Discrete Applied Mathematics 163 (2014) 361–372

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Computing the number of cubic runs in standard Sturmian words

Marcin Piątkowski [a,*], Wojciech Rytter [b,a,1]

[a] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100 Toruń, Poland

[b] Department of Mathematics, Computer Science and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

ARTICLE INFO

Article history:
Received 11 December 2011
Received in revised form 15 May 2013
Accepted 27 May 2013
Available online 17 June 2013

Keywords:
Sturmian words
Standard words
Repetitions
Cubic runs
Algorithms

ABSTRACT

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

$$\limsup_{|w|\to\infty} \frac{\rho^{(3)}(w)}{|w|} = \frac{3\Phi+2}{9\Phi+4} \approx 0.36924841,$$

where $\Phi = \frac{\sqrt{5}+1}{2}$ is the *golden ratio*, and present a sequence of strictly growing standard words achieving this limit. The exact asymptotic ratio here is irrational, contrary to the situation of squares and runs in the same class of words. Furthermore, we design an efficient algorithm for computing the number of cubic runs in standard words in linear time with respect to the size of a *directive sequence*, i.e., the compressed representation describing the word (recurrences). The explicit size of a word can be exponential with respect to this representation, and hence this is yet another example of a very fast computation on highly compressible texts.

1. Introduction

Repetitions in strings are important in combinatorics on words and in many practical applications; see for instance [6,11,19,20]. 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.

We say that a number i is a period of the word w if $w[j] = w[i+j]$ for all j with $i+j \leq |w|$. The minimal period of w is called the *period* of w, and is denoted by $period(w)$. We say that a word w is periodic if $period(w) \leq \frac{|w|}{2}$. 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 repetition in a string w is a factor of the form $u^k = u \cdot u \cdot \ldots \cdot u$, where $k \geq 2$. Such a repetition is maximal if $|u|$ is minimal possible and k is maximal possible. Some overlapping repetitions defined in that way are cyclic shifts of one another. All such cyclic shifts could be represented by one factor $u^k v$, where v is a (proper) prefix of u. Thus we allow the exponent to a be rational number. In this work, we consider maximal repetitions as repetitions with a rational exponent.

* Corresponding author. Tel.: +48 56 611 3410; fax: +48 56 611 2987.
E-mail addresses: marcin.piatkowski@mat.umk.pl (M. Piątkowski), rytter@mimuw.edu.pl (W. Rytter).

1 Fax: +48 22 55 44 200.

0166-218X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dam.2013.05.025

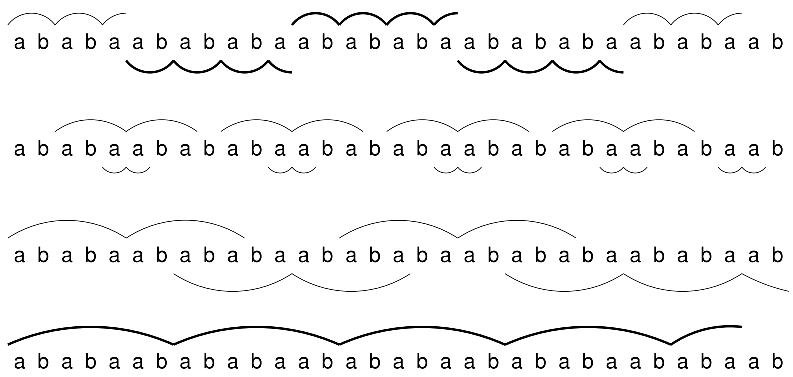

Fig. 1. The structure of repetitions in the word Sw(1, 2, 1, 3, 1). There are 19 runs and 4 cubic runs (marked in *bold*).

Formally, 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 cannot be extended (neither $w[i-1..j]$ nor $w[i..j+1]$ is a run with period $|u|$). *Cubic runs*, introduced and studied in [10], are defined in the same way, but we require that the period repeat at least three times ($k \geq 3$).

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

Example 1. Let $w = ababaababababaabababaababababaababaab$.

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

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

4 runs with period $|aba|$:

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

4 runs with period $|ababa|$:

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

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 Fig. 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 that

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

In 1999, Kolpakov and Kucherov (see [16]) showed that the number 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

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

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])

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