The Number of Cubes in Sturmian Words

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

[1] Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University, Toruń, Poland
marcin.piatkowski@mat.umk.pl

[2] Department of Mathematics, Computer Science and Mechanics,
University of Warsaw, Warsaw, Poland
rytter@mimuw.edu.pl

Abstract. We design an efficient algorithm computing the number of distinct cubes in a standard Sturmian word given by its directive sequence (the special type of recurrences). The algorithm runs in linear time with respect to the size of the compressed representation (recurrences) describing the word, though the explicit size of the word can be exponential with respect to this representation. We give the explicit compact formula for the number of cubes in any standard word derived from the structural properties of runs (maximal repetitions). Fibonacci words are the most known subclass of standard Sturmian words. It is known that the ratio of the number of cubes to the size for Fibonacci words is asymptotically equal to $\frac{1}{\phi^3} \approx 0.2361$, where $\phi = \frac{\sqrt{5}+1}{2}$. We show a class of standard Sturmian words for which this ratio is much higher and equals $\frac{3\phi+2}{9\phi+4} \approx 0.36924841$. An extensive experimentation suggests that this value is optimal.

Keywords: Standard Sturmian words, cubes, repetitions, algorithm

1 Introduction

Problems related to finding repetitions in strings are fundamental in combinatorics on words and have many practical applications (data compression, computational biology, pattern matching, etc.), see for instance [5], [8], [12] and [13]. The structure of repetitions is almost completely understood for the class of Fibonacci words, see [10], [11], [16], however it is not well understood for general words.

The most important type of repetitions are *runs* (maximal repetition), which form a compact representation of all repetitions in a word. Formally, a *run* 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 $|u|$).

In this paper we consider cubes: the nonempty words of the form $\alpha = x^3$. The length of x is called the *base* of the cube and denoted by $base(\alpha)$. 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 (min-period, in short) of w will be denoted by $period(w)$.

* 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.

Example 1. Let $\alpha = (abab)^3$ be a cube. In this case we have:

$$base(\alpha) \;=\; 4, \quad period(\alpha) \;=\; 2.$$

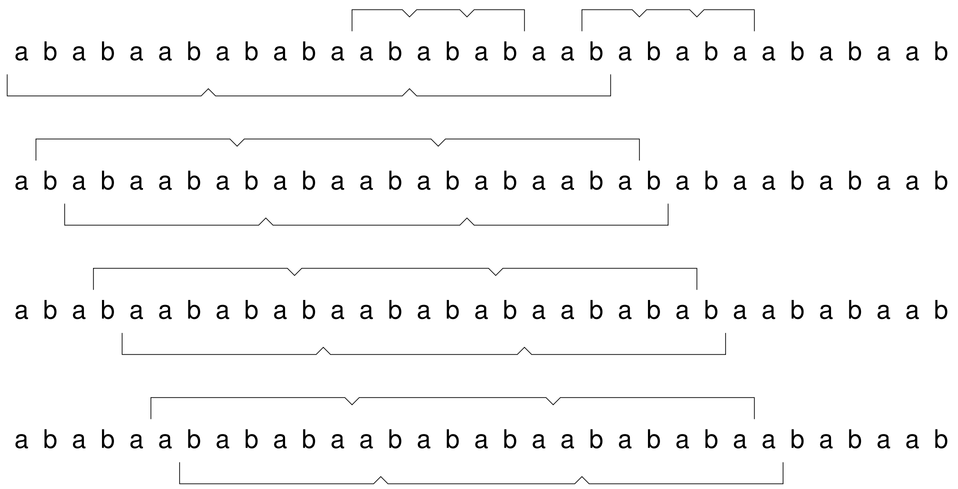

Figure 1. The structure of distinct cubes in the example binary word $w = ababaabababaabababaabababaababaab$.

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. In this paper we are interested in counting *distinct cubes*, hence we identify cubes with the same base, but perhaps multiple occurrences.

Example 2. Let w be as in Figure 1. There are 9 cubes:

$$ab \cdot ab \cdot ab, \quad ba \cdot ba \cdot ba, \quad ababaab \cdot ababaab \cdot ababaab,$$
$$babaaba \cdot babaaba \cdot babaaba, \quad abaabab \cdot abaabab \cdot abaabab,$$
$$baababa \cdot baababa \cdot baababa, \quad aababab \cdot aababab \cdot aababab,$$
$$abababa \cdot abababa \cdot abababa, \quad bababaa \cdot bababaa \cdot bababaa.$$

The *standard Sturmian* words are extensively studied in combinatorics on words. They are enough complicated to have many interesting properties and at the same time they are highly compressible. Due to their regularity, many problems are much easier for such strings compared with the general case. There are known exact formulas for the number of runs, cubic runs (i.e. runs in which the period repeats at least three times) and squares in standard words along with their *density ratio* (i.e. the asymptotic quotient of the maximal number of considered repetitions by the length of the word). See [2], [15] and [14] for details.

This paper is devoted to the investigation of the structure and the number of cubes in standard Sturmian words. Denote by $cubes(w)$ the number of cubes in a word w. We present exact formulas for $cubes(w)$ in any standard word w. We show also the algorithm, which computes the number of cubes in any standard word in linear time with respect to the size of its compressed representation – the directive sequence –

hence in time logarithmic with respect to the length of the word. We show also a class of standard words reach in cubes and prove that for this class of strings the density ratio of distinct cubes equals $\frac{3\phi+2}{9\phi+4} \approx 0.36924841$, where $\phi = \frac{\sqrt{5}+1}{2}$. An extensive computer experimentation suggests that this value is optimal.

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], [6], [12], [17], [18], [19] and references therein. They have very compact representations in terms of sequences of integers, which has many algorithmic consequences.

The *directive sequence* is the 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:

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

where $\mathrm{Sw}(\gamma) = x_{n+1}$. For simplicity we denote $q_i = |x_i|$.

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 3.
Consider the directive sequence $\gamma = (1, 2, 1, 3, 1)$. We have $\mathrm{Sw}(\gamma) = x_5$, where:

$$\begin{array}{llll}
x_{-1} = b & & & q_{-1} = 1 \\
x_0 = a & & & q_0 = 1 \\
x_1 = (x_0)^1 \cdot x_{-1} & = a \cdot b & & q_1 = 2 \\
x_2 = (x_1)^2 \cdot x_0 & = ab \cdot ab \cdot a & & q_2 = 5 \\
x_3 = (x_2)^1 \cdot x_1 & = ababa \cdot ab & & q_3 = 7 \\
x_4 = (x_3)^3 \cdot x_2 & = ababaab \cdot ababaab \cdot ababaab \cdot ababa & & q_4 = 26 \\
x_5 = (x_4)^1 \cdot x_3 & = ababaababaababaababaababa \cdot ababaab & & q_5 = 33
\end{array}$$

Without loss of generality we consider here the standard Sturmian words starting with the letter a, therefore we assume that $\gamma_0 > 0$. The words starting with the letter b can be considered similarly.

Remark 4.
The special kind of standard words are well known Fibonacci words. They are formed by repeated concatenation in the same way that the Fibonacci numbers are formed by repeated addition. By the definition Fibonacci words are standard words 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).

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.

2.1 The structure of cubes in standard words

The main idea of the computation of distinct cubes in a standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ is the partition of them into separate categories depending on the length of their periods. In this section we define the concepts of the i-partition of standard words and the generative run, which will be crucial in cubes enumeration. The following fact is a direct consequence of recurrent definition of standard words.

Fact 1
Every standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *can be represented as a sequence of concatenated words* x_i *and* x_{i-1}*, and has the form:*

$$x_i^{\alpha_1}\, x_{i-1}\, x_i^{\alpha_2}\, x_{i-1} \ldots x_i^{\alpha_s}\, x_{i-1}\, x_i \qquad \text{or} \qquad x_i^{\beta_1}\, x_{i-1}\, x_i^{\beta_2}\, x_{i-1} \ldots x_i^{\beta_s}\, x_{i-1},$$

where $\alpha_k,\ \beta_k \in \{\gamma_i,\ \gamma_i + 1\}$*, and* x_i *are as in equation (1).*

Such a decomposition of a standard word w is called the *i-partition* of w. The block x_i is then the *repeating block* and the block x_{i-1} – the *single block*.

Example 5. Recall the word $\mathrm{Sw}(1,2,1,3,1)$ from Example 3. We have then:

$\mathrm{Sw}(1,2,1,3,1)$	$ababaababababaababababaabababaababaab$
$1-$ partition	$x_1^2\, x_0\, x_1^3\, x_0\, x_1^3\, x_0\, x_1^3\, x_0\, x_1^2\, x_0\, x_1$
$2-$ partition	$x_2\, x_1\, x_2\, x_1\, x_2\, x_1\, x_2^2\, x_1$
$3-$ partition	$x_3^3\, x_2\, x_3$
$4-$ partition	$x_4\, x_3$

See Figure 2 for comparison.

The following facts characterize the possible bases of distinct cubes in standard words. Their thesis are consequence of the very special structure of the subword graphs (especially their compacted versions) of those words. For more information on the subword graphs of standard words see for instance [3] and [17].

Lemma 6 (See [9]).
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *be a standard sturmian word and* v *be a factor of* w *such that* $|x_i| \le |v| < |x_{i+1}|$*, where* x_i *are as in equation (1). Then:*

1. *There is at most one position in* x_i *(respectively* x_{i-1}*) such that any occurrence of* v *in* w *which starts in some* x_i*-block (respectively* x_{i-1}*-block) of the* i*-partition of* w *has to start at this particular position in* x_i *(respectively* x_{i-1}*).*
2. *If* v *can start at position* k *in* x_i *and at position* l *in* x_{i-1} *(*k *and* l *are uniqye by 1), then we have* $k = l$*.*

Lemma 7.
The base of each cube in the standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *has the length* $k \cdot |x_i|$, *where* $0 < k < \gamma_i$ *and* x_i*'s are as in equation (1). The min-period of each cube equals* q_i *for some* $0 \le i \le n$.

Proof. Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and $v = u^3$ be a cube in w such that $|x_i| \le |u| < |x_{i+1}|$. We denote $v = u^{(1)} \cdot u^{(2)} \cdot u^{(3)}$ to be able to refer to each occurrence of u in v. Due to Lemma 6, the factors $u^{(1)}$, $u^{(2)}$ and $u^{(3)}$ start at the same (within the block) position l of some blocks of the i-partition of w. The distance between two consecutive l position could be either $k \cdot |x_i|$ or $k \cdot |x_i| + |x_{i-1}|$. Recall that every occurrence of x_{i-1} block is separated by γ_i or $\gamma_i + 1$ occurrences of the x_i block. Since $|v| < |x_{i+1}|$ and $|x_{i+1}| = \gamma_i |x_i| + |x_{i-1}|$ we have $k < \gamma_i$ and the only possible base of v is $|u| = k \cdot |x_i|$, for $0 < k < \gamma_i$. Moreover, every standard word x_i is primitive, hence the minimal period of v has the length $|x_i| = q_i$.

We say that a cube is of *type* i if its min-period equals q_i. The number of distinct cubes of the type i in the word $\mathrm{Sw}(\gamma)$ is denoted by $\pi_i(\gamma)$.

Example 8. Let $\mathrm{Sw}(1, 2, 1, 3, 1) = ababaababababaabababaababababaababaab$ be a standard word. We have 2 cubes of the type 1 and 7 cubes of type 3, see Figure 1 and Example 2 for comparison.

For each $0 \le i \le n$ let *gen-run*(i) be the value (as a word) of the longest run with minimal period equal to q_i. It is called a *generative run* of type i (see Figure 2 for an example).

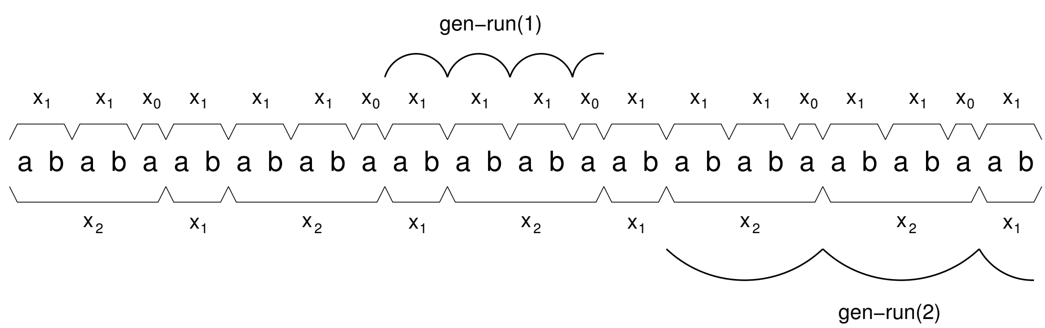

Figure 2. The 1-partition (above) and 2-partition (below) of the word $\mathrm{Sw}(1, 2, 1, 3, 1)$. We have *gen-run*$(1) = x_1^3 x_o$, *gen-run*$(2) = x_2^2 x_1$. The first generative run produces two different cubes, the second produces no cubes.

Lemma 9 (See [3]).
Each generative run of the type i *is of the form:*

$$\textit{gen-run}(i) \;=\; (x_i)^{\alpha} \cdot y,$$

where y *is a proper prefix of* x_i.

Example 10. Let $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ (see Figure 1). The generative run of type 1 has the form *gen-run*$(1) = (x_1)^3 x_0$ and generates two cubes $(ab)^3$ and $(ba)^3$. On the other hand the generative run of type 2 has the form *gen-run*$(2) = (x_2)^2 x_1$ and does not generate any cube.

3 Formula and algorithm for counting the number of cubes

In this section we present and prove formulas for the number of distinct cubes in any standard word, that depend only on its compressed representation – the directive sequence. The following zero-one function for testing the value of the remainder of the division by 3 of a nonnegative integer x will be useful to simplify those formulas:

$$\mathbf{3}_k(x) \;=\; \begin{cases} 1 & \text{if } x \bmod 3 = k \\ 0 & \text{if } x \bmod 3 \neq k \end{cases}.$$

Recall that $q_i = |x_i|$ and π_i is the number of cubes of the type i in the word $\mathrm{Sw}(\gamma)$.

Theorem 11 (Main-Formulas).
The number of cubes in standard word $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *is given by the formula:*

$$cubes(\gamma_0, \gamma_1, \ldots, \gamma_n) \;=\; \sum_{i=0}^{n} \pi_i(\gamma_0, \gamma_1, \ldots, \gamma_n),$$

where:

$$\textbf{(1)}\ (\ i \in [0, n-3]\ ) \;\Rightarrow\; \pi_i(\gamma) \;=\; \left\lfloor \frac{\gamma_i + 1}{3} \right\rfloor q_i + \boldsymbol{3}_1(\gamma_i) \cdot \Big(q_{i-1} - 1\Big)$$

$$\textbf{(2)}\ \pi_{n-2}(\gamma) \;=\; \begin{cases} \left\lfloor \dfrac{\gamma_{n-2} + 1}{3} \right\rfloor q_{n-2} + \boldsymbol{3}_1(\gamma_{n-2}) \cdot \Big(q_{n-3} - 1\Big) & if\ \ \gamma_n > 1 \\ \left\lfloor \dfrac{\gamma_{n-2}}{3} \right\rfloor \cdot q_{n-2} + \boldsymbol{3}_2(\gamma_{n-2}) \cdot \Big(q_{n-3} + 1\Big) & if\ \ \gamma_n = 1 \end{cases}$$

$$\textbf{(3)}\ \pi_{n-1}(\gamma) \;=\; \left\lfloor \frac{\gamma_{n-1}}{3} \right\rfloor \cdot q_{n-1} + \boldsymbol{3}_2(\gamma_{n-1}) \cdot \Big(q_{n-2} - 1\Big)$$

$$\textbf{(4)}\ \pi_n(\gamma) \;=\; \left\lfloor \frac{\gamma_n - 1}{3} \right\rfloor \cdot q_n + \boldsymbol{3}_0(\gamma_n) \cdot \Big(q_{n-1} + 1\Big)$$

The proof of the above theorem is a matter of Section 4. Let us see some examples.

Example 12. Let $\mathrm{Sw}(1, 2, 1, 3, 1)$ be a standard word. Using formulas from Theorem 11 we have:

$$\pi_0(1,2,1,3,1) \;=\; \pi_2(1,2,1,3,1) \;=\; \pi_4(1,2,1,3,1) \;=\; 0$$

$$\pi_1(1,2,1,3,1) \;=\; 2 \qquad\qquad \pi_3(1,2,1,3,1) \;=\; 7$$

and finally

$$cubes(1,2,1,3,1) \;=\; 9.$$

See Example 8 and Figure 1 for comparison.

The number of cubes in Fibonacci words is given by the formula

$$cubes(F_n) \;=\; f_{n-3} - n + 2,$$

where f_k denotes the k-th Fibonacci number (see [7] for the proof). As the next example we derive this formula using results from Theorem 11.

Example 13. Recall that the n-th Fibonacci word F_n is defined as:

$$F_n = \mathrm{Sw}(\underbrace{1, 1, \ldots, 1}_{n}).$$

Hence

$$(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}) \;=\; (1, 1, \ldots, 1),$$

and for each $i = 0, 1, \ldots, n-4$, we have

$$\pi_i(1, 1, \ldots, 1) \;=\; f_{i-1} - 1.$$

Moreover

$$\pi_{n-3}(1, 1, \ldots, 1) \;=\; \pi_{n-2}(1, 1, \ldots, 1) \;=\; \pi_{n-1}(1, 1, \ldots, 1) \;=\; 0.$$

Taking into account the identity

$$\sum_{i=-1}^{k} f_i \;=\; f_{k+2} - 1$$

we have

$$\begin{aligned} cubes(\underbrace{1, \ldots, 1}_{n}) &= \sum_{i=0}^{n-4}(f_{i-1} - 1) &&= \sum_{i=-1}^{n-5}(f_i - 1) \\ &= f_{n-3} - 1 - (n-3) &&= f_{n-3} - n + 2 \end{aligned}$$

Theorem 14.
The number of cubes in a standard word $\mathrm{Sw}(\gamma)$ *can be computed in linear time with respect to the length of the directive sequence* γ *(which is at least logarithmically smaller than the real length of the whole word* $\mathrm{Sw}(\gamma)$*).*

Proof.
The formulas for the number of cubes in a standard word $\mathrm{Sw}(\gamma)$ depend directly on the components of the directive sequence γ and the numbers q_i (namely $|x_i|$), see Theorem 11. Recall that, by the equation (1), we have

$$q_{i+1} \;=\; \gamma_i \cdot q_i + q_{i+1},$$

hence every number q_i can be computed by iteration of the equation (1) i times. We can compute the numbers q_0, q_1, ..., q_n consecutively and at each step i of the computation remember the number of cubes related to the value of q_i. The number of iterations performed by the algorithm corresponds directly to the length of the directive sequence, hence it has the time complexity $O(|\gamma|)$. See Algorithm 1 for details.

Algorithm 1: *Cubes*(Sw(γ))

$cubes \longleftarrow 0$;

$q_{-1} \longleftarrow 1$;
$q_0 \longleftarrow 0$;

for $k := 0$ **to** n **do**
    $q_k \;\longleftarrow\; \gamma_k\, q_{k-1} + q_{k-2}$;
    update *cubes* depending on the value of γ_k;

return *cubes*;

4 Proof of Theorem 11

Let us denote by $\widehat{w}$ the word w with two last letters removed and by $\widetilde{w}$ the word w with two last letters exchanged.

The following fact will be useful in proofs and can be shown by a simple induction, see for instance [12].

Lemma 15.
Let x_i be as in equation (1) and $i > 1$. Then:

(a) $x_{i-1} \cdot x_i \;=\; x_i \cdot \widetilde{x_{i-1}}$

(b) *The length of the longest prefix of $x_{i-1}x_i$ with period q_i equals $|x_i\widehat{x_{i-1}}|$.*

Example 16. Recall the word Sw(1, 2, 1, 3, 1) from Example 3. We have $x_2 = ababa$, $x_1 = ab$ and $\widetilde{x_1} = ba$. Therefore

$$x_1 \cdot x_2 \;=\; ab \cdot ababa \;=\; ababa \cdot ba \;=\; x_2 \cdot \widetilde{x_1}.$$

Let us fix throughout this section a standard word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$. We show each point of Theorem 11 separately.

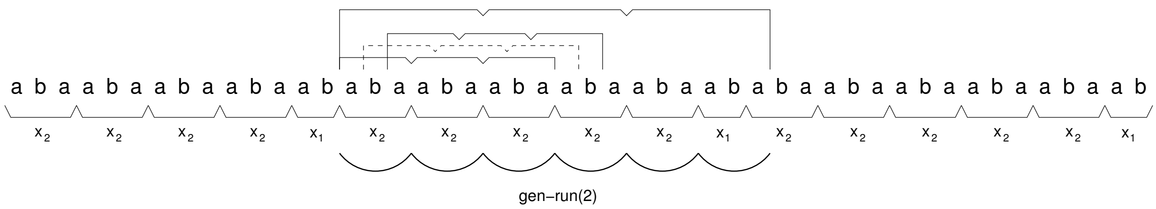

Figure 3. The illustration of Lemma 17: the structure of *gen-run*(2) and cubes of type 2 in the word Sw(1, 2, 4, 1, 2).

Lemma 17.

(a) $i \leq n - 3 \implies \textit{gen-run}(i) \;=\; (x_i)^{\gamma_i+2} \cdot \widehat{x_{i-1}}$

(b) *The point (1) from Theorem 11 is correct.*

Proof.

Point (a)

Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard word. Due to Fact 1 its i-partition has the form:

$$x_i^{\alpha_1}\, x_{i-1}\, x_i^{\alpha_2}\, x_{i-1} \ldots x_i^{\alpha_s}\, x_{i-1}\, x_i \qquad \text{or} \qquad x_i^{\beta_1}\, x_{i-1}\, x_i^{\beta_2}\, x_{i-1} \ldots x_i^{\beta_s}\, x_{i-1},$$

where $\alpha_k,\ \beta_k\ \in\ \{\gamma_i,\ \gamma_i + 1\}$. Let us consider the inner factor

$$v \ =\ (x_i)^{\gamma_i+1} \cdot x_{i-1} \cdot x_i.$$

Due to Lemma 15 the longest periodic prefix of v with period of the length $|x_i|$ (namely the generative run of type i) has the form:

$$(x_i)^{\gamma_i+2} \cdot \widehat{x_{i-1}}$$

and this concludes the proof of this point.

Point (b)

It is obvious that every cube of type i must be derived from the generative run of type i. Therefore, we have cubes with the bases: $q_i,\ 2 \cdot q_i,\ \ldots,\ \left\lfloor \frac{\gamma_i+1}{3} \right\rfloor \cdot q_i$. Each of them could be shifted to the right $q_i - 1$ times producing altogether q_i distinct cubes with the same base.

Moreover, if $\gamma_i \bmod 3\ =\ 1$, the subword $v = (x_i)^{\gamma_i+2}$ is also a cube. According to the structure of the generative run, v could be shifted to the right $q_{i-1} - 2$ times producing altogether $q_{i-1} - 1$ distinct cubes with the same base. See Figure 3 for an example of this case.

Finally the number of cubes of type i is given as:

$$\pi_i(\gamma)\ =\ \left\lfloor \frac{\gamma_i + 1}{3} \right\rfloor \cdot q_i + \mathbf{3}_1(\gamma_i) \cdot \big(q_{i-1} - 1\big).$$

This completes the proof of the lemma.

Lemma 18.

(a) *gen-run*$(n-2)\ =\ \begin{cases} (x_{n-2})^{\gamma_{n-2}+2} \cdot \widehat{x_{n-3}} & \textit{for}\ \ \gamma_n > 1 \\ (x_{n-2})^{\gamma_{n-2}+1} \cdot x_{n-3} & \textit{for}\ \ \gamma_n = 1 \end{cases}$

(b) *The point (2) from Theorem 11 is correct.*

Proof.

Point (a)

The case of $\gamma_n > 1$ folows the same argumentation as in proof of Lemma 17, hence we can assume $\gamma_n = 1$. The standard word $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_{n-1}, 1)$ has the form:

$$w\ =\ \overbrace{(\underbrace{x_{n-2} \cdots x_{n-2}}_{\gamma_{n-2}} \cdot x_{n-3}) \cdots (\underbrace{x_{n-2} \cdots x_{n-2}}_{\gamma_{n-2}} \cdot x_{n-3})}^{\gamma_{n-1}} \cdot x_{n-2} \cdot (\underbrace{x_{n-2} \cdots x_{n-2}}_{\gamma_{n-2}} \cdot x_{n-3}).$$

The longest run with the period of the length q_i (namely the generative run of type i) is the suffix of w:

$$(x_{n-2})^{\gamma_{n-2}+1} \cdot x_{n-3}$$

and this concludes the proof of this point.

Point (b)

Similarly as in the proof of Point (a) we assume $\gamma_n = 1$. Every cube of type $n-2$ is derived from the generative run of type $n-2$. Therefore we have q_{n-2} cubes for each base length: q_i, $2 \cdot q_i$, ..., $\lfloor \frac{\gamma_{n-2}}{3} \rfloor \cdot q_i$. Moreover, if $\gamma_{n-2} \bmod 3 = 2$, the factor $(x_{n-2})^{\gamma_{n-2}+1}$ is also a cube, which could be shifted q_{n-3} times. Hence we have $q_{n-3}+1$ additional cubes with the base $\frac{\gamma_{n-2}+1}{3} \cdot q_{n-2}$. See Figure 4 for an example of this case.

Finally we have

$$\pi_{n-2}(\gamma) \;=\; \left\lfloor \frac{\gamma_{n-2}}{3} \right\rfloor \cdot q_{n-2} + \mathbf{3}_2(\gamma_{n-2}) \cdot \Big(q_{n-3}+1\Big)$$

and the proof is complete.

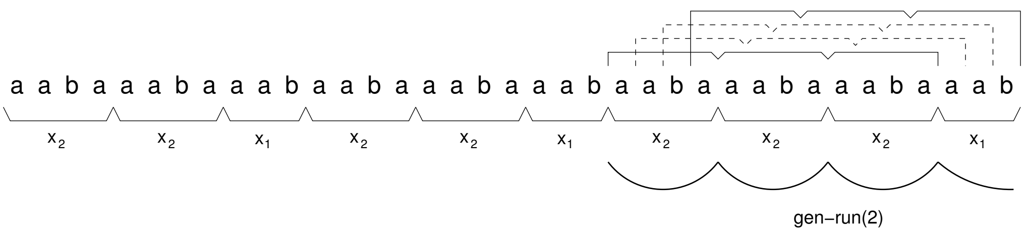

Figure 4. The illustration of the Lemma 18: the structure of *gen-run*(2) and cubes of the type 2 (i.e. type $n-2$) in the word $\mathrm{Sw}(2,1,2,2,1)$.

Lemma 19.

(a) *gen-run*$(n-1) \;=\; (x_{n-1})^{\gamma_{n-1}+1} \cdot \widehat{x_{n-2}}$

(b) *The point (3) from Theorem 11 is correct.*

Proof.

Point (a)

By definition the word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ has the form:

$$w \;=\; \overbrace{(\underbrace{x_{n-1}\cdots x_{n-1}}_{\gamma_{n-1}} \cdot x_{n-2}) \cdot (\underbrace{x_{n-1}\cdots x_{n-1}}_{\gamma_{n-1}} \cdot x_{n-2}) \cdots (\underbrace{x_{n-1}\cdots x_{n-1}}_{\gamma_{n-1}} \cdot x_{n-2})}^{\gamma_n} \cdot x_{n-1}.$$

Due to Lemma 15 the longest periodic factor of w with period of the length $|x_{n-1}|$ (namely the generative run of type $n-1$) has the form:

$$(x_{n-1})^{\gamma_{n-1}+1} \cdot \widehat{x_{n-2}}$$

and this concludes the proof of this point.

Point (b)

According to the structure of *gen-run*$(n-1)$ we have q_{n-1} cubes for each base length: q_{n-1}, $2 \cdot q_{n-1}$, $\ldots$, $\lfloor \frac{\gamma_{n-1}}{3} \rfloor \cdot q_{n-1}$. Moreover, if $\gamma_{n-1} \bmod 3 \ = \ 2$, the factor $(x_{n-1})^{\gamma_{n-1}+1}$ is also a cube, which could be shifted $q_{n-1} - 2$ times. Hence we have $q_{n-2} - 1$ additional cubes with the base $\frac{\gamma_{n-1}+1}{3} \cdot q_{n-1}$. See Figure 5 for an example of this case.

Finally we have

$$\pi_{n-1}(\gamma_0, \gamma_1, \ldots, \gamma_n) \ = \ \left\lfloor \frac{\gamma_{n-1}}{3} \right\rfloor \cdot q_{n-1} + \mathbf{3}_2(\gamma_{n-1}) \cdot \Big(q_{n-2} - 1\Big).$$

and this concludes the proof.

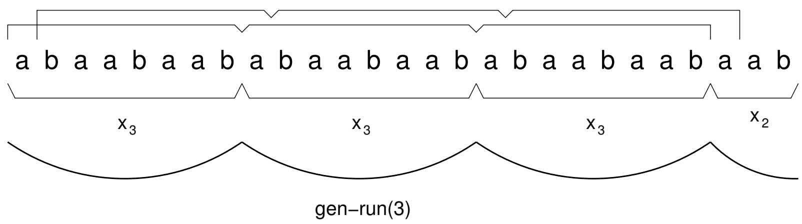

Figure 5. The illustration of the Lemma 19: the structure of *gen-run*(3) and cubes of the type 3 (i.e. type $n-1$) in the word $\mathrm{Sw}(1, 1, 2, 2, 1)$.

Lemma 20.

(a) *gen-run*$(n) \ = \ (x_n)^{\gamma_n} \cdot x_{n-1}$

(b) *The point (4) from Theorem 11 is correct.*

Proof.

Point (a)

By definition the word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ has the form:

$$w \ = \ \underbrace{x_n \cdot x_n \cdots x_n}_{\gamma_n} \cdot x_{n-1}.$$

Since x_{n-1} is the prefix of x_n, the value of generative run of type n is the whole word w.

Point (b)

According to the structure of *gen-run*(n) we have q_n cubes for each base length: q_n, $2 \cdot q_n$, $\ldots$, $\lfloor \frac{\gamma_n - 1}{3} \rfloor \cdot q_n$. Moreover, if $\gamma_n \bmod 3 \ = \ 0$, the factor $(x_n)^{\gamma_n}$ is also a cube, which could be shifted q_{n-1} times. Hence we have $q_{n-1} + 1$ additional cubes with the base $\frac{\gamma_n}{3} \cdot q_n$. See Figure 6 for an example of this case.

Finally we have

$$\pi_n(\gamma_0, \gamma_1, \ldots, \gamma_n) \ = \ \left\lfloor \frac{\gamma_n - 1}{3} \right\rfloor \cdot q_n + \mathbf{3}_0(\gamma_n) \cdot \Big(q_{n-1} + 1\Big)$$

and this completes the proof.

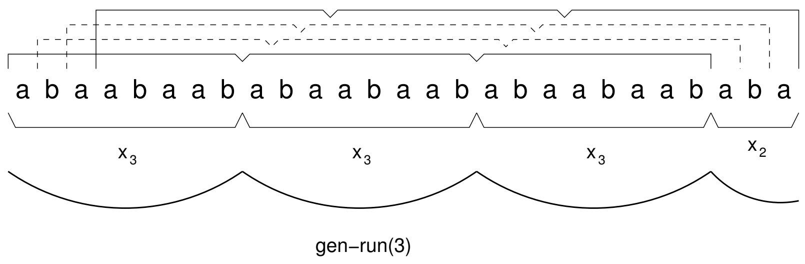

Figure 6. The illustration of Lemma 20: the structure of *gen-run*(3) and cubes of the type 3 (i.e. type n) in the word $\mathrm{Sw}(1,1,2,3)$.

Proof (of Theorem 11).
The sets of distinct cubes of type i and distinct cubes of type j are disjoint for $i \neq j$. Therefore, the thesis of Theorem 11 follows by summing up the formulas for number of cubes of all types from Lemma 17, Lemma 18, Lemma 19 and Lemma 20.

5 Standard words with large number of cubes

In this section we show the family of standard words rich in cubes. Experimental evidence shows that asymptotically this family achieves the highest ratio of the number of cubes to the length of the word.

Theorem 21.
Let $\gamma^k = (\underbrace{1,\ldots,1}_{k}, 2, 3, 1)$ *be a directive sequence and* $w_k = \mathrm{Sw}(\gamma^k)$ *be a standard word. We have:*

$$\lim_{k\to\infty} \frac{cubes(w_k)}{|w_k|} = \frac{3\phi+2}{9\phi+4} \approx 0.36924841\ldots$$

where $\phi = \frac{\sqrt{5}+1}{2}$.

Proof. Denote by f_k the k-th Fibonacci numer:

$$f_{-1} = 1; \quad f_0 = 1; \quad f_1 = 2; \quad f_2 = 3; \quad f_4 = 5; \quad \ldots$$

By definition of standard words we have

$$|\mathrm{Sw}(\gamma_k)| = 9f_k + 4f_{k-1}.$$

According to Theorem 11 we have

$$\pi_i = f_{i-1} - 1 \quad \text{for} \quad i = 0, \ldots, k-1,$$

$$\pi_k = f_{k-1} + 1, \qquad \pi_{k+1} = 2f_k + f_{k-1}, \qquad \pi_{k+2} = 0.$$

Taking into account the well known identity

$$\sum_{i=-1}^{k} f_k = f_{k+2} - 1$$

we obtain

$$\sum_{i=0}^{k-1}(f_{i-1}-1) \;=\; \sum_{i=-1}^{k-2}(f_i-1) \;=\; f_k-k-1.$$

Altogether we have

$$cubes\Big(\mathrm{Sw}(\gamma_k)\Big) = 3f_k + 2f_{k-1} - k.$$

Denote by

$$\beta_k \;=\; \frac{f_k}{f_{k-1}}.$$

Then we have

$$\lim_{k\to\infty}\beta_k \;=\; \phi.$$

Therefore

$$\begin{aligned}\lim_{k\to\infty}\frac{cubes(w_k)}{|w_k|} &= \lim_{k\to\infty}\frac{3f_k+2f_{k-1}-k}{9f_k+4f_{k-1}}\\ &= \lim_{k\to\infty}\frac{3\beta_k+2-O(1)}{9\beta_k+4}\\ &= \frac{3\phi+2}{9\phi+4}\\ &\approx 0.36924841...\end{aligned}$$

Remark 22. The extensive experimentation strongly suggests that the coefficient $\frac{3\phi+2}{9\phi+4}$ from the last theorem equals also the upper bound for the asymptotic ratio between the number of cubes and size of a standard words, see Table 1 for some examples.

Directive sequence	**Length**	**Cubes**	**Ratio**
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)$	46368	10926	0.2356366459
$(3,3,3,3,3,3,3,3,3,3,1,1,1,1,1,1,1,1,1,2,3,1)$	138069388	50878017	0.3684959985
$(5,5,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,3,1)$	1028890	379883	**0.3692163399**
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,3,1)$	125574	46349	**0.3690971061**
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,3,2)$	222491	50529	0.2271058155
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,1)$	96917	28637	0.2954796372
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,4,1)$	154231	46349	0.3005167573
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,1)$	81790	28637	0.3501283775
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,3,1)$	169358	61475	0.3629884623
$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,4,3,1)$	213142	72421	0.3397781761
$(\underbrace{1,1,1,1,1,1,1,1,1,1,\ldots,1,1,1,1,1,1,1}_{40},2,3,1)$	3073549228	1134903130	**0.3692483985**

Table 1. The example standard words with the number of cubes and cubes density.

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. 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.
5. J. Berstel and J. Karhumaki: *Combinatorics on words: a tutorial*. Bulletin of the EATCS, 79 2003, pp. 178–228.
6. 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.
7. 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.
8. M. Crochemore and W. Rytter: *Jewels of Stringology: text algorithms*, World Scientific, 2003.
9. D. Damanik and D. Lenz: *The index of sturmian sequences*. European Journal of Combinatorics, 23(1) 2002, pp. 23–29.
10. 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.
11. 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.
12. M. Lothaire: *Algebraic Combinatorics on Words*, vol. 90 of Encyclopedia of mathematics and its application, Cambridge University Press, 2002.
13. M. Lothaire: *Applied Combinatorics on Words*, vol. 105 of Encyclopedia of Mathematics and its Application, Cambridge University Press, 2005.
14. M. Piątkowski and W. Rytter: *Computing the number of cubic runs in standard Sturmian words*, in Proceedings of the 16-th Prague Stringology Conference, Czech Technical University, 2011, pp. 106–120.
15. M. Piątkowski and W. Rytter: *Asymptotic behaviour of the maximal number of squares in standard Sturmian words*. International Journal of Foundations of Computer Science, 23(2) 2012, pp. 303–321.
16. W. Rytter: *The structure of subword graphs and suffix trees of Fibonacci words*. Theoretical Computer Science, 363(2) 2006, pp. 211–223.
17. 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.
18. J. Shallit: *Characteristic words as fixed points of homomorphisms*, Tech. Rep. CS-91-72, University of Waterloo, Department of Computer Science, 1991.
19. H. Uscka-Wehlou: *Digital lines, Sturmian words, and continued fractions*, PhD thesis, Department of Mathematics, Uppsala University, 2009.