RAIRO-Theor. Inf. Appl. 42 (2008) 717–728
DOI: 10.1051/ita:2008002

Available online at:
www.rairo-ita.org

A HIERARCHY FOR CIRCULAR CODES

Giuseppe Pirillo[1, 2]

Abstract. We first prove an extremal property of the infinite Fibonacci* word f: the family of the palindromic prefixes $\{h_n \mid n \geq 6\}$ of f is not only a circular code but "almost" a comma-free one (see Prop. 12 in Sect. 4). We also extend to a more general situation the notion of a necklace introduced for the study of trinucleotides codes on the genetic alphabet, and we present a hierarchy relating two important classes of codes, the comma-free codes and the circular ones.

Mathematics Subject Classification. 68R15, 94A45.

**À la Maison Heinrich Heine à l'occasion de son 50$^{\text{e}}$ anniversaire
avec les remerciements de tout mon cœur pour le soutien
qu'elle a donné pendant les 30 dernières années,
à mes études et à mes recherches**

**Für das Heinrich-Heine-Haus aus Anlass seines 50-jährigen Bestehens
mit meinem herzlichsten Dank für die Unterstützung
meines Studiums und meiner wissenschaftlichen Arbeit
in den letzten 30 Jahren**

1. Introduction

The notion of a "code" has very different meanings in biology and theoretical computer science. In biology the "genetic code" associates the 64 trinucleotides with 20 amino acids. While in theoretical computer science a "code" is a set of uniquely decipherable words. Nevertheless, some notions in theoretical computer science, such as comma-free codes and circular codes, are useful in biology, see [1,2,6]. A theoretical method based on the notion of a necklace [20] has been

Keywords and phrases. Theory of codes, comma-free codes, circular codes.

[1] IASI CNR, Viale Morgagni 67/A, 50134 Firenze, Italy; pirillo@math.unifi.it

[2] Université de Marne-la-Vallée 5, boulevard Descartes Champs sur Marne 77454 Marne-la-Vallée Cedex 2, France.

*Bigollo (Leonardo Pisano).

Article published by EDP Sciences

useful for a complete description of the set of the trinucleotides self-complementary circular codes [21] and also for a detailed study of the trinucleotides comma-free codes [15].

We introduce the key notion of "tiling" (tessellation) of a word w which is, roughly speaking, just a sequence $w_1, w_2, \ldots, w_n$ of words such that w is a factor of $w_1 w_2 \cdots w_n$ (the sequence fills w with no overlaps and no gaps in accordance with the Latin sense of the word "tessella", a small cubical piece used to make mosaics). When the w_i belong to a finite set, even if there are potentially infinitely many tilings of a word w, only the minimal ones (see Defs. 1 and 2) are interesting. Using these notions of tilings and minimal tilings, we present here a general result, which "in nuce" was already in [20]: there is a hierarchy of codes $\mathcal{P}_0$, $\mathcal{P}_1$, $\ldots$, $\mathcal{P}_n$, $\ldots$ such that $\mathcal{P}_0$ is the class of comma-free codes and a finite code X is circular if and only if there exists a positive integer n such that X is in class $\mathcal{P}_n$.

We like to conclude this introduction, pointing out that our result (which belongs to theoretical computer science) has pratically been suggested by the concrete study of the combinatorial properties of trinucleotides (which belongs to bio-informatics). More precisely, this result is a fruit of the reflection on the fact that, under certain conditions, the Arquès and Michel code (see [1,2]) allows us to retrieve the frame 0 using a window of length 13, *i.e.*, four trinucleotide and a nucleotide.

2. Preliminary definitions and properties

We denote by A an *alphabet*, by A^* the *free monoid* on A, by A^+ the *free semigroup* on A, by ϵ the *empty word*, and, finally, by $|u|$ the *length* of a word $u \in A^*$. We consider a word u of length $k \geq 1$ as a map $u : \{1, 2, \ldots, k-1, k\} \to A$; we write $u = u(1)u(2) \cdots u(i) \cdots u(k-1)u(k)$; we denote by $\widetilde{u}$ the mirror image $u(k)u(k-1) \cdots u(2)u(1)$ of u and we say that a non-empty word v is a *palindrome* if $v = \widetilde{v}$. A word u is a *factor* of a word v if there exist two words $u', u'' \in A^*$ such that $v = u'uu''$. When $u' = \epsilon$ (resp. $u'' = \epsilon$) we say that u is a *prefix* (resp. *suffix*) of v. A *proper factor* (resp. *proper prefix*, *proper suffix*) u of v is a factor (resp. prefix, suffix) u of v such that $|u| < |v|$.

A (right) *infinite word* on A is a map q from the set of positive integers into A. We write $q = q(1)q(2) \cdots q(i) \cdots$. A word u is a *factor* of q if there exist a word u' and an infinite word q' such that $q = u'uq'$. If $u' = \varepsilon$ we say that u is a *prefix* of q. A non-empty word u may be a factor of another (finite or infinite) word w in more than one way. So it is useful to speak about occurrences. For this reason, let i, j be integers such that $1 \leq i \leq j$ (with $j \leq |w|$ if w is a finite word) and let us denote by $w(i,j)$ the factor $w(i) \cdots w(j)$ of w. We say that the pair of integers (i,j) is an *occurrence* of the factor u in the word w if $u = w(i,j)$. We denote by $F(t)$ the set of all non-empty factors of a finite or infinite word t. Given a subset X of A^* we denote by X^n the set of the words on A which are product of n words of X. We denote by A^ω the set of the infinite words on A and by X^ω the set of the infinite words on A which have the form $x_1 x_2 \cdots x_i \cdots$ with $x_i \in X$.

Hereafter β_4 is the 4-letter *genetic alphabet*. A *nucleotide* is a letter of β_4. A *domino* (resp. *trinucleotide*) is a word of length 2 (resp. 3) over β_4. We denote by β_4^3 the set of all trinucleotides over β_4. For more details, see for example, [1,2] or [20]. An ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n$ of nucleotides l_i and dominoes d_i is an n-*necklace* for a subset $X \subset \beta_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n \in X$. In [20], we proved that if X is a subset of β_4^3 then the condition "*X is a circular code*" is equivalent to the condition "*X has no* 5-*necklace*".

Now imagine writing an infinite sequence (this is a concept extraneous to biology!) of trinucleotides x_i, with $x_i \in X$, where X is a subset of β_4^3. Furthermore, imagine shifting the reading frame by one or two trinucleotides. Perhaps, we are able to read as a prefix just one trinucleotide of X in the shifted sequence (in this case X is not a comma-free code). Or maybe we are able to read as a prefix only the product of two, three, four, etc. consecutive trinucleotides of X. We might even be able to factorize the whole shifted sequence by trinucleotides of X. Anyway, if X is a circular code, then we are able to read at most four consecutive trinucleotides of X. This corresponds to the window of Arquès and Michel, which has 13 nucleotides, and is also the meaning of the above result recalled from [20].

In [20], using the notion of a necklace, we characterized the (necessarily finite) languages of trinucleotides that are circular codes. In this paper, using the notion of a tiling, we characterize all the finite languages that are circular codes and we also present a hierarchy of circular codes.

Definition 1. Given an infinite word $s = s(1)s(2)\cdots s(i)\cdots$ and factors

$$w = s(i,j),\ w_1 = s(i_1,j_1),\ w_2 = s(i_2,j_2),\ \ldots,\ w_n = s(i_n,j_n)$$

of s, we say that $w_1, w_2, \ldots, w_n$ is a *tiling* (*tessellation*) of w if $j_1 + 1 = i_2$, $j_2 + 1 = i_3, \ldots, j_{n-1} + 1 = i_n$ and $i_1 \leq i \leq j \leq j_n$. We say that $w = s(i,j)$ is the *trivial tiling* of w.

Definition 2. We say that a tiling $s(i_1,j_1)$, $s(i_2,j_2)$, $\ldots$, $s(i_n,j_n)$ of $w = s(i,j)$ is *minimal* if $s(i_2,j_2)$, $\ldots$, $s(i_{n-1},j_{n-1})$, $s(i_n,j_n)$ is not a tiling of $w = s(i,j)$ and $s(i_1,j_1)$, $s(i_2,j_2)$, $\ldots$, $s(i_{n-1},j_{n-1})$ is not a tiling of $w = s(i,j)$. In other words, $s(i_1,j_1)$, $s(i_2,j_2)$, $\ldots$, $s(i_n,j_n)$ is minimal if $i_1 \leq i < i_2$ and $j_{n-1} < j \leq j_n$.

Examples. Consider $f = f(1)f(2)\cdots f(i)\cdots = abaababaa\cdots$ and note that $f(2,4) = baa$, $f(5,7) = bab$ is a minimal tiling of $f(3,6) = aaba$. Note also that $aba = f(4,6)$ is a minimal tiling of $aba = f(4,6)$ (a factor is always a trivial tiling of itself and is clearly minimal) and, for each $n \leq 4$ and $n' \geq 6$, the factor $f(n,n')$ is again a minimal tiling of $aba = f(4,6)$. So minimal is intended in the sense that no unnecessary word is used on the left and on the right, *i.e.*, at the beginning and end of the sequence of the occurrences. In Figure 1, x_1, x_2 is a minimal tiling of y_1; x_3, x_4 is a minimal tiling of y_2; x_4, x_5 is a minimal tiling of y_3 but, for example, x_1, x_2 x_3 is not a minimal tiling of y_1 and x_2, x_3 x_4 is not a minimal tiling of y_2.

Definition 3. Let $s = s(1)s(2)\cdots s(i)\cdots$ be an infinite word and let $s(i,j)$ and $s(i',j')$ be two occurrences of the same factor w of s. Then, given a tiling

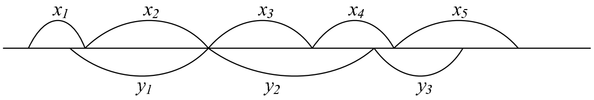

FIGURE 1. Examples of tilings.

$s(i_1, j_1)$, $s(i_2, j_2)$, ..., $s(i_n, j_n)$ of $s(i, j)$ and a tiling $s(i'_1, j'_1)$, $s(i'_2, j'_2)$, ..., $s(i'_n, j'_n)$ of $s(i', j')$, we say that these two tilings are *equivalent* if there exists factors w_1, w_2, ... , w_n of s such that $s(i_1, j_1) = s(i'_1, j'_1) = w_1$, $s(i_2, j_2) = s(i'_2, j'_2) = w_2$, ..., $s(i_n, j_n) = s(i'_n, j'_n) = w_n$ (see Fig. 2).

Definition 4. A subset X of A^+ is said to be a *code* over A if for all $n, m \geq 1$ and $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$ the condition

$$x_1 \cdots x_n = x'_1 \cdots x'_m$$

implies

$$n = m \text{ and } x_i = x'_i \text{ for } i = 1, \ldots, n.$$

Definition 5. Let $\{x_\alpha\}_{\alpha \geq 1}$ be a fixed infinite sequence of finite words from a set X and let x be an infinite sequence such that $x = x_1 x_2 \cdots x_\alpha \cdots \in X^\omega$. We say that the infinite set of integers $\{1 = i_1, i_2, \ldots, i_\alpha, \ldots\}$ satisfying $i_1 < i_2 < \cdots < i_\alpha < \cdots$ is the *natural tiling set* of x if $x_1 = x(i_1, i_2 - 1), x_2 = x(i_2, i_3 - 1), \ldots, x_\alpha = x(i_\alpha, i_{\alpha+1} - 1), \ldots$ and we denote it by $T(x)$. If $y = y_1 y_2 \cdots y_n \in X^n$ and $y_1 y_2 \cdots y_n = x(j_1, j_2 - 1) x(j_2, j_3 - 1) \cdots x(j_n, j_{n+1} - 1)$ is an occurrence of y in x, we say that the finite set of integers $\{j_1, j_2, \ldots, j_n, j_{n+1}\}$, with $j_1 < j_2 < \cdots < j_n < j_{n+1}$, is the *local tiling set* of this occurrence of y in x and, shortly, we denote it by $T(y)$.

The relationship between local tiling sets and natural tiling sets can be very different. Consider, for example, the first four occurrences of aa in an infinite word x in X^ω where X is the suffix code $\{a, aaab\}$ and x begins with $aaaaab$.

Definition 6. Let X be a subset of A^+ and let k be a non-negative integer. We say that X has the *property* $\mathcal{P}_k$ if, for each infinite sequence $x \in X^\omega$ and local tiling set $T(y)$ of an occurrence of a factor y in x, we have

$$Card(T(y) \setminus T(x)) \leq k.$$

In other words, at most k elements of $T(y)$ are not in $T(x)$. That is, the intersection of $T(y)$ with the complement of $T(x)$ has at most k elements.

Examples. For each $k \geq 1$, the subsets $\{a^k b, a\}$ and $\{ab^k, b\}$ of $\{a, b\}^+$ belong to $\mathcal{P}_k$, and the subset $\{ab^k c, b\}$ of $\{a, b\}^+$ belongs to $\mathcal{P}_{k+1}$, as one can easily verify. Moreover $\{a, b\}$ and $\{ac, b\}$ belong to $\mathcal{P}_0$. In Figure 1, an occurrence of $y = y_1 y_2 y_3$ in $x = x_1 x_2 x_3 x_4 x_5 \cdots$ has a local tiling set with 3 integers which do not belong to the natural tiling set of x.

Proposition 1. *If a subset X of A^+ has the property $\mathcal{P}_k$ for some non-negative integer k, then X is a code.*

Proof. Suppose, by way of contradiction, that X has the property $\mathcal{P}_k$, for some integer k, but X is not a code. Consider the equality $x_1 x_2 \cdots x_n = y_1 y_2 \cdots y_m$,

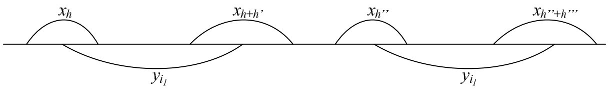

FIGURE 2. The equivalent minimal tilings of y_{i_1} and y_{i_2}.

where $x_1x_2\cdots x_n$ and $y_1y_2\cdots y_m$ are different, and suppose without loss of generality that $x_1 \neq y_1$. Consider the infinite word $x = (x_1x_2\cdots x_n)^\omega \in X^\omega$. Then the word $y = y_1y_2\cdots y_m$ has an occurrence as a prefix of x and $Card(T(y)\cap(T(x)) \geq 1$. But clearly $y^p = (y_1y_2\cdots y_m)^p$ is again a prefix of x and $Card(T(y^p)\cap(T(x)) \geq p$. Since this holds for any p, there is no positive integer k such that X has the property $\mathcal{P}_k$. Contradiction. □

Definition 7. Let X be a subset of A^+. We say that X has the *property* $\mathcal{P}$ if X has the property $\mathcal{P}_k$ for some non-negative integer k.

Proposition 2. *If a subset X of A^+ has the property $\mathcal{P}$, then X is a code.*

Proof. It is very easy. Suppose that X has the property $\mathcal{P}$. By definition, for some integer k, X has the property $\mathcal{P}_k$, and hence X is a code by Proposition 1. □

Definition 8. Given a code X over A we say that it is *prefix* (resp. *suffix*) if $u = v$ whenever $u, v \in X$ and u is a prefix (resp. suffix) of v. A code is *bifix* if it is both a prefix code and a suffix code.

The set $X = \{ab, ba\}$ does not have the property $\mathcal{P}$, as one can easily see considering the factors $(ba)^n$ of $(ab)^\omega$, but it is a bifix code.

Definition 9. A code X over A is said to be *comma-free* if, for each $y \in X$ and $u, v \in A^*$ such that $uyv = x_1\cdots x_n$ with $x_1, \ldots, x_n \in X$, we have $u, v \in X^*$.

A comma-free code X has the easiest deciphering [4]: if $w = x_1\cdots x_n \in X^*$ and $x \in F(w)\cap X$ then there is an integer i, $1 \leq i \leq n$, such that $x = x_i$. We have

Proposition 3. *A code X has the property $\mathcal{P}_0$ if and only if it is comma-free.*

Definition 10. A subset X of A^+ is a *circular code* over A if, for each $n, m \geq 1$ and for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m$ in X, $p \in A^*$ and $s \in A^+$, the conditions

$$sx_2\cdots x_np = x'_1\cdots x'_m$$

and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ and, for $i = 1, \ldots, n$, $x_i = x'_i$.

3. The infinite Fibonacci word

Let $\varphi : \{a, b\}^* \to \{a, b\}^*$ be the morphism given by $\varphi(a) = ab$, $\varphi(b) = a$. The n–th finite Fibonacci word f_n is defined in the following way: $f_0 = b$ and, for each $n \geq 0$, $f_{n+1} = \varphi(f_n)$. For each $n \geq 2$, $|f_n|$ is the n-th element F_n of the sequence of Fibonacci numbers. The infinite Fibonacci word f (see, [7–9,11–14,16,18,19,22]) is the unique infinite word over $\{a, b\}$ such that, for each n, the word f_n is a prefix of f. For each $n \geq 2$, we denote by h_n the prefix of f of length $|f_n| - 2$.

We will often consider some subsets of the family of palindromes $\{h_n \mid n \geq 3\}$ that are very useful examples of codes with the property $\mathcal{P}$.

Lemma 1 (below) is often useful in the study of properties of f. It belongs to the folklore and is very easy to prove. Item i) of the lemma is known as the "*near-commutative property*" (see [13]) and plays a central role in the combinatorics of the Fibonacci word. Note that, for each $n \geq 2$, $g_n = f_{n-2}f_{n-1}$ and a factor v of f is *special* if $va, vb \in F(f)$.

Lemma 1. *For each $n \geq 2$,*

i) $f_n = h_nxy$ and $g_n = h_nyx$, where $xy = ab$ if n is even and $xy = ba$ if n is odd;
ii) $|h_n| = F_n - 2$;
iii) h_n is a special factor;
iv) $h_{n+3} = h_{n+1}xyh_nyxh_{n+1}$, where $x, y \in \{a, b\}$, $x \neq y$;
v) h_n is a palindrome;
vi) $h_{n+2} = f_nh_{n+1} = h_{n+1}\widetilde{f}_n = h_n\widetilde{f}_{n+1} = f_{n+1}h_n$;
vii) for each integer $m \geq 0$, h_n is a prefix and a suffix of h_{n+m};
viii) $v \in F(f)$ if and only if $\widetilde{v} \in F(f)$.

Some properties of the family $\{h_n \mid n \geq 3\}$ also hold for the palindromic prefixes of a standard Sturmian word. See [5].

We will often use the following result of Karhumäki.

Proposition 4 [11]. *The Fibonacci word f is 4-power free, i.e., no factor of f has the form u^4.*

Let $u, v, w, z, z' \in F(f)$. We say that (u, v, w) is an *overlap* of z and z' if $uv = z$, $vw = z'$. When the *central component* v of an overlap (u, v, w) is non-empty, we say that the overlap is *strict*. In this case, v is a proper suffix of z and a proper prefix of z'.

Proposition 5. *Let $j \geq 3$. If (u, v, w) is a strict overlap of h_j and h_j then we have $v \in \{h_n \mid n \geq 3\}$.*

Proof. Let $v = v(1)v(2)\cdots v(k-1)v(k)$. For $1 \leq i \leq k$, we also have $v(i) = h_j(i) = h_j(|h_j| - i + 1) = v(k - i + 1)$. So $v(i) = v(k - i + 1)$ and v is a palindrome. Since v is a prefix of h_j and consequently a prefix of f, we have $v \in \{h_n \mid n \geq 3\}$ by a well-known property of the palindromic prefixes of f (the so called central words). See [5]. □

Remark 1. In [18] we studied the strict overlaps of the f_n, g_n and h_n under the supplementary condition that $uvw \in F(f)$ and we proved that, in this case, there are exactly two strict overlaps of h_n and h_n, namely $(f_{n-1}, h_{n-2}, \widetilde{f}_{n-1})$ and $(f_{n-2}, h_{n-1}, \widetilde{f}_{n-2})$.

Several properties of the Fibonacci words belong, or are in relation, to the theory of codes. We presented in [19] some relations between Fibonacci numbers, Fibonacci words and the theory of codes. For example: *if $f = u_0u_1\cdots u_i\cdots$ is the factorization of f such that $|u_i| = F_{2i+1}$ then $\{u_i \mid i \geq 0\}$ is a prefix code.* We also proved a similar result with $|u_i| = F_{2(i+1)}$ and another one concerning a factorization of f starting from the beginning and using only words of a bifix code. Related results are proved in [10,17]. Now, recall the following well-known

Proposition 6 [7]. *Each element of the family* $\{f_n \mid n \geq 1\}$ *is a primitive word.*

Proposition 7. *Each element of the family* $\{h_n \mid n \geq 3\}$ *is a primitive word, with the only exception being* $abaaba = h_5$.

Proof. Suppose, by way of contradiction, that $h_j \in \{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is not primitive. Then $h_j = v^k$ with $k \geq 2$. If $k \geq 3$ then a cube is a prefix of f. But, by a preliminary result on *fractional powers* in the Fibonacci word [16], no cube is a prefix of f. So, there remains only the possibility $k = 2$. In this case, using a result concerning squares in the Fibonacci word [22], we should have, for some $m > 0$, $h_j = f_m f_m$. Observe that $h_3 = a$, $h_4 = aba$, which clearly do not have the form $f_m f_m$ (while $h_5 = abaaba = f_3 f_3$). Now $|h_6| = 11 = F_4 + F_4 + F_3 - 2$, which is not even, and hence h_6 is not a square. In general, for $j \geq 7$, $|h_j| = F_j - 2 = 2F_{j-2} + F_{j-3} - 2$ from which it easily follows that, for $j \geq 7$, $|h_j|$ (*i.e.*, $F_j - 2$) cannot be equal to $2F_m$ for some m. So we reach a contradiction. □

Corollary. *Each singleton* $\{h_j\}$, *with* $h_j \in \{h_n \mid n \geq 3\} \setminus \{abaaba\}$, *is a comma-free code.*

Proof. Suppose, by way of contradiction, that the singleton $\{h_j\}$ with $j \geq 3$, $j \neq 5$, is not a comma-free code. Since $\{h_j\}$ is a singleton, there exists $u, v \in \{a, b\}^*$, such that h_j^α, $\alpha \geq 2$, can be written as $h_j^\alpha = u h_j v$ where $u \neq h_j^\beta$ for any positive integer β. This implies that $h_j h_j$ contains an occurrence of h_j that is different from those at the beginning and end. Hence, $h_j = ww' = w'w$ for some $w, w' \in \{a, b\}^+$. Now, by a well-known result in the theory of words, we know that if two words commute, then they are powers of a common word (*e.g.*, see [14]). Therefore, $w = v^p$ and $w' = v^q$ for some word v and positive integers p, q. But then $h_j = v^{p+q}$ where $p + q \geq 2$, contradicting the primitivity of h_j (Prop. 7). □

Lemma 2. *Each element of the family* $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ *is not the product of two or more elements of the family.*

Proof. Suppose, by way of contradiction, that, for $j \neq 5$, the palindrome h_j is the product of two or more elements of the family. This is not the case for $h_3 = a$ and $h_4 = aba$. So $j \geq 6$. Suppose $h_j = u_1 u_2 \cdots u_k$ with $k \geq 2$ and $u_i \in \{h_n \mid n \geq 3\} \setminus \{abaaba\}$ and note that a is both a prefix and a suffix of each u_i. None of the u_i can be equal to a. Otherwise, if $u_1 = a$ then h_j has aa as a prefix, if $u_k = a$ then h_j has aa as a suffix and finally if $u_i = a$, $1 < i < k$, then h_j has aaa as a factor, all of which are impossible.

So we have to prove that each element h_j of the family $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is not the product of two or more elements of the family $\{h_n \mid n \geq 4\} \setminus \{abaaba\}$.

Case $\mathbf{k = 2}$. If $u_1 = u_2 = aba$ we have $h_j = abaaba$. Contradiction. If $u_1 = aba$ and $u_2 \neq aba$ then $(aba)^3$ is a prefix of h_j. Contradiction. If $u_1 \neq aba$ and

$u_2 = aba$ then $(aba)^3$ is a suffix of h_j. Contradiction. If $u_1 \neq aba$ or $u_2 \neq aba$ then $(aba)^4$ is a factor of h_j. By Proposition 4, we have a contradiction.

Case $\mathbf{k \geq 3}$. Consider $u_1u_2u_3$. If $u_1 = u_2 = u_3 = aba$ we have $h_j = (aba)^3$. Contradiction. So at least one u_i with $1 \leq i \leq 3$ is in $\{h_n \mid n \geq 6\}$. If this happens for exactly one i, $1 \leq i \leq 3$, then $(aba)^3$ is a prefix or a suffix of h_j. If this happens for at least two values of i, $1 \leq i \leq 3$, then $(aba)^4$ is a factor of h_j. Again, we have a contradiction, by Proposition 4.

It now remains to prove that each h_j, $j \neq 5$, is not the product of two or more elements of the family $\{h_n \mid n \geq 6\}$. In this case u_iu_{i+1} (and consequently h_j) contains $(aba)^4$ as a factor. So we have a contradiction by Proposition 4. □

The family $\{f_n \mid n \geq 1\}$ is not at all a code (indeed $f_{n+2} = f_{n+1}f_n$). But we do have:

Proposition 8. *The family $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is a code.*

Proof. Suppose, by way of contradiction, that $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is not a code. Consider the equality $x_1x_2 \cdots x_n = y_1y_2 \cdots y_m$ and, without loss of generality, suppose $x_1 \neq y_1$. Suppose also, again without loss of generality, that x_1 is a prefix of y_1. Consider the smallest positive integer i, $1 \leq i \leq n$, such that y_1 is a prefix of $x_1x_2 \cdots x_i$. For some u, v we have $x_1x_2 \cdots x_i = x_1x_2 \cdots x_{i-1}uv = y_1v$. So $y_1 = x_1x_2 \cdots x_{i-1}u$. The word u is the central component of a strict overlap of y_1 and x_i and, by Proposition 5, it belongs to $\{h_n \mid n \geq 3\}$. So y_1 is the product of at least two elements of $\{h_n \mid n \geq 3\}$ which contradicts Lemma 2. □

Proposition 9. *Each subset of A^+ with at least two elements in the family $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is not a comma-free code.*

Proof. Suppose that X is such that $|X \cap (\{h_n \mid n \geq 3\} \setminus \{abaaba\})| \geq 2$ and suppose that, for some positive integers p, q, $p < q$, the palindromes h_p, h_q are in X. We claim that $\{h_p, h_q\}$ is not a comma-free code. Indeed, by Lemma 1, for some $u \in \{a, b\}^*$, $h_q = h_pu$ and if, by way of contradiction, we suppose that $\{h_p, h_q\}$ is a comma-free code, then we also have $h_q = h_pu_1u_2 \cdots u_\alpha$, for some u_1, u_2 …, $u_\alpha \in \{h_p, h_q\}^\alpha$, $\alpha \geq 1$. That is, h_q is the product of two or more elements of $\{h_n \mid n \geq 3\}$, which contradicts Lemma 2. Now, X cannot be a comma-free code because it contains a subset which is not a comma-free code. □

Proposition 10. *The family $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is a circular code.*

Proof. Suppose, by way of contradiction, that $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$ is not a circular code. Then, there exist $x_1, \ldots, x_n, x'_1, \ldots, x'_m$ in $\{h_n \mid n \geq 3\} \setminus \{abaaba\}$, $p \in \{a, b\}^+$ and $s \in \{a, b\}^+$, such that $sx_2 \cdots x_np = x'_1 \cdots x'_m$ and $x_1 = ps$. Note that s and p are central components of strict overlaps of x_1 and x'_1 and of x'_m and x_1 respectively. So, by Proposition 5, $s, p \in \{h_k \mid k \geq 3\}$. Since x_1 is the product of two elements of $\{h_k \mid k \geq 3\}$ and since $x_1 \neq abaaba$, we reach a contradiction by Lemma 2 or by Proposition 8. □

4. A hierarchy

In this section, we give a characterization of all the finite languages that are circular codes (Th. 1). A first relationship between our property $\mathcal{P}$ and some classical notions of theory of codes (limited codes and uniformly synchronous codes) is in Theorem 2. Finally, the classes $\mathcal{P}_0$, $\mathcal{P}_1$, ..., $\mathcal{P}_k$, $\mathcal{P}_{k+1}$, ... constitute a hierarchy of codes (Prop. 11).

Theorem 1. *Given an alphabet A and a finite subset X of A^+, the following conditions are equivalent:*
i) X has the property $\mathcal{P}$;
ii) X is a circular code.

Proof. **i) implies ii).** Suppose that X has the property $\mathcal{P}$. By Proposition 2, X is a code. Suppose, by way of contradiction, that X is not a circular code. Then there exist $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $p \in A^+$ and $s \in A^+$, such that $sx_2 \cdots x_n p = x'_1 \cdots x'_m$, $x_1 = ps$ and the conditions of Definition 10 are not satisfied. Consider the infinite word x:

$$x_1x_2 \cdots x_nx_1x_2 \cdots x_n \cdots x_1x_2 \cdots x_n \cdots = (x_1x_2 \cdots x_n)^\omega$$

and its natural tiling set

$$j_1 = 1 < j_2 < \cdots < j_n < \cdots$$

Note that x has a factor u which admits two factorizations

$$u = sx_2 \cdots x_n p = x'_1 \cdots x'_m.$$

Consider an occurrence of $x'_1 \cdots x'_m$ in x and its local tiling set:

$$j'_1 < j'_2 < \cdots < j'_m < j'_{m+1}.$$

For each integer $\alpha \geq 1$, the element j'_α does not belong to $\{j_1, j_2, \ldots, j_n, \ldots\}$, otherwise, since X is a code, we reach a contradiction.

So $x'_1 \cdots x'_m$ has an occurrence with $m+1$ elements of its local tiling set in the complement of the natural tiling set of $x = (x_1x_2 \cdots x_n)^\omega$.

In a similar way, for each positive integer β, the factor $(x'_1 \cdots x'_m)^\beta$ has an occurrence with $\beta m + 1$ elements of its local tiling set in the complement of the natural tiling set of $x = (x_1x_2 \cdots x_n)^\omega$.

Thus for each positive integer k, X does not have the property $\mathcal{P}_k$ and consequently does not have the property $\mathcal{P}$. Contradiction. □

Proof. **ii) implies i).** Suppose that X is a circular code and, by way of contradiction, suppose that X does not have the property $\mathcal{P}$, *i.e.*, there is no positive integer k such that X has the property $\mathcal{P}_k$.

For each positive integer k, there exist $x = x_1x_2 \cdots x_\alpha \cdots \in X^\omega$ (having $\{i_1 = 1, i_2, \ldots, i_\alpha, \ldots\}$ as a natural tiling set), a factor y of x such that $y = y_1y_2 \cdots y_n$ (with $y_i \in X$) and the local tiling set $\{j_1, j_2, \ldots, j_n, j_{n+1}\}$ of an occurrence of y in x has more than k elements in the complement of $\{i_1 = 1, i_2, \ldots, i_\alpha, \ldots\}$. This means that at least k factors y_i of y have a minimal non-trivial tiling set with the x_i.

Since X is finite and each factor of x has finitely many minimal non-trivial tiling sets with the x_i, we can choose k (for example $k = (2M+3)\delta \cdot |X| + 1$, where δ is the maximum number of minimal non-trivial tiling sets of an element $z \in X$ and M is the maximum of the lengths of elements of X) such that two y_i, say y_{i_1} and y_{i_2}, satisfy $y_{i_1} = y_{i_2} = w$ for some w. Furthermore, for suitable integers h, h', h'', h''', $h'' > h + h'$, the minimal tiling of y_{i_1} is $x_h, \ldots, x_{h+h'}$, the minimal tiling of y_{i_2} is $x_{h''}, \ldots, x_{h''+h'''}$ and $x_h = x_{h''}, \ldots, x_{h+h'} = x_{h''+h'''}$. In other words, the minimal tilings of y_{i_1} and y_{i_2} are equivalent. See Figure 2. Moreover, there exist words p, s and p', s', $p \neq \epsilon$, $p' \neq \epsilon$, such that the tilings of the occurrences y_{i_1} and y_{i_2} of w satisfy

$$\begin{gathered}
x_h = ps = x_{h''},\ x_{h+h'} = p's' = x_{h''+h'''},\\
py_{i_1}s' = x_h \cdots x_{h+h'} \quad , \quad y_{i_1} = sx_{h+1} \cdots x_{h+h'-1}p',\\
py_{i_2}s' = x_{h''} \cdots x_{h''+h'''},\\
s'x_{h+h'+1} \cdots x_{h''-1}p = y_{i_1+1} \cdots y_{i_2-1}.
\end{gathered}$$

Now, we have

$$\begin{aligned}
&sx_{h+1} \cdots x_{h+h'-1}(x_{h+h'})x_{h+h'+1} \cdots x_{h''-1}p\\
&= sx_{h+1} \cdots x_{h+h'-1}(p's')x_{h+h'+1} \cdots x_{h''-1}p\\
&= (sx_{h+1} \cdots x_{h+h'-1}p')(s'x_{h+h'+1} \cdots x_{h''-1}p)\\
&= y_{i_1}y_{i_1+1} \cdots y_{i_2-1}.
\end{aligned}$$

Since, by assumption, p is non-empty, X is not a circular code. Contradiction. □

We refer to [4] for the definitions of *limited codes* and *uniformly synchronous codes*. Combining our Theorem 1 with Theorem 2.6 of [4] we have

Theorem 2. *Given an alphabet A and a finite subset X of A^+, the following conditions are equivalent:*

i) X has the property $\mathcal{P}$;

ii) X is a circular code;

iii) X is a limited code;

iv) X is a uniformly synchronous code.

The hierarchy is justified by the following very easy

Proposition 11. *A code X in A^+ which has the property $\mathcal{P}_k$ also has the property $\mathcal{P}_{k+1}$.*

Remark 2. In the first two classes of our hierarchy there are codes whose elements are factors of the Fibonacci word. Indeed $\{ab, b\}$ and $\{a, ab\}$ belong to the class $\mathcal{P}_1$. Moreover, by Lemma 1, the family $\{h_k \mid k \geq 6\}$ is very far from being bifix, so it cannot be comma-free (see [4]). Anyway, we will see now (Prop. 12) that $\{h_k \mid k \geq 6\}$ belongs to the class $\mathcal{P}_2$ and, in this sense, it is "almost" a comma-free code. Note that the relation $\{h_k \mid k \geq 6\} \in \mathcal{P}_2$ requires only a short proof, presented hereafter. On the other hand, some relations (for example $X = \{h_n \mid n \geq 3\} \setminus \{abaaba\} \in \mathcal{P}_4$) require arguments which, as far as we know at the moment, are too long to be contained in this paper. In a forthcoming paper we will discuss these relations.

The following result shows that the family $\{h_k \mid k \geq 6\}$ is "almost" comma-free.

Proposition 12. *The code $\{h_k \mid k \geq 6\}$ belongs to the class $\mathcal{P}_2$.*

Proof. Set $X = \{h_k \mid k \geq 6\}$ and consider an infinite sequence $x = x_1 \cdots x_i \cdots \in X^\omega$, its natural tiling set $\{i_1 = 1, i_2, \ldots, i_\alpha, \ldots\}$ and a factor y of x with $y = y_1 y_2 \cdots y_n$, $y_i \in X$. Suppose, by way of contradiction, that the local tiling set $\{j_1, j_2, \ldots, j_n, j_{n+1}\}$ of an occurrence of y in x has more than 3 elements in the complement of the natural tiling set of x. Without loss of generality we can suppose that exactly three integers i, j, k, $i < j < k$, belong to this local tiling set. Consider $\beta \in \{1, \ldots, n-1\}$ such that $y_\beta = x(j', j-1)$ and $y_{\beta+1} = x(j, j'')$ for some j', j''. Set $y_\beta = v$ and $y_{\beta+1} = w$. Consider also, in the natural tiling set of x, the greatest integer i_γ that is smaller than j. Then, in the natural tiling set of x, the smallest integer that is greater than j is $i_{\gamma+1}$. We have four possibilities:

Case 1. $i_\gamma \leq j'$ and $i_{\gamma+1} \geq j''$. In this case $x(i_\gamma, i_{\gamma+1} - 1)$ contains $(aba)^4$. By Proposition 4, we have a contradiction.

Case 2. $i_\gamma > j'$ and $i_{\gamma+1} \geq j''$. In this case $x(i_\gamma, j-1)$ is the central component of a strict overlap of two elements of X and, by Proposition 5, it belongs to $\{h_k \mid k > 3\}$. If $x(i_\gamma, j-1) = a$ then w must begin with b. Contradiction. If $x(i_\gamma, j-1) = aba$ then w must begin with $ababa$. Contradiction. If $x(i_\gamma, j-1) = abaaba$ then w must begin with b. Contradiction. Finally, if $x(i_\gamma, j-1) \in \{h_k \mid k > 3\}$ then $x(i_\gamma, i_{\gamma+1})$ must contain $(aba)^4$, which is again a contradiction by Proposition 4.

Case 3. $i_\gamma \leq j'$ and $i_{\gamma+1} < j''$. In this case $x(j, i_{\gamma+1}-1)$ is the central component of a strict overlap of two elements of X and, by Proposition 5, it belongs to $\{h_k \mid k > 3\}$. If $x(j, i_{\gamma+1} - 1) = a$ then v must end with b. Contradiction. If $x(j, i_{\gamma+1} - 1) = aba$ then v must end with $ababa$. Contradiction. If $x(j, i_{\gamma+1} - 1) = abaaba$ then v must end with b. Contradiction. Finally, if $x(j, i_{\gamma+1} - 1) \in \{h_k \mid k > 3\}$ then $x(i_\gamma, i_{\gamma+1})$ must contain $(aba)^4$, which is again a contradiction by Proposition 4.

Case 4. $i_\gamma > j'$ and $i_{\gamma+1} < j''$. In this case $x(i_\gamma, j-1)$ is the central component of a strict overlap of two elements of X and, by Proposition 5, it belongs to $\{h_k \mid k > 3\}$. With the the same arguments as above, $x(j, i_{\gamma+1} - 1)$ belongs to $\{h_k \mid k > 3\}$. Consequently $x(i_\gamma, i_{\gamma+1} - 1)$ is the product of two elements of $\{h_k \mid k > 3\}$. By Lemma 2 we have a contradiction. □

The result of Proposition 12 is optimal. Indeed $h_7 = abaababaabaababaaba = abaababah_6 = h_6ababaaba$ and $h_7h_7 = abaababah_6h_6ababaaba$. So h_7^ω has an occurrence of h_6h_6 with local tiling set $9, 20, 30$. Since 9 and 30 are in the complement of $1, 20, 39, \ldots, 1 + 19n, \ldots$ we have $\{h_k \mid k \geq 6\} \notin \mathcal{P}_1$.

Acknowledgements. We thank Dipartimento di Matematica "U. Dini" for giving us a friendly hospitality, Jacques Justin for very helpful conversations, the two referees for their interesting remarks and Amy Glen for a very accurate revision of the manuscript.

References

[1] D.G. Arquès and C.J. Michel, A complementary circular code in the protein coding genes. *J. Theor. Biol.* **182** (1996) 45–58.

[2] D.G. Arquès and C.J. Michel, A circular code in the protein coding genes of mitochondria. *J. Theor. Biol.* **189** (1997) 273–290.

[3] J. Berstel, *Mots de Fibonacci.* Séminaire d'informatique théorique. LITP, Paris (1980–81) 57–78.

[4] J. Berstel and D. Perrin, *Theory of codes.* Academic Press (1985).

[5] J. Berstel and P. Seebold, *Sturmian words*, in Algebraic Combinatorics on words, edited by M. Lothaire. Cambridge University Press (2002).

[6] F.H.C. Crick, J.S. Griffith and L.E. Orgel, Codes without commas. *Proc. Natl. Acad. Sci. USA* **43** (1957) 416–421.

[7] A. de Luca, A combinatorial property of the Fibonacci words. *Inform. Process. Lett.* **12** (1981) 193–195.

[8] A. de Luca, Sturmian words: structure, combinatorics, and their arithmetics. *Theoret. Comput. Sci.* **183** (1997) 45–82.

[9] X. Droubay, Palindromes in the Fibonacci word. *Inform. Process Lett.* **55** (1995) 217–221.

[10] J. Justin and G. Pirillo, On some factorizations of infinite words by elements of codes. *Inform. Process. Lett.* **62** (1997) 289–294.

[11] J. Karhumäki, On cube–free ω–words generated by binary morphism. *Discrete Appl. Math.* **5** (1983) 279–297.

[12] D.E. Knuth, *The Art of Computer Programming.* Addison–Wesley, Reading, Mass. (1968).

[13] D.E. Knuth, J.H. Morris and V.R. Pratt, Fast pattern matching in strings. *SIAM J. Comput.* **6** (1977) 323–350.

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

[15] C.J. Michel, G. Pirillo and M.A. Pirillo, Varieties of comma-free codes. *Comput. Math. Appl.* (in press).

[16] F. Mignosi and G. Pirillo, Repetitions in the Fibonacci infinite word. *RAIRO-Theor. Inf. Appl.* **26** (1992) 199–204.

[17] G. Pirillo, Infinite words and biprefix codes. *Inform. Process Lett.* **50** 293–295 (1994).

[18] G. Pirillo, Fibonacci numbers and words. *Discrete Math.* **173** (1997) 197–207.

[19] G. Pirillo, Some factorizations of the Fibonacci word. *Algebra Colloquium* **6** (1999) 361–368.

[20] G. Pirillo, A characterization for a set of trinucleotides to be a circular code, In *Determinism, Holism, and Complexity*, edited by C. Pellegrini, P. Cerrai, P. Freguglia, V. Benci and G. Israel. Kluwer (2003).

[21] G. Pirillo and M.A. Pirillo, Growth function of self-complementary circular codes. *Biology Forum* **98** (2005) 97–110.

[22] P.P. Séébold, *Propriétés combinatoires des mots infinis engendrés par certains morphismes.* PhD. thesis, L.I.T.P., Paris. (1985).

Received January 26, 2007. Accepted November 22, 2007.