Usefulness of Directed Acyclic Subword Graphs in Problems Related to Standard Sturmian Words

Paweł Baturo[1], Marcin Piątkowski[1], and Wojciech Rytter[2,1⋆]

[1] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University
[2] Institute of Informatics, Warsaw University, Warsaw, Poland

Abstract. The class of finite Sturmian words consists of words having particularly simple compressed representation, which is a generalization of the Fibonacci recurrence for Fibonacci words. The subword graphs of these words (especially their compacted versions) have a very special regular structure. The regularity of their structure has been discovered in the context of the counting property of graphs. In this paper we investigate the structure of these subword graphs in more detail than in the previous papers. As an application we show how several syntactical properties of Sturmian words follow their graph properties. Alternative graph-based proofs of several known facts are presented. Also the neat structure of subword graphs of Sturmian words leads to algorithms computing several parameters (e.g. number of subwords, critical factorization point, short description of lexicographically maximal suffix, the structure of occurrences of subwords of a fixed length, right special factors) of standard Sturmian words in linear time with respect to the length n of the compressed representation: the directive sequence (though the words themselves can be of exponential size with respect to n). Some of the computed parameters can be of exponential size, however they have linear size grammar-based representation. This gives more examples of fast computations for highly compressed words.

1 Introduction

The *standard Sturmian words* (*standard words*, in short) are generalization of Fibonacci words and have a very simple *grammar-based* representation which has some algorithmic consequences.

Let $\mathcal{S}$ denote the set of all standard Sturmian words. These words are described by recurrences (or grammar-based representation) corresponding to so called *directive sequences*: integer sequences

$$\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n),$$

where $\gamma_0 \geq 0$, $\gamma_i > 0$ for $0 < i \leq n$. The word x_{n+1} corresponding to γ, denoted by $\mathrm{Word}(\gamma)$, is defined by recurrences:

$$x_{-1} = b, \quad x_0 = a, \quad \forall_{0 \leq i < n}\ x_{i+1} = x_i^{\gamma_i} x_{i-1} \tag{1}$$

Fibonacci words are standard Sturmian words given by directive sequences of the form

$$\gamma = (1, 1, \ldots, 1).$$

We consider here standard words starting with the letter a, hence assume $\gamma_0 > 0$. The case $\gamma_0 = 0$ can be considered similarly.

⋆ Supported by grant N206 004 32/0806 of the Polish Ministry of Science and Higher Education

Pawe l Baturo, Marcin Piątkowski, Wojciech Rytter: Usefulness of Directed Acyclic Subword Graphs in Problems Related to Standard Sturmian Words, pp. 193–207.
Proceedings of PSC 2008, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04145-1

For even $n > 0$ a standard word x_n has suffix ba, and for odd $n > 0$ it has suffix ab. The number $N = |x_{n+1}|$ is the (real) size, while $n + 1$ can be thought as the compressed size.

Example 1.
Consider directive sequence $\gamma = (1, 2, 1, 3, 1)$. We have:

$$\text{Word}(1, 2, 1, 3, 1) = abababaababababaababababaababaab$$

$$x_{-1} = b, \quad x_0 = a, \quad x_1 = x_0^1 x_{-1} = ab, \quad x_2 = x_1^2 x_0 = ababa,$$
$$x_3 = x_2^1 x_1 = abababaab, \quad x_4 = x_3^3 x_2 = abababaababababaababababa,$$
$$x_5 = x_4^1 x_3 = abababaababababaababababaababaab$$

Some of the outputs of our algorithms will be given in the grammar-compressed form which consists in giving a context-free grammar G generating a single word x. The size of G is the total length of all productions of G.

In particular each directive sequence of a standard Sturmian word corresponds to such a compression – the sequence of recurrences corresponding to the directive sequence. In this case the size of the grammar is proportional to the length of the directive sequence.

For some lexicographic properties and structure of repetitions of standard Sturmian words see [3] and [2].

2 The structure of subword graphs of standard Sturmian words

Let $Subwords(x)$ be the set of all nonempty subwords of x. We distinguish some subwords as special ones:

- A **special prefix** of x is a prefix z of x such that $za, zb \in Subwords(x)$.
- A **basic prefix** of x is a proper nonempty prefix of the type $x_k^j x_{k-1}$, where $0 \leq k \leq n$ and $0 \leq j \leq \gamma_k$.
- A **basic subword** of x is a reverse of x_k, for some k. Denote $y_k = Reverse(x_k)$.

Denote by $BP(x)$ the set of basic prefixes of x and by $SP(x)$ the set of special prefixes of x. Denote by $\hat{x}$ the prefix of x of size 2, assuming $|x| \geq 2$. Assume that $\hat{y}_0 = ab$.

Lemma 1.
Assume $x_{-1}, x_0, \ldots, x_{n+1}$ is the sequence of standard Sturmian words given by $(\gamma_0, \gamma_1, \ldots, \gamma_n)$.
(a) *For $i \geq 1$ we can represent standard word x_i as*

$$x_i = y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{i-2}^{\gamma_{i-2}} y_{i-1}^{\gamma_{i-1}-1} \hat{y}_{i-1},$$

(b) *Each special prefix of x_n has the form*

$$y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_i^j,$$

where $0 \leq j \leq \gamma_i$ for $i < n - 1$ and $0 \leq j \leq \gamma_i - 1$ for $i = n - 1$,
(c) *Each special prefix results by cutting off two last symbols from a basic prefix.*

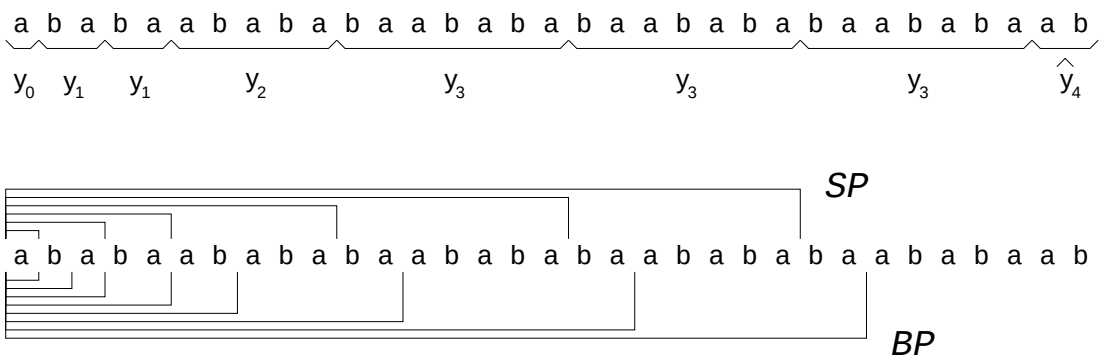

Figure 1. The structure of basic prefixes (BP), special prefixes (SP) and basic subwords of $Word(1, 2, 1, 3, 1)$.

Proof.
Point (a)
Notice that $\hat{y}_i = \hat{y}_{i+2}$ and $y_{i+1} = y_{i-1}y_i^{\gamma_i}$ for $i \geq 0$.

First we show by induction that

$$y_i = \hat{y}_i y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{i-1}^{\gamma_{i-1}-1}. \tag{2}$$

For $i = 1$ we have

$$y_1 = ba^{\gamma_0} = \hat{y}_1 y_0^{\gamma_0 - 1}$$

Assume that for $i \leq n$ the equation (2) is true. We have

$$\begin{aligned} y_{n+1} &= y_{n-1} \cdot y_n^{\gamma_n} \\ &= \left(\hat{y}_{n-1} y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{n-2}^{\gamma_{n-2}-1}\right) \cdot \left(y_{n-2} y_{n-1}^{\gamma_{n-1}} y_n^{\gamma_n - 1}\right) \\ &= \hat{y}_{n+1} y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_n^{\gamma_n - 1} \end{aligned}$$

Now we can prove equation from the point (a) using induction. For $i = 1$ we have:

$$x_1 = x_0^{\gamma_0} x_{-1} = y_0^{\gamma_0 - 1} \hat{y}_0$$

Assume that for $i \leq n$ equation from the point (a) is true. We have

$$\begin{aligned} x_{n+1} &= x_n^{\gamma_n} x_{n-1} \\ &= \left(y_0^{\gamma_0} \cdots y_{n-2}^{\gamma_{n-2}} y_{n-1}^{\gamma_{n-1}-1} \hat{y}_{n-1}\right)^{\gamma_n} \cdot y_0^{\gamma_0} \cdots y_{n-2}^{\gamma_{n-2}-1} \hat{y}_{n-2} \\ &\overset{\text{due to (2)}}{=} y_0^{\gamma_0} \cdots y_{n-1}^{\gamma_{n-1}} y_n^{\gamma_n - 1} \hat{y}_n \end{aligned}$$

Point (b).
Let $\overline{w}$ denotes here a word w with removed last two letters and assume that w contains at least two letters.

From point (a) we know that

$$z = y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_i^j$$

is a prefix of standard word x_n generated by directive sequence $(\gamma_0, \gamma_1, \ldots, \gamma_n)$, where $0 \leq j \leq \gamma_i$ for $i < n - 1$ and $0 \leq j \leq \gamma_i - 1$ for $i = n - 1$. We can also deduce, that

prefix $\overline{x_n}$ is a palindrome (see [4] for proof that every standard word x a word $\overline{x}$ is a palindrome). Hence, if z is special prefix of standard word x, then z is also suffix of $\overline{x}$.

First assume that $i < n-1$ and i is odd, the case for even i is similar.

If $0 \leq j < \gamma_i$, then z is prefix of x_{i+2} and zb is also prefix of x_{i+2} (first letter of y_i is b). Suffix of x_{i+2} is ab, hence za, as a suffix of $\overline{x_{i+2}}$, is also subword of x_{i+2}.

If $j = \gamma_i$, then z is prefix of x_{i+3} and za is also prefix of x_{i+3} (first letter of y_{i+1} is a). Suffix of x_{i+3} is is ba, hence zb, as a suffix of $\overline{x_{i+3}}$, is also subword of x_{i+3}.

Now assume that $i = n-1$. For $0 \leq j < \gamma_{n-1}$ proof is similar to the case $i < n-1$. It is obvious, due to the deduction above, that for $i = n-1$, j must be less than γ_{n-1}.

Point (c)
Notice that $\hat{y}_i = \hat{y}_{i+2}$ and $y_{i+1} = y_{i-1}y_i^{\gamma_i}$ for $i \geq 0$.

From (a) for basic prefix $x_k^j x_{k-1}$ we have:

$$\begin{aligned} x_k^j x_{k-1} &= \left(y_0^{\gamma_0} \cdots y_{k-2}^{\gamma_{k-2}} y_{k-1}^{\gamma_{k-1}-1} \hat{y}_{k-1}\right)^j \cdot y_0^{\gamma_0} \cdots y_{k-3}^{\gamma_{k-3}} y_{k-2}^{\gamma_{k-2}-1} \hat{y}_{k-2} \\ &\overset{\text{due to (2)}}{=} y_0^{\gamma_0} \cdots y_{k-1}^{\gamma_{k-1}} y_k^{j-1} \hat{y}_k \end{aligned}$$

From (b) we have that basic prefix $x_k^j x_{k-1}$ with last two letters removed ($\hat{y}_k$) is special prefix.

Example 2.
For $\text{Word}(1,2,1,3,1) = ababaababababaababababaababababaab$ we have:

$$\begin{aligned} &BP = \{x_0,\ x_1,\ x_1x_0,\ x_2,\ x_3,\ x_3x_2,\ x_3^2x_2,\ x_4\} \\ &SP = \{y_0,\ y_0y_1,\ y_0y_1^2,\ y_0y_1^2y_2,\ y_0y_1^2y_2y_3,\ y_0y_1^2y_2y_3^2\} \\ &y_0 = a \qquad y_1 = ba \qquad y_2 = ababa \qquad y_3 = baababa \\ &\text{Word}(1,2,1,3,1) = a\ \ ba\ \ ba\ \ ababa\ \ baababa\ \ baababa\ \ baababa\ \ ab \\ &\qquad\qquad\qquad = y_0\ y_1^2\ y_2\ y_3^3\ \hat{y}_4 \end{aligned}$$

The subword graph is a classical data structure representing all subwords of a given word in a succinct manner. More precisely: the *Directed Acyclic Word Graph* (dawg in short) of the word w is the minimal deterministic automaton (not necessarily complete) that accepts all suffixes of w. We refer the reader to [6] for the complete definition and more information of subword graphs.

The compacted subword graph (cdawg, in short) results from the subword graph by removing all nodes of out-degree one (except the source node and the terminal nodes) and replacing each chain by a single edge with the label representing the path label of this chain. Internal nodes of dawg of out-degree greater than one, which are copied to cdawg, are called fork nodes. In case of standard words the subword graph can be considerably compressed.

The regularity of the structure of compacted subword graphs has been discovered in [8]. The following theorem follows from the results of [8], Lemma 1 and our terminology.

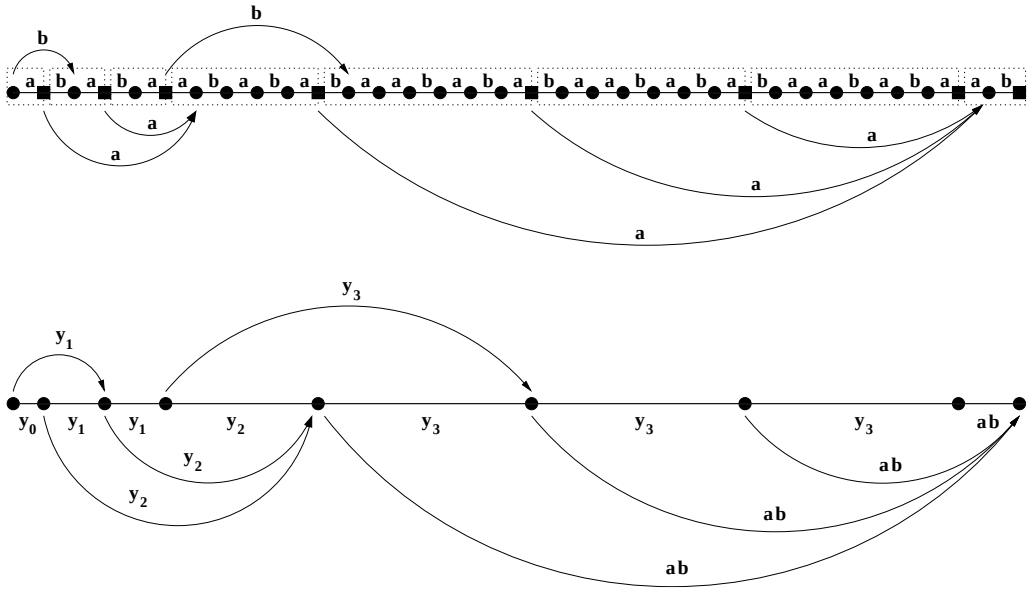

Figure 2. The structure of the subword graph (dawg) of $Word(1,2,1,3,1)$ and its compacted version (cdawg)

Theorem 2.
Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *be a standard Sturmian word.*
(1) *The labels of edges in compacted subword graph of* w *are basic subwords of* w.
(2) *The compacted subword graph of* w *has the structure illustrated on Figure 3.*

3 The number of subwords

It is known that the number of distinct subwords in the n-th Fibonacci word is

$$Subwords(Fib_{n+1}) \;=\; |Fib_n| \cdot |Fib_{n-1}| + 2 \cdot |Fib_n| - 1$$

Surprisingly essentially the same formula works generally for Sturmian words.

Theorem 3. *Let* $\gamma_n = 1$, *and* $x_{n+1} \;=\; \mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n)$, *then*

$$\big|Subwords(x_{n+1})\big| \;=\; |x_n| \cdot |x_{n-1}| + 2 \cdot |x_n| - 1$$

Proof.
Denote by v_0 the source node of the compacted subword graph for x_{n+1}. Let $t_k = |x_k|$. Define the multiplicity $mult(v)$ of a vertex v as the number of paths $v_0 \xrightarrow{*} v$, and the weights of edges as lengths of corresponding label-strings of these edges in the compacted subword graphs. Let $edges(v)$ be the sum of all weight edges outgoing from v.

Claim. Let $w \;=\; \mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n)$. Then

$$\big|Subwords(w)\big| \;=\; \sum_{v \in G} mult(v) \cdot edges(v) \tag{3}$$

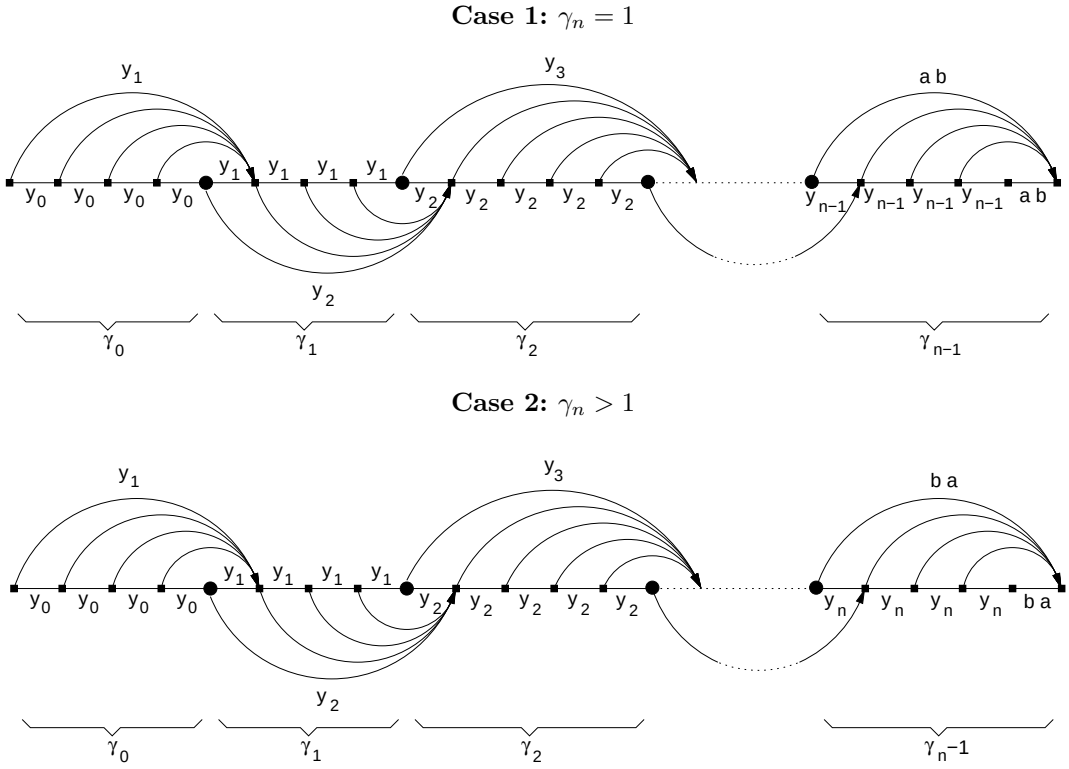

Figure 3. Compacted subwords graphs for words: Word$(\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n)$ and Word$(\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n - 1, 1)$ are isomorphic (in the sense of graph structure).

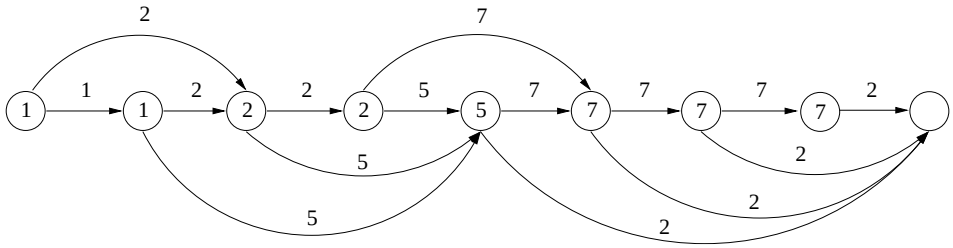

Figure 4. The structure of edge-lengths and multiplicities of nodes in the compacted subword graph of $Word(1, 2, 1, 3, 1)$. According to the Theorem 3 (and to the graph above) there are $|x_4| \cdot |x_3| + 2 \cdot |x_4| - 1 \;=\; 26 \cdot 7 + 2 \cdot 26 - 1 \;=\; 233$ subwords in our example word.

See Figure 4 for edge-lengths and node-multiplicities structure in the cdawg of example word.
We partition the set of edges into chunks, the first chunk consists of the first γ_0 consecutive vertices starting from the v_0, the second chunk contains the next γ_1 vertices, etc. The last chunk slightly differs.

The contribution of k-th internal chunk in the sum in equation (3) is

$$\big(t_{k-1} + (\gamma_k - 1)t_k\big) \cdot (t_k + t_{k+1}) \;=\; t_{k+1}^2 - t_k^2,$$

where $t_{-1} = 1$ (see Figure 5 for details).

The contribution of the last chunk is (see Figure 6)

$$(t_{n-1}+2)(t_n - t_{n-1}) + 2t_{n-1}.$$

Altogether we have

$$\sum_{k=0}^{n-2} \left(t_{k+1}^2 - t_k^2\right) + (t_{n-1}+2)(t_n - t_{n-1}) + 2t_{n-1} \;=\; t_n \cdot t_{n-1} + 2 \cdot t_n - 1$$

This completes the proof, since by definition $|x_k| = t_k$.

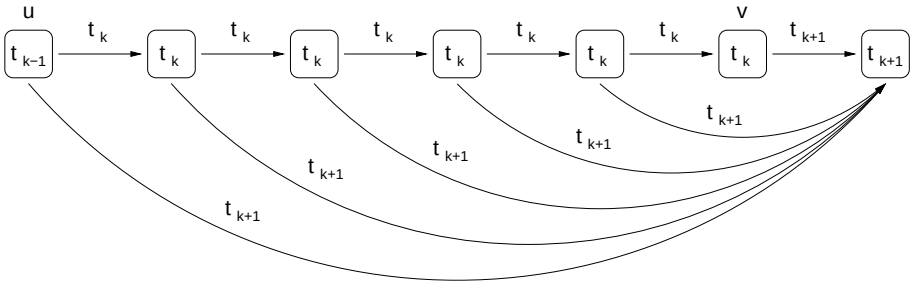

Figure 5. The k-th internal chunk G_k of the subword graph, consists of γ_k nodes from u to v (excluding u), and their outgoing edges. The multiplicity (number of path leading from v_0) of each node is written within the box corresponding to the node. The weight of the edges are the lengths of corresponding words in the cdawg.

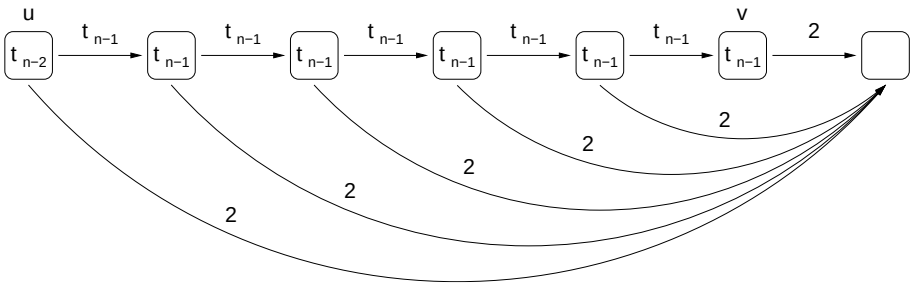

Figure 6. The final chunk G_{n-1} of the subword graph, consists of γ_{n-1} nodes from u to v, and their outgoing edges.

The case $\gamma_n > 1$ reduces to the previous case.

Theorem 4. *Let $\gamma_n > 1$. Then:*

$$\left|Subwords\big(\mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n)\big)\right| \;=\; \left|Subwords\big(\mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n - 1, 1)\big)\right|.$$

Proof.
Compacted subword graphs of $\mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n)$ and $\mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n - 1, 1)$ are isomorphic in the sense of graph structure (see Figure 3 for details). Hence we can use the result of Theorem 3 to compute $\left|Subwords\big(\mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n)\big)\right|$.

4 The structure of occurrences of subwords

In this section we are interested in the structure of first occurrences of the subwords of a given length. One type of these subwords is particularly interesting – a right special factors.

A **right special factor** of the word x is any word w such that both wa, wb are subwords of x. For each $k > 0$ there is at most one right special factor of length k of a given standard word. For standard Sturmian word x every right special factor is either special prefix or suffix of some special prefix.

Theorem 5. *Let $w = \mathrm{Word}(\gamma)$ be a standard Sturmian word. Then:*
(1) *For a given $k > 0$ the right special factor of w of length k has grammar-representation of size $O(|\gamma|)$.*
(2) *The compressed representation of the right special factor of w of length k can be computed in $O(|\gamma|)$ time.*

Proof.
Define length of the path in cdawg of w as number of edges in it and value of the path as word created by concatenation of the labels of edges in it.

Let v be an internal node in compacted subwords graph of w and z_π be a value of path π in this graph leading from root to v. It is clear that z_π is a subword of w.

Every internal node in compacted subword graph is a fork node, hence v has two outgoing edges: one with label starting with letter a and the second with label starting with letter b. This follows that $z_\pi \cdot a$ and $z_\pi \cdot b$ are also subwords of w and therefore z_π is a right special factor of w.

Observe that value of every path from root to v in cdawg of w is suffix of the value of the longest path from root to v. Moreover value of the longest path from root to v is a prefix of w, hence it is a special prefix of w. This implies that every right special factor of w is suffix of some special prefix of w.

Every right special factor of w is concatenation of some basic subwords of w. It follows easily from Lemma 1 that every right special factor of w has grammar-representation of size $O(|\gamma|)$ which can be computed in time linear to the length of directive sequence γ.

Example 3.
Let $w = \mathrm{Word}(1, 2, 1, 3, 1) = ababaababababaababababaababababaab$.
Recall that:

$$y_0 = a \qquad y_1 = ba \qquad y_2 = ababa \qquad y_3 = baababa$$

Right special factors of w with their lengths are (special prefixes are bold):

							11	$y_1^2y_3$	18	$y_1^2y_3^2$
							12	y_2y_3	19	$y_2y_3^2$
					6	y_0y_2	13	$y_0y_2y_3$	20	$y_0y_2y_3^2$
					7	y_1y_2	14	$y_1y_2y_3$	21	$y_1y_2y_3^2$
					8	$y_0y_1y_2$	15	$y_0y_1y_2y_3$	22	$y_0y_1y_2y_3^2$
		2	y_1	4	y_1^2	9	$y_1^2y_2$	16	$y_1^2y_2y_3$	23	$y_1^2y_2y_3^2$
1	$\mathbf{y_0}$	3	$\mathbf{y_0y_1}$	5	$\mathbf{y_0y_1^2}$	10	$\mathbf{y_0y_1^2y_2}$	17	$\mathbf{y_0y_1^2y_2y_3}$	24	$\mathbf{y_0y_1^2y_2y_3^2}$

See Figure 2 for the structure of cdawg of the word w.

For a set X of integers and an integer k define

$$X \oplus k \;=\; \{\; x + k \;:\; x \in X \;\}$$

Let $occ(u, w)$ be the set of first positions of occurrences of u in w, we define also the set of final positions of occurrences of a word u :

$$fin(u, w) = occ(u, w) \oplus |u| \quad \text{and} \quad \textit{first-fin}(u, w) = \mathbf{min}\,\big(fin(u, w)\big).$$

For $k \geq$ we investigate also the structure of the set

$$FIN(k, w) \;=\; \big\{\textit{first-fin}(u, w) \;:\; u \;\text{ is a subword of }\; w \;\text{ of size }\; k \;\big\}.$$

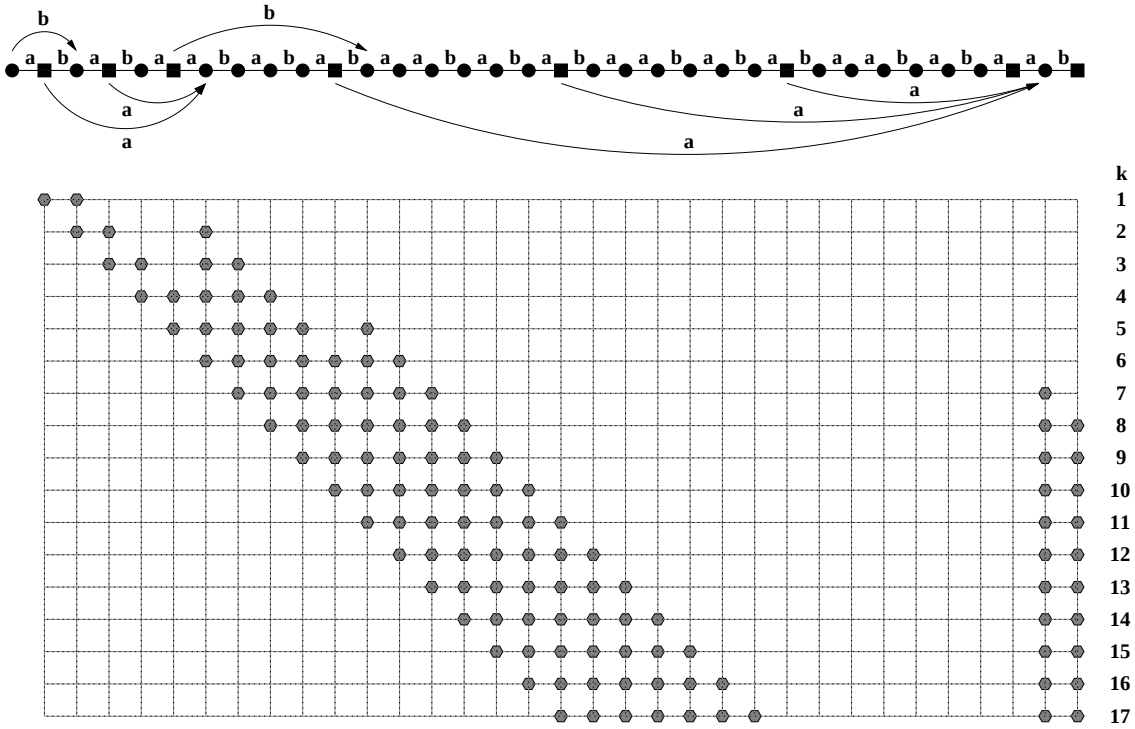

Figure 7. The subword graph of w and the structure of the sets $FIN(k, w)$ for $w =$ Word$(1, 2, 1, 3, 1)$.

Theorem 6. *Let $w =$ Word$(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word. Then:*
(1) *The set $FIN(k, w)$ consists of a single interval or of two disjoint intervals.*
(2) *For a given k we can compute the intervals representing $FIN(k, w)$ in linear time with respect to the size of the directive sequence.*

Proof.
The structure of the set $FIN(k, w)$ easily follows from the way how paths of length $k-1$ in dawg of w are extended into path of length k. Only fork nodes $i \in FIN(k-1, w)$ generate two elements of $FIN(k, w)$, each other node $i \in FIN(k-1, w)$ generates single element $i + 1$ in $FIN(k, w)$ (see Figure 7).

It is clear that the set $FIN(k+1, w)$ results from $FIN(k, w)$ by shifting each position by one to the right and adding an extra position for the fork node. Hence thesis follows from the structure of subword graphs of standard Sturmian words.

5 Relation of subword graphs to the dual Ostrovski numeration system

The dual Fibonacci numeration system has been introduced in [10], where its relation to the subword structure of Fibonacci words has been investigated. We extend these results to Sturmian words. In this case we have Ostrovski numeration system which is a generalization of Fibonacci system.

In (only) this section we consider infinite directive sequences.

For an infinite directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots)$ we introduce $[*]_\gamma$-numeration system: a version of Ostrowski's numeration system from [1] which is a generalization of the Fibonacci number system. Let us define the *base* sequence q as a sequence:

$$q \;=\; (q_0, q_1, \ldots) \;=\; \big(|x_0|, |x_1|, ...\big),$$

where x_i's are as in equation (1).

The base sequence can be defined without reference to words x_i as follows:

$$q_{-1} = q_0 = 1, \;\; q_{i+1} = q_i \cdot \gamma_i + q_{i-1} \text{ for } i \geq 0.$$

Example 4.
If $\gamma = (1, 2, 1, 2, \ldots)$, then the base sequence is:

$$q = (1, 2, 5, 7, 19, \ldots)$$

If $\gamma = (1, 2, 1, 1, 1, \ldots)$, then the base sequence is:

$$q = (1, 2, 5, 7, 12, 19, \ldots)$$

Define:

$$\mathrm{val}_\gamma(\alpha_0, \alpha_1, \ldots, \alpha_n) \;=\; \alpha_0 \cdot q_0 + \alpha_1 \cdot q_1 + \ldots + \alpha_n \cdot q_n$$

For $0 \leq i < |x_n|$ the representation of i in Ostrovski numeration system is defined as follows:

$$[i]_\gamma \;=\; (\alpha_0, \alpha_1, \ldots, \alpha_n),$$

where we require:

(1) $\mathrm{val}_\gamma(\alpha_0, \alpha_1, \ldots, \alpha_n) = i$
(2) $\forall_{0 \leq j < n}\ \alpha_j \leq \gamma_j$
(3) $\alpha_{j+1} = \gamma_{j+1} : \alpha_j = 0$

In other words in the representation of a number i we take at most γ_k numbers $|x_k|$, for each k, and if we take exactly γ_k numbers $|x_k|$ then we take zero numbers $|x_{k-1}|$.

Example 5.
Let $\gamma = (1, 2, 1, 3, 1, \ldots)$. Then

$$q = \big(|x_0|, |x_1|, ...\big) = (1, 2, 5, 7, 26, 33, \ldots)$$

We have $[29]_\gamma = (1, 1, 0, 0, 1)$, because

$$29 = 1 \cdot 1 + 1 \cdot 2 + 0 \cdot 5 + 0 \cdot 7 + 1 \cdot 26$$

We have $[58]_\gamma = (0, 2, 0, 3, 0, 1)$, because

$$58 \;=\; 0 \cdot 1 + 2 \cdot 2 + 0 \cdot 5 + 3 \cdot 7 + 0 \cdot 26 + 1 \cdot 33$$

For $0 \leq i < |x_n|$ we define representation of i in the dual Ostrovski numeration system as:

$$\hat{[i]}_\gamma = (\alpha_0, \alpha_1, \ldots, \alpha_n),$$

where:

(1) $\mathrm{val}_\gamma(\alpha_0, \alpha_1, \ldots, \alpha_n) = i$
(2) $\forall_{0 \leq j < n}\ \alpha_j \leq \gamma_j$
(3) $\big(\ \alpha_j < \gamma_j$ and $\exists\ (i > j)\ \alpha_i > 0\ \big)\ : \alpha_{j+1} > 0$

In other words in the representation of a number i in numeration system defined above we take at most γ_k numbers $|x_k|$, and if we take $\alpha_k < \gamma_k$ numbers $|x_k|$ and α_k is not the last one component of this representation then we must take at least one number $|x_{k+1}|$.

Example 6.
Let $\gamma = (1, 2, 1, 3, 1, \ldots)$. Then

$$q = \big(|x_0|, |x_1|, ...\big) = (1, 2, 5, 7, 26, 33, \ldots)$$

We have $\hat{[29]}_\gamma = (1, 1, 1, 3)$, because

$$29 = 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 5 + 3 \cdot 7$$

We have $\hat{[58]}_\gamma = (0, 2, 0, 3, 0, 1)$, because

$$58 \;=\; 0 \cdot 1 + 2 \cdot 2 + 0 \cdot 5 + 3 \cdot 7 + 0 \cdot 26 + 1 \cdot 33$$

Uniqueness of representation in Ostrovski numeration system was proved in [1]. Uniqueness of representation in dual Ostrovski numeration system was proved in [8].

Let $\mathcal{G}_\infty$ be the infinite compacted subword graph corresponding to a given directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots)$.

The following fact is an interpretation of the corresponding result in [8] in terms of the dual Ostrovski numeration system.

Theorem 7.

(1) *Let π be a path from the root to another node of $\mathcal{G}_\infty$. Let* $\mathrm{rep}(\pi) = (h_0, h_1, \ldots)$, *where h_i is the number of edges of weight q_i on the path π. Then* $\mathrm{rep}(\pi)$ *is the representation of the length $|\pi|$ of this path in the dual Ostrovski numeration system corresponding to the directive sequence of $\mathcal{G}_\infty$.*
(2) *For each $k > 1$ there is exactly one fork-path of length k in $\mathcal{G}_\infty$.*

Proof.
Point (1)
Let π be a path from root to some node v in $\mathcal{G}_\infty$ – infinite compacted subwords graph corresponding to directive sequence $(\gamma_0, \gamma_1, \gamma_2, \ldots)$, and let $\mathrm{rep}(\pi) = (h_0, h_1, \ldots)$ be

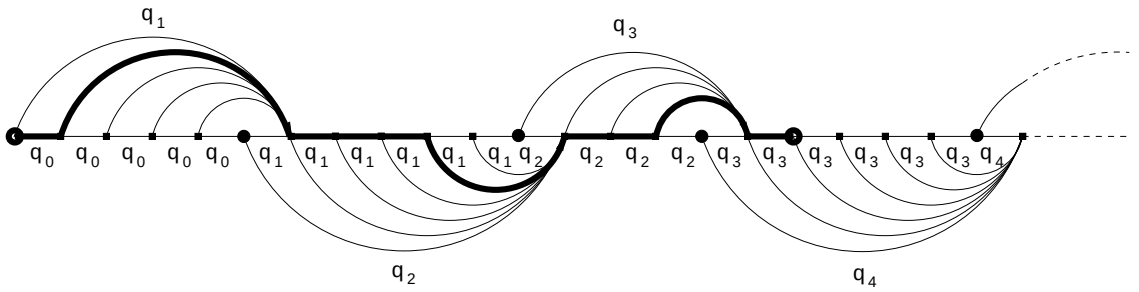

Figure 8. The illustration of the point (1) of Theorem 7. In this case representation of the length of the path π in dual Ostrovski numeration system is given by: $\mathrm{rep}(\pi) = (1, 4, 3, 2)$ and $|\pi| = 1 \cdot |q_0| + 4 \cdot |q_1| + 3 \cdot |q_2| + 2 \cdot |q_3|$.

defined as above. It is sufficient to check if requirements of definition of dual Ostrovski numeration system are satisfied.

Construction of π implies that

$$|\pi| = h_0 \cdot q_0 + h_1 \cdot q_1 + h_2 \cdot q_2 + \cdots$$

and $\forall_i\ 0 \leq h_i \leq \gamma_i$. Moreover from $\mathcal{G}_\infty$ structure (see Figure 8) it is obvious that if $h_i < \gamma_i$ (we have taken q_i less than γ_i times) and h_i is not the last non zero element in $\mathrm{rep}(\pi)$ then $h_{i+1} > 0$ (we must take at least one q_{i+1} to continue construction of π). This concludes the proof of point (1).

Point (2) follows directly from point (1) and uniqueness of representation in dual Ostrovski numeration system.

Ostrovski automata

For a directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ we define $SD(\gamma)$ as the set of representations $(i_0, i_1, \ldots, i_n)$ in the dual Ostrovski numeration system of all numbers not exceeding the number written as γ in this representation.

Remark.
Observe that for any symbol a the value of a^0 is an empty word.

Denote

$$L(\gamma) = \{a_0^{i_0} a_1^{i_1} \ldots a_n^{i_n} \ : \ (i_0, i_1, \ldots, i_n) \in SD(\gamma)\}$$

for alphabet $\Sigma = \{a_0, a_1, \ldots, a_n\}$.

The minimal deterministic finite automaton accepting language $L(\gamma)$ is called the Ostrovski automaton and denoted by $OA(\gamma)$.

Theorem 8. *The minimal Ostrovski automaton for γ, without the dead state, is isomorphic as a graph to the compact directed acyclic subword graph of* $\mathrm{Word}(\gamma)$.

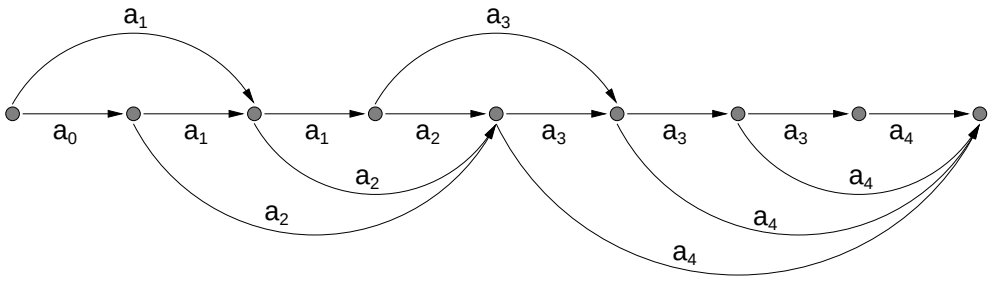

Figure 9. Minimal deterministic automaton (without dead state) $OA(1,2,1,3,1)$ accepting the set of strings $a_0^{i_0}a_1^{i_1}a_2^{i_2}a_3^{i_3}a_4^{i_4}$, where $(i_0, i_1, \ldots, i_4)$ is a representation in the dual Ostrovski numeration system of a natural number.

6 Critical factorization and maximal suffixes

The **minimal local period** in a word w at position k is a positive integer p such that $w[i-p] = w[i]$ for every $k \leq i < k+p$, where $w[i]$ and $w[i-p]$ are defined.

The **critical factorization point** in a word w is position k in w for which minimal local period at k equals the (global) minimal period of w. We refer the reader to [6] for the more detailed definition of the critical factorization point.

The following nontrivial fact was shown by Crochemore and Perrin [5].

Fact 1
The critical factorization point of w is given as the starting position of a lexicographically maximal suffix, maximized over all possible orders of the alphabet.

Example 7
Let $w = \text{Word}(1,2,1,3,1) = ababaababababaababababaababababaab$.
Minimal local periods of w are as follows:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	⋯⋯
	a	b	a	b	a	a	b	a	b	a	b	a	a	b	a	b	a	⋯⋯
$p(i)$	1	2	2	2	5	1	7	2	2	2	2	7	1	7	2	2	2	2 ⋯⋯

i	⋯⋯	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
	⋯⋯	b	a	a	b	a	b	a	b	a	a	b	a	b	a	a	b
$p(i)$	⋯⋯	2	7	1	7	2	2	2	4	**33**	1	5	2	2	5	1	3	1

where i denotes position in w and $p(i)$ – minimal local period at position i in w.
Hence critical factorization point is at position $i = 25$.

For a word w define $\pi_a(w)$ as a path in the dawg of w which starts in the root, ends in the sink, and in which we use the letter a whenever we have a choice (in every fork node). Similarly define $\pi_b(w)$. Path $\pi_a(w)$ ($\pi_b(w)$ respectively) can be also defined for cdawg of w: in every fork node we choose the edge with label starting with letter a (letter b respectively). Length of the path, denoted $|\pi|$, is is defined as length of the word given by π.

It is easily seen that lexicographically maximal suffix of w with respect to the letter ordering "$a < b$" is given by $\pi_b(w)$ and the lexicographically maximal suffix of w with respect to the letter ordering "$a > b$" is given by $\pi_a(w)$.

Lemma 9.
Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word and $\pi_a(w)$, $\pi_b(w)$ be defined as above. Then:

$$\pi_a(w) = y_0^{\gamma_0} y_2^{\gamma_2} \cdots y_{2k}^{\gamma_{2k}} \cdot \hat{y}_{n-1}$$
$$\pi_b(w) = y_1^{\gamma_1} y_3^{\gamma_3} \cdots y_{2l+1}^{\gamma_{2l+1}} \cdot \hat{y}_{n-1}$$

where $k = \lfloor \frac{n-1}{2} \rfloor$ and $l = \lfloor \frac{n-2}{2} \rfloor$.

Proof.
Recall that definition of basic subwords follows that y_i starts with letter a for even i and y_i starts with letter b for odd i.

We are constructing path $\pi_a(w)$ in cdawg of w by choosing edge with label starting with letter a whenever it is possible. From structure of cdawgs of standard Sturmian words (see Figure 3) we have that every fork node has two outgoing edges: one with label y_{2i} (starting with letter a) and second with label y_{2i+1} (starting with letter b).

To construct $\pi_a(w)$ we have to choose γ_0 times edge with label y_0, then γ_2 times edge with label y_2, and so on up to y_{2k}, where $k = \lfloor \frac{n-1}{2} \rfloor$. Finally, by Lemma 1, it suffices to add $\hat{y}_{n-1}$, the last two letters of w.

The same proof works for path $\pi_b(w)$.

Construction of paths $\pi_a(w)$ and $\pi_b(w)$ implies the following fact.

Theorem 10.
Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word. Then:
(1) *The critical factorization point of w is at position*

$$k = |w| - \boldsymbol{min}\left\{\ |\pi_a(w)|,\ |\pi_b(w)|\ \right\}$$

(2) *The critical factorization point of w can be computed in linear time with respect to the size of the directive sequence.*

Proof.
The proof is immediate by Fact 1 and recalling that $\pi_a(w)$ and $\pi_b(w)$ corresponds to lexicographically maximal suffixes of w with respect to letter orderings "$a > b$" and "$a < b$" respectively.

Example 8.
Let $w = \mathrm{Word}(1, 2, 1, 3, 1) = abababaababababaababababaababaab$.
See Figure 2 for its subword graph structure.
We have

$$\pi_a(w) = y_0 y_2 ab = a\ ababa\ ab$$
$$\pi_b(w) = y_1^2 y_3^3 ab = ba\ ba\ baababa\ baababa\ baababa\ ab$$

Hence the position

$$i = |w| - |y_0 y_2 ab| = 33 - 8 = 25$$

is the critical factorization point of w.

Similar computations were given in [7,9] for Fibonacci words. The paths $\pi_a(w)$ and $\pi_b(w)$ have regular structure, consequently the words represented by them are well compressible. This implies the following fact.

Theorem 11. *Let $w = \mathrm{Word}(\gamma)$ be a standard Sturmian word. Then:*
(1) *The lexicographically maximal suffix of w has grammar-based representation of size $O\big(|\gamma|\big)$.*
(2) *The compressed representation of the lexicographically maximal suffix of w can be computed in $O\big(|\gamma|\big)$ time.*

Proof.
The lexicographically maximal suffix of a standard Sturmian word w is given either by path $\pi_a(w)$ or by path $\pi_b(w)$ (depending on which letter ordering was chosen). The thesis follows directly from the structure of $\pi_a(w)$, $\pi_b(w)$ and the subword graph of w (see Lemma 9).

References

1. J. Allouche and J. Shallit: *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge University Press, 2003.
2. P. Baturo, M. Piątkowski, and W. Rytter: *The number of runs in Sturmian words*, CIAA, 2008.
3. P. Baturo and W. Rytter: *Occurrence and lexicographic properties of standard Sturmian words*, LATA, 2007.
4. J. Berstel and P. Seebold: *Sturmian words, in: M. Lothaire, Algebraic combinatorics on words, (Chapter 2), vol. 90 of Encyclopedia of Mathematics and its Applications*, Cambridge University Press, 2002.
5. M. Crochemore and D. Perrin: *Two-Way String Matching*, J. ACM 38(3): 651-675, 1991.
6. M. Crochemore and W. Rytter: *Jewels of stringology: text algorithms*, World Scientific, 2003.
7. T. Harju and D. Nowotka: *On the density of critical factorizations*, ITA 36(3): 315-327, 2002.
8. F. Mignosi, J. Shallit, and I. Venturini: *Sturmian Graphs and a Conjecture of Moser*, Lecture Notes in Computer Science 3340, 175-187, 2004.
9. W. Rytter: *The structure of subword graphs and suffix trees of Fibonacci words*, Theoretical Computer Science Volume 363, Issue 2, 211 - 223, 2006.
10. W. Rytter: *The number of runs in a string*, Information and Computation Volume 205, Issue 9, 1459-1469, 2007.