Not so many runs in strings

Mathieu Giraud

Not so many runs in strings*

Mathieu Giraud

CNRS, LIFL, INRIA, Université Lille 1
mathieu.giraud@lifl.fr

Abstract. Since the work of Kolpakov and Kucherov in [5, 6], it is known that $\rho(n)$, the maximal number of runs in a string, is linear in the length n of the string. A lower bound of $3/(1+\sqrt{5})n \sim 0.927n$ has been given by Franek and al. [3, 4], and upper bounds have been recently provided by Rytter, Puglisi and al., and Crochemore and Ilie ($1.6n$) [8, 7, 1]. However, very few properties are known for the $\rho(n)/n$ function. We show here by a simple argument that $\lim_{n\mapsto\infty} \rho(n)/n$ exists and that this limit is never reached. Moreover, we further study the asymptotic behavior of $\rho_p(n)$, the maximal number of runs with period at most p. We provide a new bound for some microruns : we show that there is no more than $0.971n$ runs of period at most 9 in binary strings. Finally, this technique improves the previous best known upper bound, showing that the total number of runs in a binary string of length n is below $1.52n$.

1 Introduction

The study of repetitions is an important field of research, both for word combinatorics theory and for practice, with applications in domains like computational biology or cryptanalysis. The notion of *run* (also called maximal repetition or m-repetition [5]) allows a compact representation of the set of all tandem periodicities, even fractional, in a string. The proper counting of those runs is important for all algorithms dealing with repetitions.

Since the work of Kolpakov and Kucherov in [5, 6], it is known that $\rho(n)$, the maximal number of runs in a string, is linear in the length n of the string. They gave the first algorithm computing all runs in a linear time, but without an actual constant.

Upper bounds have been recently provided by Rytter ($5n$) [8] and Puglisi, Simpson, and Smyth ($3.48n$) [7]. The best upper bound known today, $1.6n$, was obtained by Crochemore and Ilie [1]. They count separately the *microruns*, that is the runs with short periods, and the runs with larger ones. Crochemore and Ilie show that the number of microruns with period at most 9 verifies $\rho_9(n) \leq n$. For larger runs, they prove that

$$\rho_{\geq p}(n) \leq \frac{2}{p}\left(\sum_{i=0}^{\infty}\left(\frac{2}{3}\right)^i\right)n = \frac{6}{p}\cdot n$$

* Int. Conf. on Language and Automata Theory and Applications (LATA 2008), 2008

n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
$\rho(n)$	2	3	4	5	5	6	7	8	8	10	10	11	12	13	14	15	15	16	17	18	19	20	21	22	23	24	25

Table 1. Values of $\rho(n)$ for small values of n for binary strings, from [5].

A lower bound of αn, with $\alpha = 3/(1+\sqrt{5}) = 0.927...$, has been given by [3] then [4]. In [3], Franek, Simpson and Smyth propose a sequence of strings (x_n) with increasing lengths such that $\lim_{n\mapsto\infty} r(x_n)/|x_n| = \alpha$, where $r(x)$ is the number of runs in the string x. In [4], Franek and Yang show that α is an asymptotic lower bound by showing that there exists a whole family of asymptotic lower bounds arbitrarily close to α.

In fact, very few properties are known for the $\rho(n)/n$ function [4, 9]. In this paper, after giving some definitions (Section 2), we show by a simple rewriting argument that $\ell = \lim_{n\mapsto\infty} \rho(n)/n$ exists and that this limit is never reached (Section 3), proving that

$$\frac{\rho(n)}{n} \leq \ell - \frac{1}{4n}$$

Section 4 proves the convergence of $\rho(n)/n$ even in the case of a fixed alphabet, for example for binary strings. Moreover, we further study the asymptotic behavior of $\rho_p(n)$, the number of runs with short periods (Section 5), showing that $\ell_p = \lim_{n\mapsto\infty} \rho_p(n)/n$ exists and that, for some constant z_p,

$$\ell_p - \frac{z_p}{n} \leq \frac{\rho_p(n)}{n} \leq \ell_p \leq \ell$$

Practically, this inequality implies that the count of some microruns is *below* n, and this improves the upper bound of [1] to $1.52n$ for binary strings. Section 6 gives some concluding remarks.

2 Definitions

Let $x = x_1 x_2 \dots x_n$ be a string over an alphabet. Let $p \geq 1$ be an integer. We say that x has a *period* p if for any i with $1 \leq i \leq n-p$, $x_{i+p} = x_i$. We denote by $x[i...j]$ the substring $x_i x_{i+1} \dots x_j$. A *run* is a substring $x[i...j]$:

- which has a period $p \leq \lfloor (j-i+1)/2 \rfloor$,
- that is maximal : if they exist, neither $x_{i-1} = x_{i-1+p}$, nor $x_{j+1} = x_{j+1-p}$,
- and such that $x[i...i+p-1]$ is primitive : it is not an integer power of another string.

We define by $r_p(x)$ the number of runs of period $\leq p$ in x, called *microruns* in [1], and by $r(x) = r_{\lfloor |x|/2 \rfloor}(x)$ the total number of runs in x. For example, the four runs of $x = \mathsf{atattatt}$ are $x[4,5] = \underline{\mathsf{tt}}$, $x[7,8] = \underline{\mathsf{tt}}$, $x[1,4] = \underline{\mathsf{at}}\mathsf{at}$ and $x[2,8] = \underline{\mathsf{tat}}\mathsf{tatt}$, and thus $r_1(x) = 2$, $r_2(x) = 3$, and $r_3(x) = r(x) = 4$.

Given an integer $n \geq 2$, we now consider all strings of length n. We define as

$$\rho_p(n) \;=\; \max\{r_p(x) \mid |x| = n\}$$

the maximal number of runs of period $\leq p$ in a string of length n. Then we define as

$$\rho(n) \;=\; \max\{r(x) \mid |x| = n\} \;=\; \rho_{\lfloor n/2 \rfloor}(n)$$

the maximal total number of runs in a string of length n. Kolpakov and Kucherov gave in [6] some values for $\rho(n)$ (Table 1). Table 3, at the end of this paper, shows some values for $\rho_p(n)$. Note that $r(x) = \rho(|x|)$ does not imply that $r_p(x) = \rho_p(|x|)$ for all p : for example, $r(\mathsf{aatat}) = 2 = \rho(5)$ but $r_1(\mathsf{aatat}) = 1 < \rho_1(5) = 2$.

Finally, we can define values $r_{\geq p}(x)$ and $\rho_{\geq p}(n)$ for *macroruns*, that is runs with a period at least p. Again, $r(x) = \rho(|x|)$ does not imply that $r_{\geq p}(x) = \rho_{\geq p}(|x|)$. For example, $r_{\geq 2}(\mathsf{aatt}) = 0 < \rho_{\geq 2}(4) = 1 = r_{\geq 2}(\mathsf{atat})$.

3 Asymptotic behavior of the number of runs

Franek and al. [3, 4] list some known properties for $\rho(n)$:

- For any n, $\rho(n+2) \geq \rho(n) + 1$
- For any n, $\rho(n+1) \leq \rho(n) + \lfloor n/2 \rfloor$
- For some n, $\rho(n+1) = \rho(n)$
- For some n, $\rho(n+1) = \rho(n) + 2$

We add the following two simple properties.

Proposition 1. *The function ρ is superadditive : for any m and n, we have $\rho(m+n) \geq \rho(m) + \rho(n)$.*

Proof. Take two strings x et y of respective lengths m and n such that $r(x) = \rho(m)$ and $r(y) = \rho(n)$. Let $\bar{y}$ be a rewriting of y with characters not present in x. Then $x\bar{y}$ is a string of length $m+n$ containing exactly the runs of x and the rewritten runs of y. Thus $\rho(m+n) \geq r(x\bar{y}) = r(x) + r(y) = \rho(m) + \rho(n)$.

Proposition 2. *For any n, $\rho(4n) \geq 4\rho(n) + 1$*

Proof. Take a string x of length n with $r(x) = \rho(n)$. Let $\bar{x}$ be a rewriting of x with characters not present in x. Then $r(x\bar{x}x\bar{x}) \geq 4r(x) + 1$.

We have in particular $\rho(tn) \geq t\rho(n)$, giving our main result :

Theorem 1. *$\rho(n)/n$ converges to its upper limit ℓ. Moreover, the limit is never reached, as for any n we have*

$$\frac{\rho(n)}{n} \;\leq\; \ell - \frac{1}{4n}$$

Proof. Let ℓ be the upper limit of $\rho(n)/n$. This limit is finite because of [6]. Given ε, there is a n_0 such that $\rho(n_0)/n_0 \geq \ell - \varepsilon/2$. For any $n \geq n_0$, let be $t = \lfloor n/n_0 \rfloor$. Then we have $\rho(n)/n \geq \rho(tn_0)/n \geq t\rho(n_0)/n$ by Proposition 1, thus $\rho(n)/n \geq t/(t+1) \cdot \rho(n_0)/n_0$. Let be t_0 such that $t_0/(t_0+1) \cdot \rho(n_0)/n_0 \geq \rho(n_0)/n_0 - \varepsilon/2$. Then, for any $n \geq t_0 n_0$, we have $\rho(n)/n \geq \ell - \varepsilon$, thus $\ell = \lim_{n \mapsto \infty} \rho(n)/n$. Finally, Proposition 2 gives $\ell \geq \rho(4n)/4n \geq \rho(n)/n + \frac{1}{4n}$.

The proof of convergence of $f(n)/n$ when f is superadditive is known as Fekete's Lemma [2, 10]. This convergence result was an open question of [4]. In fact, the motivation of [4] was the remark that *"the sequence $|x_i|$ (of [3]) is only "probing" the domain of the function $\rho(n)$ and $r(x_i)$ is "pushing" the value of $\rho(n)$ above αn in these "probing" points"*. Then Franek and Yang [4] prove that every $\alpha - \varepsilon$ is an actual asymptotic lower bound by building specific sequences. With Theorem 1, it is now sufficient to study bounds on any $(\rho(n_i)/n_i)$ sequence (for a growing sequence (n_i)) to give bounds on $\rho(n)/n$.

Note that this convergence does not imply monotonicity. In fact, if $\ell < 1$, then $\rho(n)/n$ is asymptotically non monotonic, as there will be in this case an infinity of n's such that $\rho(n+1) = \rho(n)$. Note also that, although Proposition 1 and 2 require to double the alphabet size, the alphabet remains finite : the proof of Theorem 1 only requires to double once this alphabet size. Moreover, it is possible to prove Proposition 1 without rewriting in a larger alphabet, thus proving the convergence of $\rho(n)/n$ when considering only binary strings. This second proof, more elaborated, is given in the next section.

The bound $\ell - \frac{1}{4n}$ can be improved. For example, with a rewriting similar to the one used in Proposition 2, it can be shown that $\rho(2n^2) \geq (2n+1)\rho(n)$, giving by successive iterations $\rho(n)/n \leq \ell - \frac{1}{2n}$. This has not been reported here to keep the proof simple.

Concerning microruns with period at most p, Proposition 1 still holds :

Proposition 3. *For any p, m, and n, we have $\rho_p(m+n) \geq \rho_p(m) + \rho_p(n)$. Thus for any p, $\rho_p(n)/n$ converges to its upper limit ℓ_p.*

The proof is the same as above. On the contrary, Proposition 2 may be not true for microruns. For example, for any n, $\rho_1(n) = \lfloor n/2 \rfloor$, and thus for any even n, we have $\rho_1(n)/n = \ell_1 = 1/2$.

Finally, Theorem 1 is fully valid for macroruns, and as a rewriting argument shows that $\rho_{\geq p}(n) \geq \rho(\lfloor n/p \rfloor)$, we have $\ell_{\geq p} \geq \ell/p$.

4 A proof of Proposition 1 for fixed alphabets

Here we prove Proposition 1 without rewriting in a larger alphabet, thus proving the convergence of $\rho(n)/n$ when considering only binary strings. This proof is borrowed and simplified from one part of a proof of Franek and al. (Theorem 2 of [3]). A key observation is that some runs of x and y are merged in xy only when a word z^2 is both a suffix of x and a prefix of y (case a_2 on Figure 1). We first have this property :

Proposition 4. *Let Σ be an alphabet with $|\Sigma| \geq 2$, and let x and y be strings on Σ such that $|y| \geq 1$. Then there exists strings x' and y' on Σ such that $|x'| + |y'| = |x| + |y|$, $|y'| < |y|$ and $r(x') + r(y') \geq r(x) + r(y)$.*

Proof. Let w be the largest string, eventually empty, such that w is a suffix of x and a prefix of y. Thus $x = uw$ and $y = wv$ for some strings u and v. Let $x' = uwv$ and $y' = w$. Clearly $|x'| + |y'| = |x| + |y|$ and $|y'| \leq |y|$. Without loss of generality, we assume that y is not a suffix of x. (If it is not the case, we rewrite y into $\bar{y}$ using an isomorphism of Σ onto itself.) Thus $|y'| < |y|$. Now we count the runs of x and y. The runs of period p that have $2p$ characters ("a square") completely included in w were once in x and once in y. Such runs can be found again once in x' and once in y'. By definition of w, all the others runs of x and y are found exactly once in x', without being merged.

To prove Proposition 1, we take two strings x_0 and y_0 of respective lengths m and n such that $r(x_0) = \rho(m)$ and $r(y_0) = \rho(n)$. Applying recursively Proposition 4 gives a finite sequence of pairs of strings $(x_0, y_0), (x_1, y_1), \ldots (x_t, y_t)$ with $r(x_i) + r(y_i) \geq r(x_{i-1}) + r(y_{i-1})$ and $|y_0| > |y_1| > \ldots > |y_t| = 0$ for some t. Finally $|x_t| = |x_0| + |y_0| = m + n$, and thus $\rho(m+n) \geq r(x_t) \geq r(x_0) + r(y_0) = \rho(m) + \rho(n)$, proving Proposition 1.

Note that the proof of Franek and al. in [3] was in a different context, and that no result leading to our Proposition 1 was stated as such in their paper.

5 On the number of microruns

In this section, we study the asymptotic behavior of the number of microruns beyond the result of Proposition 3. Additionally, we provide a new bound on the number of some microruns (see the end of the section).

The idea to bound the number of microruns is to count the *new runs* created by the concatenation of two strings. Let x and y be two strings, and s be a run of xy with period q. Then s is exactly in one of the following two cases (Figure 1) :

- a) s has a substring that is a run (with the same period q) completely included in x, or in y, or in both;
- b) s has strictly less than 2 periods in x and in y.

We call the runs in the case b) the *new runs* between x and y, and we denote by $z_p(x, y)$ the number of such runs. Then $r_p(xy) \leq r_p(x) + r_p(y) + z_p(x, y)$, the inequality coming from the fact that a run from x can be merged with a run from y (case a_2 on Figure 1). We can bound the number of new runs, and thus have an upper bound on $r_p(xy)$:

Proposition 5. *Let $z_p = \max\{z_p(x, y) \mid |x| = |y| = 2p-1\}$ the maximal number of new runs between words of length $2p - 1$. Then, for every strings x and y of any length, we have $z_p(x, y) \leq z_p$.*

Proposition 6. *For any p, m, and n, $\rho_p(m + n) \leq \rho_p(m) + \rho_p(n) + z_p$.*

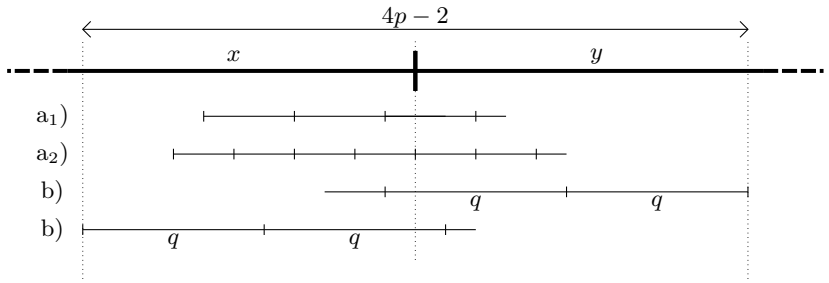

Fig. 1. a$_1$) Runs with a substring that is a run included in x. a$_2$) Runs with two substrings that are runs both in x and y. b) "New runs" between x and y. To count the new runs with period $q \leq p$, it is sufficient to consider words of length $4p-2$.

Proof. (Proposition 5.) Any new run with period $q \leq p$ has less than $2q-1 \leq 2p-1$ characters in x, and in y (Figure 1). (Proposition 6.) Let x and y be two strings such that $|x| = m$, $|y| = n$, and $r_p(xy) = \rho_p(m+n)$. Then $\rho_p(m+n) = r_p(xy) \leq r_p(x) + r_p(y) + z_p(x,y) \leq \rho_p(m) + \rho_p(n) + z_p$.

Table 2 provides actual values of z_p for small values of p. An immediate bound on z_p is $z_p \leq z_{p-1} + 2$. Knowing bounds on z_p helps to characterize the asymptotic behavior of the number of microruns :

Theorem 2. *For any p and n, we have $\ell_p \leq \rho_p(n)/n + z_p/n$, and thus*

$$\ell_p - \frac{z_p}{n} \leq \frac{\rho_p(n)}{n} \leq \ell_p \leq \ell$$

Proof. By Proposition 6, for any $t \geq 1$, we have $\rho_p(tn) \leq t\rho_p(n) + (t-1)z_p$. Thus $\rho_p(tn)/tn \leq \rho_p(n)/n + \frac{t-1}{t} z_p/n$. Taking this inequality to the limit gives the result.

Thus we know that the convergence of $\rho_p(n)/n$ to ℓ_p is faster than z_p/n. Note that we do not have a similar result for $\rho(n)$, as we do not have a convenient way to bound $\rho(m+n)$ like in Proposition 5.

With Theorem 2, one can show that the number of some microruns is *below* n. For example, we have for binary strings $z_9 = 7$ and $\rho_9(34) = 26$, thus $\ell_9 \leq 33/34 = 0.970...$. Thus *there are less than* $0.971n$ *runs of period* ≤ 9 *in any binary string of length* n. This result is better than Lemma 2 of [1] which proved the n bound by the count of amortizing positions for centers of runs.

Using the result of Crochemore and Ilie's Proposition 1 [1] for large runs, we get an upper bound on $\rho(n)/n$. The best bound we obtain, with $z_{10} = 7$ and $\rho_{10}(34) = 26$, gives $\ell_{10} \leq 33/34$, and finally, with Crochemore and Ilie's Proposition 1 :

$$\ell \leq \ell_{10} + \frac{6}{11} = 1.516...$$

p	z_p	example
1, 2	1	t t
3	2	ttat ta
4	4	ataaata attaat
5, 6, 7	5	ttatatta taatataa
8	6	ttttattattttat taattattaa
9, 10	7	ttatatatattatata taatatatataa

Table 2. Values for z_p for binary strings with worst-case examples of length $\leq 4p-2$.

Thus *the number of runs in a binary string of length n is not more than* $1.52n$. This result could be further extended by choosing other periods for the count of microruns, by computation or by other techniques.

6 Perspectives

The results on the asymptotic behavior of the functions ρ and ρ_p of Theorems 1 and 2 simplify the research on lower and upper bounds. We hope that these results will bring a better understanding of the number of runs and be a step towards proving the conjecture of [5] ($\ell \leq 1$) or the stronger conjecture of [3] ($\ell = 0.927...$).

A side result of Theorem 2 was a new upper bound for some microruns, and thus an upper bound for the total number of runs. This upper bound can be lowered again by doing a more precise analysis, theoretical or computational, of the z_p values. This would require large evaluations of some z_p and $\rho_p(n)$ values. Other techniques could provide better bounds for microruns. For example, it should be possible to push the idea of Crochemore and Ilie by finding more amortizing positions than the number of centers of runs in a given interval of positions. Again, when the number of possible positions grows, the complexity of their method increases.

For the lower bound, it remains to be shown if one can find strings with more runs than those of [3, 4]. Although Theorem 1 also provides a way to have a lower bound on $\rho(n)/n$, all the computations we ran gave not better bounds than the 0.927 bound of [3, 4].

Now an important question is if the actual value of ℓ can be found with such a separation between microruns and macroruns. The inequality $\ell \leq \ell_p + \ell_{\geq p+1}$ may be strict for some p, as the values $\ell_1 = 1/2$ and $\ell_{\geq 2} > 1/2$ may suggest. If this inequality is strict for several p's, the conjecture may be impossible to prove by this way if one choose a bad splitting period p.

Another open question is if one of the constants $\ell_p = \lim_{n \mapsto \infty} \rho_p(n)/n$ is equal to ℓ, or if, more probably, the limit ℓ is obtained by considering asymptotically runs with any period. Finally, it remains to be proven if strings on binary alphabets can always achieve the highest number of runs.

n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
$\rho(n)$	2	3	4	5	5	6	7	8	8	10	10	11	12	13	14	15	15	16	17	18	19	20	21	22	23	24	25	26	27	27	28
1, 2	**2**	**3**	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	12	13	13	14	14	15	15	16	16	17	17
3		3	**4**	4	**5**	**6**	6	7	**8**	9	9	10	11	12	12	13	14	15	15	16	17	18	18	19	20	21	21	22	23	24	24
4			**5**	5	6	**7**	**8**	8	9	9	10	11	12	13	13	14	15	16	16	17	18	18	19	20	21	21	22	23	24	24	
5					6	7	8	8	9	**10**	**11**	11	12	13	14	14	15	16	17	18	18	19	20	21	21	22	23	24	25	25	
6							8	8	9	10	11	11	12	13	14	14	15	16	17	18	18	19	20	21	21	22	23	24	25	25	
7									**10**	10	11	**12**	**13**	**14**	14	**15**	15	16	**18**	18	19	20	21	21	22	23	24	25	26	26	
8											11	12	13	14	**15**	15	**16**	**17**	18	18	19	20	21	22	23	24	25	26	26	27	
9													13	14	15	15	16	17	18	18	19	20	21	22	23	24	25	26	26	27	
10															15	15	16	17	18	18	19	20	21	22	23	24	25	26	26	27	
11																	16	17	18	**19**	19	**21**	21	22	23	24	25	26	26	27	
12																				18	19	19	21	21	22	23	24	25	26	26	27
13																						**20**	21	**22**	**23**	**24**	**25**	**26**	**27**	**27**	**28**
14																								22	23	24	25	26	27	27	28

Table 3. Values of $\rho_p(n)$ for small values of n and p for binary strings. For each n, the value in bold shows the smallest p such that $\rho_p(n) = \rho(n)$.

References

1. Maxime Crochemore and Lucian Ilie. Maximal repetitions in strings. *J. Comput. Systems Sci.*, 2007. (in press).
2. M. Fekete. Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Mathematische Zeitschrift*, 17:228–249, 1923.
3. Frantisek Franek, R. J. Simpson, and W. F. Smyth. The maximum number of runs in a string. In *Proceedings of the 2003 Australasian Workshop on Combinatorial Algorithms (AWOCA 03)*, pages 26–35, 2003.
4. Frantisek Franek and Qian Yang. An asymptotic lower bound for the maximal-number-of-runs function. In *Prague Stringology Conference 2006*, pages 3–8, 2006.
5. Roman Kolpakov and Gregory Kucherov. Maximal repetitions in words or how to find all squares in linear time. Technical Report 98-R-227, LORIA, 1998.
6. Roman Kolpakov and Gregory Kucherov. On maximal repetitions in words. *Journal on Discrete Algorithms*, 1(1):159–186, 2000.
7. S. J. Puglisi, J. Simpson, and B. Smyth. How many runs can a string contain? *Theoretical Computer Science*, 2008. (in press).
8. Wojciech Rytter. The number of runs in a string : improved analysis of the linear upper bound. *Information and Computation*, 205(9):1459–1469, 2007.
9. Bill Smyth. The maximum number of runs in a string. In *International Workshop on Combinatorial Algorithms (IWOCA 2007), Problems Session*, 2007.
10. J. L. van Lint and R. M. Wilson. *A course in combinatorics*. Cambridge University Press, 1992.