# The maximal number of runs in standard Sturmian words

Paweł Baturo [‡] Marcin Piątkowski [‡*] Wojciech Rytter [†, ‡]

[‡] Faculty of Mathematics and Computer Science
Nicolaus Copernicus University

[†] Institute of Informatics
Warsaw University

{pbaturo, marcin.piatkowski, rytter}@mat.umk.pl

Submitted: …; Accepted: …; Published: XX
Mathematics Subject Classification: 05C88

**Abstract**

Problems related to repetitions are central in the area of combinatorial algorithms on strings. In this paper we investigate these problems for a very special class $\mathcal{S}$ of strings called the standard Sturmian words. They have very compact representations in terms of sequences of integers. For a sequence $\gamma$ denote by $\mathrm{Sw}(\gamma)$ the standard word generated by $\gamma$. Usually the size of this word is exponential with respect to the size of $\gamma$, hence we are dealing here with repetition problems in compressed strings. In this paper we derive an explicit formula for the number $\rho(w)$ of runs in a standard word $w$, show that $\frac{\rho(w)}{|w|} \leqslant \frac{4}{5}$ for each $w \in \mathcal{S}$, and there is an infinite sequence of strictly growing words $w_k \in \mathcal{S}$ such that $\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} = \frac{4}{5}$. 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).

## 1 Introduction

A *run* (a *maximal repetition*) is a non-extendable (with the same period) periodic segment in a string, in which the period repeats at least twice. Runs are important in combinatorics on words and many practical applications: data compression, computational biology,

*The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL „Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

pattern-matching and so on. The structure of repetitions is almost completely understood for the class of Fibonacci words, see [9], [12], [7]. In this paper we investigate the structure of runs in class $\mathcal{S}$ of standard Sturmian words and give the exact formula and the tight asymptotic bound for the number of maximal repetitions.

We continue here the work of [4], 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 linear time with respect to the size of the whole word. We show the algorithm, which computes the number of runs in any standard word in linear time with respect to the size of its compressed representation – the directive sequence – hence in logarithmic time with respect to the length of the word.

Throughout the paper we use the standard notions of combinatorics on words. In particular, words are finite sequences over a finite set $\Sigma$ of letters, called the alphabet. For a word $w = w_1 w_2 \dots w_n$, by $w[i]$ we denote its $i$-th letter (namely $w_i$), by $w[i..j]$ the subword $w_i w_{i+1} \dots w_j$, by $w$ its length and by $|w|_a$ the number of letters $a$ occurring in $w$.

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

A *maximal repetition* (a *run*, in short) in a word $w$ is an interval $\alpha = [i..j]$ such that $w[i..j] = u^k v$ $(k \geqslant 2)$ is a nonempty periodic subword of $w$, where $u$ is of the minimal length and $v$ is a proper prefix (possibly empty) of $u$, that can not be extended (neither $w[i-1..j]$ nor $w[i..j+1]$ is a run with the period $|u|$).

A run $\alpha$ can be properly included as an interval in another run $\beta$, but in this case $period(\alpha) < period(\beta)$. The value of the run $\alpha = [i...j]$ is the factor $val(\alpha) = w[i...j]$. When it creates no ambiguity we identify sometimes run with its value and the period of the run $\alpha = [i...j]$ with the subword $w[i..period(w)]$ – called also the *generator* of the repetition. The meaning will always be clear from the context. Observe that two different runs could correspond to the identical subwords, if we disregard their positions. Hence runs are also called the maximal *positioned* repetitions.

**Example 1.** Let $w = ababaabababaabababaabababaababaab$.
There are 5 runs with the period $|a|$:

$$w[5..6] = a^2, \quad w[12..13] = a^2, \quad w[19..20] = a^2,$$
$$w[26..27] = a^2, \quad w[31..32] = a^2,$$

5 runs with the period $|ab|$:

$$w[1..5] = (ab)^2 a, \quad w[6..12] = (ab)^3 a, \quad w[13..19] = (ab)^3 a,$$
$$w[20..26] = (ab)^3 a, \quad w[27..31] = (ab)^2 a,$$

4 runs with the period $|aba|$:

$$w[3..8] = (aba)^2, \qquad w[10..15] = (aba)^2,$$
$$w[17..22] = (aba)^2, \qquad w[24..29] = (aba)^2,$$

4 runs with the period $|ababa|$:

$$w[1..10] = (ababa)^2, \qquad w[8..17] = (ababa)^2,$$
$$w[15..24] = (ababa)^2, \qquad w[22..33] = (ababa)^2ab,$$

and 1 run with the period $|ababaab|$:

$$w[1..31] = (ababaab)^4aba.$$

Alltogether we have 19 runs, see Figure 1 for comparison.

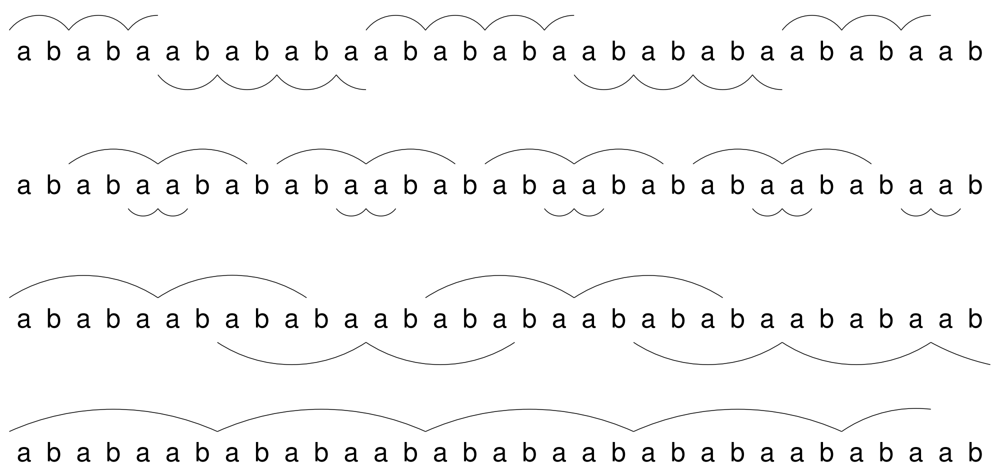


***Figure 1:*** *The structure of maximal repetitions for an example binary word.*

Denote by $\rho(w)$ the number of runs in the word $w$ and by $\rho(n)$ the maximal number of runs in the words of length $n$. The most interesting and open conjecture about maximal repetitions is:

$$\rho(n) < n.$$

In 1999 Kolpakov and Kucherov (see [8]) showed that the number $\rho(w)$ of runs in a string $w$ is $O\big(|w|\big)$, but the exact multiplicative constant coefficient is still unknown. The best known results related to the value of $\rho(n)$ are

$$0.944542\, n \leqslant \rho(n) \leqslant 1.048\, n.$$

The upper bound is by [1], [2] and the lower bound is by [5], [6], [10]. Table 1 shows the maximal number of runs and the repetition ratio in all binary words compared to standard words for small values of $n$.

## 2 Standard words

The *directive sequence* is the integer sequence: $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$, where $\gamma_0 \geqslant 0$ and $\gamma_i > 0$ for $i = 1, 2, \ldots, n$. The standard word corresponding to $\gamma$, denoted by $x_{n+1} = \mathrm{Sw}(\gamma)$, is described by the recurrences of the form:

$$x_{-1} = b, \qquad x_0 = a, \qquad \ldots, \qquad x_n = x_{n-1}^{\gamma_{n-1}} x_{n-2}, \qquad x_{n+1} = x_n^{\gamma_n} x_{n-1}. \tag{1}$$

The sequence of words $\{x_i\}_{i=0}^{n+1}$ is called the standard sequence. Every word occurring in a standard sequence is a standard word, and every standard word occurs in some standard sequence. We assume that the standard word given by the empty directive sequence is $a$ and $\mathrm{Sw}(0) = b$. The class of all standard words is denoted by $\mathcal{S}$.

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

$$\begin{array}{lclcl}
x_{-1} & = & b & & \\
x_0 & = & a & & \\
x_1 & = & (x_0)^1 \cdot x_{-1} & = & a \cdot b \\
x_2 & = & (x_1)^2 \cdot x_0 & = & ab \cdot ab \cdot a \\
x_3 & = & (x_2)^1 \cdot x_1 & = & ababa \cdot ab \\
x_4 & = & (x_3)^3 \cdot x_2 & = & ababaab \cdot ababaab \cdot ababaab \cdot ababa \\
x_5 & = & (x_4)^1 \cdot x_3 & = & ababaababaababaababaababa \cdot ababaab
\end{array}$$

and finally

$$\mathrm{Sw}(1, 2, 1, 3, 1) \;=\; ababaababaababaababaababaababaab.$$

For $\gamma_0 > 0$ we have standard words starting with the letter $a$ and for $\gamma_0 = 0$ we have standard words starting with the letter $b$. In fact the word $\mathrm{Sw}(0, \gamma_1, \ldots, \gamma_n)$ can be obtained from $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ by switching the letters $a$ and $b$.

Observe that for even $n > 0$ the standard word $x_n$ has the suffix $ba$, and for odd $n > 0$ it has the suffix $ab$. Moreover, every standard word consists either of repeated occurrences of the letter $a$ separated by single occurrences of the letter $b$ or repeated occurrences of the letter $b$ separated by single occurrences of the letter $a$. Those letters are called the *repeating letter* and the *single letter*, respectively. If the repeating letter is $a$ (letter $b$ respectively), the word is called the Sturmian word of the type $a$ (type $b$ respectively), see the definition 6.1.4 in [11] for comparison.

*Remark* 3. Without loss of generality we consider in this paper the standard Sturmian words of the type $a$, therefore we assume that $\gamma_0 > 0$. The words of the type $b$ can be considered similarly and all the results hold.

*Remark* 4. Standard words are generalization of Fibonacci words, the well known family of strings. By definition Fibonacci words are standard words given by directive sequences of the form $\gamma = (1, 1, \ldots, 1)$ (the $n$-th Fibonacci word $F_n$ corresponds to a sequence of $n$ ones).

# 3 Morphic reduction of standard words

The recurrent definition of standard words from the previous section leads to the simple characterization by the composition of morphisms.

Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a directive sequence. We associate with $\gamma$ a sequence of morphisms $\{h_i\}_{i=0}^n$, defined as

$$h_i : \begin{cases} a \longrightarrow a^{\gamma_i} b \\ b \longrightarrow a \end{cases} \qquad \text{for } 0 \leqslant i \leqslant n. \tag{2}$$

**Lemma 5.**
*For $0 \leqslant i \leqslant n$ the morphism $h_i$ transforms a standard word into another standard word, and we have:*

$$\begin{aligned} \mathrm{Sw}(\gamma_n) &= h_n(a), \\ \mathrm{Sw}(\gamma_i, \gamma_{i+1}, \ldots, \gamma_n) &= h_i\big(\mathrm{Sw}(\gamma_{i+1}, \gamma_{i+2}, \ldots, \gamma_n)\big). \end{aligned}$$

*Proof.* The induction on the length of the directive sequence.

Recall that the standard word given by the empty directive sequence is $a$. For $|\gamma| = 1$ we have, by definition of standard words and the morphism $h_n$,

$$\mathrm{Sw}(\gamma_n) = a^{\gamma_n} b = h_n(a).$$

Assume now that $|\gamma| = k \geqslant 2$ and for directive sequences shorter than $k$ the thesis holds. We have then:

$$\begin{aligned} \mathrm{Sw}(\gamma_i, \ldots, \gamma_n) &= \big[\mathrm{Sw}(\gamma_i, \ldots, \gamma_{n-1})\big]^{\gamma_n} \cdot \mathrm{Sw}(\gamma_i, \ldots, \gamma_{n-2}) \\ &\overset{ind.}{=} \Big[h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-1})\Big)\Big]^{\gamma_n} \cdot h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-2})\Big) \\ &= h_i\Big(\big[\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-1})\big]^{\gamma_n} \cdot \mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-2})\Big) \\ &= h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_n)\Big), \end{aligned}$$

which concludes the proof. □

*Remark* 6. As a direct conclusion from Lemma 5 we have that for the directive sequence $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_n)$

$$\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n) \;=\; h_0 \circ h_1 \circ \dots \circ h_n(a).$$

**Example 7.** Consider the directive sequence $\gamma = (1, 2, 1, 3, 1)$.
We have:

$$\begin{aligned}
\mathrm{Sw}(1) &= h_4(a) && = ab\\
\mathrm{Sw}(3,1) &= h_3\big(\mathrm{Sw}(1)\big) && = aaaba\\
\mathrm{Sw}(1,3,1) &= h_2\big(\mathrm{Sw}(3,1)\big) && = abababaab\\
\mathrm{Sw}(2,1,3,1) &= h_1\big(\mathrm{Sw}(1,3,1)\big) && = aabaaabaaabaaabaaba\\
\mathrm{Sw}(1,2,1,3,1) &= h_0\big(\mathrm{Sw}(2,1,3,1)\big) && = ababaababababaababababaabababaababaab.
\end{aligned}$$

Compare with Example 2.

## Reduction sequence

Observe that the inverse morphism $h_i^{-1}$ can be seen as a reduction of the word $w_i = \mathrm{Sw}(\gamma_i, \dots, \gamma_n)$ to $w_{i+1} = \mathrm{Sw}(\gamma_{i+1}, \dots, \gamma_n)$. Using this approach we can reduce the computation of runs in $w_i$ to the same computation in $w_{i+1}$. Our concept is similar to the one shown in [4], but is more closely related to the combinatorial structure of standard words.

Recall that $|w|_a$ denotes the number of occurrences of the letters $a$ in the word $w$. In the rest of this paper we use the following notation:

$$\begin{aligned}
N_\gamma(k) &= \big|\mathrm{Sw}(\gamma_k, \gamma_{k+1}, \dots, \gamma_n)\big|_a,\\
M_\gamma(k) &= \big|\mathrm{Sw}(\gamma_k, \gamma_{k+1}, \dots, \gamma_n)\big|_b,
\end{aligned} \tag{3}$$

which enables us to simplify the formulas for the number of runs.

*Remark* 8. As a direct conclusion from the above definition, the equation (1) and the equation (2) we have that the numbers $N_\gamma(k)$ and $M_\gamma(k)$ satisfy:

$$\begin{aligned}
N_\gamma(k) &= \gamma_k\, N_\gamma(k+1) + N_\gamma(k+2),\\
M_\gamma(k) &= N_\gamma(k+1).
\end{aligned} \tag{4}$$

*Remark* 9. Observe that for Fibonacci word $F_n$ the number of the letters $a$ in $F_n$ equals the length of the word $F_{n-1}$, and therefore

$$N_\gamma(k) = |F_{n-k-1}| \qquad \text{and} \qquad M_\gamma(k) = |F_{n-k-2}|.$$

**Example 10.** Let $\gamma = (1, 2, 1, 3, 1)$ be a directive sequence. We have then:

| $\gamma$ | $\mathrm{Sw}(\gamma)$ | $\lvert\mathrm{Sw}(\gamma)\rvert_a$ | $\lvert\mathrm{Sw}(\gamma)\rvert_b$ |
|---|---|---|---|
| $(1)$ | $ab$ | $N_\gamma(4) = 1$ | $M_\gamma(4) = 1$ |
| $(3, 1)$ | $aaaba$ | $N_\gamma(3) = 4$ | $M_\gamma(3) = 1$ |
| $(1, 3, 1)$ | $abababaab$ | $N_\gamma(2) = 5$ | $M_\gamma(2) = 4$ |
| $(2, 1, 3, 1)$ | $aabaaabaaabaaabaaba$ | $N_\gamma(1) = 14$ | $M_\gamma(1) = 5$ |
| $(1, 2, 1, 3, 1)$ | $ababaababababaabababaababababaababaab$ | $N_\gamma(0) = 19$ | $M_\gamma(0) = 14$ |

The following lemma enables us to express the length of any standard word in terms of the numbers $N_\gamma(k)$ and $M_\gamma(k)$.

**Lemma 11.**
*Let* $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$, $A = N_\gamma(2)$ *and* $B = N_\gamma(3)$. *Then*

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

*Proof.* By the definition of $N_\gamma(k)$ and $M_\gamma(k)$ we have

$$|w| \;=\; N_\gamma(0) + M_\gamma(0) \qquad \text{and} \qquad M_\gamma(0) \;=\; N_\gamma(1).$$

By repeated application of the formulas from the equation (3) we obtain:

$$\begin{aligned}
|w| &= N_\gamma(0) + N_\gamma(1) \\
&= (\gamma_0 + 1)\,N_\gamma(1) + N_\gamma(2) \\
&= \big((\gamma_0 + 1)\,\gamma_1 + 1\big)\,N_\gamma(2) + (\gamma_0 + 1)\,N_\gamma(3) \\
&= \big((\gamma_0 + 1)\,\gamma_1 + 1\big)\,A + (\gamma_0 + 1)\,B
\end{aligned}$$

and the proof is complete. □

# 4 The formula and the algorithm

In this section we present the formula for the number of runs in standard words and investigate its asymptotic behaviour. The proof of its correctness is the aim of the section 6.

We begin with the definition of some useful zero-one functions for testing the parity of a nonnegative integer $i$:

$$even(i) \;=\; \begin{cases} 1 & \text{for even } i \\ 0 & \text{for odd } i \end{cases} \qquad \text{and} \qquad odd(i) \;=\; \begin{cases} 1 & \text{for odd } i \\ 0 & \text{for even } i \end{cases},$$

and for testing if a positive integer $i$ equals 1:

$$unary(i) \;=\; \begin{cases} 1 & \text{for} \quad i=1 \\ 0 & \text{for} \quad i>1 \end{cases}.$$

These functions will be used to simplify the formula for the number of runs in standard words.

**Theorem 12** (Formula for the number of runs)**.**
*Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $n \geqslant 3$. The number of runs in a standard word $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ is given by the following formula:*

$$\rho(w) \;=\; \begin{cases} 2A + 2B + \Delta(\gamma) - 1 & \textit{for} \quad \gamma_0 = \gamma_1 = 1 \\ (\gamma_1 + 2)A + B + \Delta(\gamma) - odd(n) & \textit{for} \quad \gamma_0 = 1;\ \gamma_1 > 1 \\ 2A + 3B + \Delta(\gamma) - even(n) & \textit{for} \quad \gamma_0 > 1;\ \gamma_1 = 1 \\ (2\gamma_1 + 1)A + 2B + \Delta(\gamma) & \textit{for} \quad \gamma_0 > 1;\ \gamma_1 > 1 \end{cases}, \tag{5}$$

*where:*

$$\begin{aligned} A &= N_\gamma(2) = |\mathrm{Sw}(\gamma_2, \gamma_3, \ldots, \gamma_n)|_a, \\ B &= N_\gamma(3) = |\mathrm{Sw}(\gamma_3, \gamma_4, \ldots, \gamma_n)|_a, \\ \Delta(\gamma) &= n - 1 - (\gamma_1 + \ldots + \gamma_n) - unary(\gamma_n). \end{aligned}$$

The detailed proof of the above theorem is shown in Section 6.

The formula presented above leads to a simple and fast algorithm to compute the number of runs in standard words.

**Theorem 13.**
*We can count the number of runs in any standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ in linear time with respect to the length of the directive sequence $|\gamma|$.*

*Proof.* The formula for the number of runs in standard words from Theorem 12 depends directly on the components of the directive sequence $\gamma$ and the numbers $N_\gamma(2)$ and $N_\gamma(3)$. It is sufficient to prove that we can compute the numbers $N_\gamma(k)$ for $k = 1, 2, 3$ in time $O(n)$. For this purpose we can iterate the equation (1):

**Algorithm 1:** to compute $N_\gamma(k)$

**Input**: $\gamma = (\gamma_0, \ldots, \gamma_n)$
**Output**: $N_\gamma(k)$

$x \longleftarrow 1$ $y \longleftarrow 0$
**for** $i := n$ **to** $k$ **do**
  $(x, y) \longleftarrow (\gamma_i \cdot x + y, x)$
**return** $x$

Using the above algorithm and applying the formula from the equation (5) we can count the number of runs in any standard word in linear time with respect to the size of the directive sequence (logarithmic with respect to the length of the whole word). □

Now we can use the formula from the equation (5) to compute the number of runs in some example standard words.

**Example 14.** Let $\gamma = (1, 2, 1, 3, 1)$ be a directive sequence. We have $n = 4$ and

$$\mathrm{Sw}(\gamma) \;=\; abababaababababaabababaabababaababaab.$$

In this case

$$A = N_\gamma(2) = 5, \quad B = N_\gamma(3) = 4, \quad \Delta = (4-1) - 7 - 1 = -5, \quad odd(4) = 0.$$

Theorem 12 implies:

$$\begin{aligned} \rho(w) &= (\gamma_1 + 2)\; A + B + \Delta - odd(4) \\ &= 4\; A + B - 5 \\ &= 4 \cdot 5 + 4 - 5 \\ &= 19, \end{aligned}$$

see Figure 1 and Example 1 for comparison.

It is known that the number of runs in the $n$-th Fibonacci word $F_n$ is given by the formula

$$\rho(F_n) \;=\; 2\; F_{n-2} + 3,$$

see [9] for the proof. As the next example we derive this formula using Theorem 12.

**Example 15.** Recall that $F_n = \mathrm{Sw}(1, 1, \ldots 1)$ ($n$ ones) and in this case $N_\gamma(k) = F_{n-k-1}$. According to the formula from the equation (5) we have:

$$\begin{aligned} \rho(F_n) &= 2\; N_\gamma(2) + 2\; N_\gamma(3) + n - 1 - n - 1 - 1 \\ &= 2\; F_{n-3} + 2\; F_{n-4} - 3 \\ &= 2\; F_{n-2} - 3. \end{aligned}$$

# 5 Asympthotic behaviour of the number of runs

The following lemma gives us the bound for the number of runs in standard words described by the directive sequences of the length at most 2.

**Lemma 16** (Estimation for short $\gamma$)**.**
*Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence, $w = \mathrm{Sw}(\gamma)$ and $n \leqslant 2$. Then $\rho(w) < \frac{4}{5}\,|w|$.*

*Proof.* Recall that the standard word given by the empty directive sequence is $a$ and does not include any repetition. Therefore, we have to consider two cases: $|\gamma| = 1$ and $|\gamma| = 2$.

**Case 1:** First assume that $|\gamma| = 1$.
Then, $\gamma = (\gamma_0)$ and

$$w = \mathrm{Sw}(\gamma_0) = a^{\gamma_0}b \qquad \text{and} \qquad |w| = \gamma_0 + 1.$$

There is one run for $\gamma_0 > 1$, no run for $\gamma_0 = 1$ and obviously $\rho(w) < \frac{4}{5}|w|$.

**Case 2:** Assume now that $|\gamma| = 2$.
We have $\gamma = (\gamma_0, \gamma_1)$ and

$$w \;=\; \mathrm{Sw}(\gamma_0, \gamma_1) \;=\; \left(a^{\gamma_0}b\right)^{\gamma_1}a \qquad \text{and} \qquad |w| \;=\; (\gamma_0 + 1)\gamma_1 + 1.$$

The number of runs in $w$ depends on the values of $\gamma_0$ and $\gamma_1$ as follows:

$$\rho(w) = \begin{cases} 0 & \text{for} \quad \gamma_0 = 1,\; \gamma_1 = 1 \\ 1 & \text{for} \quad \gamma_0 > 1,\; \gamma_1 = 1 \\ 1 & \text{for} \quad \gamma_0 = 1,\; \gamma_1 > 1 \\ \gamma_1 + 1 & \text{for} \quad \gamma_0 > 1,\; \gamma_1 > 1 \end{cases}.$$

In each case we have

$$\rho(w) \;<\; \frac{4}{5}\Big((\gamma_0 + 1)\gamma_1 + 1\Big) \;=\; \frac{4}{5}|w|$$

and the proof is complete. □

Now we are ready to estimate the asymptotic bound for the number of runs in all standard Sturmian words.

**Theorem 17** (Upper bound)**.**
*For each standard Sturmian word $w$ we have $\rho(w) \leqslant \frac{4}{5}\,|w|$.*

*Proof.* Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, ..., \gamma_n)$ be a standard word. Recall the formula (5) from Theorem 12 and observe that $\Delta(\gamma) \leqslant 0$.

The case when $n \leqslant 2$ follows from the lemma 16. It is sufficient to prove the thesis for $n \geqslant 3$. We consider four cases depending on the values of $\gamma_0$ and $\gamma_1$.

**Case 1:** $\gamma_0 = \gamma_1 = 1$.

We have, due to Lemma 11 and the equation (5):

$$|w| \;=\; 3A + 2B \qquad \text{and} \qquad \rho(w) \;\leqslant\; 2\,A + 2\,B.$$

Hence

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

due to inequalities $A \geqslant B \geqslant 1$. This completes the proof of this case.

**Case 2:** $\gamma_0 = 1;\ \gamma_1 > 1$.

We have, due to Lemma 11 and the equation (5):

$$|w| \ = \ (2\,\gamma_1 + 1)\ A + 2B \qquad \text{and} \qquad \rho(w) \leqslant (\gamma_1 + 2)\ A + B.$$

Consequently:

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

because $\gamma_1 \geqslant 2$ and $\frac{\gamma_1+2}{2\,\gamma_1+1} \leqslant \frac{4}{5}$.

**Case 3:** $\gamma_0 > 1;\ \gamma_1 = 1$.

Due to the equation (5) and Lemma 11, we have:

$$\rho(w) \ \leqslant \ 2A + 3B,$$

$$|w| \ = \ \Big((\gamma_0 + 2)\ A + (\gamma_0 + 1)\ B\Big) \ \geqslant \ 4A + 3B,$$

and consequently:

$$\frac{\rho(w)}{|w|} \ \leqslant \ \frac{2A + 3B}{4A + 3B} \ \leqslant \ \frac{3A + 2B}{4A + 3B} \ \leqslant \ \frac{3}{4} \ < \ \frac{4}{5}.$$

**Case 4:** $\gamma_0 > 1;\ \gamma_1 > 1$.

In this case, due to the equation (5) and Lemma 11, we have:

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

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

and consequently

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

because

$$\frac{2\,\gamma_1 + 1}{3\,\gamma_1 + 1} \leqslant \frac{4}{5}.$$

This completes the proof of the theorem. □

The above results give us the asymptotic bound for the number of runs in standard words. Below we construct a strictly growing sequence of standard words to show that this estimation is tight.

**Theorem 18.**
*For the class $\mathcal{S}$ of standard words we have:*

$$\sup \left\{ \frac{\rho(w)}{|w|} \ : \ w \in \mathcal{S} \right\} \ = \ 0.8.$$

*Proof.* Let $\gamma = (1, 2, k, k)$ and $w_k \ = \ \mathrm{Sw}(1, 2, k, k)$. By definition we have

$$w_k \ = \ \Big( (ababa)^k \ ab \Big)^k ababa \qquad \text{and} \qquad |w_k| \ = \ 5k^2 + 2k + 5,$$

and due to the equation (5):

$$\big|\rho(w_k)\big| \ = \ 4k^2 - k + 3,$$

see Figure 2 for the case $k = 3$. Consequently

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

which completes the proof. $\square$

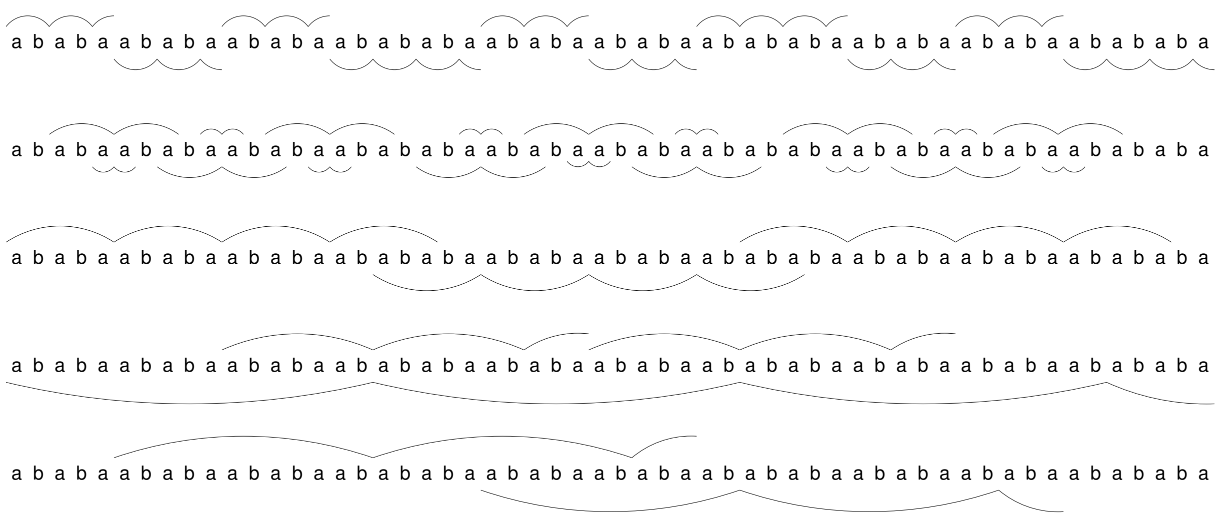


***Figure 2:*** *The structure of runs in the standard word* $\mathrm{Sw}(1, 2, k, k)$ *for* $k = 3$. *There are* $4k^2 - k + 3 \ = \ 36$ *runs.*

# 6 The proof of Theorem 12 (the formula)

This section is devoted to the proof of Theorem 12. We begin with the characterization of the structure of the possible periods of the maximal repetitions in standard words. Recall their recurrent definition given by the equation (1) and the words $x_i$ defined there.

The following lemma is a version of Theorem 1 in [3] using slightly different notation.

**Lemma 19** (Structural Lemma)**.**
*The period of each maximal repetition in the standard word* $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *is of the form* $x_i^j x_{i-1}$*, where* $0 \leqslant j < \gamma_i$*.*

To prove the above lemma it is sufficient to show that no factor of the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$, that does not satisfy the condition given there, could be the generator of some repetition, see the proof of Theorem 1 in [3] for more details.

The main idea of the proof of the correctness of the formula given in the equation (5) is the partition of the set of all maximal repetitions in the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ into three separate categories depending on the length of their periods. We say that a run is:

**short** – if the length of its period does not exceed $|x_1|$,

**large** – if the length of its period exceeds $|x_2|$,

**medium** – otherwise.

Denote by $\rho_S(w)$, $\rho_M(w)$ and $\rho_L(w)$ the number of short, medium and large runs in the word $w$, respectively.

**Example 20.** Recall the word $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ from Example 1 and the set of its maximal repetitions. In this case we have:

- 10 short runs (periods $a$ and $ab$),
- 8 medium runs (periods $aba$ and $ababa$),
- 1 large run (the period $ababaab$),

see Figure 1 for comparison.

## Counting short runs

We start with the computation of the number of the *short* runs. These are the runs with the periods of the form $a$ or $a^+b$. Their number depends on the value of $\gamma_0$ and $\gamma_1$.

**Lemma 21** (Short Runs)**.**
*Let* $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *be a standard word. The number of* short *runs in* $w$ *is given by the formula:*

$$\rho_S(w) = \begin{cases} N_\gamma(2) + N_\gamma(3) - 1 & \text{for} \quad \gamma_0 = 1,\ \gamma_1 = 1 \\ 2\, N_\gamma(2) - odd(n) & \text{for} \quad \gamma_0 = 1,\ \gamma_1 > 1 \\ N_\gamma(1) + N_\gamma(3) - even(n) & \text{for} \quad \gamma_0 > 1,\ \gamma_1 = 1 \\ N_\gamma(1) + N_\gamma(2) & \text{for} \quad \gamma_0 > 1,\ \gamma_1 > 1 \end{cases}.$$

*Proof.* Short runs are the runs with the periods of the form $a$ or $a^k b$. We estimate the number of runs with the periods of each type separately.

**Case 1:** runs with the periods of the form $a$.

First assume that $\gamma_0 > 0$. Every run with the period $a$ in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ equals $a^{\gamma_0}$ or $a^{\gamma_0+1}$ and is followed by the single letter $b$. Due to Lemma 5 every such run in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ corresponds to the letter $a$ in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$. Hence in this case we have $N_\gamma(1)$ runs with the period $a$.

Assume now that $\gamma_0 = 1$. In this case the word $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ consists of the blocks of the two types: $ab$ or $aab$ and only the blocks of the second type include the runs with the period $a$. Due to Lemma 5 every such run in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ corresponds to the letter $b$ followed by the letter $a$ in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$, hence the number of such runs equals the number of blocks $ba$ in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$.

Recall that for an even length of the directive sequence $|(\gamma_1, \dots, \gamma_n)|$ ($n$ is even) the word $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$ ends with $ba$ and in this case the number of runs with the period $a$ in in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$ equals the number of the letters $b$ in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$, hence $N_\gamma(2)$. For an odd length of the directive sequence $|(\gamma_1, \dots, \gamma_n)|$ ($n$ is odd) the word $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$ ends with $ab$ and the last letter $b$ does not correspond to a run in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$. In this case, the number of runs with the period $a$ in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ is one less than the number of the letters $b$ in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$, hence $N_\gamma(2) - 1$. Finally the whole case can be summarized as:

$$\begin{cases} N_\gamma(2) - odd(n) & \text{for} \quad \gamma_0 = 1 \\ N_\gamma(1) & \text{for} \quad \gamma_0 > 1 \end{cases} .$$

**Case 2:** runs with the periods of the form $a^k b$.

Notice that, due to the equation (2) and Lemma 5, the runs with the periods $a^{\gamma_0} b$ and $a^{\gamma_0+1} b$ in in the word $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ correspond to the runs with the periods $a$ in the word $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$. Similar reasoning as above shows that the number of such in the word $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ equals:

$$\begin{cases} N_\gamma(3) - even(n) & \text{for} \quad \gamma_1 = 1 \\ N_\gamma(2) & \text{for} \quad \gamma_1 > 1 \end{cases} .$$

Combining the results from the two above cases we conclude the thesis of the lemma. □

## Counting medium runs

Recall that *medium* runs are the maximal repetitions with the periods $x_1^k x_0$ for $0 < k < \gamma_1$ and $x_2$, where $x_i$ are as in the equation (1). Observe that the medium runs appear in the standard words generated by the directive sequences of the length at least 3. We have to consider two cases: the directive sequences of the length 3 and the longer directive sequences. The value of $\gamma_0$ does not affect the number of the medium runs, hence to simplify the calculations we assume in further proofs that $\gamma_0 = 1$. We start with counting the medium runs in the standard words generated by the directive sequences of the length greater than 3.

**Lemma 22** (Medium runs, $\mathbf{n \geqslant 3}$)**.**
*Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and $n \geqslant 3$. The number of* medium *runs in $w$ is given by the formula:*

$$\rho_M(w) \;=\; N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1.$$

The thesis of the lemma is the consequence of the following stronger claim:

**Claim 23.** *Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. There are:*

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

**(2)** *$N_\gamma(3)$ runs with the period $x_2$.*

*Proof.* **Point (1)**
The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ has the form:

$$(ab)^{\alpha_1} a (ab)^{\alpha_2} a \ldots (ab)^{\alpha_s} a\; ab \qquad \text{or} \qquad (ab)^{\alpha_1} a (ab)^{\alpha_2} a \ldots (ab)^{\alpha_s} a,$$

where $0 < \alpha_i \leqslant (\gamma_1 + 1)$ and $s = N_\gamma(2)$, because, due to Lemma 5, every factor $(ab)^{\alpha_i} a$ in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ corresponds to the letter $a$ in $\mathrm{Sw}(\gamma_2, \ldots, \gamma_n)$. For example, see Figure 3, the word $\mathrm{Sw}(1, 4, 2, 2)$ has the form

$$\mathrm{Sw}(1, 4, 2, 2) \;=\; (ab)^4 a (ab)^4 a (ab)^5 a (ab)^4 a (ab)^5 a.$$

Each pair of neighboring factors: $(ab)^{\alpha_i} a \cdot (ab)^{\alpha_{i+1}} a$ produces $\gamma_1 - 1$ runs with the period $(ab)^k a$ for each $0 < k < \gamma_1$. In the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ we have $N_\gamma(2) - 1$ such pairs and therefore $(N_\gamma(2) - 1)(\gamma_1 - 1)$ medium runs with the periods $x_1^k x_0$ for $0 < k < \gamma_1$.

**Point (2)**
The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ can be represented as a sequence of concatenated words $x_1$ and $x_2$ and has the form:

$$(a): \quad x_2^{\alpha_1} x_1 x_2^{\alpha_2} x_1 \ldots x_2^{\alpha_s} x_1 x_2 \qquad \text{or} \qquad (b): \quad x_2^{\beta_1} x_1 x_2^{\beta_2} x_1 \ldots x_2^{\beta_s} x_1.$$

For example the word $\mathrm{Sw}(1, 4, 2, 2)$ has the decomposition $x_2^2 x_1 x_2^2 x_1 x_2$, see Figure 3.

First assume the case $(a)$. Each run with the period $x_2$ has the form $x_2^k x_1$. By the definition of standard words the factor $x_1 x_2$ has $x_2$ as a prefix. Therefore, the number of such runs in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ equals the number of factors $x_1$ in the decomposition mentioned above, which, due to Lemma 5, corresponds to the number of the letters $b$ in $\mathrm{Sw}(\gamma_2, \ldots, \gamma_n)$, namely $N_\gamma(3)$.

Assume now the case $(b)$. The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ has the suffix $x_1$ but in this case we have $\beta_s \geqslant 2$. Hence the number of runs with the period $x_2$ is the same as in the previous case. □

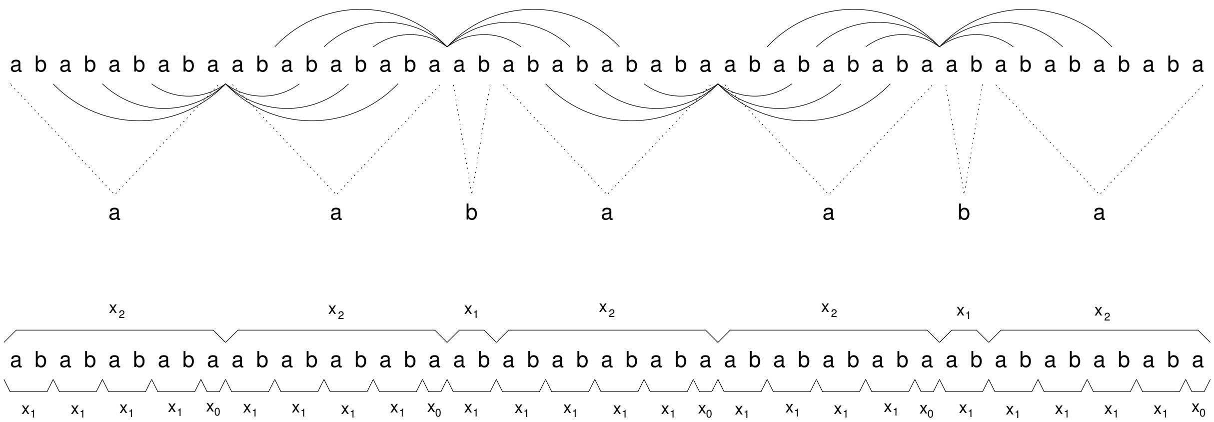


***Figure 3:*** *The structure of runs with the periods* $|x_1| < p < |x_2|$ *for the word* $\mathrm{Sw}(1, 4, 2, 2)$ *and its decomposition into words* $x_1$*,* $x_2$ *and* $x_0$*,* $x_1$*.*

*Proof of the Lemma 22.* Summing up the formulas from the points (1) and (2) of Claim 23 we obtain:

$$\begin{aligned} \rho_M &= \big(N_\gamma(2) - 1\big)\ (\gamma_1 - 1) + N_\gamma(3) \\ &= \big(\gamma_1 N_\gamma(2) + N_\gamma(3)\big)\ -\ N_\gamma(2) - \gamma_1 + 1 \\ &= N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1 \end{aligned}$$

and this completes the proof of the lemma. □

The structure of the medium runs in standard words defined by the directive sequences of the length 3 is slightly different.

**Lemma 24** (Medium runs, n=2)**.**
*Let* $w = \mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2)$ *be a standard word. The number of* medium *runs in* $w$ *is given by the formula:*

$$\rho_M(w) \;=\; N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1 - unary(\gamma_2)$$

*Proof.* The proof for the case $\gamma_2 > 1$ follows the same argumentation as in the proof of Lemma 22.

In the case $\gamma_2 = 1$ the word $\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2)$ has the decomposition

$$\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2) \;=\; \big(a^{\gamma_0} b\big) a \cdot a^{\gamma_0} b \;=\; x_2 \cdot x_1 .$$

There is no run with the period $x_2$, and we have to subtract 1 from the number of the factors $x_1$ in this case. □

## The recurrence for large runs

Recall that the run is called *large* if it has the period of the length greater than $|x_2|$, where $x_2$ is as in the equation (1). We reduce the problem of counting the large runs to counting the medium runs, using the morphic representation of the standard words.

Let $h$ be a morphism and let $v = a_1 a_2 \dots a_k$ be the word of the length $k$. The morphism $h$ defines the partition of the word $w = h(v)$ into segments $h(a_1)$, $h(a_2)$,..., $h(a_t)$. These segments are called the *h-blocks*. We say that a factor $x$ of the word $w$ is *synchronized* with the morphism $h$ in $w$ if and only if each occurrence of $x$ in $w$ starts at the beginning of some $h$-block and ends at the end of some $h$-block. Observe that every factor in $w$ that is synchronized with $h$ corresponds to some factor in $v$, hence the morphism $h$ preserves the structure of the factors that are synchronized with it.

**Example 25.** Let $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ and $v = \mathrm{Sw}(2, 1, 3, 1)$ be standard words and $h_0$ be the morphism defined as:

$$h_0 : \begin{cases} a \longrightarrow ab \\ b \longrightarrow a \end{cases} .$$

Recall that

$$\begin{aligned} \mathrm{Sw}(1, 2, 1, 3, 1) &= h_0\big(\mathrm{Sw}(2, 1, 3, 1)\big), \\ \mathrm{Sw}(1, 2, 1, 3, 1) &= ababaababababaababababaabababaababaab, \\ \mathrm{Sw}(2, 1, 3, 1) &= aabaaabaaabaaabaaba. \end{aligned}$$

The factors $w[6..8] = aba$ and $w[13..17] = abaab$ are not synchronized with the morphism $h_0$, because both of them ends in the middle some $h_0$-block. The factor $w[22..28]$ starts at the beginning of some $h_0$-block and ends at the end of some $h_0$-block, hence is synchronized with the morphism $h_0$. Moreover it corresponds with the factor $v[13..16] = aaba$, see Figure 4 for comparison.

**Lemma 26** (Synchronization Lemma)**.**
*The periods of the* large *runs in the word* $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ *are synchronized with the morphism* $h_0$.

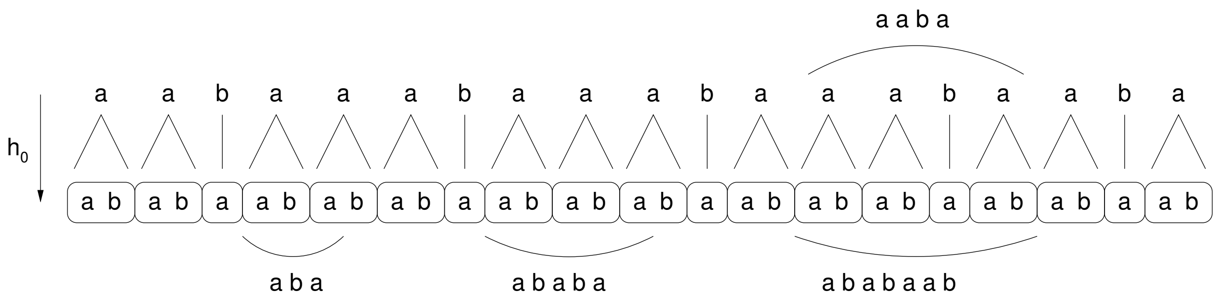


***Figure 4:*** *The periods of the medium runs $x_1x_0 = aba$ and $x_2 = ababa$ do not synchronize with the morphism $h_0$ in the word* $\mathrm{Sw}(1,2,1,3,1)$*, while the period of the large run $x_3 = ababaab$ is synchronized with $h_0$ and corresponds to the medium run with the period $x_1x_0 = aaba$ in the word* $\mathrm{Sw}(2,1,3,1)$*.*

*Proof.* Let $h_0$ be the morphism defined as

$$h_0 : \begin{cases} a & \longrightarrow & a^{\gamma_0}b \\ b & \longrightarrow & a \end{cases} .$$

Due to Lemma 5 we have

$$\mathrm{Sw}(\gamma_0,\ldots,\gamma_n) \; = \; h_0\big(\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)\big).$$

Moreover, $h_0$ determines the partition of $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ into $h_0$-blocks of the form $a^{\gamma_0}b$ and $a$, see Figure 4 for the partition of $\mathrm{Sw}(1,2,1,3,1)$.

Recall that the period of each large run in the word $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ is of the form $x_i^k x_{i-1}$, where $0 \leqslant k < \gamma_i$ and $i \geqslant 2$. By the definition of standard words the factor $x_i^k x_{i-1}$ starts with $a^{\gamma_0}b$, hence at the beginning of some $h_0$-block.

For even $i \geqslant 2$ the subword $x_i^k x_{i-1}$ ends with $x_1 = a^{\gamma_0}b$, hence at the end of some $h_0$-block, and is obviously synchronized with $h_0$.

For odd $i \geqslant 2$ the factor $x_i^k x_{i-1}$ ends with

$$x_3 \cdot x_2 \; = \; x_2^{\gamma_2}x_1 \cdot x_1^{\gamma_1}x_0 \; = \; x_2^{\gamma_2} \cdot (a^{\gamma_0}b)^{\gamma_1+1}a.$$

First assume that $x_i^k x_{i-1}$ is followed by the block $a^{\gamma_0}b$. The single letter $a$ at the end of $x_i^k x_{i-1}$ is then the whole $h_0$-block and $x_i^k x_{i-1}$ is synchronized with the morphism $h_0$.

Assume now that $x_i^k x_{i-1}$ ends with $(a^{\gamma_0}b)^{\gamma_1+1}a$ and is followed by $(a^{\gamma_0-1}b)$, namely it ends in the middle of some $h_0$-block. In this case we have the occurrence of the factor $(a^{\gamma_0}b)^{\gamma_1+2}$ in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$, which is reduced by the morphism $h_0^{-1}$ to the factor $a^{\gamma_1+2}b$ in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$. By the definition the standard word $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$ can include only the blocks of the two types: the short block – $a^{\gamma_1}b$ and the long block – $a^{\gamma_1+1}b$, hence we have the contradiction and the proof is complete. □

The following lemma, which is a direct conclusion from Synchronization lemma, allows us to reduce the problem of counting the large runs in the word $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ to those in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$.

**Lemma 27** (Recurrence Lemma)**.**
*Let $w = \mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ and $v = \mathrm{Sw}(\gamma_1, \dots, \gamma_n)$ be standard words. The number of* large *runs in $w$ is given by the recurrence*

$$\rho_L(w) \;=\; \rho_L(v) \;+\; \rho_M(v).$$

*Proof.* The synchronization lemma implies that the morphism defined as in the equation (2) preserve the structure of the long runs in standard words. Recall that $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ is reduced by $h_0^{-1}$ to $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$. Therefore, every large run $\alpha$ in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ corresponds to some run $\beta$ in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$.

Due to Lemma 19 the period of the run $\alpha$ is of the form $x_i^k x_{-1}$, where $0 < k \leqslant \gamma_i$ and $i \geqslant 2$. The corresponding run $\beta$ is either large (for $i = 2$) or medium (for $i = 2$). Hence to compute all large runs in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ it is sufficient to compute all large and medium runs in $\mathrm{Sw}(\gamma_1, \dots, \gamma_n)$. □

The thesis of the next lemma gives us the compact formula for the number of the medium and the large runs in standard words.

**Lemma 28** (Large Runs)**.**
*Let $w = \mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ be a standard word. We have*

$$\rho_L(w) \;+\; \rho_M(w) \;=\; N_\gamma(1) \;+\; n - 1 \;-\; (\gamma_1 + \ldots + \gamma_n) \;-\; unary(\gamma_n).$$

*Proof.* Due to the formulas from Lemma 22 and Lemma 24 and the recurrence from Lemma 27 we have

$$\begin{aligned}
\rho_L(w) \;+\; \rho_M(w) \;&=\; \sum_{i=0}^{n-2} \rho_M\big(\mathrm{Sw}(\gamma_i, \dots, \gamma_n)\big) \\
&=\; \Big(N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1\Big) \;+ \\
&\qquad\qquad \vdots \\
&\quad\; \Big(N_\gamma(n-2) - N_\gamma(n-1) - \gamma_{n-2} + 1\Big) \;+ \\
&\quad\; \Big(N_\gamma(n-1) - N_\gamma(n) - \gamma_{n-1} + 1 - unary(\gamma_n)\Big).
\end{aligned}$$

Taking into account that $N_\gamma(n) = \gamma_n$ the above formula can be written as

$$\rho_L(w) \;+\; \rho_M(w) \;=\; N_\gamma(1) \;+\; (n-1) \;-\; (\gamma_1 + \ldots + \gamma_n) \;-\; unary(\gamma_n),$$

which concludes the thesis. □

Now we are ready to prove the formula for the number of runs in standard words given by the equation (5).

*Proof of Theorem 12.* Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. Recall that we divide the set of all runs in $w$ into three disjoint subsets of *short*, *medium* and *large* runs, depending on the length of their periods. The number $\rho(w)$ of runs in $w$ is then given as:

$$\rho(w) \;=\; \rho_S(w) \;+\; \rho_M(w) \;+\; \rho_L(w),$$

where $\rho_S(w)$, $\rho_M(w)$ and $\rho_L(w)$ denote the number of short, medium and large runs in $w$ respectively.

Due to Lemma 21 the number of short runs in $w$ is given by the formula:

$$\rho_S(w) = \begin{cases} N_\gamma(2) + N_\gamma(3) - 1 & \text{for} \quad \gamma_0 = 1,\ \gamma_1 = 1 \\ 2\,N_\gamma(2) - odd(n) & \text{for} \quad \gamma_0 = 1,\ \gamma_1 > 1 \\ N_\gamma(1) + N_\gamma(3) - even(n) & \text{for} \quad \gamma_0 > 1,\ \gamma_1 = 1 \\ N_\gamma(1) + N_\gamma(2) & \text{for} \quad \gamma_0 > 1,\ \gamma_1 > 1 \end{cases}.$$

Moreover, due to Lemma 28, the number of large and medium runs in $w$ is given as:

$$\rho_L(w) \;+\; \rho_M(w) \;=\; N_\gamma(1) \;+\; n-1 \;-\; (\gamma_1 + \ldots + \gamma_n) \;-\; unary(\gamma_n).$$

Finally, combining the above formulas, we have:

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

where:

$$\begin{aligned} A &= N_\gamma(2) = |\mathrm{Sw}(\gamma_2, \gamma_3, \ldots, \gamma_n)|_a, \\ B &= N_\gamma(3) = |\mathrm{Sw}(\gamma_3, \gamma_4, \ldots, \gamma_n)|_a, \\ \Delta(\gamma) &= n - 1 - (\gamma_1 + \ldots + \gamma_n) - unary(\gamma_n). \end{aligned}$$

This completes the proof of the theorem. □

# 7 Final remarks

The aim of this paper was to study problems related to maximal repetitions for one of the most thoroughly investigated class of strings in combinatorics on words – the standard Sturmian words. We have presented the formulas for the numbers of runs along with the detailed analysis of their asymptotic behaviour. The complete understanding of their combinatorial structure for a large class of complicated words is a step towards a better understanding of this problem in general.

The maximal repetition ratio 0.8 for standard words has been first discovered by us doing experiments with very long strings. Similarly, we were tuning many intermediate formulas with the assistance of the computer. Our algorithms for computing the number of runs in standard words is an example of the very fast computation on highly compressed texts in linear time with respect to the size of their compressed representation.

| n | All binary words | | | Standard words | | |
|---|---|---|---|---|---|---|
| | $\rho(n)$ | $\rho(n)/n$ | Example word | $\rho(n)$ | $\rho(n)/n$ | Example word |
| 5 | 2 | 0.4 | aaabb | 1 | 0.2 | ababa |
| 6 | 3 | 0.5 | aabaab | 1 | 0.1667 | aaaaba |
| 7 | 4 | 0.5714 | aabaabb | 3 | 0.4286 | aabaaba |
| 8 | 5 | 0.625 | aabbaabb | 3 | 0.375 | abaababa |
| 9 | 5 | 0.5555 | aaabbaabb | 3 | 0.3333 | aaabaaaba |
| 10 | 6 | 0.6 | aabaabbaab | 4 | 0.4 | aabaabaaba |
| 11 | 7 | 0.6364 | aabaabbaabb | 5 | 0.4545 | aabaaabaaba |
| 12 | 8 | 0.6667 | aabaabbaabaa | 6 | 0.5 | ababaababaab |
| 13 | 8 | 0.6154 | aaabaabbaabaa | 7 | 0.5385 | abaababaababa |
| 14 | 10 | 0,7143 | aabaabbaabaabb | 5 | 0.3571 | abaabaabaababa |
| 15 | 10 | 0.6667 | aaabaabbaabaabb | 6 | 0.4 | aabaaabaaabaaba |
| 16 | 11 | 0.6875 | aabaabbaabaabbaa | 7 | 0.4375 | abababaababababaab |
| 17 | 12 | 0.7059 | aabaababbabaababb | 9 | 0.5294 | ababaababaababaab |
| 18 | 13 | 0.7222 | aabaabbaabaabbaabb | 10 | 0.5556 | aabaaabaabaaabaaba |
| 19 | 14 | 0.7368 | aabaabbaabaabbaabaa | 11 | 0.5789 | ababaababababaababaab |
| 20 | 15 | 0.75 | aababaababbabaababaa | 8 | 0.4 | abababababaabababababaab |
| 21 | 15 | 0.7143 | aaababaababbabaababaa | 13 | 0.619 | abaababaabaababaababa |
| 22 | 16 | 0.7273 | aabaababaababbabaababb | 12 | 0.5454 | ababaababaababaababaab |
| 23 | 17 | 0.7391 | aabaababaababbabaababaa | 13 | 0.5652 | abaababaababaababaabaab |
| 24 | 18 | 0.75 | aabaabbaabaabbabbaabbabb | 13 | 0.5417 | aabaabaaabaabaaabaabaaba |
| 25 | 19 | 0.76 | aabaabbaabaaabaabbaabaabb | 14 | 0.56 | aabaaabaabaaabaabaaabaaba |
| 26 | 20 | 0.7692 | aababaababbabaababaababbab | 15 | 0.5769 | ababaababababaabababaababaab |
| 27 | 21 | 0.7778 | aabaababaababbabaababaababb | 16 | 0.5926 | aabaabaaabaabaabaaabaabaaba |
| 28 | 22 | 0.7857 | aababaababbabaababaababbabaa | 16 | 0.5714 | abaababaababaababaababaabaab |
| 29 | 23 | 0.7931 | aababaababbabaababaababbababb | 18 | 0.6207 | ababaababaababababaababaababaab |
| 30 | 24 | 0.8 | aababbabaababbababbabaababbaba | 18 | 0.6 | abaabaababaabaabaababaabaabaab |
| 31 | 25 | 0.8065 | aababaababbabaababaababbabaabab | 20 | 0.6452 | ababaababababaababaababababaababaab |

***Table 1:*** *The comparison of the maximal number of runs and the repetition ratio for the class of all binary words and the class of standard words of a given length.*

## References


[1] M. Crochemore and L. Ilie. Analysis of maximal repetitions in strings. In *Proceedings of the 32nd International Conference on Mathematical Foundations of Computer Science*, volume 4708 of *Lecture Notes in Computer Science*, pages 465–476. Springer, 2007.

[2] M. Crochemore, L. Ilie, and L. Tinta. Towards a solution of the "runs" conjecture. In *Proceedings of the 19th annual symposium on Combinatorial Pattern Matching*, volume 5029 of *Lecture Notes in Computer Science*, pages 290–302. Springer, 2008.

[3] D. Damanik and D. Lenz. Powers in Sturmian sequences. *European Journal of Combinatorics*, 24(4):377–390, 2003.

[4] F. Franek, A. Karaman, and W. F. Smyth. Repetitions in Sturmian strings. *Theoretical Computer Science*, 249(2):289–303, 2000.

[5] F. Franek, R. J. Simpson, and W. F. Smyth. The maximum number of runs in a string. In *Proceedings of 14th Australian Workshop on Combinatiorial Algorithms*, pages 26–35, 2003.

[6] F. Franek and Q. Yang. An asymptotic lower bound for the maximal number of runs in a string. *Jnternational Journal of Foundations of Computer Science*, 19(1):195–203, 2008.

[7] 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.

[8] R. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 596–604. IEEE Computer Society, 1999.

[9] R. M. Kolpakov and G. Kucherov. On maximal repetitions in words. In *Proceedings of 12th International Symposium on Fundamentals of Computation Theory*, volume 1684 of *Lecture Notes in Computer Science*, pages 374–385. Springer, 1999.

[10] K. Kusano, W. Matsubara, A. Ishino, H. Bannai, and A. Shinobara. New lower bound for the maximum number of runs in a string. *Computing Research Repository*, abs/0804.1214, 2008.

[11] N. Pytheas Fogg. *Substitutions in Dynamics, Arithmetics and Combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer, 2002.

[12] W. Rytter. The structure of subword graphs and suffix trees of Fibonacci words. *Theoretical Computer Science*, 363(2):211–223, 2006.