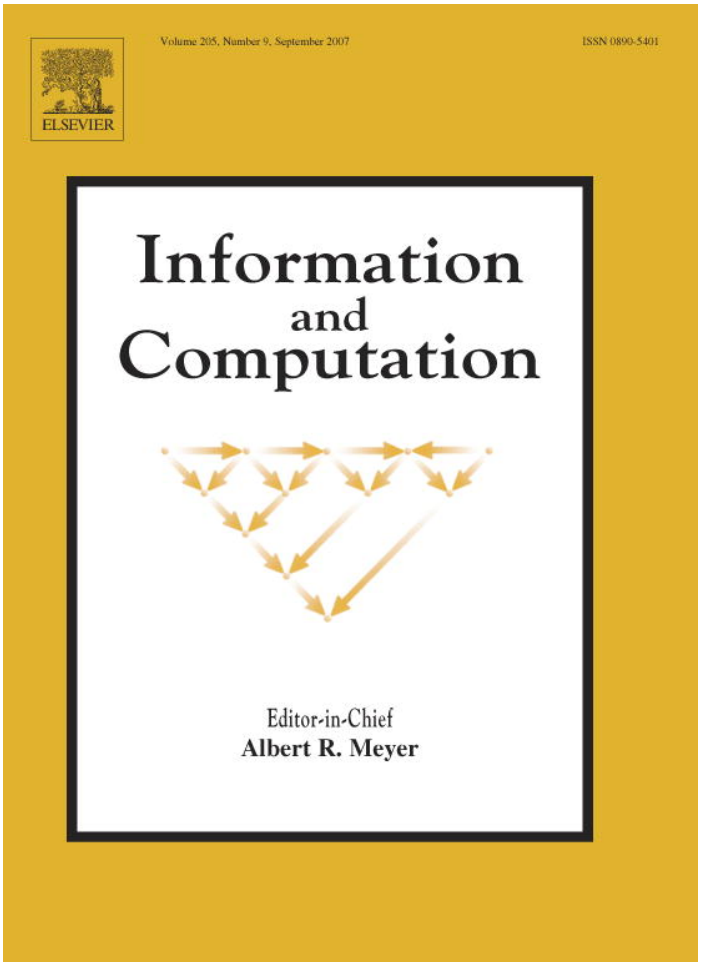

Available online at www.sciencedirect.com
ScienceDirect
Information and Computation 205 (2007) 1459–1469
Information and Computation
www.elsevier.com/locate/ic

The number of runs in a string☆

Wojciech Rytter[1]

Institute of Informatics, Warsaw University, 02–097 Warsaw, Banacha 2, Poland

Received 14 June 2006; revised 15 January 2007
Available online 8 February 2007

Abstract

A *run* in a string is a nonextendable (with the same minimal period) periodic segment in a string. The set of runs corresponds to the structure of internal periodicities in a string. Periodicities in strings were extensively studied and are important both in theory and practice (combinatorics of words, pattern-matching, computational biology). Let $\rho(n)$ be the maximal number of runs in a string of length n. It has been shown that $\rho(n) = O(n)$, the proof was very complicated and the constant coefficient in $O(n)$ has not been given explicitly. We demystify the proof of the linear upper bound for $\rho(n)$ and propose a new approach to the analysis of runs based on the properties of subperiods: the periods of periodic parts of the runs We show that $\rho(n) \leq 3.44\, n$ and there are at most $O.67n$ runs with periods larger than 87. This supports the conjecture that the number of all runs is smaller than n. We also give a completely new proof of the linear bound and discover several new interesting "periodicity lemmas".

Keywords: Run; String; Periodicity

1. Introduction

We consider finite strings over a finite arbitrary alphabet. The set of all runs in a string corresponds to the structure of its regularities. Initial interest was mostly in repetitions of the type xx (so called *squares*), [1,12]. The number of squares, with *primitive* x, is $\Omega(n \log n)$, hence the number of periodicities of this type is not linear. Then, it has been discovered that the number of runs (also called maximal repetitions or repeats) is linear and consequently linear time algorithms for runs were investigated [10,9]. The result of [10] was one of the deepest results related to combinatorics and algorithmics of strings. However, the most intriguing question remained the asymptotically tight bound for the number of runs. The first bound was quite complicated and has not given any *concrete* constant coefficient in $O(n)$ notation. This subject has been studied in [15,16,2]. A beautiful construction showing the lower bound of approximately $0.927\, n$ has been given in [2].

☆ The preliminary version of this paper appeared in [14] Symposium on Theoretical Aspects of Computer Science, STACS 2006, with the weaker result.

Email address: rytter@mimuw.edu.pl

[1] The research was supported by the Grant KBN N206 004 32/0806.

0890-5401/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ic.2007.01.007

The exact number of runs has been considered for special strings: *Fibonacci words* and (more generally) *Sturmian words*, [8,7,13]. In this paper, we make a step towards better understanding of the structure of runs. The proof of the linear upper bound is simplified and small *explicit* constant coefficient is given in $O(n)$ notation.

A period p of a word w is any positive integer p such that $w[i] = w[i+p]$ whenever both sides of this equation are defined. Let $per(w)$ denote the size of the smallest period of w. We say that a word w is **periodic** iff $per(w) \leqslant \frac{|w|}{2}$. A word w is said to be *primitive* iff w is not of a form z^k, where z is a finite word and $k \geqslant 2$ is a natural number.

A **run** in a string w is an interval $\alpha = [i \ldots j]$ such that $w[i \ldots j]$ is a periodic word with the period $p = per(w[i \ldots j])$ and this period is not extendable to the left or to the right of $[i \ldots j]$. In other words, $[i \ldots j]$ is a run iff $|j - i + 1| \geqslant 2p$, $i = 1$ or $w[i-1] \neq w[i-1+p]$ and $j = n$ or $w[j+1] \neq w[j+1-p]$. A run α can be properly included as an interval in another run β, but in this case $per(\alpha) < per(\beta)$.

The value of the run $\alpha = [i \ldots j]$ is $val(\alpha) = w[i \ldots j]$. When it creates no ambiguity we identify sometimes runs with their values although two different runs could correspond to the identical subwords, if we disregard positions of these runs. Hence runs are also called maximal *positioned* repetitions.

Denote by $RUNS(w)$ the set of runs of w, see Fig. 1 for an example.

Denote: $\rho(n) = \max\{|RUNS(w)| : |w| = n\}$. The most interesting and open conjecture about the runs is: $\rho(n) < n$.

We make a small step towards proving validity of this conjecture and show that $\rho(n) \leqslant 3.44\, n$. The proof of linear upper bound in [10] does not give any explicit constant coefficient at all.

Components of a run.

Each value of the run α is a string $x^k y = w[i \ldots j]$, where $|x| = per(\alpha) \geqslant 1$, $k \geqslant 2$ is an integer and y is a proper prefix of x (possibly empty).

The subword x is called the periodic part of the run and denoted by $PerPart(\alpha) = x$. Denote

$$SquarePart(\alpha) = w[i \ldots i + 2\ per(\alpha) - 1],\ center(\alpha) = i + |x|$$

The position i is said to be the *occurrence* of this run and is denoted by $first(\alpha)$. We write $\alpha \prec \beta$ iff $first(\alpha) < first(\beta)$.

Define also $dist(\alpha, \beta) = |first(\alpha)) - first(\beta)|$.

Example. In Fig. 2 we have: $first(\alpha) = 2$, $first(\beta) = 4$, $PerPart(\gamma) = (aba)^4 ab$; and $center(\alpha) = 22$, $center(\beta) = center(\gamma) = 21$

In the paper, the crucial role is played by the runs α with highly periodic $PerPart(\alpha)$. Denote **subperiod** $(\alpha) = per(PerPart(\alpha))$.

Example. In Fig. 2 we have:

$$subperiod(\alpha) = subperiod(\beta) = subperiod(\gamma) = 3.$$

We say that a word w is **highly periodic** (*h-periodic*) if $per(w) \leqslant \frac{|w|}{4}$. A word which is not highly periodic is said to be *weakly periodic*.

Observe that a word can be periodic but at the same time weakly periodic. Also, according to the definition, weakly periodic word can be not periodic.

Algorithmic aspects An efficient algorithm for the computation of all runs was given in [10]. Its basic component is a special decomposition of the string into blocks using a version of Lempel-Ziv compression (see [3,4]) Essentially

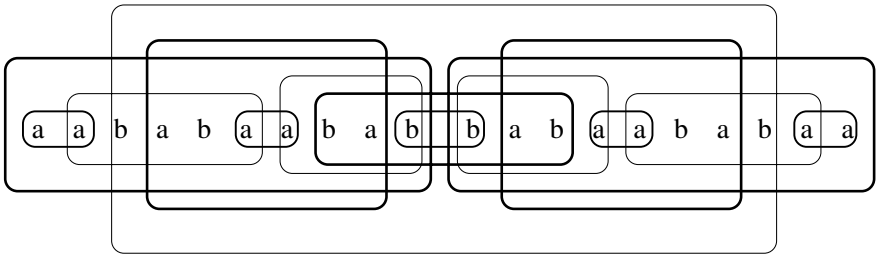

Fig. 1. The structure of $RUNS((aabab)^2(babaa)^2)$.

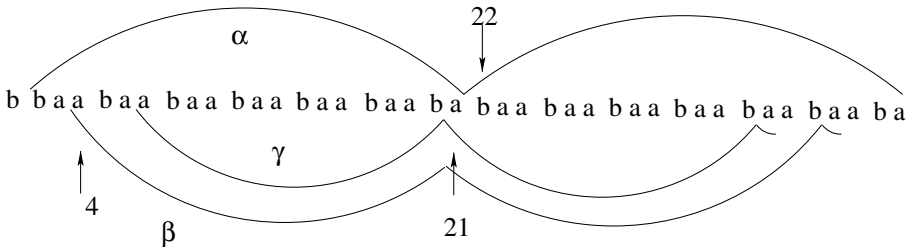

Fig. 2. Example of three runs $\alpha = [2..26]$, $\beta = [4..28]$, $\gamma = [7..31]$ with subperiod 3. We have $\alpha \prec \beta \prec \gamma$ and $dist(\alpha, \beta) = 2$. The runs β, γ are left-periodic (the subperiod 3 continues to the left), α is not. The runs α, β (as well as β, γ) are "neighbors" in sense of Lemma 1. The occurrences (starting positions) of very large runs can be very close. The periodic parts are indicated by the arcs.

the algorithm does not differ very much from the algorithm computing squares in strings. The main problem in the analysis of the efficieny of this algorithm is the estimation of the number of runs, which is the subject of this paper.

2. Two classes of runs and their sparsity properties

Our main approach is to explore sparsity properties of runs, however two very long runs can start at very close positions, so the set of all runs do not have good sparsity property. The key idea is to partition the set of all runs into two classes, for each of them we will have a sparsity property which will be *good* in different senses.

A run α is said to be a **highly periodic run** (an *hp-run*, in short) iff $PerPart(\alpha)$ is h-periodic. The run which is not h-periodic is called a **weakly-periodic** run (*wp-run*).

In Fig. 2 α, β, γ are hp-runs, in Fig. 1 all runs are wp-runs.

Example. In order to understand better the structure of hp-runs we show that the number of hp-runs is $\Omega(\frac{1}{2}n - o(n))$, though we are interested mainly in the upper bound. Let $x_{k,l} = ((01)^k 0)^l$, see Fig. 3. The string $x_{k,k}$ has approximately $\frac{1}{2}n$ hp-runs. The exact formula for its number of hp-runs is $(k-1)(k-4)+1$ with the length of $x_{k,k}$ equal to $n = (2k+1) \cdot k$. Possibly it is asymptotically the maximal number of hp-runs in a string of length n.

Denote $\Delta = \frac{5}{4}$. We define two partitions of the set of runs of a given word:

$$\mathcal{L}(k) = \{first(\alpha) : \alpha \text{ is a wp-run of } w,\ \Delta^k \leqslant per(\alpha) < \Delta^{k+1}\}$$

$$\mathcal{R}(k) = \{first(\alpha) : \alpha \text{ is a hp-run of } w,\ k \leqslant subperiod(\alpha) < 2k\ \}$$

We say that a set $X \subseteq \{1, 2, \ldots, n\}$ is p-**sparse** iff in any interval of size at most p there are at most **two** positions in X.

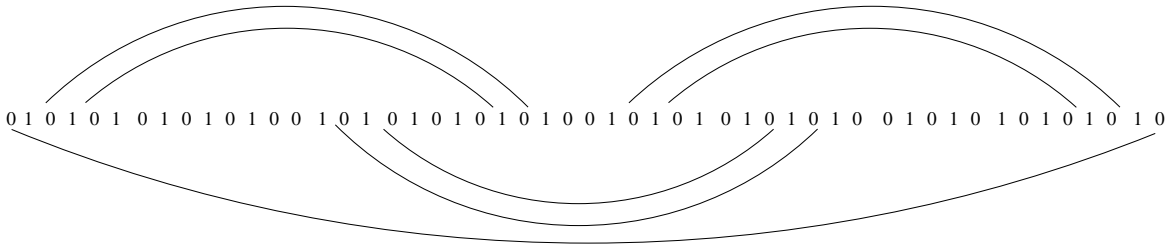

Fig. 3. The string $x_{k,l} = ((01)^k 0)^l$ for $k = 6, l = 4$, and its 7 hp-runs (the general fomula or the number of hp-runs of $x_{k,l}$ is $(l-1)(k-4)+1$).

We assume, to abbreviate the terminology, that we have a fixed word w of length n. Hence, in the notation we omit the dependance on the input word. A tedious proof of the following lemma is given in the last two sections of the paper.

Lemma 1. [Key-Lemma]

(A) *For each* $k \geqslant 1$ *the set* $\mathcal{L}(k)$ *is* $\lceil \frac{1}{4}\Delta^k \rceil$*-sparse.*
(B) *For each* $p > 1$ *the set* $\mathcal{R}(p)$ *is* p*-sparse.*

Example.

(a) In Fig. 5 the runs α, β are in $\mathcal{L}(11)$ (since $\lceil \Delta^{11} \rceil = 12$ and the lengths of periodic parts of α, β are 12, 13, respectively). At the same time they start in the same interval of length $\lceil \frac{1}{4}\Delta^{11} \rceil = 3$.
(b) In Fig. 2 the runs α, β are in $\mathcal{R}(3)$ (since both have subperiod 3) and they start in the same interval of length 3.

Denote by $\mathbf{HP}(n, p)$ the maximal number of hp-runs α with $subperiod(\alpha) \geqslant p$, and by $\mathbf{WP}(n, p)$ the maximal number of wp-runs α with $period(\alpha) \geqslant p$, maximized over strings of length n.

The Key-Lemma implies in a simple way the following fact.

Lemma 2.

$$HP(n, p) \leqslant \tfrac{4}{p} \cdot n; \quad WP(n, \lceil \Delta^r \rceil) \leqslant 40 \cdot \Delta^{-r} \cdot n$$

Proof.

Point 1. Denote by $\mathbf{hp}(n, p)$ the maximal number of hp-runs α with $p \leqslant subperiod(\alpha) < 2p$, maximized over strings of length n.

It follows directly from Lemma 1 that $hp(n, p) \leqslant \frac{2}{p}\, n$. Hence the number of hp-runs with subperiod at least p is bounded from above by:

$$hp(n, p) + hp(n, 2p) + hp(n, 4p) + hp(n, 8p) + \ldots \leqslant 2n \cdot \left(\frac{1}{p} + \frac{1}{2p} + \frac{1}{4p} + \frac{1}{8p} + \ldots \right) \leqslant \frac{4}{p} n$$

Point 2. It follows directly from Lemma 1 that

$$|\mathcal{L}(k)| \leqslant 2 \cdot \left(1/\Delta^k \cdot \frac{1}{4} \right) \cdot n = 8\Delta^{-k} \cdot n$$

Consequently, we have

$$WP(n, \lceil \Delta^r \rceil) \leqslant \sum_{k=r}^{\infty} |\mathcal{L}(k)| \leqslant \sum_{k=r}^{\infty} 8 \cdot \Delta^{-k} \cdot n = 8\Delta^{-r} \times \frac{1}{1 - \Delta^{-1}} \cdot n = 40 \cdot \Delta^{-r} \cdot n \quad \square$$

Observe that $HP(n, 1) = 0$ and $\rho(n) = HP(n, 2) + WP(n, \lceil \Delta^0 \rceil)$. Consequently, we have directly the following corollary of Lemma 1.

Corollary 1. $\rho(n) \leqslant 42n$.

In this way we have a very *simple* (if we disregard technicalities of the proof of Lemma 2.1) proof of a linear upper bound for $\rho(n)$ with an explicit coefficient.

In the next section, we reduce the coefficient from 42 to 3.44. The reduction is done by estimating separately runs with small periodic part.

3. Estimating number of runs with small periodic part

We say that a run α is **large** iff $per(\alpha) > 86$. Otherwise the run is called a **small run**. We estimate now the number of small runs. Let $\rho(n,k)$ be the maximal number of all runs α with $per(\alpha) \leqslant k$, in a string of length n. We estimate the number of runs with small $PerPart(\alpha)$ in a rather naive way using the following lemma.

Lemma 3. *For any given $k \geqslant 1$ there are at most $\frac{1}{k+1}$ n runs with $per(\alpha) = k$ or $per(\alpha) = 2k$.*

Proof. The proof of the following simple fact is illustrated in Fig. 4.

Claim 1. *If u, v are primitive words and $|u| = 2|v|$, then vv is not contained in uu as a subword.*

Assume that $\alpha \prec \beta$ are two different runs with periods k or $2k$.

If $per(\alpha) = per(\beta) = k$ then α, β can have an overlap of size at most $k-1$, otherwise α, β could be merged into a single run. Hence, $first(\beta) - first(\alpha) \geqslant k+1$.

If $per(\alpha) = k$ and $per(\beta) = 2k$ then it is possible that $first(\beta) - first(\alpha) = 1$. Due to the claim the distance from $first(\beta)$ to the occurrence of the next run γ with period k or $2k$ is at least $2k+1$. Then, two consecutive distances give together $(first(\beta) - first(\alpha) + (first(\gamma) - first(\beta)) \geqslant 2k+2$, and "on average" the distance is $k+1$. Therefore, there are at most $\frac{n}{k+1}$ runs with a period k or $2k$. □

The last lemma motivates the introduction of the infinite set Φ, generated by the following algorithm (which never stops).

$\Phi := \emptyset$; $\Psi := \{1, 2, 3, \ldots\}$;
repeat forever
    $k :=$ min Ψ; insert k into Φ;
    remove k and $2k$ from Ψ;

Define the set $\Phi(p) = \{k \in \Phi : k \leqslant p\}$. For example:

$$\Phi(34) = \{1, 3, 4, 5, 7, 9, 11, 12, 13, 15, 16, 17, 19, 20, 21, 23, 25, 27, 28, 29, 31, 33\}$$

For $p \geqslant 1$ define the numbers:

$$\mathcal{H}(p) = \sum_{k \in \Phi(p)} \frac{1}{k+1}.$$

The next lemma follows directly from Lemma 3 and from the structure of the set Φ. We have, by straightforward brute-force calculations, that $\mathcal{H}(86) \leqslant 2.77$.

Lemma 4.

(a) $\rho(n,p) \leqslant \mathcal{H}(p) \times n$.
(b) *There are at most* $2.77n$ *small runs.*

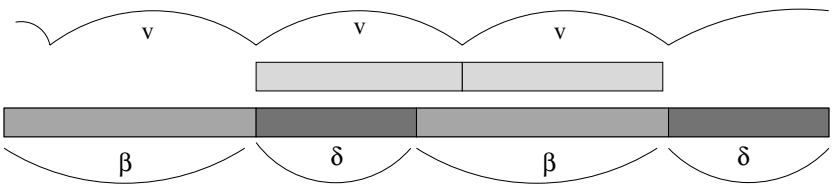

Fig. 4. An occurrence of a square text vv inside a square text uu, where $u = \beta\delta$. The condition $|u| = 2|v|$ implies that $|v|$ is a period of uu, consequently u is not primitive.

4. Estimating the number $\rho(n)$ of all runs

The analysis of the number of large runs is based on the *sparsity* properties of *hp-runs* and *wp-runs*. We estimate separately the number of runs in four disjoint classes:

- large hp-runs with subperiod larger than 21 (at most $\frac{2}{11}n$);
- large hp-runs with subperiod smaller than 22 (at most $\frac{1}{45}n$);
- large wp-runs; (at most $0.4612 \cdot n$)
- small runs (at most $2.77n$, due to Lemma 4).

Denote by $\mathbf{hp}(n, p)$ the maximal number of hp-runs α with $p \leqslant subperiod(\alpha) < 2p$, maximized over strings of length n.

We use the following well known fact, see for example [5] for the proof.

Lemma 5. [Periodicity Lemma] *Let p and q be two periods of the word x. If $p + q \leqslant |x|$, then the greatest common divisor of p, q is also a period of x.*

Lemma 6.

(a) *There are at most $\frac{2}{11}n$ hp-runs with subperiod larger than 21.*
(b) *There are at most $\frac{1}{45}n$ large hp-runs with subperiod smaller than 22.*

Proof. Point (a) follows directly from Lemma 2.

We show now Point (b). Two occurrences of two hp-runs with subperiods not exceeding 21 and with periods larger than 86 have to at least at distance $87 - 2 \cdot 21 = 45$, otherwise they have large overlap implying that these two runs merge into a single one (due to Lemma 5). Hence we have at most $\frac{1}{45}n$ large hp-runs with subperiod smaller than 22. □

Lemma 7. *There are at most $0.4612 \cdot n$ large wp-runs.*

Proof. If we choose $r = 20$ then $\lfloor \Delta^r \rfloor = 86$ and $\lceil \Delta^r \rceil = 87$.

Now it follows from Lemma 2 that the number of large runs is bounded from above by

$$40 \cdot \left(\frac{5}{4}\right)^{-20} \cdot n \leqslant 0.4612 \cdot n \quad \square$$

4.1. Main Result

We can now combine all estimation together and proof the main result which gives a *concrete* constant coefficient in $O(n)$ notation for $\rho(n)$.

Theorem 1. [Main Result]

(1) $\rho(n) \leqslant 3.44\, n$;
(2) *There are at most $0.67n$ large runs in a string of size n.*

Proof. According to Lemma 7, Lemmas 4 and 6 we have at most 2.77 small runs, and at most $\left(\frac{2}{11} + \frac{1}{45} + 0.4612\right) \cdot n \leqslant 0.67n$ large runs.

Putting all together we get: $\rho(n) \leqslant (2.77 + 0.67n) \cdot n = 3.44n$. □

5. The proof of point A of Lemma 2.1

We introduce a useful terminology of *neighboring runs*. We say that two different runs α, β are **neighbors** iff there is a positive number η such that:

$$dist(\alpha,\beta) \leqslant \frac{1}{4}\eta \text{ and } \eta \leqslant per(\alpha), per(\beta) \leqslant \Delta\ \eta$$

Informally, two runs are neighbors iff they have similar periods and are positioned close to each other relatively to their sizes, in particular this means that

$$per(\alpha), per(\beta) \geqslant 4\ |dist(\alpha,\beta)|.$$

Example. In Fig. 5 we have two runs α, β which are neighbors with $\eta = 12$, $per(\alpha) = 13$, $per(\beta) = 12$ and $dist(\alpha,\beta) = 2$.

5.1. The three meighbors lemma

If $\alpha \prec \beta$ and the *square part* of β is not contained in the *square part* of α then we write $\alpha \prec\prec \beta$ (see Fig. 7). More formally:

$$\alpha \sqsupset \beta \quad \text{iff} \quad SquarePart(\beta) \text{ is contained in } SquarePart(\alpha) \text{ as an interval}$$

$$\alpha \prec\prec \beta \quad \text{iff} \quad [\ \alpha \prec \beta \text{ and } \textit{not}\ (\alpha \sqsupset \beta)\ ]$$

Lemma 8.

(a) *If $\alpha \sqsupset \beta$ are distinct neighbors then β is an hp-run.*
(b) *If $\alpha \prec\prec \beta$ are distinct neighbors then the prefix of β of size $per(\alpha) - \delta$ has a period $|q - p|$, where $\delta = first(\beta) - first(\alpha)$ and $p = per(\alpha)$, $q = per(\beta)$.*

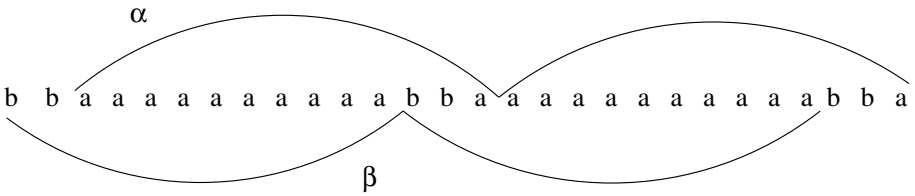

Fig. 5. Two weakly periodic runs α, β which are neighbors with $\eta = 12$. We have: $\alpha, \beta \in \mathcal{L}(11)$ and $dist(\alpha,\beta) < \lceil \frac{1}{4}\Delta^{11} \rceil = 3$.

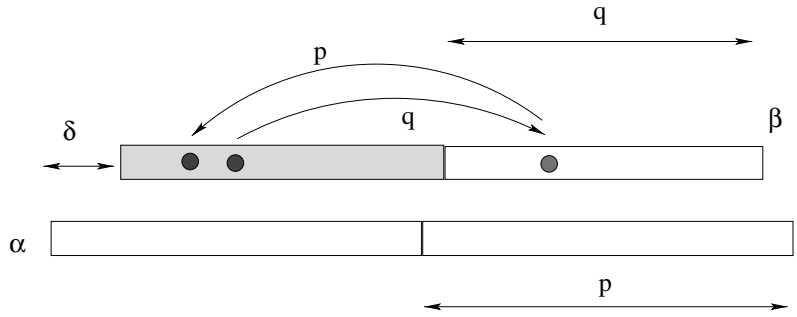

Fig. 6. Two neighbors with $\alpha \sqsupset \beta$, a case $center(\beta) > center(\alpha)$. The *square part* of β is contained in the *square part* of α. The periodic part of β is h-periodic, so it should have a period $p - q$, where $p = per(\alpha)$, $q = per(\beta)$.

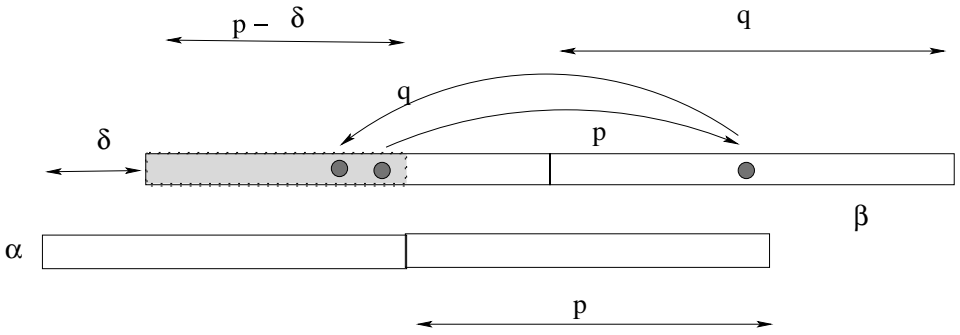

Fig. 7. Two neighbors with $\alpha \prec\prec \beta$, the case $p < q$. The shaded part has the period $|q - p|$, where $p = per(\alpha)$, $q = per(\beta)$.

Proof.

Point (a) We refer the reader to Fig. 6, where the case $center(\beta) > center(\alpha)$ is illustrated. Obviously $p > q$. It is easy to see that the whole $PerPart(\beta)$ has a period $per(\alpha) - per(\beta)$.

Let η be the constant from the definition of neighbors, then

$$per(\alpha) - per(\beta) \leqslant \frac{1}{4}\eta \text{ and } |PerPart(\beta)| \geqslant \eta\,,$$

hence $PerPart(\beta)$ is h-periodic. The case $center(\beta) \leqslant center(\alpha)$ can be considered similarly.

Point (b) We refer to Fig. 7, when only the case $p < q$ is shown. For each position i in the shaded area we have $w[i] = w[i + p] = w[i + p - q]$. The opposite case $p > q$ can be considered similarly. This completes the proof. □

Lemma 9. [The three-neighbors lemma] *If we have three distinct runs $\alpha_1, \alpha_2, \alpha_3$ which are pairwise neighbors with the same number η then at least one of them is h-periodic.*

Proof. If for some i, j $\alpha_i \sqsupset \alpha_j$ then the point (a) of the previous lemma can be applied.

Assume now that we have three runs $\alpha_1 \prec \alpha_2 \prec \alpha_3$ which are pairwise neighbors, with periods $p1, p2, p3$, respectively. Let $\delta_1 = first(\alpha_2) - first(\alpha_1)$, and$\delta_2 = first(\alpha_3) - first(\alpha_2)$. Then, due to Lemma 8 the periodic part of

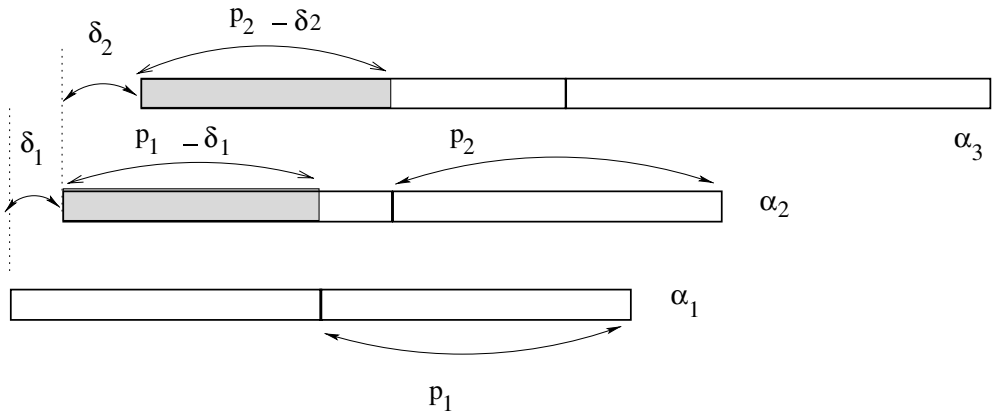

Fig. 8. The three-neighbors lemma, a situation when $\alpha_1 \prec\prec \alpha_2 \prec\prec \alpha_3$. α_2 should be h-periodic, since both its large suffix and large prefix have small periods.

the "middle" run α_2 has a suffix γ_2 of size $p_2 - \delta_2$ with a period $|p_3 - p_2|$ and a prefix γ_1 of size $p_1 - \delta_1$ with a period $|p_2 - p_1|$, see Fig. 8.

Let η be the number from the definition of neighbors. We have

$$\delta_1 + \delta_2 \leqslant \tfrac{1}{4}\eta,\ p_1 \geqslant \eta, \text{ and } |\gamma_1 \cup \gamma_2| = p_2.$$

Hence:

$$|\gamma_1 \cap \gamma_2| \ \geqslant\ (p_2 - \delta_2) + (p_1 - \delta_1) - p_2 \ =\ p_1 - \delta_1 - \delta_2 \ \geqslant\ \frac{3}{4}\eta$$

We have $|p_3 - p_2|, |p_2 - p_1| \leqslant \frac{1}{4}\eta$, hence $per(\gamma_1), per(\gamma_2) \leqslant \frac{1}{4}\eta$. Due to the periodicity lemma $\gamma_1 \cap \gamma_2$ has a period which divides periods of γ_1 and γ_2, and the periodic part of α_2 has a period of size not larger than $\frac{1}{4}\eta$. Consequently, the run α_2 is h-periodic. This completes the proof. □

5.2. The proof of point A of lemma 21

If we take $\eta = \lceil \Delta^k \rceil$ then, as a direct corollary of Lemma 9 we obtain Point A of Lemma 1.

6. The proof of point B of Lemma 2.1.

First we prove the following lemma.

Lemma 10. *Assume we have two distinct hp-runs α, β with the same subperiod p and such that periodic part of one of them is a prefix of the periodic part of another. Then, $dist(\alpha, \beta) \geqslant p$.*

Proof. The proof is by contradiction. Assume that $id = dist(\alpha, \beta) < p$. Denote by w the common prefix of these runs of size $4p$. Such a prefix exists due to high periodicity of these runs. Consequently, w has periods p and $d < p$. Then, due to the *periodicity lemma* (Lemma 5), w has a period which is a proper divisor of p. This implies that the periodic parts of these runs would have subperiod smaller than p, which contradicts the assumption that p is the smallest subperiod. □

We say that a hp-run $\alpha = [i \ldots j]$ of a string w is **left-periodic** iff $w[i-1] = w[i-1+subperiod(\alpha)]$. The runs β, γ in Fig. 2 are left-periodic. We also say that a position i in a word w *breaks* period p iff $w[i] \neq w[i+p]$. Hence a *hp-run* α of a word w is *left-periodic* iff $first(\alpha) - 1$ does not break $subperiod(\alpha)$. In other words the subperiod of $PerPart(\alpha)$ continues to the left.

Example. In Fig. 2 the runs α, β, γ are shown, the first one is not left periodic and the other two are. The position $center(\beta) - 1 = center(\gamma) - 1 = 20$ breaks subperiod 3. The periodic part of β is a prefix of a periodic part of γ.

Lemma 11. *Assume two neighbors α, β with the same subperiod are left-periodic. Then, $center(\alpha) = center(\beta)$.*

Proof. We first prove that positions $center(\alpha) - 1, center(\beta) - 1$ break $subperiod(\alpha)$, see Fig. 9. The proof is by contradiction. If it is not true then one of these runs can be extended one position to the left. This contradicts the definition of the run as a left non-extendible segment. The positions $center(\alpha)$ and $center(\beta)$ are positions in the same h-periodic segment λ, see Fig. 9. They should be equal to the first position of this segment, because the next position to the left breaks the period. Hence, they should be the same position, consequently $center(\alpha) = center(\beta)$. □

Lemma 12. *If α, β are two hp-runs of a string w and satisfy for a given $p > 1$ the inequality*

$$dist(\alpha, \beta) < p \text{ and } \alpha, \beta \in \mathcal{R}(p)$$

then $subperiod(\alpha) = subperiod(\beta)$.

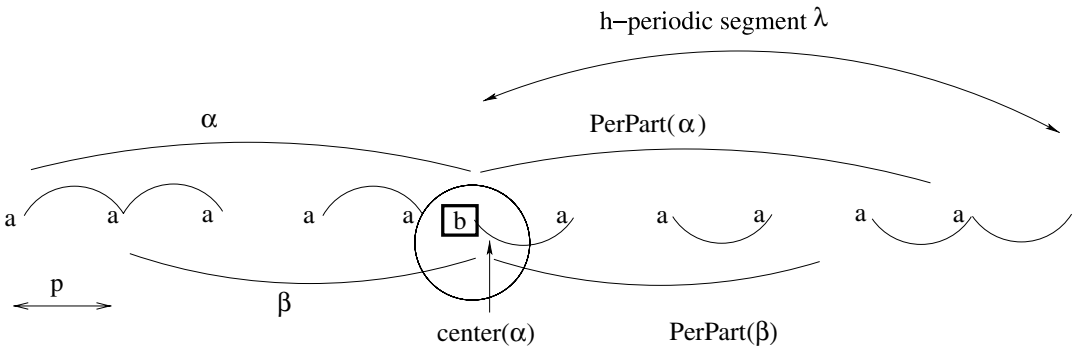

Fig. 9. Two left-periodic runs. The position $center(\alpha)-1=center(\beta)-1$ breaking subperiod p is placed in a small square. $subperiod(\alpha)=subperiod(\beta)=p$, $center(\alpha)=center(\beta)$. The second occurrences of periodic parts of α and β start at the same position $center(\alpha)$, consequently, $PerPart(\beta)$ is a prefix of $PerPart(\alpha)$.

Proof. Assume that $first(\alpha) \leqslant first(\beta)$.
If $dist(\alpha,\beta) < p$ and $p \leqslant subperiod(\alpha), subperiod(\beta) < 2p$ then periodic parts of hp-runs α,β have an overlap of size at least $subperiod(\alpha)+subperiod(\beta)$. Then, due to the periodicity lemma the periodic parts $PerPart(\alpha)$, $PerPart(\beta)$ have the same minimal period. Consequently, $subperiod(\alpha)=subperiod(\beta)$. □

6.1. The proof of point B of lemma 2.1.

Due to Lemma 12 the hp-runs Lemma is reduced to a slightly weaker statement:

For a given $p>1$ there are at most two occurrences of hp-runs with subperiod p in any interval of length p.

The proof of this fact is by contradiction. Assume we have three distinct hp-runs $\alpha_1 \prec \alpha_2 \prec \alpha_3$ with the same subperiod p such that $dist(\alpha_i,\alpha_j) \leqslant p$ for $1 \leqslant i,j \leqslant 3$. Then, all of them are neighbors. We show that $\alpha_2=\alpha_3$. Both α_2,α_3 should be left-periodic since their subperiods extend to the left at least to $first(\alpha_1)$.

Therefore, α_2,α_3 are hp-runs and they are neighbors. Due to Lemma 11 we have:

$$center(\alpha_2)=center(\alpha_3).$$

Consequently, periodic parts of α_2 and α_3 have occurrences starting at the same position $center(\alpha_2)$. If two words start at a same position then one should be a prefix of another. Consequently, $PerPart(\alpha_3)$ is a prefix of $PerPart(\alpha_2)$. Now, due to Lemma 10, if $\alpha_2 \neq \alpha_3$ then $first(\alpha_3)-first(\alpha_2) \geqslant p$. However, $first(\alpha_3)-first(\alpha_2) < p$. This implies that all of $\alpha_1,\alpha_2,\alpha_3$ cannot be pairwise distinct. This contradicts the assumption and completes the proof of this fact and of Point B of Lemma 2.1.

Acknowledgement

The author thanks an anonymous referee for numerous helpful comments.

References

[1] M. Crochemore, An optimal algorithm for computing the repetitions in a word, Inf. Proc. Lett. 42 (5) (1981) 244–250.
[2] F. Franek, R.J. Simpson, W.F. Smyth, The maximum number of runs in a string, in: M. Miller, K. Park (Eds.) Proceeding of 14th Australian Workshop on Combinatorial Algorithms, (2003), 26–35.
[3] M. Crochemore, Recherche linéaire d'un carré dans un mot, Comptes Rendus Acad. Sci. Paris Sér. I Math 296 (1983) 781–784.
[4] M. Crochemore, Transducers and repetitions, Theoret. Comput. Sci. 45 (1) (1986) 63–86.
[5] M. Crochemore, W. Rytter, Jewels of stringology: text algorithms, World Scientific (2003).
[6] Kangmin Fan, William F. Smyth, R.J. Simpson, A new periodicity lemma, CPM (2005) 257–265.
[7] F. Franek, A. Karaman, W.F. Smyth, Repetitions in Sturmian strings, TCS 249-2 (2000) 289–303.
[8] C. Iliopoulos, D. Moore, W.F. Smyth, A characterization of the squares in a Fibonacci string, TCS 172 (1997) 281–291.

[9] R. Kolpakov, G. Kucherov, On maximal repetitions in words, J. Discr. Algorithms 1 (2000) 159–186.
[10] R. Kolpakov, G. Kucherov, Finding maximal repetitions in a word in linear time, FOCS 1 (1999) 596–604.
[11] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, 2003.
[12] M.G. Main, R.J. Lorentz, An $O(n \log n)$ algorithm for finding all repetitions in a string, J. Algorithms 5 (1984) 422–432.
[13] W. Rytter, The structure of subword graphs and suffix trees of fibonacci words, in colloquium on implementation and application of automata, TCS 363–2 (2006) 211–223.
[14] W. Rytter, The number of runs in a string: improved analysis of the linear upper bound, STACS 2006, LNCS 3884.
[15] W.F. Smyth, Computing Patterns in Strings, Addison-Wesley, 2003.
[16] W.F. Smyth, Repetitive perhaps, but certainly not boring, TCS 249-2 (2000) 343–355.