RAIRO-Theor. Inf. Appl. 44 (2010) 151–158
DOI: 10.1051/ita/2010010 Available online at:
www.rairo-ita.org

BINARY WORDS AVOIDING THE PATTERN AABBCABBA

Pascal Ochem[1, 2]

Abstract. We show that there are three types of infinite words over the two-letter alphabet $\{0,1\}$ that avoid the pattern $AABBCABBA$. These types, P, E_0, and E_1, differ by the factor complexity and the asymptotic frequency of the letter 0. Type P has polynomial factor complexity and letter frequency $\frac{1}{2}$. Type E_0 has exponential factor complexity and the frequency of the letter 0 is at least 0.45622 and at most 0.48684. Type E_1 is obtained from type E_0 by exchanging 0 and 1.

Mathematics Subject Classification. 68R15.

1. Introduction

This paper deals with pattern avoidability [4,7]. Let Σ_s denote the s-letter alphabet $\{0,1,\ldots,s-1\}$. A *pattern* is a finite word over the alphabet of capital letters $\{A,B,\ldots\}$. An *occurrence* of a pattern is obtained by replacing each alphabet letter by a non-empty word. For example, the word 0111010011 is an occurrence of the pattern $ABBA$ where $A \mapsto 011$ and $B \mapsto 10$; it also contains another occurrence of this pattern (*i.e.* 1001) as a factor. A word *avoids* a pattern P if it contains no occurrence of P as a factor. The *avoidability index* $\lambda(P)$ of the pattern P is the smallest alphabet size over which an infinite word avoiding P exists. Patterns such as A, ABC, ABA, $ABACBA$ cannot be avoided with any finite alphabet. These patterns such that $\lambda(P)=\infty$ are said to be *unavoidable* and have been characterized by Zimin [11].

Let t_n be the number of words of length n in a language. If that language is closed under taking factors, which is the case for words avoiding a pattern, then t_n is sub-multiplicative and the *growth rate* $\lim_{n\to\infty}(t_n)^{\frac{1}{n}}$ is well-defined.

Keywords and phrases. Combinatorics on words, letter frequency, factor complexity.

[1] CNRS, Lab. J.V. Poncelet, Moscow; ochem@lri.fr
[2] LRI, Bât 490 Université Paris-Sud 11, 91405 Orsay Cedex, France.

Article published by EDP Sciences

See the survey of Berstel [3] for more information on the growth rate. For a given pattern P, once its avoidability index is known, it is interesting to consider the factor complexity of words avoiding P over $\Sigma_{\lambda(P)}$, in order to know whether P is "barely" or "easily" avoided over $\Sigma_{\lambda(P)}$. For example, it is known that $\lambda(ABDACEBAFCAGCB) = 4$ and that there are only polynomialy many words over Σ_4 avoiding that pattern [1], so their growth rate is 1. On the other hand, $\lambda(AA) = 3$ and there are exponentially many ternary square-free words, since their growth rate is > 1.30125 [6].

In this paper, we show that binary words avoiding $AABBCABBA$ can be classified into three disjoint types P, E_0, and E_1. Type E_1 is obtained from type E_0 by exchanging 0 and 1. There are polynomialy many words of type P and the asymptotic frequency of the letter 0 in words of type P is $\frac{1}{2}$. There are exponentially many words of type E_0 but their growth rate is small. When it is defined, the frequency of the letter 0 in an infinite word of type E_0 is between 0.45622 and 0.48684. Type E_1 is obtained from type E_0 by exchanging 0 and 1.

2. Three types of words avoiding AABBCABBA

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. We are interested in infinite binary recurrent words avoiding the pattern $AABBCABBA$. Such words equivalently avoid the formula $AABB.ABBA$ (see [4,5] for more on formulas). This means that for every occurrence of $AABB$ (*e.g.*, 000011) that appears, the corresponding occurrence of $ABBA$ (so, 001100) does not appear, or *vice versa*. To see this, suppose that both an occurrence of $AABB$ and the corresponding occurrence of $ABBA$ appear in an infinite recurrent word w. Since these two occurrences are recurrent factors in w, then w must contain, from left to right, the mentioned occurrence of $AABB$, followed by one letter, and then an infinite suffix that has to contain the corresponding occurrence of $ABBA$. This creates an occurrence of $AABBCABBA$.

Remark 2.1. An infinite recurrent word avoiding $AABBCABBA$ also avoids the patterns $AABBA$ and $AAAA$.

This remark is a straightforward consequence of the property on formulas mentioned above. An occurrence of $AABBA$ contains an occurrence of $AABB$ and the corresponding occurrence of $ABBA$. An occurrence of $AAAA$ is both an occurrence of $AABB$ such that $A = B$ and the corresponding occurrence of $ABBA$.

Figure 1 is a graph whose vertices are the occurrences of length 4 of $AABB$ or $ABBA$ that might be recurrent in an infinite binary word avoiding $AABBCABBA$. The factors 0000 and 1111 have been ruled out since they are occurrences of $AAAA$ (see Rem. 2.1). An edge stands for an incompatibility between an occurrence of $AABB$ and the corresponding occurrence of $ABBA$: two factors associated to adjacent vertices cannot be recurrent in a same infinite word avoiding $AABBCABBA$. So, given an infinite binary recurrent word w avoiding

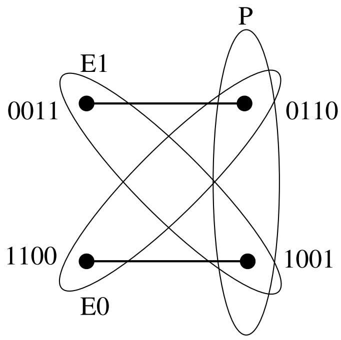

Figure 1. Graph of incompatibilities between factors of length 4.

$AABBCABBA$, we can associate the set of vertices of the graph that appear as factors in w. Moreover, this set is an independent set.

Let us check that neither an independent set of size at most one nor $\{0011, 1100\}$ can be associated to an infinite binary recurrent word avoiding $AABBCABBA$. By symmetry and maximality, we only need to consider the case of the sets $\{0110\}$ and $\{0011, 1100\}$. In the case of the set $\{0110\}$ (resp. $\{0011, 1100\}$), we can enumerate lexicographically all binary words avoiding the patterns $AABBCABBA$, $AABBA$, and $AAAA$, and the factors 0011, 1100, and 1001 (resp. the factors 0110 and 1001).

There remain three potential types for an infinite binary recurrent word avoiding $AABBCABBA$, that we call P, E_0, and E_1. These three types respectively contain factors in $\{0110, 1001\}$, $\{1100, 0110\}$, and $\{0011, 1001\}$. Notice that by exchanging 0 and 1, type P stays unchanged, type E_0 becomes type E_1, and type E_1 becomes type E_0.

3. Type P has polynomial growth

Let t be the fixed point of the morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$, and let h be the morphism defined by

$$\begin{aligned} 0 &\mapsto 00101101110111101001, \\ 1 &\mapsto 00101101101001, \\ 2 &\mapsto 00010. \end{aligned}$$

In this section, we give a characterization of words of type P:

Theorem 3.1. *The set of factors of type P is the set of factors of $h(t)$.*

The following lemma about t is needed in the proof of Theorem 3.1.

Lemma 3.2. *If w is an infinite recurrent ternary square-free word, the following assertions are equivalent:*

- *w has the same set of factors as t;*
- *w contains neither 010 nor 212 as a factor;*
- *w does not contain factors of the form $0v1v0$ with $v \in \Sigma_3^*$.*

Proof. The equivalence of the first and the second assertion is a well known result of Thue (see [2] for a translation). Let us prove the equivalence of the second and third assertion, which is that when considering recurrent languages of ternary square-free word, avoiding factors of the form $0v1v0$ with $v \in \Sigma_3^*$ is equivalent to avoiding the factors 010 and 212. Because of square-freeness, avoiding $0v1v0$ is equivalent to avoiding 010, 02120, and $02v'212v'20$ with $v' \in \Sigma_3^*$. Because it is a recurrent language, avoiding 02120 is equivalent to avoiding 212, since 02120 is the only possible extension of 212 that does not create a square. □

Let us prove one direction of Theorem 3.1, namely that $h(t)$ contains only factors of type P. Since t is recurrent, so is $h(t)$. Since $h(t)$ contains 0110 and 1001, it remains to check that $h(t)$ avoids $AABBCABBA$. First, we show that $h(t)$ contains no square xx with $|x| > 4$. It is easy to check that no such *large square* appears in the h-image of a factor of t of length at most two. Notice also that for every letter $i \in \Sigma_3$, the factor $h(i)$ appears only in $h(t)$ as the h-image of the letter i. This implies that any large square would be a factor of a word of the form $h(pvmvs)$ with $v \in \Sigma_3^*$, $p, m, s \in \Sigma_3$, $p \neq m$, and $m \neq s$. So there would be a large square also in $h(pms)$, which happens only in the case $pms = 010$. Since t contains no factors of the form $0v1v0$ by Lemma 3.2, $h(t)$ contains no square xx with $|x| > 4$. So we can list all the occurrences of the pattern $AABB$ in $h(t)$, because their length is at most 16. Then we can check that for every occurrence of the pattern $AABB$ in $h(t)$, the corresponding occurrence of $ABBA$ is not a factor of $h(t)$.

Now, we prove the other direction of Theorem 3.1, namely that every factor of type P is a factor of $h(t)$. First, we check that a factor of type P is a factor of the h-image of some ternary word. We consider the language P' of binary words avoiding 0011, 1100, $AAAA$, $AABBA$, and $AABBCABBA$. It contains P by Remark 2.1. We compute the set of factors in P' of length $|h(0)| + |h(1)| = 33$ and remove from this set factors that are not prolongable in P'. This can be done with the method described in Section 4, until this set becomes equal to the set of factors of $h(t)$ of length 33. In this set, every factor with prefix $h(i)$ for some $i \in \Sigma_3$ is such that the prefix $h(i)$ is followed by either $h((i+1) \pmod 3)$ or $h((i+2) \pmod 3)$. Thus, a factor of type P is a factor of the h-image of some ternary word.

Let $L \subset \Sigma_3^*$ denote the language of words whose h-image is of type P. Since factors of type P are recurrent, words in L are bi-prolongable in L. Let $u \in \Sigma_3^+$. We suppose now that L contains a square occurrence uu. Because of the prolongability, this implies that L contains a factor $puus$ for some $p, s \in \Sigma_3$. Since 00 is a common proper prefix of $h(1)$, $h(2)$, and $h(3)$, we can write $h(u) = 00r$ for some $r \in \Sigma_2^+$.

The following three cover every possible values of p and s. Each case is ruled out because it contains an occurrence of $AABBA$, which is forbidden by Remark 2.1.

- If $s = 2$, then $h(uu2) = 00r00r00010$ contains an occurrence of $AABBA$ with $A = 0$ and $B = r00$.
- If $p = 2$, then $h(2uus)$ has a prefix $0001000r00r00$ that contains an occurrence of $AABBA$ with $A = 0$ and $B = 0r0$.
- If $p, s \in \{0, 1\}$, then $h(puus)$ contains a factor $0100100r00r0010$ because 01001 is a common suffix of $h(0)$ and $h(1)$, and 0010 is a common prefix of $h(0)$ and $h(1)$. This factor is an occurrence of $AABBA$ with $A = 010$ and $B = 0r0$.

This shows that the language L contains square-free words only.

Factors of the form $0v1v0$ with $v \in \Sigma_3^*$ are not in L since their image by h contains the factor $1101001h(v)00101101101001h(v)0010110111$ which is an occurrence of $AABBA$ with $A = 1$ and $B = 01001h(v)001011011$.

To summarize, every factor of type P is a factor of the h-image of some recurrent ternary square-free word avoiding factors of the form $0v1v0$ with $v \in \Sigma_3^*$. By Lemma 3.2, every factor of type P is thus a factor of $h(t)$. This concludes the proof of Theorem 3.1.

As a corollary of Theorem 3.1, words of type P have polynomial growth.

4. Types E_0 and E_1 have exponential growth

Theorem 4.1. *The growth rate for words of type E_0 is between* 1.002584956 *and* 1.02930952.

Proof. For the lower bound, we extend the result [7] that the image of any ternary $\frac{7}{4}^+$-free word by the following 102-uniform morphism k avoids $AABBCABBA$.

$$\begin{aligned} 0 &\mapsto w0010110111011100010110001000101101110 \\ 1 &\mapsto w1100010110111011100010110001000101101 \\ 2 &\mapsto w1110001011000100010110111011000101101 \end{aligned}$$

with $w =$110001011011101110001011011101100010110111000101100010001011 01110.

These words avoiding $AABBCABBA$ are actually of type E_0 since they are recurrent and contain the factors 1100 and 0110.

Kolpakov [6] has shown that the growth rate of ternary $\frac{7}{4}^+$-free (resp. square-free) words is at least 1.245 (resp. 1.30125).

Ternary $\frac{7}{4}^+$-free words were used [7] as pre-image for k in order to have simple and standardized proofs. To get the lower bound of Theorem 4.1, we need the stronger statement that the k-image of any ternary square-free word avoids $AABBCABBA$. We can prove this by checking that the k-image of any ternary square-free word of length 3 contains no square xx with $|x| > 26$. Then again, for each occurrence of $AABB$ in the k-image of some ternary square-free word, we can check that the corresponding occurrence of $ABBA$ does not appear. The growth rate of words of type E_0 is thus at least $1.30125^{1/102} > 1.0025849$.

For the upper bound, we basically use our method [9] that gave an upper bound on the growth rate of ternary square-free words. We have noticed that the notion of prolongability is much more important for words of type E_0 than for ternary square-free words (maybe because the growth rate is much lower). For example, in a ternary square-free word pws such that $|w| = 50$ and $|p| = |s| = 15$, the factor w is very probably a recurrent factor in some infinite ternary square-free word. This is not the case for type E_0. We take this behavior into account by computing iteratively a set of words of some length avoiding $AABBCABBA$, 0011, and 1001 from another such set. These sets contain all words of type E_0 of the specified length but maybe also other words that are not prolongable. Let $f(n, e, S, k)$ be the function that computes the set of words w such that pws avoids $AABBCABBA$, 0011, and 1001, $|w| = n$, $|p| = |s| = e$, and every factor of length k of pws belongs to S, where S is a previously computed set of words of length k. For example (with fictional values), we can first compute a set of words of length 40 from scratch: $S_1 \leftarrow f(40, 5, \emptyset, 0)$. Then a set of words of length 50 from S_1: $S_2 \leftarrow f(50, 10, S_1, 40)$. Then another set of words of length 50 from S_2: $S_3 \leftarrow f(50, 10, S_2, 50)$. Of course, we have that $S_3 \subseteq S_2$ and hope that $S_3 \subset S_2$. Maybe even the set of prefixes of length 40 of words in S_3 is smaller than the initial set S_1. The user thus computes sets of words of increasing size and obtain a set of words that are prolongable by at least e letters, where e is the second parameter in the final call. Cassaigne [4] described a similar method using Rauzy graphs. We have obtained a set S of words of length 360 that are prolongable by 40 letters to the left and to the right.

The upper bound in Theorem 4.1 has been obtained by applying the transfer matrix method [9] with parameters $k = 359$ and $l = 101$. That is, we constructed a matrix M such that $M[i, j]$ is the number of factors of length $k + l = 460$ whose prefix (resp. suffix) is the ith (resp. jth) factor of length k. Then the upper bound is obtained by taking the lth root of the largest eigenvalue of M. Compared to the calculation described in [9], we made the following modifications: we used an adjacency list representation, because the matrix here is much sparser, and we required that only the words w of length $k+l$ such that every factor of w of length 360 belongs to S are taken into account in the matrix. Shur [10] presented another method for upper bounds on the growth rate that gives a better result for ternary square-free words. It would be interesting to check if his method also gives a better bound for words of type E_0. □

5. Letter frequencies

Let $|v|_i$ denote the number of occurrences of the letter i in the finite word v.

Theorem 5.1. *Let w be an infinite recurrent word avoiding $AABBCABBA$. For all $\varepsilon > 0$, there exists an integer n_ε such that the frequency $\frac{|v|_0}{|v|}$ of the letter 0 in* *every finite factor v of w with length at least n_ε is in*

- $\left[\frac{1}{2}-\varepsilon, \frac{1}{2}+\varepsilon\right]$ *if w is of type P;*
- $\left[\frac{271}{594}-\varepsilon, \frac{37}{76}+\varepsilon\right]$ *if w is of type E_0;*
- $\left[\frac{39}{76}-\varepsilon, \frac{323}{594}+\varepsilon\right]$ *if w is of type E_1.*

Proof. Let us check that infinite words of type P have letter frequency $\frac{1}{2}$. It is well-known (and easy to check) that the letters of Σ_3 have equal frequencies in the fixed point t of the morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$. Now, by Theorem 3.1, words of type P are factors of the image of t by a morphism h that satisfies $|h(0)|_0+|h(1)|_0+|h(2)|_0 = |h(0)|_1+|h(1)|_1+|h(2)|_1$.

For types E_0 and E_1, we only have to compute lower bounds, since if x is a lower bound on the frequency of the letter 0 for type E_i, then $(1-x)$ is an upper bound on the frequency of the letter 0 for type E_{1-i}. These lower bounds were obtained using our method [8] with a "suffix cover". A suffix cover C of a language L is a set of factors such that every large enough and prolongable enough word in L has a suffix that belongs to C. We used the suffix cover $C_0 = \{00, 1101110001011000100010, 1100010, 110, 1\}$ for type E_0, and the suffix cover $C_1 = \{00111010010001000, 01110111010010001000, 0100011101001000, 010010001001110100100, 0111011101001000, 0100, 010, 01110, 1\}$ for type E_1.

To check that C_0 is a suffix cover of E_0, it is sufficient to verify that every word in the set S computed in Section 4 has a suffix in C_0, because S contains every factor of type E_0 of length 360. We also check that the complement of every word in S has a suffix in C_1. Now, to prove for example that the asymptotic frequency of the letter 0 is at least $\frac{271}{594}$ in an infinite word of type E_0, we verify with backtracking that, for every $u \in C_0$, there exists no right infinite binary word w such that uw is of type E_0 and $\frac{|p|_0}{|p|} < \frac{271}{594}$ for every finite prefix p of w. □

It is noticeable that these three sets of potential frequencies are disjoint: if w is an infinite binary recurrent word avoiding $AABBCABBA$ with defined letter frequencies, then the frequency of 0 is in $\left[\frac{271}{594}, \frac{37}{76}\right] \cup \left\{\frac{1}{2}\right\} \cup \left[\frac{39}{76}, \frac{323}{594}\right] = [0.45622\ldots, 0.48684\ldots] \cup \{0.5\} \cup [0.51315\ldots, 0.54377\ldots]$. The infinite words of type E_0 obtained by the construction in [7] and in Section 4 are of type E_0 and the frequency of the letter 0 is $\frac{48}{102} = \frac{8}{17} = 0.47058\ldots$

6. Conclusion

Infinite binary recurrent words avoiding $AABBCABBA$ split into three types when considering the factors of length 4. Informally, such splittings happen because the letter C appears only once in the pattern, but is not necessarily related to the length of factors. Nothing prevents a priori from further sub-splittings into sub-types when considering larger factor lengths. Type P obviously cannot be split. Since types E_0 and E_1 are symmetrical, we can focus on type E_0 and consider the set S of words of type E_0 of length 360 discussed in Section 4. We have checked that for every two (distinct) words $w_1, w_2 \in S$, and for every occurrence of $AABB$ appearing in w_1, the corresponding occurrence of $ABBA$ does not appear in w_2. This means that no sub-splitting happens for length 360. We leave as an open question whether such a sub-splitting exists.

We do not know how to prove a negative answer. A positive answer could be obtained by constructing an infinite word of type E_0 containing a particular occurrence of $AABB$ (as a recurrent factor) and another one containing the corresponding occurrence of $ABBA$.

References

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

[2] J. Berstel, Axel Thue's papers on repetitions in words: a translation. *Publications du LaCIM*, Département de mathématiques et d'informatique **95**, Université du Québec à Montréal (1995). http://www-igm.univ-mlv.fr/~berstel/Articles/1994ThueTranslation.pdf

[3] J. Berstel, Growth of repetition-free words - a review. *Theoret. Comput. Sci.* **340** (2005) 280–290.

[4] J. Cassaigne, *Motifs évitables et régularité dans les mots.* Thèse de Doctorat, Université Paris VI (1994).

[5] R.J. Clark, *Avoidable formulas in combinatorics on words.* Ph.D. thesis, University of California, Los Angeles (2001).

[6] R. Kolpakov, Efficient lower bounds on the number of repetition-free words. *J. Integer Sequences* **10** (2007) Article 07.3.2.

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

[8] P. Ochem, Letter frequency in infinite repetition-free words. *Theoret. Comput. Sci.* **380** (2007) 388–392.

[9] P. Ochem and T. Reix, Upper bound on the number of ternary square-free words, *Proceedings of the Workshop on Words and Automata (WOWA'06)* (St Petersburg, June 2006). http://www.lri.fr/~ochem/morphisms/wowa.ps

[10] A.M. Shur, Combinatorial complexity of regular languages. *CSR 2008. Lect. Notes Comput. Sci.* **5010** (2008) 289–301.

[11] A.I. Zimin, Blocking sets of terms. *Math. USSR Sbornik* **47** (1984) 353–364. English translation.