On the Maximal Sum of Exponents of Runs in a String

Maxime Crochemore[a,c], Marcin Kubica[b], Jakub Radoszewski[b,1], Wojciech Rytter[b,d], Tomasz Waleń[b]

[a] *King's College London, London WC2R 2LS, UK*
[b] *Dept. of Mathematics, Computer Science and Mechanics, University of Warsaw, Warsaw, Poland*
[c] *Université Paris-Est, France*
[d] *Dept. of Math. and Informatics, Copernicus University, Toruń, Poland*

Abstract

A run is an inclusion maximal occurrence in a string (as a subinterval) of a repetition v with a period p such that $2p \leq |v|$. The exponent of a run is defined as $|v|/p$ and is greater or equal 2. We show new bounds on the maximal sum of exponents of runs in a string of length n. Our upper bound of $4.1\,n$ is better than the best previously known proven bound of $5.6\,n$ by Crochemore & Ilie (2008). The lower bound of $2.035\,n$, obtained using a family of binary words, contradicts the conjecture of Kolpakov & Kucherov (1999), that the maximal sum of exponents of runs in a string of length n is smaller than $2n$.

Keywords: runs, repetitions in strings, exponents of repetitions.

1. Introduction

Repetitions and periodicities in strings are one of the fundamental topics in combinatorics on words [1, 2]. They are also important in other areas: lossless compression, word representation, computational biology, etc. In this paper we consider bounds on the sum of exponents of repetitions that a string of a given length may contain. In general, repetitions are studied also from other points of view, like: the classification of words (both finite and infinite) not containing repetitions of a given exponent, efficient identification of factors being repetitions of different types and computing the bounds on the number of various types of repetitions occurring in a string. More results and motivation can be found in a survey by Crochemore et al. [3].

Email addresses: maxime.crochemore@kcl.ac.uk (Maxime Crochemore), kubica@mimuw.edu.pl (Marcin Kubica), jrad@mimuw.edu.pl (Jakub Radoszewski), rytter@mimuw.edu.pl (Wojciech Rytter), walen@mimuw.edu.pl (Tomasz Waleń)

[1] Some parts of this paper were written during the author's Erasmus exchange at King's College London.

The concept of runs (also called maximal repetitions) has been introduced to represent all repetitions in a string in a succinct manner. The crucial property of runs is that their maximal number in a string of length n (denoted as $\rho(n)$) is $O(n)$, see Kolpakov & Kucherov [4]. This fact is the cornerstone of any algorithm computing all repetitions in strings of length n in $O(n)$ time. Due to the work of many people, much better bounds on $\rho(n)$ have been obtained. The lower bound $0.927\,n$ was first proved by Franek & Yang [5]. Afterwards, it was improved by Kusano et al. [6] to $0.944565\,n$ employing computer experiments, and very recently by Simpson [7] to $0.944575712\,n$. On the other hand, the first explicit upper bound $5\,n$ was settled by Rytter [8], afterwards it was systematically improved to $3.48\,n$ by Puglisi et al. [9], $3.44\,n$ by Rytter [10], $1.6\,n$ by Crochemore & Ilie [11, 12] and $1.52\,n$ by Giraud [13]. The best known result $\rho(n) \leq 1.029\,n$ is due to Crochemore et al. [14], but it is conjectured [4] that $\rho(n) < n$. Some results are known also for repetitions of exponent higher than 2. For instance, the maximal number of cubic runs (maximal repetitions with exponent at least 3) in a string of length n (denoted $\rho_{cubic}(n)$) is known to be between $0.406\,n$ and $0.5\,n$, see Crochemore et al. [15].

A stronger property of runs is that the maximal sum of their exponents in a string of length n (notation: $\sigma(n)$) is linear in terms of n, see final remarks in Kolpakov & Kucherov [16]. This fact has applications to the analysis of various algorithms, such as computing branching tandem repeats: the linearity of the sum of exponents solves a conjecture of [17] concerning the linearity of the number of maximal tandem repeats and implies that all can be found in linear time. For other applications, we refer to [16]. The proof that $\sigma(n) < cn$ in Kolpakov and Kucherov's paper [16] is very complex and does not provide any particular value for the constant c. A bound can be derived from the proof of Rytter [8] but the paper mentions only that the obtained bound is "unsatisfactory" (it seems to be $25\,n$). The first explicit bound $5.6\,n$ for $\sigma(n)$ was provided by Crochemore and Ilie [12], who claim that it could be improved to $2.9\,n$ employing computer experiments. As for the lower bound on $\sigma(n)$, no exact values were previously known and it was conjectured [18, 16] that $\sigma(n) < 2n$.

In this paper we provide an upper bound of $4.1\,n$ on the maximal sum of exponents of runs in a string of length n and also a stronger upper bound of $2.5\,n$ on the maximal sum of exponents of cubic runs in a string of length n. As for the lower bound, we bring down the conjecture of Kolpakov & Kucherov, that $\sigma(n) < 2n$, by providing an infinite family of binary strings for which the sum of exponents of runs is greater than $2.035\,n$.

2. Preliminaries

We consider *words* (*strings*) u over a finite alphabet Σ, $u \in \Sigma^*$; the empty word is denoted by ε; the positions in u are numbered from 1 to $|u|$. For $u = u_1 u_2 \dots u_m$, let us denote by $u[i\,..\,j]$ a *factor* of u equal to $u_i \dots u_j$ (in particular $u[i] = u[i\,..\,i]$). Words of the form $u[1\,..\,i]$ are called prefixes of u, and words of the form $u[i\,..\,|u|]$ suffixes of u.

If u and v are words, then the concatenation of u and v is denoted by uv. If k is a non-negative integer, then the k^{th} power of the word u, denoted by u^k, is the word:

$$u^k = \underbrace{uu\ldots u}_{k \text{ times}} .$$

A *square* is the 2^{nd} power of some non-empty word, and a *cube* is the 3^{rd} power of some non-empty word.

We say that an integer p is the (shortest) *period* of a word $u = u_1 \ldots u_m$ (notation: $p = \mathsf{per}(u)$) if p is the smallest positive integer, such that $u_i = u_{i+p}$ holds for all $1 \leq i \leq m - p$. The *primitive root* of a word u, denoted $\mathsf{root}(u)$, is the shortest word w such that $w^k = u$ for some positive integer k. We call a word u *primitive* if $\mathsf{root}(u) = u$, otherwise it is called *non-primitive*. We say that words u and v are cyclically equivalent (or that one of them is a cyclic rotation of the other) if $u = xy$ and $v = yx$ for some $x, y \in \Sigma^*$. It is a simple observation that if u and v are cyclically equivalent then $|\mathsf{root}(u)| = |\mathsf{root}(v)|$.

We say that $v \in \Sigma^*$ is a *Lyndon word* if it is primitive and minimal or maximal in the class of words that are cyclically equivalent to it. It is known (see [2]) that a Lyndon word has no non-trivial prefix that is also its suffix.

A *run* (also called a maximal repetition) in a string u is an interval $[i\,..\,j]$, such that:

- the period p of the associated factor $u[i\,..\,j]$ satisfies $2p \leq j - i + 1$,
- the interval cannot be extended to the right nor to the left, without violating the above property, that is, $u[i-1] \neq u[i+p-1]$ and $u[j-p+1] \neq u[j+1]$, provided that the respective letters exist.

A *cubic run* is a run $[i\,..\,j]$ for which the shortest period p satisfies $3p \leq j-i+1$. For simplicity, in the rest of the text we sometimes refer to runs and cubic runs as to occurrences of the corresponding factors of u. The (fractional) *exponent* of a run v, denoted $\mathsf{exp}(v)$, is defined as $(j - i + 1)/p$.

For a given word $u \in \Sigma^*$, we introduce the following notation:

- $\rho(u)$ and $\rho_{cubic}(u)$ are the numbers of runs and cubic runs in u respectively,
- $\sigma(u)$ and $\sigma_{cubic}(u)$ are the sums of exponents of runs and cubic runs in u respectively.

For a non-negative integer n, we use the same notations $\rho(n)$, $\rho_{cubic}(n)$, $\sigma(n)$ and $\sigma_{cubic}(n)$ to denote the maximal value of the respective function for a word of length n.

3. Upper Bounds for $\sigma(n)$ and $\sigma_{cubic}(n)$

In this section we utilize the concept of *handles* of runs as defined in [15]. The original definition refers only to cubic runs, but here we extend it also to ordinary runs.

Let $u \in \Sigma^*$ be a word of length n. Let us denote by $P = \{p_1, p_2, \ldots, p_{n-1}\}$ the set of inter-positions in u that are located *between* pairs of consecutive letters of u. We define a function H assigning to each run v in u a set of some inter-positions within v (called later on *handles*) — H is a mapping from the set of runs occurring in u to the set 2^P of subsets of P. Let v be a run with period p and let w be the prefix of v of length p. Let $w_{\min}$ and $w_{\max}$ be the minimal and maximal words (in lexicographical order) cyclically equivalent to w.

Example 1. *If $w = abaab$ then $w_{\min} = aabab$, $w_{\max} = babaa$.*

$H(v)$ is defined as follows:

a) if $w_{\min} = w_{\max}$ then $H(v)$ contains all inter-positions within v,
b) if $w_{\min} \neq w_{\max}$ then $H(v)$ contains inter-positions from the middle of any occurrence of $w_{\min}^2$ or $w_{\max}^2$ in v.

Note that $H(v)$ can be empty for a non-cubic-run v.

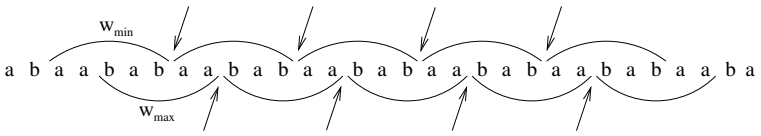

Figure 1: An example of a run $v = (abaab)^6 a$ with exponent $\mathsf{exp}(v) = 6.2$. It contains $\lfloor 6.2 \rfloor - 1 = 5$ occurrences of each of the words $w_{\min} = aabab$ and $w_{\max} = babaa$. The set $H(v)$ contains $2 \cdot (\lfloor 6.2 \rfloor - 2) = 8$ inter-positions, pointed by arrows in the figure

Lemma 3.1. *$w_{\min}$ and $w_{\max}$ are Lyndon words.*

Proof. By the definition of $w_{\min}$ and $w_{\max}$, it suffices to show that both words are primitive. This follows from the fact that, due to the minimality of p, w is primitive and that $w_{\min}$ and $w_{\max}$ are cyclically equivalent to w. □

Lemma 3.2. *Case (a) in the definition of $H(v)$ implies that $|w_{\min}| = 1$.*

Proof. $w_{\min}$ is primitive, therefore if $|w_{\min}| \geq 2$ then $w_{\min}$ contains at least two distinct letters, $a = w_{\min}[1]$ and $b = w_{\min}[i] \neq a$. If $b < a$ ($b > a$) then the cyclic rotation of $w_{\min}$ by $i - 1$ letters would be lexicographically smaller (greater) than $w_{\min}$, so $w_{\min} \neq w_{\max}$. □

Lemma 3.3. *$H(v_1) \cap H(v_2) = \emptyset$ for any two distinct runs v_1 and v_2 in u.*

Proof. Assume, to the contrary, that $p_i \in H(v_1) \cap H(v_2)$ is a handle of two different runs v_1 and v_2. By the definition of the handle and Lemma 3.1, p_i is located in the middle of two squares of Lyndon words: w_1^2 and w_2^2, where $|w_1| = \mathsf{per}(v_1)$ and $|w_2| = \mathsf{per}(v_2)$. Note that $w_1 \neq w_2$, since otherwise runs v_1 and v_2 would be the same. Without the loss of generality, we can assume that $|w_1| < |w_2|$. So, the word w_1 is both a prefix and a suffix of w_2 (see Fig. 2), what contradicts the fact that w_2 is a Lyndon word. □

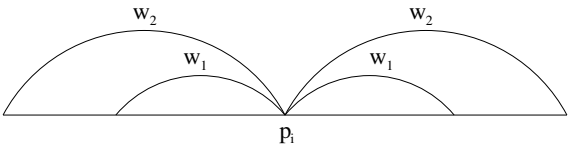

Figure 2: A situation where p_i is in the middle of two squares of different Lyndon words w_1^2 and w_2^2.

To prove the upper bound for $\sigma(n)$, we need to state an additional property of handles of runs. Let $\mathcal{R}(u)$ be the set of all runs in a word u, and let $\mathcal{R}_1(u)$ and $\mathcal{R}_{\geq 2}(u)$ be the sets of runs with period 1 and at least 2, respectively.

Lemma 3.4. *If $v \in \mathcal{R}_1(u)$ then $\mathsf{exp}(v) = |H(v)| + 1$.*
If $v \in \mathcal{R}_{\geq 2}(u)$ then $\lceil \mathsf{exp}(v) \rceil \leq \frac{|H(v)|}{2} + 3$.

Proof. For the case of $v \in \mathcal{R}_1(u)$, the proof is straightforward from the definition of handles. Assume now that $v \in \mathcal{R}_{\geq 2}(u)$ and let w be a prefix of v of length $\mathsf{per}(v)$. Then the word w^k for $k = \lfloor \mathsf{exp}(v) \rfloor$ is a prefix of v, and therefore both words $w_{\min}^{k-1}$ and $w_{\max}^{k-1}$ are factors of v. Each of the words provides $k-2$ distinct handles for v. Hence:

$$|H(v)| \geq 2 \cdot (\lfloor \mathsf{exp}(v) \rfloor - 2) \ .$$

□

Now we are ready to prove the upper bound for $\sigma(n)$. In the proof we use the bound $\rho(n) \leq 1.029\, n$ on the number of runs from [14].

Theorem 3.5. *The sum of the exponents of runs in a string of length n is less than $4.1\, n$.*

Proof. Let u be a word of length n. Using Lemma 3.4, we obtain:

$$\begin{aligned}
\sigma(u) &= \sum_{v \in \mathcal{R}_1(u)} \mathsf{exp}(v) + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \mathsf{exp}(v) \ \leq \\
&\leq \sum_{v \in \mathcal{R}_1(u)} (|H(v)| + 1) + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \left(\frac{|H(v)|}{2} + 3 \right) \ = \\
&= \sum_{v \in \mathcal{R}_1(u)} |H(v)| + |\mathcal{R}_1(u)| + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \frac{|H(v)|}{2} + 3 \cdot |\mathcal{R}_{\geq 2}(u)| \ \leq \\
&\leq 3 \cdot |\mathcal{R}(u)| + \sum_{v \in \mathcal{R}_1(u)} |H(v)| + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \frac{|H(v)|}{2} \ \leq \\
&\leq 3 \cdot |\mathcal{R}(u)| + \sum_{v \in \mathcal{R}(u)} |H(v)| \qquad (1)
\end{aligned}$$

Due to the disjointness of handles of runs (Lemma 3.3), $\sum_{v\in\mathcal{R}(u)} |H(v)| < n$. Combining it with (1), we obtain:

$$\sigma(u) \;<\; 3\cdot|\mathcal{R}(u)| + n \;\le\; 3\cdot\rho(n) + n \;\le\; 3\cdot 1.029\,n + n \;<\; 4.1\,n\ .$$

□

A similar approach for cubic runs, this time using the bound of $0.5\,n$ for $\rho_{cubic}(n)$ from [15], enables us to immediately provide a stronger upper bound for the function $\sigma_{cubic}(n)$.

Theorem 3.6. *The sum of the exponents of cubic runs in a string of length n is less than $2.5\,n$.*

Proof. Let u be a word of length n. Let $\mathcal{R}_{cubic,1}$ be the subset of $\mathcal{R}_1$ restricted to cubic runs only. Similarly, let $\mathcal{R}_{cubic,\ge 2}$ be the subset of $\mathcal{R}_{\ge 2}$ restricted to cubic runs only. From Lemma 3.4, we obtain:

$$\begin{aligned}
\sigma_{cubic}(u) \;&=\; \sum_{v\in\mathcal{R}_{cubic,1}(u)} \mathsf{exp}(v) + \sum_{v\in\mathcal{R}_{cubic,\ge 2}(u)} \mathsf{exp}(v) \;\le \\
&\le\; \sum_{v\in\mathcal{R}_{cubic,1}(u)} (|H(v)|+1) + \sum_{v\in\mathcal{R}_{cubic,\ge 2}(u)} \left(\frac{|H(v)|}{2}+3\right) \;= \\
&=\; \sum_{v\in\mathcal{R}_{cubic,1}(u)} |H(v)| + |\mathcal{R}_{cubic,1}(u)| + \\
&\quad + \sum_{v\in\mathcal{R}_{cubic,\ge 2}(u)} \frac{|H(v)|}{2} + 3\cdot|\mathcal{R}_{cubic,\ge 2}(u)| \;\le \\
&\le\; 3\cdot|\mathcal{R}_{cubic}(u)| + \sum_{v\in\mathcal{R}_{cubic,1}(u)} |H(v)| + \sum_{v\in\mathcal{R}_{cubic,\ge 2}(u)} \frac{|H(v)|}{2} \;\le \\
&\le\; 3\cdot|\mathcal{R}_{cubic}(u)| + \sum_{v\in\mathcal{R}_{cubic}(u)} |H(v)| \qquad (2)
\end{aligned}$$

Due to the disjointness of handles of runs (Lemma 3.3), $\sum_{v\in\mathcal{R}_{cubic}(u)} |H(v)| < n$. Combining it with (2), we obtain:

$$\sigma_{cubic}(u) \;<\; 3\cdot|\mathcal{R}_{cubic}(u)| + n \;\le\; 3\cdot\rho_{cubic}(n) + n \;<\; 3\cdot\frac{n}{2} + n \;= 2.5\,n\ .$$

□

4. Lower Bound for $\sigma(n)$

Let us start by investigating the sums of exponents of runs for words of two known families that contain a large number of runs. We consider first the words defined by Franek & Yang [5], then the Padovan words defined by Simpson [7].

i	$\lvert x_i\rvert$	$\rho(x_i)/\lvert x_i\rvert$	$\sigma(x_i)$	$\sigma(x_i)/\lvert x_i\rvert$
1	6	0.3333	4.00	0.6667
2	27	0.7037	39.18	1.4510
3	116	0.8534	209.70	1.8078
4	493	0.9047	954.27	1.9356
5	2090	0.9206	4130.66	1.9764
6	8855	0.9252	17608.48	1.9885
7	37512	0.9266	74723.85	1.9920
8	158905	0.9269	316690.85	1.9930
9	673134	0.9270	1341701.95	1.9932

Table 1: Number of runs and sum of exponents of runs in Franek & Yang's [5] words x_i.

They give large sums of exponents, however below $2n$. Then we construct a new family of words which breaks the barrier of $2n$.

Let $\circ$ be a special concatenation operator defined as:

$$x[1\,..\,n] \circ y[1\,..\,m] = \begin{cases} x[1\,..\,n]y[2\,..\,m] = x[1\,..\,n-1]y[1\,..\,m] & \text{if } x[n] = y[1], \\ x[1\,..\,n-1]y[2\,..\,m] & \text{if } x[n] \neq y[1]. \end{cases}$$

Also let g be a morphism defined as:

$$g(x) = \begin{cases} 010010 & \text{if } x = 0, \\ 101101 & \text{if } x = 1, \\ g(x[1\,..\,n]) = g(x[1]) \circ g(x[2]) \circ \ldots \circ g(x[n]) & \text{if } |x| > 1. \end{cases}$$

Then $x_i = g^i(0)$ is the family of words described by Franek and Yang [5], which gives the lower bound $\rho(n) \geq 0.927\,n$, conjectured for some time to be optimal. The sums of exponents of runs of several first terms of the sequence x_i are listed in Table 1.

Define a mapping $\delta(x) = R(f(x))$, where $R(x)$ is the reverse of x and f is the morphism

$$f(a) = aacab,\ f(b) = acab,\ f(c) = ac\ .$$

Let y'_i be a sequence of words defined for $i > 5$ recursively using $y'_{i+5} = \delta(y'_i)$. The first 5 elements of the sequence y'_i are:

$$b,\ a,\ ac,\ ba,\ aca\ .$$

The strings y'_i are called modified Padovan words. If we apply the following morphism h:

$$h(a) = 10100101100101001011010 0,$$

$$h(b) = 1010010110100, \qquad h(c) = 10100101$$

to y'_i, we obtain a sequence of run-rich strings y_i defined by Simpson [7], which gives the best known lower bound $\rho(n) \geq 0.944575712\,n$. Table 2 lists the sums of exponents of runs of selected words from the sequence y_i.

i	$\lvert y_i\rvert$	$\rho(y_i)/\lvert y_i\rvert$	$\sigma(y_i)$	$\sigma(y_i)/\lvert y_i\rvert$
1	13	0.6154	16.00	1.2308
6	69	0.7971	114.49	1.6593
11	287	0.8990	542.72	1.8910
16	1172	0.9309	2303.21	1.9652
21	4781	0.9406	9504.38	1.9879
26	19504	0.9434	38903.64	1.9946
31	79568	0.9443	158862.94	1.9966
36	324605	0.9445	648270.74	1.9971
41	1324257	0.9446	2644879.01	1.9973

Table 2: Number of runs and sum of exponents of runs in Simpson's [7] modified Padovan words y_i.

The values in Tables 1 and 2 have been computed experimentally. They suggest that for the families of words x_i and y_i the maximal sum of exponents could be less than $2n$. We show, however, a lower bound for $\sigma(n)$ that is greater than $2n$.

Theorem 4.1. *There are infinitely many binary strings w such that:*

$$\frac{\sigma(w)}{|w|} > 2.035 \ .$$

Proof. Let us define two morphisms $\phi : \{a, b, c\} \mapsto \{a, b, c\}$ and $\psi : \{a, b, c\} \mapsto \{0, 1\}$ as follows:

$$\phi(a) = baaba, \quad \phi(b) = ca, \quad \phi(c) = bca$$

$$\psi(a) = 01011, \quad \psi(b) = \psi(c) = 01001011 \ .$$

We define $w_i = \psi(\phi^i(a))$. Table 3 and Figure 3 show the sums of exponents of runs in words $w_1, \ldots, w_{10}$, computed experimentally.

Clearly, for any word $w = (w_8)^k$, $k \geq 1$, we have:

$$\frac{\sigma(w)}{|w|} > 2.035 \ .$$

□

5. Relating the Upper Bound for $\sigma(n)$ to Semicubic Runs

Recall that $1.029\,n$ is the best known upper bound for $\rho(n)$ [14]. On the other hand, the best known corresponding upper bound for cubic runs, for which the exponent is at least 3, is much smaller: $0.5\,n$ [15].

This suggests that the upper bound for the maximal number of runs with an intermediate exponent, e.g. at least 2.5, in a string of length n could be smaller than the general bound for exponent at least 2. Let us call such runs with exponent at least 2.5 *semicubic* runs.

i	$\lvert w_i\rvert$	$\sigma(w_i)$	$\sigma(w_i)/\lvert w_i\rvert$
1	31	47.10	1.5194
2	119	222.26	1.8677
3	461	911.68	1.9776
4	1751	3533.34	2.0179
5	6647	13498.20	2.0307
6	25205	51264.37	2.0339
7	95567	194470.30	2.0349
8	362327	737393.11	2.0352
9	1373693	2795792.39	2.0352
10	5208071	10599765.15	2.0353

Table 3: Sums of exponents of runs in words w_i.

Observation. The number of semicubic runs in Fibonacci strings is relatively small, it can be proved that in case of these strings every semicubic run is also cubic (if exponent is at least 2.5 then it is at least 3).

Let $\rho_{semic}(u)$ be the number of semicubic runs in the string u and let $\rho_{semic}(n)$ denote the maximum of $\rho_{semic}(u)$ over all strings u of length n.

Using extensive computer experiments, we have found the following family of binary words $z_i = \nu(\mu^i(a))$, where:

$$\begin{array}{lcllcl} \mu(a) & = & ba & \nu(a) & = & 0010100010 \\ \mu(b) & = & aba & \nu(b) & = & 001010 \end{array}$$

that contain, for sufficiently large i, at least $0.52\,n$ semi-cubic runs, see Table 4.

Using the same program we managed to construct strings with $\rho(n) \geq 0.944575\,n$, that is, very close to the best known lower bound for this function and also close to the best known upper bound. This suggests that the data from the program are good approximation for semicubic runs as well. Therefore we conjecture the following.

Conjecture 5.1 (Semicubic-Runs Conjecture)**.** $\rho_{semic}(n) \leq 0.6\,n$.

If the above conjecture holds, it lets us instantly improve the upper bound for $\sigma(n)$.

Theorem 5.2. *If Conjecture 5.1 is true then* $\sigma(n) \leq 3.9\,n$.

Proof. Let $\mathcal{R}_{semic}(u)$ denote the set of all semicubic runs in a string u, and let $\mathcal{R}_{semic,1}(u) = \mathcal{R}_{semic}(u) \cap \mathcal{R}_1(u)$ and $\mathcal{R}_{semic,\geq 2}(u) = \mathcal{R}_{semic}(u) \cap \mathcal{R}_{\geq 2}(u)$. Conjecture 5.1 allows us to improve the part of the proof of Theorem 3.5 related

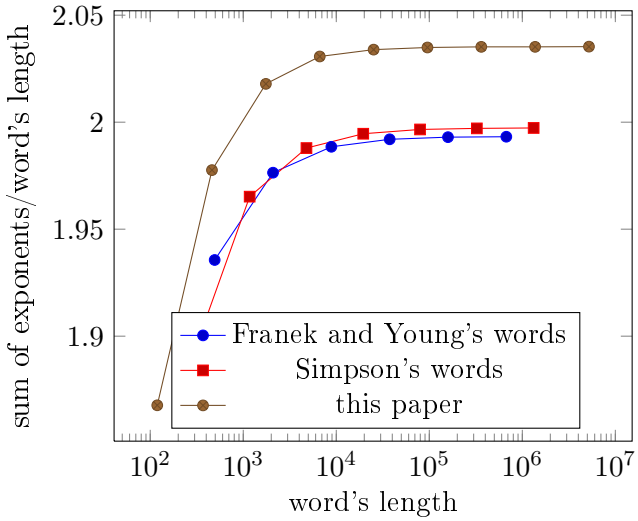

Figure 3: Comparison of the sum of exponents of runs in selected families of words.

to the term $\sum_{v\in\mathcal{R}_{\geq 2}(u)} \mathsf{exp}(v)$:

$$
\begin{aligned}
\sigma(u) \;&=\; \sum_{v\in\mathcal{R}_1(u)} \mathsf{exp}(v) + \sum_{v\in\mathcal{R}_{\geq 2}(u)} \mathsf{exp}(v) \;= \\
&=\; \sum_{v\in\mathcal{R}_1(u)} \mathsf{exp}(v) + \sum_{v\in\mathcal{R}_{\geq 2}(u)\setminus\mathcal{R}_{semic,\geq 2}(u)} \mathsf{exp}(v) + \sum_{v\in\mathcal{R}_{semic,\geq 2}(u)} \mathsf{exp}(v) \;\leq \\
&\leq\; \sum_{v\in\mathcal{R}_1(u)} (|H(v)|+1) + \sum_{v\in\mathcal{R}_{\geq 2}(u)\setminus\mathcal{R}_{semic,\geq 2}(u)} 2.5 + \\
&\quad + \sum_{v\in\mathcal{R}_{semic,\geq 2}(u)} \left(\frac{|H(v)|}{2}+3\right) \;\leq \\
&\leq\; \sum_{v\in\mathcal{R}_1(u)} |H(v)| + |\mathcal{R}_1(u)| + 2.5\cdot|\mathcal{R}_{\geq 2}(u)\setminus\mathcal{R}_{semic,\geq 2}(u)| + \\
&\quad + \sum_{v\in\mathcal{R}_{semic,\geq 2}(u)} \frac{|H(v)|}{2} + 3\cdot|\mathcal{R}_{semic,\geq 2}(u)| \;\leq \\
&\leq\; \sum_{v\in\mathcal{R}(u)} |H(v)| + |\mathcal{R}_1(u)| + 2.5\cdot|\mathcal{R}_{\geq 2}(u)\setminus\mathcal{R}_{semic,\geq 2}(u)| + \\
&\quad + 3\cdot|\mathcal{R}_{semic,\geq 2}(u)| \;\leq \\
&\leq\; n + |\mathcal{R}_1(u)| + 2.5\cdot(|\mathcal{R}_{\geq 2}(u)| - |\mathcal{R}_{semic,\geq 2}(u)|) + 3\cdot|\mathcal{R}_{semic,\geq 2}(u)| \;\leq \\
&\leq\; n + |\mathcal{R}_1(u)| + 2.5\cdot|\mathcal{R}_{\geq 2}(u)| + 0.5\cdot|\mathcal{R}_{semic,\geq 2}(u)| \;\leq \\
&\leq\; n + 2.5\cdot|\mathcal{R}(u)| + 0.5\cdot|\mathcal{R}_{semic}(u)| \;\leq \\
&\leq\; n + 2.5\cdot 1.029n + 0.5\cdot 0.6n \;\leq\; 3.9n
\end{aligned}
$$

i	$\lvert z_i\rvert$	$\rho_{semic}(z_i)$	$\rho_{semic}(z_i)/\lvert z_i\rvert$
1	16	5	0.3125
2	42	17	0.4048
3	100	46	0.4600
4	242	118	0.4876
5	584	296	0.5069
6	1410	724	0.5135
7	3404	1762	0.5176
8	8218	4266	0.5191
9	19840	10316	0.5200
10	47898	24920	0.5203
11	115636	60182	0.5204
12	279170	145310	0.5205
13	673976	350832	0.5205
14	1627122	847004	0.5206

Table 4: The number of semicubic runs in the words $z_i = \nu(\mu^i(a))$.

□

6. Conclusions

In this paper we have provided an upper bound of $4.1\,n$ on the maximal sum of exponents of runs in a string of length n and also a stronger upper bound of $2.5\,n$ for the maximal sum of exponents of cubic runs in a string of length n. As for the lower bound, we bring down the conjecture Kolpakov & Kucherov [18, 16], that $\sigma(n) < 2n$, by providing an infinite family of binary strings for which the sum of exponents of runs is greater than $2.035\,n$.

A natural open problem is to tighten these bounds. One of the possible directions for this improvement, presented in this paper, consists in finding bounds for the maximal number of runs with exponent at least f, where $f \in (2, 3)$, in a string of length n.

References

[1] J. Berstel, J. Karhumaki, Combinatorics on words: a tutorial, Bulletin of the EATCS 79 (2003) 178–228.

[2] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA., U.S.A., 1983.

[3] M. Crochemore, L. Ilie, W. Rytter, Repetitions in strings: Algorithms and combinatorics, Theor. Comput. Sci. 410 (50) (2009) 5227–5235.

[4] R. M. Kolpakov, G. Kucherov, Finding maximal repetitions in a word in linear time, in: Proceedings of the 40th Symposium on Foundations of Computer Science, 1999, pp. 596–604.

[5] F. Franek, Q. Yang, An asymptotic lower bound for the maximal number of runs in a string, Int. J. Found. Comput. Sci. 19 (1) (2008) 195–203.

[6] K. Kusano, W. Matsubara, A. Ishino, H. Bannai, A. Shinohara, New lower bounds for the maximum number of runs in a string, CoRR abs/0804.1214.

[7] J. Simpson, Modified Padovan words and the maximum number of runs in a word, Australasian J. of Comb. 46 (2010) 129–145.

[8] W. Rytter, The number of runs in a string: Improved analysis of the linear upper bound, in: B. Durand, W. Thomas (Eds.), STACS, Vol. 3884 of Lecture Notes in Computer Science, Springer, 2006, pp. 184–195.

[9] S. J. Puglisi, J. Simpson, W. F. Smyth, How many runs can a string contain?, Theor. Comput. Sci. 401 (1-3) (2008) 165–171.

[10] W. Rytter, The number of runs in a string, Inf. Comput. 205 (9) (2007) 1459–1469.

[11] M. Crochemore, L. Ilie, Analysis of maximal repetitions in strings, in: L. Kucera, A. Kucera (Eds.), MFCS, Vol. 4708 of Lecture Notes in Computer Science, Springer, 2007, pp. 465–476.

[12] M. Crochemore, L. Ilie, Maximal repetitions in strings, J. Comput. Syst. Sci. 74 (5) (2008) 796–807.

[13] M. Giraud, Not so many runs in strings, in: C. Martín-Vide, F. Otto, H. Fernau (Eds.), LATA, Vol. 5196 of Lecture Notes in Computer Science, Springer, 2008, pp. 232–239.

[14] M. Crochemore, L. Ilie, L. Tinta, Towards a solution to the "runs" conjecture, in: P. Ferragina, G. M. Landau (Eds.), CPM, Vol. 5029 of Lecture Notes in Computer Science, Springer, 2008, pp. 290–302.

[15] M. Crochemore, C. S. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, T. Walen, On the maximal number of cubic runs in a string, in: A. H. Dediu, H. Fernau, C. Martín-Vide (Eds.), LATA, Vol. 6031 of Lecture Notes in Computer Science, Springer, 2010, pp. 227–238.

[16] R. M. Kolpakov, G. Kucherov, On the sum of exponents of maximal repetitions in a word, Tech. Report 99-R-034, LORIA.

[17] D. Gusfield, J. Stoye, Simple and flexible detection of contiguous repeats using a suffix tree (preliminary version), in: M. Farach-Colton (Ed.), CPM, Vol. 1448 of Lecture Notes in Computer Science, Springer, 1998, pp. 140–152.

[18] R. M. Kolpakov, G. Kucherov, On maximal repetitions in words, J. of Discr. Alg. 1 (1999) 159–186.