Vesa Halava | Tero Harju | Tomi Kärki

Overlap-freeness in infinite partial words

Turku Centre *for* Computer Science

TUCS Technical Report
No 888, February 2008

Overlap-freeness in infinite partial words

Vesa Halava
Department of Mathematics and
TUCS - Turku Centre for Computer Science
University of Turku, FI-20014 Turku, Finland
vehalava@utu.fi
Tero Harju
Department of Mathematics and
TUCS - Turku Centre for Computer Science
University of Turku, FI-20014 Turku, Finland
harju@utu.fi
Tomi Kärki
Department of Mathematics and
TUCS - Turku Centre for Computer Science
University of Turku, FI-20014 Turku, Finland
topeka@utu.fi

TUCS Technical Report
No 888, February 2008

Abstract

We prove that there exist infinitely many infinite overlap-free binary partial words with one hole. Moreover, we show that there exist infinitely many binary partial words with an infinite number of holes which are 3-overlap-free, i.e., they are cube-free and they do not contain a factor of the form $xyxyx$ where the length of x is at least three and y is nonempty.

Keywords: Repetition-freeness, k-free, overlap, partial words, Thue-Morse word, infinite words

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

Repetitions, i.e., consecutive occurrences of words within a word and especially repetition-freeness have been fundamental research subjects in combinatorics on words since the seminal papers of Thue [13, 14] in the beginning of the 20th century. Thue showed that there exists an infinite word w over a 3-letter alphabet, which does not contain any squares xx, where x is a nonempty word in w. Moreover, he constructed an infinite binary word t which does not contain any overlaps $xyxyx$ for any nonempty words x and y. This celebrated word is nowadays called the Thue-Morse word, which has many surprising and remarkable properties; see [16]. As an example, we mention applying t for designing an unending play of chess [5, 10] and for solving the Burnside problem for groups [1] and semigroups [11, 12].

In [9] Manea and Mercaş considered repetition-freeness of partial words. Partial words are words with "do not know"-symbols called holes and they were first introduced by Berstel and Boasson in [2]. Motivation for the study of partial words comes from applications in word algorithms and molecular biology, in particular; see [3] for using partial words in DNA sequencing and DNA comparison. The theory of partial words has developed rapidly in the recent years and many classical topics in combinatorics on words have been revisited. Topics such as periodicity, primitivity, unbordered word, codes and equations have been considered in the first book on partial words authored by Blanchet-Sadri in 2007 [4]. See also related works by Shur and Gamzova [15], Leupold [7] and Lischke [8]. As another approach for modeling missing or uncertain information in words we want to mention word relations, a generalization of the compatibility of partial words introduced in [6].

It was shown in [9] that there exist infinitely many cube-free binary partial words containing an infinite number of holes. In this paper we show that this result can be improved by giving three theorems with short and simple proofs. First, we prove that there exist infinitely many infinite overlap-free binary partial words with one hole. Secondly, we show that an infinite overlap-free binary partial word cannot contain more than one hole. However, we prove that a binary partial word with an infinite number of holes can be "almost overlap-free". More precisely, there exist infinitely many cube-free binary partial words with an infinite number of holes which do not contain a factor of the form $xyxyx$ where the length of x is at least three and y is nonempty.

2 Preliminaries

Let $\mathcal{A}$ be a finite alphabet. The elements of $\mathcal{A}$ are called *letters*. A word $w = a_1a_2\cdots a_n$ of length n over the alphabet $\mathcal{A}$ is a mapping $w\colon \{1, 2, \ldots, n\} \to \mathcal{A}$ such that $w(i) = a_i$. The length of a word w is denoted by $|w|$, and ε is the

empty word of length zero. By a (right) infinite word $w = a_1a_2a_3\cdots$ we mean a mapping w from the positive integers $\mathbb{N}_+$ to the alphabet $\mathcal{A}$ such that $w(i) = a_i$. The set of all finite words is denoted by $\mathcal{A}^*$, infinite words are denoted by $\mathcal{A}^\omega$ and $\mathcal{A}^+ = \mathcal{A}^* \setminus \{\varepsilon\}$. A finite word v is a *factor* of w if $w = xvy$, where $x \in \mathcal{A}^*$ and $y \in \mathcal{A}^* \cup \mathcal{A}^\omega$. The set of factors of w is denoted by $F(w)$. If $x = \varepsilon$, then v is a *prefix* of w. A prefix of w of length n is denoted by $\mathrm{pref}_n(w)$. If $v \in \mathcal{A}^* \cup \mathcal{A}^\omega$ and $w = xv$, then v is called a *suffix* of w.

A partial word u of length n over the alphabet $\mathcal{A}$ is a partial function $u\colon \{1, 2, \ldots, n\} \to \mathcal{A}$. The *domain* $D(u)$ is the set of positions $i \in \{1, 2, \ldots, n\}$ where $u(i)$ is defined. The set $H(u) = \{1, 2, \ldots, n\} \setminus D(u)$ is called the set of *holes*. If $H(u)$ is empty, then u is a (full) word. As for full words, we use the notation $|u| = n$ for the length of partial words. Similarly to finite words, we define that infinite partial words are partial functions from $\mathbb{N}_+$ to $\mathcal{A}$.

Let $\diamond$ be a symbol that does not belong to $\mathcal{A}$. For a partial word u, we define its *companion* to be the full word $u_\diamond$ over the augmented alphabet $\mathcal{A}_\diamond = \mathcal{A} \cup \{\diamond\}$ such that $u_\diamond(i) = u(i)$, if $i \in D(u)$, and $u_\diamond(i) = \diamond$, otherwise. The symbol $\diamond$ represents the holes, and the sets $\mathcal{A}_\diamond^*$ and $\mathcal{A}_\diamond^\omega$ correspond to the sets of finite and infinite partial words, respectively.

A partial word u is said to be *contained* in v, denoted by $u \subset v$, if $|u| = |v|$, $D(u) \subseteq D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. Two partial words u and v are *compatible*, denoted by $u \uparrow v$, if there exists a partial word z such that $u \subset z$ and $v \subset z$. Using the companions this means that $u_\diamond(i) = v_\diamond(i)$ whenever neither $u_\diamond(i)$ nor $v_\diamond(i)$ is a hole $\diamond$. Factors, prefixes and suffixes of partial words are defined naturally using the one-to-one correspondence between partial words and their companions.

A morphism on $\mathcal{A}^*$ is a mapping $\varphi\colon \mathcal{A}^* \to \mathcal{A}^*$ satisfying $\varphi(xy) = \varphi(x)\varphi(y)$ for all $x, y \in \mathcal{A}^*$. Note that φ is completely defined by the values $\varphi(a)$ for every letter a on $\mathcal{A}$. A morphism is called *prolongable on a letter* a if $\varphi(a) = aw$ for some word $w \in \mathcal{A}^+$ such that $\varphi^n(w) \neq \varepsilon$ for all integers $n \geq 1$. By the definition, $\varphi^n(a)$ is a prefix of $\varphi^{n+1}(a)$ for all integers $n \geq 0$ and the sequence $(\varphi^n(a))_{n\geq 0}$ converges to the unique infinite word

$$\varphi^\omega(a) := \lim_{n\to\infty} \varphi^n(a) = aw\varphi(w)\varphi^2(w)\cdots,$$

which is a fixed point of φ.

As an example, consider the morphism $\tau\colon \{0,1\}^* \to \{0,1\}^*$, where $\tau(0) = 01$ and $\tau(1) = 10$. The word

$$t := \lim_{n\to\infty} \tau^n(0) = 01101001100101101 0\cdots$$

obtained by iterating the morphism τ is called the *Thue-Morse word*. Note that, by the construction, the Thue-Morse word has a unique decomposition into blocks $\tau(0) = 01$ and $\tau(1) = 10$. In a block ab the letter a is called *on-beat* and the letter b is called *off-beat*. For other definitions and properties of the famous word t, see [16].

3 Overlap-free infinite partial words

A kth *power* of a word $u \neq \varepsilon$ is a word $u^k = \mathrm{pref}_{k\cdot|u|}(u^\omega)$, where u^ω denotes the infinite catenation of the word u, k is a rational number and $k \cdot |u|$ is an integer. A word w is called *k-free* if there does not exist a word x such that x^k is a factor of w. If $k = 2$ or $k = 3$, then we talk about square-free or cube-free words, respectively. An *overlap* is a word of the form $xyxyx$ where $x, y \in \mathcal{A}^+$. A word is called *overlap-free* or *2^+-free* if it does not contain kth powers for any $k > 2$. Hence, it can contain squares but it cannot contain any longer repetitions such as overlaps or cubes.

It is easy to verify that there does not exist a square-free infinite word over a binary alphabet. On the other hand, Thue proved the following theorem.

Theorem 1 ([13, 14]). *The Thue-Morse word is overlap-free.*

A partial word u is *k-free* if, for any nonempty factor v of u, there does not exists a word x such that v is contained in the kth power of x, i.e., $v \subset x^k$. Similarly, a partial word u is *overlap-free* if it is k-free for every $k > 2$. Thus, an overlap-free partial word u cannot contain an *overlap* v such that $v \subset xyxyx$ for any nonempty words x and y.

In [9] Manea and Mercaş proved that there exist infinitely many cube-free binary partial words containing exactly one hole. We give a short proof of an improvement of this result.

Theorem 2. *There exist infinitely many overlap-free binary partial words containing exactly one hole.*

Proof. Consider a suffix t' of the Thue-Morse word t such that t' begins with the factor 010011 of $\tau^4(0)$. We claim that the infinite partial word $w = \diamond t'$ is overlap-free. By Theorem 1, the word t' is overlap-free. Hence, if there is an overlap in w, it must be a prefix of w. Without loss of generality, we may assume that $w = uu'a\cdots$, where $u \uparrow u'$ and $a = \mathrm{pref}_1(u')$. Since small prefixes of w do not contain overlaps, we must have $|u| = |u'| \geq 7$ and, consequently, $u = \diamond 010011\cdots$ and $u' = a010011\cdots$. Here the factor 11 is synchronizing, i.e., the first 1 is off-beat and the second 1 is on-beat. This implies that u' begins on-beat and so $|u| = |u'|$ is even. Hence, $a = 1$ and the letter in w after the prefix $uu'a$ must be an off-beat 0. In other words, we have $w = \diamond 010011x1010011x10$ for some $x \in \mathcal{A}^+$. Thus, it follows that t' contains an overlap $0y0y0$, where $y = 10011x1$. This is a contradiction. Since the Thue-Morse word is clearly recurrent, i.e., every factor occurs infinitely many times, there exist infinitely many different suffixes t' of the word t such that $\diamond t'$ is overlap-free. □

However, we cannot avoid overlaps if a binary infinite partial word contains several holes.

Theorem 3. *If an infinite binary partial word contains more than one hole, it is not overlap-free.*

Proof. Assume that an infinite binary overlap-free word w contains at least two holes. Then the gap between any two holes must be at least two. Otherwise, there is a cube of the form $\diamond\diamond a$ or $\diamond a\diamond$ in the word. Hence, there exists a position $i \geq 3$ such that $w_{i-2}w_{i-1}w_i = ab\diamond$. Since w is cube-free, the letters a and b must be different. For the same reason, $w_{i+1}w_{i+2} = ab$, but w_{i+3} is either a or b. However, the words $ab\diamond abaa$, $ab\diamond abab$, $ab\diamond abba$ and $ab\diamond abbb$ contain overlaps or cubes. At least one of them must be a factor of w, which is a contradiction. □

4 Partial words and 3-overlap-freeness

By the above considerations, it is clear that there are no overlap-free binary partial words containing infinitely many holes. On the other hand, it was proved in [9] that there exist infinitely many binary cube-free words with infinitely many holes. This result is improved in the following theorem which also has a shorter proof. For the result we need a definition of k-overlap-freeness.

Definition 1. A partial word w is *k-overlap-free* if it is cube-free and, for any factor v of w, there is no overlap $xyxyx$ such that $v \subset xyxyx$ and $|x| \geq k$.

Note that this means that a k-overlap-free partial word does not contain repetitions of the form $xyx'y'x''$ such that x, x', x'', and respectively y, y', are pairwise compatible nonempty partial words and $|x| \geq k$. For example, the word $01\diamond 0100110$ is 3-overlap-free but not 2-overlap-free since it contains the factor $01\diamond 01001$, where $x = x' = x'' = 01$ and $y = \diamond \uparrow 0 = y'$. By the definition, it is evident that any k-overlap-free word is also k'-overlap-free for $k' \geq k$. Note that a word is 1-overlap-free if and only if it is overlap-free. Now we may state the following theorem.

Theorem 4. *There exist infinitely many 3-overlap-free binary partial words containing infinitely many holes.*

Proof. Let T be an infinite partial word obtained from the Thue-Morse word t by replacing every occurrence of 01011010011 by $0101\diamond 010011$. By the recurrence of t, the word T has an infinite number of holes. Note also that

$$01010, 00100 \notin F(t), \tag{1}$$

since t is overlap-free and has a decomposition into blocks 01 and 10.

Assume now that T is not 3-overlap-free. By Theorem 1, it is evident that the word T cannot contain cubes w^3, where the length of w is one, two or three. Hence, T must have a factor $w = xyx'y'x''$ such that x, x', x'', and respectively y, y', are pairwise compatible nonempty partial words and $|x| = |x'| = |x''| \geq 3$. Since

t is overlap-free, there must be at least one hole in w such that replacing the hole by the letter 1 the word w would not be an overlap. Such a hole is here called *necessary* and denoted by $\hat{\diamond}$. In other words, there must be a position i either in x or in y such that one of the letters $x(i), x'(i), x''(i)$ or, respectively $y(i), y'(i)$, is a necessary hole $\hat{\diamond}$ and another one is a zero, denoted by $\hat{0}$, corresponding to the hole.

Assume that $|xy| \geq 8$. We divide our considerations into cases:

(i) Assume that there is a necessary hole $\hat{\diamond}$ in y. Then the zero $\hat{0}$ corresponding to that hole is in y'. Since $|xy| = |x'y'| = |y'x''| \geq 8$, there is either $0101\hat{\diamond} \in F(xy)$ and $01010\hat{} \in F(x'y')$ or $\hat{\diamond}0100 \in F(yx')$ and $\hat{0}0100 \in F(y'x'')$. Hence, either 01010 or 00100 occurs in t contradicting (1).

(ii) Assume that a necessary hole $\hat{\diamond}$ occurs in x'. By the assumption on the length of xy, this means that $0101\hat{\diamond}0100$ is a factor of $yx'y'$. If $\hat{0}$ is in x, then $\hat{0}0100 \in F(xy)$. Otherwise, $\hat{0}$ is in x'' and $01010\hat{} \in F(y'x'')$. Both cases contradict (1).

(iii) Assume that $x(i) = \hat{\diamond}$. If $0101\hat{\diamond}$ is a factor of x, then 01010 is a factor of either x' or x''. This is a contradiction. Hence, $\hat{\diamond}0100$ must occur in xy. By (1), the factor $\hat{0}0100$ cannot occur in $x'y'$, and therefore $x'(i) = \diamond$ and $x''(i) = \hat{0}$. Hence, also the hole $x'(i)$ is necessary and a contradiction follows from Case (ii).

The case where a necessary hole occurs in y' is symmetric to Case (i) and the case where $x''(i) = \hat{\diamond}$ is symmetric to Case (iii). Thus, we may conclude that T does not contain any 3-overlaps of the form $xyx'y'x''$ where $|xy| \geq 8$.

Hence, it suffices to consider overlaps of the form $xyx'y'x''$ where $|xy| \leq 7$ and the overlap contains at least one hole. By the overlap-freeness and $\{01, 10\}$-decomposition of t, it is easy to show that every $\diamond$ in T occurs in a factor corresponding to $\tau^5(1)$ of t and every such factor contains exactly one hole. Hence, every hole in T occurs in

$$v = 10010110011010010 1\diamond 0100110010110.$$

Consider repetitions of u in T such that $|u| \leq 7$ and u contains a hole. By the structure of $v \in F(T)$, we may easily verify that the greatest kth power of u occurring in T has $k = 2 + 2/|u|$. Namely, the words $(10100)(101\diamond 0)10$ and $(01\diamond)(010)01$ are factors of v. However, we have shown that T is 3-overlap-free. Since every suffix of T is also 3-overlap-free, the statement is proved. □

It remains an open question whether there exist 2-overlap-free infinite binary partial words. As a final remark, we mention that such words cannot be obtained by "punching" the Thue-Morse word. More precisely, if an infinite word w has a decomposition into blocks 01 and 10, then by replacing letters with holes we cannot obtain a 2-overlap-free word w'. Suppose that in such word w a letter 1 is replaced by a hole. This hole must occur either in a factor $01\diamond 01$ or in a factor $10\diamond 10$ of w'. Otherwise, the modified word w' contains a cube. Assume that the hole occurs in $01\diamond 01$. The other case is symmetric. By the block decomposition

of the original w we know that $\diamond$ must be in on-beat position. Hence, $01\diamond010$ is a factor of the modified word. However, the next block cannot be 10, since then the cube $(\diamond 0)(10)(10)$ occurs. On the other hand, it cannot be 01 either. In this case, the modified word has a 2-overlap $(01\diamond)(010)01$. The case where a letter 0 is replaced by a hole is symmetric.

References

[1] S. I. Adian, The Burnside problem and identities in groups, Ergebnisse der Mathematik und ihrer Grenzgebiete [Results in Mathematics and Related Areas 95], Springer-Verlag, Berlin, 1979.

[2] J. Berstel, L. Boasson, Partial words and a theorem of Fine and Wilf, Theoret. Comput. Sci. 218 (1999) 135–141.

[3] F. Blanchet-Sadri, Codes, orderings, and partial words, Theoret. Comput. Sci. 329 (2004) 177–202.

[4] F. Blanchet-Sadri, Algorithmic Combinatorics on Partial Words, Chapman & Hall/CRC Press, Boca Raton, FL, 2007.

[5] M. Euwe, Mengentheoretische Betrachtungen über das Schachspiel, Proc. Konin. Acad. Wetenschappen, Amsterdam 32 (1929) 633–642.

[6] V. Halava, T. Harju and T. Kärki, Relational codes of words, Theoret. Comput. Sci. 389 (2007) 237–249.

[7] P. Leupold, Partial words for DNA coding, Lecture Notes in Comput. Sci. 3384 (2005) 224–234.

[8] G. Lischke, Restorations of punctured languages and similarity of languages, MLQ Math. Log. Q. 52 (2006) 20–28.

[9] F. Manea, R. Mercaş, Freeness of partial words, Theoret. Comput. Sci. 389 (2007) 265–277.

[10] M. Morse, Abstract 360: a solution of the problem of infinite play in chess, Bull. Amer. Math. Soc. 44 (1938) 632.

[11] M. Morse, G.A. Hedlund, Symbolic dynamics, Amer. J. Math 60 (1938) 815–866.

[12] M. Morse, G.A. Hedlund, Unending chess, symbolic dynamics and a problem in semigroups, Duke Math. J. 11 (1944) 1–7.

[13] A. Thue, Über unendliche Zeichenreihen, Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania 7 (1906) 1–22.

[14] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania 1 (1912) 1–67.

[15] A.M. Shur, Yu.V. Gamzova, Partial words and the interaction property of periods, Izv. Math. 68 (2004) 405–428.

[16] J.-P. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, in: C. Ding. T. Helleseth, H. Niederreiter (Eds.), Sequences and Their Applications: Proceedings of SETA '98, Springer-Verlag, 1999, pp. 1–16.

Turku
Centre *for*
Computer
Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Sciences

ISBN 978-952-12-2075-3
ISSN 1239-1891