The number of runs in Sturmian words

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

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

Abstract. Denote by $\mathcal{S}$ the class of *standard Sturmian* words. It is a class of highly compressible words extensively studied in combinatorics of words, including the well known Fibonacci words. The suffix automata for these words have a very particular structure. This implies a simple characterization (described in the paper by the Structural Lemma) of the periods of runs (maximal repetitions) in Sturmian words. Using this characterization we derive an explicit formula for the number $\rho(w)$ of runs in words $w \in \mathcal{S}$, with respect to their *recurrences* (*directive sequences*). We show that $\frac{\rho(w)}{|w|} \leq \frac{4}{5}$ for each $w \in \mathcal{S}$, and there is an infinite sequence of strictly growing words $w_k \in \mathcal{S}$ such that $\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} = \frac{4}{5}$. The complete understanding of the function ρ for a large class $\mathcal{S}$ of complicated words is a step towards better understanding of the structure of runs in words. We also show how to compute the number of runs in a standard Sturmian word in linear time with respect to the size of its compressed representation (recurrences describing the word). This is an example of a very fast computation on texts given implicitly in terms of a special grammar-based compressed representation (usually of logarithmic size with respect to the explicit text).

1 Introduction

The runs (maximal repetitions) in strings are important in combinatorics on words and in practical applications: data compression, computational biology, pattern-matching. A run is a non-extendable (with the same period) periodic segment in a string in which the period repeats at least twice. In 1999 Kolpakov and Kucherov [10] showed that the number $\rho(w)$ of runs in a string w is $O(|w|)$, but the exact multiplicative constant coefficient is unknown, recent bounds are given in [11, 5]. In order to better understand the behavior of the function ρ for general words we give **exact** estimations for a class $\mathcal{S}$ of highly compressible words: the standard Sturmian words (standard words, in short). The class $\mathcal{S}$ of standard Sturmian words is of particular interest due to their importance in combinatorics on words, [2, 3]. The standard words are a generalization of Fibonacci words and, like Fibonacci words, are described by recurrences.

The recurrence for a standard word is related to so called *directive sequence* - an integer sequence of the form

$\gamma = (\gamma_0, \gamma_1, ..., \gamma_n)$, where $\gamma_0 \geq 0$, $\gamma_i > 0$ for $0 < i \leq n$.

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

The standard word corresponding to γ, denoted by $S(\gamma) = x_{n+1}$, is defined by recurrences:

$$x_{-1} = b,\ x_0 = a,\ x_1 = x_0^{\gamma_0} x_{-1},\ x_2 = x_1^{\gamma_1} x_0, \tag{1}$$

$$x_3 = x_2^{\gamma_2} x_1,\ \ldots,\ x_n = x_{n-1}^{\gamma_{n-1}} x_{n-2},\ x_{n+1} = x_n^{\gamma_n} x_{n-1} \tag{2}$$

For example the recurrence for the 4-th Fibonacci word is

$$fib_{-1} = b,\ fib_0 = a,\ fib_1 = fib_0^1 b,\ fib_2 = fib_1^1 fib_0,$$

$$fib_3 = fib_2^1 fib_1,\ fib_4 = fib_3^1 fib_2.$$

$$fib_4 = abaababa = S(\gamma_0, \gamma_1, \gamma_2, \gamma_3) \text{ where } (\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (1, 1, 1, 1)$$

We consider here standard words starting with the letter a, hence assume $\gamma_0 > 0$. The case $\gamma_0 = 0$ can be considered similarly. For even $n > 0$ a word x_n has suffix ba, and for odd n it has suffix ab.

The number $N = |x_{n+1}|$ is the (real) size, while n can be thought of as the compressed size.

Example 1. Consider more complicated example (used later to demonstrate counting of runs), let $\gamma = (1, 2, 1, 3, 1)$, we have

$$S(\gamma) = abababaababaababababaababababaababaab$$

The corresponding recurrence is

$$x_{-1} = b;\ x_0 = a,\ x_1 = x_0^1 x_{-1},\ x_2 = x_1^2 x_0, x_3 = x_2^1 x_1,\ x_4 = x_3^3 x_2,\ x_5 = x_4^1 x_3.$$

A number i is a period of the word w iff $w[j] = w[i + j]$ for all i with $i + j \leq |w|$. The minimal period of w will be denoted by $period(w)$. We say that a word w is **periodic** iff $period(w) \leq \frac{|w|}{2}$. A word w is said to be *primitive* iff w is not of the form z^k, where z is a finite word and $k \geq 2$ is a natural number.

A **run** in a string w is an interval $\alpha = [i...j]$ such that $w[i...j]$ is a periodic word with the period $p = period(w[i...j])$ and this period is not extendable to the left or to the right of $[i...j]$. In other words, $[i...j]$ is a run iff $j - i + 1 \geq 2p$, $i = 1$ or $w[i-1] \neq w[i-1+p]$ and $j = n$ or $w[j+1] \neq w[j+1-p]$.

A run α can be properly included as an interval in another run β, but in this case $period(\alpha) < period(\beta)$. The value of the run $\alpha = [i...j]$ is $val(\alpha) = w[i...j]$

When it creates no ambiguity we identify sometimes runs with their values, although two different runs could correspond to identical subwords, if we disregard positions of these runs. Hence runs are also called maximal *positioned* repetitions.

Let $\rho(w)$ be the number of runs in a word w. The most interesting and open conjecture about runs is: $\rho(|w|) < |w|$. The first linear bound was given by Kolpakov and

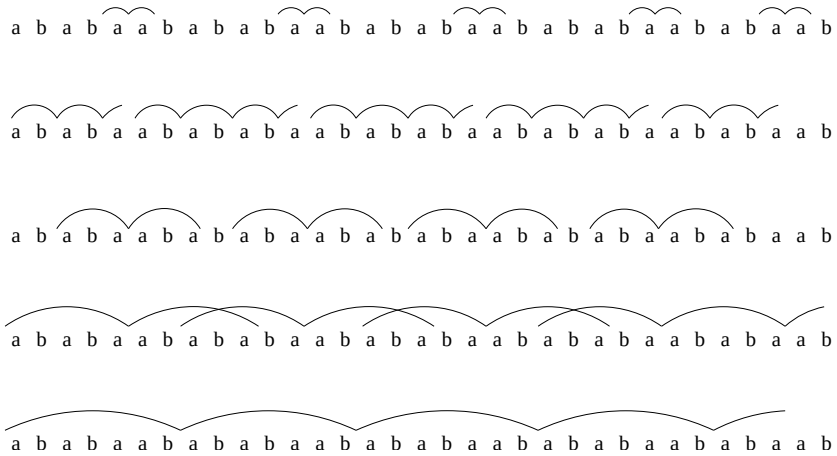

Fig. 1. The structure of runs of $S(1,2,1,3,1)$.There are 5 runs with period $|a|$, 5 with period $|ab|$. We have 10 *short* runs (period of size at most $|x_1| = |ab|$), 8 medium (with period $|x_1| < p \leq |x_2| = 5$, and 1 large run. Consequently $\rho(1,2,1,3,1) = 19$.

Kucherov [10], the best upper bound is by [6, 5] and the best lower bound is by [5, 7]. The structure of runs and squares is almost completely understood for the class of Fibonacci words, see [9, 13, 4]. We continue the work of [8], where it was shown how to compute the number of runs for block-complete Sturmian words (not all standard Sturmian words have this property) in time linear with respect to the size of the whole word (while our algorithm is linear with respect to the size of compressed representation). A similar approach as in [8] is used in this paper – a kind of a reduction sequence, however our reductions are different than those in [8] and correspond closely to the structure of the recurrences (directive sequences). Also our aim is different – derivation of a simple formula for $\rho(w)$ and asymptotic behavior of $\rho(w)$.

Our results We show that $\sup_{w \in \mathcal{S}} \frac{\rho(w)}{|w|} = 0.8$ and provide an easily computable formula for the number of runs. We give also a fast algorithm computing $\rho(w)$ in time linear with respect to the length of the directive sequence defining w: this gives an algorithm efficient with respect to the compressed size of the input.

2 Morphic representations and the numbers $N_\gamma(k)$

Essentially we use an idea of a *reduction sequence* introduced in [8]. The computation of runs in $S(\gamma_0, \gamma_1, \ldots, \gamma_n)$ is reduced to a computation for $S(\gamma_1, \gamma_2, \ldots, \gamma_n)$.
The relation between $S(\gamma_0, \gamma_1, \ldots, \gamma_n)$ and $S(\gamma_1, \gamma_2, \ldots, \gamma_n)$ is described in terms of morphisms transforming one of them to the other.

For $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ define the sequence of morphisms:

$$h_i(a) = a^{\gamma_i} b, \quad h_i(b) = a\,, \quad \text{for } 0 \le i \le n$$

Lemma 1. *Assume $0 \le i < n$. We have*

$$S(\gamma_n) = h_n(a), \;\; S(\gamma_i, \gamma_{i+1} \ldots, \gamma_n) = h_i(S(\gamma_{i+1}, \gamma_{i+2} \ldots, \gamma_n)).$$

Let $|w|_r$ denote the number of occurrences of a letter $r \in \{a,\ b\}$ in the word w. Denote

$$N_\gamma(k) = |S(\gamma_k, \gamma_{k+1}, \ldots \gamma_n)|_a, \;\; M_\gamma(k) = |S(\gamma_k, \gamma_{k+1}, \ldots \gamma_n)|_b$$

The numbers $N_\gamma(k),\ M_\gamma(k)$ satisfy the equation:

$$N_\gamma(k) = \gamma_k\ N_\gamma(k+1) + N_\gamma(k+2); \;\; M_\gamma(k) = N_\gamma(k+1) \tag{3}$$

Observation. In case of the directive sequence $(1, 1, \ldots, 1)$ describing the Fibonacci word the numbers $N_\gamma(k)$ are Fibonacci numbers, since the number of letters a in fib_n equals the size of fib_{n-1}.

Example. For the word $S(1, 2, 1, 3, 1) = ababaababababaababababaababababaababaab$ from Figure 1 we have $\gamma = (1, 2, 1, 3, 1)$ and:

$$S(1) = ab,\ S(3, 1) = aaaba,\ S(1, 3, 1) = (ab)^3 a\ ab,$$

$$N_\gamma(3) = |S(3, 1)|_a = 4,\ N_\gamma(2) = |S(1, 3, 1)|_a = 5$$

Lemma 2. *Let $A = N_\gamma(2),\ B = N_\gamma(3)$ and $w = S(\gamma_0, \gamma_1, \ldots, \gamma_n)$. Then*

$$|w| = ((\gamma_0 + 1)\ \gamma_1 + 1)\ A + (\gamma_0 + 1)\ B$$

Proof. We have $|w| = N_\gamma(0) + M_\gamma(0)$ and $M_\gamma(0) = N_\gamma(1)$.
Hence $|w| = N_\gamma(0) + N_\gamma(1)$ and by equation (3):

$$|w| = \gamma_0 N_\gamma(1) + (\gamma_1 + 1) N_\gamma(2) + N_\gamma(3).$$

Now Equation (3) directly implies the thesis.

For our example word $A = 5,\ B = 4,\ \gamma_0 = 1,\ \gamma_1 = 2$. The formula gives the number $(4 + 1)\ 5 + 8 = 33$, which is the correct length of $S(1, 2, 1, 3, 1)$.

3 Counting runs and repetition ratios in Standard Words

We introduce a zero-one function $unary$ testing if the number equals 1,
if $x = 1$ then $unary(x) = 1$ else $unary(x) = 0$.
Similarly define zero-one functions $even(k)$ and $odd(k)$ with the value equal 1 iff k is even (odd), respectively.

We use the following notation in this section:

$$A \;=\; N_\gamma(2) \;=\; |S(\gamma_2, \gamma_3 \ldots, \gamma_n)|_a,\; B \;=\; N_\gamma(3) \;=\; |S(\gamma_3, \gamma_4 \ldots, \gamma_n)|_a$$

$$\Delta(\gamma) = n - 1 - (\gamma_1 + \ldots + \gamma_n) - unary(\gamma_n).$$

The following theorem will be proven later.

Theorem 1. [Formula for the number of runs]
Let $n \geq 3$ and $\gamma \;=\; (\gamma_0, \ldots, \gamma_n)$. Then the number of runs in $S(\gamma)$ equals

$$\rho(\gamma) \;=\; \begin{cases} 2\;A + 2\;B + \Delta(\gamma) - 1 & \textit{if}\;\; \gamma_0 = \gamma_1 = 1 \\ (\gamma_1 + 2)\;A + B + \Delta(\gamma) - odd(n) & \textit{if}\;\; \gamma_0 = 1;\; \gamma_1 > 1 \\ 2A + 3B + \Delta(\gamma) - even(n) & \textit{if}\;\; \gamma_0 > 1;\; \gamma_1 = 1 \\ (2\;\gamma_1 + 1)\;A \;+\; 2\;B \;+\; \Delta(\gamma) & Otherwise \end{cases},$$

Example 2. We now show how to compute $\rho(1, 2, 1, 3, 1)$, using our formula, for the word shown in Figure 1. In this case

$$\gamma \;=\; (\gamma_0, \gamma_1\gamma_2, \gamma_3, \gamma_4) \;=\; (1, 2, 1, 3, 1) \;\text{ and }\; n = 4$$

$$A = N_\gamma(2) = 5,\; B = N_\gamma(3) = 4,\; \Delta = (4 - 1) - 7 = 4,\; even(n) = 1$$

Theorem 1 implies correctly (see Figure 1):

$$\rho(\gamma) \;=\; (\gamma_1 + 2)A + B + \Delta - even(4) \;=\; 4A + B - 4 - 1 \;=\; 4 \cdot 5 + 4 - 4 - 1 \;=\; 19.$$

Example 3. As the next example derive the formula for the number of runs in Fibonacci word $fib_n \;=\; S(1, 1, \ldots, 1)$ (n ones) for $n \geq 3$. Let F_n be the n-th Fibonacci number. In this case $\;N_\gamma(k) = F_{n-k-1}$. According to formula from Theorem 1 we have

$$\rho(fib_n) \;=\; 2N_\gamma(2) + 2N_\gamma(3) + n - 1 - n - 1 - 1$$

$$=\; 2\;F_{n-3} + 2\;F_{n-4} - 3 \;=\; 2\;F_{n-2} - 3.$$

Theorem 2. $\rho(w) \leq \frac{4}{5}\;|w|$ *for each* $w \in \mathcal{S}$

Proof. The easy when $n \leq 2$ can be considered separately, we omit a simple proof for this case. Assume now that $n \geq 3$ and consider 4 cases.
Let $w \;=\; S(\gamma_0, ..., \gamma_n)$. Observe that $\Delta(\gamma) \leq 0$.

Case 1: $\gamma_0 = \gamma_1 = 1$. We have, due to Lemma 2: $|w| \;=\; 3A + 2B$.
According to Theorem 1 we have $\rho(\gamma) \leq 2\;A + 2\;B$. Then

$$\frac{\rho(w)}{|w|} \;\leq \frac{2A + 2B}{3A + 2B} \;\leq \frac{4}{5}$$

due to inequalities $A \geq B \geq 1$. This completes the proof in this case.

Case 2: $\gamma_0 = 1;\ \gamma_1 > 1$. We have, due to Lemma 2:

$$|w| \ = \ (2\ \gamma_1 + 1)\ A + 2B$$

We have also, due to Theorem 1, that $\rho(w) \leq (\gamma_1 + 2)\ A + B$. Consequently:

$$\frac{\rho(w)}{|w|} \leq \frac{(\gamma_1 + 2)\ A + B}{(2\ \gamma_1 + 1)\ A + 2B} \leq \frac{4}{5}$$

because $\gamma_1 \geq 2$ and $\frac{\gamma_1+2}{2\ \gamma_1+1} \leq \frac{4}{5}$.

Case 3: $\gamma_0 > 1;\ \gamma_1 = 1$. In this case we have $\rho(w) \ \leq \ 2A + 3B$, due to Theorem 1, and , due to Lemma 2,

$$|w| \ = \ (\gamma_0 + 2)\ A \ + \ (\gamma_0 + 1)\ B \ \geq \ 4A + 3B$$

Consequently we have

$$\frac{\rho(w)}{|w|} \leq \frac{2A + 3B}{4A + 3B} \leq \frac{3A + 2B}{4A + 3B} \leq \frac{3}{4}$$

Case 4: $\gamma_0 > 1;\ \gamma_1 > 1$. In this case, due to Theorem 1 and Lemma 2, we have

$$\rho(w) \ \leq \ (2\ \gamma_1 + 1)\ A \ + \ 2\ B,$$

$$|w| \ = \ ((\gamma_0 + 1)\ \gamma_1 + 1)\ A \ + \ (\gamma_0 + 1)\ B.$$

We have

$$\frac{\rho(w)}{|w|} \ \leq \ \frac{(2\ \gamma_1 + 1)\ A \ + \ 2\ B}{((\gamma_0 + 1)\ \gamma_1 + 1)\ A \ + \ (\gamma_0 + 1)\ B} \ \leq \ \frac{(2\ \gamma_1 + 1)\ A \ + \ 2\ B}{(3\ \gamma_1 + 1)\ A \ + \ 3\ B} \ \leq \ \frac{4}{5}$$

because

$$\frac{2\ \gamma_1 + 1}{3\ \gamma_1 + 1} \leq \frac{4}{5}$$

This completes the proof.

Theorem 3.
For the class $\mathcal{S}$ of standard words we have
$\sup\ \{\ \frac{\rho(w)}{|w|}\ :\ w \in \mathcal{S}\ \}\ =\ 0.8.$

Proof. Let

$$w_k = S(1, 2, k, k) = \Big((ababa)^k\ ab\Big)^k\ ababa,$$

see the figure 2 for the case $k = 3$. We have $|w_k| = 5k^2 + 2k + 5$.
Theorem 1 implies that $|\rho(1, 2, k, k)| = 4k^2 - k + 3$. Consequently

$$\lim_{k\to\infty}\ \frac{\rho(w_k)}{|w_k|} \ = \ \lim_{k\to\infty}\ \frac{4k^2 - k + 3}{5k^2 + 2k + 5} \ = \ 0.8$$

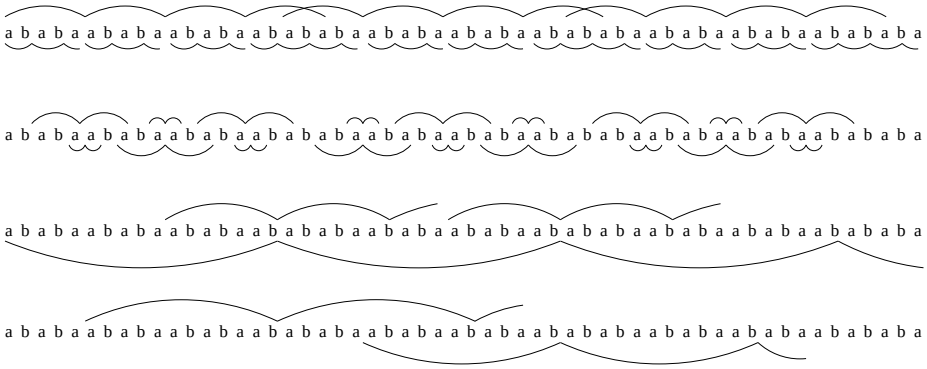

Fig. 2. The structure of runs of $S(1, 2, k, k)$ for $k = 3$, there are $4k^2 - k + 3 = 36$ runs.

Theorem 4.
We can count number of runs in standard word $S(\gamma_0, \ldots, \gamma_n)$ *in time* $O(n)$.

Proof. We need only to compute in $O(n)$ time the numbers $N_\gamma(k)$ for $k = 1, 2, 3$. We can compute it iterating Equation 2.

Algorithm Compute $N_\gamma(k)$;
$x := 1$; $y := 0$;
for $i := n$ **downto** k **do**
    $(x, y) := (\gamma_i \cdot x + y,\ x)$
return x;

4 The proof of Theorem 1

We assume now that x_i' are as given by recurrences described in Equations 1,2. The structure of subword graphs for standard Sturmian words is very special [14, 13], in particular it implies the following fact.

Lemma 3. [Structural Lemma]
The period of each run of $S(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *is of the form* $x_i^j x_{i-1}$, *where* $0 \leq j < \gamma_i$.

We say that a run is **short** if the length of its period does not exceed $|x_1|$, **large** if the period exceeds $|x_2|$, and **medium** otherwise. Denote by $\rho_{short}(\gamma),\ \rho_{med}(\gamma),\ \rho_{large}(\gamma)$ the number of short, medium and large runs in $S\gamma$, respectively. For example we have 10 short, 8 medium and 1 large run in Figure 1.

Lemma 4. [Short Runs] *The number of short runs in $S(\gamma)$ is*

$$\rho_{short}(\gamma) = \begin{cases} N_\gamma(2) + N_\gamma(3) - 1 & \textit{if } \gamma_0 = \gamma 1 = 1 \\ 2\, N_\gamma(2) - odd(n) & \textit{if } \gamma_0 = 1;\ \gamma 1 > 1 \\ N_\gamma(1) + N_\gamma(3) - even(n) & \textit{if } \gamma_0 > 1;\ \gamma 1 = 1 \\ N_\gamma(1) + N_\gamma(2) & \textit{otherwise} \end{cases}$$

Proof. We estimate separately numbers of runs with periods x_0 and x_1

Claim. Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be directive sequence. There are:
- **(a)** $N_\gamma(1)$ runs with period x_0 if $\gamma_0 > 1$,
- **(b)** $M_\gamma(1)$ runs with period x_0 if $\gamma_0 = 1$,
- **(c)** $N_\gamma(2)$ runs with period x_1 if $\gamma_1 > 1$,
- **(d)** $M_\gamma(2)$ runs with period x_1 if $\gamma_1 = 1$.

Point (a). Let us define morphism $h(a) = a^{\gamma_0} b$ and $h(b) = a$. Every run with period x_0 in $S(\gamma)$ is equal to a^{γ_0} or a^{γ_0+1}. Every such run is separated by the letter b and corresponds to the letter a in $h^{-1}(S(\gamma_0, \ldots, \gamma_n)) = S(\gamma_1, \ldots, \gamma_n)$.

Point (b). The proof of this point is similar to (a).

Points (c,d). A run with the period x_1 in $S(\gamma)$ corresponds to a run with the period x_0 in $h^{-1}(S(\gamma))$ and now validity of this case follows from points (a) and (b). This completes the proof of the claim and the lemma.

Lemma 5. [Medium Runs, $\mathbf{n \geq 3}$] *If $n \geq 3$ then*

$$\rho_{med}(\gamma) \ = \ N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1$$

Proof. The thesis follows directly from the following stronger claim (the proof is omitted in this version)

Claim. Let $\gamma = (\gamma_0, \ldots, \gamma_n)$. There are:

- **(a)** $N_\gamma(2) - 1$ runs with period $x_1^i x_0$ for each $0 < i < \gamma_1$.
- **(b)** $N_\gamma(3)$ runs with period x_2.

The claim of the lemma follows by summing formulas from the points (a) and (b). We have

$$(N_\gamma(2) - 1)\ (\gamma_1 - 1) + N_\gamma(3) \ =$$
$$(\gamma_1 N_\gamma(2) + N_\gamma(3)) \ - \ N_\gamma(2) - \gamma_1 + 1 \ = \ N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1$$

This completes the proof of the lemma.

Lemma 6. [Medium Runs, n=2] *If $n = 2$ then*

$$\rho_{med}(\gamma) \ = \ N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1 - unary(\gamma_n)$$

Proof. The proof for the case $\gamma_n > 1$ is similar to the one for Lemma 5. In the case $\gamma_n = 1$ there are no intermediate runs, and we have to subtract $unary(\gamma_n) = 1$ in this case.

We reduce the problem of counting large runs to the one for counting medium runs, using the morphic representation of $S\gamma$. Let h be a morphism and let $y = a_1 a_2 \dots a_t$ be a word of length t.
The morphism partitions $x = h(y)$ into segments $h(a_1),\ h(a_2) \dots h(a_t)$. These segments are called here h-blocks.

We say that a subword w of x is **synchronized** with h in x iff each occurrence of w in x starts at the beginning of some h-block and ends at the end of some h-block. Figure 3 shows examples of synchronized and non-synchronized subwords with the morphism $h_0 : S(2,1,3,1) \to S(1,2,1,3,1)$ related to the morphic structure of $S(1,2,1,3,1)$. Recall that $h_0(a) = a^{\gamma_0} b,\ h_0(b) = a$.

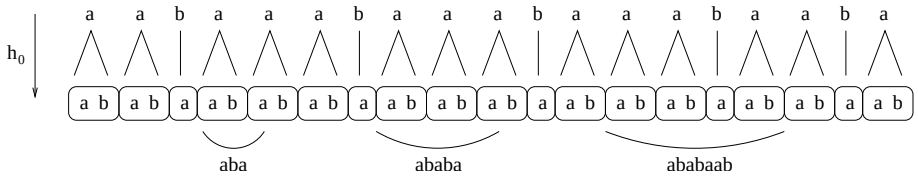

Fig. 3. The medium run-periods $x_1 x_0 = aba$ and $x_2 = ababa$ do not synchronize with h_0 on the string from Figure 1, while the large run-period $x_3 = ababaab$ is synchronized with h_0.

Lemma 7. [Synchronization Lemma]
The large run-periods are synchronized with h_0 in $S(\gamma_0, \dots, \gamma_n)$

Proof. We omit the proof of the following *syntactical* fact.

Claim.
(a) If $i \geq 2$ then $x_i x_{i-1}$ ends with $a^{\gamma_0} b$ or with $(a^{\gamma_0} b)^{\gamma_1 + 1} a$
(b) $a^{\gamma_1 + 2}$ is not a sub-word in $S(\gamma_1, \dots, \gamma_n)$

In the inverse morphism h_0^{-1} the block $a^{\gamma_0} b$ goes to a and the block a goes to b. If the word starts and ends with $a^{\gamma_0} b$ then it is obviously synchronized with the morphism. The word $x_i x_{i-1}$, for $i \geq 2$, starts with $a^{\gamma_0} b$. The only problem is when it ends with a and this occurrence of a is followed by $a^{\gamma_0 - 1} b$. However, due to the point (a) of the claim, we have an occurrence of the sequence $(a^{\gamma_0} b)^{\gamma_1 + 2}$ in $S(\gamma_1, \dots, \gamma_n)$. After applying the inverse of h_0 this sequence goes to $a^{\gamma_1 + 2}$ in $S(\gamma_1, \dots, \gamma_n)$. However this is impossible due to point (b) of the claim. This completes the proof.

The following fact is implied by synchronization lemma.

Lemma 8. [Recurrence Lemma]

$$\rho_{large}(\gamma_0, \gamma_1, \dots, \gamma_n) = \rho_{large}(\gamma_1, \gamma_2, \dots, \gamma_n) + \rho_{med}(\gamma_1, \gamma_2, \dots, \gamma_n).$$

4.1 Completing the proof of Theorem 1

The claim of the next lemma follows from Lemma 5 and the recurrence from Lemma 8.

Lemma 9. [Large Runs]

$$\rho_{large} + \rho_{med} \;=\; N_\gamma(1) \;+\; n-1 \;-\; (\gamma_1 + ... + \gamma_n) \;- unary(\gamma_n)$$

Proof. According to Lemma 5 we have

$$\rho_{large} + \rho_{med} \;=\; (N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1) +$$

$$(N_\gamma(2) - N_\gamma(3) - \gamma_2 + 1) \;+\; \ldots + \; (N_\gamma(n-1) - N_\gamma(n) - \gamma_{n-1} + 1 - unary(\gamma_n))$$

$$=\; N_\gamma(1) \;+\; n-1 \;-\; (\gamma_1 + ... + \gamma_n) \;- unary(\gamma_n),$$

since $N_\gamma(n) = \gamma_n$. This completes the proof.

Now the formula in Theorem 1 results by combining the formulas for ρ_{short} and for the sum $\rho_{large} + \rho_{med}$ using the equalities

$$\rho(\gamma) \;=\; \rho_{short}(\gamma) + \rho_{med}(\gamma) + \rho_{large}(\gamma), \;\text{ and }\; N_\gamma(1) \;=\; \gamma_1 N_\gamma(2) + N\gamma(3).$$

References

1. P. Baturo, W.Rytter, Occurrence and lexicographic properties of standard Sturmian words, LATA 2007
2. J. Berstel, 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, Cambridge (2002) 45-110
3. J. Berstel, J. Karhumäki, Combinatorics on words - a tutorial. Bull. EATCS 79 (2003), pp 178-228.
4. C. Iliopoulos, D. Moore, W.F. Smyth, Characterization of the Squares in a Fibonacci String. Theor. Comput. Sci. 172(1-2): 281-291 (1997)
5. M. Crochemore and L. Ilie, Analysis of Maximal Repetitions in Strings MFCS 2007, 465-476
6. M. Crochemore, L. Ilie, I. Tinta, Towards a solution to the "runs" conjecture, to be published in CPM 2008
7. F. Franek, R.J. Simpson, W.F. Smyth, The maximum number of runs in a string, in: M. Miller, K. Park (Eds.) Proceeding of 14th Australian Workshop on Combinatorial Algorithms, (2003), 26-35.
8. F. Franek, A. Karaman, W. F. Smyth, Repetitions in Sturmian strings, Theoretical Computer Science 249-2 (2000) 289-303.
9. R. Kolpakov, G. Kucherov, On Maximal Repetitions in Words. FCT 1999: 374-385
10. R. Kolpakov, G. Kucherov, Finding Maximal Repetitions in a Word in Linear Time. FOCS 1999: 596-604
11. W. Rytter, The number of runs in a string, Information and Computation Volume 205, Issue 9, (2007), 1459-1469.
12. W. Rytter, Grammar Compression, LZ-Encodings, and String Algorithms with Implicit Input. ICALP 2004: 15-27
13. W. Rytter, The structure of subword graphs and suffix trees of Fibonacci words, Theoretical Computer Science Volume 363, Issue 2, (2006), 211 - 223.
14. M. Sciortino, L. Zamboni, Suffix Automata and Standard Sturmian Words, DLT'07, 382-398
15. B. Smyth, Computing patterns in strings, Addison Wesley 2003