Freeness of Partial Words

Florin Manea[1,3] Robert George Mercaş[2]

13 July 2007

Abstract

The paper approaches the classical combinatorial problem of freeness of words, in the more general case of partial words. First, we propose an algorithm that tests efficiently whether a partial word is k-free or not. Then, we show that there exist arbitrarily many cube-free infinite partial words containing an infinite number of holes, over binary alphabets; thus, there exist arbitrarily many k-free infinite partial words containing an infinite number of holes for $k \geq 3$. Moreover, we present an efficient algorithm for the construction of a cube-free partial word with n holes. In the final section of the paper, we show that there exists an infinite word, over a four-symbol alphabet, in which we can substitute randomly one symbol with a hole, and still obtain a cube-free word; we show that such a word does not exist for alphabets with less symbols. Further, we prove that in this word we can replace arbitrarily many symbols with holes, such that each two consecutive holes are separated by at least two symbols, and obtain a cube-free partial word. This result seems interesting because any partial word containing two holes with less than two symbols between them is not cube-free. Finally, we modify the previously presented algorithm to construct, over a four-symbol alphabet, a cube-free partial word with exactly n holes, having minimal length, among all the possible cube-free partial words with at least n holes.

Keywords:Infinite words; Partial words; Thue-Morse word; k-freeness; Overlap-freeness; etc.

1 Introduction

The area of Combinatorics on Words took birth at the beginning of the last century, when Thue initiated a systematic study of words in a series of

[1]Faculty of Mathematic and Computer Science, University of Bucharest Str. Academiei 14, 010014, Bucharest, Romania, Email: flmanea@funinf.cs.unibuc.ro

[2]Research Group in Mathematical Linguistics, Rovira i Virgili University Pl. Imperial Tarraco 1, 43005, Tarragona, Spain, Email: robertgeorge.mercas@estudiants.urv.cat

[3]Corresponding author. Work partially supported by the Research Grant no. ET75/2005 of the Romanian National Authority for Scientific Research

papers ([22, 23]). In these papers, there were considered several combinatorial problems that arose in the study of the sequences of symbols, problems that were solved with the usual tools of discrete mathematics. One of the most important results obtained by Thue regarded the repetitions (consecutive occurrences of a factor) inside a word (see [22, 23], or "Section 1.6: Repetitions in words" from [1]).

Nowadays, the interest in the study of Combinatorics on Words is increasing, since this field finds applications in several areas such as: computer science (language theoretic properties, algorithms on strings, data compression, data communication, model checking - see, for example, [1, 2, 9, 10, 13]), biology and bio-inspired computing (DNA analysis, bio-inspired computing models, molecular biology - see, for example, [11, 12, 14, 16]), etc., starting from the premise that the data used in these areas can be easily represented as words over some particular alphabet.

Having as motivation many intriguing practical problems that appear as applications of the central topics in the field of Combinatorics on Words, such as gene comparison, Berstel and Boasson suggested the usage of partial words in this context (see [3]). Partial words, a canonical extension of the classical words, are sequences that, besides regular symbols, may have a number of unknown symbols, called "holes" or "wild cards". Molecular biology, in particular, has stimulated a considerable interest in the study of combinatorics on partial words; for example, the alignment of the DNA sequences is conceived as a construction of two compatible partial words (see [16]).

Until now, there have been investigated several combinatorial properties of the partial words such as: periodicity, conjugacy and primitivity, i.e., Fine and Wilf's Theorem, Defect Theorem, Critical Factorization Theorem (see [4, 5, 6, 7, 20, 21]). Also, in [5], the author made a first step in investigating languages of partial words by introducing the concept of pcodes, sets of partial words fulfilling a code-like property. A new approach from this point of view was given by Leupold in [15], where he obtained languages of partial words by puncturing the classical ones. As basic properties, he studies the finiteness of a language's root and analyzes the conditions in which such a language is a code. Languages of partial words obtained by puncturing languages of full words are approached in [17], as well; in this paper, the author describes, in terms of similarity of languages, the restorations of punctured languages (i.e., languages that may have produced the given language of partial words, by puncturing), provided that the number of unknown positions (holes) in a word or, in a more general case, the proportion of unknown positions per word, respectively, is bounded by a positive constant.

Since in many applications the length of the words investigated can be arbitrarily large, it is natural to study infinite words (words of infinite length). In this paper, the concept of partial word is extended to that of infinite partial word. In this framework, we study the problem of identifying and con-

structing k-free partial words, i.e., words that do not contain k consecutive factors which are pairwise compatible. Our study is aimed in two directions: we are interested in both combinatorial and algorithmic aspects regarding the k-freeness of infinite partial words.

The structure of the paper is as follows: first, we present the notions and main results that we use in our approach; further, we propose an algorithm that efficiently decides whether a given partial word is k-free or not, for $k \in \mathbb{N}$, and, in the case when the word is not k-free, it outputs a factor of the input word that violates the k-freeness property. In the last two sections we present several results that state the existence of infinite k-free partial words (with $k \geq 3$), over alphabets with 2 and 4 symbols; these results are proved by effectively constructing such words. Moreover, we present two efficient algorithms for the construction of cube-free partial words over binary or four-symbol alphabets, respectively, whose number of holes is given as input data.

2 Preliminaries

In this section we present the main definitions and results that are to be used throughout the paper. For a more detailed presentation of these aspects, as well as for the proofs of the results cited here, we refer to [3, 4, 5, 6, 7, 9, 13, 18].

In the following, we denote by $\mathbb{N}$ the set of natural numbers, and by $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$ the set of positive integers. For $i, j \in \mathbb{N}$ we denote by $\{i, \ldots, j\}$ the set $\{k \mid k \in \mathbb{N} \text{ such that } i \leq k \leq j\}$.

2.1 Finite and infinite words

Let A be a non-empty finite set, called alphabet. An element a from A is usually called symbol or letter; if A has k elements it is called k-symbol alphabet.

A finite word w over the alphabet A is a finite sequence of symbols from A; usually, a finite word is depicted as $w = a_1 \ldots a_n$. The sequence with 0 symbols, or the empty word, is denoted by λ. Observe that a finite word $w = a_1 \ldots a_n$ can be defined as a mapping $w : \{1, \ldots .n\} \to A$, with $w(i) = a_i$.

Similarly, a one-way infinite word is depicted as: $w = a_1 a_2 a_3 \ldots$, and can be formally defined as a mapping from $\mathbb{N}_+$ to A, that associates to each position of the word the symbol that is present on that position.

We denote by A^* the set of finite words over the alphabet A, by A^+ the set of finite and non-empty words over A, and by A^ω the set of one-way infinite words over the same alphabet. It is not hard to see that A^* is the free monoid generated by A, under the operation of catenation of words (the catenation of two words u and v is defined as the string uv); the unit element

in this monoid is represented by the empty word λ. We stress out the fact that we can also apply catenation to pairs consisting of a finite word and a one-way infinite word, given that the left factor is finite.

The length of a finite word w over the alphabet A, denoted by $|w|$, is defined as the number of occurrences of the symbols from A in that word. A finite word u is said to be a factor of the (infinite) word w if $w = xuy$, where x is a finite word. Moreover, u is a prefix of w if $w = uy$ and u is finite; u is a suffix of w if $w = xu$ where x is a finite word (note that u is infinite if and only if w is infinite).

A morphism is a mapping $h : A^* \to B^*$ that satisfies $h(xy) = h(x)h(y)$, for all $x, y \in A^*$; since A^* is the free monoid generated by A, h is completely defined by the values $h(a)$, for all $a \in A$, and $h(\lambda) = \lambda$. Given a morphism h we can canonically define how this morphism works for infinite words: for $w = a_1a_2a_3 \ldots \in A^\omega$, we have $h(w) = h(a_1)h(a_2)h(a_3) \ldots$.

Since A^ω is an uncountable set, hence there is no effective way to define its elements, we focus on infinite words that can be described through some precise method. The most frequently used method to define infinite words (as stated in the survey [13]) is that of iterating a morphism. More precisely, we assume that $h : A^* \to A^*$ is a morphism such that it verifies the following relation: there exists a symbol $a \in A$ verifying $h(a) = a\alpha$, with $\alpha \in A^+$. Because a is a prefix of $h(a)$ it follows that $h^i(a)$ is a prefix of $h^{i+1}(a)$. Consequently, the limit (called the infinite word defined by iterating the morphism h) $w = \lim_{i\to\infty} h^i(a)$ exists. This infinite word is a fixed point of the morphism h, i.e., $h(w) = w$. In the following we present an example of an infinite word defined using this method.

Example 1 *(The Thue-Morse word) Let $h : \{a, b\}^* \to \{a, b\}^*$ be a morphism defined by $h(a) = ab$ and $h(b) = ba$. We define $t_0 = a$ and $t_i = h^i(a)$. Remark that $t_{i+1} = h(t_i)$ and that $t_{i+1} = t_i\overline{t}_i$, where $\overline{x}$ is the word obtained from x by replacing each occurrence of a with b and each occurrence of b with a. We define the Thue-Morse word by: $t = \lim_{i\to\infty} t_i = \lim_{i\to\infty} h^i(a)$. The Thue-Morse word t is a fixed point for the morphism h, i.e., $h(t) = t$.*

We say that an (infinite) word w is k-free if there does not exist a word x such that x^k is a factor of w. Also, a (infinite) word is called overlap-free if it does not contain any factor of the form $xyxyx$ with $x \neq \lambda$. It is clear that any overlap-free word w is $k-$free, for $k \geq 3$, and, as well, any 2-free word is overlap-free. For simplicity, a 2-free word is said to be square-free, and a 3-free word is said to be cube-free.

A result that will be used throughout the paper, regarding the Thue Morse infinite word t, defined in Example 1, is the following:

Theorem 1 *(Thue Theorem)[22, 23] The Thue-Morse word t is overlap-free.*

Remark 1 *As a consequence of Theorem 1, it follows that the Thue-Morse word t is k-free for all $k \geq 3$.*

2.2 Partial words

A partial word of length n over the alphabet A is a partial function $u : \{1, \ldots, n\} \overset{\circ}{\to} A$. For $i \in \{1, \ldots, n\}$, if $u(i)$ is defined we say that i belongs to the domain of u (denoted by $i \in D(u)$), otherwise we say that i belongs to the set of holes of u (denoted by $i \in H(u)$). A partial word whose set of holes is empty is called full word.

Let $\diamond$ be a symbol that does not belong to A. If u is a partial word of length n over A, then the companion of u is the total function (or the full word) $u_\diamondsuit : \{1, \ldots, n\} \to A \cup \{\diamondsuit\}$ defined by:

$$u_\diamondsuit = \begin{cases} u(i), \text{ if } i \in D(u) \\ \diamondsuit, \text{ otherwise} \end{cases}$$

For convenience, finite partial words are seen as full words over the extended alphabet $A \cup \{\diamond\}$ (see [4, 5, 6, 7]); this permits us to speak, for example, about the partial word $a\diamond bba$ instead of the partial word with the companion $a\diamond bba$. Usually, a partial word u of length n is depicted as $u = a_1 \ldots a_n$, where $a_i = u_\diamond(i)$. In this way, one can easily define the catenation of partial words, as the catenation of the corresponding full words over $A \cup \{\diamond\}$, and the length of partial words, as the length of the corresponding full words over $A \cup \{\diamond\}$.

The partial words u and v are said to be equal if u and v have the same length, $D(u) = D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. If u and v are two partial words of equal length, then u is said to be contained in v, $u \subset v$, if all the elements of $D(u)$ are contained in $D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. We say that u is properly contained in v, $u \sqsubset v$, if $u \subset v$ and $u \neq v$. Note that for a full word u and a partial word v, with $|u| = |v|$, if $u \subset v$ then $H(v) = \emptyset$ and $u = v$.

Similarly to the case of full words, we say that the partial word u is a factor of the partial word w if there exist partial words x and y such that $w = xuy$. If $x = \lambda$ we say that u is a prefix of w, and if $y = \lambda$ we say that u is a suffix of w. If $w = a_1 \ldots a_n$, we denote by $w[i..j]$ the factor $a_i \ldots a_j$ of w, and by $w[i]$ the symbol a_i; we say that $w[i]$ is the symbol placed on the i^{th} position in the partial word w.

We say that two partial words u and v are compatible, denoted by $u \uparrow v$, if there exists a partial word w such that $u \subset w$ and $v \subset w$.

The notion of one-way infinite partial word extends the notion of partial word in a natural way. A one-way infinite partial word over the alphabet A is a partial function $u : \mathbb{N}_+ \overset{\circ}{\to} A$. As in the case of finite partial words, for $i \in \mathbb{N}_+$, such that $u(i)$ is defined, we say that i belongs to the domain of u ($i \in D(u)$); otherwise we say that i belongs to the set of holes of u

($i \in H(u)$). The infinite partial words that do not contain any hole are called infinite full words. The companion function of the infinite partial word u is the total function (the full infinite word) $u_\diamond : \mathbb{N}_+ \to A \cup \{\diamond\}$, defined by the same relation as in the case of finite partial words; we say that the symbol $u_\diamond(i)$ is the symbol placed on the i^{th} position in the infinite partial word u. Intuitively, the companion function associates with each position of the word the symbol (from $A \cup \{\diamond\}$) appearing on that position. One-way infinite partial words are seen as elements of $(A \cup \{\diamond\})^\omega$: an infinite partial word is usually depicted as $u = a_1 a_2 a_3 \ldots$, with $a_i \in A \cup \{\diamond\}$.

The infinite partial words we describe in this paper are obtained from infinite full words by applying the finite transduction defined by a deterministic generalized sequential machine of particular type. In the following we describe formally this strategy.

Recall that a deterministic generalized sequential machine (dgsm) is a 6-tuple $M = (Q, V, U, q_0, F, f)$ where Q is a set of states, $q_0 \in Q$ is the initial state, $F \subset Q$ is the set of final states, V and U are finite sets of symbols, namely, the set of input symbols and, respectively, the set of output symbols, and the transition-output function $f : Q \times V \to Q \times U^*$; this function is extended canonically to $Q \times V^*$. The finite transduction defined by M is the function $T_M : V^* \to U^*$, defined by: $T_M(v) = u$ if and only if $f(q_0, v) = (q, u)$ and $q \in F$.

In this paper we will use a particular type of dgsms in which we consider that each state is final and that $\diamond$ is contained in the set of output symbols. If M is such a dgsm, it is not hard to see that if u is a prefix of v, then $T_M(u)$ is also a prefix of $T_M(v)$. Let w be a infinite full word, and denote by w_n the prefix of length n of this word. One can obtain an infinite partial word w' from w by taking: $w' = \lim_{n\to\infty} T_M(w_n)$.

A partial word $w \in (A \cup \{\diamond\})^*$ is said to be k-free if for any non-empty factor $x_1 \ldots x_k$ of w, there does not exist a partial word u, such that $x_i \subset u$ for all $i \in \{1, \ldots, k\}$.

Remark 2 *It is rather simple to note that any partial word w over A, with $|w| \geq 2$ and $H(w) \neq \emptyset$, cannot be square-free, since it contains at least one of the factors $a\diamond$ or $\diamond a$, where $a \in A \cup \{\diamond\}$. Also, if w is n-free, then w is m-free for $m \geq n$.*

3 An efficient algorithm for deciding if a partial word is k-free

In this section we propose an algorithm that, given a finite partial word w and a natural number k, decides whether w is k-free or not. Moreover, if w is not k-free, the algorithm computes a non-empty factor $x_1 \ldots x_k$ of the input word w and a partial word u such that $x_i \subset u$, for all $i \in \{1, \ldots, k\}$.

We analyze the soundness and the time complexity of this algorithm (on the random access machine model). In the following we assume that the input partial word w is over the alphabet A, and $*$ is a symbol not contained in A; also we denote by n the length of w.

First, we define the two-dimensional array $\uparrow [\,][\,]$, with n rows, $\lfloor n/k \rfloor$ columns and with elements from $A \cup \{\diamond\}$, as follows:

$$(\mathbf{1}).\ \uparrow[i][l] = \begin{cases} a, \text{ if there exists a symbol } a \text{ such that } w[i+hl] \subset a \text{ for all} \\ \quad h \in \{0, \ldots, k-1\}, \text{ and, for any other symbol } b, \text{ such that} \\ \quad w[i+hl] \subset b \text{ for all } h \in \{0, \ldots, k-1\}, \text{ we have } a \subset b \\ *, \text{ otherwise} \end{cases}$$

The usage of the symbol $\uparrow$ to denote this array is motivated by the fact that $\uparrow[i][l] \neq *$ if and only if every two symbols a and b in the set $\{w[i], \ldots, w[i+(k-1)l]\}$ are compatible (therefore, $a \uparrow b$), and both a and b are contained in $\uparrow[i][l]$.

The values stored in this array can be computed using the following relation:

$$(\mathbf{2}).\ \uparrow[i+l][l] = \begin{cases} \uparrow[i][l], \text{ if } w[i+lk] \subset \uparrow[i][l] \text{ and there exists } h \in \{1,\ldots,k-1\} \\ \quad \text{such that } w[i+lh] \neq \diamond; \\ w[i+lk], \text{ if } w[i+lh] = \diamond, \text{ for all } h \in \{1,\ldots,k-1\}; \\ *, \text{ otherwise.} \end{cases}$$

An algorithm that effectively computes this array consists basically in the following two steps:

- for each possible value of l and for each i, such that $i \leq l$, we compute $\uparrow[i][l]$, using the definition **(1)** of the array $\uparrow[\,][\,]$.
- we use the relation **(2)**, defined above, to compute recursively the elements $\uparrow[i+l][l], \ldots, \uparrow[i+l\,\lfloor (n-i-(k-1)l)/l \rfloor][l]$.

By a careful implementation of the above strategy, the time needed to compute all the elements of the array $\uparrow[\,][\,]$, on an input consisting of the partial word w, with $|w| = n$, and the natural number k, is $\mathcal{O}(n^2/k)$.

Further we show how this array can be used to decide whether a given partial word is k-free or not.

Algorithm 1

```
function Free(w, k);
begin
let n := |w|;
compute the array ↑[ ][ ] as explained above;
for l = 1 to ⌊n/k⌋ do
   let counter := 0, s := 0;
   for i = 1 to n − l ∗ k + 1 do
```

```
    if ↑ [i][l] ≠ ∗ then
      let counter := counter + 1;
      if s = 0 then
        let s := i;
      endif
    else let counter := 0, s := 0;
    endif ;
    if counter = l then
      output: w is not k-free. We have x_j = w[s+jl..s+(j+1)l−1], j ∈ {1,...,k},
              and u =↑ [s][l] ... ↑ [s + l − 1][l]
      return Free(w, k) := False (the algorithm stops)
    else counter := 0;
    endif ;
  endfor
endfor
return Free(w, k) := True; (the algorithm stops)
end.
```

In order to prove the soundness of Algorithm 1, we remark the following immediate facts:

- For two partial words x and u, both of length l, we have $x \subset u$ if and only if $x[i] \subset u[i]$, for all $i \in \{1, \ldots, l\}$.
- Consequently, for a non-empty factor $x = x_1 \ldots x_k$ of the partial word w there exists a partial word u, of length $l > 0$, such that $x_i \subset u$ if and only if the string $u' =\uparrow [r][l] \uparrow [r+1][l] \ldots \uparrow [r+l-1][l]$ does not contain the symbol $*$, where r is the first position of the factor x in w.
- If the input word w is not k-free, and, by definition, there exists a factor $x_1 \ldots x_k$ of w and a non-empty partial word u, such that $x_i \subset u$, for all $i \in \{1, \ldots, k\}$, then the length of a factor x_i is bounded by $\lfloor n/k \rfloor$.

The algorithm we propose identifies, if any, a non-empty factor $x_1 \ldots x_k$ of the input word w and a partial word u such that $x_i \subset u$, for all $i \in \{1, \ldots, k\}$. According to the facts presented above, we remark that w is not k-free if and only if the sequence $\uparrow [1][l], \ldots, \uparrow [n - lk + 1][l]$ contains l consecutive positions that differ from $*$; moreover, if this sequence contains l such consecutive positions, starting from the position s, it follows that a possibility to choose the k factors $x_1, \ldots, x_k$ of w and the partial word u, proving that the input word is not k-free, is to set $u =\uparrow [s][l] \ldots \uparrow [s+l-1][l]$ and $x_i = w[s + (i-1)l..s + il - 1]$, for $i \in \{1, \ldots, k\}$. Once such a possibility is discovered, the algorithm stops and concludes that w is not k-free; if no such possibility is identified for any $l \leq \lfloor n/k \rfloor$, the algorithm stops, and decides that w is k-free.

The overall time complexity of Algorithm 1 is clearly $\mathcal{O}(n^2/k)$, where $n = |w|$, since the most time consuming operation is the computation of the array $\uparrow [\,][\,]$. The space needed by this algorithm is also $\mathcal{O}(n^2/k)$.

Finally, we remark that Algorithm 1 can be applied, as well, for full words. However, in this case, an algorithm working in time $\mathcal{O}(n \log n)$ can be developed using suffix arrays ([10, 19]).

4 A generalization of the Thue theorem

The main result we present in this section is that for $k \geq 3$ there exist k-free infinite partial words, containing an arbitrary number of holes, over binary alphabets. Moreover, we present an algorithm that, given a natural number n as input, constructs in $\mathcal{O}(n)$ time a cube-free partial word that contains exactly n holes.

The first result that we propose is the following:

Proposition 1 *There exist arbitrarily many cube-free infinite partial words, containing exactly one hole, over a binary alphabet.*

Proof. The proof of this property is based on the following approach: we find the symbols of the Thue-Morse word t (described in Example 1) that can be replaced by a hole, such that the infinite partial word that we obtain remains cube-free.

Assume that we replace an arbitrary position in t with a hole; let t' be the infinite partial word that we obtain in this manner. We will prove that for a non-empty factor $x_1x_2x_3$ of t', and a partial word u such that $x_i \subset u$, for all $i \in \{1, 2, 3\}$, we have $|x_i| < 4$ and $|x_i| \neq 2$, for all $i \in \{1, 2, 3\}$.

Indeed, if none the factors x_1, x_2, x_3 contains the hole inserted in t, the result is an immediate consequence of Remark 1. Hence, we may assume that the hole is contained in one of the words x_1, x_2 or x_3.

Assume that there exist a non-empty factor $x_1x_2x_3$ of t' and a partial word u, such that:
– one of the factors x_1, x_2 and x_3 contains a hole,
– $x_i \subset u$, for all $i \in \{1, 2, 3\}$,
– $|x_i| \geq 4$ or $|x_i| = 2$, for all $i \in \{1, 2, 3\}$.
Without loss of generality, we may assume that the hole was placed in x_1 (the other cases can be approached similarly). Also, let y_1 be the factor of t in which a hole was inserted in order to obtain x_1; note that x_2, x_3 and $y_1x_2x_3$ are factors of t, and we have $x_2 = x_3 = u$ and $y_1 \neq u$.

There are several cases to be analyzed:

1: $|x_1| = 2k, k \geq 1$, and the first symbol of x_1 is placed on an odd position in t'. Since $t = h(t)$ (as shown in Example 1), t is cube free and in t' was inserted exactly one hole, it follows that $x_2 = x_3 = u = h(a_1 \dots a_k)$, for some $a_j \in \{a, b\}$, for all $j \in \{1, \dots, k\}$ and $y_1 =$

$h(a_1 \dots a_{l-1} a' a_{l+1} \dots a_k)$, where $a' \neq a_l$, for an index l, with $l \leq k$; moreover, one of the two symbols of $h(a')$ was replaced with a hole to obtain x_1. If $a' = b$ it follows that $a_l = a$; since $h(a') = ba$ and $h(a_l) = ab$ it follows that any partial word that can be obtained from $h(a')$ by replacing one of its symbols with a hole cannot be contained in $h(a_l)$. The same argument holds in the case when $a' = a$ and $a_l = b$. Thus, we have obtained a contradiction.

2: $|x_1| = 2k, k \geq 1$, and the first symbol of x_1 is placed on an even position in t'. It follows that $x_2 = x_3 = u = b_1 h(a_1 \dots a_k) b_2$, for some $b_1, b_2, a_j \in \{a, b\}$, for all $j \in \{1, \dots, k\}$; remark that $b_1 \neq b_2$, since $b_2 b_1 = h(o)$, for some $o \in \{a, b\}$. The word y_1 may have one of the following forms:
– $y_1 = b_1' h(a_1 \dots a_k) b_2$ with $b_1' \neq b_1$, or
– $y_1 = b_1 h(a_1 \dots a_k) b_2'$ with $b_2' \neq b_2$, or
– $y_1 = b_1 h(a_1 \dots a_{l-1} a' a_{l+1} \dots a_k) b_2$, for some index l and $a_l \neq a'$.
The last possibility leads to a contradiction similarly to the case 1.
If $y_1 = b_1' h(a_1 \dots a_k) b_2$ with $b_1' \neq b_1$, since $b_1 \neq b_2$, it follows that t contains the factor: $h(a_1 \dots a_k) b_2 x_2 x_3 = h(a_1 \dots a_k) b_2\ b_1 h(a_1 \dots a_k) b_2\ b_1 h(a_1 \dots a_k) b_2$, a contradiction to the fact that t is overlap-free.
If $y_1 = b_1 h(a_1 \dots a_k) b_2'$, with $b_2' \neq b_2$, it follows $b_2' = b_1$, and hence $b_1 b_1 = h(o)$, for some $o \in \{a, b\}$, again a contradiction.

3: $|x_1| = 2k + 1, k \geq 2$, and the first symbol of x_1 is placed on an odd position in t'. It follows that $x_2 = q h(a_1 \dots a_k)$, and $x_3 = h(b_1 \dots b_k) o$ for $q, o, a_j, b_j \in \{a, b\}$, for all $j \in \{1, \dots, k\}$. We can easily observe that $a_j \neq b_j$, for $j \in \{1, \dots, k\}$, and $a_{j-1} \neq b_j$, for $j \in \{2, \dots, k\}$. Consequently, $a_j = a_{j+1}$ and $b_j = b_{j+1}$, for all $j \in \{1, \dots, k-1\}$. Since t is cube-free, it follows that $k = 2$. We may assume, without loss of generality, that $q = a$; hence: $x_2 = x_3 = ababa$. But this is a contradiction to the fact that t is overlap-free.

4: $|x_1| = 2k + 1, k \geq 2$, and the first symbol of x_1 is placed on an even position in t'. It follows that $x_2 = h(a_1 \dots a_k) q$, and $x_3 = o h(b_1 \dots b_k)$ for $q, o, a_j, b_j \in \{a, b\}$, for all $j \in \{1, \dots, k\}$. As in the former case, we observe that $a_j \neq b_j$, for all $j \in \{1, \dots, k\}$ and $a_{j+1} \neq b_j$, for all $j \in \{1, \dots, k-1\}$. Consequently, $a_j = a_{j+1}$ and $b_j = b_{j+1}$, for all $j \in \{1, \dots, k-1\}$. In particular, we obtain that $a_1 \neq b_k$; thus the first symbol of $h(a_1)$ and last symbol of $h(b_k)$ coincide. Since q equals the last symbol of $h(b_k)$ and o equals the first symbol of $h(a_1)$, it follows that $q = o$. This is not possible, since $qo = h(e)$, for some symbol $e \in \{a, b\}$.

All the cases lead to a contradiction. Consequently, we have proved that for a non-empty factor $x_1 x_2 x_3$ of t', and a partial word u, such that $x_i \subset u$, for all $i \in \{1, 2, 3\}$, we have $|x_1| < 4$ and $|x_1| \neq 2$.

Therefore, if we want to replace a symbol of the infinite word t with a hole, and obtain an infinite cube-free partial word t', we should only verify that this replacement does not cause the apparition in t' of a non-empty factor $x_1x_2x_3$, with $|x_1| = |x_2| = |x_3|$, $|x_1| \in \{1, 3\}$, for which there exists a partial word u such that $x_i \subset u$, for all $i \in \{1, 2, 3\}$.

We observe that there exist positions in t where a substitution, respecting the restrictions described above, can be performed. For example, in the word $t_5 = abbabaabbaabab\underline{b}babaababbaabbabaab$, which is a prefix of t (see Example 1), the underlined symbol can be replaced with a hole, and the partial word we obtain remains cube-free.

Also, we observe that t_5 has an infinite number of occurrences as a factor of t. For each such occurrence, we can construct a cube-free infinite partial word with exactly one hole, by replacing the 14^{th} symbol in t_5 with $\diamond$ (and obtain an infinite word of the form $xabbabaabbaaba\diamond babaababbaabbabaaby$, with $x \in \{a, b\}^*$ and $y \in \{a, b\}^\omega$, contained in t).

In conclusion, we have proved that there exist infinitely many cube-free infinite partial words, containing exactly one hole.

□

Remark 3 *We can extend the definition of the overlap-freeness to partial words, similarly to the case of k-freeness. A partial word w is overlap-free if for any factor $x_1y_1x_2y_2x_3$ of w one cannot find two partial words x and y, with $|x| > 0$, such that $x_i \subset x$, for $i \in \{1, 2, 3\}$, and $y_j \subset y$, for $j \in \{1, 2\}$. Observe that if we substitute any symbol of the word $t_3 = abbabaab$ with a hole, we obtain a partial word that is not overlap-free; obviously, the same holds for $\bar{t}_3$. Since t can be written as the catenation of an infinite number of words t_3 and $\bar{t}_3$, it follows that any infinite partial word that can be obtained from t by substituting several of its symbols with holes is not overlap-free. Consequently, any infinite partial word obtained from t using the procedure described in the proof of Proposition 1 is not overlap-free. We state as an open problem the task of constructing overlap-free infinite partial words over a binary alphabet.*

Since any cube-free infinite partial word is k-free, for $k \geq 3$ (as noted in Remark 2), we obtain, as a corollary of Proposition 1, the following result:

Corollary 1 *For $k \geq 3$, there exist infinitely many k-free infinite partial words, containing exactly one hole, over a binary alphabet.*

We also obtain, as another consequence, an already known result (see [8, 18]):

Corollary 2 *For $k \geq 3$, there exist infinitely many k-free infinite full words, over a binary alphabet.*

Proof. Let t' be one of the infinite k-free partial word constructed in the proof of Proposition 1. We replace the hole in t' with an a symbol; it is clear that the word obtained in this manner is a k-free infinite full word, for $k \geq 3$. This procedure can be applied to each of the infinite partial words constructed in the proof of Proposition 1 and obtain an infinite full word; each two of these newly obtained infinite full words are different. This proves the corollary.

□

Next, we extend the result stated in Proposition 1 in order to obtain infinite cube-free partial words, with arbitrarily many holes.

First, remark that:

Remark 4 *The word $t_k, k \geq 1$, has an infinite number of non-overlapping occurrences in t, with its first symbol placed on an odd position. To begin with, t_k has one occurrence in t, with the first symbol placed on the position 1. Also, since $t_{i+1} = t_i\bar{t}_i$, thus $t_{i+2} = t_i\bar{t}_i\bar{t}_it_i$, and $|t_i| = 2^i$, for all $i \geq 1$, it can be easily proved by induction that t_k occurs at least 2^l times in t_{k+l+1}, all these occurrences having their first symbol placed on an odd position.*

Now we can prove:

Proposition 2 *There exists a cube-free infinite partial word, containing an infinite number of holes, over a binary alphabet.*

Proof. From Remark 4 it follows that in the Thue-Morse word t there exist an infinite number of non-overlapping occurrences of the word t_5, each having its first symbol placed on an odd position. Further, for each of these occurrences of t_5, we replace its 14^{th} symbol (the underlined symbol in the factor *abbabaabbaabab*$\underline{b}$*babaabbabbaabbabaab*) with a hole, in this manner resulting an infinite partial word, with an infinite number of holes, t'. It is clear that t' can be obtained from t using the finite transduction defined by a dgsm. We claim that the partial word t' is cube-free.

Note that if there exist a non-empty factor $x_1x_2x_3$ of t' and a partial word u such that $x_i \subset u$, for $i \in \{1, 2, 3\}$, only a finite number of holes are contained in this factor; let n be this number. Consequently, to prove our claim it is sufficient to show that any word obtained by replacing n symbols of t with holes, on some of the aforementioned positions, is cube-free, for all $n \in \mathbb{N}$.

We prove this result by induction on n: for $n = 1$ it was already shown to be true in the proof of Proposition 1. We assume the statement holds for all $k < n$, and prove it for n.

Let $t^{\{n\}}$ be a word obtained by replacing n symbols of t with holes, on n of the positions already defined. Assume, for the purpose of contradiction, that $t^{\{n\}}$ contains a non-empty factor $x_1x_2x_3$ and there exists a partial word u such that $x_i \subset u$, for $i \in \{1, 2, 3\}$. All the holes are contained in the factor

$x_1x_2x_3$; otherwise, we obtain, using the procedure described above, a non-cube-free infinite partial word with less than n holes, a contradiction to the induction hypothesis.

Note that in the infinite partial word $t^{\{n\}}$ there are at least 31 symbols between two distinct holes. Moreover, since $n \geq 2$, it follows that the factor $x_1x_2x_3$, whose length is divisible by 3, has at least 33 symbols, and, consequently, $|x_i| \geq 11$, for all $i \in \{1, 2, 3\}$. Also, remark that any the hole appearing in $t^{\{n\}}$ replaces a b symbol, and, consequently, the position that corresponds in u to that hole is occupied by an a symbol (otherwise, the hole is not necessary, and, again, we obtain a contradiction to the induction hypothesis). Finally, remark that all the holes are placed on an even position in t.

There are several cases to be analyzed:

(1) x_1 contains at least one hole;

(2) x_3 contains at least one hole, and x_1 does not contain any hole;

(3) x_2 contains at least one hole, and both x_1 and x_3 do not.

In the first case, we have $x_1 = w_{11}o_1w_{12}$, $x_2 = w_{21}o_2w_{22}$, and $x_3 = w_{31}o_3w_{32}$, $o_1 = \diamond$ and $u = u_1au_2$, where $w_{ij} \subset u_j$, for $i \in \{1, 2, 3\}$ and $j \in \{1, 2\}$. Again, there are two cases to be discussed:
– $o_2 = a$, and,
– $o_2 = \diamond$.
Note that o_2 and o_3 cannot be simultaneously equal to $\diamond$, because, otherwise, none of the holes o_1, o_2 and o_3 is necessary, contradiction to the induction hypothesis.

We will only describe how the first case leads to a contradiction, since the other one can be treated similarly. We remarked that $|x_1| \geq 11$; therefore, we have $|w_{11}| + |w_{12}| \geq 10$, so at least one of the words $|w_{11}|$ and $|w_{12}|$ is of length greater or equal to 5. If $|w_{11}| \geq 5$, it follows that $baaba\diamond$ is a factor of x_1, and, consequently, $baabaa$ is a factor of u. Thus $baabaa$ or a partial word contained in $baabaa$, with exactly one $\diamond$ replacing one of the a symbols, is a factor of x_2; but this leads to a contradiction. Indeed, in the case when no hole appears in this factor, since $t^{\{n\}}$ was obtained by substituting some of the symbols of $t = h(t)$ with holes, it follows that at least one of the two groups aa should be the image of a symbol through the morphism h, which is impossible. In the other case, when a hole replaces an a symbol, considering the way we introduce holes in t, it follows that $\diamond$ can replace only the last a in the sequence, which coincides with the symbol denoted by o_2. This is a contradiction to the assumption that $o_2 \neq \diamond$. If $|w_{11}| < 5$ and $|w_{12}| \geq 6$, it follows that $\diamond babaab$ is a factor of x_1. If $5 > |w_{11}| \geq 1$, it follows that $aababaab$ is a factor of u. Consequently $aababaab$ (or a partial word contained in $aababaab$, with exactly one $\diamond$ replacing an a symbol) is

a factor of x_2, again a contradiction, from the same reason as above. If $|w_{11}| = 0$, it follows that *ababaababba* is a factor of u. Hence *ababaababba* (or a partial word contained in *ababaababba*, with exactly one $\diamond$ replacing one of the a symbols, other than the first) is a prefix of x_2. Remark that no partial word contained in *ababaababba*, with a hole instead of an a other than the first one, can be obtained by the procedure that we use, since any hole should be followed by the factor *babaab* or preceded by an a symbol. Therefore, *ababaababba* is a prefix of x_2. Also, note that the first symbol of x_2 is on an even position in $t^{\{n\}}$ (otherwise, the group aa would have been the image of a symbol through the morphism h, a contradiction). In this case, since x_1 starts with $\diamond$, and a hole can be placed only on even positions, we obtain that the length of the string w_{12} is odd. Since the first symbol of x_2 is a, it follows that the last symbol of w_{12} is b, as well as the last symbol of u. This implies that the last symbols of x_2 and x_3 are b symbols. Since the last symbol of x_3 is placed on an odd position, it follows that the symbol placed exactly after x_3 in $t^{\{n\}}$ is an a. Consequently, $y_1 = w_{12}o_2$, $y_2 = w_{22}o_3$, $y_3 = w_{32}a$ and $y_1y_2y_3$ are factors of $t^{\{n\}}$, $u' = u_2a$ is a partial word, such that $y_i \subset u'$, for $i \in \{1, 2, 3\}$, and $y_1y_2y_3$ contains $n - 1$ holes, which is a contradiction to the induction hypothesis.

Further, we assume that x_1 does not contain any hole, and analyze the other cases. If $x_2 = w_{21}\diamond w_{22}$, with $w_{21} \neq \lambda$, or $x_3 = w_{31}\diamond w_{32}$, with $w_{31} \neq \lambda$, we can apply similar arguments to the above reasoning, and reach the same conclusion. If none of these cases occur, it follows that holes may replace only the symbols placed on the first positions of x_2 and x_3; therefore, $n \leq 2$. But, due to the induction hypothesis, we have $n \geq 2$, and, thus, we obtain $n = 2$. Hence, we have $x_2 = \diamond w$ and $x_3 = \diamond w$, and no other $\diamond$ exists in $t^{\{n\}}$; moreover, $x_1 = u = aw$. It follows that $aw\diamond w\diamond w$ is a factor of $t^{\{2\}}$, where w is a non-empty word that does not contain any hole. Thus, $wbwbw$ is a factor of t, a contradiction to the fact that t is overlap-free.

Since all the cases lead to a contradiction, we conclude that the assumption we made is false. This concludes our proof.

□

Considering that there are infinitely many non-overlapping occurrences of t_5 in t, having their first symbols placed on odd positions, it follows that we can obtain an infinite number of cube-free infinite partial words with an infinite number of holes. This can be done by choosing, randomly, an infinite number of such occurrences of t_5 and substitute, in each of them, the 14^{th} symbol with a hole, as we have described in the proof of Proposition 2; it is clear that all the infinite partial words obtained in this manner are cube-free.

The following corollary is immediate:

Corollary 3 *For $k \geq 3$, there exist arbitrarily many k-free infinite partial words, containing an infinite number of holes, over a binary alphabet.*

The proof of Proposition 2 provides an efficient solution to the following algorithmic problem: given the natural number n find a k-free partial word (for some $k \geq 3$) containing exactly n holes. In the following, we propose an algorithm that constructs a cube-free partial word with exactly n holes, offering, thus, a solution for this problem.

As stated in Remark 4, the word t_{n+6}, with $n \geq 1$, contains at least 2^n non-overlapping occurrences of t_5 having the first symbol on an odd position. Also, note that both the computational time and space needed to construct t_n are $\mathcal{O}(2^n)$. Thus, $t_{\lceil \log_2 n \rceil + 6}$ has $\mathcal{O}(n)$ symbols and can be constructed in $\mathcal{O}(n)$ time; also, it contains n non-overlapping occurrences of t_5, each having its first symbol on an odd position. According to the proof of Proposition 2, the following algorithm constructs a cube-free partial word:

Algorithm 2
function $Construct - cube - free - word(n)$;
begin
construct $t_{\lceil \log_2 n \rceil + 6}$;
identify n *non-overlapping occurrences of* t_5 *in* $t_{\lceil \log_2 n \rceil + 6}$, *having their first symbols on odd positions*
for *each of these occurrence* do
  substitute *its* 14^{th} *symbol with* $\diamond$; endfor;
denote *by* $t'_{\lceil \log_2 n \rceil + 6}$ *the word obtained after the* n *substitutions were performed;*
return $Construct - cube - free - word(n) := t'_{\lceil \log_2 n \rceil + 6}$; *(the algorithm stops)*
end.

The running time of the above algorithm is clearly $\mathcal{O}(n)$. Indeed, we have already stated that the step where $t_{\lceil \log_2 n \rceil + 6}$ is constructed can be performed in linear time; also, the identification of the occurrences of t_5 as well as the step where the 14^{th} symbol of each of these strings is substituted with a hole can be completed in $\mathcal{O}(n)$ steps, since these operations can be easily implemented using a dgsm.

We remark that it is impossible to solve this problem with an algorithm that requires less than n steps, since the string we construct must have at least n symbols, the holes.

5 Infinite k-free partial words over four-symbol alphabets

Although the result in Proposition 2 exhibits a method for the construction of infinite k-free partial words over an alphabet with at least two symbols, we study in this section the existence of infinite k-free partial words over a four-symbol alphabet. The motivation for this study comes from the following two points.

First, a theoretical motivation. The constructions we present in this section have properties that seem interesting to us: we define an infinite word in which we can replace any symbol with a hole and it still remains cube-free; we show that it is impossible to define such a word in the case of alphabets with less than 4 symbols. The same infinite word verifies the property that we can replace randomly an arbitrarily large (infinite) number of its symbols with holes, such that each two consecutive holes are separated by at least two symbols, and still obtain a cube-free infinite partial word. This property can be regarded as an optimal result, since any partial word, that contains two holes with less than two symbols between them, is not cube-free; note, though, that such partial words may still be k-free, for some $k > 3$. Nevertheless, this condition provides the elements needed to construct a cube-free partial word that contains exactly n holes and has the minimum length among all the possible cube-free partial words with n holes, regardless of the alphabet over which these words are constructed.

Second, there have been studied applications of both partial and infinite words in the processing and analysis of DNA strings ([11, 16]), which are encoded over the four-symbol alphabet $\{a, c, g, t\}$. Therefore, it seems interesting to us to analyze the existence and construction of k-free partial words that contain effectively 4 symbols.

To begin with, we define the morphism $\phi : \{a, b\}^* \to \{a, b, c, d\}^*$, that works as follows: $\phi(a) = abcd$ and $\phi(b) = badc$. Let $w = \phi(t)$ be the infinite word obtained by applying ϕ to the Thue-Morse word t. We observe that if we delete the c and d symbols from w, we obtain t; also, if we delete the a and b symbols, we obtain the Thue-Morse word in which a is replaced by c and b by d, respectively. To keep the exposure simple, assume that the distance between two symbols of w, placed on the positions n_1 and n_2 of w, respectively, is defined as $|n_1 - n_2|$. Note that the distance between two identical symbols of w can be $4s, 4s + 1$ or $4s + 3$, for some $s \in \mathbb{N}_+$.

It is not hard to see that w is cube-free. To prove this, assume, for the purpose of contradiction, that w contains a factor xxx, with $x \in \{a, b, c, d\}^+$. Also, assume that x has an a as its first symbol. First, it follows that $|x| = 4k$, $|x| = 4k + 1$ or $|x| = 4k + 3$, with $k \in \mathbb{N}$, since $|x|$ equals the distance between the first symbol of the first x factor and the first symbol of the second x factor, which are identical. If $|x| = 4k$ it follows that if we delete the c and d symbols from xxx we obtain a non-empty factor sss, $s \in \{a, b\}^+$, contained in the Thue-Morse word t; but, this would mean that t is not cube-free, a contradiction to Theorem 1. If $|x|$ is odd it follows that the distance between the first symbol of the first factor x and the first symbol of the third factor x is $4p + 2$, for some $p \in \mathbb{N}$, a contradiction. If x has as first symbol a b, a c or a d, similar arguments lead to a contradiction.

Also, we observe that any word that can be obtained from w by substituting one of its symbol with a hole is still cube-free. Let w' be an infinite word obtained by replacing a symbol of w with a hole; also, assume that w'

contains a non-empty factor $x_1x_2x_3$ and there exists a partial word u such that $x_i \subset u$ for all $i \in \{1,2,3\}$. First, remark that $|x_1| > 1$. If $|x_1| = 4k$, with $k \in \mathbb{N}_+$, and $\diamond$ replaces a c or a d symbol, then we proceed as above and delete the c and the d symbols, as well as the $\diamond$, and obtain that t is not cube-free, a contradiction; the same strategy is applied for the case when $\diamond$ replaces an a or a b symbol, but now the deleted symbols are a, b and $\diamond$. If $|x_1| = 4k+1$ or $|x_1| = 4k+3$, with $k \in \mathbb{N}$, it follows, from the proof of the fact that w is cube free, that one of the first symbols of x_1, x_2 or x_3 is a hole (otherwise the distance between two identical symbols of w is $4p+2$, for some $p \in \mathbb{N}$). But, this would mean that the symbols on the second positions of each of these factors coincide. Hence, the distance between the second symbol of x_1 and the second symbol of x_3, which are identical, is $4p+2$ with $p \in \mathbb{N}$, a contradiction. Finally, with the same arguments, $|x_1| \neq 4k+2$. Consequently, the assumption that we made is false, and by replacing any symbol of w by a hole in w we still obtain a cube-free word.

Remark that in the case of 3-symbol alphabets it is impossible to construct an infinite word z in which we can substitute randomly one of its symbols with a hole and obtain a cube-free word in all the cases. Indeed, if such a word exists it follows that the number of symbols between two identical symbols of z is at least 2; but the words that verify this condition are of the form $l_1l_2l_3l_1l_2l_3l_1l_2l_3\ldots$, where l_1, l_2 and l_3 are different symbols. A word having this form is not cube-free, and, thus, the partial word obtained by replacing one of its symbols with a hole is not cube-free, as well.

The strategy of replacing randomly a symbol of w with a hole produces a cube-free infinite partial word in all the cases, but it does not necessarily produce an overlap-free infinite partial word. Indeed, one can prove that w is overlap-free in the same way as we have proved that it is cube-free. However, w contains a factor *cabcdabcd*, in which the first symbol can be replaced with a hole, and, in this manner, obtain an infinite partial word that is not overlap-free. The problem of finding an the symbols of w that can be replaced with a hole such that the word we obtain is overlap-free seems interesting to us.

The main result that we propose in this section is the following:

Proposition 3 *If w' is an infinite partial word obtained from $w = \phi(t)$ by replacing an infinite number of its symbols with holes, such that each two consecutive holes are separated by at least two symbols, then w' is cube-free.*

Proof. Assume, for the purpose of contradiction, that w' contains a non-empty factor $x_1x_2x_3$ and there exists a partial word u, such that $x_i \subset u$, for all $i \in \{1,2,3\}$.

It is not hard to see that $|x_1| \geq 3$. Note that there exists $k \leq |x_1|$ such that the k^{th} symbols of x_1 and x_3 are both different from $\diamond$; this holds because, otherwise, it is impossible to have at least two symbols between every two consecutive holes. This remark proves that the length of $|x_1|$ is

even (otherwise, the distance between the two identical symbols placed on the k^{th} position of x_1 and x_3 is a number of the form $4m+2$, a contradiction). In a similar fashion we can show that there exists $l \leq |x_1|$ such that the l^{th} symbols of x_1 and x_2 are both different from $\diamond$; combined with the fact that $|x_1|$ is even, this leads to the fact that $|x_1| = 4m$, for some $m \in \mathbb{N}$.

Let y_1, y_2 and y_3 be the factors of w from which x_1, x_2 and, respectively, x_3 were obtained, by replacing some of their symbols with holes. Since $|y_i| = |x_i| = 4m$, for $i \in \{1, 2, 3\}$, it follows that these words have the following form: $y_i = w_{i1}b_{i,1} \dots b_{i,m-1}w_{i2}$, such that: $b_{i,j} = \phi(l_{i,j})$, with $l_{i,j} \in \{a, b\}$, $|w_{11}| = |w_{21}| = |w_{31}|$, $|w_{12}| = |w_{22}| = |w_{32}|$, $w_{12}w_{21} = \phi(l_1)$, and $w_{22}w_{31} = \phi(l_2)$, with $l_1, l_2 \in \{a, b\}$. We assume that, for $i \in \{1, 2, 3\}$, we have $x_i = u_{i1}c_{i,1} \dots c_{i,m-1}u_{i2}$, where $c_{i,j}$ was obtained from $b_{i,j}$, u_{i1} from w_{i1}, and u_{i2} from w_{i2}, for all i and j, respectively, by replacing some of their symbols with holes.

If there exist $j \in \{1, \dots, m-1\}$ and $i, k \in \{1, 2, 3\}$ such that $i \neq k$ and $b_{i,j} \neq b_{k,j}$, it follows that $b_{i,j}$ and $b_{k,j}$ differ on every position, i.e., the o^{th} symbols of $b_{i,j}$ differs from the o^{th} symbol of $b_{k,j}$, for all $o \in \{1, 2, 3, 4\}$. Consequently, at least 4 symbols in both these words must be substituted with holes in order to obtain $c_{i,j}$ and $c_{k,j}$, which are both included in the same partial word. But this is impossible, since two consecutive holes are separated by at least two symbols. Thus, we obtain that $b_{1,j} = b_{2,j} = b_{3,j}$, for all $1 \leq j \leq m$. This proves that $y_i = w_{i1}xw_{i2}$, for $1 \leq i \leq 3$, and $x = \phi(x')$, for some factor x' of t.

In the same manner, we can show that $w_{12}w_{21} = w_{22}w_{31}$, and, as a consequence $l_1 = l_2 = l$. Note that if $x \neq \lambda$, we deduce that w contains the factor $x\phi(l)x\phi(l)x$, and, since $x = \phi(x')$, it follows that t is not overlap-free (having as a factor the word $x'lx'lx'$), a contradiction. Hence, we may assume, for the rest of the proof, that $x = \lambda$.

Moreover, we can obtain, similarly, that if $|w_{11}| \geq 3$ then $w_{11} = w_{21} = w_{31}$; but this would prove that w contains the factor $vy_1y_2y_3$, where $v \in \{a, b, c, d\} \cup \{\lambda\}$ such that $vw_{11} = w_{12}w_{21} = w_{22}w_{31} = \phi(l)$. Consequently, w contains the factor $vw_{11}w_{12}w_{21}w_{22}w_{31}$, a contradiction to the fact that w is cube-free. Analogously, the case when $|w_{12}| \geq 3$ leads to a contradiction.

Thus, the only possibility left to be analyzed is when we have $|w_{i1}| = |w_{i2}| = 2$, for all $i \in \{1, 2, 3\}$. If $w_{11} = w_{21}$ or $w_{32} = w_{12}$, we obtain again, easily, a contradiction. Hence, we have $w_{11} \neq w_{21}$ (which implies that they differ on every position) and $w_{32} \neq w_{12}$ (also implying that they differ on every position). Since u_{ij} was obtained from w_{ij} by substituting some of their symbols with holes, for all $i \in \{1, 2, 3\}$ and $j \in \{1, 2\}$, it follows that the strings $u_{32}u_{11}$ and $u_{12}u_{21}$ are both contained in the same word (which consists in the last two symbols of u followed by the first two symbols of u). Again, this means that at least 4 symbols in the strings $w_{32}w_{11}$ and $w_{12}w_{21}$ were substituted with holes. But this is impossible because each two consecutive holes are separated by at least two symbols.

We have shown that all the cases lead to a contradiction, and, consequently, the assumption that we have made, namely that w' is not cube-free, is false. This concludes our proof.

□

The following corollary results immediately:

Corollary 4 *If w' is an infinite partial word obtained from $w = \phi(t)$ by inserting an infinite number of holes, such that each two consecutive holes are separated by at least two symbols, then w' is k-free, for every $k \geq 3$.*

This time, an algorithm that produces a k-free word with n holes, for $k \geq 3$, can be obtained more easily: we construct the prefix of length $3n-2$ of $\phi(t)$ and replace n of its symbols with holes, such that the number of symbols between two consecutive holes is 2. In this way we obtain the cube-free partial word (thus, k-free partial word) with exactly n holes.

Algorithm 3
function $Construct-cube-free-word-4-symbols(n)$;
begin
construct $t_{\lceil \log_2 \lceil 3n/4 \rceil \rceil}$ *(this word has* $\lceil 3n/4 \rceil$ *symbols);*
construct $u = \phi(t_{\lceil \log_2 \lceil 3n/4 \rceil \rceil})$ *(this word has at least* $3n$ *symbols);*
construct v *as the prefix of length* $3n-2$ *of* u*;*
construct v' *from* v *by replacing the symbols on the positions* $1, 4, \ldots, 3n-2$ *with holes;*
return $Construct-cube-free-word-4-symbols(n) := v'$; *(the algorithm stops)*
end.

The time complexity of this algorithm, as in the case of Algorithm 2, is clearly $\mathcal{O}(n)$, as well as its space complexity. Remark, also, that the partial word produced by this algorithm has the minimal length that a cube-free partial word containing n holes can have. Indeed, if the length of a partial word w is less than $3n-2$ it follows that in this word one can find two holes that are separated by at most one symbol, and, consequently, it has at least one factor of the form $\diamond\diamond l$, $\diamond l\diamond$ or $l\diamond\diamond$, for some symbol l. In all these cases w is not cube-free.

6 Conclusions

We state, as a possible sequel of the work presented here, the open problem of the existence of overlap-free infinite partial words, as well as the problem of designing efficient algorithms for the construction of such words. Also, from the algorithmic point of view, there are several problems that seem interesting to us, and we were not able to solve efficiently nor to prove that they are intractable. First, given a k-free full word, what is the maximum number symbols that can be replaced with holes in this word, such that

the partial word we obtain is k-free? Second, given a partial word over an alphabet V, find, if any, a possibility to replace each hole with a symbol from V such that the word obtained in this fashion is k-free; is there a method to compute the number of all these possibilities efficiently?

Acknowledgments

We would like to thank the anonymous referees for their useful comments and suggestions. Also, we thank to Francine Blanchet-Sadri and Victor Mitrana for their suggestions and careful reading of our paper.

References

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

[2] J.-P. Allouche, J. Shallit, *Sums of digits, overlaps, and palindromes*, DMTCS 4(1): 1–10, 2000.

[3] J. Berstel, L. Boasson, *Partial words and a theorem of Fine and Wilf*, Theor. Comput. Sci. 218(1): 135–141, 1999.

[4] F. Blanchet-Sadri, R. A. Hegstrom, *Partial words and a theorem of Fine and Wilf revisited*, Theor. Comput. Sci. 270(1-2): 401–419, 2002.

[5] F. Blanchet-Sadri, *Codes, orderings, and partial words*, Theor. Comput. Sci. 329(1-3): 177–202, 2004.

[6] F. Blanchet-Sadri, S. Duncan, *Partial words and the critical factorization theorem*, J. Comb. Theory, Ser. A 109(2): 221–245, 2005.

[7] F. Blanchet-Sadri, *Primitive partial words*, Discr. Appl. Math. 148(3): 195-213, 2005.

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

[9] C. Choffrut, J. Karhumäki, *Combinatorics on words*, in G. Rozenberg and A. Salomaa (Eds.), Handbook of Formal Languages, Vol. 1, Ch. 6, 329-438, Springer-Verlag, Berlin, 1997.

[10] M. Crochemore, W. Ryter, *Jewels of Stringology*, World Scientific Publishing Co., 2002.

[11] T. Harju, *Combinatorics on words*, in Z. Esik, C. Martin-Vide, V. Mitrana (Eds.): Recent Advances in Formal Languages and Applications, Ch. 19, 381–392, Springer-Verlag, Berlin, 2006.

[12] T. Head, G. Paun, D. Pixton, *Language theory and molecular genetics*, in G. Rozenberg and A. Salomaa (Eds.), Handbook of Formal Languages, Vol. 2, Ch. 7, 295-360, Springer-Verlag, Berlin, 1997..

[13] J. Karhumäki A. Lepistö, *Combinatorics on infinite words*, in Z. Esik, C. Martin-Vide, V. Mitrana (Eds.): Recent Advances in Formal Languages and Applications, Ch. 20, 393-410, Springer-Verlag, Berlin, 2006.

[14] L. Kari, G. Rozenberg, A. Salomaa, *L systems*, in G. Rozenberg and A. Salomaa (Eds.), Handbook of Formal Languages, Vol. 1, Ch. 7, 253-328, Springer-Verlag, Berlin, 1997.

[15] P. Leupold, *Languages of partial Words - how to obtain them and what properties they have*, Formal Grammars, 7: 179–192, 2004.

[16] P. Leupold, *Partial words for DNA coding*, Lecture Notes Comp. Sci. 3384: 224–234, 2005.

[17] G. Lischke, *Restoration of punctured languages and similarity of languages*, Math. Logic Quart. 52: 20–28, 2005.

[18] M. Lothaire, *Combinatorics on words*, Encyclopedia of Mathematics, Vol. 17, Addison-Wesley, 1983.

[19] U. Manber, G. Myers, *Suffix arrays: a new method for on-line string searches*, SIAM J. Comput., 22(5): 935–948, 1993.

[20] A.M. Shur, Y.V. Gamzova, *Partial Words and the periods' interaction property*, Izvestiga RAN 68: 199–222, 2004.

[21] A.M. Shur, Y.V. Konovalova, *On the periods of partial words*, Lecture Notes Comp. Sci. 2136: 657-665, 2001.

[22] A. Thue, *Über unendliche Zeichenreihen*, Norske Vid. Selsk. Skr. I, Mat. Nat. Kl. Christiana 7: 1-22, 1906. Reprinted in T. Nagell, A. Selberg, S. Selberg, and K. Thalberg (Eds.), Selected Mathematical Papers of Axel Thue. Oslo, Norway: Universitetsforlaget, 139–158, 1977.

[23] A. Thue, *Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen*, Norske Vid. Selsk. Skr. I, Mat. Nat. Kl. Christiana 1: 1–67, 1912. Reprinted in T. Nagell, A. Selberg, S. Selberg, and K. Thalberg (Eds.), Selected Mathematical Papers of Axel Thue. Oslo, Norway: Universitetsforlaget, 413-478, 1977.