Understanding maximal repetitions in strings

Maxime Crochemore, Lucian Ilie

Symposium on Theoretical Aspects of Computer Science 2008 (Bordeaux), pp. 11-16
www.stacs-conf.org

UNDERSTANDING MAXIMAL REPETITIONS IN STRINGS

MAXIME CROCHEMORE [1] AND LUCIAN ILIE [2]

[1] King's College London, Strand, London WC2R 2LS, United Kingdom
and Institut Gaspard-Monge, Université Paris-Est, France
E-mail address: maxime.crochemore@kcl.ac.uk

[2] Department of Computer Science, University of Western Ontario
N6A 5B7, London, Ontario, Canada
E-mail address: ilie@csd.uwo.ca

Abstract. The cornerstone of any algorithm computing all repetitions in a string of length n in $\mathcal{O}(n)$ time is the fact that the number of runs (or maximal repetitions) is $\mathcal{O}(n)$. We give a simple proof of this result. As a consequence of our approach, the stronger result concerning the linearity of the sum of exponents of all runs follows easily.

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. One of the most important problems in this area was finding an algorithm for computing all repetitions in linear time. A major obstacle was encoding all repetitions in linear space because there can be $\Theta(n \log n)$ occurrences of squares in a string of length n (see [1]). All repetitions are encoded in runs (that is, maximal repetitions) and Main [9] used the s-factorization of Crochemore [1] to give a linear-time algorithm for finding all leftmost occurrences of runs. What was essentially missing to have a linear-time algorithm for computing all repetitions, was proving that there are at most linearly many runs in a string. Iliopoulos et al. [4] showed that this property is true for Fibonacci words. The general result was achieved by Kolpakov and Kucherov [7] who gave a linear-time algorithm for locating all runs in [6].

Kolpakov and Kucherov proved that the number of runs in a string of length n is at most cn but could not provide any value for the constant c. Recently, Rytter [10] proved that $c \leq 5$. The conjecture in [7] is that $c = 1$ for binary alphabets, as supported by computations for string lengths up to 31. Using the technique of this note, we have proved [2] that it is smaller than 1.6, which is the best value so far.

1998 ACM Subject Classification: F.2.2 Nonnumerical Algorithms and Problems; G.2.1 Combinatorics.
Key words and phrases: combinatorics on words, repetitions in strings, runs, maximal repetitions, maximal periodicities, sum of exponents.
This work has been done during the second author's stay at Institut Gaspard-Monge. The same author's research was supported in part by NSERC.

Both proofs in [6] and [10] are very intricate and our contribution is a simple proof of the linearity. On the one hand, the search for a simple proof is motivated by the very importance of the result – this is the core of the analysis of any optimal algorithm computing all repetitions in strings. None of the above-mentioned proofs can be included in a textbook. We believe that the simple proof shows very clearly why the number of runs is linear. On the other hand, a better understanding of the structure of runs could pave the way for simpler linear-time algorithms for finding all repetitions. For the algorithm of [6] (and [9]), relatively complicated and space-consuming data structures are needed, such as suffix trees.

The technical contribution of the paper is based on the notion of δ-close runs (runs having close centers), which is an improvement on the notion of neighbors (runs having close starting positions) introduced by Rytter [10].

On top of that, our approach enables us to derive easily the stronger result concerning the linearity of the sum of exponents of all runs of a string. Clearly this result implies the first one, but the converse is not obvious. The second result was given another long proof in [7]; it follows also from [10].

Finally, we strongly believe that our ideas in this paper can be further refined to improve significantly the upper bound on the number of runs, if not to prove the conjecture. The latest refinements and computations (December 2007) show a $1.084n$ bound.

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, by $w[i]$ its ith letter, and by $w[i\mathinner{.\,.}j]$ its factor $w[i]w[i+1]\cdots w[j]$. We say that w has period p iff $w[i] = w[i+p]$, for all $1 \le i \le |w| - p$. The smallest period of w is called *the period* of w and 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 integer (two or more) power of another string. Any nonempty string can be uniquely written as an integer power of a primitive string, called its *primitive root*. It can also be uniquely written in the form $u^e v$ where $|u|$ is its (smallest) period, e is the integral part of its exponent, and v is a proper prefix of u.

The following well-known *synchronization* property will be useful: 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 [8] for all concepts used here.

For a string $w = w[1\mathinner{.\,.}n]$, a *run*[1] (or maximal repetition) is an interval $[i\mathinner{.\,.}j]$, $1 \le i < j \le n$, such that (i) the factor $w[i\mathinner{.\,.}j]$ is periodic (its exponent is 2 at least) and (ii) both $w[i-1\mathinner{.\,.}j]$ and $w[i\mathinner{.\,.}j+1]$, if defined, have a strictly higher (smallest) period. As an example, consider $w = \mathsf{abbababbaba}$; $[3\mathinner{.\,.}7]$ is a run with period 2 and exponent 2.5; we have $w[3\mathinner{.\,.}7] = \mathsf{babab} = (\mathsf{ba})^{2.5}$. Other runs are $[2\mathinner{.\,.}3], [7\mathinner{.\,.}8], [8\mathinner{.\,.}11], [5\mathinner{.\,.}10]$ and $[1\mathinner{.\,.}11]$. For a run starting at i and having period $|x| = p$, we shall call $w[i\mathinner{.\,.}i+2p-1] = x^2$ the *square* of the run (this is the only part of a run we can count on). Note that x is primitive

[1]Runs were introduced in [9] under the name *maximal periodicities*; the are called *m-repetitions* in [7] and *runs* in [4].

and the square of a run cannot be extended to the left (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 $e_x = i_x + 2p - 1$, and its *center* by $c_x = i_x + p$.

3. Linear number of runs

We describe in this section our proof of the linear number of runs. The idea is to partition the runs by grouping together those having close centers and similar periods. To this aim, for any $\delta > 0$, we say that two runs having squares x^2 and y^2 are *δ-close* if (i) $|c_x - c_y| \le \delta$ and (ii) $2\delta \le |x|, |y| \le 3\delta$. We prove that there cannot be more than three mutually δ-close runs. (There is one exception to this rule – case (vi) below – but then, even fewer runs are obtained.) This means that the number of runs with the periods between 2δ and 3δ in a string of length n is at most $\frac{3n}{\delta}$. Summing up for values $\delta_i = \frac{1}{2}\left(\frac{3}{2}\right)^i$, $i \ge 0$, all periods are considered and we obtain that the number of runs is at most

$$\sum_{i=0}^{\infty} \frac{3n}{\delta_i} = \sum_{i=0}^{\infty} \frac{3n}{\frac{1}{2}\left(\frac{3}{2}\right)^i} = 18n. \tag{3.1}$$

For this purpose, we start investigating what happens when three runs in a string w are δ-close. Let us denote their squares by x^2, y^2, z^2, their periods by $|x| = p$, $|y| = q$, $|z| = r$, and assume $p \le q \le r$. We discuss below all the ways in which x^2 and y^2 can be positioned relative to each other and see that long factors of both runs have small periods which z^2 has to synchronize. This will restrict the beginning of z^2 to only one choice as otherwise some run would be left extendable. Then a fourth run δ-close to the previous three cannot exist.

Notice that, for cases (i)-(v) we assume the centers of the runs are different; the case when they coincide is covered by (vi).

(i) $(i_y < i_x <) c_y < c_x < e_x \le e_y$. Then x and the suffix of length $e_y - c_x$ of y have period $q - p$; see Fig. 1(i). We may assume the string corresponding to this period is a primitive string as otherwise we can make the same reasoning with its primitive root.

Since z^2 is δ-close to both x^2 and y^2, it must be that $c_z \in [c_x - \delta \mathinner{.\,.} c_y + \delta]$. Consider the interval of length $q - p$ that ends at the leftmost possible position for c_z, that is, $I = [c_x - \delta - (q-p) \mathinner{.\,.} c_x - \delta - 1]$. It is included in the first period of z^2, that is, $[i_z \mathinner{.\,.} c_z - 1]$, and in $[i_x \mathinner{.\,.} c_y]$. Thus $w[I]$ is primitive and equal, due to z^2, to $w[I + r]$ which is a factor of $w[c_x \mathinner{.\,.} e_y]$. 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: $w = \mathsf{baababababaabababababaab}$, $x^2 = w[5 \mathinner{.\,.} 14] = (\mathsf{ababa})^2$, $y^2 = w[1 \mathinner{.\,.} 14] = (\mathsf{baababa})^2$, and $z^2 = w[3 \mathinner{.\,.} 20] = (\mathsf{abababaab})^2$.

We have already, due to z^2, that $x = \rho^\ell \rho'$, where $|\rho| = q - p$ and ρ' a prefix of ρ. A fourth run δ-close to the previous three would have to have the same beginning as z^2 and the length of its period would have to be also q plus a multiple of $q - p$. This would imply an equation of the form $\rho^m \rho' = \rho' \rho^m$ and then ρ and ρ' are powers of the same string, a contradiction with the primitivity of x.

(ii) $(i_y < i_x <) c_y < c_x < e_y \le e_x$; this is similar with (i); see Fig. 1(ii). Here the prefix of length $e_y - c_x$ of x is a suffix of y and has period $q - p$.

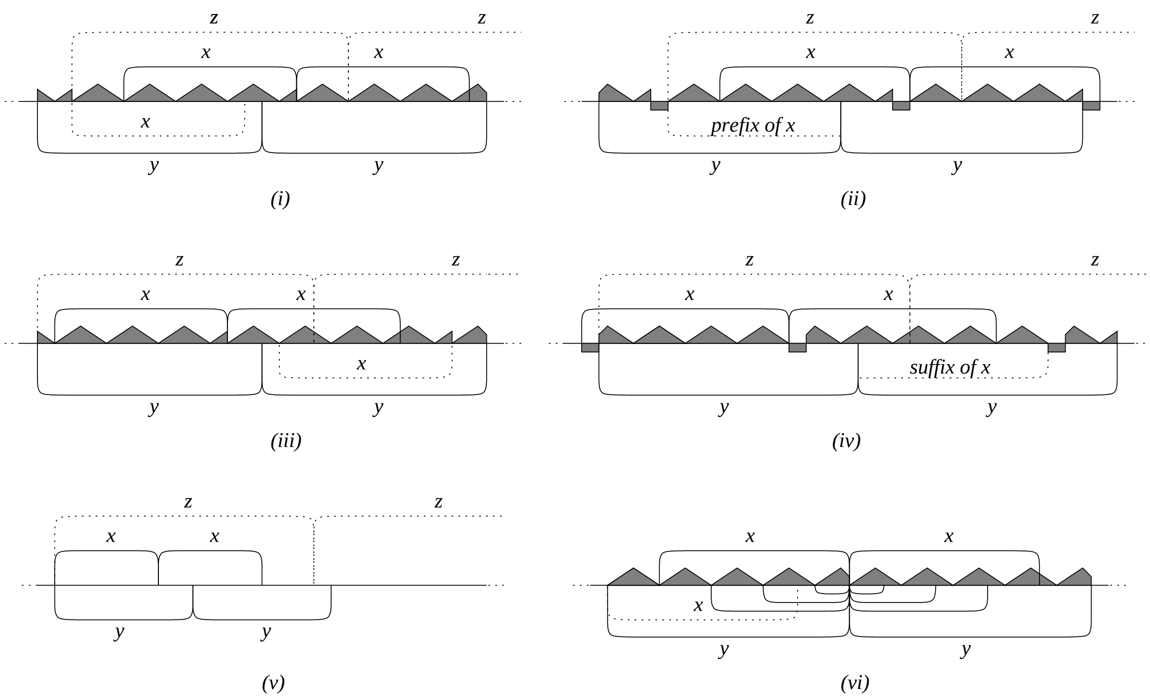

Figure 1: Relative position of x^2 and y^2.

(iii) $i_y < i_x < c_x < c_y (< e_x < e_y)$. Here x and the prefix of length $c_x - i_y$ of y have period $q - p$; see Fig. 1(iii). As above, a third δ-close run z^2 would have to share the same beginning with 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, because of the three-prefix-square lemma[2] of [3], since p is primitive, it would have a period at least $q + r$, which is impossible.

(iv) $i_x < i_y (< c_x < c_y < e_x < e_y)$; this is similar with (iii); see Fig. 1(iv). A third run would begin at the same position as y^2 and there is no fourth run.

(v) $i_x = i_y$; see Fig. 1(v). Here not even a third δ-close run exists because of the three-square lemma that implies $r \geq p + q$.

(vi) $c_x = c_y$. This case is significantly different from the other ones, as we can have many δ-close runs here. However, the existence of many runs with the same center implies very strong periodicity properties of the string which allow us to count the runs globally and obtain even fewer runs than before.

In this case both x and y have the same small period $\ell = q - p$; see Fig. 1(vi). If we note $c = c_y$ then we have h runs $x_j^{\alpha_j}$, $1 \leq j \leq h$, beginning at positions $i_{x_j} = c - ((j-1)\ell + \ell')$, where ℓ' is the length of the suffix of x that is a prefix of the period.

We show that in this case we have less runs than as counted in the sum (3.1). For $h \leq 9$ there is nothing to prove as no four of our $x_j^{\alpha_j}$ runs are counted for the same δ. Assume $h \geq 10$. There exists δ_i such that $\frac{\ell}{2} \leq \delta_i \leq \frac{3\ell}{4}$, that is, this δ_i is considered in (3.1). Then it is not difficult to see that there is no run in w with period between ℓ and $\frac{9}{4}\ell$ and center inside $J = [c + \ell + 1 \mathinner{.\,.} c + (h-2)\ell + \ell']$. But $\ell \leq 2\delta_i < 3\delta_i \leq \frac{9}{4}\ell$ and the length of J is

[2]For three words u, v, w, it states that if uu is a prefix of vv, vv is a prefix of ww, and u is primitive, then $|u| + |v| \leq |w|$.

$(h-3)\ell + \ell' \geq (h+1)\delta_i$. This means that at least h intervals of length δ_i in the sum (3.1) are covered by J and therefore at least $3h$ runs in (3.1) are replaced by our h runs.

We need also mention that these h intervals of length δ_i are not reused by a different center with multiple runs since such centers cannot be close to each other. Indeed, if we have two centers c_j with the above parameters h_j, ℓ_j, $j = 1, 2$, then, as soon as the longest runs overlap over $\ell_1 + \ell_2$ positions, we have $\ell_1 = \ell_2$, due to Fine and Wilf's lemma. Then, the closest positions of J_1 and J_2 cannot be closer than $\ell_1 = \ell_2 \geq \delta_i$ as this would make some of the runs non-primitive, a contradiction. Thus the bound in (3.1) still holds and we proved

Theorem 3.1. *The number of runs in a string of length n is $\mathcal{O}(n)$.*

4. The sum of exponents

Using the above approach, we show in this section that the sum of exponents of all runs is also linear. The idea is to prove that the sum of exponents of all runs with the centers in an interval of length δ and periods between 2δ and 3δ is less than 8. (As in the previous proof, there are exceptions to this rule, but in those cases we get a smaller sum of exponents.) Then a computation similar to (3.1) gives that the sum of exponents is at most $48n$.

To start with, Fine and Wilf's periodicity lemma can be rephrased as follows: For two primitive strings x and y, any powers x^α and y^β 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.

Next consider two δ-close runs, x^α and y^β, $\alpha, \beta \in \mathbb{Q}$. It cannot be that both α and β are 2.5 or larger, as this would imply an overlap of length at least $|x| + |y|$ between the two runs, which is forbidden by Fine and Wilf's lemma since x and y are primitive. Therefore, in case we have three mutually δ-close runs, two of them must have their exponents smaller than 2.5. If the exponent of the third run is less than 3, we obtain the total of 8 we were looking for. However, the third run, say z^γ, $\gamma \in \mathbb{Q}$, may have a larger exponent. If it does, that affects the runs in the neighboring intervals of length δ. More precisely, if $\gamma \geq 3$, then there cannot be any center of run with period between 2δ and 3δ in the next (to the right) interval of length δ. Indeed, the overlap between any such run and z^γ would imply, as above, that their roots are not primitive, a contradiction. In general, the following $\lfloor 2(\gamma - 2.5) \rfloor$ intervals of length δ cannot contain any center of such runs. Thus, we obtain a smaller sum of exponents when this situation is met.

The second exception is given by case (vi) in the previous proof, that is, when many runs share the same center; we use the same notation as in (vi). We need to be aware of the exponent of the run $x_1^{\alpha_1}$, with the smallest period, as α_1 can be as large as ℓ (and unrelated to h, the number of runs with the same center). We shall count α_1 into the appropriate interval of length δ_i; notice that $x_1^{\alpha_1}$ and $x_2^{\alpha_2}$ are never δ-close, for any δ, because $|x_2| > 2|x_1|$. For $2 \leq j \leq h-1$, the period $|x_j|$ cannot be extended by more than ℓ positions to the right past the end of the initial square, and thus $\alpha_j \leq 2 + \frac{1}{j}$. Therefore, their contribution to the sum of exponents is less than $3(h-2)$. They replace the exponents of the runs with centers in the interval J and periods between ℓ and $\frac{9}{4}\ell$ which otherwise would contribute at least $6h$ to the sum of exponents. The run with the longest period, $x_h^{\alpha_h}$, can have an arbitrarily high exponent but the replaced runs in J need to account only for a fraction (3 units) of it since $\alpha_h \geq 3$ implies new centers with multiple runs and hence new J intervals (precisely $\lfloor \alpha_h - 2 \rfloor$) that account for the rest. We proved

Theorem 4.1. *The sum of exponents of the runs in a string of length n is $\mathcal{O}(n)$.*

References

[1] M. Crochemore, An optimal algorithm for computing the repetitions in a string, *Inform. Proc. Letters* **12** (1981) 244 – 250.
[2] M. Crochemore and L. Ilie. Maximal repetitions in strings. *Journal of Computer and System Sciences*, 2007. In press.
[3] M. Crochemore and W. Rytter, Squares, cubes, and time-space efficient string searching, *Algorithmica* **13** (1995) 405 – 425.
[4] C.S. Iliopoulos, D. Moore, W.F. Smyth, A characterization of the squares in a Fibonacci string, *Theoret. Comput. Sci.* **172** (1997) 281 – 291.
[5] R. Kolpakov and G. Kucherov, On the sum of exponents of maximal repetitions in a word, Tech. Report 99-R-034, LORIA, 1999.
[6] 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.
[7] R. Kolpakov and G. Kucherov, On maximal repetitions in words, *J. Discrete Algorithms* **1**(1) (2000) 159 – 186.
[8] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge Univ. Press, 2002.
[9] M.G. Main, Detecting lefmost maximal periodicities, *Discrete Applied Math.* **25** (1989) 145 – 153.
[10] 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, 2006, 184 – 195.