Maximal repetitions in strings

Maxime Crochemore, Lucian Ilie

Maximal repetitions in strings ⋆

Maxime Crochemore [a,1], Lucian Ilie [b,*,2]

[a] *Department of Computer Science, King's College London, London WC2R 2LS, UK and Institut Gaspard-Monge, Université Paris-Est, F-77454 Marne-la-Vallée Cedex 2, FRANCE*

[b] *Department of Computer Science, University of Western Ontario N6A 5B7, London, Ontario, CANADA*

Abstract

The cornerstone of any algorithm computing all repetitions in strings of length n in $\mathcal{O}(n)$ time is the fact that the number of maximal repetitions (runs) is linear. Therefore, the most important part of the analysis of the running time of such algorithms is counting the number of runs. Kolpakov and Kucherov [FOCS'99] proved it to be cn but could not provide any value for c. Recently, Rytter [STACS'06] proved that $c \leq 5$. His analysis has been improved by Puglisi et al. to obtain 3.48 and by Rytter to 3.44 (both submitted). The conjecture of Kolpakov and Kucherov, supported by computations, is that $c = 1$. Here we improve dramatically the previous results by proving that $c \leq 1.6$ and show how it could be improved by computer verification down to 1.18 or less. While the conjecture may be very difficult to prove, we believe that our work provides a good approximation for all practical purposes.

For the stronger result concerning the linearity of the sum of exponents, we give the first explicit bound: $5.6n$. Kolpakov and Kucherov did not have any and Rytter considered "unsatisfactory" the bound that could be deduced from his proof. Our bound could be as well improved by computer verification down to $2.9n$ or less.

Key words: combinatorics on words, repetitions in strings, runs, maximal repetitions, maximal periodicities, sum of exponents
MSC: 68R15, 68W40

⋆ This work has been done during the second author's stay at Institut Gaspard-Monge.

* Corresponding author
Email addresses: maxime.crochemore@kcl.ac.uk (Maxime Crochemore), ilie@csd.uwo.ca (Lucian Ilie).

[1] Research supported in part by CNRS.
[2] Research supported in part by CNRS and NSERC.

1. Introduction

Repetitions in strings constitute one of the most fundamental areas of string combinatorics with very important applications to text algorithms, data compression, or analysis of biological sequences. They have been study already in the papers of Axel Thue [20], considered as having founded stringology. While Thue was interested in finding long sequences with few repetitions, one of the most important problems from the algorithmic point of view was finding all repetitions fast. A major obstacle for a linear-time algorithm was finding a way to encode all repetitions in linear space. The problem was studied first by Crochemore [2] where maximal (non-extendable) integer powers were introduced and an (optimal) $\mathcal{O}(n \log n)$ algorithm for finding them all was given. Moreover, the bound was shown to be optimal as it is reached by the Fibonacci strings.

The next step was to consider occurrences of fractional repetitions, of right-maximal (non-extendable to the right) repetitions by Apostolico and Preparata in [1] and then of maximal (non-extendable both ways, called *runs* for the rest of the paper) repetitions by Main [15] who gave a linear-time algorithm for finding all leftmost occurrences of runs.

Iliopoulos et al. [9] showed that for Fibonacci strings the number of maximal repetitions is linear. Even if their result applies to a particular class of strings, it is important since the Fibonacci strings were known to contain many repetitions. The breakthrough came in the paper of Kolpakov and Kucherov [12], where it was finally proved that encoding all occurrences of repetitions into runs was the right way to obtain a linear-sized output. They modified Main's algorithm to compute all runs in linear time; see [11]. For more details on various algorithms computing repetitions, see Chapter 8 of [14].

Kolpakov and Kucherov [12] showed that the number of runs in a string of length n is at most cn but their proof could not provide any value for the constant c. Another breakthrough came very recently, when Rytter [17] proved that $c \leq 5$. Puglisi et al. [16] improved Rytter's analysis to bring the constant down to 3.48 and then Rytter [18] produced his own version of the improved analysis with a constant factor of 3.44. The fact that the two bounds are so close may show that the ideas in Rytter's initial paper have been well squeezed.

Based on the results in Table 1, Kolpakov and Kucherov [12] conjectured that $c = 1$ for binary alphabets. A stronger conjecture is proposed in [6] where a family of strings is given with the number of runs equal to $\frac{3}{2\phi} = 0.927\ldots$ (ϕ is the golden ratio), thus proving $c \geq 0.927\ldots$. The authors of [6] conjectured that this bound is optimal. Some reasons which might indicate that the optimal bound may be less than n are discussed in Section 7.

Table 1
Maximum number of runs in binary strings of length $n, 5 \leq n \leq 31$ (from [12]).

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
max. no. of runs	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

The proof of [12] is extremely long and complex and the one of [17] is still very intricate. (The two improvements in [16] and [18] only make a more careful analysis of the ideas of [17].) A simple proof for the linearity is given by the authors in [3] where an improvement of the notion of neighbors of [17] is introduced.

It is interesting to notice that the number of runs having the same starting point is logarithmic, due to the three-square lemma of [4], that is, they are not uniformly distributed, which makes proving linearity hard. The situation is even worse for centers (beginning of the second period of the run – see next section for details) where linearly many runs can share the same center. However, while Rytter [17] counted runs by their beginnings, we count them by centers and obtain much better results. A more detailed comparison of the two approaches is included in Section 3.

In this paper we improve significantly the previous results by proving the bound $1.6n$ on the number of runs in a string of length n. This bound can be lowered a bit by extra effort for the runs with short periods, but we show also how it could be improved by computer verification down to $1.18n$ or even further. Notice that the bound on the number of runs has an important direct impact on the running time of all algorithms computing all repetitions since it says how many runs we expect the algorithm to output. It is important as well from mathematical point of view, that is, to find the best upper bound on the number of runs, and from algorithm-design point of view, as it may lead to simpler algorithms for finding all repetitions (the algorithm of [11] uses relatively complicated data structures such as suffix trees). While the conjecture may be very difficult to solve, we believe that our work provides a good approximation for all practical purposes.

The approach in [3] is used also to give a simple proof for the stronger result concerning the linearity of the sum of exponents. This result has been proved by Kolpakov and Kucherov [10]. (It follows also from Rytter's construction in [17].) It has applications to the analysis of various algorithms, such as computing branching tandem repeats: the linearity of the sum of exponents solves a conjecture of [19] concerning the linearity of the number of maximal tandem repeats and implies that all can be found in linear time. For other applications we refer to [10].

But the proof of [10] is very complex and could not provide a constant. A bound can be derived from the proof of Rytter [17] but he mentioned only that the bound that he obtains is "unsatisfactory." It seems to be $25n$. The improved analysis in [18] does not mention the sum of exponents at all. We provide here the first explicit bound, which is $5.6n$. As with the other bound, extra effort for the runs with short periods can lower the bound, but we show how it could be improved by computer verification down to $2.9n$ or further. As mentioned in [10], computations seem to indicate a $2n$ bound.

The paper is organized as follows. In the next section we give the basic definitions needed in the paper. The new bound is given in the following section which contains also a comparison between our approach and the one of Rytter [17,18]. Our approach is more accurate for both long and short periods. The division into two subsets according to period length is natural because, as explained in Section 3, no approach seems to work well for both. For long ones, Rytter [18] proves the bound $0.67n$ for runs with periods 87 or higher. For comparison sake, we could easily deduce the corresponding bound using our approach: it is $0.06897n$, that is, roughly ten times better. The analyses needed for the new bound are presented in Section 4, for runs with arbitrarily long periods, and Section 5, for runs with periods 9 or less. The sum of exponents in discussed in Section 6. Some comments on both bounds, as well as ways to improve them further by computer verification are included in Section 7. We conclude with a discussion on several other related problems.

2. Definitions

Let A be an alphabet and A^* the set of all finite strings over A. We denote by $|w|$ the length of a string w, its ith letter by $w[i]$ and the factor $w[i]w[i+1]\dots w[j]$ by $w[i\,..\,j]$. We say that w has period p iff $w[i] = w[i+p]$, for all $i, 1 \le i \le |w| - p$. The shortest period of w is called *the period* of w. The ratio between the length and the period of w is called the *exponent* of w.

For a positive integer n, the nth power of w is defined inductively by $w^1 = w$, $w^n = w^{n-1}w$. A string is *primitive* if it cannot be written as a proper (two or more) integer power of another string. Any string can be uniquely written as an integer power of a primitive string, called its *primitive root*. The following well-known *synchronization* property will be useful for us: if w is primitive, then w appears as a factor of ww only as a prefix and as a suffix (not in-between). Another property we use is *Fine and Wilf's periodicity lemma*: If w has periods p and q and $|w| \ge p + q$, then w has also period $\gcd(p, q)$. (This is a bit weaker than the original lemma which works as soon as $|w| \ge p + q - \gcd(p, q)$, but it is good enough for our purpose.) We refer the reader to [13,14] for further information on all concepts used here.

For a string w, a *run*[3] (or maximal repetition) is an interval $[i\,..\,j]$ such that both (i) the factor $w[i\,..\,j]$ has its shortest period at most $\frac{j-i+1}{2}$ and (ii) $w[i-1\,..\,j]$ and $w[i\,..\,j+1]$, if defined, have a strictly higher shortest period. As an example, consider $w = \mathsf{abbababbaba}$; $[3\,..\,7]$ is a run with period 2 and exponent 2.5; we have $w[3\,..\,7] = \mathsf{babab} = (\mathsf{ba})^{2.5}$. Other runs are $[2\,..\,3], [7\,..\,8], [8\,..\,11], [5\,..\,10]$ and $[1\,..\,11]$.

By definition, a run is a maximal occurrence of a repetition of exponent at least two. Therefore, it starts with a square and continues with the same period. But the square is the only part of the run we can count on. Therefore, for a run starting at i and having period $|x| = p$, we shall call $w[i\,..\,i+2p-1] = x^2$ the *square* of the run. Notice that x is primitive and the square of a run is not left-extendable (with the same period) but may be extendable to the right. The *center* of the run is the position $c = i + p$. We shall denote the *beginning* of the run by $i_x = i$, the *end* of its square by $j_x = i_x + 2p - 1$, and its center by $c_x = c$.

3. The bound

The idea is to partition the runs by grouping together those having close centers and similar periods and then prove that we have only one on the average in each group. For any $\delta > 0$, we say that two runs having squares x^2 and y^2 are *δ-close* if both (i) $|c_x - c_y| \le \delta$ and (ii) $2\delta \le |x|, |y| \le 3\delta$. Abusing the language, we shall sometimes say that the squares, instead of the runs, are δ-close. Another notion that we shall use frequently is that of runs with the periods between 2δ and 3δ; we shall call those *δ-runs*.

We prove (Section 4) that the number of runs is highest when any interval of length δ contains only one center of a δ-run. That means that the number of δ-runs in a string of length n is at most $\frac{n}{\delta}$. We could then sum up for values of δ which would cover all possible periods but we make one further improvement. Since any bound for arbitrarily

[3] Runs were introduced in [15] under the name *maximal periodicities*; they are called *m-repetitions* in [12] and *runs* in [9].

long runs performs poorly on runs with short periods, we bound separately the number of runs with short periods; precisely we prove that there are at most n runs with period at most 9 in any string of length n (Section 5).

Summing up the above for all values $\delta_i = \frac{10}{2}\left(\frac{3}{2}\right)^i$, $i \geq 0$, to cover all periods greater than 9, we obtain the following upper bound for the number of runs in a string of length n:

$$n + \sum_{i=0}^{\infty} \frac{n}{\delta_i} = n + \left(\frac{2}{10} \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^i \right) n = 1.6n \ . \tag{1}$$

Our main result is

Theorem 1 *The number of runs in a string of length n is less than $1.6n$.*

Our approach differs from the one of Rytter [17] in several respects. First, our notion of δ-closeness is different from his notion of *neighbors*. We consider the case when the centers of the runs are close to each other as opposed to beginnings as this gives us a better overlap between the runs. Thus we can choose a better interval length for the periods. Second, we make a combinatorially finer analysis of the close runs which enables us to count all runs together; [17] splits them into weekly and highly periodic. Doing so, the proof becomes conceptually simpler. For runs with long periods we can say that our approach is about ten times better than Rytter's. He explicitly states that the number of runs with periods larger than 87 is at most $0.67n$. With our approach, this number is about ten times smaller:

$$\left(\frac{2}{87} \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^i \right) n \leq 0.06897n \ .$$

Third, our approach for runs with short periods is different from the one of [17]. We essentially verify that the conjecture is true up to a certain threshold for the periods of the runs. Due to the complexity of the analysis, we restricted this threshold to 9 but it can be checked automatically for higher thresholds, every time improving the bound. More on this is in Section 7.

4. Runs with close centers

In this section we show that, for a given δ, each interval of positions of length δ contains at most 1 center of a δ-run on the average. The result is used for runs having a period greater than 9 in the sum (1).

We investigate what happens when two (or, sometimes, three) runs in a string w are δ-close. Denote their squares by x^2, y^2, z^2, their root lengths by $|x| = p$, $|y| = q$, $|z| = r$, and assume $p \leq q \leq r$.

We discuss below all ways in which occurrences of x^2 and y^2 can be positioned relative to each other and see that long factors of both runs have short periods. When we have only two δ-close runs, synchronization properties show that the next (to the right) interval of length δ (as counted in (1)) does not contain any center of a δ-run.

When we have three δ-close runs, z^2 has to synchronize the short periods mentioned above, which restricts the beginning of z^2 to only one choice as otherwise some run would be left extendable (which is not possible). Stronger periodicity properties are implied by the existence of the third run and we can find an interval of length at least 4δ which

contains no other center of δ-runs. Such an interval covers at least three intervals of length δ no matter how the decomposition of $[1\,..\,n]$ into such intervals is done. Thus, less runs than in (1) are obtained.

It is also possible to have arbitrarily many δ-close runs, that is, when they all have the same center; case (i). A similar global counting approach is performed in this case. The existence of such runs implies strong periodicity properties of a factor of w and we exhibit a long interval without any center of runs with certain periods. In total, less runs than in (1) are obtained.

There can exist several δ-close runs such that some of them only share the same center. Therefore, we shall discuss the case of many runs having the same center first. It helps solving some situations in the other cases.

(i) $c_x = c_y$. First, both x and y have the same small period $\ell = q - p$; see Fig. 1. If we denote $c = c_y$ then we have h runs $x_j^{\alpha_j}$, $2 \le \alpha_j \in \mathbb{Q}$, for $1 \le j \le h$, having period $|x_j| = (j-1)\ell + \ell'$ and beginning at $i_{x_j} = c - ((j-1)\ell + \ell')$. If we set $x_j = u^{j-1}u'$, with $|u| = \ell, |u'| = \ell'$, then the last letters of u and u' are different, as otherwise x would be left-extendable. As an example, take $w = (\mathsf{ab})^6\mathsf{aa}(\mathsf{ba})^6$, where $c = 14, h = 7, \ell = 2$, $\ell' = 1$, $u = \mathsf{ab}$, and $u' = \mathsf{a}$.

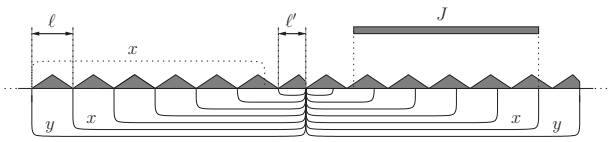

Fig. 1. The runs with the same center in case (i).

We show that for $h \ge 6$ we have less runs than in (1). Notice that only for $h \ge 7$ we can have three of the $x_j^{\alpha_j}$s mutually δ-close. Therefore, we may assume for the other cases ((ii)–(v)) that there are no three δ-close runs with the same center.

There exists δ_{i_0} such that $\frac{\ell}{2} \le \delta_{i_0} \le \frac{3\ell}{4}$, that is, this δ_{i_0} is considered in (1). The periods corresponding to δ_{i_0} are between ℓ and $\frac{9}{4}\ell$.

We claim that there is no run in w with period between ℓ and $\frac{9}{4}\ell$ and center inside the interval $J = [c + \ell + 1\,..\,c + (h-2)\ell + \ell']$. Indeed, assume there is a run with the initial square t^2, $c_t \in J$. If $i_t \ge c$, then the prefixes of length ℓ of the first and second occurrences of t, respectively, must synchronize. If $i_t > c$, then t^2 is left-extendable, a contradiction. If $i_t = c$, then ℓ divides $|t|$ and hence t is not primitive, a contradiction. If $i_t < c$, then synchronization is obtained (either the prefixes or the suffixes of length ℓ of the two occurrences of t synchronize) and we get that the last letters of u and u' are the same, a contradiction.

Then, the length of J is larger than $(h-3)\ell$ which in turn is larger than $(h-2)\delta_{i_0}$ (since $3\ell \ge 4\delta_{i_0}$ and $h \ge 6$). Thus J covers at least $h-3$ intervals of length δ_{i_0} that would contain, if considered in (1), $h-3$ runs. This is enough as we need to account, for each δ from (1), for the extra runs, that is, all but one. At least three δs are needed for all our h runs, so we need to account for at most $h-3$ runs, which we already did.

We need also mention that these h intervals of length δ_{i_0} are not reused by a different center with multiple runs since such centers cannot be close to each other. Indeed, assume we have two centers c_j with the above parameters h_j, ℓ_j, $j = 1, 2$. Then the periods satisfy $\frac{\ell_j}{2} \le \delta_{i_0} \le \frac{3\ell_j}{4}$, $j = 1, 2$, and so $\ell_j \le \frac{3}{2}\ell_{3-j}$, $j = 1, 2$. As soon as the longest runs with

centers at c_1 and c_2, respectively, overlap over $\ell_1 + \ell_2$ positions, we have $\ell_1 = \ell_2$, due to Fine and Wilf's lemma. Thus, the closest positions of J_1 and J_2 cannot be closer than $\ell_1 = \ell_2 \geq \delta_{i_0}$ as otherwise some of the runs become non-primitive, a contradiction.

(ii) $(i_y < i_x) < c_y \leq c_x < e_x \leq e_y$. Then x and the suffix of length $c_y - i_x + (q-p)$ of y have period $q-p$; see Fig. 2. (Only the periods to the left of e_y are implied by x^2 and y^2, the ones to the right of e_y are obtained in the case when a third run, z^2, exists – see below.) We may assume this period is a primitive string as otherwise we can make the same reasoning with its primitive root.

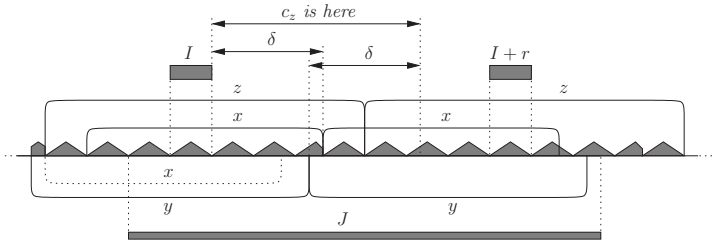

Fig. 2. Relative position of x^2, y^2 and z^2 in case (ii).

It is not difficult to see that no δ-run can have its center in the interval $[c_y+\delta \,..\, c_y+2\delta]$ as it would be left extendable. This argument is sufficient for the case when no third δ-close run exists.

If there is a third run, z^2, then we need a stronger argument to account for the three centers in the same interval of length δ. Since z^2 is δ-close to both x^2 and y^2, it must be that $c_z \in [c_x - \delta \,..\, c_y + \delta]$. Consider the interval of length $q-p$ that ends at the leftmost possible position for c_z, that is, $I = [c_x - \delta - (q-p) \,..\, c_x - \delta - 1]$; see Fig. 2. The following arguments show that it is included in the first period of z^2, that is, $[i_z \,..\, c_z - 1]$, and in $[i_x \,..\, c_y - 1]$. It is clear that I ends before both $c_z - 1$ and $c_y - 1$. The other two inequalities are proved as follows. First $i_z = c_z - r \leq c_y + \delta - q \leq c_x + p - \delta - q$. Then $i_x = c_x - p \leq c_x - \delta - q + p$. Subsequently, all such inequalities can be proved similarly and are left as exercises.

Thus $w[I]$ is primitive and equal, due to z^2, to $w[I+r]$ which is contained in $[c_x \,..\, e_y]$. Therefore, the periods inside the former must synchronize with the ones in the latter. It follows, in the case $i_z > i_x - (q-p)$, that $w[i_z - 1] = w[c_z - 1]$, that is, z^2 is left extendable, a contradiction. If $i_z < i_x - (q-p)$, then $w[c_x - 1] = w[i_x - (q-p) - 1] = w[i_x - 1]$, that is, x^2 is left extendable, a contradiction. The only possibility is that $i_z = i_x - (q-p)$ and r equals q plus a multiple of $q-p$. Here is an example that this is indeed possible: $w = \mathsf{baababababaababababaab}$, $x^2 = w[5 \,..\, 14]$, $y^2 = w[1 \,..\, 14]$, and $z^2 = w[3 \,..\, 20]$.

The existence of z^2 implies that the period of the second occurrence of y extends past e_y, as seen in Fig. 2. Consider the interval $J = [i_z + 2(q-p) \,..\, e_z - 2(q-p)]$. The existence of a fourth δ-run with center inside J but different from c_x would imply that either x^2 or z^2 is left extendable, a contradiction. (Notice that we allowed the length of two $(q-p)$-periods at either end of J so that the hiccup of the period ending at c_x does not cause any problems.) On the other hand, such a run can have c_x as its center. If so, then all such runs are accounted for by case (i) since we have at least three periods at c_x between 2δ and 3δ: q, $q-(q-p)$, and $q-2(q-p)$. The length of J is

$2r - 4(q-p) \geq 2(q+q-p) - 4(q-p) = 2p \geq 4\delta$ and therefore it covers at least three intervals of length δ. In total, we have at most the number of runs as counted in (1).

(iii) $(i_y < i_x) < c_y < c_x < e_y \leq e_x$. This case is similar with (ii); see the top part of Fig. 3.

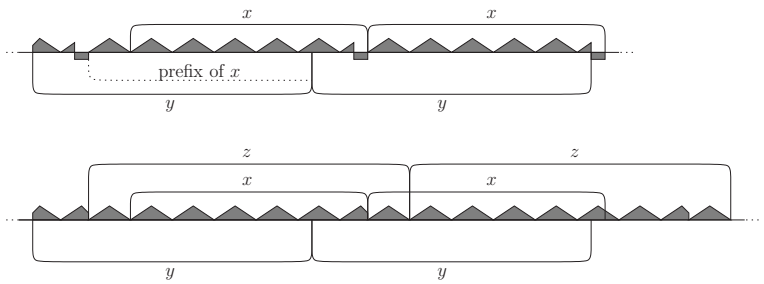

Fig. 3. Relative position of x^2, y^2 and z^2 in case (iii).

Again, no δ-run can have its center in the interval $[c_y + \delta \,..\, c_y + 2\delta]$. For the case when a third run exists, denote $\varepsilon = c_x - c_y - (q-p)$. The synchronizing interval is ε positions to the left compared to the one at (ii), that is, $I = [c_x - \delta - (q-p) - \varepsilon \,..\, c_x - \delta - \varepsilon - 1]$. A third δ-close run would have to start again at $i_z = i_x - (q-p)$; see the bottom part of Fig. 3. Notice that z^2 smoothes out the non-periodic factors of length equal to ε (the small rectangles below the line in the bottom part of Fig. 3).

The interval with no other center of δ-runs is again $J = [i_z + 2(q-p) \,..\, e_z - 2(q-p)]$.

(iv) $i_y \leq i_x < c_x < c_y (< e_x < e_y)$. Here x and the prefix of length $c_x - i_y$ of y have period $q-p$. As in case (ii) (the synchronizing interval I is the same) a third δ-close run z^2 would have to have the same beginning as y^2, otherwise one of y^2 or z^2 would be left extendable. A fourth δ-close run would have to start at the same place and we can take here the same interval J; see Fig. 4. (The extra drawings show that we have the two $(q-p)$-periods we need at the end of z.) One thing is a bit different. There can be δ-runs with center c_x. They are accounted for as before, using case (i). Notice, however, that the period q does not extend completely to the left of c_x. Still, the missing part is too small to affect the reasoning.

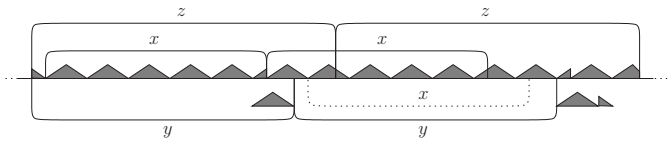

Fig. 4. Relative position of x^2, y^2 and z^2 in case (iv).

(v) $i_x < i_y (< c_x < c_y < e_x < e_y)$. We have here a synchronizing interval as in (ii); see Fig. 5. A third δ close run z^2 would have to start, as in (iv), at the same place as y^2 and have the period q plus a multiple of $q-p$. It would imply that $w[i_x \,..\, i_y - 1] = w[c_y - (i_y - i_x) \,..\, c_y - 1]$ which would make y^2 left-extendable, a contradiction; see Fig. 5. Therefore, there cannot be a third run in this case.

We proved

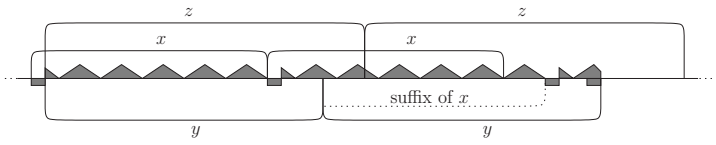

Fig. 5. Relative position of x^2, y^2 and z^2 in case (v).

Proposition 1 *There is at most 1 center of a δ-run on average in each interval of length δ.*

5. Microruns

We prove in this section that the number of runs with periods at most 9, which we call *microruns*,[4] in a string of length n is at most n. All runs we are talking about in this proof have periods at most 9.

The idea of the proof is as follows. We pick an arbitrary position i and consider all possible microruns with center at i. Then we show that the number of microruns with centers in an interval $[i-j\,..\,i]$ is at most j where j can vary but is always less than 5. Put otherwise, any position with two or more centers of microruns is amortized within up to 4 previous positions to its left.

The number of possible subsets of periods at i is very high – $2^9 = 512$ – but we have the following lemma to help. It not only reduces significantly the number of cases but it helps with the analysis of each case as well.

Assume we have a string w. We shall say that *w has p at i* if there is a run with period p and center at position i in w. Denote also $C(i) = \{p \mid p \text{ at } i\}$. (This set depends also on w but we shall consider w implicit and omit it.)

Note that there are two differences between (a run of period) p (and center) at i and a *period p centered at i*: the former is not left-extendable and its root is primitive whereas the latter may have none of the two properties.

Lemma 1 *Consider a string w and the periods p and $p-\ell$, $0 < \ell < p$. Let h be the smallest integer such that $h\ell \geq p$ ($h = \lceil p/\ell \rceil$).*

(i) (periods) If w has the period $p-\ell$ at i and the period p at $i+j$ or $i-j$ with $j \leq \ell$, then w has the also periods $p-k\ell$, $2 \leq k \leq h-1$, at i.

(ii) (runs) If w has $p-\ell$ at i and either (a) p at $i+j$ with $j \leq \ell-1$, or (b) p at $i-j$ with $j \leq \ell$, then w has $p-k\ell$ at i, for $2 \leq k \leq h-3$ (that is, all but the shortest two).

Proof. (i) Assume p at $i-j$; the other case is completely symmetric. Assume also $\ell < p/2$ since otherwise there is nothing to prove. Then $w[i\,..\,i+p-\ell-1] = w[i-p\,..\,i-\ell-1]$ and the overlap between $w[i-p\,..\,i-\ell-1]$ and $w[i-(p-\ell)\,..\,i-1]$ gives that $w[i-(p-\ell)\,..\,i-1] = w[i\,..\,i+p-\ell-1]$ has period ℓ. All periods at i claimed in the statement follow immediately.

(ii) Assume again p at $i-j$ but now j is strictly less than ℓ. We have the periods as claimed by (i). First, if ℓ divides p, then $w[i-(p-\ell)\,..\,i-1]$ is not primitive, a contradiction. Therefore, ℓ does not divide p. Then, any $w[i-(p-k\ell)\,..\,i-1]$, for $2 \leq k \leq h-3$, must be primitive since otherwise, Fine and Wilf's lemma would imply that

[4] By analogy with the *microsatellites* in bioinformatics; these correspond to the concatenation of short DNA sequences (1 to 4 nucleotides) that are similar.

$w[i-(p-\ell)\,..\,i-1]$ is not primitive, a contradiction. (The two shortest periods are not long enough to apply Fine and Wilf's lemma and indeed, they need not be primitive.)

For the non-left-extendability, we have $w[i-(p-k\ell)-1] = w[i-(p-\ell)-1] \neq w[i-1]$. Here we see the difference between (a) and (b): at (a) we need $i-(p-\ell)-1 \geq i+j-p$, that is, $j \leq \ell - 1$. □

Another useful remark, with an obvious proof, is next.

Remark 1 *If we have p at i, then we cannot have p at j for any $j, i-p \leq j \leq i+p$, $j \neq i$.*

It is also obvious how Remark 1 is used. As far as Lemma 1 is concerned, it can be used in many ways. The first is, as already announced, to reduce as much as possible sets of periods of microruns with the same center. For instance, if we do not have periods 1,2,3 at i but do have 5, then we cannot have anything else: having 4 would imply having 1,2,3; 6 implies 1,2,3,4; 7 implies 1,3; 8 implies 2; 9 implies 1. This way our potential 512 cases are reduced to 26.

The lemma helps also with the analysis of each case, as seen in the proof of Lemma 2 below.

Lemma 2 *The number of runs with periods at most 9 in a string of length n is bounded by n.*

Proof. We shall discuss in detail the analysis of one case and then give a list of all possible cases and the corresponding amortizing positions.

Consider, for example, the case when $C(i) = \{1, 3\}$. We shall use many times Lemma1 without mentioning. What we know so far about w is that $w[i-4\,..\,i+2] = \overline{\mathsf{a}}\mathsf{abaaba}$, where $\mathsf{a} \neq \mathsf{b}$ and $\overline{\mathsf{a}}$ means any letter different from a. The smallest element of $C(i-1)$ is 5. If we have 5 at $i-1$, then $w[i-7\,..\,i+3] = \overline{\mathsf{b}}\mathsf{aababaabab}$. Thus, $C(i-1) = \{5\}$ and $C(i-2) = \emptyset$, which means that the two centers at i are amortized, in this case, within the previous two positions, since the total number of centers inside the interval $[i-2\,..\,i]$ is 3. If there is not 5 at $i-1$, then the next that can be is 7 and the reasoning is identical. If there is not 7 at $i-1$, the next can be 8. If so, then $C(i-1) = \{8\}$ and the only possible (but not necessary) candidate at $i-2$ is 2. If there is 2 at $i-2$ then $C(i-2) = \{2\}$, $C(i-3) = \emptyset$, and in this case the two centers at i are amortized within the previous three positions.

The reasoning continues like this until all possibilities are completely analyzed. Actually the case $C(i) = \{1, 3\}$ has the longest analysis and there are very few comparable ones. This is the reason why we proved the result for periods up to 9. For higher numbers it gets quickly too complicated.

We give in Table 2 a list of tuples which consider all possible sets of periods of microruns with centers at an arbitrary position i and the corresponding possible sets of periods of microruns at the positions to the left of i, as many as needed to amortize them. Other sets are impossible due to Lemma 1. Thus, if the tuple contains, say, j elements, that means the tuple represents $(C(i-j+1), C(i-j+2), \ldots, C(i))$. Here is the list, which the reader can verify himself using a similar reasoning as the one above; the list gives the pairs in the order they result from the proof, that is, increasing lexicographical order where the components corresponding to higher positions are more significant. □

Table 2
The amortizing positions for the proof of Lemma 2.

$i-3$	$i-2$	$i-1$	i
	$\emptyset$	$\{5\}$	$\{1,3\}$
	$\emptyset$	$\{7\}$	$\{1,3\}$
$\emptyset$	$\{2\}$	$\{8\}$	$\{1,3\}$
	$\emptyset$	$\{8\}$	$\{1,3\}$
$\emptyset$	$\{2\}$	$\{9\}$	$\{1,3\}$
	$\emptyset$	$\{9\}$	$\{1,3\}$
		$\emptyset$	$\{1,3\}$
$\emptyset$	$\{1\}$	$\{7\}$	$\{1,4\}$
	$\emptyset$	$\{7\}$	$\{1,4\}$
$\emptyset$	$\{1\}$	$\{9\}$	$\{1,4\}$
	$\emptyset$	$\{9\}$	$\{1,4\}$
		$\emptyset$	$\{1,4\}$
$\emptyset$	$\{1\}$	$\{9\}$	$\{1,5\}$
	$\emptyset$	$\{9\}$	$\{1,5\}$
		$\emptyset$	$\{1,5\}$
		$\emptyset$	$\{1,6\}$
	$\emptyset$	$\{3\}$	$\{1,7\}$
		$\emptyset$	$\{1,7\}$
	$\emptyset$	$\{3\}$	$\{1,8\}$
		$\emptyset$	$\{1,8\}$
	$\emptyset$	$\{3\}$	$\{1,9\}$
$\emptyset$	$\{1\}$	$\{4\}$	$\{1,9\}$
	$\emptyset$	$\{4\}$	$\{1,9\}$
		$\emptyset$	$\{1,9\}$

$i-3$	$i-2$	$i-1$	i
$\emptyset$	$\{1\}$	$\{8\}$	$\{2,5\}$
	$\emptyset$	$\{8\}$	$\{2,5\}$
		$\emptyset$	$\{2,5\}$
		$\emptyset$	$\{2,6\}$
		$\emptyset$	$\{2,7\}$
		$\emptyset$	$\{2,8\}$
		$\emptyset$	$\{2,9\}$
	$\emptyset$	$\{1\}$	$\{3,7\}$
		$\emptyset$	$\{3,7\}$
	$\emptyset$	$\{1\}$	$\{3,8\}$
		$\emptyset$	$\{3,8\}$
	$\emptyset$	$\{1\}$	$\{3,9\}$
		$\emptyset$	$\{3,9\}$
	$\emptyset$	$\{1\}$	$\{4,9\}$
		$\emptyset$	$\{4,9\}$
	$\emptyset$	$\{1\}$	$\{5,8\}$

$i-4$	$i-3$	$i-2$	$i-1$	i
	$\emptyset$	$\emptyset$	$\{7\}$	$\{1,3,5\}$
	$\emptyset$	$\{7\}$	$\emptyset$	$\{1,3,5\}$
		$\emptyset$	$\emptyset$	$\{1,3,5\}$
		$\emptyset$	$\emptyset$	$\{1,3,7\}$
	$\emptyset$	$\{2\}$	$\emptyset$	$\{1,3,8\}$
		$\emptyset$	$\emptyset$	$\{1,3,8\}$
	$\emptyset$	$\{2\}$	$\emptyset$	$\{1,3,9\}$
		$\emptyset$	$\emptyset$	$\{1,3,9\}$
	$\emptyset$	$\{1\}$	$\emptyset$	$\{1,4,7\}$
		$\emptyset$	$\emptyset$	$\{1,4,7\}$
	$\emptyset$	$\{1\}$	$\emptyset$	$\{1,4,9\}$
		$\emptyset$	$\emptyset$	$\{1,4,9\}$
	$\emptyset$	$\{1\}$	$\emptyset$	$\{1,5,9\}$
		$\emptyset$	$\emptyset$	$\{1,5,9\}$
	$\emptyset$	$\{1\}$	$\emptyset$	$\{2,5,8\}$
		$\emptyset$	$\emptyset$	$\{2,5,8\}$
$\emptyset$	$\emptyset$	$\emptyset$	$\{9\}$	$\{1,3,5,7\}$
$\emptyset$	$\emptyset$	$\{9\}$	$\emptyset$	$\{1,3,5,7\}$
$\emptyset$	$\{9\}$	$\emptyset$	$\emptyset$	$\{1,3,5,7\}$
	$\emptyset$	$\emptyset$	$\emptyset$	$\{1,3,5,7\}$
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{1,3,5,7,9\}$

6. Sum of exponents

We give in this section our bound on the sum of exponents which relies heavily on the results we have proved so far. The strategy is similar. We show that the sum of exponents of runs with periods four or less is at most $2n$.

Lemma 3 *The sum of exponents of runs with periods at most 4 in a string of length n is bounded by $2n$.*

Proof. The idea is similar to the one in the proof of Lemma 2 except that here it is not clear how many positions we need to check. The problem is that exponents can be arbitrarily large and therefore no fixed-size interval can amortize them. Therefore, we need some changes. First, we shall choose the intervals to the right, as that is the direction in which the exponents increase. Second, we shall amortize periods and not runs, that is, in our arguments we shall not use the non-left-extendability of runs. We

need to do this in order to amortize different parts of the same run separately.

We shall say that w *has period* (p, e) *at* i if $w[i-p\,..\,i+(e-1)p-1]$ has period p, $e \geq 2$, and $w[i-p\,..\,i-1]$ is primitive. (What is missing from having a run is the non-extendability.) Also, w has period $(p, \_)$ at i if it has (p, e) for some e.

Assume also that the maximum period allowed for the microruns in this section is MAX_PER $= 4$. We shall make the reasoning for an arbitrary MAX_PER though, so that it becomes clear that the whole procedure can be automatized.

Given a string w and a position i, we say that $[i\,..\,j]$ is an *amortizing interval for* i *in* w if

$$\sum_{k=i}^{j} \sum_{\substack{p \leq \text{MAX\_PER} \\ w \text{ has period } (p,\_) \text{ at } k}} \max\{e \mid w \text{ has } (p,e) \text{ at } k, k+p(e-2) \leq j\} \leq 2(j-i+1)\ .$$

The idea is to consider only those exponents which correspond to the parts of the runs that need to be amortized by the current interval $[i\,..\,j]$, that is, those parts that do not stretch past j more than $p-1$ positions. If a run stretches more than p positions past j, then there is at least a full square with center outside the interval and it will be amortized, if needed, by a disjoint interval.

All we need to prove is that there exists a fixed upper bound on the lengths of amortizing intervals for all strings and all positions. The general strategy consists of considering all possibilities of periods at a given position for which the sum of exponents is larger than 2. Then we look for an amortizing interval to the right. The exponents are considered according to the above formula and updated, if needed, when the interval is increased.

For MAX_PER $= 4$, we give in Table 3 all possibilities of periods and exponents that can be encountered. Each line gives the pairs (period, exponent) corresponding to all positions in the amortizing interval $[i\,..\,j]$, where $j \leq$ MAX_PER $- 1 = 3$. Recall that the exponents represent only the amortized part; the period may continue past the end of the interval arbitrarily. Some exponents are ranges of the form $\frac{s}{p}\,..\,\frac{s+o}{p}$, which means that the same entry is obtained for any exponent $\frac{t}{p}$, for $s \leq t \leq s+o$.

Just to give an example, let us discuss the case when both periods 1 and 3 appear at i. We have the factor $w[i-3\,..\,i+2] =$ abaaba. Notice that the exponent for period 1 at i is maximal. There is no exponent at $i+1$ which is enough to amortize the exponents at i provided that the one for 3 does not extend to the right; we obtain the entry $(\{(1,2),(3,2)\},\emptyset)$. However, if the exponent for period 3 at i goes up to $\frac{7}{3}$, the fact that there is no exponent at $i+1$ is no longer sufficient. We look therefore at position $i+2$ where we find no exponent. This is enough to amortize the exponents at i (and $i+1$) even in the case when the exponent for period 3 at i is $\frac{8}{3}$. We obtain the entry $(\{(1,2),(3,\frac{7}{3}\,..\,\frac{8}{3})\},\emptyset,\emptyset)$. The length of the amortizing interval is 3, which means we need not consider the exponent $\frac{9}{3}$ for the period 3 at i. If it exists, then the exponent $\frac{6}{3}$ or larger corresponding to the period 3 at $i+3$ will be dealt with outside the amortizing interval $[i\,..\,i+2]$ we used for i; in such a case, $\frac{5}{3}$ units of the exponent will be reamortized. □

For runs with periods higher than 4, we shall use the discussion in Section 4 and Fine and Wilf's lemma. The lemma can be rephrased as follows: For two primitive strings x and y, any powers x^α and y^β, $\alpha \geq 2$ and $\beta \geq 2$, cannot have a common factor longer than $|x|+|y|$ as such a factor would have also period $\gcd(|x|,|y|)$, contradicting the primitivity of x and y.

Table 3
The amortizing intervals for the proof of Lemma 3.

i	$i+1$	$i+2$	$i+3$	interval size
$(1,3)$	$\emptyset$			2
$(1,3)$	$(4,\frac{8}{4}..\frac{9}{4})$	$\emptyset$		3
$(2,\frac{5}{2})$	$\emptyset$			2
$(3,\frac{7}{3})$	$\emptyset$			2
$(3,\frac{7}{3}..\frac{8}{3})$	$(1,2)$	$\emptyset$		3
$(4,\frac{9}{4})$	$\emptyset$			2
$(4,\frac{9}{4}..\frac{10}{4})$	$(1,2)$	$\emptyset$		3
$(4,\frac{9}{4}..\frac{10}{4})$	$(1,3)$	$\emptyset$		3
$(1,2),(3,2)$	$\emptyset$			2
$(1,2),(3,\frac{7}{3}..\frac{8}{3})$	$\emptyset$	$\emptyset$		3
$(1,2),(4,2)$	$\emptyset$			2
$(1,2),(4,\frac{9}{4}..\frac{10}{4})$	$\emptyset$	$\emptyset$		3
$(1,2),(4,\frac{9}{4}..\frac{11}{4})$	$\emptyset$	$(1,2)$	$\emptyset$	4
$(1,3),(4,\frac{8}{4}..\frac{10}{4})$	$\emptyset$	$\emptyset$		3

Next consider a fixed δ and two δ-runs, x^α and y^β, $\alpha,\beta\in\mathbb{Q}$, and denote their periods $|x|=p$ and $|y|=q$. The strings x^α and y^β cannot overlap more than $2.5\min(p,q)$ as otherwise Fine and Wilf's lemma would imply that x and y are not primitive, a contradiction. Therefore, their suffixes $x^{\alpha-2.5}$ and $y^{\beta-2.5}$ (assuming the exponents large enough) cannot overlap at all. Therefore, the sum of exponents of δ-runs is at most 2.5 times the number of runs plus whatever exponent is left of each run after removing the prefix of exponent 2.5. For x^α, that means $\alpha-2.5=\frac{|x^{\alpha-2.5}|}{|x|}\leq\frac{|x^{\alpha-2.5}|}{2\delta}$ and when summing up all these, as they cannot overlap, we obtain $\frac{n}{2\delta}$.

Assuming that the number of runs as above is at most $\frac{n}{\delta}$ and using Lemma 3, we obtain the following bound on the sum of exponents, where $\delta_i=\frac{5}{2}\left(\frac{3}{2}\right)^i$, $i\geq 0$:

$$2n+\sum_{i=2}^{\infty}\left(2.5\frac{n}{\delta_i}+\frac{n}{2\delta_i}\right)=2n+\left(3\frac{2}{5}\sum_{i=2}^{\infty}\left(\frac{2}{3}\right)^i\right)n=5.6n\ . \tag{2}$$

Notice however, that in our analysis from Section 4, for the case (i) of many runs with the same center, we accounted for some of the runs using other runs with periods belonging to a different δ-class. That means the number of runs for each δ need not be $\frac{n}{\delta}$. Still our bound is exact because the runs we account for in case (i) have very small exponents. Recall that we need to account, for each δ, for all runs but one. Using the notation from case (i), any run $x_j^{\alpha_j}$, $2\leq j\leq h-1$, cannot extend its period $|x_j|$ by more than ℓ positions to the right past the end of the initial square, and therefore has $\alpha_j\leq 2+\frac{1}{j}\leq 2.5$. The runs with the shortest and the longest periods, $x_1^{\alpha_1}$ and $x_h^{\alpha_h}$, respectively, may have arbitrarily large exponents but we need not account for either one. The bound in (2) therefore holds and we proved

Theorem 2 *The sum of exponents of runs in a string of length n is less than $5.6n$.*

7. Comments

A small improvement of our main result in Theorem 1 can be rather easily obtained as follows. We can make a better choice of the δ_is to cover all periods:
$\delta_0 = \frac{10}{2}$ – covers the periods between 10 and 15,
$\delta_1 = \frac{16}{2}$ – covers the periods between 16 and 24,
$\delta_i = \frac{25}{2}\left(\frac{3}{2}\right)^{i-2}$, for all $i \geq 2$ – cover all periods larger than or equal to 25.
The bound becomes:

$$n + \left(\frac{2}{10} + \frac{2}{16} + \frac{2}{25}\sum_{i=2}^{\infty}\left(\frac{2}{3}\right)^{i-2}\right)n = 1.565n\ .$$

The method of choosing values can be extended in the same manner to all δ_is but the improvements to the final bound are less and less significant. One would have to modify the proof of Theorem 1 to accommodate these changes, which is not difficult to do, but we preferred to keep the proof simpler.

One could also try to improve the interval $[2\delta\,..\,3\delta]$ in the definition of δ-closeness but the reasoning becomes more complicated.

The proof technique in Section 5 can be automatized so that larger periods for microruns can be considered. If one can prove it, for instance, for microruns with periods up to 32, then the bound improves to $1.18n$ (here we kept the same interval $[2\delta\,..\,3\delta]$ but included the improvement described above in this section, using better choice of the δ_is). A similar computer-aided approach can be applied to the bound on the sum of exponents which could be improved down to $2.9n$, assuming one can verify that the result in Lemma 3 holds for runs with periods up to 20.

Actually solving the conjecture using the above approach may be possible. For instance, one could attempt to verify by computer that the number of runs with periods less than 40 is at most $0.85n$ (the remaining ones are less than $0.15n$ by our reasoning). An efficient implementation of our implicit algorithm, based on Lemma 1, that we used in the proof of Lemma 2 is necessary.

We need to comment a bit more here. Our approach essentially approximates the number of runs, as it is very difficult, if not impossible, to account for all runs in this way. Therefore, the fact that we can attempt solving the conjecture shows, on the one hand, that our approach must be very close to reality, that is, the approximation we obtain is very good, yet, on the other, we believe that the bound n is not optimal as we seem to be able to get too close to it. Recall however, that it has to be more than $0.92n$, according to the lower bound of [6], that means, not too far away.

Another promising approach is extending Lemma 1 to cover all periods. Of course, removing the bound on the length of periods of microruns in Lemma 1 makes it identical to the conjecture but we believe that the proof supporting the result can be extended. Precisely, we conjecture that each position with two or more centers can be amortized within at most half of the length of the longest possible period of a run, that is, at most a quarter of the length of the string.

8. Further research

We discuss here several related problems. The first one is old but the others are proposed here.

Squares. As the number of all square occurrences in a string may be quadratic and that of primitively rooted square occurrences can still be superlinear, as already mentioned in the Introduction, it is natural to count squares, that is, each square is counted only once no matter how many occurrences it has. As proved by Fraenkel and Simpson [5], there are at most $2n$ squares in a string of length n. A simple proof has been given by Ilie [7]. Based on the numerical evidence, it has been conjectured that this number is actually less than n; see also Chapter 8 of [14]. The best bound to date is $2n - \Theta(\log n)$ due to Ilie [8].

Runs. In this paper we counted in fact occurrences of repetitions because the runs are defined as intervals. Inspired by the square problem, we may look at their associated strings and count only the number of runs associated with different strings. Notice that the number of nonequivalent runs and that of squares do not seem to be obviously related to each other. The same run may contain several distinct squares (e.g., ababa contains the squares abab and baba) but we can have also different runs corresponding to a single squares (e.g., aa and aaa can be different runs but only the square aa is involved).

$(2+\varepsilon)^+$*-repetitions.* A way to weaken the conjecture on the number of squares is to increase the exponent of the repetition. Given a non-negative ε, one could count only the number of repetitions of exponent $2+\varepsilon$ or higher. We need first to make it precise what we are talking about. We count primitively rooted repetitions of exponent at least $2+\varepsilon$ and having distinct roots. That is, x^α and y^β, x and y primitive, $\alpha \geq 2+\varepsilon$, $\beta \geq 2+\varepsilon$, are different if and only if $x \neq y$.

This conjecture might be easier to prove. At least for $2+\varepsilon = 1+\phi$, where ϕ is the golden ratio, we can prove it immediately. We count each square at the position where its rightmost occurrence starts and show that no two distinct squares can have the same rightmost starting position. Assume $x^{1+\phi}$ is a prefix of $y^{1+\phi}$ and denote $|x| = p < q = |y|$. Then necessarily $|x^{1+\phi}| = (1+\phi)p > \phi q = |y^\phi|$ as otherwise $x^{1+\phi}$ would have another occurrence to the right. That means $\phi^2 p = (1+\phi)p > \phi q$, or $\phi p > q$. Therefore, the overlap between the two runs has the length $|x^{1+\phi}| = (1+\phi)p = p+\phi p > p+q$. By Fine and Wilf's lemma, this means x and y are powers of the same string and therefore not primitive, a contradiction.

$(2-\varepsilon)^+$*-repetitions.* This is similar to the previous problem except that now we consider repetitions of exponent $2-\varepsilon$ or higher. Is the number of such maximal repetitions still linear? If this is false for any $\varepsilon > 0$, then 2 is the optimal threshold. Otherwise, the optimal threshold needs to be found.

Acknowledgements

We would like to thank Liviu Tinta for pointing out a couple of errors in Table 2.

References

[1] A. Apostolico and F. Preparata, Optimal off-line detection of repetitions in a string, *Theoret. Comput. Sci.* **22**(3) (1983) 297 – 315.

[2] M. Crochemore, An optimal algorithm for computing the repetitions in a word, *Inform. Proc. Letters* **12** (1981) 244 – 250.

[3] M. Crochemore and L. Ilie, A simple proof that the number of runs in a word is linear, manuscript, 2006.

[4] M. Crochemore and W. Rytter, Squares, cubes, and time-space efficient string searching, *Algorithmica* **13** (1995) 405 – 425.

[5] A.S. Fraenkel and J. Simpson, How many squares can a string contain?, *J. Combin. Theory, Ser. A,* **82** (1998) 112 – 120.

[6] F. Franek, R.J. Simpson, and W.F. Smyth, The maximum number of runs in a string, *Proc. 14th Australasian Workshop on Combinatorial Algorithms*, M. Miller and K. Park (eds.) (2003) 26 – 35.

[7] L. Ilie, A simple proof that a word of length n has at most $2n$ distinct squares, *J. Combin. Theory, Ser. A*, **112**(1) (2005) 163 – 164.

[8] L. Ilie, A note on the number of squares in a word, *Theoret. Comput. Sci.* 380(3) (2007) 373 – 376.

[9] C.S. Iliopoulos, D. Moore, and W.F. Smyth, A characterization of the squares in a Fibonacci string, *Theoret. Comput. Sci.* **172** (1997) 281 – 291.

[10] R. Kolpakov and G. Kucherov, On the sum of exponents of maximal repetitions in a word, Tech. Report 99-R-034, LORIA, 1999.

[11] R. Kolpakov and G. Kucherov, Finding maximal repetitions in a word in linear time, *Proc. of FOCS'99*, IEEE Computer Society Press, 1999, 596 – 604.

[12] R. Kolpakov and G. Kucherov, On maximal repetitions in words, *J. Discrete Algorithms* **1**(1) (2000) 159 – 186.

[13] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge Univ. Press, 2002.

[14] M. Lothaire, *Applied Combinatorics on Words*, Cambridge Univ. Press, 2005.

[15] M.G. Main, Detecting lefmost maximal periodicities, *Discrete Applied Math.* **25** (1989) 145 – 153.

[16] S.J. Puglisi, J. Simpson, and B. Smyth, How many runs can a string contain?, submitted, 2006.

[17] W. Rytter, The number of runs in a string: improved analysis of the linear upper bound, in: B. Durand and W. Thomas (eds.), *Proc. of STACS'06*, Lecture Notes in Comput. Sci. **3884**, Springer-Verlag, Berlin, Heidelberg, 2006, 184 – 195.

[18] W. Rytter, The number of runs in a string, submitted, 2006.

[19] J. Stoye and D. Gusfield, Simple and flexible detection of contiguous repeats using a suffix tree, in: M. Farach-Colton, ed., *Proc. of the 9th CPM*, Lecture Notes in Comput. Sci., **1448**, Springer, 1998, 140 – 152.

[20] Thue, A., Über unendliche Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl., Cristiana* **7**, 1906.