The Number of Runs in a String: Improved Analysis of the Linear Upper Bound*

Wojciech Rytter

Instytut Informatyki, Uniwersytet Warszawski,
Banacha 2, 02–097, Warszawa, Poland
Department of Computer Science, New Jersey Institute of Technology
rytter@mimuw.edu.pl.

Abstract. A *run* (or a *maximal repetition*) in a string is an inclusion-maximal periodic segment in a string. Let $\rho(n)$ be the maximal number of runs in a string of length n. It has been shown in [8] that $\rho(n) = O(n)$, the proof was very complicated and the constant coefficient in $O(n)$ has not been given explicitly. We 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 $\rho(n) \leq 5\ n$. Our proof is inspired by the results of [4], where the role of new periodicity lemmas has been emphasized.

1 Introduction

Periodicities in strings were extensively studied and are important both in theory and practice (combinatorics of words, pattern-matching, computational biology). The set of all runs in a string corresponds to the structure of its repetitions. Initial interest was mostly in repetitions of the type xx (so called *squares*), [1, 10]. 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 [8, 7]. 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 [12, 13, 2]. The lower bound of approximately $0.927\ n$ has been given in [2]. The exact number of runs has been considered for special strings: *Fibonacci words* and (more generally) *Sturmian words*, [6, 5, 11]. 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.

Let $period(w)$ denote the size of the smallest period of w. We say that a word w is *periodic* iff $period(w) \leq \frac{|w|}{2}$.

* Research supported by the grants 4T11C04425 and CCR-0313219.

B. Durand and W. Thomas (Eds.): STACS 2006, LNCS 3884, pp. 184–195, 2006.
© Springer-Verlag Berlin Heidelberg 2006

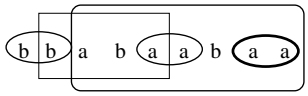

Fig. 1. $RUNS(\ b\ b\ a\ b\ a\ a\ b\ a\ a\ ) = \{[1,2],[2,5],[3,9],[5,6],[8,9]\}$

A *run* in a string w is an inclusion-maximal interval $\alpha = [i...j]$ such that the substring $w[i...j] = w[i]w[i+1]...w[j]$ is periodic. Denote by $RUNS(w)$ the set of runs of w. For example we have 5 runs in an example string in Figure 1.
Denote: $\rho(n) = \max\{|RUNS(w)| : |w| = n\}$.

The most interesting conjecture about $\rho(n)$ is: $\rho(n) < n$.

We make a small step towards proving validity of this conjecture and show that $\rho(n) \leq 5\ n$. The proof of linear upper bound in [8] does not give any explicit constant coefficient at all.

The value of the run $\alpha = [i...j]$ is $val(\alpha) = w[i...j]$. 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.

Each value of the run α is a string $x^k y = w[i...j]$, where $|x| = period(\alpha) \geq 1$, $k \geq 2$ 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 $SquarePart(\alpha) = [i \dots i + 2\, period(\alpha) - 1]$.

We also introduce terminology for the starting position of the second occurrence of periodic part: $center(\alpha) = i + |x|$.

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

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

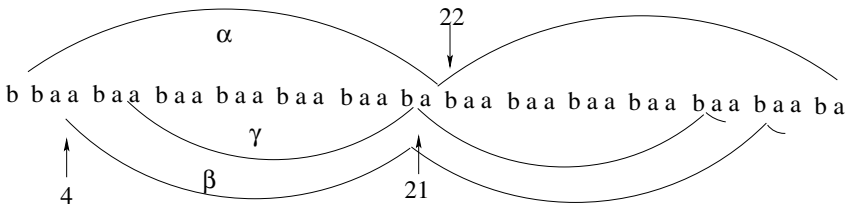

Fig. 2. Example of three highly periodic runs $\alpha \prec \beta \prec \gamma$ with subperiod 3. The runs β, γ are left-periodic (the subperiod 3 continues to the left), α is not. The runs α, β (as well as β, γ) are "neighbors" in sense of Lemma 1. The occurrences (starting positions) of very large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

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

$$\textbf{subperiod}(\alpha) \;=\; period(PerPart(\alpha)).$$

In Figure 2 we have: $subperiod(\alpha) = subperiod(\beta) = subperiod(\gamma) = 3$.

We say that a word w is **highly periodic** (*h-periodic*) if $period(w) \leq \frac{|w|}{4}$. A run 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 Figure 2 α, β, γ are a highly periodic runs.

Denote $\boldsymbol{\Delta} = \frac{\mathbf{5}}{\mathbf{4}}$. We say that two different runs α, β are **neighbors** iff there is a positive number η such that:

$$|first(\alpha) - first(\beta)| \leq \frac{1}{4}\eta \;\text{ and }\; \eta \leq period(\alpha), period(\beta) \leq \Delta\, \eta$$

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

$$period(\alpha), period(\beta) \geq 4\, |first(\alpha) - first(\beta)|.$$

It is *"intuitively obvious"* that if we have many neighbors gathered together then such situation forces one of them to be highly periodic. The tedious proof of the following key-lemma is given in Section 3.

Lemma 1 [The Three-Neighbors]. *Lemma] If we have three distinct runs which are pairwise neighbors with the same number η then at least one of them is h-periodic.*

We cannot replace Three-Neighbors Lemma with *Two-Neighbors Lemma*, see Figure 3.

We show that *hp-runs* are also *sparse* in a certain sense. Another tedious proof of the following lemma is given in Section 4. Figure 2 shows that "two" cannot be replaced by "single", the runs α, β have subperiod 3 and start in the interval $[2 \ldots 4]$ of size 3.

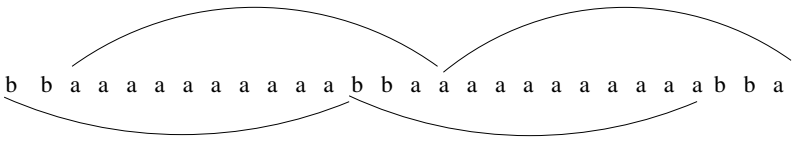

Fig. 3. Two weakly-periodic runs which are neighbors

Lemma 2 [HP-Runs Lemma]. *For a given $p > 1$ there are at most two occurrences of hp-runs with subperiod p in any interval of length p.*

2 Estimating the Number $\rho(n)$

The analysis is based on the *sparsity* properties of *hp-runs* and *wp-runs* expressed by Lemmas 1 and 2.

Denote by $\mathbf{WP}(n,k)$ the maximal number of wp-runs α in a string of length n with $period(\alpha) \geq k$.

Let $\mathbf{HP}(n)$ be the maximal number of all hp-runs in a string of length n. It can be shown that $HP(n) \geq \frac{1}{3}n - c_0$, where c_0 is a constant (take $w = (ab)^m b(ab)^m b(ab)^m$). However we are interested in the upper bound.

Let $\rho(n,k)$ be the maximal number of all runs α with $period(\alpha) \leq k$, in a string of length n. We separately estimate the numbers $WP(n,k)$, $HP(n)$, $\rho(n,k)$.

2.1 Estimating the Number of Weakly Periodic Runs

We group wp-runs into groups of potential neighbors. Denote

$$\mathcal{G}(k) \;=\; \{\alpha \;:\; \alpha \text{ is a weakly periodic run of } w,\ \Delta^k \leq period(\alpha) < \Delta^{k+1}\};$$

Lemma 3. $WP(n, \lceil \Delta^r \rceil) \leq 40\Delta^{-r} \times n$.

Proof. Let w be a string of length n. If $\alpha, \beta \in \mathcal{G}(k)$ for the same k, and $|first(\alpha) - first(\beta)| \leq \Delta^k/4$ then α, β are neighbors with $\eta = \Delta^k$.

Now Lemma 1 can be reformulated as follows: $|\mathcal{G}(k)| \leq 2 \cdot (1/(\Delta^k \cdot \frac{1}{4}) \cdot n = 8\Delta^{-k} \cdot n$.

The last inequality follows directly from Lemma 1, which implies that there are at most two elements of $\mathcal{G}(k)$ in any interval of size $\frac{1}{4}\Delta^k$.

Consequently we have

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

2.2 Estimating the Number of Highly Periodic Runs

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

Lemma 4. *If* $p \geq 2$ *then* $hp(n,p) \leq \frac{2}{p}\, n$.

Proof. It is easy to see the following claim (using the periodicity lemma).

Claim. If α, β are two hp-runs which satisfy
$|first(\alpha) - first(\beta)| < p$ and $p \leq subperiod(\alpha), subperiod(\beta) \leq 2p$,
then $subperiod(\alpha) = subperiod(\beta)$.

It follows from the claim and Lemma 2 that for any interval of length p there are at most two hp-runs occurring in this interval and having subperiods in $[p \dots 2p]$, since such hp-runs should have the same subperiod $p' \geq p$. Therefore there are at most $\frac{2}{p'}\ n \leq \frac{2}{p}\ n$ hp-runs with subperiods in $[p \dots 2p]$. This completes the proof.

Lemma 5. *$HP(n) \leq 1.75\ n$.*

Proof. Observe that there are no hp-runs with subperiod 1.
According to Lemma 4 we have:

$$HP(n) \leq hp(n,2) + hp(n,5) + hp(n,11) + hp(n,23) + hp(n,47) + hp(n,95) + \dots$$

$$= 2\ n\ \times (\frac{1}{2} + \frac{1}{5} + \frac{1}{11} + \frac{1}{23} + \frac{1}{47} + \dots) \times n\ =\ 2\ n\ \times\ \sum_{k=1}^{\infty}\ \frac{1}{p_k},$$

where $p_k = 2^k + 2^{k-1} - 1$. A rough estimation gives:

$$2 \times \sum_{k=1}^{\infty}\ \frac{1}{p_k}\ <\ 1.75$$

Hence $HP(n) \leq 1.75\ n$.

2.3 The Runs with Periods Bounded by a Constant

We estimate the number of runs with small periods in a rather naive way.

Lemma 6. *For any given $k \geq 1$ there are at most $\frac{1}{k+1}\ n$ runs with $period(\alpha) = k$ or $period(\alpha) = 2k$.*

Proof. We omit the proof of the following simple fact.

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

If $period(\alpha) = k$ and $period(\beta) = 2k$ 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 hp-run γ with period k or $2k$ is at least $2k+1$. Then two consecutive distances give together $(first(\beta) - dirst(\alpha) + (first(\gamma) - first(\beta))\ \geq\ 2k+2$, 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;
    remove k and $2k$ from Ψ;
    insert k into Φ;

Define the set $\Phi(p) = \{k \in \Phi : k \leq p\}$. 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 $p \geq 1$ define the numbers:

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

The next lemma follows directely from Lemma 6 and from the structure of the set Φ.

Lemma 7. $\rho(n, p) \leq \mathcal{H}(p) \times n$.

2.4 Estimating the Number of all Runs

Our main result is a *concrete* constant coefficient in $O(n)$ notation for $\rho(n)$.

Theorem 1. $\rho(n) \leq 5\ n$.

Proof. Obviously, for each $r \geq 1$ we have:

$$\rho(n) \leq HP(n) + WP(n, \lceil \Delta^r \rceil) + \rho(n, \lfloor \Delta^r \rfloor)$$

$$\leq (1.75 + 40\ \Delta^{-r} + \mathcal{H}(\lceil \Delta^r \rceil) \times n.$$

If we choose $r = 20$, then

$$\lfloor \Delta^{20} \rfloor = 86, \quad \mathcal{H}(86) \leq 2.77, \quad 40\Delta^{-20} \leq 0.4612.$$

Due to Lemmas 5,6,7 we have:

$$\rho(n) \leq (1.75 + \mathcal{H}(86) + 40\Delta^{-20}) \times n \leq$$

$$(1.75 + 2.77 + 0.4612) \times n < 5\ n.$$

This completes the proof of the main result.

3 The Proof of Lemma 1

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 Figure 5). More formally:

$\alpha \sqsupset \beta$ iff $SquarePart(\beta)$ is contained in $SquarePart(\alpha)$ 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 highly periodic.*
(b) *If $\alpha \prec\prec \beta$ are distinct neighbors then the prefix of β of size $period(\alpha)-\delta$ has a period $|q-p|$, where $\delta = first(\beta)-first(\alpha)$ and $p = period(\alpha),\ q = period(\beta)$.*

Proof. **Point (a).** We refer the reader to Figure 4, 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 $period(\alpha) - period(\beta)$.

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

$$period(\alpha) - period(\beta) \leq \frac{1}{4}\eta \ \text{ and } \ |PerPart(\beta)| \geq \eta\ ,$$

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

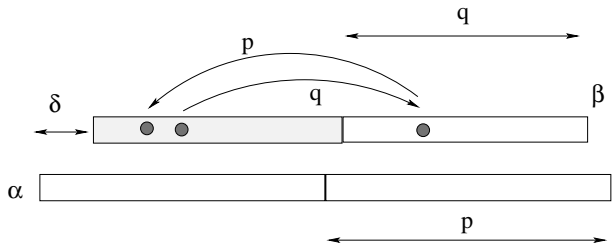

Fig. 4. Two neighbors with $\alpha \sqsupset \beta$, a case $center(\beta) > center(\alpha)$. The *square part* of β is contained in the *square part* of α. The periodic part of β is h-periodic, so it should have a period $p - q$, where $p = period(\alpha),\ q = period(\beta)$.

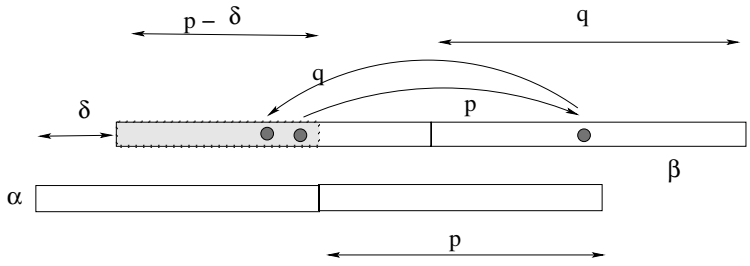

Fig. 5. Two neighbors with $\alpha \prec\prec \beta$, the case $p < q$. The shaded part has the period $|q - p|$, where $p = period(\alpha),\ q = period(\beta)$.

Point (b). We refer to Figure 5, 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.

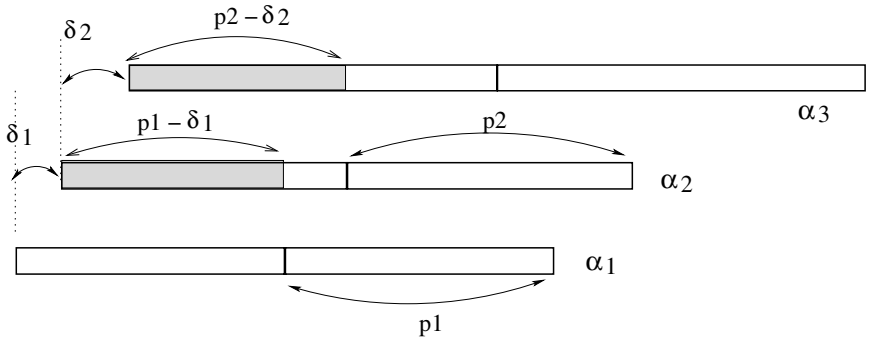

Fig. 6. The Three-Neighbors Lemma, a situation when $\alpha_1 \prec\prec \alpha_2 \prec\prec \alpha_3$. α_2 should be h-periodic, since both its large suffix and large prefix have small periods.

The Proof of the Three-Neighbors Lemma
Assume we have 3 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 "middle" run α_2 has a suffix $\gamma 2$ of size $p_2 - \delta_2$ with a period $|p3 - p2|$ and a prefix $\gamma 1$ of size $p1 - \delta 1$ with a period $|p2 - p1|$, see Figure 6.

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

$$\delta_1 + \delta_2 \leq \tfrac{1}{4}\eta, \;\; p1 \geq \eta, \text{ and } \;\; |\gamma 1 \cup \gamma 2| = p_2.$$

Hence:

$$|\gamma_1 \cap \gamma 2| \;\geq\; (p_2 - \delta_2) + (p1 - \delta 1) - p2 \;=\; p1 - \delta 1 - \delta 2 \;\geq\; \frac{3}{4}\eta$$

We have $|p3 - p2|, |p2 - p1| \leq \frac{1}{4}\eta$, hence $period(\gamma 1), period(\gamma 2) \leq \frac{1}{4}\eta$. Due to the periodicity lemma $\gamma_1 \cap \gamma 2$ has a period which divides periods of $\gamma 1$ and $\gamma 2$, and the whole $\alpha_2 = \gamma_1 \cup \gamma 2$ has a period of size not larger than $\frac{1}{4}\eta$. Consequently, the run α_2 is h-periodic. This completes the proof of our key lemma.

4 The Proof of Lemma 2

The proof is based on the following simple lemma.

Lemma 9. *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* $|first(\alpha) - first(\beta)| \geq p$.

Proof. If $|first(\alpha) - first(\beta)| < p$ then, due to *periodicity lemma* [9, 3, 12], the periodic part of one of the runs would have subperiod smaller than p, which contradicts the assumption that p is the smallest subperiod.

We say that a hp-run $\alpha = [i \dots j]$ of a string w is **left-periodic** iff $w[i-1] = w[i-1+subperiod(\alpha)]$. The runs β, γ in Figure 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 Figure 2 the runs α, β, γ are shown, the first one is not left periodic and the other two are. The position $center(\beta) - 1 = center(\gamma) - 1 = 21$ breaks subperiod 3. The periodic part of β is a prefix of a periodic part of γ.

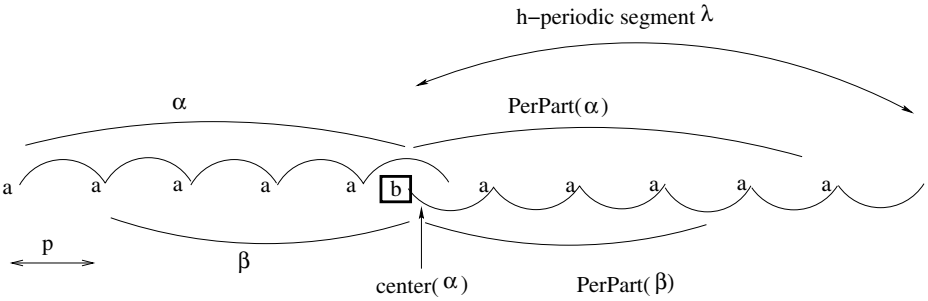

Fig. 7. Two left-periodic runs. The position $center(\alpha) - 1 = center(\beta) - 1$ breaking subperiod p is placed in a small square. $subperiod(\alpha) = subperiod(\beta) = p$, $center(\alpha) = center(\beta)$. The second occurrences of periodic parts of α and β start at the same position $center(\alpha)$, consequently $PerPart(\beta)$ is a prefix of $PerPart(\alpha)$.

Lemma 10. *Assume two neighbors α, β are left-periodic and h-periodic. Then $center(\alpha) = center(\beta)$.*

Proof. We first prove that positions $center(\alpha) - 1, center(\beta) - 1$ break $subperiod(\alpha)$, see Figure 7. 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 Figure 7. 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 $center(\alpha) = center(\beta)$.

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

Proof. The proof is by contradiction. Assume we have three distinct highly periodic runs $\alpha_1 \prec \alpha_2 \prec \alpha_3$ with the same subperiod p such that $|first(\alpha_i) - first(\alpha_j)| \leq p$ for $1 \leq i, j \leq 3$. 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 the runs α_2, α_3 are h-periodic and they are neighbors. Due to Lemma 10 $center(\alpha_2) = center(\alpha_3)$. 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 9, if $\alpha_2 \neq \alpha_3$ then $first(\alpha_3) - first(\alpha_2) \geq p$. 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.

5 The Sum of Exponents of Periodicities

We define the *exponent of periodicity* of a run α as $exp(\alpha) = |\alpha|/period(\alpha)$.

The linear bound on $\rho(n)$ gives, almost automatically, a linear upper bound on the sum of exponents of periodicities. The run α is called a *long* run iff $exp(\alpha) \geq 4$. Denote by $Exp(w)$ the sum of exponents of periodicity of all runs of w, and by *L-Exp*(w) the sum of exponents of all long runs of w.

Let $\mu(n)$ be maximum $Exp(w)$ and $\mu(n, 4)$ be maximum *L-Exp*(w) of a string w of length n. Denote by $\gamma(n)$ the maximum number of long runs in a string of size n.

Lemma 11 **(a)** $\mu(n, 4) \leq 5\, n$; **(b)** $\gamma(n) \leq 1.25\, n$; **(c)** $\mu(n) \leq \mu(n, 4) + 4\, \rho(n)$.

Proof. Denote

$$\mathcal{G}'(k) = \{\alpha : 2^k \leq period(\alpha) < 2^{k+1},\ exp(\alpha) \geq 4\}$$

If $\alpha = [i...j]$ then denote $\Gamma(\alpha) = [i + 3\, period(\alpha) - 1 \ \ldots \ j]$.

Claim. If $\alpha \neq \beta$ are in a same $\mathcal{G}'(k)$, for some integer k, then $\Gamma(\alpha) \cap \Gamma(\beta) = \emptyset$.

Proof (of the claim). The following inequality follows from the *periodicity lemma*:

$$|\alpha \cap \beta| \leq \min\ \{3\, period(\alpha),\ 3\, period(\beta)\}$$

The claim follows easily from this inequality.
Observe now that $|\Gamma(\alpha)| = (exp(\alpha) - 3)\, period(\alpha)$.

Denote by $\mathcal{L}$ the set of long runs with $period(\alpha) > 1$. In other words $\mathcal{L} = \sum_{k>0} \mathcal{G}'(k)$. Due to the claim and the inequality $period(\alpha) \geq 2^k$ we have:

$$\sum_{\alpha \in\ \mathcal{G}'(k)} (exp(\alpha) - 3)\, period(\alpha) \leq n, \quad \text{hence} \quad \sum_{\alpha \in \mathcal{G}'(k)} (exp(\alpha) - 3) \leq \frac{n}{2^k} \quad \text{and}$$

$$\sum_{\alpha\ \in\ \mathcal{L}} (exp(\alpha) - 3) \ \leq\ n \sum_{k=1}^{\infty} \frac{1}{2^k} \ \leq\ n. \tag{1}$$

We have that $exp(\alpha) - 3 \geq 1$, hence $|\mathcal{L}| \leq n$, and we have at most n long runs with $period(\alpha) > 1$. There are at most $\frac{1}{4}\, n$ long runs with period 1. Altogether we have $\gamma(n) \leq 1.25\, n$. This proves point (b).

We now prove point (a). Due to Equation 1 we have:

$$\sum_{\alpha\, \in\, \mathcal{L}} exp(\alpha) \leq n \;+\; \sum_{\alpha\, \in\, \mathcal{L}} 3 \;\leq\; n + 3\,|\mathcal{L}| \;\leq\; 4n$$

On the other hand all runs with period 1 are pairwise disjoint, so the sum of exponents of these runs is at most n. Hence the total sum of exponents of all long α's is at most $n + 4\, n = 5\, n$. This completes the proof of point (a). Point (c) follows directly from definitions.

6 Final Remarks

We gave an estimation $\rho(n) \leq 5\, n$. The important part of our contribution is also a new approach based on subperiods. The proof is completely different from the one in [8], where the proof was by induction on n. The only complicated parts of our proof are the proofs of Lemma 1 and Lemma 2, which can be viewed as *new periodicity lemmas* of independent interest. The proofs of these lemmas are tedious but the lemmas are intuitively almost obvious. In a certain sense we demystified the whole proof of the linear upper bound for $\rho(n)$. The point (c) of Lemma 11 gives directly linear bound on $\mu(n)$ (the sum of exponents of periodicities of all runs), though the constant coefficient is still not satisfactory. Experimental evidence suggests $\mu(n) \leq 2n$. One should possibly rewrite the whole proof of Theorem 1, proving the linear bound on $\rho(n)$ in terms of $\mu(n)$, to improve the coefficient in the linear bound for $\mu(n)$. However this would hideously obscure the proof of Theorem 1.

References

1. *M. Crochemore*, An optimal algorithm for computing the repetitions in a word, Inf. Proc. Letters 42:5(1981) 244-250
2. *F. Franek, R.J.Simpson, W.F.Smyth*, The maximum number of runs in a string, Proc. 14-th Australian Workshop on Combinatorial Algorithms, M.Miller, K. Park (editors) (2003) 26-35
3. *M. Crochemore, W.Rytter*, Jewels of stringology: text algorithms, World Scientific 2003
4. *Kangmin Fan, William F. Smyth, R. J. Simpson*: A New Periodicity Lemma. CPM 2005: 257-265
5. *F. Franek, A. Karaman, W.F.Smyth*, Repetitions in Sturmian strings, TCS 249-2 (2000) 289-303
6. *C. Iliopoulos, D. Moore, W.F.Smyth*, A characterization of the squares in a Fibonacci string, TCS 172 (1997) 281-291
7. *R.Kolpakov, G.Kucherov*, On maximal repetitions in words, Journal of Discr. Algorithms 1 (2000) 159-186

8. *R.Kolpakov, G.Kucherov*, Finding maximal repetitions in a word in linear time, FOCS (1999) 596-604
9. *Lothaire*, Algebraic combinatorics on words, Cambridge University Press
10. *M.G.Main, R.J.Lorentz*, An $O(n \log n)$ algorithm for finding all repetitions in a string, Journal of Algorithms 5 (1984) 422-432
11. *W.Rytter*, The structure of subword graphs and suffix trees of Fibonacci words, in Colloquium on Implementation and Application of Automata, CIAA (2005)
12. *W.F.Smyth*, Computing patterns in strings, Addison-Wesley (2003)
13. *W.F.Smyth*, Repetitive perhaps, but certainly not boring, TCS 249-2 (2000) 343-355.