Computing the Number of Cubic Runs in Standard Sturmian Words

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

[1] Department of Mathematics, Computer Science and Mechanics,
University of Warsaw, Warsaw, Poland
rytter@mimuw.edu.pl
[2] Faculty of Mathematics and Informatics,
Nicolaus Copernicus University, Toruń, Poland
martinp@mat.umk.pl

Abstract. The *standard Sturmian* words are extensively studied in combinatorics of words. They are enough complicated to have many interesting properties and at the same time they are highly compressible. In this paper we present compact formulas for the number $\rho^{(3)}$ of cubic runs in any standard word. We show also that

$$\lim_{|w|\to\infty} \frac{\rho^{(3)}(w)}{|w|} = \frac{5\Phi+3}{13\Phi+9} \approx 0.36924841$$

and present the sequence of strictly growing standard words achieving this limit. The exact asymptotic ratio is here irrational, contrary to the situation of squares and runs in the same class of words. Furthermore we design an efficient algorithm computing the number of cubic runs in standard words in linear time with respect to the size of the compressed representation (recurrences) describing the word. The explicit size of the word can be exponential with respect to this representation. This is yet another example of a very fast computation on highly compressible texts.

Keywords: standard Sturmian words, repetitions, cubic runs, algorithms

1 Introduction

Repetitions in strings are important in combinatorics on words and many practical applications, see for instance [6], [11], [19] and [20]. The structure of repetitions is almost completely understood for the class of Fibonacci words, see [15], [17], [24], however it is not well understood for general words.

Runs are repetitions in which the period repeats at least twice. Highly repetitive segments, in which the repetitions ratio is at least 3, called the *cubic runs*, were introduced and studied in [10].

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

A *maximal repetition* (a *run*, in short) in a word w is an interval $\alpha = [i..j]$ such that $w[i..j] = u^k v$ $(k \geq 2)$ 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 $|u|$). *Cubic runs*

⋆ The research supported by Ministry of Science and Higher Education of Poland, grant N N206 258035.

Figure 1. The structure of repetitions in the word $\mathrm{Sw}(1,2,1,3,1)$. There are 19 runs and 4 cubic runs (marked in **bold**).

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

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

5 runs with the period $|ab|$ (including 3 cubic runs):

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

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 $|ababa|$:

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

and 1 (cubic) run with the period $|ababaab|$:

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

All together we have 19 runs and 4 cubic runs, see Figure 1 for comparison.

Denote by $\rho(w)$ and $\rho^{(3)}(w)$ the number of runs and cubic runs in the word w, and by $\rho(n)$ and $\rho^{(3)}(n)$ the maximal number of runs and cubic runs in the words of length n respectively. The most interesting and open conjecture about maximal repetitions is:

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

In 1999 Kolpakov and Kucherov (see [16]) showed that the number $\rho(w)$ of runs in a string w is $O(|w|)$, but the exact multiplicative constant coefficient is still unknown. The best known results related to the value of $\rho(n)$ are

$$0.944575712\, n \;\leq\; \rho(n) \leq\; 1.029\, n.$$

The upper bound is by [8], [9] and the lower bound is by [13], [14], [18], [27]. The best known results related to $\rho^{(3)}(n)$ are (due to [10]):

$$0.41\, n \;\leq\; \rho^{(3)}(n) \;\leq\; 0.5\, n.$$

For the class $\mathcal{S}$ of standard Sturmian words there are known exact formulas for the number of runs and squares and their asymptotic behavior, see [2] and [22] for details. In this case we have

$$\lim_{n\to\infty} \frac{\rho(n)}{n} \;=\; 0.8.$$

This paper is devoted to the investigation of the structure of cubic runs in standard Sturmian words. We present the exact recurrence formulas for the number $\rho^{(3)}(w)$. Next we derive the algorithm computing $\rho^{(3)}(w)$ for any word $w \in \mathcal{S}$ in linear time with respect to the compressed representation of w, hence logarithmic time with respect to the length of the whole word w. We show also, that for any standard word w, we have

$$\rho^{(3)}(w_k) \;\leq\; 0.36924841\, |w|,$$

and construct the sequence $\{w_k\}$ of strictly growing standard words, for which we have

$$\lim_{k\to\infty} \frac{\rho^{(3)}(w_k)}{|w_k|} \;=\; \frac{5\Phi+3}{13\Phi+9} \;\approx\; 0.36924841.$$

Some useful applets related to problems considered in this paper can be found on the web site: `http://www.mat.umk.pl/~martinp/stringology/applets/`

2 Standard Sturmian words

Standard Sturmian words (standard words in short) are one of the most investigated class of strings in combinatorics on words, see for instance [1], [4], [5], [7], [19], [25], [26], [28] and references therein. They have very compact representations in terms of sequences of integers, which has many algorithmic consequences.

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

$$\begin{aligned} x_{-1} &= b, & x_0 &= a, \\ x_1 &= x_0^{\gamma_0} x_{-1}, & x_2 &= x_1^{\gamma_1} x_0, \\ &\vdots & &\vdots \\ x_n &= x_{n-1}^{\gamma_{n-1}} x_{n-2}, & x_{n+1} &= x_n^{\gamma_n} x_{n-1}, \end{aligned} \tag{1}$$

where $\mathrm{Sw}(\gamma) = x_{n+1}$.

A 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:

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

and finally

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

The special case of standard words are Fibonacci words, which are formed by repeated concatenation in the same way that the Fibonacci numbers are formed by repeated addition. By definition the are given by directive sequences of the form $\gamma = (1, 1, \ldots, 1)$ (n-th Fibonacci word F_n corresponds to a sequence of n ones).

Observe that 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 also 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 [23] for comparison.

Remark 3. Without loss of generality we consider here 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.

The number $N = |\mathrm{Sw}(\gamma)|$ is the (real) size of the word, while $(n+1) = |\gamma|$ can be thought as its compressed size. Observe that, by the definition of standard words, N is exponential with respect to n. Each directive sequence corresponds to a *grammar-based compression*, which consists in describing a given word by a context-free grammar G generating this (single) word. The size of the grammar G is the total length of all productions of G. In our case the size of the grammar is proportional to the length of the directive sequence.

3 Morphic reduction of standard words

The recurrent definition of standard words 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 γ a sequence of morphisms $\{h_i\}_{i=0}^{n}$, defined as:

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

Lemma 4. *For $0 \leq i \leq n$ 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. We will prove the above lemma by 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 $|\gamma| = k \geq 2$ and for directive sequences shorter than k the thesis holds. We have then:

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

which concludes the proof. □

Remark 5. As a direct conclusion from Lemma 4 we have that the standard word corresponding to the directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ is given as:

$$\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n) = h_0 \circ h_1 \circ \cdots \circ h_n(a). \tag{3}$$

The inverse morphism h_i^{-1} can be seen as a reduction of the word $\mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$ to the word $\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_n)$ and allows us to reduce the computation of cubic runs in $\mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$ to the same computation in $\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_n)$.

Denote by $|w|_a$ the number of occurrences of the letters a in the word w. We define the function, which will be useful in the rest of this paper. For a directive sequence $\gamma = (\gamma_0, \ldots, \gamma_n)$ and an integer $0 \le k \le n+1$ we define

$$N_\gamma(k) = |\mathrm{Sw}(\gamma_k, \gamma_{k+1}, \ldots, \gamma_n)|_a . \tag{4}$$

Moreover, for $k > n+1$, we define $N_\gamma(k) = 0$.

Remark 6. As a direct conclusion from the above definition, the equation (1) and the equation (2) we have that for $0 \le k \le n$ the numbers $N_\gamma(k)$ satisfy:

$$N_\gamma(k) = \gamma_k \, N_\gamma(k+1) + N_\gamma(k+2). \tag{5}$$

Example 7. Let $\gamma = (1, 2, 1, 3, 1)$ be a directive sequence. We have then

$$\begin{aligned}
\mathrm{Sw}(1,2,1,3,1) &= ababaabababaabababaababababaababaab & \quad N_\gamma(0) &= 19,\\
\mathrm{Sw}(2,1,3,1) &= aabaaabaaabaaabaaba & N_\gamma(1) &= 14,\\
\mathrm{Sw}(1,3,1) &= abababaab & N_\gamma(2) &= 5,\\
\mathrm{Sw}(3,1) &= aaaba & N_\gamma(3) &= 4,\\
\mathrm{Sw}(1) &= ab & N_\gamma(4) &= 1,\\
\mathrm{Sw}(\varepsilon) &= a & N_\gamma(5) &= 1.
\end{aligned}$$

Remark 8. In case of Fibonacci words the numbers $N_\gamma(k)$ are Fibonacci numbers:

$$N_\gamma(k) \;=\; |F_{n-k-1}| \;=\; f_{n-k-1}. \tag{6}$$

4 Formulas for the number of cubic runs

In this section we present and prove formulas for the number of cubic runs in any standard word, that depend only on its compressed representation – the directive sequence. The following zero-one functions for testing the parity of a nonnegative integer i will be useful to simplify those formulas:

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

We begin with the characterization of possible periods of cubic runs in standard words. The following lemma is a consequence of the very special structure of subword graphs (especially their compacted versions) of those words. See [3] and [25] for more information.

Lemma 9. *The period of each cubic run in the standard word* $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *is of the form* x_i*, where* x_i*'s are as in equation (1).*

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 a cubic run. We can use similar argumentation as in proof of Theorem 1 in [12]. The details are omitted in this version.

The main idea of the computation of cubic runs in a standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ is the partition of them into three separate categories depending on the length of their periods. We say that a cubic run is:

short – if it has the period of the form a or $a^k b$,
medium – if it has the period of the form x_2,
large – if it has the period of the form x_i, for $i > 2$.

Denote by $\rho_S^{(3)}(w)$, $\rho_M^{(3)}(w)$ and $\rho_L^{(3)}(w)$ the number of short, medium and large cubic runs in the word w, respectively. We will consider each type separately.

Example 10. Recall the word $w = \mathrm{Sw}(1,2,1,3,1)$ from Example 1. We have:

– 3 short cubic runs (period ab),
– no medium cubic run,
– 1 large cubic run (period $ababaab$),

see Figure 1 for comparison.

4.1 Short runs

We start with the computation of the *short* cubic runs. These are the cubic runs with the periods of the form a or $a^k b$. Their number depends on the values of γ_0 and γ_1.

Lemma 11. *The number $\rho_{S_1}^{(3)}$ of cubic runs with the period a in the standard word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ equals:*

$$\rho_{S_1}^{(3)}(w) \;=\; \begin{cases} 0 & \text{for } \gamma_0 = 1 \\ N_\gamma(2) - odd(n) & \text{for } \gamma_0 = 2 \\ N_\gamma(1) & \text{for } \gamma_0 > 2 \end{cases} . \qquad (7)$$

Proof. First assume that $\gamma_0 > 2$. Every cubic run with the period a in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ equals a^{γ_0} or a^{γ_0+1} and is followed by the single letter b. Due to Lemma 4 every such cubic run in $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ corresponds to the letter a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Hence in this case we have $N_\gamma(1)$ cubic runs with the period a.

Assume now that $\gamma_0 = 2$. In this case the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ consists of the blocks of the two types: aab and $aaab$. Only the blocks of the second type include the cubic run with the period a. Due to Lemma 4 every such cubic run in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ corresponds to the letter b followed by the letter a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Hence the number of such cubic runs equals the number of blocks ba in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$.

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

Finally assume that $\gamma_0 = 1$. In this case the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ consists of the blocks of the two types: ab and aab. None of them includes a cubic run with the period a, and this completes the proof. $\square$

Lemma 12. *The number $\rho^{(3)}_{S_2}$ of cubic runs with the period $a^k b$ in the standard word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ equals:*

$$\rho^{(3)}_{S_2}(w) = \begin{cases} 0 & \text{for } \gamma_1 = 1 \\ N_\gamma(3) - even(n) & \text{for } \gamma_1 = 2 \\ N_\gamma(2) & \text{for } \gamma_1 > 2 \end{cases}. \tag{8}$$

Proof. Notice that, due to equation (2) and Lemma 4, cubic runs with the periods of the form $a^k b$ in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ correspond to cubic runs with the period a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Similar reasoning as above establishes the desired formula. □

4.2 Medium runs

Recall that a cubic run is called *medium* if it has the period of the form x_2. Observe that medium cubic runs appear in standard words generated by 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 values of γ_0 and γ_1 does not affect the number of medium cubic runs, hence to simplify the calculations we can assume in further proofs that $\gamma_0 = \gamma_1 = 1$.

We start with counting medium runs in standard words generated by directive sequences of the length greater than 3.

Lemma 13. *Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and $n \geq 3$. The number of medium cubic runs in w equals:*

$$\rho^{(3)}_{M}(w) = \begin{cases} N_\gamma(4) - 1 & \text{for } \gamma_2 = 1 \\ N_\gamma(3) & \text{for } \gamma_2 \geq 2 \end{cases}. \tag{9}$$

Proof. We start with the assumption that $\gamma_2 > 2$. In this case every factor of the form $x_3 = x_2^{\gamma_2} x_1$ includes one cubic runs with the period x_2. Hence the number of such cubic runs equals the number of factors x_3 in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$, namely $N_\gamma(3)$ (due to Lemma 4).

Assume now that $\gamma_2 = 2$. The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ can be represented as a sequence of concatenated words x_3 and x_2 and has the form:

$$x_3^{\alpha_1} x_2 x_3^{\alpha_2} x_2 \cdots x_3^{\alpha_s} x_2 x_3 \qquad \text{or} \qquad x_3^{\beta_1} x_2 x_3^{\beta_2} x_2 \cdots x_3^{\beta_s} x_2.$$

Observe that $x_3 = x_2 x_2 x_1$ and every occurrence of x_3 in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ either follows some occurrence of x_2 or is followed by some occurrence of x_2. In the first case we have $x_2 \cdot x_3 = x_2 \cdot x_2 x_2 x_1$ and there is a cubic run with period x_2. In the second case we have $x_3 \cdot x_2 = x_2 x_2 x_1 \cdot x_2$, and there is also a cubic run with period x_2, since x_1 is a prefix of x_2. Therefore the number of medium cubic runs in this case equals the number of the factors x_3 in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$, namely $N_\gamma(3)$.

Finally assume that $\gamma_2 = 1$. The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ can be represented as a sequence of concatenated words x_3 and x_4 and has the form:

$$x_4^{\alpha_1} x_3 x_4^{\alpha_2} x_3 \cdots x_4^{\alpha_s} x_3 x_4 \qquad \text{or} \qquad x_4^{\beta_1} x_3 x_4^{\beta_2} x_3 \cdots x_4^{\beta_s} x_3.$$

We have $x_3 = x_2 x_1$ and $x_4 = x_2 x_1 \cdots x_2 x_1 \cdot x_2$. Therefore only the last one occurrence of x_4 in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ does not correspond to a cubic run with the period x_2 and we have $N_\gamma(4) - 1$ such cubic runs in this case. This completes the proof. □

Lemma 14. *The number of medium cubic runs in the word* $w = \mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2)$ *equals:*

$$\rho_M^{(3)}(w) \;=\; \begin{cases} 1 & \text{for } \gamma_2 > 2 \\ 0 & \text{for } \gamma_2 \le 2 \end{cases} . \tag{10}$$

Proof. We have $\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2) \;=\; x_2^{\gamma_2} x_1$. Hence there is only one medium run (with the period x_2) if $\gamma_2 > 2$ and no medium run otherwise. □

4.3 Large runs

Recall that a cubic run is called *large* if it has the period of the form x_i for $i > 2$, where x_i are as in the equation (1). We reduce the problem of counting large cubic runs to the one for counting medium cubic runs, using morphic representation of standard words introduced in previous section.

Let h be a morphism, $v = a_1 a_2 \cdots a_k$ be a word of the length k and let $w = h(v)$. The morphism h defines the partition of w 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 15. Let $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ and $v = \mathrm{Sw}(2, 1, 3, 1)$ be standard words and h_0 be a morphism defined as:

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

Recall that

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

The factors $w[6..8] = aba$ and $w[13..17] = abaab$ are not synchronized with h_0, because both of them end in the middle some h_0-block. From the other hand, the factor $w[22..28] = abábaab$ and all its occurrences in w (namely $w[1..7]$, $w[8..14]$ and $w[15..21]$) start at the beginning of some h_0-block and end at the end of some h_0-block. Hence the factor $w[22..28]$ is synchronized with the morphism h_0. Moreover it corresponds with the factor $v[13..16] = aaba$, see Figure 2 for comparison.

Lemma 16. *The periods of* large *cubic runs in the standard word* $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *are synchronized with the morphism* h_0.

Proof. Let h_0 be the morphism defined as

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

Figure 2. The factors *aba* and *ababa* do not synchronize with the morphism h_0 in the word $\mathrm{Sw}(1,2,1,3,1)$, while the factor *ababaab* (in fact the period of the large cubic run) is synchronized with h_0 and corresponds to the factor *aaba* in the word $\mathrm{Sw}(2,1,3,1)$.

Due to Lemma 4 we have $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n) = h_0(\mathrm{Sw}(\gamma_1,\ldots,\gamma_n))$. 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 2 for the partition of $\mathrm{Sw}(1,2,1,3,1)$.

Recall that the period of each large cubic run in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ is of the form x_i, where $i \geq 3$. By definition of standard words the factor x_i starts with $a^{\gamma_0}b$, hence at the beginning of some h_0-block.

For odd $i \geq 3$ the subword x_i 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 even $i \geq 3$ the factor x_i 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 is followed by the block $a^{\gamma_0}b$. The single letter a at the end of x_i is then the whole h_0-block and x_i is synchronized with the morphism h_0.

Assume now that x_i 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 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 Lemma 16, allows us to reduce the problem of counting large cubic runs in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ to counting large and medium cubic runs in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$.

Lemma 17. *Let $w = \mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ and $v = \mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$ be standard words. The number of* large *cubic runs in w is given by the recurrence*

$$\rho_L^{(3)}(w) \;=\; \rho_L^{(3)}(v) \;+\; \rho_M^{(3)}(v).$$

Proof. Lemma 16 implies that the morphism defined in the equation (2) preserves the structure of long cubic runs in standard words. Recall that the word $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ is reduced by h_0^{-1} to the word $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$. Therefore, every large cubic run α in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ corresponds to some cubic run β in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$.

Due to Lemma 9 the period of the cubic run α is of the form x_i, where $i \geq 3$. The corresponding cubic run β is either large (for $i > 3$) or medium (for $i = 3$). Hence to compute all large cubic runs in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ it is sufficient to compute all large and medium cubic runs in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$. □

The following theorem summarizes all the formulas developed above.

Theorem 18. *Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence, $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ and $w_i = \mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$, for $0 \le i \le n$, be standard words. The number of cubic runs in w is given as:*

$$\rho^{(3)}(w) \;=\; \rho^{(3)}_{S_1}(w) \;+\; \rho^{(3)}_{S_1}(w) \;+\; \sum_{i=0}^{n-2} \rho^{(3)}_{M}(w_i). \tag{11}$$

Proof. The thesis of the theorem follows by combining the formulas (7), (8), the formula (9) repeated $n-2$ times, and finally the formula (10). □

Example 19. Consider a directive sequence $\gamma = (1, 2, 1, 3, 1)$. We compute the number of cubic runs in $\mathrm{Sw}(1, 2, 1, 3, 1)$ using the formulas mentioned above. We have:

short cubic runs with period a:	0
short cubic runs with period $a^k b$:	$\lvert aaaba\rvert_a - 1 = 3$
medium cubic runs:	$\lvert ab\rvert_a - 1 = 0$
large cubic runs:	$\rho^{(3)}_M(2, 1, 3, 1) + \rho^{(3)}_M(1, 3, 1) = \lvert ab\rvert_a + 0 = 1$

Altogether there are 4 cubic runs, see Example 1 and Figure 1 for comparison.

4.4 Algorithm for computation of the number of cubic runs

The formulas investigated above allow us to develop an efficient algorithm computing the number of cubic runs in any standard Sturmian word.

Theorem 20. *Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $\mathrm{Sw}(\gamma)$ be a standard word. We can count the number of cubic runs in $\mathrm{Sw}(\gamma)$ in linear time with respect to the length of the directive sequence $|\gamma|$ (logarithmic time with respect to the length of the whole word $|\mathrm{Sw}(\gamma)|$).*

Proof. The formulas (7), (8), (9) and (10) for the number of cubic runs in a standard word $\mathrm{Sw}(\gamma)$ depend directly on the components of the directive sequence γ and the numbers $N_\gamma(k)$. We can compute the numbers $N_\gamma(n)$, $N_\gamma(n-1)$, …, $N_\gamma(1)$ by consecutive iteration of the equation (5). In each step i of the computation we remember the number of cubic runs related to the value of the γ_i. The number of iterations performed by the algorithm correspond directly to the length of the directive sequence, hence it has the time complexity $\mathrm{O}(|\gamma|)$. See Algorithm 1 for more details. □

5 Asymptotic behaviour of the number of cubic runs

This section is devoted to the computation of the asymptotic limit

$$\lim_{|w|\to\infty} \frac{\rho^{(3)}(w)}{|w|} \tag{12}$$

for $w \in \mathcal{S}$.

Algorithm 1: *Counting-Cubic-Runs*(Sw(γ))

$(x, y, cr) \longleftarrow (1, 0, 0)$;
if $\gamma_n > 2$ **then** $cr \longleftarrow cr + 1$;
for $i := n$ **to** 3 **do**
    $(x, y) \longleftarrow (\gamma_i \cdot x + y, x)$;
    if $\gamma_{i-1} \geq 2$ **then** $cr \longleftarrow cr + x$;
    else $cr \longleftarrow cr + y - 1$;
if $\gamma_1 = 2$ **then**
    $cr \longleftarrow cr + x$;
    if n *is even* **then** $cr \longleftarrow cr - 1$;
$(x, y) \longleftarrow (\gamma_2 \cdot x + y, x)$;
if $\gamma_1 > 2$ **then** $cr \longleftarrow cr + x$;
if $\gamma_0 = 2$ **then**
    $cr \longleftarrow cr + x$;
    if n *is odd* **then** $cr \longleftarrow cr - 1$;
if $\gamma_0 > 2$ **then** $cr \longleftarrow cr + \gamma_1 \cdot x + y$;
return cr;

Theorem 21. *Let $w = \mathrm{Sw}(\gamma)$ be a standard word. Then we have*

$$\rho^{(3)}(w) \leq 0.36924841\,|w|.$$

Moreover there is infinite and strictly growing sequence of standard words achieving this asymptotic limit.

Proof. To prove the above theorem we will construct a directive sequence corresponding to a standard word for which the number of cubic runs will be maximal in relation to their length.

Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma)$ be a standard word. The number of cubic runs with the period of the form a in w corresponds directly to the value γ_0, see equation (7). The word w consists of blocks of the two types: $a^{\gamma_0}b$ and $a^{\gamma_0+1}b$. For $\gamma_0 \geq 3$ every such block contains a desired cubic run, for $\gamma_0 = 2$ only the second type of blocks contains a short cubic run, and for $\gamma_0 < 2$ there is no short cubic run in w. Moreover, for $\gamma_0 > 3$ the number of considered cubic runs does not change while the length of the word increases significantly.
For $\gamma_0 = 2$ we have, by the equations (5) and (7):

$$|w| = (\gamma_1 + 1)\, N_\gamma(2) + 3\, \gamma_1\, N_\gamma(3) \qquad \text{and} \qquad \rho^{(3)}_{S_1}(w) = N_\gamma(2) \pm 1,$$

and for $\gamma_0 = 3$ we have:

$$|w| = (4\, \gamma_1 + 1)\, N_\gamma(3) + 4\, N_\gamma(3) \qquad \text{and} \qquad \rho^{(3)}_{S_1}(w) = \gamma_1\, N_\gamma(2) + N_\gamma(3).$$

Therefore for the change of the value γ_0 from 2 to 3 the increase of $\rho^{(3)}_{S_1}$ (namely: $(\gamma_1 - 1)\, N_\gamma(2) + N_\gamma(3)$) is significant in relation to the increase of the length of the whole word (namely: $\gamma_1\, N_\gamma(2) + \gamma_1\, N_\gamma(3)$). Hence $\gamma_0 = 3$ is the optimal value. It does not affect the number of cubic runs with longer periods, hence we assume in further discussion its optimal value.

The number of cubic runs with the period of the form $a^k b$ in w depends on the value of γ_1, see equation (8). Similar argumentation as above shows that γ_1 must be greater than 1 and no more than 3. For $\gamma_1 = 2$ we have, by the equations (5) and (8):

$$|w| = (9\gamma_2 + 4)N_\gamma(3) + 9N_\gamma(4) \qquad \text{and} \qquad \rho^{(3)}_{S_2}(w) = N_\gamma(3) \pm 1,$$

and for $\gamma_1 = 3$ we have:

$$|w| = (13\gamma_2 + 4)N_\gamma(3) + 13N_\gamma(4) \qquad \text{and} \qquad \rho^{(3)}_{S_2}(w) = \gamma_2 N_\gamma(3) + N_\gamma(4).$$

Therefore the change of the value of γ_1 from 2 to 3 increases the number of cubic runs by: $(\gamma_2 - 1)N_\gamma(3) + N_\gamma(4) \pm 1$ and at the same time increases the length of the word by: $4\,\gamma_2\,N_\gamma(3) + 4\,N_\gamma(4)$. Hence we conclude that $\gamma_1 = 2$ is the optimal value.

The number of medium cubic runs in the word w corresponds to the value of γ_2. It is easy to see that γ_2 must be at most 2, otherwise the length of the word increases significantly and the value $\rho^{(3)}_M(w)$ does not change.
For $\gamma_2 = 1$ we have, by the equations (5) and (9):

$$|w| = (13\gamma_3 + 9)N_\gamma(4) + 13N_\gamma(5) \qquad \text{and} \qquad \rho^{(3)}_M(w) = N_\gamma(4) - 1.$$

and for $\gamma_2 = 2$ we have:

$$|w| = (22\gamma_3 + 9)N_\gamma(4) + 22N_\gamma(5) \qquad \text{and} \qquad \rho^{(3)}_M(w) = \gamma_3 N_\gamma(4) + N_\gamma(5).$$

Therefore the change of the value of γ_2 from 1 to 2 increases the number of cubic runs by: $(\gamma_3 - 1)\,N_\gamma(4) + N_\gamma(5) + 1$ and at the same time increases the length of the word by: $9\,\gamma_3\,N_\gamma(4) + 9\,N_\gamma(5)$. Hence we conclude that $\gamma_2 = 1$ is the optimal value.

We compute large cubic runs by reduction of them to medium runs, see Lemma 17. Applying $n-2$ times the above argumentation for the medium cubic runs we conclude that optimal value of $\gamma_3, \gamma_4, \ldots, \gamma_{n-1}$ is also 1. Similarly for $\gamma_n > 1$ there is one additional long run whereas the length of the word increases more than two times.

We have shown above, that the maximal value of the quotient of the number of cubic runs to the length of the word is achieved by the standard words generated by directive sequences of the form $\gamma = (3, 2, 1, 1, \ldots, 1)$. Now we are ready to compute the value of the asymptotic limit from the equation (12).

Let us consider a sequence of standard words:

$$w_k \;=\; (3, 2, \underbrace{1, 1, 1, \ldots, 1}_{k}).$$

We have by definition of standard words and Remark 8:

$$|w_k| \;=\; 13\,N_\gamma(3) + 9\,N_\gamma(4) \;=\; 13\,f_{k-1} + 9\,f_{k-2},$$

and by Theorem 18 and Remark 8:

$$\rho^{(3)}(w_k) \;=\; 5\,N_\gamma(3) + 3\,N_\gamma(4) - k \pm 1 \;= 5\,f_{k-1} + 3\,f_{k-2}.$$

We have also that:

$$\lim_{n\to\infty} \frac{f_n}{f_{n-1}} \;=\; \Phi \;\approx 1.61803390,$$

hence

$$\lim_{k\to\infty} \frac{\rho^{(3)}(w_k)}{|w_k|} \;\approx\; \frac{5\Phi + 3}{13\Phi + 9} \;\approx\; 0.36924841,$$

and this completes the proof. ☐

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*, in Proceedings of the 13th international conference on Implementation and Applications of Automata, vol. 5148 of Lecture Notes in Computer Science, Springer, 2008, pp. 252–261.
3. P. Baturo, M. Piątkowski, and W. Rytter: *Usefulness of directed acyclic subword graphs in problems related to standard Sturmian words*. International Journal of Foundations of Computer Science, 20(6) 2009, pp. 1005–1023.
4. P. Baturo and W. Rytter: *Compressed string-matching in standard Sturmian words*. Theoretical Computer Science, 410(30–32) 2009, pp. 2804–2810.
5. J. Berstel: *Sturmian and Episturmian words: a survey of some recent results*, in Proceedings of the 2nd international conference on Algebraic informatics, vol. 4728 of Lecture Notes in Computer Science, Springer, 2007, pp. 23–47.
6. J. Berstel and J. Karhumaki: *Combinatorics on words: a tutorial*. Bulletin of the EATCS, 79 2003, pp. 178–228.
7. J. Berstel, A. Lauve, C. Reutenauer, and F. Saliola: *Combinatorics on Words: Christoffel Words and Repetitions in Words*, CRM monograph series, Providence, R.I: American Mathematical Society, 2009.
8. M. Crochemore and L. Ilie: *Analysis of maximal repetitions in strings*, in Proceedings of the 32nd International Conference on Mathematical Foundations of Computer Science, vol. 4708 of Lecture Notes in Computer Science, Springer, 2007, pp. 465–476.
9. 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, vol. 5029 of Lecture Notes in Computer Science, Springer, 2008, pp. 290–302.
10. M. Crochemore, C. S. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, and T. Walen: *On the maximal number of cubic runs in a string*, in Proceedings of the International Conference on Implementation and Applications of Automata, 2010, pp. 227–238.
11. M. Crochemore and W. Rytter: *Jewels of Stringology: text algorithms*, World Scientific, 2003.
12. D. Damanik and D. Lenz: *Powers in Sturmian sequences*. European Journal of Combinatorics, 24(4) 2003, pp. 377–390.
13. 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, 2003, pp. 26–35.
14. 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) 2008, pp. 195–203.
15. C. S. Iliopoulos, D. Moore, and W. F. Smyth: *A characterization of the squares in a Fibonacci string*. Theoretical Computer Science, 172(1–2) 1997, pp. 281–291.
16. 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, IEEE Computer Society, 1999, pp. 596–604.
17. R. M. Kolpakov and G. Kucherov: *On maximal repetitions in words*, in Proceedings of 12th International Symposium on Fundamentals of Computation Theory, vol. 1684 of Lecture Notes in Computer Science, Springer, 1999, pp. 374–385.
18. 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.
19. M. Lothaire: *Algebraic Combinatorics on Words*, vol. 90 of Encyclopedia of mathematics and its application, Cambridge University Press, 2002.
20. M. Lothaire: *Applied Combinatorics on Words*, vol. 105 of Encyclopedia of Mathematics and its Application, Cambridge University Press, 2005.
21. M. Piątkowski: *Stringological applets* . http://www.mat.umk.pl/˜martinp/stringology/applets.
22. M. Piątkowski and W. Rytter: *Asymptotic behaviour of the maximal number of squares in standard Sturmian words*, in Proceedings of the 14-th Prague Stringology Conference, Czech Technical University, 2009, pp. 237–248, accepted to International Journal of Foundations of Computer Science.

23. N. Pytheas Fogg: *Substitutions in Dynamics, Arithmetics and Combinatorics*, vol. 1794 of Lecture Notes in Mathematics, Springer, 2002.
24. W. Rytter: *The structure of subword graphs and suffix trees of Fibonacci words.* Theoretical Computer Science, 363(2) 2006, pp. 211–223.
25. M. Sciortino and L. Zamboni: *Suffix automata and standard Sturmian words*, in Proceedings of the 11th International Conference on Developments in Language Theory, vol. 4588 of Lecture Notes in Computer Science, Springer, 2007, pp. 382–398.
26. J. Shallit: *Characteristic words as fixed points of homomorphisms*, Tech. Rep. CS-91-72, University of Waterloo, Department of Computer Science, 1991.
27. J. Simpson: *Modified Padovan words and the maximum number of runs in a word.* Australian Journal of Combinatorics, 46 2010, pp. 129–145.
28. H. Uscka-Wehlou: *Digital lines, Sturmian words, and continued fractions*, PhD thesis, Department of Mathematics, Upspsala University, 2009.