On the maximal number of highly periodic runs in a string

Maxime Crochemore\textsuperscript{1,3}, Costas Iliopoulos\textsuperscript{1,4}, Marcin Kubica\textsuperscript{2}, Jakub Radoszewski\textsuperscript{2}, Wojciech Rytter\textsuperscript{2,5}⋆⋆, and Tomasz Waleń\textsuperscript{2}⋆

\textsuperscript{1} Dept. of Computer Science, King’s College London, London WC2R 2LS, UK [maxime.crochemore,csi]@kcl.ac.uk
\textsuperscript{2} Dept. of Mathematics, Computer Science and Mechanics, University of Warsaw, Warsaw, Poland [kubica,jrad,rytter,walen]@mimuw.edu.pl
\textsuperscript{3} Université Paris-Est, France
\textsuperscript{4} Digital Ecosystems & Business Intelligence Institute, Curtin University of Technology, Perth WA 6845, Australia
\textsuperscript{5} Dept. of Math. and Informatics, Copernicus University, Toruń, Poland

Abstract. A run is a maximal occurrence of a repetition \(v\) with a period \(p\) such that \(2p \leq |v|\). The maximal number of runs in a string of length \(n\) was studied by several authors and it is known to be between 0.944\(n\) and 1.029\(n\). We investigate highly periodic runs, in which the shortest period \(p\) satisfies \(3p \leq |v|\). We show the upper bound 0.5\(n\) on the maximal number of such runs in a string of length \(n\) and construct a sequence of words for which we obtain the lower bound 0.406\(n\).

1 Introduction

Repetitions and periodicities in strings are one of the fundamental topics in combinatorics on words \([2, 13]\). They are also important in other areas: lossless compression, word representation, computational biology etc. Repetitions are studied from different directions: classification of words not containing repetitions of a given exponent, efficient identification of factors being repetitions of different types and finally computing the bounds of the number of repetitions of a given exponent that a string may contain, which we consider in this paper. Both the known results in the topic and a deeper description of the motivation can be found in the survey by Crochemore et al. \([5]\).

The concept of runs (also called maximal repetitions) has been introduced to represent all repetitions in a string in a succinct manner. The crucial property of runs is that their maximal number in a string of length \(n\) (denoted as \(\text{runs}(n)\)) is \(O(n)\) \([10]\). Due to the work of many people, much better bounds on \(\text{runs}(n)\) have

⋆ Research supported in part by the Royal Society, UK.
⋆⋆ Supported by grant N206 004 32/0806 of the Polish Ministry of Science and Higher Education.
been obtained. The lower bound $0.927n$ was first proved in [8]. Afterwards it was improved by Kusano et al. [12] to $0.944n$ employing computer experiments and very recently by Simpson [18] to $0.94575712n$. On the other hand, the first explicit upper bound $5n$ was settled in [15], afterwards it was systematically improved to $3.44n$ [17], $1.6n$ [3, 4] and $1.52n$ [9]. The best known result $\text{run}(n) \leq 1.029n$ is due to Crochemore et al. [6], but it is conjectured [10] that $\text{run}(n) < n$.

The maximal number of runs was also studied for special types of strings and tight bounds were established for Fibonacci strings [10, 16] and more generally Sturmian strings [1].

The combinatorial analysis of runs in strings is strongly related to the problem of estimation of the maximal number of occurrences of squares in a string. In the latter the gap between the upper and lower bound is much larger than for runs [5, 7]. However, a recent paper [11] by some of the authors shows that introduction of exponents larger than 2 can lead to obtaining tighter bounds for the number of corresponding occurrences.

In this paper we introduce and study the concept of highly periodic runs (hp-runs) in which the period is at least three times shorter than the run. We show the following bounds on the number $\text{hp-run}(n)$ of such runs in a string of length $n$:

$$0.406n \leq \text{hp-run}(n) \leq \frac{n - 1}{2}$$

The upper bound is achieved by analyzing prime words (i.e. words that are primitive and minimal/maximal in the class of their cyclic equivalents) that appear as periods of hp-runs. As for the lower bound, we give a simple argument that leads to $0.4n$ bound and then describe a family of words that improves this bound to $0.406n$.

2 Definitions

We consider words over a finite alphabet $A$, $u \in A^*$; by $\varepsilon$ we denote an empty word; the positions in a word $u$ are numbered from 1 to $|u|$. By $\text{Alph}(u)$ we denote the set of all letters of $u$. For $u = u_1u_2 \ldots u_m$, by $u[i..j]$ we denote a factor of $u$ equal to $u_i \ldots u_j$ (in particular $u[i] = u[i..i]$). Words $u[1..i]$ are called prefixes of $u$, and words $u[i..m]$ — suffixes of $u$. We say that positive integer $p$ is the (shortest) period of a word $u = u_1 \ldots u_m$ (notation: $p = \text{per}(u)$) if $p$ is the smallest number such that $u_i = u_{i+p}$ holds for all $1 \leq i \leq m - p$.

If $w^k = u$ ($k$ is a non-negative integer) then we say that $u$ is the $k^{th}$ power of the word $w$. A square is the $2^{nd}$ power of some word. The primitive root of a word $u$, denoted $\text{root}(u)$, is the shortest such word $w$ that $w^k = u$ for some positive $k$. We call a word $u$ primitive if $\text{root}(u) = u$, otherwise it is called nonprimitive. We say that words $u$ and $v$ are cyclically equivalent (or that one of them is a cyclic rotation of the other) if $u = xy$ and $v = yx$ for some $x, y \in A^*$. It is a simple observation that if $u$ and $v$ are cyclically equivalent then $\text{root}(u) = \text{root}(v)$.

Let us assume that $A$ is totally ordered by $\leq$ what induces a lexicographical order in $A^*$, also denoted by $\leq$. We say that $u \in A^*$ is a prime word if it
is primitive and minimal or maximal in the class of words that are cyclically equivalent to it. It can be proved [13] that a prime word $u$ cannot have a proper (i.e. non-empty and different than $u$) prefix that would also be its suffix.

A run (also called a maximal repetition) in a string $u$ is an interval $[i \ldots j]$ such that both the associated factor $u[i \ldots j]$ has period $p$, $2p \leq j - i + 1$, and the property cannot be extended to the right nor to the left: $u[i - 1] \neq u[i + p - 1]$ and $u[j - p + 1] \neq u[j + 1]$ when the letters are defined. A highly periodic run (hp-run) is a run $[i \ldots j]$ for which the shortest period $p$ satisfies $3p \leq j - i + 1$. For simplicity, in the further text we sometimes refer to runs or hp-runs as to occurrences of corresponding factors of $u$.

### 3 Upper bound

Let $u \in A^*$ be a word of length $n$. By $P = \{p_1, p_2, \ldots, p_{n-1}\}$ we denote the set of inter-positions of $u$ that are located between pairs of consecutive letters of $u$.

We define a function $F$ that assigns to each hp-run $v$ in a string the set of handles among all inter-positions within $v$. Hence, $F$ is a mapping from the set of hp-runs occurring in $u$ to the set $2^P$ of subsets of $P$. Let $v$ be a hp-run with period $p$ and let $w$ be the prefix of $v$ of length $p$. By $w_{\min}$ and $w_{\max}$ we denote words cyclically equivalent to $w$ that are minimal and maximal in lexicographical order. We define $F(v)$ as follows:

a) if $w_{\min} \neq w_{\max}$ then $F(v)$ contains inter-positions between consecutive occurrences of $w_{\min}$ and between consecutive occurrences of $w_{\max}$ within $v$

b) if $w_{\min} = w_{\max}$ then $F(v)$ contains all inter-positions within $v$.

**Lemma 1.** $w_{\min}$ and $w_{\max}$ are prime words.

**Proof.** By the definition of $w_{\min}$ and $w_{\max}$, it suffices to show that both words are primitive. This follows from the fact that, due to the minimality of $p$, $w$ is primitive and that $w_{\min}$ and $w_{\max}$ are cyclically equivalent to $w$. 

**Lemma 2.** Case b) from the above definition implies that $|w_{\min}| = 1$.

**Proof.** $w_{\min}$ is primitive, therefore if $|w_{\min}| \geq 2$ then $w_{\min}$ would contain at least two distinct letters, $a = w_{\min}[1]$ and $b = w_{\min}[i] \neq a$. If $b < a$ ($b > a$) then the cyclic rotation of $w_{\min}$ by $i - 1$ letters would be lexicographically smaller (greater) than $w_{\min}$ — a contradiction.

Note that in case b) of the definition of $F$ obviously $F(v)$ contains at least two distinct handles. The following lemma concludes that the same property also holds in case a).

**Lemma 3.** Each of the words $w_{\min}^2$ and $w_{\max}^2$ is a factor of $v$.

**Proof.** Recall that $3p \leq |v|$, where $p = \text{per}(v)$. By Lemma 2, this concludes the proof in case b). As for the proof in case a), it suffices to note that the first occurrences of each of the words $w_{\min}$, $w_{\max}$ within $v$ start non-further than $p$ positions from the beginning of $v$. 

3
We now show a crucial property of $F$.

**Lemma 4.** $F(v_1) \cap F(v_2) = \emptyset$ for every two distinct hp-runs $v_1, v_2$ in $u$.

**Proof.** Assume to the contrary that $p_i \in F(v_1) \cap F(v_2)$ is a handle of two different runs $v_1$ and $v_2$. By Lemmas 1 and 3, $p_i$ is located in the middle of two squares $w_1^2$ and $w_2^2$ of prime words, where $|w_1| = \text{per}(v_1)$ and $|w_2| = \text{per}(v_2)$. $w_1 \neq w_2$, since in the opposite cases runs $v_1$ and $v_2$ would be the same. W.l.o.g. assume that $|w_1| < |w_2|$. Then, word $w_1$ is both a prefix and a suffix of $w_2$ (see fig. 2), what contradicts the primality of $w_2$. ☐

**Fig. 2.** A situation where $p_i$ is in the middle of two different squares $w_1^2$ and $w_2^2$.

The following theorem concludes the analysis of the upper bound.
Theorem 1. A word \( u \in A^* \) of length \( n \) may contain at most \( \frac{n-1}{2} \) runs.

Proof. Due to Lemma 3, for each hp-run \( v \) within \( u \), \( |F(v)| \geq 2 \). Since \( |P| = n-1 \), Lemma 4 implies the conclusion of the theorem. \( \square \)

4 Lower bound

Lemma 5. Let \( s \) be a word and denote:

\[
\begin{align*}
    r &= \text{hp-runs}(s), \\
    \ell &= |s|
\end{align*}
\]

There exists a sequence of words \( (s_n)_{n=0}^{\infty} \), \( s_0 = s \), such that

\[
\begin{align*}
    r_n &= \text{hp-runs}(s_n), \\
    \ell_n &= |s_n| \quad \text{and} \quad \lim_{n \to \infty} \frac{r_n}{\ell_n} = \frac{r}{\ell} + \frac{1}{5\ell}
\end{align*}
\]

Proof. We define the sequence \( s_n \) recursively. Denote \( A = \text{Alph}(s_n) \) and let \( \overline{A} \) be a disjoint copy of \( A \). By \( \overline{A} \) we denote the word obtained from \( s_n \) by substituting letters from \( A \) with the corresponding letters from \( \overline{A} \). We define \( s_{n+1} = (s_n \overline{s_n})^3 \).

Recall that \( \ell_0 = \ell \), \( r_0 = r \) and note that for \( n \geq 1 \)

\[
\ell_n = 6\ell_{n-1}, \quad r_n = 6r_{n-1} + 1
\]

By simple induction this concludes that

\[
\frac{r_n}{\ell_n} = \frac{r}{\ell} + \frac{1}{5\ell} \sum_{i=1}^{n} \frac{1}{6^i} = \frac{r}{\ell} + \frac{1}{5\ell} \left( 1 - \frac{1}{6^{n+1}} \right)
\]

Taking \( n \to \infty \) in the above formula we obtain the conclusion of the lemma. \( \square \)

Starting with the 3-letter word \( s = a^3 \) for which \( r/\ell = 1/3 \), from Lemma 5 we obtain the bound \( 0.4n \). This bound is, however, not optimal — we will show an example of a sequence of words for which we obtain the bound \( 0.406n \).

Let \( A = \{a, b\} \). We denote:

\[
X = (a^3b^3)^3, \quad Y = a^4b^3a, \quad \alpha = XY, \quad \beta = Xa
\]

Lemma 6. A couple of important properties of words \( \alpha \) and \( \beta \):

\[
\begin{align*}
    &\text{– } XYX \text{ introduces a new hp-run with the period 7. Hence, each of the pairs} \\
    &\alpha\alpha \text{ and } \alpha\beta \text{ introduces a new hp-run.} \\
    &\text{– } \beta \text{ is a prefix of } \alpha. \text{ Hence, } \alpha\beta\alpha\beta\alpha \text{ introduces the hp-run } (\alpha\beta)^3. \\
    &\text{– } Y \text{ is a prefix of } aX, \text{ therefore } \alpha \text{ is a prefix of } \beta\alpha. \text{ Hence, } \alpha\alpha\beta\alpha \text{ introduces} \\
    &\text{the hp-run } \alpha^3.
\end{align*}
\]

Now we will also be dealing with a new alphabet \( A' = \{\alpha, \beta\} \). We define the Fibonacci morphism \( h \) as:

\[
h(\alpha) = \alpha\beta, \quad h(\beta) = \alpha
\]

Let

\[
\begin{align*}
    f_n &= h^n(\alpha), \quad r_n = \text{hp-runs}(f_n), \quad \ell_n = |f_n|
\end{align*}
\]
| n | \( r_n \) | \( \ell_n \) | \( r_n/\ell_n \) | \( f_n \) |
|---|---|---|---|---|
| 0 | 9 | 26 | 0.3462 | \( \alpha \) |
| 1 | 17 | 45 | 0.3778 | \( \alpha \beta \) |
| 2 | 26 | 71 | 0.3662 | \( \alpha \beta \alpha \) |
| 3 | 45 | 116 | 0.3879 | \( \alpha \beta \alpha \beta \) |
| 4 | 71 | 187 | 0.3796 | \( \alpha \beta \alpha \beta \alpha \beta \alpha \) |
| 5 | 119 | 303 | 0.3927 | \( \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \) |
| 6 | 192 | 490 | 0.3918 | \( \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \) |

Table 1: A first few words of the sequence \( f_n \) with the corresponding terms of sequences \( r_n \) and \( \ell_n \).

**Theorem 2.**

\[
\lim_{n \to \infty} \frac{r_n}{\ell_n} > 0.406
\]

In particular,

\[
\frac{r_{19}}{\ell_{19}} \geq \frac{103,664}{255,329} > 0.406
\]

**Proof.** We start with the values \( \ell_n, r_n \) for \( n \leq 4 \) that are precomputed in Table 1 and show that for \( n \geq 5 \) the following recursive formulas hold:

\[
\ell_n = \ell_{n-1} + \ell_{n-2}
\]

(1)

\[
r_n \geq r_{n-1} + r_{n-2} + n - 4 \quad \text{if} \quad 2 \mid n
\]

(2)

\[
r_n \geq r_{n-1} + r_{n-2} + n - 2 \quad \text{if} \quad 2 \nmid n
\]

(3)

The “in particular” part of the lemma is a straightforward consequence of the formulas.

(1) is obvious, therefore we concentrate on the inequalities for \( r_n \). The recursive part of each of them \((r_{n-1} + r_{n-2})\) is a consequence of the formula \( f_n = f_{n-1} f_{n-2} \) and the fact that Fibonacci words contain repetitions of exponent at most \( 2 + \Phi < 4 \), see [14]. Due to Lemma 6, for even values of \( n \) a new hp-run is introduced upon concatenation — see the example for \( n = 6 \):

\[
\alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta
\]

and for odd values of \( n \), three more hp-runs appear, as in the following example for \( n = 5 \):

\[
\alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta
\]

\[
\alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta
\]

and for odd values of \( n \), three more hp-runs appear, as in the following example for \( n = 5 \):

\[
\alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta
\]

\[
\alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta \alpha \beta
\]
Apart from that, since
\[
h(\alpha\beta\alpha\beta\alpha) = \alpha\beta\alpha\beta\alpha\beta\alpha\beta
\]
contains a hp-run \(f_2^3\), word \(f_n\) introduces \(n - 5\) new hp-runs composed form \(f_2^3, f_3^3, \ldots, f_n^{n-4}\), each created by iterating \(h^i(\alpha\beta\alpha\beta\alpha)\) — see the example for \(n = 7\):
\[
\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta
\]
\[
\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta
\]
In total, we obtain \(n - 4\) new hp-runs for even \(n\) and \(n - 2\) for odd \(n\), what concludes the proof of the inequalities. \(\square\)

References

1. P. Baturo, M. Piatkowski, and W. Rytter. The number of runs in sturmian words. In O. H. Ibarra and B. Ravikumar, editors, CIAA, volume 5148 of Lecture Notes in Computer Science, pages 252–261. Springer, 2008.
2. J. Berstel and J. Karhumaki. Combinatorics on words: a tutorial. Bulletin of the EATCS, 79:178–228, 2003.
3. M. Crochemore and L. Ilie. Analysis of maximal repetitions in strings. In L. Kucera and A. Kucera, editors, MFCS, volume 4708 of Lecture Notes in Computer Science, pages 465–476. Springer, 2007.
4. M. Crochemore and L. Ilie. Maximal repetitions in strings. J. Comput. Syst. Sci., 74(5):796–807, 2008.
5. M. Crochemore, L. Ilie, and W. Rytter. Repetitions in strings: algorithms and combinatorics. Theoret. Comput. Sci. (to appear).
6. M. Crochemore, L. Ilie, and L. Tinta. Towards a solution to the ”runs” conjecture. In P. Ferragina and G. M. Landau, editors, CPM, volume 5029 of Lecture Notes in Computer Science, pages 290–302. Springer, 2008.
7. M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. Algorithmica, 13(5):405–425, 1995.
8. F. Franek and Q. Yang. An asymptotic lower bound for the maximal number of runs in a string. Int. J. Found. Comput. Sci., 19(1):195–203, 2008.
9. M. Giraud. Not so many runs in strings. In C. Martín-Vide, F. Otto, and H. Fernau, editors, LATA, volume 5196 of Lecture Notes in Computer Science, pages 232–239. Springer, 2008.
10. R. M. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. In Proceedings of the 40th Symposium on Foundations of Computer Science, pages 596–604, 1999.
11. M. Kubica, J. Radoszewski, W. Rytter, and T. Walen. On the maximal number of cubic subwords in a string. In Proceedings of the 20th International Workshop on Combinatorial Algorithms (to appear), 2009.
12. K. Kusano, W. Matsubara, A. Ishino, H. Bannai, and A. Shinohara. New lower bounds for the maximum number of runs in a string. CoRR, abs/0804.1214, 2008.
13. M. Lothaire. Combinatorics on Words. Addison-Wesley, Reading, MA., U.S.A., 1983.
14. F. Mignosi and G. Pirillo. Repetitions in the fibonacci infinite word. ITA, 26:199–204, 1992.
15. W. Rytter. The number of runs in a string: Improved analysis of the linear upper bound. In B. Durand and W. Thomas, editors, STACS, volume 3884 of Lecture Notes in Computer Science, pages 184–195. Springer, 2006.
16. W. Rytter. The structure of subword graphs and suffix trees in fibonacci words. Theor. Comput. Sci., 363(2):211–223, 2006.
17. W. Rytter. The number of runs in a string. Inf. Comput., 205(9):1459–1469, 2007.
18. J. Simpson. Modified padovan words and the maximum number of runs in a word. Australasian Journal of Combinatorics (to appear).