comput. Sci. Tech., A-51, North-Holland, Amsterdam, 1994

[6] C. F. Du, H. Mousavi, L. Schaeffer, and J. Shallit. Decision algorithms for Fibonacci-automatic words, with applications to pattern avoidance. `arXiv:1406.0670`

[7] T. Harju and D. Nowotka. Binary words with few squares. *Bull. EATCS* **89** (2006), 164–166.

[8] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theoret. Informatics Appl.* **40** (2006), 427–441.

[9] P. Ochem. Binary words avoiding the pattern AABBCABBA. *RAIRO - Theoret. Informatics Appl.* **44(1)** (2010), 151–158.

[10] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[11] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.