Problems in between words and abelian words: k-abelian avoidability [1]

Mari Huova, Juhani Karhumäki and Aleksi Saarela

Department of Mathematics and TUCS
University of Turku
20014 Turku, FINLAND
email: {mari.huova, karhumak, aleksi.saarela}@utu.fi

Abstract

We consider a recently defined notion of k-abelian equivalence of words in connection with avoidability problems. This equivalence relation, for a fixed natural number k, takes into account the numbers of occurrences of the different factors of length k and the prefix and the suffix of length $k-1$. We search for the smallest alphabet in which k-abelian squares and cubes can be avoided, respectively. For 2-abelian squares this is four – as in the case of *abelian words*, while for 2-abelian cubes we have only strong evidence that the size is two – as it is in the case of *words*. However, we are able to prove this optimal value only for 8-abelian cubes.

Keywords: combinatorics on words, k-abelian equivalence, avoidability

1. Introduction

In Combinatorics on Words the theory of avoidability is one of the oldest and most studied topics, and Axel Thue can be referred to as the initiator of this area with the first results at the beginning of 20th century [Th1, Th2]. Thue proved, for example, the existence of an infinite binary word not containing any factor of the word three times consecutively, that is the existence of a cube-free word. Similarly, he showed that squares can be avoided in infinite ternary words.

Avoidability questions have been studied since late 1960s for *abelian equality*, the commutative variant of equality where only the number of each letter counts and not their order. Evdokimov [Ev] showed apparently as the first nontrivial result that commutative squares could be avoided in infinite words over a 25-letter alphabet. The size of the alphabet was reduced to five by Pleasant [Pl] and the optimal value, four, was found by Keränen [Ke]. Dekking [De] managed

[1]Supported by the Academy of Finland under the grant 121419.

to prove already earlier that the optimal value for the size of the alphabet where abelian cubes are avoidable is three.

We introduce in this note new variants of the problems by defining repetitions via new equivalence relations which lie properly in between equality and abelian equality. For this relation we use a notion *k-abelian equivalence*, where $k \geq 1$ is a natural number. We notice that 1-abelian equivalence means the usual abelian equivalence. Modifying the above Thue's problems we ask for what are the smallest alphabets where k-abelian squares and cubes can be avoided. A goal of this note is to point out that these problems are not trivial even for small values of k and that in some cases the behaviour of k-abelian equivalence is similar to the word equivalence and in some cases to the abelian equivalence. The main result is to show the existence of an infinite 8-abelian cube-free word over binary alphabet.

2. Basic notions and problems

For the basic terminology of words as well as avoidability we refer to [Lo] and [CK]. Here we define only our basic notions and problems. Our basic notion is *k-abelian equivalence* of words, see [KSZ]. Let $k \geq 1$ be a natural number. We say that words u and v in Σ^+ are *k-abelian equivalent*, in symbols $u \equiv_{a,k} v$, if

1. $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ and $\mathrm{suf}_{k-1}(u) = \mathrm{suf}_{k-1}(v)$, and
2. for all $w \in \Sigma^k$, the number of occurrences of w in u and v coincide, i.e. $|u|_w = |v|_w$.

Here pref_{k-1} (resp. suf_{k-1}) is used to denote the prefixes (resp. suffixes) of length $k-1$.

It is straightforward to see that $\equiv_{a,k}$ is an equivalence relation. Because of the first condition, it is also a congruence, that is $u \equiv_{a,k} u'$ and $v \equiv_{a,k} v'$ imply $uv \equiv_{a,k} u'v'$.

The first condition also makes the relation a sharpening of abelian equality. It allows us to count the number of occurrences of each letter by counting the occurrences of the letters in each factor of length k and taking into account the special role of the letters of the prefix and the suffix of length $k-1$. Indeed, words *aba* and *bab* have the same number of occurrences of the factors of length 2 but they are not abelian equivalent.

More generally, k-abelian equality implies $(k-1)$-abelian equality, so

$$u = v \Rightarrow u \equiv_{a,k} v \Rightarrow u \equiv_{a,k-1} v \Rightarrow \cdots \Rightarrow u \equiv_{a,2} v \Rightarrow u \equiv_a v,$$

where $\equiv_a$ denotes the abelian equivalence (which is the same as 1-abelian equivalence). Also

$$u = v \Leftrightarrow u \equiv_{a,k} v \quad \forall\, k \geq 1.$$

Now, notions like *k-abelian repetitions* are naturally defined. For instance, $w = uv$ is a *k-abelian square* if and only if $u \equiv_{a,k} v$. The basic problem we are considering is k-abelian avoidability. We ask what is the size of the smallest alphabet where k-abelian squares or cubes can be avoided for a fixed k. We recall that an alphabet avoids, for example, k-abelian squares if there exists such an infinite word over this alphabet that it does not contain any k-abelian squares as a factor.

3. Preliminaries

Before concentrating on k-abelian avoidability we mention examples of two morphisms in classical area. We use a special technique to prove that these particularly simple morphisms generate a binary cube-free and a ternary square-free word. There are some similarities with the technique used in the proof of main theorem, i.e., the form of the morphism restricts in an obvious way the possible lengths of squares and gives an easy way to determine the positions of some factors in the infinite word.

Example 1. (Due to a solution of P. Sarvamaa in a course on Combinatorics on Words at University of Turku.) Consider the morphism

$$h : \begin{cases} a \mapsto aab \\ b \mapsto abb \end{cases},$$

that is $h(x) = axb$ for $x \in \{a, b\}$. It defines a cube-free word, as is particularly simple to see. Let

$$w = h^{\infty}(a) = \prod_{i=1}^{\infty} ax_i b,$$

where each $x_i \in \{a, b\}$ and, moreover,

$$w = h^{-1}(w) = x_1 x_2 \dots . \tag{1}$$

Now, assume that w contains a shortest cube $u_1u_2u_3$ with $u_1 = u_2 = u_3 = u$. If $|u|$ would not be divisible by three, then the first letters of u_1, u_2 and u_3 would be in different positions inside the factors ax_ib and hence both a and b would be among these first letters. So $|u|$ must be divisible by three. Denote by v_i, for $i = 1, 2, 3$, the scattered subword of u_i formed by the occurrences of x_j in u_i. Since $|u|$ is divisible by three, necessarily $v_1 = v_2 = v_3$, and hence by (1) w contains a shorter cube $v_1v_2v_3$, which is a contradiction.

Example 2. Let h and w be as in Example 1. The critical exponent of w is 3. This can be seen as follows. If w has a factor $uauau$, then it also has a factor $v = buauaub$, and thus a factor $h(v) = abbh(u)aabh(u)aabh(u)abb$. Now $h(v)$ contains a factor $u'au'au'$, where $u' = bh(u)a$. Because aa is a factor of w, it follows by induction that w has factors of the form $uauau$ for arbitrarily long words u.

Example 3. Consider the morphism

$$g : \begin{cases} a \mapsto abcbacbcabcba \\ b \mapsto bcacbacabcacb \\ c \mapsto cabacbabcabac \end{cases} .$$

Let

$$t = g^{\infty}(a) = \prod_{i=1}^{\infty} x_i s_i,$$

where $x_i \in \{a, b, c\}$, $x_i s_i = g(x_i)$ and, moreover,

$$t = g^{-1}(t) = x_1 x_2 \dots . \tag{2}$$

It was proved by Leech [Le] that the infinite word t is square-free. The words $g(a)$, $g(b)$ and $g(c)$ have equal length, are palindromes and can be obtained from each other by cyclically permuting the three letters. The morphism g is the simplest square-free morphism with these symmetry properties, but there are shorter uniform morphisms that are square-free [Ze].

The square-freeness can be proved in a way that is quite similar to the proof in Example 1. However, there are more details that need to be checked: it must be verified that t does not contain a square of a word of length less than eight, and that the *starting position* of every factor of length eight is uniquely determined modulo $|g(a)| = 13$. These two conditions can be checked mechanically.

Now, assume that t contains a shortest square $u_1 u_2$ with $u_1 = u_2 = u$. Then $|u| \geq 8$. If $|u|$ would not be divisible by 13, then the prefixes of u_1 and u_2 of length eight would be in different positions modulo 13, and hence they would be different. So $|u|$ must be divisible by 13. Denote by v_i, for $i = 1, 2$, the scattered subword of u_i formed by the occurrences of x_j in u_i. Since $|u|$ is divisible by 13, necessarily $v_1 = v_2$, and hence by (2) t contains a shorter square $v_1 v_2$, which is a contradiction.

Example 4. The word t in Example 3 has a repetition of order 15/8: $g(aba)$ contains the factor $ag(b)a = abcacbacabcacba$. The proof in Example 3 can be modified to show that there are no higher powers.

Research on k-abelian equivalences was initiated and, for example, the characterizations of 2- and 3-abelian equivalence classes over binary alphabet were given in [HKSS]. From these characterizations it is possible to conclude that the number of equivalence classes of binary 2-abelian words of length n is $n^2 - n + 2$ and $\Theta(n^4)$ in the case of 3-abelian words, see [HKSS]. In the general case the number of k-abelian equivalence classes of words of length n is polynomial in n but the degree of the polynomial grows exponentially in k, see [HKSS] and [KSZ].

Next we discuss k-abelian avoidability and concentrate on the case $k = 2$. We try to find the size of the smallest alphabet avoiding 2-abelian squares and

2-abelian cubes, respectively. Before going into our problems we recall the following Table 1 which summarizes the results we mentioned at the beginning and at the same time tells the limits of our problems.

Avoidability of squares				Avoidability of cubes			
	type of rep.				type of rep.		
size of the alph.	$=$	$\equiv_{a,2}$	$\equiv_a$	size of the alph.	$=$	$\equiv_{a,2}$	$\equiv_a$
2	$-$	$-$	$-$	2	$+$	?	$-$
3	$+$	?	$-$	3	$+$	$+$	$+$
4	$+$	$+$	$+$				

Table 1: Avoidability of different types of repetitions in infinite words.

Our next example shows that the ordinary method of iterating a morphism might not give answers to our problems.

Example 5. In each of the following known cases where a repetition free infinite word is obtained by iterating a morphism, a 2-abelian cube is found fairly early from the beginning. If anything else is not mentioned see [AS] for the reference of the following words.

- Infinite overlap-free Thue-Morse word (morphism: $0 \to 01$, $1 \to 10$): $01\,\overbrace{101001}\,\overbrace{100101}\,\overbrace{101001}\,011...$
- Cube-free infinite word (morphism: $0 \to 001$, $1 \to 011$): $001001\,\overbrace{011001}\,\overbrace{001011}\,\overbrace{001011}\,011...$
- Morphism $0 \to 001011$, $1 \to 001101$, $2 \to 011001$ maps ternary cube-free words to binary cube-free words, see [Br], but $001011 \equiv_{a,2} 001101 \equiv_{a,2} 011001$, thus images of all words mapped with this morphism contain 2-abelian cubes.
- A binary sequence called Kolakoski sequence is cube-free, see [Ca] and [Lep], but not 2-abelian cube-free: $122\,\overbrace{112122}\,\overbrace{122112}\,\overbrace{112212}\,112...$ (It is an open question whether the Kolakoski sequence is a morphic word.)
- A binary overlap-free word w can also be gained in form $w = c_0c_1c_2\ldots$, where c_n means the number of zeros (mod 2) in the binary expansion of n. Again, a 2-abelian cube of length 6 begins as early as from the fifth letter: $w = 0010\,\overbrace{011010}\,\overbrace{010110}\,\overbrace{011010}\,011...$

By computer checking we were able to decide the size of the smallest alphabet avoiding 2-abelian squares. This solves our first question mark in Table 1.

Theorem 1. *The longest ternary word which is 2-abelian square-free has length 537, which shows that there does not exist an infinite 2-abelian square-free word* *over a ternary alphabet. This longest word is unique up to the permutations of the letters.*

We generated this longest ternary word avoiding 2-abelian squares with a computer program, for the explicit word, see [HKSS] and [HK]. The correctness of the result has been checked by two independent computer programs.

We also counted the number of ternary 2-abelian square-free words of each length. The result of this computer based analysis is shown in Figure 1. The shape of the function is a bit surprising here!

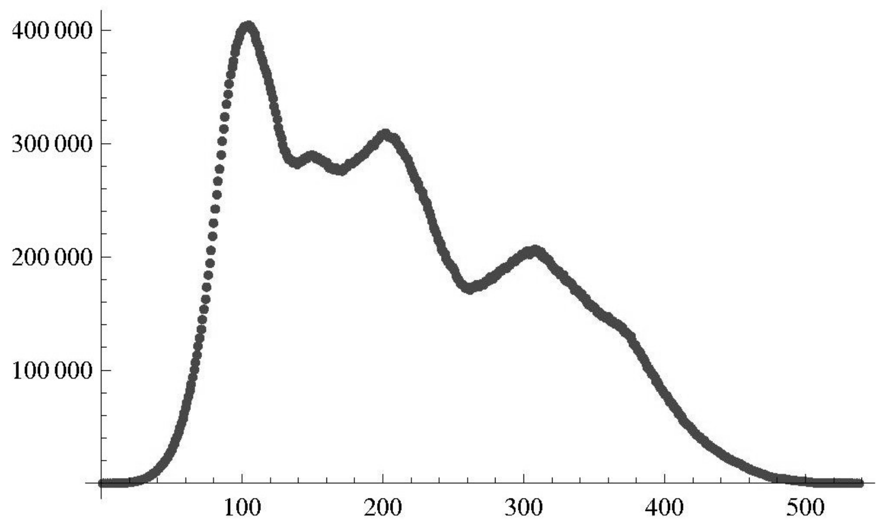

Figure 1: The number of 2-abelian square-free words with respect to their lengths.

To solve the other question mark in Table 1 we also did some computer checking – and obtained evidence that the answer is likely to be different compared to the first one.

Example 6. With a computer we were able to construct a binary word of more than 100 000 letters that still avoids 2-abelian cubes. This shows that there exist, at least, very long binary 2-abelian cube-free words.

The idea of the program is the same as in the program generating the longest ternary word avoiding 2-abelian squares. The correctness of the program generating the word of Example 6 is supported by several independent tests. It has been checked independently for a few words of lengths up to 10 000 letters generated by the program that they are indeed 2-abelian cube-free. In addition, we also checked 2-abelian cube-freeness of the mirror image of the word of length

100 000 letters and the word obtained by exchanging the letters a and b in this original word. The results were as expected and this supports the correctness of the original program. By changing a letter in the word generated by the program and causing a 2-abelian cube we have checked that the program will find 2-abelian cubes.

Example 7. We searched the number of binary 2-abelian cube-free words of a given length by computer analysis. The numbers of the words with lengths from 1 to 60 grow approximately with a factor 1.3 at each increment of the length. So that the number of binary 2-abelian cube-free words of length 60 is already 478 456 030. And already, with length 12 there exist more binary 2-abelian cube-free words (254) than ternary 2-abelian square-free words (240).

We also chose some binary 2-abelian cube-free prefixes and counted the number of suitable extensions for these, i.e. the number of binary 2-abelian cube-free words having these fixed prefixes. As a result we found examples of binary 2-abelian cube-free words with a property that the number of their extensions grows again approximately with a factor 1.3 when increasing the length of extensions by one.

These examples support the conjecture that there would exist an infinite binary word that avoids 2-abelian cubes. However, we were not able to prove this, but as shown in the next section, we were able to conclude this for 8-abelian cubes.

4. Main result

In the case $k = 8$ we have an affirmative answer to our problem: 8-abelian cubes can be avoided in two letter alphabet. It is known that abelian squares can be avoided in 4-letter alphabet, see [Ke] − a result being far from trivial. We show that starting from such a word and mapping with a uniform morphism we can produce an infinite binary 8-abelian cube-free word.

We need the following notation. If $u = a_0 \dots a_{n-1}$, where a_i are letters and $0 \le i \le j \le n$, then we let $u[i..j] = a_i \dots a_{j-1}$.

Theorem 2. *Let $w \in \{0, 1, 2, 3\}^\omega$ be an abelian square-free word. Let $k \le n$ and $h : \{0, 1, 2, 3\}^* \to \{0, 1\}^*$ be an n-uniform morphism that satisfies the following conditions:*

1. *if $u \in \{0, 1, 2, 3\}^4$ is square-free, then $h(u)$ is k-abelian cube-free,*
2. *if $u \in \{0, 1, 2, 3\}^*$ and v is a factor of $h(u)$ of length $2k - 2$, then every occurrence of v in $h(u)$ has the same starting position modulo n,*
3. *there is a number i such that $0 \le i \le n - k$ and for at least three letters $x \in \{0, 1, 2, 3\}$, $v = h(x)[i..i + k]$ satisfies $|h(u)|_v = |u|_x$ for every $u \in \{0, 1, 2, 3\}^*$.*

Then $h(w)$ is k-abelian cube-free.

Proof. The first condition prohibits short k-abelian cubes in $h(w)$. If $h(w)$ contained a k-abelian cube of length less than $3k$, then this cube would be a factor of $h(u)$ for some $u \in \{0,1,2,3\}^4$, where u is a factor of w and thus square-free.

The second condition restricts the length of every k-abelian cube in $h(w)$ to be divisible by $3n$. If $h(w)$ contained a k-abelian cube pqr, where $|p| = |q| = |r| = m \geq k$, then

$$p[m-k+1..m] \cdot q[0..k-1] = q[m-k+1..m] \cdot r[0..k-1].$$

and the starting positions of these factors would differ by m. Now m is divisible by n showing that the length of every k-abelian cube in $h(w)$ is divisible by $3n$.

By using the third condition we show that a k-abelian cube in $h(w)$ would lead to an abelian square in w. Let $a'a_1 \dots a_s b' b_1 \dots b_s c' c_1 \dots c_s d'$ be a factor of w, where $a_j, b_j, c_j, a', b', c' \in \{0,1,2,3\}$, so that pqr is a k-abelian cube in $h(w)$ with

$$p = p_1 h(a_1 \dots a_s) p_2, \quad q = q_1 h(b_1 \dots b_s) q_2, \quad r = r_1 h(c_1 \dots c_s) r_2,$$

where p_1 is a suffix of $h(a')$, $p_2 q_1 = h(b')$, $q_2 r_1 = h(c')$, r_2 is a prefix of $h(d')$, $|p_1| = |q_1| = |r_1|$ and $|p_2| = |q_2| = |r_2|$. Let i be the number and a, b, c the three letters in condition 3. Let $|p| = m$, $|p_2| = j$ and $v_x = h(x)[i..i+k]$ for $x \in \{a,b,c\}$. There are three cases.

If $j \leq i$, then p_2 is too short to contain v_x and $h(a')$ contains v_x if and only if p_1 contains v_x for $x \in \{a,b,c\}$. Similarly for q_2, $h(b')$ and q_1. This gives by condition 3

$$|a'a_1 \dots a_s|_x = |h(a'a_1 \dots a_s)|_{v_x} = |p|_{v_x} = |q|_{v_x} = |h(b'b_1 \dots b_s)|_{v_x} = |b'b_1 \dots b_s|_x$$

for $x \in \{a,b,c\}$. Thus $a'a_1 \dots a_s$ and $b'b_1 \dots b_s$ are abelian equivalent, which contradicts the abelian square-freeness of w.

If $j \geq i+k$, then respectively

$$|a_1 \dots a_s b'|_x = |h(a_1 \dots a_s b')|_{v_x} = |p|_{v_x} = |q|_{v_x} = |h(b_1 \dots b_s c')|_{v_x} = |b_1 \dots b_s c'|_x$$

for $x \in \{a,b,c\}$, so $a_1 \dots a_s b'$ and $b_1 \dots b_s c'$ are abelian equivalent, which is a contradiction.

If $i < j < i+k$, then any of $p_1,\ p_2,\ q_1$ or q_2 cannot contain v_x for $x \in \{a,b,c\}$, which gives

$$|a_1 \dots a_s|_x = |h(a_1 \dots a_s)|_{v_x} = |p|_{v_x} = |q|_{v_x} = |h(b_1 \dots b_s)|_{v_x} = |b_1 \dots b_s|_x$$

for $x \in \{a,b,c\}$. Further, $v_{b'}$ is a factor of $t = p[m-k+1..m]q[0..k-1]$ and $v_{c'}$ is a factor of $q[m-k+1..m]r[0..k-1]$, which is the same word as t. Now $v_{b'}$ and $v_{c'}$ have the same starting positions in t, so $v_{b'} = v_{c'}$, and $b' = c'$ by condition 3. Thus $a_1 \dots a_s b'$ and $b_1 \dots b_s c'$ are abelian equivalent. This contradiction completes the proof. $\square$

Now we are ready for our main theorem.

Theorem 3. *Let $w \in \{0,1,2,3\}^\omega$ be an abelian square-free word. Let $h : \{0,1,2,3\}^* \to \{0,1\}^*$ be the morphism defined by*

$$h(0) = 00101\ 0\ 011001\ 0\ 01011,$$
$$h(1) = 00101\ 0\ 011001\ 1\ 01011,$$
$$h(2) = 00101\ 1\ 011001\ 0\ 01011,$$
$$h(3) = 00101\ 1\ 011001\ 1\ 01011.$$

Then $h(w)$ is 8-abelian cube-free.

Proof. Condition 1 of Theorem 2 is satisfied for $k \geq 4$.

Condition 2 of Theorem 2 is satisfied for $k \geq 6$.

Condition 3 of Theorem 2 is satisfied for $k = 8$ and $i = 5$. The three letters are 0, 1 and 3, and the corresponding factors are 00110010, 00110011 and 10110011.

The claim now follows from Theorem 2. □

The satisfiability of the three conditions in the previous proof has been checked by computer as well as by paper and pencil. The computations are not tedious.

We remark that for the case $k = 2$ the requirements of the last condition are too strict because there exist only four different binary words of length 2. Thus Theorem 2 cannot be applied for the case $k = 2$.

It is interesting to note that the infinite word avoiding 8-abelian cubes is a morphic word, that is obtained as a morphic image (in fact by a uniform morphism) of a word defined by iterating a morphism. Indeed, the solution of Keränen uses a method of iterating a morphism. Even more strongly our solution is given by an automatic sequence, see [AS].

As a conclusion, we know that 2-abelian square-freeness behaves like abelian square-freeness and it seems that 2-abelian cube-freeness would behave like ordinary cube-freeness. Main theorem also shows that in the case $k = 8$ the behaviour of k-abelian cube-freeness is similar to the behaviour of cube-freeness of words. These results reinforce the impression of k-abelian equality to represent an equivalence in between equality and abelian equality. In some cases the k-abelian equivalence resembles more the ordinary word equivalence and in some cases the abelian one.

References

[AS] J.-P. Allouche, J. Shallit: *Automatic Sequences: Theory, Applications, Generalizations*. Cambridge University Press, Cambridge (2003).

[Br] F.-J. Brandenburg: *Uniformly growing k-th power-free homomorphisms*. Theoret. Comput. Sci. 23, 69–82 (1983).

[Ca] A. Carpi: *On repeated factors in C^∞-words*. Inform. Process. Lett., vol. 52(6), 289–294 (1994).

[CK] C. Choffrut, J. Karhumäki: *Combinatorics of words*. In: G. Rozenberg, A. Salomaa (eds.) Handbook of Formal Languages, vol. 1, 329—438. Springer, Heidelberg (1997)

[De] F. M. Dekking: *Strongly non-repetitive sequences and progression-free sets*. J. Combin. Theory Ser. A 27(2), 181–185 (1979).

[Ev] A. A. Evdokimov: *Strongly asymmetric sequences generated by a finite number of symbols*. Dokl. Akad. Nauk SSSR 179, 1268–1271 (1968); English translation in Soviet Math. Dokl. 9, 536–539 (1968).

[HK] M. Huova, J. Karhumäki: *Observations and Problems on k-abelian avoidability*. In Dagstuhl Seminar Combinatorial and Algorithmic Aspects of Sequence Processing, arXiv:1104.4273, (2011).

[HKSS] M. Huova, J. Karhumäki, A. Saarela, K. Saari: *Local squares, periodicity and finite automata*. In: C.S. Calude, G. Rozenberg, A. Salomaa (eds.) Rainbow of Computer Science, LNCS, vol. 6570, 90—101. Springer, Heidelberg (2011).

[KSZ] J. Karhumäki, A. Saarela, L. Zamboni: *Generalized abelian equivalence*. Manuscript to be published.

[Ke] V. Keränen: *Abelian squares are avoidable on 4 letters*. In: W. Kuich (ed.) ICALP 1992, LNCS, vol. 623, 41–52. Springer, Heidelberg (1992).

[Le] J. Leech: *A problem on strings of beads*. Math. Gazette 41, 277–278 (1957).

[Lep] A. Lepistö: *Repetitions in Kolakoski sequence*. In: G. Rozenberg, A. Salomaa (eds.) Developments in Language Theory "At the Crossroads of Mathematics, Computer Science and Biology", World Scientific, 130–143 (1994).

[Lo] M. Lothaire: *Combinatorics on words*. Addison-Wesley, Reading (1983).

[Pl] P. A. B. Pleasant: *Non-repetitive sequences*. Proc. Cambridge Philos. Soc. 68, 267–274 (1970).

[Th1] A. Thue: *Über unendliche Zeichenreihen*. Norske vid. Selsk. Skr. Mat. Nat. Kl. 7, 1–22 (1906).

[Th2] A. Thue: *Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen*. Norske vid. Selsk. Skr. Mat. Nat. Kl. 1, 1–67 (1912).

[Ze] T. Zech: *Wiederholungsfreie Folgen*. Z. Angew. Math. Mech. 38, 206–209 (1958).