An asymptotic lower bound for the maximal-number-of-runs function

Frantisek Franek* and Qian Yang

Department of Computing & Software
Faculty of Engineering
McMaster University
Hamilton, Ontario
Canada L8S 4K1
franek@mcmaster.ca yangq6@univmail.cis.mcmaster.ca

Abstract. An asymptotic lower bound for the maxrun function $\boldsymbol{\rho}(n) = \max\{\textit{number of runs in string } \boldsymbol{x} \mid \textit{all strings } \boldsymbol{x} \textit{ of length } n\}$ is presented. More precisely, it is shown that for any $\varepsilon > 0$, $(\alpha-\varepsilon)n$ is an asymptotic lower bound, where $\alpha = \frac{3}{1+\sqrt{5}} \approx 0.927$. A recent construction of an increasing sequence of binary strings "rich in runs" is modified and extended to prove the result.

Key words: run, lower bound, asymptotic lower run, maximum number of runs

1 Introduction

An important structural characteristic of a string over an alphabet is its periodicity. Repetitions (tandem repeats) have always been in the focus of the research into periodicities. The notion of runs captures maximal repetitions which themselves are not repetitions and allows for a succinct notation ([5]). Even though it had been known that there could be $O(n \log n)$ of repetitions in a string of length n ([1]), it was shown in 2000 by Kolpakov and Kucherov that number of runs was linear in the length of the input string ([4]). Their proof was existential and thus did not specify the constants of linearity. The behaviour of the **maxrun function** $\boldsymbol{\rho}(n) = \max\{\textit{number of runs in string } \boldsymbol{x} \mid \textit{all strings } \boldsymbol{x} \textit{ of length } n\}$ became an interest of study to many. Smyth et al. (e.g. [3], [6], [2]) presented a set of conjectures about $\boldsymbol{\rho}(n)$:

1. $\boldsymbol{\rho}(n) < n$,
2. $\boldsymbol{\rho}(n+1) \leq \boldsymbol{\rho}(n)+2$,
3. $\boldsymbol{\rho}(n) = \boldsymbol{\rho}_2(n)$, the maxrun function for binary strings.

Just recently, Rytter improved the upper bound of $\boldsymbol{\rho}(n)$ to $6.3n$ (see [7]).

[3] introduced a construction of an increasing sequence $\{\boldsymbol{x}_n \ : \ n < \infty\}$ of binary strings "rich in runs" so that $\lim_{n\to\infty} \frac{\boldsymbol{r}(\boldsymbol{x_n})}{|\boldsymbol{x_n}|} = \alpha$, where $\alpha = \frac{3}{1+\sqrt{5}} \approx 0.927$ and $\boldsymbol{r}(\boldsymbol{x}) = \textit{number of runs in } \boldsymbol{x}$. Although any such sequence does not establish a lower bound (not even an asymptotic one), it has been "viewed" as such. The assumption underneath that view is that $\boldsymbol{\rho}(n)$ behaves "reasonably", i.e. that $\boldsymbol{\rho}(n)$ does not exhibit wild jumps up, or equivalently, that $\frac{\boldsymbol{\rho}(n)}{n}$ does not exhibit wild oscillations,

* Supported in part by a research grant from the Natural Sciences and Engineering Research Council of Canada.

which is generally expected to be the case (cf. the second conjecture). However, since the "reasonable behaviour" of $\boldsymbol{\rho}(n)$ is yet to be established, we modify and extend the method from [3] to provide formally a family of true asymptotic lower bounds arbitrarily close to αn by proving

Theorem: *For any $\varepsilon > 0$ there is a positive integer N so that for any $n \geq N$, $\boldsymbol{\rho}(n) \geq (\alpha - \varepsilon)n$.*

2 Basic notation, facts, and methods

A **run** $\mathcal{R}$ in a string $\boldsymbol{x}$ is a four-tuple of positive integers (s, p, e, t), where

1. s is the **starting position** of $\mathcal{R}$.
2. p is the **size of its period**.
3. $e \geq 2$ is its **exponent**, i.e. the maximal value e so that $\boldsymbol{x}[s..s+p-1] = \boldsymbol{x}[s+p..s+2p-1] = \cdots = \boldsymbol{x}[s+(e-1)p..s+ep-1]$.
4. The **period** of $\mathcal{R}$, $\boldsymbol{x}[s..s+p-1]$ itself is not a repetition.
5. The **square part** of the run $\mathcal{R}$, $\boldsymbol{x}[s..s+p-1] = \boldsymbol{x}[s+p..s+2p-1]$ is **left-maximal**, i.e. $\boldsymbol{x}[s-1..s+p-2] \neq \boldsymbol{x}[s+p-1..s+2p-2]$.
6. t is the **tail** of $\mathcal{R}$ and indicates how far to the right the run can be extended, i.e. t is a maximal number so that for any $0 < t' \leq t$, $\boldsymbol{x}[s+t'..s+t'+p-1] = \boldsymbol{x}[s+t'+p..s+t'+2p-1] = \cdots = \boldsymbol{x}[s+t'+(e-1)p..s+t'+ep-1]$.

Not too much is known about the behaviour of the maxrun function:

- *For any n, $\boldsymbol{\rho}(n+2) \geq \boldsymbol{\rho}(n)+1$.*
  Take a string $\boldsymbol{x}$ of length n with $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{\rho}(n)$. Take a letter c that does not occur in $\boldsymbol{x}$. Then $\boldsymbol{x}$cc is a string of length $n+2$ and $\boldsymbol{\rho}(n+2) \geq \boldsymbol{r}(\boldsymbol{x}cc) = \boldsymbol{r}(\boldsymbol{x})+1 = \boldsymbol{\rho}(n)+1$.
- *For any n, $\boldsymbol{\rho}(n+1) \leq \boldsymbol{\rho}(n)+\lfloor \frac{n}{2} \rfloor$.*
  Take a string $\boldsymbol{x}$ of length $n+1$ with $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{\rho}(n+1)$. There can be at most $\lfloor \frac{n}{2} \rfloor$ squares starting at position 1. Then $\boldsymbol{\rho}(n) \geq \boldsymbol{r}(\boldsymbol{x}[2..n+1]) \geq \boldsymbol{r}(\boldsymbol{x}) - \lfloor \frac{n}{2} \rfloor \geq \boldsymbol{\rho}(n+1) - \lfloor \frac{n}{2} \rfloor$.
- *For some n, $\boldsymbol{\rho}(n+1) = \boldsymbol{\rho}(n)$.*
  Established by computations, it is not clear if this as an asymptotic property (for instance, $\boldsymbol{\rho}(33) = 27$ while $\boldsymbol{\rho}(34) = 27$).
- *For some n, $\boldsymbol{\rho}(n+1) = \boldsymbol{\rho}(n)+2$.*
  Established by computations, it is not clear if this as an asymptotic property (for instance, $\boldsymbol{\rho}(13) = 8$ while $\boldsymbol{\rho}(14) = 10$).

Note that the function $\frac{\boldsymbol{\rho}(n)}{n}$ may thus not be monotonic. It is not even clear whether $\lim_{n\to\infty} \frac{\boldsymbol{\rho}(n)}{n}$ exists, as $\frac{\boldsymbol{\rho}(n)}{n}$ may be oscillating with a non-decreasing magnitude.

In [3] a special concatenation operator $\circ$ for binary strings was introduced:

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

Morphism g was defined by

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

The strings 010010 and 101101 were selected as they provide in the concatenation one extra run:
$\boldsymbol{r}(g(0) \circ g(0)) = 6 = 2\boldsymbol{r}(g(0))+2$, the same for $g(1) \circ g(1)$, $\boldsymbol{r}(g(0) \circ g(1)) = 5 = \boldsymbol{r}(g(0))+\boldsymbol{r}(g(1))+1$, the same for $\boldsymbol{r}(g(1) \circ g(0))$. Let us remark that a computer search carried to the length of 20 did not discover any better pair of strings with such properties.

An important aspect of the morphism is that it "preserves" the runs in $\boldsymbol{x}$: it is a bit tedious to prove and thus not included in the paper, but any left-maximal square in $\boldsymbol{x}$ induces a square in $g(\boldsymbol{x})$. It follows that every run in $\boldsymbol{x}$ induces a run in $g(\boldsymbol{x})$. It is also important to show that two distinct runs in $\boldsymbol{x}$ do not get "glued" together by g.

Let us fix a string $\boldsymbol{x}$. Let $\lambda(\boldsymbol{x})$ denote the number of pairs 00 or 11 in $\boldsymbol{x}$. We can calculate the length of $g(\boldsymbol{x})$:

$$|g(\boldsymbol{x})| = 6|\boldsymbol{x}|-\lambda(\boldsymbol{x})-2(|\boldsymbol{x}|-\lambda(\boldsymbol{x})-1) = 4|\boldsymbol{x}|+\lambda(\boldsymbol{x})+2 \tag{2}$$

the number of pairs 00 or 11 in $g(\boldsymbol{x})$:

$$\lambda(g(\boldsymbol{x})) = |\boldsymbol{x}| \tag{3}$$

the number of runs in $g(\boldsymbol{x})$:

$$\boldsymbol{r}(g(\boldsymbol{x})) = \boldsymbol{r}(\boldsymbol{x})+2|\boldsymbol{x}|+(|\boldsymbol{x}|-1) = \boldsymbol{r}(\boldsymbol{x})+3|\boldsymbol{x}|-1 \tag{4}$$

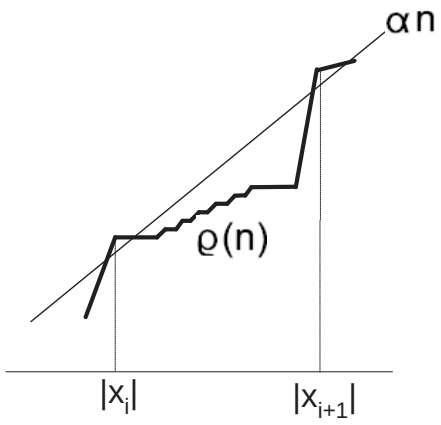

Figure 1. $\boldsymbol{\rho}(n)$ function between $|\boldsymbol{x}_i|$ and $|\boldsymbol{x}_{i+1}|$

In [3] a sequence of strings was generated inductively from a starting string, for instance: $\boldsymbol{x}_0 = 0$, $\boldsymbol{x}_1 = g(0) = 010010$, and $\boldsymbol{x}_{i+1} = g(\boldsymbol{x}_i)$ for $i \geq 1$. Then $|\boldsymbol{x}_{i+1}| = 4|\boldsymbol{x}_i|+|\boldsymbol{x}_{i-1}|+2$ according (2) and $\boldsymbol{r}(\boldsymbol{x}_{i+1}) = \boldsymbol{r}(\boldsymbol{x}_i)+3|\boldsymbol{x}_i|-1$ according to (4). It is not

hard to show that the limit $\lim_{i\to\infty}\frac{|\boldsymbol{x}_i|}{|\boldsymbol{x}_{i+1}|}$ exists and $\beta = \lim_{i\to\infty}\frac{|\boldsymbol{x}_i|}{|\boldsymbol{x}_{i+1}|} = -2+\sqrt{5}$. The limit $\lim_{i\to\infty}\frac{\boldsymbol{r}(\boldsymbol{x}_i)}{|\boldsymbol{x}_i|}$ also exists and $\alpha = \lim_{i\to\infty}\frac{\boldsymbol{r}(\boldsymbol{x}_i)}{|\boldsymbol{x}_i|} = \beta(\alpha+3)$ giving $\alpha = \frac{3}{1+\sqrt{5}}$.

The sequence $\{|\boldsymbol{x}_i| \ : \ i < \infty\}$ is only "probing" the domain of the function $\boldsymbol{\rho}(n)$ and $\boldsymbol{r}(\boldsymbol{x}_i)$ is "pushing" the value of $\boldsymbol{\rho}(n)$ above αn in these "probing" points (see Figure 1). Since the size of $\boldsymbol{x}_{i+1}$ is more than 4 times the size of $\boldsymbol{x}_i$, the gaps between $|\boldsymbol{x}_i|$ and $|\boldsymbol{x}_{i+1}|$ are getting bigger and bigger.

The basic idea of establishing an asymptotic lower bound for $\boldsymbol{\rho}(n)$ is to start similar sequences from various "starting" strings to cover the domain of $\boldsymbol{\rho}(n)$ densely enough with the "probing" points to get any n close to some "probing" point and hence the value of $\boldsymbol{\rho}(n)$ close to αn. To be able to do so, we must change a bit the way the sequences are generated. The details of this are in the next section.

3 The proof of the theorem

Let $\varepsilon > 0$ be given. We have to find N so that for any $n \geq N$, $\boldsymbol{\rho}(n) \geq (\alpha-\varepsilon)n$.

Firs we will chose and fix three parameters k, δ, and R that we will use throughout the proof. These parameters depend on the given ε: choose and fix a positive integer k so that $\frac{\alpha}{k+1} < \varepsilon$; choose and fix a positive real δ so that $\delta \leq \frac{k+1}{k}(\varepsilon-\frac{\alpha}{k+1})$. It follows that $\frac{k}{k+1}(\alpha-\delta) \geq \alpha-\varepsilon$. Let R be the smallest integer so that $\left(\frac{k+1}{k}\right)^R \geq 5$.

Consider an increasing sequence $\mathcal{S}_{a,b}$ of positive integers with two integer parameters a and b defined by $n_0(a,b) = a$, $n_1(a,b) = 4a+b$, and $n_{i+2}(a,b) = 4n_{i+1}(a,b)+n_i(a,b)$ for $i \geq 0$. It is not hard to show that $lim_{i\to\infty}\frac{n_i(a,b)}{n_{i+1}(a,b)}$ exists and that

$$lim_{i\to\infty}\frac{n_i(a,b)}{n_{i+1}(a,b)} = -2+\sqrt{5}$$

Importantly, ranges of such sequences are "tied" together based on the parameters, i.e. for any integer $t \geq 1$ and any i

$$n_i(ta,tb) = tn_i(a,b). \tag{5}$$

For $0 \leq j < R$, set

$$a(j) = 3(k+1)^j k^{(R-j)} \text{ and } b(j) = \frac{a(j)}{3} = (k+1)^j k^{(R-j)}. \tag{6}$$

It follows that $\frac{k+1}{k}a(j) = a(j+1)$, $\frac{k+1}{k}b(j) = b(j+1)$, and $b(j) \geq 3$.

Based on the morphism $g(\boldsymbol{x})$ (see (1)) we define a new morphism $h(\boldsymbol{x})$ by removing the last 2 letters from $g(\boldsymbol{x})$:

$$\text{if } g(\boldsymbol{x}) = y[1..n], \text{then } h(\boldsymbol{x}) = \boldsymbol{y}[1..n-2] \tag{7}$$

We use the term *string* $\boldsymbol{s}$ *ends with a square* to indicate that $\boldsymbol{s}$ has a left-maximal square as its suffix. We call a string *good* if it ends with at most one square.

Claim: (a) if $\boldsymbol{x}$ *is good, then* $h(\boldsymbol{x})$ *is good*

(b) if $\boldsymbol{x}$ *ends with 011, then* $h(\boldsymbol{x})$ *ends with 011*

(c) if $\boldsymbol{x}$ *is good, then* $\boldsymbol{r}(g(\boldsymbol{x})) \geq \boldsymbol{r}(h(\boldsymbol{x})) \geq \boldsymbol{r}(g(\boldsymbol{x})) - 2$.
(*the claim will be proven after completing the proof of the theorem*)

Now we are in the position to define the "probing" sequences.
For any $0 \leq j < R$ we define a sequence of binary strings $\{\boldsymbol{x}_i(j) \ : \ i < \infty\}$ by:

$$\boldsymbol{x}_0(j) = (011)^{b(j)}$$

and for any $i \geq 0$,

$$\boldsymbol{x}_{i+1}(j) = h(\boldsymbol{x}_i(j))$$

where $b(j)$ is defined in (6). From (2) and (4) it follows that for any $i \geq 0$,

$$|\boldsymbol{x}_0(j)| = 3b(j) = a(j),$$

$$|\boldsymbol{x}_1(j)| = 4a(j)+b(j), \text{and}$$

$$|\boldsymbol{x}_{i+2}(j)| = 4|\boldsymbol{x}_{i+1}(j)|+|\boldsymbol{x}_i(j)|.$$

Thus, the sequence $\{|\boldsymbol{x}_i(j)| \ : \ i < \infty\}$ is the $\mathcal{S}_{a(j),b(j)}$ sequence and so $\lim_{i\to\infty} \frac{|\boldsymbol{x}_i(j)|}{|\boldsymbol{x}_{i+1}(j)|} = -2+\sqrt{5}$.

Since our starting string $\boldsymbol{x}_0(j)$ is good as it equals $(011)^{b(j)}$ and $b(j) \geq 3$, according to the *Claim*, every $\boldsymbol{x}_i(j)$ is good and ends with 011, and

$$\boldsymbol{r}(g(\boldsymbol{x}_i(j))) \geq \boldsymbol{r}(\boldsymbol{x}_{i+1}(j)) \geq \boldsymbol{r}(g(\boldsymbol{x}_i(j)))-2$$

and so

$$\lim_{i\to\infty} \frac{\boldsymbol{r}(\boldsymbol{x}_i(j))}{|\boldsymbol{x}_i(j)|} = \alpha.$$

Therefore, for any $0 \leq j < R$ there is a positive integer I_j so that for any $i \geq I_j$,

$$\frac{\boldsymbol{\rho}(|\boldsymbol{x}_i(j)|)}{|\boldsymbol{x}_i(j)|} \geq \frac{\boldsymbol{r}(\boldsymbol{x}_i(j))}{|\boldsymbol{x}_i(j)|} \geq \alpha-\delta.$$

Let $I = \max\{I_j \ : \ 0 \leq j < R\}$. Then for any $i \geq I$ and any $0 \leq j < R$,

$$\frac{\boldsymbol{\rho}(|\boldsymbol{x}_i(j)|)}{|\boldsymbol{x}_i(j)|} \geq \frac{\boldsymbol{r}(\boldsymbol{x}_i(j))}{|\boldsymbol{x}_i(j)|} \geq \alpha-\delta. \tag{8}$$

From (5) and (6) it follows, that for any i and any $0 \leq j < R$,

$$n_i(a(j), b(j)) = \big(\frac{k+1}{k}\big) n_i(a(j-1), b(j-1)) = \cdots = \big(\frac{k+1}{k}\big)^j n_i(a(0), b(0)).$$

Set $N = \max\{n_I(a(j), b(j)) \ : \ 0 \leq j < R\}$. This is the N we were searching for.

If $n \geq N$, then for some $i \geq I$,

$$n_i(a(0), b(0)) < n \leq n_{i+1}(a(0), b(0)).$$

Then there is $0 \le j < R-1$ so that

$$\left(\frac{k+1}{k}\right)^j n_i(a(0),b(0)) < n \le \left(\frac{k+1}{k}\right)^{j+1} n_i(a(0),b(0))$$

$\big[$ since $\left(\frac{k+1}{k}\right)^R \ge 5$, then $\left(\frac{k+1}{k}\right)^R n_i(a(0),b(0)) \ge n_{i+1}(a(0),b(0))$ $\big]$.
It follows that

$$n_i(a(j),b(j)) < n \le \frac{k+1}{k} n_i(a(j),b(j)).$$

Now we can estimate the value of $\frac{\boldsymbol{\rho}(n)}{n}$ using (8):

$$\frac{\boldsymbol{\rho}(n)}{n} \ge \frac{\boldsymbol{\rho}(n_i(a(j),b(j)))}{n} \ge \frac{k}{k+1} \frac{\boldsymbol{\rho}(n_i(a(j),b(j)))}{n_i(a(j),b(j))} \ge \frac{k}{k+1}(\alpha-\delta) \ge \alpha-\varepsilon.$$

Thus $\boldsymbol{\rho}(n) \ge (\alpha-\varepsilon)n$. □

Proof. **of** *Claim*
Let us assume that $\boldsymbol{x}$ ends with 011. Then $g(\boldsymbol{x})$ ends with 010010110101101, and so $h(\boldsymbol{x})$ ends with 0100101101011. Consider all runs in $g(\boldsymbol{x})$ that may be "destroyed" by removing the last 2 letters from $g(\boldsymbol{x})$:
(a) if $\boldsymbol{x}$ ends with a square, then the square may induce a left-maximal square in $g(\boldsymbol{x})$ and it will be "destroyed". Since $\boldsymbol{x}$ is good, there may be at most 1 such run destroyed.
(b) $g(\boldsymbol{x})$ ends with square 101|101 that will get destroyed.
(c) The run 01011|01011|01 in $g(\boldsymbol{x})$ becomes a left-maximal square suffix in $h(\boldsymbol{x})$.
No other runs in $g(\boldsymbol{x})$ are affected. Hence $h(\boldsymbol{x})$ is good and at most 2 runs in $g(\boldsymbol{x})$ are destroyed.

4 Conclusion and further research

We showed that the expectation of αn being a lower bound for the maxrun function $\boldsymbol{\rho}(n)$ is valid by proving that there is a whole family of asymptotic lower bounds arbitrarily close to αn. The further research will include trying to push the lower bound higher up to see whether the conjecture $\boldsymbol{\rho}(n) < n$ holds. This will involve finding novel ways of creating strings "rich in runs" as the approach with concatenation $\circ$ seems to give as much as it could.

References

[1] M. Crochemore: *An optimal algorithm for computing the repetitions in a word*. Inform. Process. Lett., 5(5) 1981, pp. 297–315.
[2] Fan Kangmin and W. F. Smyth: *A new periodicity lemma*. to appear in SIAM J. of Discr. Math.
[3] F. Franek, J. Simpson, and W. F. Smyth: *The maximum number of runs in a string*, in Proceedings of 14th Australasian Workshop on Combinatorial Algorithms AWOCA 2003, Seoul National University, Seoul, Korea, July 13-16 2003.
[4] R. Kolpakov and G. Kucherov: *On maximal repetitions in words*. J. of Discrete Algorithms, (1) 2000, pp. 159–186.
[5] M. G. Main: *Detecting leftmost maximal periodicities*. Discrete Applied Maths., (25) 1989, pp. 145–153.
[6] S. J. Puglisi, W. F. Smyth, and A. Turpin: *Some restrictions on periodicity in strings*, in Proceedings of 16th Australasian Workshop on Combinatorial Algorithms AWOCA 2005, University of Ballarat, Victoria, Australia, September 18-21 2005, pp. 263–268.
[7] W. Rytter: *The number of runs in a string: Improved analysis of the linear upper bound*, in Proceedings of 23rd Annual Symposium on Theoretical Aspects of Computer Science STACS 2006, Marseille, France, February 23-25 2006, pp. 184–195.