Asymptotic behavior of the numbers of runs and microruns

Mathieu Giraud

Asymptotic behavior of the numbers of runs and microruns[☆]

Mathieu Giraud[a,b]

[a] *CNRS, LIFL, Université Lille 1, 59 655 Villeneuve d'Ascq cedex, France*
[b] *INRIA Lille Nord-Europe, 59 650 Villeneuve d'Ascq, France*

Abstract

The notion of *run* (also called maximal repetition) allows a compact representation of the set of all tandem periodicities, even fractional, in a string. Since the work of Kolpakov and Kucherov in [8, 9], it is known that $\rho(n)$, the maximum number of runs in a string, is linear in the length n of the string. Lower bounds haven been provided by Franek et al. and Matsubara et al. (0.9445...) [5, 6, 10], and upper bounds have been provided by Rytter, Puglisi et al., and Crochemore and Ilie ($1.048n$) [12, 11, 1, 2]. 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. We further study the asymptotic behavior of $\rho_p(n)$, the maximal number of runs with period at most p. Finally, we provide the first exact limits for some microruns. For example, we have $\lim_{n\mapsto\infty} \rho_{14}(n)/n = 15/17$.

Key words: word combinatorics, maximal repetitions, runs, asymptotic behavior, maximum number of runs

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 [8]) 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 [8, 9], it is known that $\rho(n)$, the maximum 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.

[☆] A preliminary version of this paper appeared in [3]. All the results in Section 4.2, with the exact limits given in Table 3, are new for this extended article.

Email address: mathieu.giraud@lifl.fr (Mathieu Giraud)

n	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\rho(n)$	2	3	4	5	5	6	7	8	8	10	10	11	12	13

n	19	20	21	22	23	24	25	26	27	28	29	30	31
$\rho(n)$	14	15	15	16	17	18	19	20	21	22	23	24	25

Table 1: Values of $\rho(n)$ for binary strings, from [8].

Upper bounds have been recently provided by Rytter ($5n$) [12] and Puglisi, Simpson, and Smyth ($3.48n$) [11]. A $1.6n$ bound 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. They show that the number of microruns with period at most 9 satisfies $\rho_9(n) \leq n$. For larger runs, they prove that

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

Crochemore, Ilie, and Tinta extended those results with massive computations, bringing down the upper bound from $1.6n$ to $1.048n$ [2].

A lower bound of αn, with $\alpha = 3/(1+\sqrt{5}) = 0.927...$, has been given by [5] then [6]. In [5], 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 [6], 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. Recently, Matsubara et al. provided an $174719/184973 = 0.9445..$ lower bound by repeating a large run-rich string [10].

In fact, very few properties are known for the $\rho(n)/n$ function [6, 13]. 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 3.1), proving that

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

Section 3.2 proves the convergence of $\rho(n)/n$ even in the case of a fixed alphabet, for example for binary strings. In Section 4.1, we further study the asymptotic behavior of $\rho_p(n)$, the number of runs with short periods, showing that $\ell_p = \lim_{n\mapsto\infty} \rho_p(n)/n$ exists and that, for some constant z_p,

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

Moreover, we provide a simple way to exactly count some microruns (Section 4.2). We give in Table 3 the first exact limits ℓ_p for microruns on binary strings with $p \leq 14$. Section 5 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 an interval $[i\,..\,j]$:

- such that $x[i\,..\,j]$ has period $p \leq (j-i+1)/2$,
- 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 $[4\,..\,5]\,(\underline{\mathsf{tt}})$, $[7\,..\,8]\,(\underline{\mathsf{tt}})$, $[1\,..\,4]\,(\underline{\mathsf{at}}\mathsf{at})$ and $[2\,..\,8]\,(\underline{\mathsf{tat}}\mathsf{tatt})$, 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 maximum 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 maximum number of runs in a string of length n. Kolpakov and Kucherov gave in [9] some values for $\rho(n)$ (Table 1). Table 4, 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. On the number of runs

3.1. Rewritings and asymptotic behavior of the number of runs

Franek et al. [5, 6] 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 and 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. (See below for a discussion on the size of the alphabet.) Then $x\bar{y}$ is a string of length $m+n$ containing exactly the runs of x and the rewritten runs of y. Thus $\rho(m+n) \geq r(x\bar{y}) = r(x) + r(y) = \rho(m) + \rho(n)$.

For any $t \geq 1$, we have in particular $\rho(tn) \geq t\rho(n)$.

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 can now state 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 [9]. 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 [4, 14]. This convergence result was an open question of [6]. In fact, the motivation of [6] was the remark that *"the sequence $|x_i|$ (of [5]) 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 [6] prove that every $\alpha - \varepsilon$ is an actual asymptotic lower bound by building specific sequences. With Propositions 1 and 2 and 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. Moreover, taking the following inequality to the limit, we have $\ell_{\geq p} \geq \ell/p$.

Proposition 4. *For any p and n, $\rho_{\geq p}(pn) \geq \rho(n)$.*

Proof. Take a string $x = x_1 x_2 \dots x_n$ of length n with $r(x) = \rho(n)$. Let $\bar{x}$ be the string $x_1^p x_2^p \dots x_n^p$ of length pn. Then $r_{\geq p}(\bar{x}) \geq r(x)$.

3.2. 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 et al. (Theorem 2 of [5]). A key observation is that some runs of x and y are merged in xy only when a string 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 5. *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 longest 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 consider the runs of period p that were counted in $r(x) + r(y)$. The runs with $2p$ characters ("a square") completely included in w were counted once in $r(x)$ and once in $r(y)$. Such runs are counted again once in $r(x')$ and once in $r(y')$. By definition of w, all the others runs counted in $r(x)$ and $r(y)$ are counted exactly once in $r(x')$, 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 5 gives a finite sequence of pairs of strings $(x_0, y_0), (x_1, y_1), \dots (x_t, y_t)$ with $r(x_i) + r(y_i) \geq r(x_{i-1}) + r(y_{i-1})$ and $|y_0| > |y_1| > \dots > |y_t| = 0$ for some t.

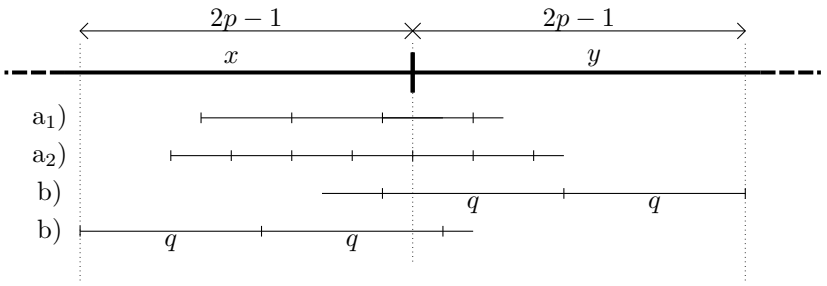

Figure 1: a$_1$) Run with at least two periods included in x. a$_2$) Run with at least two periods included in x, and at least two periods included in y. b) "New runs" between x and y. To bound the new runs with period $q \leq p$, it is sufficient to consider strings of length $4p-2$. Note that $4p$ characters would be required to exactly count the new runs.

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 et al. in [5] was in a different context, and that no result leading to our Proposition 1 was stated as such in their paper.

4. On the number of microruns

In the following sections, p is fixed and we study the asymptotic behavior of the number of microruns $\rho_p(n)/n$ beyond the result of Proposition 3. In section 4.1, the idea is to bound the *new runs* created by the concatenation of two strings. In section 4.2, the idea is to count exactly the new runs created by the concatenation of a string and a character. Both sections provide new bounds or exact limits on the number of some microruns.

4.1. New runs obtained by string concatenation

Let x and y be two strings, and s be a run of xy with period $q \leq p$. Then s is exactly in one of the following two cases (Figure 1):

- a) s has at least two periods included in x, or at least two periods included in y;
- b) s has strictly less than two periods included in x, and strictly less than two periods included in y.

We call the runs in the case b) the *new runs* between x and y, and we denote by $\mathsf{NR}_p(x,y)$ the number of such runs. Then $r_p(xy) \leq r_p(x) + r_p(y) + \mathsf{NR}_p(x,y)$, 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 6. *Let $z_p = \max\{\mathsf{NR}_p(x,y) \mid |x| = |y| = 2p-1\}$ the maximum number of new runs with period $q \leq p$ between strings of length $2p-1$. Then, for every strings x and y of any length, we have $\mathsf{NR}_p(x,y) \leq z_p$.*

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 $\leq 4p-2$.

PROOF. Any new run with period $q \leq p$ has at most $2q-1 \leq 2p-1$ characters in x, and in y (Figure 1).

Proposition 7. *For any m and n, $\rho_p(m+n) \leq \rho_p(m) + \rho_p(n) + z_p$.*

PROOF. Let x and y be two strings such that $|x| = m$, $|y| = n$, and $r_p(xy) = \rho_p(m+n)$. Then $\rho_p(m+n) = r_p(xy) \leq r_p(x) + r_p(y) + \mathsf{NR}_p(x,y) \leq \rho_p(m) + \rho_p(n) + z_p$.

Table 2 provides some values of z_p for binary strings. 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 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 7, 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, as t goes to infinity, 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 6.

As a side result of Theorem 2, we have new bounds of the number of some microruns. For example, brute-force computations give for binary strings $z_9 = 7$ and $\rho_9(34) = 26$, thus $\ell_9 \leq 33/34 = 0.970$. For binary strings, this result is better than Lemma 2 of [1] which proved the n bound by the count of amortizing positions for centers of runs. The next section further improves this bound and finds exact values for some ℓ_p's.

4.2. The exact number of microruns

In this section, we propose to count exactly the number of microruns, by considering the concatenation of a string and *a single character*. Let x be a string, $\alpha \in \Sigma$ a character, and s be a run of $x\alpha$ with period $q \leq p$. Then s is exactly in one of the following two cases (Figure 2) :

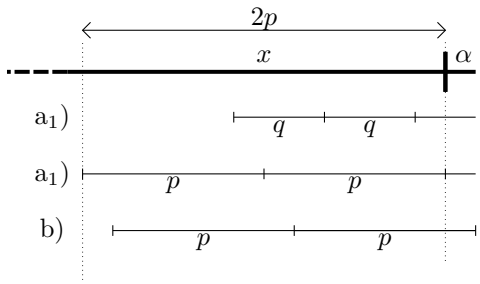

Figure 2: a$_1$) Runs of x, extended in α. b) "New runs" between x and α. Unlike in Fig. 1, there is no case a$_2$) where some runs are merged. To count the new runs with period $q \leq p$, it is sufficient to consider the suffix of x of length $2p$.

- a) s has at least two periods included in x;
- b) s has strictly less than two periods included in x.

As in the previous section, we call the runs in the case b) the *new runs* between x and α, and we denote by $\mathsf{NR}_p(x, \alpha)$ the number of such runs. As there is here no merging of runs, we have $r_p(x\alpha) = r_p(x) + \mathsf{NR}_p(x, \alpha)$. In fact, the last $2p$ characters of x are sufficient to know $\mathsf{NR}_p(x, \alpha)$:

Proposition 8. *Let x be a string with $|x| = n \geq 2p$. Then we have $\mathsf{NR}_p(x, \alpha) = \mathsf{NR}_p(x[n - 2p + 1 \dots n], \alpha)$.*

Proof. Any new run with period $q \leq p$ has at most $2q - 1 \leq 2p - 1$ characters in x. Any run of x with period $q \leq p$ extending in α has at a least 2 periods in the last $2p$ characters of x (Figure 2). Knowing the last $2p$ characters of x is thus sufficient to tell apart the two kinds of runs.

If v is a string of length $2p$ and $n \geq 2p$, we define $f_p^n(v) = \max_{|x|+|v|=n} r_p(xv)$ as the maximum number of runs of all the strings ending with the suffix v. The function f_p^{n+1} can be entirely determined from the functions f_p^n and NR_p:

Proposition 9. *If $|w| = 2p - 1$, $\alpha \in \Sigma$, and $n \geq 2p$, then*

$$f_p^{n+1}(w\alpha) = \max_{\beta \in \Sigma} \left(f_p^n(\beta w) + \mathsf{NR}_p(\beta w, \alpha) \right).$$

Proof. To compute $f_p^{n+1}(w\alpha) = \max_{|x|+|w\alpha|=n+1} r_p(xw\alpha)$, we suppose, without loss of generality, that the string x is of length at least one and we write $x = y\beta$, where y is a string and $\beta \in \Sigma$.

$$\begin{aligned}
f_p^{n+1}(w\alpha) &= \max_{|y\beta|+|w\alpha|=n+1} r_p(y\beta w\alpha) \\
&= \max_{|y\beta|+|w\alpha|=n+1} \left(r_p(y\beta w) + \mathsf{NR}_p(y\beta w, \alpha) \right) \\
&= \max_{|y\beta|+|w\alpha|=n+1} \left(r_p(y\beta w) + \mathsf{NR}_p(\beta w, \alpha) \right) && \text{(Proposition 8)} \\
&= \max_{\beta \in \Sigma} \left(\max_{|y|+|\beta w|=n} r_p(y\beta w) + \mathsf{NR}_p(\beta w, \alpha) \right) \\
&= \max_{\beta \in \Sigma} \left(f_p^n(\beta w) + \mathsf{NR}_p(\beta w, \alpha) \right)
\end{aligned}$$

Once all the $|\Sigma|^{2p+1}$ values of NR_p are computed, the above equation can be used to recursively determine any f_p^n function in $O(n \cdot |\Sigma|^{2p+1})$ time. Then any number of microruns $\rho_p(n) = \max_{|v|=2p} f_p^n(v)$ follows. When n grows, an additive periodic behavior can emerge:

Theorem 3. *If for some n_0, k and s with $2p \leq n_0 < n_0 + k$, we have $f_p^{n_0+k} = f_p^{n_0} + s$ then, for any $n \geq n_0$, we have $f_p^{n+k} = f_p^n + s$, and $\ell_p = s/k$.*

Proof. If the property $f_p^{n+k} = f_p^n + s$ is true for some n, then

$$\begin{aligned} f_p^{n+1+k}(w\alpha) &= \max_{\beta\in\Sigma} \left(f_p^{n+k}(\beta w) + \mathsf{NR}_p(\beta w, \alpha)\right) \\ &= \max_{\beta\in\Sigma} \left((f_p^n(\beta w) + s) + \mathsf{NR}_p(\beta w, \alpha)\right) \\ &= f_p^{n+1}(w\alpha) + s \end{aligned}$$

and the property is true for $n+1$. By induction, it is true for every $n \geq n_0$. In particular, for every $t \geq 0$, we have $f_p^{n_0+tk} = f_p^{n_0} + ts$, that is $\rho_p(n_0 + tk) = \rho_p(n_0) + ts$ and finally $\lim_{t\mapsto\infty} \rho_p(n_0+tk)/(n_0+tk) = s/k$. As $\rho_p(n)/n$ converges (Proposition 3), the Theorem is proved.

By computing f_p^n functions for successive n and by checking the additive periodicity condition of Theorem 3, one can have exact values of ℓ_p for small p's. Table 3 lists results for $p \leq 14$ on binary strings. Note that the periodicity on $\rho_p(n)$ can appear before the periodicity on f_p^n. For example, as soon as $n \geq 35$, $\rho_9(n + 13) = \rho_9(n) + 11$, but the periodicity on f_9^n only starts at $n_0 = 51$.

Using the result of Crochemore and Ilie's Proposition 1 [1] for large runs, we get an upper bound on $\rho(n)/n$. For binary strings, the exact value $\ell_{14} = 15/17$ gives:

$$\ell \leq \ell_{14} + \ell_{\geq 15} \leq \frac{15}{17} + \frac{6}{15} = 1.282...$$

Thus *the number of runs in a binary string of length n is not more than $1.29n$.* This result was better than the $1.6n$ bound published in [1], but the better bound of $1.048n$ has now been published [2]. Nevertheless, the values we give in the Table 3 are the first known exacts limits for such microruns.

5. 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. Moreover, the application of the Theorem 3 provides the first exact limits for the number of some microruns. We hope that these results will bring a better understanding of the number of runs and be a step towards proving the conjecture of [8] ($\ell \leq 1$).

As Theorem 2 and 3 provide upper bounds or exact limits for some microruns, they can be used to bound the total number of runs. In both cases, this would require large evaluations of z_p or $f_p^n(w)$ values that could be improved

$$
\begin{array}{rll}
\ell_1 = \ell_2 = & \dfrac{1}{2} = 0.5 & f_1^4 = f_1^2 + 1 \\
 & & f_2^7 = f_2^5 + 1 \\
\ell_3 = \ell_4 = & \dfrac{3}{4} = 0.75 & f_3^{16} = f_3^{12} + 3 \\
 & & f_4^{45} = f_4^{41} + 3 \\
\ell_5 = \ell_6 = & \dfrac{7}{9} = 0.777... & f_5^{36} = f_5^{27} + 7 \\
 & & f_6^{47} = f_6^{38} + 7 \\
\ell_7 = & \dfrac{4}{5} = 0.8 & f_7^{62} = f_7^{57} + 4 \\
\ell_8 = \ell_9 = \ell_{10} = \ell_{11} = \ell_{12} = & \dfrac{11}{13} = 0.846... & f_8^{62} = f_8^{49} + 11 \\
 & & f_9^{64} = f_9^{51} + 11 \\
 & & f_{10}^{69} = f_{10}^{56} + 11 \\
 & & f_{11}^{120} = f_{11}^{107} + 11 \\
 & & f_{12}^{145} = f_{12}^{132} + 11 \\
\ell_{13} = \ell_{14} = & \dfrac{15}{17} = 0.882... & f_{13}^{113} = f_{13}^{96} + 15 \\
 & & f_{14}^{104} = f_{14}^{87} + 15
\end{array}
$$

Table 3: Exact limits of $\rho_p(n)/n$ for binary strings, obtained by successive applications of the equation in Proposition 9 until the additive periodicity condition of Theorem 3 is true. Each assertion on the right gives the smallest $n \geq 2p$ such that $f_p^{n+k} = f_p^n + s$ for some s and $k > 0$. The value ℓ_{13} required three hours of computation on a standard 2 GHz workstation. This time is almost entirely spent in the initial computation of the 2^{2p+1} values of the function NR_p, obtained by aggregate calls to `mreps` [7]. The successive computations of f_p^n are done in a few seconds.

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

Table 4: Values of $\rho_p(n)$ for binary strings. For each n, the value in bold shows the smallest period p such that $\rho_p(n) = \rho(n)$.

by a more precise analysis, for example by taking inspiration from the methods of Crochemore and Ilie. Moreover, a better analysis could improve their $6/p$ bound on the number of macroruns $\ell_{\geq p}$.

For the lower bound, it remains to be shown if one can find strings with more runs than those of [5, 6]. 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 0.9445... bound of [10].

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. 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.

References

[1] Maxime Crochemore and Lucian Ilie. Maximal repetitions in strings. *J. Comput. Systems Sci.*, 74(5):796–807, 2008.

[2] Maxime Crochemore, Lucian Ilie, and Liviu Tinta. Towards a solution to the runs conjecture. In *Combinatorial Pattern Matching (CPM 2008)*, volume 5029 of *LNCS*, pages 290–302, 2008.

[3] Mathieu Giraud. Not so many runs in strings. In *Language, Automata Theory and Applications (LATA 2008)*, volume 5196 of *LNCS*, pages 290–302, 2008.

[4] M. Fekete. Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Mathematische Zeitschrift*, 17:228–249, 1923.

[5] Frantisek Franek, R. J. Simpson, and W. F. Smyth. The maximum number of runs in a string. In *Australasian Workshop on Combinatorial Algorithms (AWOCA 03)*, pages 26–35, 2003.

[6] 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.

[7] Roman Kolpakov, Ghizlane Bana, and Gregory Kucherov. mreps: Efficient and flexible detection of tandem repeats in dna. *Nucleic Acids Res*, 31(13):3672–3678, 2003.

[8] 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.

[9] Roman Kolpakov and Gregory Kucherov. On maximal repetitions in words. *Journal on Discrete Algorithms*, 1(1):159–186, 2000.

[10] Wataru Matsubara, Kazuhiko Kusano, Akira Ishino, Hideo Bannai, and Ayumi Shinohara. New lower bounds for the maximum number of runs in a string. In Jan Holub and Jan Žďárek, editors, *Prague Stringology Conference 2008 (PSC 08)*, pages 140–145, 2008.

[11] Simon J. Puglisi, Jamie Simpson, and Bill Smyth. How many runs can a string contain? *Theoretical Computer Science*, 401(1-3):165–171, 2008.

[12] Wojciech Rytter. The number of runs in a string : improved analysis of the linear upper bound. *Information and Computation*, 205(9):1459–1469, 2007.

[13] Bill Smyth. The maximum number of runs in a string. In *International Workshop on Combinatorial Algorithms (IWOCA 2007), Problems Session*, 2007.

[14] J. L. van Lint and R. M. Wilson. *A course in combinatorics*. Cambridge University Press, 1992.