AUSTRALASIAN JOURNAL OF COMBINATORICS
Volume **57** (2013), Pages 49–64

Overlapping factors in words

Manolis Christodoulakis

University of Cyprus
P.O. Box 20537, 1678 Nicosia
Cyprus
christodoulakis.manolis@ucy.ac.cy

Michalis Christou   Maxime Crochemore[*]

King's College London
London WC2R 2LS
U.K.
michalis.christou@kcl.ac.uk   maxime.crochemore@kcl.ac.uk

Costas S. Iliopoulos[†]

King's College London
London WC2R 2LS
U.K.
csi@dcs.kcl.ac.uk

Abstract

The concept of quasiperiodicity is a generalization of the notion of periodicity where in contrast to periodicity the quasiperiods of a quasiperiodic string may overlap. A lot of research has been concentrated around algorithms for the computation of quasiperiodicities in strings while not much is known about bounds on their maximum number of occurrences in words. We study the overlapping factors of a word as a means to provide more insight into quasiperiodic structures of words. We propose a linear time algorithm for the identification of all overlapping factors of a word, we investigate the appearance of overlapping factors in Fibonacci words and we provide some bounds on the maximum number of distinct overlapping factors in a word.

[*] Also at: Université Paris-Est, France.

[†] Also at: Digital Ecosystems & Business Intelligence Institute, Curtin University, GPO Box U1987 Perth WA 6845, Australia

1 Introduction

The notion of periodicity in strings is well studied in many fields like combinatorics on words, pattern matching, data compression and automata theory (see [23, 24]), because it is of paramount importance in several applications, not to talk about its theoretical aspects.

The concept of quasiperiodicity is a generalization of the notion of periodicity, and was defined by Apostolico and Ehrenfeucht in [3]. In a periodic repetition the occurrences of the single periods do not overlap. In contrast, the quasiperiods of a quasiperiodic string may overlap. It was shown that maximal quasiperiodic substrings in a string can be reported in $O(n \log^2 n)$ [3], an algorithm later improved to $O(n \log n)$ time ([6, 18]), shown to be optimal by Groult et al. [13]. We call a border u of a non-empty string y a cover of y, if every letter of y is within some occurrence of u in y. Finding all the covers of a string y of length n can be done in linear time [22, 25]. An $O(\log(\log n))$-time work-optimal parallel algorithm was given later by Iliopoulos and Park in [19]. Apostolico and Breslauer gave an online linear-time algorithm for computing the minimal cover array of a string in [2, 5]. Additionally, Li and Smyth [22] provided a linear time algorithm for the computation of the maximal cover array of a given string; this algorithm gives also all the covers for every prefix of the string.

Seeds are regularities of strings strongly related to the notion of cover, as a seed is a cover of a superstring of the word. An $O(n \log n)$ algorithm for the corresponding problem on seeds was given in [17] and later a linear time algorithm has been proposed by Kociumaka et al. [20]. Recently Christou et al. gave a linear-time algorithm for computing the minimal/maximal left-seed array of y [10], an $O(n \log n)$-time algorithm for computing the minimal right-seed array of y, and a linear-time solution for computing the maximal right-seed array of y [8]. A quadratic time algorithm for the computation of the seed array has been proposed in [10] while Christodoulakis et al. gave some polynomial time algorithms for the approximate seed problem [7].

A lot of research has been concentrated around algorithms for the computation of quasiperiodicities in strings while not much is known about bounds on their number of occurrences in words. The infinite Thue-Morse word was the first example of an infinite overlap-free word [1, 28, 29], proven to be the only infinite binary word having that property [26]. Additionally, it was shown that there is an infinite number of overlap-free infinite partial words with one hole, but none in infinite words with more than one hole [14]. Fibonacci strings are important in many concepts [4] and are often cited as a worst case example for many string algorithms or as a lower bound for the number of periodicities/quasiperiodicities in a string. Over the years much relevant scientific work has been done on them, e.g. identifying all covers of a circular Fibonacci string [16], identifying all maximal quasiperiodicities in Fibonacci words [13], identifying all covers and seeds of a Fibonacci word and covers of a circular Fibonacci string [9], etc.

In this paper we study the overlapping factors of a word as a means to provide more insight into quasiperiodic structures of words. We propose a linear time algorithm for the identification of all overlapping factors of a word, we investigate the appearance of overlapping factors in Fibonacci words and we provide some bounds on the maximum number of distinct overlapping factors in a word.

The rest of the paper is structured as follows. In Section 2, we present the basic definitions used throughout the paper. In Section 3, we prove and also quote some properties for periodicities in words, periodicities and factor structure of Fibonacci words as well as some facts regarding the Fibonacci sequence that will prove useful later on the solution of the problems that we are considering. In the next sections we propose a linear time algorithm for the identification of all overlapping factors of a word (Section 4), we investigate the appearance of overlapping factors in Fibonacci words (Section 5) and we provide some bounds on the maximum number of distinct overlapping factors in a word (Section 6). Finally, we give some future proposals and a brief conclusion in Section 7.

2 Definitions

Throughout this paper we consider a string y of length $|y| = n$, where $n \in \mathbb{Z}^*$, on a fixed alphabet. It is represented as $y[1\,..\,n]$ when $n > 0$. A string w is

- a *factor* of y if $y = uwv$ for two strings u and v;
- a *prefix* of y if $y = wv$ for some string v;
- a *suffix* of y if $y = uw$ for some string u;
- a *border* of y if it is both a prefix and a suffix of y.

A *proper* factor of y is a factor which is not equal to y itself; *proper* prefixes, suffixes and borders are defined similarly. We denote the longest common prefix of two strings x and y as $LCP(x, y)$ and the length of the longest proper border of y as *Border*(y). A non empty string u is a *period* of y if y is a prefix of u^k for some positive integer k, or equivalently if y is a prefix of uy. *Period*(y), is the length of the shortest period of y. A run is a maximal (non-extendable) occurrence of a repetition of exponent at least two. This means that the factor $y[i\,..\,j]$ is a run if:

- $y[i\,..\,j]$ has period p
- $j - i + 1 \geq 2p$
- $y[i-1] \neq y[i+p-1]$ (if $y[i-1]$ is defined), $y[j+1] \neq y[j-p+1]$ (if $y[j+1]$ is defined)
- $y[i\,..\,i+p-1]$ is primitive, that is, it is not a proper integer power (2 or larger) of another string

$$a\ b\ a\quad \underbrace{\overline{a\ b\ \underline{a}}\,\underline{\ b\ a}}_{\text{quasiperiodic square}}$$

Figure 1: Quasiperiodic squares and overlapping factors in F_5

For any non-negative integer ℓ, a string w is called a *superposition* of u and v with an *overlap* of length ℓ if there exist strings x, y, z such that

$$w = xyz, u = xy, v = yz \text{ and } |y| = \ell.$$

Note that a *concatenation* is a superposition with overlap of length 0. The concatenation of two copies of some word x is called a *square* (e.g. $w = xx$), while the concatenation of three copies of x is called a *cube* (e.g. $w = xxx$).

A string w is a *quasiperiodic square* of y if it is a factor of y and $w = xv = ux$, where x and v are non empty words and $|x| > |v|$. A factor x of y is an overlapping factor of y if there exists a quasiperiodic square $w = xv = ux$, where x and v are non empty words and $|x| > |v|$, in y.

A string x is a *cover* of y if y is a superposition of one or more copies of x. Similarly a string x is a *seed* of y if y is a factor of a superposition of one or more copies of x. A left seed of y is a seed of y that is also a prefix of y, and a right seed is a seed of y that is also a suffix of y.

The n^{th} Fibonacci number denoted by f_n is defined by the recurrence and initial conditions:

$$f_0 = 1, \quad f_1 = 1, \quad f_n = f_{n-1} + f_{n-2} \quad n \in \{2, 3, 4, \dots\}.$$

The first few terms are: $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \dots$ We define a (finite) Fibonacci string F_k, $k \in \{0, 1, 2, \dots\}$, as follows:

$$F_0 = b, \quad F_1 = a, \quad F_n = F_{n-1}F_{n-2} \quad n \in \{2, 3, 4, \dots\}.$$

Notice that $|F_n| = f_n$, the n^{th} Fibonacci number.

3 Properties

In this section we prove and also quote some properties for periodicities in words, periodicities and factor structure of Fibonacci words, as well as some facts regarding the Fibonacci sequence that will prove useful later on the solution of the problems that we are considering.

The following property of borders, which we give without proof, is quoted in almost every publication regarding periodicity.

Lemma 1 *[27] Let u be a border of x and let z be a string such that $|z| \leq |u|$. Then z is a border of x if and only if z is a border of u.*

The following lemma gives the relation between covers of a string and a way of finding them.

Lemma 2 *[25] Let u be a cover of x and let $z \neq u$ be a factor of x such that $|z| \leq |u|$. Then z is a cover of x if and only if z is a cover of u.*

Proof. Clearly if z is a cover of u and u is a cover of x then z is a cover of x. Suppose now that both z and u are covers of x. Then z is a border of x and hence of u ($|z| \leq |u|$); thus z must also be a cover of u. □

The following lemmas provide some insight in the periodicities and factor structure of Fibonacci words, developing the background for further investigation of those words.

Lemma 3 *[11] For $n \geq 1$, the set of nonempty borders of F_n is*

$$\{F_n, F_{n-2}, F_{n-4}, F_{n-6}, \dots F_k\}$$

where $k = 1$ if n is odd and $k = 2$ if n is even.

Lemma 4 *[15] $F_k = P_k\delta_k$, where $P_k = F_{k-2}F_{k-3} \mathinner{..} F_1$, $k \geq 2$ and $\delta_k = ab$ if k is even or $\delta_k = ba$ otherwise.*

Proof. Easily proved by induction. □

It is sometimes useful to consider the expansion of a Fibonacci string as a concatenation of two Fibonacci substrings. We define the F_m, F_{m-1} expansion of F_n, where $n \in \{2, 3, \dots\}$ and $m \in \{1, 2, \dots, n-1\}$, as follows:

- Expand F_n using the recurrence formula as $F_{n-1}F_{n-2}$.
- Expand F_{n-1} using the recurrence formula as $F_{n-2}F_{n-3}$.
- Keep expanding as above until F_{m+1} is expanded.

Lemma 5 *[9] The F_m, F_{m-1} expansion of F_n, where $n \in \{2, 3, \dots\}$ and $m \in \{1, 2, \dots, n-1\}$ is unique.*

Proof. Easily proved by induction. □

Lemma 6 *There is no $F_{m-1}F_{m-1}$ in the F_m, F_{m-1} expansion of F_n, where $n \in \{2, 3, \dots\}$ and $m \in \{1, 2, \dots, n-1\}$.*

Proof. Any F_{m-1} in the expansion comes from an expanded $F_{m+1} = F_mF_{m-1}$. Therefore, any F_{m-1} in the F_m, F_{m-1} expansion of F_n must be preceded by an F_m. □

Lemma 7 *There is no $F_mF_mF_m$ in the F_m, F_{m-1} expansion of F_n, where $n \in \{2, 3, \dots\}$ and $m \in \{1, 2, \dots, n-1\}$.*

Proof. The above statement holds for $m = n-1$, since $|F_n| < 3|F_{n-1}|$. For $m < n-1$, Lemma 6 shows that any F_m in the F_{m+1}, F_m expansion of F_n, where $n \in \{2, 3, \dots\}$ and $m \in \{1, 2, \dots, n-2\}$, is preceded by an F_{m+1} and followed by an F_{m+1} or nothing, giving the following cases:

- Expanding $F_{m+1}F_mF_{m+1}$ to $F_mF_{m-1}F_mF_mF_{m-1}$ gives no $F_mF_mF_m$ in the expansion as $F_mF_{m-1}F_mF_mF_{m-1}$ contains no $F_mF_mF_m$ and it can be preceded by at most one F_m (Lemma 6).
- Expanding $F_{m+1}F_m$ to $F_mF_{m-1}F_m$ gives no $F_mF_mF_m$ in the expansion as $F_mF_{m-1}F_m$ contains no $F_mF_mF_m$ and it can be preceded by at most one F_m (Lemma 6).

□

Lemma 8 *[9] The string F_m occurs in F_n precisely at the positions where either F_m or F_{m-1} occurs in the F_m, F_{m-1} expansion of F_n, with $n \in \{5, 6, \dots\}$ and $m \in \{3, 4, \dots, n-2\}$, except for the last F_{m-1} if it is a suffix of F_n.*

Proof. Using the recurrence relation we can get the F_m, F_{m-1} expansion of F_n as shown earlier:

$$F_n = F_mF_{m-1}F_mF_mF_{m-1}F_mF_{m-1}F_m \dots$$

In this expansion, apart from the obvious occurrences of F_m (e.g. the 1st, 3rd, 4th, etc. factors are F_m), there are also occurrences starting where each F_{m-1} starts, if that F_{m-1} is not a suffix of F_n. Note that, by Lemma 6 all F_{m-1} factors are either followed by F_m or they are not followed by any string (F_{m-1} is a suffix of F_n). In the former case, by further expanding the F_m that follows the F_{m-1} as $F_{m-2}F_{m-3}F_{m-2}$, an occurrence of F_m is revealed starting at the position where F_{m-1} occurs: $\underline{F_{m-1}F_{m-2}}F_{m-3}F_{m-2}$ (see also Figure 2).

Next, we prove that no other occurrences of F_m exist in F_n. Notice that, any remaining occurrence should have one of the following forms:

- xy, where x is a non empty suffix of F_m and y a non empty prefix of F_{m-1}. Then both x and y are borders of F_m. It holds that $|x| + |y| = |F_m| = |F_{m-1}| + |F_{m-2}|$, but $|x| \leq |F_{m-2}|$, $|y| \leq |F_{m-2}|$ and so there exists no such occurrence of F_m in F_n.
- xy, where x is a non empty suffix of F_{m-1} and y a non empty prefix of F_m. Then y is also a border of F_m and so belongs to $\{F_{m-2}, F_{m-4}, \dots, F_3\}$, if n is odd, or to $\{F_{m-2}, F_{m-4} \dots F_4\}$, otherwise. But $|x| + |y| = |F_m|$ and $0 < |x| \leq |F_{m-1}|$ so in either case the only solution is $x = F_{m-1}$ and $y = F_{m-2}$ giving the occurrences of F_m at the starting positions of F_{m-1} in the above expansion. (See also Figure 2.)

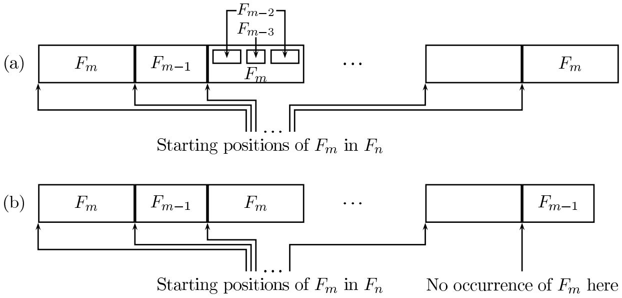

Figure 2: The string F_m occurs as a factor in F_n, starting at the positions where either F_m or F_{m-1} occurs in the F_m, F_{m-1} expansion of F_n (as shown in (a)), except for the case where the F_m, F_{m-1} expansion ends in F_{m-1} (when $n \equiv m \bmod 2$), in which case the last F_{m-1} is not a starting position for F_m (as shown in (b)).

- $xF_{m-1}y$, where x is a non empty suffix of F_m and y a non empty prefix of F_m. Then both x and y are borders of F_m. It holds that $|x| + |y| = |F_{m-2}|$, but as both x and y are non empty $|x| \leq |F_{m-4}|$, $|y| \leq |F_{m-4}|$ and so there exists no such occurrence of F_m in F_n.
- xy, where x is a non empty suffix of F_m and y a non empty prefix of F_m. Then both x and y are borders of F_m. It holds that $|x| + |y| = |F_m|$, but as both x and y are non empty $|x| \leq |F_{m-2}|$ and $|y| \leq |F_{m-2}|$. The only case that allows for that is when $m = 2$, i.e. $F_m = 01$. However $F_2 \neq F_0F_0 = 11$ and so there exists no such occurrence of F_m in F_n.

□

The following lemmas show some facts regarding the Fibonacci sequence that will prove useful later on the solution of the problems that we are considering.

Lemma 9 *[12] The sum of the first n Fibonacci numbers is the $(n+2)^{th}$ Fibonacci number minus 1, i.e. $\sum_{i=1}^{n} f_i = f_{n+2} - 1$.*

Lemma 10 *[12] The sum of the squares of the first n Fibonacci numbers is the product of the n^{th} and the $(n+1)^{st}$ Fibonacci numbers, i.e. $\sum_{i=1}^{n} f_i^2 = f_n f_{n+1}$.*

Lemma 11 *[12] The ratio of successive Fibonacci numbers $\frac{f_n}{f_{n-1}}$ approaches $\phi = \frac{1+\sqrt{5}}{2}$, the golden ratio, as n approaches infinity, i.e.* $\lim_{n \to +\infty} \frac{f_n}{f_{n-1}} = \phi$.

Lemma 12 *[12]* $\sum_{i=1}^{2n} f_i f_{i-1} = f_{2n}^2$.

Lemma 13 *[12]* $\sum_{i=1}^{2n+1} f_i f_{i-1} = f_{2n+1}^2 - 1$.

4 An algorithm for the computation of all overlapping factors of a word

In this section we show the relation between runs and overlapping factors of a word, thus giving a linear time algorithm for their computation.

The following lemma sheds some light on the periodic structure of quasiperiodic squares.

Lemma 14 *A factor w of a string y is a quasiperiodic square if and only if $Period(w) < \frac{|w|}{2}$.*

Proof. ($\Rightarrow$) Let $w = y[i\,..\,j] = (z)^k$, where $k > 2$. Then $w = z^{k-1}v = uz^{k-1}$, where z^{k-1} and v are non empty words and $|z^{k-1}| > |v| = |z|$ and hence w is a quasiperiodic square.

($\Leftarrow$) Let $w = y[i\,..\,j] = xu = vx$, where x and v are non empty words and $|x| > |v|$. As both x and v are prefixes of w and $|x| > |v|$ then v is also a prefix of x. Therefore $w = xu = vx = vvp$, where p is a non empty word. Using similar arguments $w = v^{\frac{|w|}{|v|}}$, where $\frac{|w|}{|v|} > 2$. □

Runs are maximal periodicities, therefore by knowing all the runs of a word we can identify all its periodic factors, its quasiperiodic squares and overlapping factors. Therefore, the following lemma will prove quite useful in the identification of all overlapping factors in a word.

Lemma 15 *[21] There exists a linear time algorithm that identifies all runs in a string y.*

Using the above lemma we are now in a position to describe a linear time algorithm for the identification of all overlapping factors in a word.

Theorem 16 *There exists a linear time algorithm that identifies all overlapping factors in a string y.*

Proof. Immediate consequence of Lemma 14 and Lemma 15.

Any quasiperiodic square of period $|z|$ that appears in y implies the appearance of a run of period $|z|$ in y which must include it and vice versa. Hence, suppose $y[i\,..\,j] = z^k$ is a run, then every factor of $y[i\,..\,j]$ of length ℓ, where $2|z| < \ell \leq j-i+1$, is a quasiperiodic square with period at most $|z|$. If $y[k\,..\,k+\ell-1]$ is such a quasiperiodic square then $y[k\,..\,k+\ell-|z|-1]$ and $y[k+|z|\,..\,k+\ell-1]$ are some of its overlapping factors. Any longer ones imply the appearance of a run with a shorter period (and will be identified with its help). Suppose $y[k\,..\,k+m-1]$ and $y[k+\ell-m\,..\,k+\ell-1]$, with $0 < m < \ell-|z|$, are shorter overlapping factors of this quasiperiodic square. Then $y[k\,..\,k+m-1]$ and $y[k+\ell-m\,..\,k+\ell-1]$ reappear as a prefix of $y[k+|z|\,..\,k+\ell-1]$ and a suffix of $y[k\,..\,k+\ell-|z|-1]$ respectively. Hence, they are overlapping factors of the quasiperiodic squares $y[k\,..\,k+m+|z|-1]$ and $y[k+\ell-m-|z|\,..\,k+\ell-1]$ respectively which are indicated by this run. □

5 Overlapping factors in Fibonacci words

In this section we investigate the appearance of overlapping factors in Fibonacci words. By considering the expansion of a Fibonacci string as a concatenation of two consecutive Fibonacci substrings we are able to identify Fibonacci subwords in a Fibonacci word and the positions of their occurrences, thus being able to derive information for the overlapping factors of Fibonacci words.

The following lemmas identify some overlapping factors in Fibonacci words by looking in the expansion of Fibonacci strings as a concatenation of two consecutive Fibonacci substrings.

Lemma 17 *Factors of F_n of form xF_my, where x is a non empty suffix of F_m, y is a non empty prefix of F_{m-1}, $1 \leq |x| \leq |F_m| - 1$, $1 \leq |y| \leq |F_{m-1}| - 2$ and $3 \leq m \leq n-4$, are overlapping factors of F_n for $n \geq 6$.*

Proof. We consider the F_m, F_{m-1} expansion of F_n. Suppose that there are two occurrences of the required factor (see also Figure 3) that overlap. There are three cases:

- The F_m that appears in the second factor starts k positions $(1 \leq k \leq |F_m| - 1)$ after the F_m that appears in the first factor. As of Lemma 8 k can only be $|F_{m-1}|$. Then the factors $x_1F_my_1$ and $x_2F_my_2$ form a qusiperiodic square of form $x_1F_{m-1}Fmy_2$, which is a contradiction as x_2 is a suffix of F_{m-1}.
- The F_m that appears in the first factor and the F_m that appears in the second factor are consecutive. Hence overlapping factors of this form must lie in a quasiperiodic square of form $F_mF_mF_{m-1}F_m$. Then y can be up to:

$$\begin{aligned} LCP(F_{m-1}, F_{m-3}F_{m-2}) &= LCP(P_{m-1}\delta_{m-1}, F_{m-3}P_{m-2}\delta_{m-2}) \\ &= LCP(P_{m-1}\delta_{m-1}, P_{m-1}\delta_{m-2}) \\ &= P_{m-1} \end{aligned}$$

  Clearly, $1 \leq |x| \leq |F_m| - 1$ and $1 \leq |y| \leq |F_{m-1}| - 2$.
- The F_m that appears in the first factor and the F_m that appears in the second factor are separated by a non-empty factor z. Hence overlapping factors of this form must lie in a factor u of form xF_mzF_my, where z is a non empty word. Then $z = vF_m$ which forces y to be of form F_mw or $F_{m-1}s$ which contradicts the length of y (v, w, s are non empty words).

□

Lemma 18 *Factors of F_n of form xF_my, where x is a non empty suffix of F_{m-1} and y is a non empty prefix of F_{m-1}, with $1 \leq |x| \leq |F_{m-1}| - 1$, $1 \leq |y| \leq |F_{m-1}| - 1$, $|x| + |y| > |F_{m-1}|$ and $3 \leq m \leq n-5$, are overlapping factors of F_n for $n \geq 8$.*

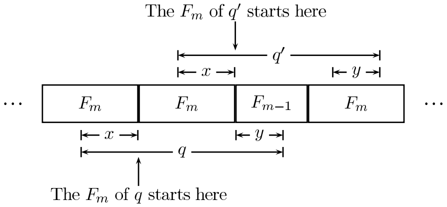

Figure 3: Factors of F_n, q and q' (with $q = q'$), of the form xF_my, where x is a non empty suffix of F_m and y is a non empty prefix of F_{m-1}. Notice that the occurrences of F_m in q and q' are consecutive.

Proof. Lemmas 8, 6 and 7 suggest that quasiperiodic squares containing overlapping factors of the required form must lie in a factor of form $F_{m-1}F_mF_{m-1}F_mF_{m-1}$ in the F_mF_{m-1} expansion of F_n. Clearly, $1 \leq |x| \leq |F_m| - 1$ and $1 \leq |y| \leq |F_{m-1}| - 1$ with the restriction that $|x| + |y| > |F_{m-1}|$, so that the factors overlap. □

Lemma 19 *Factors of F_n of form xF_my, where x is a non empty suffix of F_{m-1}, y is a non empty prefix of F_m and $2 \leq m \leq n-4$, are not overlapping factors of F_n for $n \geq 6$.*

Proof. Lemmas 8, 6 and 7 suggest that quasiperiodic squares containing overlapping factors of the required form must lie in a factor of form $F_{m-1}F_mF_mF_{m-1}F_mF_m$ in the F_mF_{m-1} expansion of F_n but this causes no overlap of the squares. □

Lemma 20 *Factors of F_n of form $xF_{m-1}y$, where x is a non empty suffix of F_m and y is a non empty prefix of F_{m+1}, with $1 \leq |x| \leq |F_m|-1$, $|F_{m-2}|+1 \leq |y| \leq |F_{m+1}|-1$, $|x| + |y| > |F_{m-1}|$ and $3 \leq m \leq n-5$, are overlapping factors of F_n for $n \geq 8$. Factors of F_n of form $xF_{n-5}y$, where x is a non empty suffix of F_{n-4} and y is a non empty prefix of F_{n-3}, such that either ($1 \leq |x| \leq |F_{n-4}| - 1$, $|F_{n-6}| + 1 \leq |y| \leq |F_{n-4}|-2$ and $|x|+|y| > |F_{n-5}|$) or ($1 \leq |x| \leq |F_{n-4}|-1$, $|F_{n-6}|+1 \leq |y| \leq |F_{n-3}|-1$, $|x| + |y| > |F_{n-3}|$), are overlapping factors of F_n for $n \geq 6$.*

Proof. Lemmas 8, 6 and 7 suggest that quasiperiodic squares containing overlapping factors of the required form must lie in a factor of the F_mF_{m-1} expansion of F_n having one of the following forms:

- $F_mF_{m-1}F_mF_{m-1}F_mF_m$, where $3 \leq m \leq n-5$.
  Clearly, $1 \leq |x| \leq |F_m|-1$ and $|F_{m-2}|+1 \leq |y| \leq |F_{m+1}|-1$ with the restriction that $|x| + |y| > |F_{m-1}|$.

- $F_mF_{m-1}F_mF_mF_{m-1}$, where $3 \leq m \leq n-4$.
  Clearly, $1 \leq |x| \leq |F_m| - 1$ and $|F_{m-2}| + 1 \leq |y| \leq |F_m| - 2$ with the restriction that $|x| + |y| > |F_{m-1}|$.

- $F_mF_{m-1}F_mF_mF_{m-1}F_mF_{m-1}$, where $3 \leq m \leq n-4$.
  Clearly, $1 \leq |x| \leq |F_m|-1$ and $|F_{m-2}|+1 \leq |y| \leq |F_{m+1}|-1$ with the restriction that $|x|+|y| > |F_{m+1}|$. □

The following lemma gives all the overlapping factors in a Fibonacci word by identifying factors in the F_m, F_{m-1} expansion of F_n, where $n \in \{2, 3, \dots\}$ and $m \in \{1, 2, \dots, n-1\}$, and by using the results of Lemmas 17, 18 and 20.

Lemma 21 *The overlapping factors of F_n are:*

- *$\emptyset$, if $n = \{0, 1, 2, 3\}$;*
- *factors of form F_my, where y is a possibly empty prefix of F_{m-1}, $0 \leq |y| \leq |F_{m-1}|-1$ and $3 \leq m \leq n-4$, factors of F_n of form xF_m, where x is a possibly empty prefix of F_m, $0 \leq |x| \leq |F_m|-1$ and $3 \leq m \leq n-4$ and factors mentioned in Lemmas 17, 18 and 20, if $n \geq 6$.*

Proof. It is easy to see that the lemma is valid for $0 \leq n \leq 5$.

For greater n, we observe that any overlapping factors are of form xF_my, such that F_m is the leftmost occurrence of the longest Fibonacci word present in the factor, with $m \in \{3, 4, \dots, n-1\}$, $|x| \geq 0$ and $|y| \geq 0$. Clearly, for $m = n-1$ there are no factors of the above form that compose a quasiperiodic square. (Lemma 8 shows that F_{n-1} occurs in F_n only as a prefix of it.)

For $m = n-2$, consider the F_{n-2}, F_{n-3} expansion of F_n: $F_n = F_{n-2}F_{n-3}F_{n-2} = F_{n-2}F_{n-2}F_{n-5}F_{n-4}$. Then $x = 0$ and y can be up to

$$\begin{aligned} LCP(F_{n-2}, F_{n-5}F_{n-4}) &= LCP(P_{n-3}\delta_{n-3}, F_{n-5}P_{m-4}\delta_{n-4}) \\ &= LCP(P_{n-3}\delta_{n-3}, P_{n-3}\delta_{n-4}) \\ &= P_{n-3} \end{aligned}$$

For $m = n-3$, consider the F_{n-3}, F_{n-4} expansion of F_n:

$$F_n = F_{n-3}F_{n-4}F_{n-3}F_{n-3}F_{n-4} = F_{n-3}F_{n-3}F_{n-6}F_{n-5}F_{n-3}F_{n-4}.$$

As before, $x = 0$ and y can be up to $LCP(F_{n-4}, F_{n-6}F_{n-5}) = P_{n-4}$. Considering the second occurrence of F_{n-3} we get quasiperiodic squares composed by factors of form $xF_{n-3}y$, where x is a possibly empty suffix of F_{n-3} and y is a possibly empty prefix of F_{n-4}, with $0 \leq |x| \leq |F_{n-3}|-1$, $0 \leq |y| \leq |F_{n-4}|-1$ and $|x|+|y| > |F_{n-4}|$.

For $m \leq n-4$, we get the following cases:

- Factors of F_n of form F_my, where y is a possibly empty prefix of F_{m-1}, $0 \leq |y| \leq |F_{m-1}|-1$ and $3 \leq m \leq n-4$ form quasiperiodic squares of F_n.
- Factors of F_n of form xF_m, where x is a possibly empty prefix of F_m, $0 \leq |x| \leq |F_m|-1$ and $3 \leq m \leq n-4$, form quasiperiodic squares of F_n.

- Factors of form xF_my, where $|x| > 0$ and $|y| > 0$ are given by Lemmas 17, 18, 19 and 20.

□

The following theorem counts the number of distinct overlapping factors in a Fibonacci word F_n, denoted by $OF(F_n)$, and shows that its limit over the square of its length tends to a constant number.

Theorem 22 $\lim_{n\to+\infty} \frac{OF(F_n)}{|F_n|^2} = 0.0597933994\ldots$

Proof. By summing the number of overlapping factors given by Lemma 21 and considering only the quadratic terms (as Lemma 9 shows that the sum of linear terms gives a linear term) we get:

$$\begin{aligned}\lim_{n\to+\infty} \frac{\mathrm{OF}(F_n)}{|F_n|^2} = \lim_{n\to+\infty} \frac{1}{|F_n|^2}(f_{n-4}f_{n-5} + \frac{1}{2}f_{n-4}^2 + 2\sum_{i=0}^{n-5} f_if_{i-1}\\ + f_{n-4}f_{n-5} + \frac{1}{2}f_{n-5}^2 + \frac{1}{2}\sum_{i=0}^{n-6} f_i^2)\end{aligned}$$

where the first two terms come from the case of $m = n-3$ in Lemma 21, the next three terms come from Lemmas 17 and 20 and the final term comes from Lemma 18. Therefore:

$$\begin{aligned}\lim_{n\to+\infty} \frac{\mathrm{OF}(F_n)}{|F_n|^2} &= \lim_{n\to+\infty} \frac{1}{f_n^2}(2\sum_{i=0}^{n-4} f_if_{i-1} + \frac{1}{2}\sum_{i=0}^{n-4} f_i^2)\\ &= \lim_{n\to+\infty} \frac{1}{f_n^2}(2f_{n-4}^2 + \frac{1}{2}f_{n-4}f_{n-3})\\ &\text{(by applying Lemmas 10, 12 and 13)}\\ &= \frac{2}{\phi^8} + \frac{1}{2\phi^7} \text{ (by applying Lemma 11)}\\ &= 0.0597933994\ldots\end{aligned}$$

□

6 Bounds on the maximum number of distinct overlapping factors in a word

In this section we investigate the maximum number of overlapping factors in a word. We show a simple upper bound by considering basic combinatorial properties. Fibonacci words give a first lower bound via the analysis given in the previous section. Then we provide a more complex example, giving a better lower bound.

$$\begin{aligned}
y_0 &= a_0\\
y_1 &= \underbrace{a_0}_{y_0}\underbrace{a_0}_{y_0}\underbrace{a_0}_{y_0}\underbrace{a_0}_{y_0}a_1\\
y_2 &= \underbrace{a_0a_0a_0a_0a_1}_{y_1}\underbrace{a_0a_0a_0a_0a_1}_{y_1}\underbrace{a_0a_0a_0a_0a_1}_{y_1}\underbrace{a_0a_0a_0a_0a_1}_{y_1}\ \underbrace{a_0a_0a_0a_0}_{y_1[1\mathinner{..}|y_1|-1]}\ a_2\\
&\vdots
\end{aligned}$$

Figure 4: The family of words described in Theorem 24

Theorem 23 *The number of distinct overlapping factors in a word y, denoted by $OF(y)$ is bounded above by $\frac{n^2-n}{4}$, where $|y| = n$.*

Proof. The number of factors of y of length at least 2 is $\binom{n}{2}$. When a factor overlaps with another means it appears at least twice in y. Therefore, $OF(y) \leq \frac{1}{2}\binom{n}{2} = \frac{n^2-n}{4}$ □

Fibonacci words give a first lower bound via the analysis given in the previous section. It is easy to observe that periodicity causes the appearance of many overlapping factors. The following example makes use of that observation to give a better lower bound on $OF(y)$. Using simple calculus one can verify that a period of $\frac{|y|}{5}$ gives maximum number of overlapping factors. We then allow periods to have similar structure to get a higher lower bound as shown in the following theorem.

Theorem 24 *The maximum number of distinct overlapping factors in a word y over the square of its length $|y| = n$, i.e. $\frac{OF(y)}{n^2}$, is at least $\frac{5}{48} = 0.1041666667\ldots$.*

Proof. Consider the following family of words (see also Figure 4):

$$y_i = y_{i-1}^4 y_{i-1}[1\mathinner{..}|y_{i-1}|-1]a_i \quad \text{with } y_0 = a_0$$

Then:

$$\lim_{i\to+\infty}\frac{\mathrm{OF}(y_i)}{|y_i|^2} = \lim_{i\to+\infty}\frac{1}{n^2}\Big(\sum_{j=1}^{i}\text{number of overlapping factors with length } \ell, \text{ where } |y_{j-1}| \leq \ell < |y_j|\Big)$$

In order to search for overlapping factors with length ℓ, where $|y_{j-1}| \leq \ell < |y_j|$, it is enough to restrict to the word y_j (as we can't include $y_j[|y_j|]$ in such a factor). Then counting such factor from the second occurrence of y_{j-1} in y_j we get $3|y_{j-1}|-1$ factors starting from position $y_j[|y_{j-1}|+1]$, $3|y_{j-1}|-2$ factors starting from position

$y_j[|y_{j-1}|+2]$, … and $2|y_{j-1}|$ factors starting from position $y_j[2|y_{j-1}|-1]$. Overall:

$$\begin{aligned}\lim_{i\to+\infty}\frac{\mathrm{OF}(y_i)}{|y_i|^2} &= \lim_{i\to+\infty}\frac{1}{n^2}\sum_{j=1}^{i}(\sum_{k=2|y_{j-1}|}^{3|y_{j-1}|-1} k)\\ &= \lim_{i\to+\infty}\frac{1}{n^2}\sum_{j=1}^{i}(\sum_{k=1}^{3|y_{j-1}|-1} k - \sum_{k=1}^{2|y_{j-1}|-1} k)\\ &= \lim_{i\to+\infty}\frac{1}{n^2}\sum_{j=1}^{i}\frac{5|y_{j-1}|^2-|y_{j-1}|}{2}\\ &= \lim_{i\to+\infty}\frac{1}{2n^2}(\sum_{j=1}^{i}5(\frac{n}{5^{i-j+1}})^2-\sum_{j=1}^{i}\frac{n}{5^{i-j+1}})\end{aligned}$$

$$= \lim_{i\to+\infty}\frac{5}{2n^2}\sum_{j=1}^{i}(\frac{n}{5^{i-j+1}})^2 = \lim_{i\to+\infty}\frac{5}{2}\sum_{j=1}^{i}\frac{1}{25^{i-j+1}}$$

$$= \lim_{i\to+\infty}\frac{5}{2}\sum_{j=1}^{i}\frac{1}{25^j} = \frac{5}{2}\frac{1/25}{24/25} = \frac{5}{48} = 0.1041666667\ldots$$

□

7 Conclusion and Future Work

The concept of quasiperiodicity is a generalization of the notion of periodicity, and was defined by Apostolico and Ehrenfeucht in [3]. In a periodic repetition the occurrences of the single periods do not overlap, while the quasiperiods of a quasiperiodic string may overlap. In this paper we investigated the overlapping factors of a word as a means to provide more insight in quasiperiodic structures of words. First, we proposed a linear time algorithm for the identification of all overlapping factors of a word. We then investigated the appearance of overlapping factors in Fibonacci words and bounded the maximum number of distinct overlapping factors in a word between $\frac{5}{48}n^2$ and $\frac{1}{4}n^2$. Future research might concentrate on deriving better such bounds.

References

[1] J. Allouche and J. Shallit. *Automatic sequences: theory, applications, generalizations*. Cambridge University Press, 2003.

[2] A. Apostolico and D. Breslauer. Of periods, quasiperiods, repetitions and covers. In J. Mycielski, G. Rozenberg, and A. Salomaa, editors, *Structures in Logic and Computer Science*, volume 1261 of *Lecture Notes in Computer Science*, pages 236–248. Springer, 1997.

[3] A. Apostolico and A. Ehrenfeucht. Efficient detection of quasiperiodicities in strings. *Theor. Comput. Sci.*, 119(2):247–265, 1993.

[4] J. Berstel. Fibonacci words-a survey. *The book of L*, pages 13–27, 1986.

[5] D. Breslauer. An on-line string superprimitivity test. *Inf. Process. Lett.*, 44(6):345–347, 1992.

[6] G. Brodal and C. Pedersen. Finding maximal quasiperiodicities in strings. In *Combinatorial Pattern Matching*, pages 397–411. Springer, 2000.

[7] M. Christodoulakis, C. Iliopoulos, K. Park, and J. Sim. Approximate seeds of strings. *Journal of Automata, Languages and Combinatorics*, 10(5/6):609–626, 2005.

[8] M. Christou, M. Crochemore, O. Guth, C. S. Iliopoulos, and S. Pissis. On the right-seed array of a string. *Computing and Combinatorics*, pages 492–502, 2011.

[9] M. Christou, M. Crochemore, and C. S. Iliopoulos. Quasiperiodicities in Fibonacci strings. In *Local Proceedings of International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, 2012.

[10] M. Christou, M. Crochemore, C. S. Iliopoulos, M. Kubica, S. Pissis, J. Radoszewski, W. Rytter, B. Szreder, and T. Waleń. Efficient seeds computation revisited. In *Combinatorial Pattern Matching*, pages 350–363. Springer, 2011.

[11] L. J. Cummings, D. Moore, and J. Karhumäki. Borders of Fibonacci strings. *Journal of Combinatorial Mathematics and Combinatorial Computing*, 20:81–88, 1996.

[12] R. Dunlap. *The golden ratio and Fibonacci numbers.* World Scientific Pub Co Inc, 1997.

[13] R. Groult and G. Richomme. Optimality of some algorithms to detect quasiperiodicities. *Theoretical Computer Science*, 411(34-36):3110–3122, 2010.

[14] V. Halava, T. Harju, T. Kärki, and P. Séébold. Overlap-freeness in infinite partial words. *Theoretical Computer Science*, 410(8-10):943–948, 2009.

[15] C. S. Iliopoulos, D. Moore, and W. F. Smyth. A characterization of the squares in a Fibonacci string. *Theoretical Computer Science*, 172(1-2):281–291, 1997.

[16] C. S. Iliopoulos, D. Moore, and W. F. Smyth. The covers of a circular Fibonacci string. *Journal of Combinatorial Mathematics and Combinatorial Computing*, 26:227–236, 1998.

[17] C. S. Iliopoulos, D. W. G. Moore, and K. Park. Covering a string. *Algorithmica*, 16:289–297, Sept. 1996.

[18] C. S. Iliopoulos and L. Mouchard. An $O(n \log n)$ algorithm for computing all maximal quasiperiodicities in strings. In C. S. Calude and M. J. Dinneen, editors, *Proceedings of DMTCS'99 and CATS'99*, pages 13–25, Singapore, 1999. Springer-Verlag.

[19] C. S. Iliopoulos and K. Park. A work-time optimal algorithm for computing all string covers. *Theoretical Computer Science*, 164(1-2):299–310, 1996.

[20] T. Kociumaka, M. Kubica, J. Radoszewski, W. Rytter, and T. Waleń. A linear time algorithm for seeds computation. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 1095–1112. SIAM, 2012.

[21] R. Kolpakov, G. Kucherov, et al. Finding maximal repetitions in a word in linear time. In *Symposium on Foundations of Computer Science-FOCS*, volume 99, pages 596–604, 1999.

[22] Y. Li and W. F. Smyth. Computing the cover array in linear time. *Algorithmica*, 32(1):95–106, 2002.

[23] M. Lothaire, editor. *Algebraic Combinatorics on Words.* Cambridge University Press, 2002.

[24] M. Lothaire, editor. *Applied Combinatorics on Words.* Cambridge University Press, 2005.

[25] D. Moore and W. F. Smyth. An optimal algorithm to compute all the covers of a string. *Information Processing Letters*, 50(5):239–246, 1994.

[26] P. Séébold. Overlap-free sequences. *Automata on Infinite Words*, pages 207–215, 1985.

[27] B. Smyth. *Computing patterns in strings.* ACM Press Bks. Pearson/Addison-Wesley, 2003.

[28] A. Thue. Über unendliche zeichenreihen. *Norske Vid. Skrifter I Mat.-Nat. Kl.*, 1:1–22, 1906.

[29] A. Thue. Über die gegenseitige lage gleicher teile gewisser zeichenreihen. *Norske Vid. Skrifter I Mat.-Nat. Kl.*, 7:1–67, 1912.

(Received 5 Apr 2012; revised 24 Feb 2013, 21 Apr 2013)