On Morphisms Generating Run-Rich Strings

Kazuhiko Kusano, Kazuyuki Narisawa, and Ayumi Shinohara

Graduate School of Information Sciences, Tohoku University, Japan
{kusano@shino., narisawa@, ayumi@}ecei.tohoku.ac.jp

Abstract. A run in a string is a periodic substring which is extendable neither to the left nor to the right with the same period. Strings containing many runs are of interest. In this paper, we focus on the series of strings $\{\psi(\phi^i(\mathtt{a}))\}_{i\geq 0}$ generated by two kinds of morphisms, $\phi : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{a},\mathtt{b},\mathtt{c}\}^*$ and $\psi : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{0},\mathtt{1}\}^*$. We reveal a simple morphism ϕ_r plays a critical role to generate run-rich strings. Combined with a morphism ψ', the strings $\{\psi'(\phi_r^i(\mathtt{a}))\}_{i\geq 0}$ achieves *exactly the same* lower bound as the current best lower bound for the maximum number $\rho(n)$ of runs in a string of length n. Moreover, combined with another morphism ψ'', the strings $\{\psi''(\phi_r^i(\mathtt{a}))\}_{i\geq 0}$ give a new lower bound for the maximum value $\sigma(n)$ of the sum of exponents of runs in a string of length n.

Keywords: run, sum of exponents, repetition, morphic word

1 Introduction

Repetitions are one of the most fundamental topics in stringology, and they are also important for practical areas, such as string processing, data compression and bioinformatics. A *run* (or *maximal repetition*) in a string is a periodic substring which is extendable neither to the left nor to the right with the same period. All repetitions in a string can be succinctly represented by runs. Strings containing many runs (we call them *run-rich strings*) are of interest to researchers. In 1999, Kolpakov and Kucherov [12] showed that the maximum number $\rho(n)$ of runs in a string of length n is $\rho(n) \leq cn$ for some constant c. Since then, a great deal of efforts have been devoted to estimate the constant c [8,17,9,18,4,3,10,16,15,19,14,13,2,5], while it is conjectured that $c < 1$. The current best upper bound for $\rho(n)/n$ is 1.029 due to Crochemore *et al.* [5] in 2011, and the current best lower bound is 0.9445757 due to Simpson [19] in 2010.

The maximum value $\sigma(n)$ of sum of exponents in runs in a string of length n is of another concern. Clearly $2\rho(n) \leq \sigma(n)$, since each exponent in a run is at least 2. The current best upper bound 4.087 and the best lower bound 2.035257 for $\sigma(n)/n$ are both given by Crochemore *et al.* [6] in 2011.

In order to provide lower bounds for $\rho(n)$ and $\sigma(n)$, various kinds of run-rich strings are shown in the literature. In 2003, Franek *et al.* [8,7] defined an ingenious run-rich strings to show a lower bound $3/(1+\sqrt{5}) = 0.9270509$ for $\rho(n)/n$. In 2008, Matsubara *et al.* [15] found a more run-rich string of length 184973 which contains 174719 runs by computer experiments, that provided a better lower bound 0.9445648. They improved it in [14] to 0.9445756 by defining a series $\{t_i\}_{i\geq 0}$ of strings. In 2010, Simpson [19] provided another series $\{s_i\}_{i\geq 0}$ of strings based on the modified Padovan words, that gives the current best lower bound 0.9445757. We note that $\{t_i\}$ also gives exactly the same bound, assuming that the recurrence formula conjectured in [14] is correct. In 2011, Crochemore *et al.* [6] showed the current best lower bound 2.035257 for $\sigma(n)/n$ by defining the strings $\{\psi_c(\phi_c^i(\mathtt{a}))\}_{i\geq 0}$ using two morphisms $\phi_c : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{a},\mathtt{b},\mathtt{c}\}^*$ and $\psi_c : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{0},\mathtt{1}\}^*$.

Kazuhiko Kusano, Kazuyuki Narisawa, Ayumi Shinohara: On Morphisms Generating Run-Rich Strings, pp. 35–47.
Proceedings of PSC 2013, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-05330-0

$$\phi_r(\mathtt{a}) = \mathtt{abac} \qquad \phi_r(\mathtt{b}) = \mathtt{aac} \qquad \phi_r(\mathtt{c}) = \mathtt{a}$$

$$\begin{cases} h(\mathtt{a}) = \mathtt{101001011001010010110100} \\ h(\mathtt{b}) = \mathtt{1010010110100} \\ h(\mathtt{c}) = \mathtt{10100101} \end{cases} \qquad \begin{cases} \psi_e(\mathtt{a}) = \mathtt{101001010010} \\ \psi_e(\mathtt{b}) = \mathtt{110100} \\ \psi_e(\mathtt{c}) = \mathtt{1} \end{cases}$$

$$u_i = h(\phi_r^i(\mathtt{a})) \qquad v_i = \psi_e(\phi_r^i(\mathtt{a}))$$

$$\rho(u_i)/|u_i| \to 0.9445757 \ (i \to \infty) \qquad \sigma(v_{12})/|v_{12}| = 2.036982$$

$$\sigma\big((v_{12})^k\big)/\big|(v_{12})^k\big| \to 2.036992 \ (k \to \infty)$$

Figure 1. Two morphisms ϕ_r and ψ_e we discovered, and the summary of the results.

In this paper, we focus on the strings defined by the same form $\{\psi(\phi^i(\mathtt{a}))\}_{i\geq 0}$, and try to find better ones by computer experiments. We report two morphisms ϕ_r and ψ_e in Fig. 1 that we discovered. These morphisms are effective for defining run-rich strings from the following two viewpoints:

1. The strings $\{h(\phi_r^i(\mathtt{a}))\}_{i\geq 0}$ achieve *exactly the same* lower bound for $\rho(n)/n$ with the current best lower bound 0.9445757. Here, h is the morphism proposed by Simpson [19] to define the run-rich strings $\{h(p_i)\}_{i\geq 0}$ based on the *modified Padovan words* $\{p_i\}_{i\geq 0}$, and $\{h(p_i)\}_{i\geq 0}$ are the very strings that achieve the current best lower bound.
2. The strings $\{\psi_e(\phi_r^i(\mathtt{a}))\}_{i\geq 0}$ give a new lower bound 2.036992 for $\sigma(n)/n$; that is better than the current best lower bound 2.035257.

Therefore, the simple morphism ϕ_r plays a critical role to generate run-rich strings, both for the number $\rho(n)$ of runs and the sum $\sigma(n)$ of exponents of runs. Another attractive feature of ϕ_r is its simplicity, compared to the definition of the modified Padovan words.

The rest of this paper is organized as follows. In Section 2, we introduce some notations on runs. Section 3 reviews three series of strings that appeared in the literature [14,19,6], that give the current best lower bounds for $\rho(n)$ and $\sigma(n)$. We then explain in Section 4, a simple search strategy based on enumerations for finding good morphisms. In Section 5, for the strings $u_i = h(\phi_r^i(\mathtt{a}))$, we prove $\rho(u_i)/|u_i| \to 0.9445757$, that exactly equals to the current best lower bound for $\rho(n)/n$. In Section 6, we show that the lower bound for $\sigma(n)/n$ is improved to be 2.036992 by the string $\psi_e(\phi_r^{12}(\mathtt{a}))$. Section 7 concludes and discusses some future work. In Appendix, we supply some lemmas and remarks easily verified by Mathematica, for convenience.

2 Preliminaries

Let Σ be an alphabet. We denote by Σ^n the set of all strings of length n over Σ, and $|w|$ denotes the length of a string w. We denote by $w[i]$ the ith letter of w, and $w[i..j]$ is a substring $w[i]w[i+1]\cdots w[j]$ of w.

For a string w of length n and a positive integer $p \leq n$, we say that p is a *period* of w if $w[i] = w[i+p]$ holds for any $1 \leq i \leq n-p$. A string may have several periods. For instance, string $\mathtt{abaababa}$ has three periods 5, 7 and 8. A string w is *primitive* if w cannot be written as $w = u^k$ by any string u and any integer $k \geq 2$. A *run* (also called a maximal repetition) in a string w is an interval $[i..j]$, such that:

(1) the smallest period p of $w[i..j]$ satisfies $2p \leq j-i+1$,
(2) either $i = 1$ or $w[i-1] \neq w[i+p-1]$,

(3) either $j = n$ or $w[j+1] \neq w[j-p+1]$.

That is, *run* is a maximal repetition which is extendable neither to the left nor to the right. The (fractional) *exponent* of the run $[i..j]$ is defined as $\frac{j-i+1}{p}$. We often represent the run $[i..j]$ by a triplet $\langle i, j-i+1, p\rangle$ of the initial position, length, and the shortest period, for convenience. We denote by $Run(w)$ the set of all runs in string w. For instance, let us consider a string $w = \texttt{aabaababababab}$. It contains 4 runs; $Run(w) = \{\langle 1,2,1\rangle, \langle 4,5,1\rangle, \langle 1,7,3\rangle, \langle 5,12,2\rangle\}$. On the other hand, $\langle 1,6,3\rangle$ is not a run in w since the repetition can be extended to the right. Neither is $\langle 5,12,4\rangle$, since the smallest period of $w[5..12]$ is 2, but not 4.

We denote by $\rho(w)$ the number of runs contained in string w, and by $\sigma(w)$ the sum of exponents of all runs in string w.

Example 1. For a string $w = \texttt{aabaabaaaacaacac}$, we have

$$Run(w) = \{\langle 1,2,1\rangle, \langle 4,2,1\rangle, \langle 7,4,1\rangle, \langle 12,2,1\rangle, \langle 13,4,2\rangle, \langle 1,8,3\rangle, \langle 9,7,3\rangle\}.$$

Thus, $\rho(w) = 7$, and $\sigma(w) = \frac{2}{1} + \frac{2}{1} + \frac{4}{1} + \frac{2}{1} + \frac{4}{2} + \frac{8}{3} + \frac{7}{3} = 17$.

For a non-negative integer n, we denote by $\rho(n)$ the maximum number of runs in a string of length n, and by $\sigma(n)$ the maximum value of the sum of exponents of runs in a string of length n. That is,

$$\rho(n) = \max\{\rho(w) \mid w \in \Sigma^n\} \quad \text{and} \quad \sigma(n) = \max\{\sigma(w) \mid w \in \Sigma^n\}.$$

3 Previously Known Series of Run-Rich Strings

This section briefly reviews three series of strings containing many runs, which are defined by recursions,

The first one is due to Simpson [19], which gives the current best lower bound for the maximum number $\rho(n)$ of runs in a string of length n.

Definition 2 ([19]). *The modified Padovan words $\{p_i\}$ are defined by*

$$p_1 = \texttt{b}, \quad p_2 = \texttt{a}, \quad p_3 = \texttt{ac}, \quad p_4 = \texttt{ba}, \quad p_5 = \texttt{aca}, \ \textit{and}\ p_i = R(f(p_{i-5}))\ \textit{for}\ i > 5,$$

where $R(w)$ is the reverse of w, and $f : \{\texttt{a},\texttt{b},\texttt{c}\} \to \{\texttt{a},\texttt{b},\texttt{c}\}^*$ is a morphism

$$f(\texttt{a}) = \texttt{aacab}, \quad f(\texttt{b}) = \texttt{acab}, \quad f(\texttt{c}) = \texttt{ac}.$$

Simpson's words $\{s_i\}$ are defined by $s_i = h(p_i)$, where $h : \{\texttt{a},\texttt{b},\texttt{c}\} \to \{\texttt{0},\texttt{1}\}^*$ is a morphism

$$\begin{aligned} h(\texttt{a}) &= \texttt{1010010110010100101101 00}, \\ h(\texttt{b}) &= \texttt{1010010110100}, \\ h(\texttt{c}) &= \texttt{10100101}. \end{aligned} \tag{1}$$

Theorem 3 ([19]). $\lim_{n\to\infty} \dfrac{\rho(n)}{n} \geq \lim_{i\to\infty} \dfrac{\rho(s_i)}{|s_i|} = \eta > 0.9445757$,
where η is the real root of $2693z^3 - 7714z^2 + 7379z - 2357 = 0$.

Proof. Simpson [19] proved that $\lim_{i\to\infty} \frac{\rho(s_i)}{|s_i|} = \frac{11\kappa^2+7\kappa-6}{11\kappa^2+8\kappa-6}$, where κ is the real root of $z^3 - z - 1 = 0$. We can verify $\frac{11\kappa^2+7\kappa-6}{11\kappa^2+8\kappa-6} = \eta$ easily (Lemma 16 in Appendix). $\square$

The second one is proposed by Matsubara *et al.* [14].

Definition 4 ([14]). *Matsubara* et al.*'s words* $\{t_i\}$ *are defined by*

$$
\begin{aligned}
t_0 &= \mathtt{1001010010110100101},\\
t_1 &= \mathtt{1001010010110},\\
t_2 &= \mathtt{1001010010110100101001011010010100101},\\
t_k &= t_{k-1}\,t_{k-2} \qquad (k \bmod 3 = 0,\ k > 2),\\
t_k &= t_{k-1}\,t_{k-4} \qquad (k \bmod 3 \neq 0,\ k > 2).
\end{aligned}
$$

Interestingly, these strings $\{t_i\}$ give *exactly* the same lower bound as $\{s_i\}$ for $\rho(n)$.

Theorem 5 ([14][1]). $\lim_{n\to\infty} \frac{\rho(n)}{n} \geq \lim_{i\to\infty} \frac{\rho(t_i)}{|t_i|} = \eta > 0.9445757$,
where η *is the real root of* $2693z^3 - 7714z^2 + 7379z - 2357 = 0$.

Proof. We can verify that the value $\lim_{i\to\infty} \rho(t_i)/|t_i|$ shown in the proof of Theorem 6 in the paper [14] is *exactly* equal to η (Lemma 17 in Appendix). □

The third one is introduced by Crochemore *et al.* [6], which gives the current best lower bound for the maximum value $\sigma(n)$ of the sum of exponents of runs.

Definition 6 ([6]). *Crochemore* et al.*'s words* $\{c_i\}$ *are defined by* $c_i = \psi_c(\phi_c^i(\mathtt{a}))$ *using two morphisms* $\phi_c : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{a},\mathtt{b},\mathtt{c}\}^*$ *and* $\psi_c : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{0},\mathtt{1}\}^*$ *such that*

$$
\begin{aligned}
&\phi_c(\mathtt{a}) = \mathtt{baaba}, \qquad \phi_c(\mathtt{b}) = \mathtt{ca}, \qquad \phi_c(\mathtt{c}) = \mathtt{bca},\\
&\psi_c(\mathtt{a}) = \mathtt{01011}, \quad \psi_c(\mathtt{b}) = \psi_c(\mathtt{c}) = \mathtt{01001011}.
\end{aligned}
$$

Theorem 7 ([6]). $\lim_{n\to\infty} \frac{\sigma(n)}{n} \geq \frac{\sigma(c_{10})}{|c_{10}|} \geq \frac{10599765.15}{5208071} > 2.035257.$

4 Searching for Better Morphisms

Inspired by a simple and elegant definition of Crochemore's words, we are interested in finding other series of strings defined by similar recursions, that hopefully contain more runs or larger sum of exponents.

We focus on the series $\{w_i\}$ of strings defined by $w_i = \psi(\phi^i(\mathtt{a}))$ using two morphisms $\phi : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \mapsto \{\mathtt{a},\mathtt{b},\mathtt{c}\}^*$ and $\psi : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \mapsto \{\mathtt{0},\mathtt{1}\}^*$, and try to find good pair of ϕ and ψ, in the sense that either $\rho(w_i)$ or $\sigma(w_i)$ is large enough.

Various approaches are possible to search for good pairs. For instance, even a simple random search might be usable. We chose the following two-phase strategy, as the search space is huge (needless to say, infinite) and we observed that inappropriate choices of ψ would never succeed to find good ϕ's.

In the first phase, we search for ϕ by fixing ψ to h defined in Eq. (1) in Definition 2. We enumerate every possible morphism ϕ in increasing order with respect to the sum $|\phi(\mathtt{a})| + |\phi(\mathtt{b})| + |\phi(\mathtt{c})|$, and compute all runs in the string $h(\phi^i(\mathtt{a}))$ whose length is reasonably long. If a good ϕ yielding many runs is found, report it. A pseudo-code is shown in Algorithm 1. At this point, we succeeded to find a good morphism ϕ_r,

[1] Strictly speaking, the general formula of $\rho(t_i)$ in the paper is derived from a recurrence formula, which is verified for $i = 0, 1, \ldots, 14$, but not formally proved.

i	$\|u_i\|$	$\rho(u_i)$	$\rho(u_i)/\|u_i\|$	i	$\|s_i\|$	$\rho(s_i)$	$\rho(s_i)/\|s_i\|$
0	**24**	**16**	**0.66666**	2	**24**	**16**	**0.66666**
1	69	56	0.81159	7	93	79	0.84946
2	218	193	0.88532	12	380	345	0.90789
3	667	616	0.92353	17	1552	1450	0.93427
4	2057	1925	0.93582	22	**6333**	**5963**	**0.94157**
5	**6333**	**5963**	**0.94157**	27	25837	24383	0.94372
6	19504	18400	0.94340	32	105405	99538	0.94433
7	60064	56711	0.94417	37	430010	406149	0.94451
8	184973	174693	0.94442	42	**1754267**	**1657007**	**0.94455**
9	569642	538041	0.94452	47	7156700	6760011	0.94457
10	**1754267**	**1657005**	**0.94455**				

Table 1. Comparison of $u_i = h\left(\phi_r^i(\mathtt{a})\right)$ with Simpson's words $s_i = h\left(p_i\right)$. Rows holding the same lengths are highlighted in **bold**, for clarity.

which achieves the same lower bound for $\rho(n)$ as the current best one. We will fully explain it in Section 5.

In the second phase, we fix ϕ to the best ϕ_r found in the first phase, and enumerate every ψ in the same way (see Algorithm 2 for a pseudo-code). We finally found a good morphism ϕ_e so that $\sigma(\psi_e(\psi_r^8(\mathtt{a})))/|\psi_e(\psi_r^8(\mathtt{a}))| = 2.03632$ clearly exceeds the current best lower bound 2.035257 for $\sigma(n)/n$. We will describe the new lower bounds in Section 6.

5 Simpler Morphism Achieving the Current Best Lower Bound for $\rho(n)$

We obtained the following morphism $\phi_r : \{\mathtt{a},\mathtt{b},\mathtt{c}\} \to \{\mathtt{a},\mathtt{b},\mathtt{c}\}^*$,

$$\phi_r(\mathtt{a}) = \mathtt{abac}, \quad \phi_r(\mathtt{b}) = \mathtt{aac}, \quad \phi_r(\mathtt{c}) = \mathtt{a}. \tag{2}$$

Combined with the morphism h in Definition 2, we now have another good series $\{u_i\}$ of run-rich strings, defined by $u_i = h(\phi_r^i(\mathtt{a}))$. Table 1 compares $\{u_i\}$ with Simpson's words $\{s_i\}$ with respect to the length and the number of runs. While the definition of our strings $\{u_i\}$ is much simpler than that of Simpson's words $\{s_i\}$, the numbers of runs are *almost* the same; note that it is *not exactly* the same, since $|u_{10}| = |s_{42}| = 1754267$ and $\rho(u_{10}) = 1757005 < 1757007 = \rho(s_{42})$. More interestingly, however, the asymptotic value of the ratio $\rho(u_i)/|u_i|$ *exactly* coincides with that of $\rho(s_i)/|s_i|$, as we will see in Theorem 10.

We begin by obtaining a general formula representing the length $|u_i|$.

Lemma 8. *Let $L(z) = \sum_{i=0}^{\infty} |u_i| z^i$ be the ordinary generating function of the sequence $\{|u_i|\}_{i\geq 0}$ of lengths of u_i's. Then*

$$L(z) = \frac{-8z^2 - 21z - 24}{z^3 + 3z^2 + 2z - 1}.$$

Proof. Let $|w|_a$ denote the number of occurrences of a in string w. Then for any $w \in \{\mathtt{a},\mathtt{b},\mathtt{c}\}^*$, the length $|w|$ is calculated by the sum $|w|_\mathtt{a} + |w|_\mathtt{b} + |w|_\mathtt{c}$. Let M be the *incidence matrix* (see e.g. Chapter 8.2 in [1]) of the morphism ϕ_r defined by

$$M = \begin{pmatrix} |\phi_r(\mathtt{a})|_\mathtt{a} & |\phi_r(\mathtt{b})|_\mathtt{a} & |\phi_r(\mathtt{c})|_\mathtt{a} \\ |\phi_r(\mathtt{a})|_\mathtt{b} & |\phi_r(\mathtt{b})|_\mathtt{b} & |\phi_r(\mathtt{c})|_\mathtt{b} \\ |\phi_r(\mathtt{a})|_\mathtt{c} & |\phi_r(\mathtt{b})|_\mathtt{c} & |\phi_r(\mathtt{c})|_\mathtt{c} \end{pmatrix} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

Then for any string $w \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}^*$, it holds that

$$\begin{pmatrix} |\phi_r(w)|_{\mathtt{a}} \\ |\phi_r(w)|_{\mathtt{b}} \\ |\phi_r(w)|_{\mathtt{c}} \end{pmatrix} = M \begin{pmatrix} |w|_{\mathtt{a}} \\ |w|_{\mathtt{b}} \\ |w|_{\mathtt{c}} \end{pmatrix},$$

which induces the recurrence formula $|u_i| = 2|u_{i-1}| + 3|u_{i-2}| + |u_{i-3}|$ for $i \geq 3$, since the characteristic polynomial of M is $-x^3 + 2x^2 + 3x + 1$. Taking into account the initial values $|u_0| = 24$, $|u_1| = 69$ and $|u_2| = 218$, we obtain the generating function $L(z)$ of the sequences $|u_i|$'s as we stated (see e.g. [11] for handling generating functions). See also Remark 18 in Appendix. □

Lemma 9. *Let $R(z) = \sum_{i=0}^{\infty} \rho(u_i)\, z^i$ be the ordinary generating function of the sequence $\{\rho(u_i)\}_{i \geq 0}$ of the numbers of runs in u_i's. Then*

$$R(z) = \frac{-16 - 8z + 7z^2 - 5z^3 - 3z^4 - z^5 + z^6}{(1-z)^2(1+z)(-1+2z+3z^2+z^3)}.$$

Proof. By observing the sequence $\rho(u_0),\ \rho(u_1), \ldots, \rho(u_{10})$, we found a recurrence formula would hold, as in Table 2:

$$\begin{aligned} a_{i+2} - a_i &= 25, \quad (i \geq 1), \\ a_1 = 58, &\quad a_2 = 72, \end{aligned} \tag{3}$$

where a_i is defined[2] by

$$a_i = \rho(u_{i+3}) - 2\rho(u_{i+2}) - 3\rho(u_{i+1}) - \rho(u_i)\,. \tag{4}$$

Assuming that Eq. (3) holds for any $i \geq 1$ (in this sense, the proof is incomplete yet), we can get the general term of a_i as

$$\begin{aligned} a_i &= \frac{3}{4}(-1)^i + \frac{25i}{2} + \frac{185}{4} \qquad (i \geq 1), \\ a_0 &= 46. \end{aligned}$$

Combined with Eq. (4), we get the generating function $R(z)$ of $\rho(u_i)$ as stated. See Remark 18 in Appendix. □

Theorem 10. $\displaystyle\lim_{i \to \infty} \frac{\rho(u_i)}{|u_i|} = \eta,$
where η is the real root of $2693z^3 - 7714z^2 + 7379z - 2357 = 0$.

Proof. By Lemma 8 and 9, we have the generating functions $L(z)$ and $R(z)$ for $|u_i|$ and $\rho(u_i)$, respectively. Lemma 19 in Appendix completes the rest. □

[2] Based on the fact that the characteristic polynomial of M is $-x^3 + 2x^2 + 3x + 1$.

i	$\rho(u_i)$	a_i	$a_{i+2} - a_i$	$a_{i+1} - a_i$
0	16	46	26	12
1	56	58	25	14
2	193	72	25	11
3	616	83	25	14
4	1925	97	25	11
5	5963	108	25	14
6	18400	122		11
7	56711	133		
8	174693			
9	538041			
10	1657005			

Table 2. Observation on the series $\{\rho(u_i)\}$ for $u_i = h\left(\phi_r^i(\mathtt{a})\right)$. If we define a_i as in Eq. (4), the difference sequence $a_{i+2} - a_i$ of order 2 seems to be a constant 25 except the initial value $a_2 - a_0 = 26$. Note also that the difference sequence $a_{i+1} - a_i$ of order 1 has alternating values 14 and 11.

6 New Lower Bounds for $\sigma(n)$

In the second phase of search, we obtained the morphism $\psi_e : \{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \to \{\mathtt{0}, \mathtt{1}\}^*$,

$$\psi_e(\mathtt{a}) = \mathtt{101001010010}, \quad \psi_e(\mathtt{b}) = \mathtt{110100}, \quad \psi_e(\mathtt{c}) = \mathtt{1}.$$

Combined with the morphism ϕ_r in Eq. (2), let us define $v_i = \psi_e(\phi_r^i(\mathtt{a}))$. In this section, we will show that the strings $\{v_i\}$ give a better lower bound of the maximum sum $\sigma(n)$ of exponents of runs.

Table 3 shows the length $|v_i|$, the number $\rho(v_i)$ of runs, and the sum $\sigma(v_i)$ of exponents, together with their ratios to the length. First let us notice that the strings $\{v_i\}$ do *not* contain so many runs. In fact, we can verify $\lim_{i\to\infty} \rho(v_i)/|v_i| = 0.923118$ assuming that a similar recurrence relation as Eqs. (3), (4) holds (see Lemma 20 in Appendix for confidence), that is strictly inferior to the current best lower bound $\lim_{i\to\infty} \rho(u_i)/|u_i| = 0.9445757$.

However, on the other hand, the sum $\sigma(v_i)$ of exponents of runs in the string v_i is very large. Figure 2 illustrates the comparison of our words $v_i = \psi_e\left(\phi_r^i(\mathtt{a})\right)$ with Crochemore et al.'s words $c_i = \psi_c\left(\phi_c^i(\mathtt{a})\right)$. Apparently, $\sigma(v_i)$ for $i \geq 8$ exceeds the current best lower bound $\sigma(c_{10}) = 2.035257$.

Theorem 11. *There exist infinitely many strings w such that:*

$$\frac{\sigma(w)}{|w|} > 2.03698.$$

Proof. In Table 3, we see that $\sigma(v_{12})/|v_{12}| = 15389914.96/7555252 > 2.03698$. Thus, for any string $w = (v_{12})^k$, $k \geq 1$, we have

$$\frac{\sigma(w)}{|w|} = \frac{\sigma\left((v_{12})^k\right)}{|(v_{12})^k|} \geq \frac{k \cdot \sigma(v_{12})}{k \cdot |v_{12}|} > 2.03698,$$

since $\sigma(xy) \geq \sigma(x) + \sigma(y)$ holds for any strings x and y. $\square$

In the rest of this section, we further push up the lower bound for $\sigma(n)$ by estimating the behavior of $\sigma(v_i)$ more carefully. It would be preferable to get a general

i	$\lvert v_i\rvert$	$\rho(v_i)$	$\frac{\rho(v_i)}{\lvert v_i\rvert}$	$\sigma(v_i)$	$\frac{\sigma(v_i)}{\lvert v_i\rvert}$	$\frac{\sigma(v_i^3)-\sigma(v_i^2)}{\lvert v_i\rvert}$
0	12	7	0.583333	14.90	1.24166	1.70238
1	31	23	0.741935	49.70	1.60322	1.94014
2	99	83	0.838384	180.88	1.82707	1.99612
3	303	268	0.884488	590.11	1.94756	2.02682
4	934	849	0.908994	1869.94	2.00208	2.03278
5	2876	2638	0.917246	5818.98	2.02329	2.03581
6	8857	8158	0.921079	17997.22	2.03197	2.03657
7	27276	25157	0.922313	55509.41	2.03510	2.03686
8	83999	77518	0.922844	171049.01	2.03632	2.03694
9	258683	238768	0.923014	526871.76	2.03674	2.03697
10	796639	735364	0.923083	1622679.68	2.03690	2.03698
11	2453326	2264678	0.923105	4997332.12	2.03696	2.03699152
12	7555252			15389914.96	2.03698	2.03699251

Table 3. Numbers of runs, and sums of exponents in runs in strings $v_i = \psi_e\left(\phi_r^i(\mathtt{a})\right)$

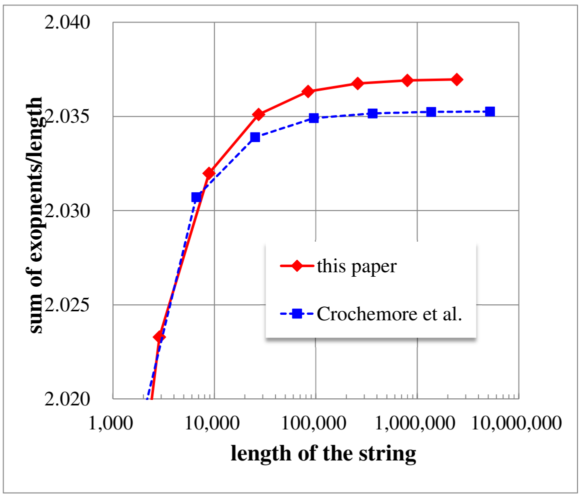

Figure 2. Comparison of the sum of exponents of runs in $v_i = \psi_e(\phi_r^i(\mathtt{a}))$ and Crochemore *et al.*'s $c_i = \psi_c(\phi_c^i(\mathtt{a}))$

formula of $\sigma(v_i)$, as similar to $\rho(u_i)$ in Section 5. Unfortunately, however, we failed to guess recurrence formulas on $\sigma(v_i)$ up to now. A part of the difficulty comes from the fact that $\sigma(v_i)$ is a fractional number, while $\rho(u_i)$ is an integer.

As an alternative approach, we consider a series of strings $\{w^k\}_{k\geq 1}$ of a run-rich string w, and compute a simple general formula for $\sigma\left(w^k\right)$. We first recall a property on runs in a string of the form w^k.

Lemma 12 ([15]). *Let $r = \langle i, l, p\rangle$ be a run in a string w^k for $k \geq 3$. If $l \geq 2|w|$, then $i = 1$ and $l = kn$, that is, $r = w^k$.*

Lemma 13. *For any string w and any $k \geq 2$,*

$$\sigma\left(w^k\right) = \left(\sigma\left(w^3\right) - \sigma\left(w^2\right)\right) \cdot k - \left(2\sigma\left(w^3\right) - 3\sigma\left(w^2\right)\right).$$

Proof. By Lemma 12, for any $k \geq 3$, the set $Run\left(w^k\right)$ consists of a single long run $\langle 1, \left|w^k\right|, p\rangle$ that covers the whole w^k, and many (possibly zero) short runs whose lengths are at most $2|w|$. Thus, we can verify that $\sigma\left(w^{k+1}\right) - \sigma\left(w^k\right) = \sigma(w^3) - \sigma(w^2)$ for any $k \geq 2$. By solving it, we get the general formula of $\sigma\left(w^k\right)$ as stated. □

Theorem 14. *For any string w and any $\varepsilon > 0$, there exists a positive integer N such that for any $n \geq N$,*

$$\frac{\sigma(n)}{n} > \frac{\sigma(w^3) - \sigma(w^2)}{|w|} - \varepsilon.$$

Proof. By Lemma 13, $\sigma\big(w^k\big) = A \cdot k - B$, where $A = \sigma(w^3) - \sigma(w^2)$ and $B = 2\sigma(w^3) - 3\sigma(w^2)$. For any given $\varepsilon > 0$, we choose $N > \frac{A+B}{\varepsilon}$. For any $n \geq N$, let k be the integer satisfying $k > \frac{n}{|w|} \geq k - 1$. Notice that $k > \frac{n}{|w|} \geq \frac{N}{|w|} > \frac{A+B}{|w|\varepsilon}$. Since $\sigma(i+1) \geq \sigma(i)$ for any i, and $\left|w^{k-1}\right| = |w|(k-1)$, we have

$$\frac{\sigma(n)}{n} > \frac{\sigma(|w|(k-1))}{|w|k} \geq \frac{\sigma\big(w^{k-1}\big)}{|w|k} = \frac{A(k-1) - B}{|w|k} = \frac{A}{|w|} - \frac{A+B}{|w|k} > \frac{A}{|w|} - \varepsilon.$$

□

We now have a slightly better lower bound for $\sigma(n)$ compared to Theorem 11.

Theorem 15. *For any $\varepsilon > 0$ there exists a positive integer N such that for any $n \geq N$, $\frac{\sigma(n)}{n} > 2.036992 - \varepsilon$*

Proof. From Theorem 14 and the fact shown in Table 3, we have the bound. □

7 Concluding Remarks

We provided a new lower bound $2.036992n$ for the maximum value $\sigma(n)$ of the sum of exponents in runs in a string of length n, by exhibiting the series $\{\psi_e(\phi_r^i(\mathtt{a}))\}_{i\geq 0}$ of strings. Moreover, we also showed that the current best lower bound $0.9445757n$ for the number $\rho(n)$ of runs in a string of length n can be achieved by yet another series $\{h(\phi_r^i(\mathtt{a}))\}_{i\geq 0}$ of strings than Simpson's words $\{s_i\}_{i\geq 0}$ and Matsubara *et al.*'s words $\{t_i\}_{i\geq 0}$.

We note that the proof for Lemma 9 is incomplete for the moment, because the recurrence formula Eq. (3) is not formally proved yet for $i \geq 6$, in Table 2. We are also interested in obtaining a general formula of $\sigma(\psi_e(\phi_r^i(\mathtt{a})))$, which will yield a slightly better lower bound for $\sigma(n)$. Recall that for the standard Sturmian words, the number of runs in them can be exactly and directly computed from their *directive sequences* [3]. Similarly, it would be wonderful if we could develop a general technique to evaluate $\rho(\psi(\phi^i(\mathtt{a})))$ and $\sigma(\psi(\phi^i(\mathtt{a})))$ directly from the definition of ψ and ϕ. A natural extension of our experimental approach is to enlarge the domain of the morphism ϕ. For instance, can we get more run-rich strings $\{\psi(\phi^i(\mathtt{a}))\}_{i\geq 0}$ if we consider $\phi : \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\} \to \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}^*$?

Acknowledgments

This work was supported by KAKENHI 24106010, 23300051, 23220001, and 25560067.

Algorithm 1 find good morphism $\phi : \{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \to \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}^*$ by enumeration

$maxNum := 0$
$maxExp := 0$
for $N := 3$ to ∞ **do**
  for $\ell_\mathtt{a} := 1$ to $N-2$ **do**
    for $\ell_\mathtt{b} := 1$ to $N - \ell_\mathtt{a} - 1$ **do**
      $\ell_\mathtt{c} := N - \ell_a - \ell_\mathtt{a}$
      for $n_\mathtt{a} := 0$ to $3^{\ell_\mathtt{a}} - 1$ **do**
        for $n_\mathtt{a} := 0$ to $3^{\ell_\mathtt{b}} - 1$ **do**
          for $n_\mathtt{a} := 0$ to $3^{\ell_\mathtt{c}} - 1$ **do**
            Let $x_\mathtt{a}$ (resp. $x_\mathtt{b}$, $x_\mathtt{c}$) be the ternary representation of $n_\mathtt{a}$ (resp. $n_\mathtt{b}$, $n_\mathtt{c}$) in $\ell_\mathtt{a}$ (resp. $\ell_\mathtt{b}$, $\ell_\mathtt{c}$) digits over $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$
            Let $\phi(\mathtt{a}) = x_\mathtt{a}$, $\phi(\mathtt{b}) = x_\mathtt{b}$ and $\phi(\mathtt{c}) = x_\mathtt{c}$
            Let w be the prefix of $h(\phi^k(\mathtt{a}))$ of length 10000, where k is the minimum integer satisfying $\left|h(\phi^k(\mathtt{a}))\right| \geq 10000$
            if $\rho(w) > maxNum$ **then**
              $maxNum := \rho(w)$ and **report** ϕ
            if $\sigma(w) > maxExp$ **then**
              $maxExp := \sigma(w)$ and **report** ϕ

Algorithm 2 find good morphism $\psi : \{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \to \{\mathtt{0}, \mathtt{1}\}^*$ by enumeration

$maxNum := 0$
$maxExp := 0$
for $N := 3$ to ∞ **do**
  for $\ell_\mathtt{a} := 1$ to $N-2$ **do**
    for $\ell_\mathtt{b} := 1$ to $N - \ell_\mathtt{a} - 1$ **do**
      $\ell_\mathtt{c} := N - \ell_\mathtt{a} - \ell_\mathtt{b}$
      for $n_\mathtt{a} := 0$ to $2^{\ell_\mathtt{a}} - 1$ **do**
        for $n_\mathtt{b} := 0$ to $2^{\ell_\mathtt{b}} - 1$ **do**
          for $n_\mathtt{c} := 0$ to $2^{\ell_\mathtt{c}} - 1$ **do**
            Let $y_\mathtt{a}$ (resp. $y_\mathtt{b}$, $y_\mathtt{c}$) be the binary representation of $n_\mathtt{a}$ (resp. $n_\mathtt{b}$, $n_\mathtt{c}$) in $\ell_\mathtt{a}$ (resp. $\ell_\mathtt{b}$, $\ell_\mathtt{c}$) digits over $\{\mathtt{0}, \mathtt{1}\}$.
            Let $\psi(\mathtt{a}) = y_\mathtt{a}$, $\psi(\mathtt{b}) = y_\mathtt{b}$ and $\psi(\mathtt{c}) = y_\mathtt{c}$
            Let w be the prefix of $\psi(\phi_r^k(\mathtt{a}))$ of length 10000, where k is the minimum integer satisfying $\left|\psi(\phi_r^k(\mathtt{a}))\right| \geq 10000$
            if $\rho(w) > maxNum$ **then**
              $maxNum := \rho(w)$ and **report** ψ
            if $\sigma(w) > maxExp$ **then**
              $maxExp := \sigma(w)$ and **report** ψ

References

1. J.-P. Allouche and J. Shallit: *Automatic Sequences*, Cahmbridge University Press, 2003.
2. A. Baker, A. Deza, and F. Franek: *A computational framework for determining run-maximal strings*. Journal of Discrete Algorithms, 2012.
3. P. Baturo, M. Piatkowski, and W. Rytter: *The number of runs in Sturmian words*, in Proc. CIAA 2008, 2008, pp. 252–261.
4. M. Crochemore, L. Ilie, and L. Tinta: *Towards a solution to the "runs" conjecture*, in Proc. CPM 2008, vol. 5029 of LNCS, 2008, pp. 290–302.
5. M. Crochemore, L. Ilie, and L. Tinta: *The "runs" conjecture*. Theoretical Computer Science, 412 2011, pp. 2931–2941.
6. M. Crochemore, M. Kubica, J. Radoszewski, W. Rytter, and T. Waleń: *On the maximal sum of exponents of runs in a string*. Journal of Discrete Algorithms, 2011.
7. F. Franek and Q. Yang: *An asymptotic lower bound for the maximal number of runs in a string*. International Journal of Foundations of Computer Science, 19(01) 2008, pp. 195–203.
8. F. Franěk, R. Simpson, and W. Smyth: *The maximum number of runs in a string*, in Proc. AWOCA2003, 2003, pp. 26–35.
9. F. Franěk and Q. Yang: *An asymptotic lower bound for the maximal-number-of-runs function*, in Proc. Prague Stringology Conference (PSC'06), 2006, pp. 3–8.
10. M. Giraud: *Not so many runs in strings*, in Proc. LATA 2008, 2008, pp. 245–252.
11. R. L. Graham, D. E. Knuth, and O. Patashnik: *Concrete Mathematics: A Foundation for Computer Science*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd ed., 1994.
12. R. M. Kolpakov and G. Kucherov: *Finding maximal repetitions in a word in linear time*, in Proc. FOCS'99, 1999, pp. 596–604.
13. K. Kusano, K. Narisawa, and A. Shinohara: *Computing maximum number of runs in strings*, in String Processing and Information Retrieval, Springer, 2012, pp. 318–329.
14. W. Matsubara, K. Kusano, H. Bannai, and A. Shinohara: *A series of run-rich strings*, in Proc. LATA 2009, 2009, pp. 578–587.
15. W. Matsubara, K. Kusano, A. Ishino, H. Bannai, and A. Shinohara: *New lower bounds for the maximum number of runs in a string*, in Proc. PSC2008, 2008, pp. 140–145.
16. S. J. Puglisi, J. simpson, and W. F. Smyth: *How many runs can a string contain?* Theoretical Computer Science, 401(1–3) 2008, pp. 165–171.
17. W. Rytter: *The number of runs in a string: Improved analysis of the linear upper bound*, in Proc. STACS 2006, vol. 3884 of LNCS, 2006, pp. 184–195.
18. W. Rytter: *The number of runs in a string*. Inf. Comput., 205(9) 2007, pp. 1459–1469.
19. J. Simpson: *Modified Padovan words and the maximum number of runs in a word*. Australasian Journal of Combinatorics, 46 2010, pp. 129–146.

A Appendix

We note some lemmas and remarks verified by Mathematica 9.0.1.

Lemma 16. $(11\kappa^2 + 7\kappa - 6)/(11\kappa^2 + 8\kappa - 6) = \eta$, *where* κ *is the real root of* $z^3 - z - 1 = 0$, *and* η *is the real root of* $2693z^3 - 7714z^2 + 7379z - 2357 = 0$.

Proof. We can verify it as follows.

kappa = Solve[z^3 − z − 1 == 0, z][[1]]

$$\left\{z \to \frac{1}{3}\left(\frac{27}{2} - \frac{3\sqrt{69}}{2}\right)^{1/3} + \frac{\left(\frac{1}{2}\left(9+\sqrt{69}\right)\right)^{1/3}}{3^{2/3}}\right\}$$

FullSimplify[(11z^2 + 7z − 6)/(11z^2 + 8z − 6)/.kappa]

$\text{Root}\left[-2357 + 7379\#1 - 7714\#1^2 + 2693\#1^3\&, 1\right]$

□

Lemma 17. *The real root* η *of* $2693z^3 - 7714z^2 + 7379z - 2357 = 0$ *is*

$$\frac{7714 - 109145\sqrt[3]{\frac{2}{-27669823+9298929\sqrt{69}}} + \sqrt[3]{\frac{-27669823+9298929\sqrt{69}}{2}}}{8079} = 0.9445757124$$

Proof. We can easily verify it as follows.

eta = Solve[2693x^3 − 7714x^2 + 7379x − 2357 == 0][[1]]

$$\left\{x \to \frac{7714-109145\left(\frac{2}{-27669823+9298929\sqrt{69}}\right)^{1/3}+\left(\frac{1}{2}\left(-27669823+9298929\sqrt{69}\right)\right)^{1/3}}{8079}\right\}$$

***N*[%, 10]**

$\{x \to 0.9445757124\}$

□

Remark 18. The following instructions would give a confidence that $L(z)$ (resp. $R(z)$) in Lemma 8 (resp. Lemma 9) is a generating function of $|u_i|$ (resp. $\rho(u_i)$) in Table 1.

Table[SeriesCoefficient[(−8z^2 − 21z − 24)/(z^3 + 3z^2 + 2z − 1), {z, 0, n}], {n, 0, 10}]

$\{24, 69, 218, 667, 2057, 6333, 19504, 60064, 184973, 569642, 1754267\}$

Table[SeriesCoefficient[(−16 − 8z + 7z^2 − 5z^3 − 3z^4 − z^5 + z^6)/ ((1 − z)^2 ∗ (1 + z) ∗ (−1 + 2z + 3z^2 + z^3)), {z, 0, n}], {n, 0, 10}]

$\{16, 56, 193, 616, 1925, 5963, 18400, 56711, 174693, 538041, 1657005\}$

Lemma 19. *Assume that* $\displaystyle\sum_{i=0}^{\infty} |u_i| z^i = \frac{-8z^2 - 21z - 24}{z^3 + 3z^2 + 2z - 1}$, *and*
$$\sum_{i=0}^{\infty} \rho(u_i)\, z^i = \frac{-16 - 8z + 7z^2 - 5z^3 - 3z^4 - z^5 + z^6}{(1-z)^2(1+z)(-1+2z+3z2+z3)}.$$
Then $\displaystyle\lim_{i\to\infty} \frac{\rho(u_i)}{|u_i|} = \eta$, *where* η *is the real root of* $2693z^3 - 7714z^2 + 7379z - 2357 = 0$.

Proof. We can verify it as follows.

leng[n_]:=SeriesCoefficient $\left[\frac{-24-21z-8z^2}{-1+2z+3z^2+z^3}, \{z, 0, n\}\right]$
run[n_]:=SeriesCoefficient $\left[\frac{-16-8z+7z^2-5z^3-3z^4-z^5+z^6}{(-1+z)^2(1+z)(-1+2z+3z^2+z^3)}, \{z, 0, n\}\right]$

Table[leng[n], {n, 0, 10}]
{24, 69, 218, 667, 2057, 6333, 19504, 60064, 184973, 569642, 1754267}

Table[run[n], {n, 0, 10}]
{16, 56, 193, 616, 1925, 5963, 18400, 56711, 174693, 538041, 1657005}

FullSimplify[Limit[run[n]/leng[n], $n \to$ Infinity]]
Root $\left[-2357 + 7379\#1 - 7714\#1^2 + 2693\#1^3\&, 1\right]$

□

Lemma 20. *Assume* $\displaystyle\sum_{i=0}^{\infty} |v_i| z^i = \frac{-12 - 7z - z^2}{-1 + 2z + 3z^2 + z^3}$, *and*
$$\sum_{i=0}^{\infty} \rho(v_i)\, z^i = \frac{-7 - 2z - 8z^3 - 8z^4 - 2z^5 + z^6 + z^7}{(-1+z)^2(1+z)\,(-1+2z+3z^2+z^3)}.$$
Then $\displaystyle\lim_{i\to\infty} \frac{\rho(v_i)}{|v_i|} = 0.9231182492\ldots$ *is the real root of* $175z^3 - 344z^2 + 397z - 211 = 0$.

Proof. We can easily verify it as follows.

leng[n_]:=SeriesCoefficient $\left[\frac{-12-7z-z^2}{-1+2z+3z^2+z^3}, \{z, 0, n\}\right]$
run[n_]:=SeriesCoefficient $\left[\frac{-7-2z-8z^3-8z^4-2z^5+z^6+z^7}{(-1+z)^2(1+z)(-1+2z+3z^2+z^3)}, \{z, 0, n\}\right]$

Table[leng[n], {n, 0, 10}]
{12, 31, 99, 303, 934, 2876, 8857, 27276, 83999, 258683, 796639}

Table[run[n], {n, 0, 10}]
{7, 23, 83, 268, 849, 2638, 8158, 25157, 77518, 238768, 735364}

FullSimplify[Limit[run[n]/leng[n], $n \to$ Infinity]]
Root $\left[-211 + 397\#1 - 344\#1^2 + 175\#1^3\&, 1\right]$

N[%, 10]
0.9231182492

□