A New Periodicity Lemma*

Kangmin Fan[1], W. F. Smyth[1,2], and R. J. Simpson[3]

[1] Algorithms Research Group, Department of Computing & Software
McMaster University, Hamilton, Ontario, Canada L8S 4K1
fank@mcmaster.ca
www.cas.mcmaster.ca/cas/research/groups.shtml

[2] Department of Computing, Curtin University, GPO Box U1987
Perth WA 6845, Australia
smyth@computing.edu.au

[3] Department of Mathematics & Statistics, Curtin University
GPO Box U1987, Perth WA 6845, Australia
simpson@math.curtin.edu.au

April 11, 2005

Abstract. Given a string $\boldsymbol{x} = \boldsymbol{x}[1..n]$, a ***repetition*** of period p in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r = \boldsymbol{x}[i..i{+}rp{-}1]$, $p = |\boldsymbol{u}|$, $r \geq 2$, where neither $\boldsymbol{u} = \boldsymbol{x}[i..i{+}p{-}1]$ nor $\boldsymbol{x}[i..i{+}(r{+}1)p{-}1]$ is a repetition. The maximum number of repetitions in any string $\boldsymbol{x}$ is well known to be $\Theta(n \log n)$. A ***run*** or ***maximal periodicity*** of period p in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r\boldsymbol{t} = \boldsymbol{x}[i..i+rp+|\boldsymbol{t}|-1]$ of $\boldsymbol{x}$, where $\boldsymbol{u}^r$ is a repetition, $\boldsymbol{t}$ a proper prefix of $\boldsymbol{x}$, and no repetition of period p begins at position $i{-}1$ of $\boldsymbol{x}$ or ends at position $i{+}rp{+}|\boldsymbol{t}|$. In 2000 Kolpakov and Kucherov showed that the maximum number $\rho(n)$ of runs in any string $\boldsymbol{x}$ is $O(n)$, but their proof was nonconstructive and provided no specific constant of proportionality. At the same time, they presented experimental data strongly suggesting that $\rho(n) < n$. In this paper, as a first step toward proving this conjecture, we present a periodicity lemma that establishes limitations on the number of squares, and their periods, that can occur over a specified range of positions in $\boldsymbol{x}$. We then apply this result to specify corresponding limitations on the occurrence of runs.

1 Introduction

The study of strings began with an investigation of periodicity properties [15], and periodicity of various kinds still remains today a central theme, important both in theory and practice — for example, in data compression, pattern-matching, computational biology, and many other

* The work of the first and second authors was supported in part by grants from the Natural Sciences & Engineering Research Council of Canada.

areas. In this paper we present results that specify restrictions on the nature and extent of periodic behaviour in strings. Although these results are theoretical, their importance is very much a product of their practical application, as we explain below.

It will be convenient throughout to represent strings in boldface (for example, $\boldsymbol{x} = \boldsymbol{x}[1..n]$) and their lengths in italics (for example, $x = |\boldsymbol{x}|$).

A ***repetition*** in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r = \boldsymbol{x}[i..i+ru-1]$, $r \geq 2$, where neither $\boldsymbol{x}[i..i+u-1]$ nor $\boldsymbol{x}[i..i+(r+1)u-1]$ is a repetition. We call $\boldsymbol{u}$ the ***generator*** and u the ***period*** of the repetition, and we represent it economically by an integer triple (i, u, r). In the early 1980s three quite different $O(x \log x)$ algorithms were published [2, 1, 12] for the computation of all the repetitions in a given string $\boldsymbol{x}$. In a sense these algorithms were all asymptotically optimal, since in [2] it was shown that in fact a Fibonacci string $\boldsymbol{f_n}$ contains $\Theta(f_n \log f_n)$ repetitions.

In [11] a more compact encoding of repetitions was introduced: a ***run*** or ***maximal periodicity*** of period u in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r\boldsymbol{t} = \boldsymbol{x}[i..i+ru+t-1]$ of $\boldsymbol{x}$, where $\boldsymbol{u}^r$ is a repetition, $\boldsymbol{t}$ a proper prefix of $\boldsymbol{u}$, and no repetition of period u begins at position $i-1$ of $\boldsymbol{x}$ or ends at position $i+ru+t$. $\boldsymbol{u}$ is called the ***generator*** of the run, $\boldsymbol{t}$ its ***tail***, and a run is economically represented by a 4-tuple (i, u, r, t). Computing all the runs in $\boldsymbol{x}$ permits all the repetitions in $\boldsymbol{x}$ to be listed in an obvious way. Main [11] showed how to compute all the "leftmost" runs in $\boldsymbol{x}$ in time $\Theta(x)$, provided that the suffix tree [16, 13] and the Lempel-Ziv factorization [9] of $\boldsymbol{x}$ were both available. In [3] it was shown that a suffix tree could be computed in linear time on an ***indexed*** (bounded integer) alphabet; since the LZ factorization is computable in linear time from the suffix tree, this meant that the overall worst-case time requirement of Main's algorithm was $\Theta(x)$ on an indexed alphabet. In [8] Kolpakov & Kucherov took matters a step further by extending Main's algorithm to also compute non-leftmost runs in $\boldsymbol{x}$ in time proportional to their number, and then by showing that the maximum number $\rho(x)$ of runs in any string $\boldsymbol{x}$ was at most

$$k_1 x - k_2 \log_2 x \sqrt{x}, \tag{1}$$

where k_1 and k_2 are positive constants. Thus, at least in principle, all the runs in $\boldsymbol{x}$ could be determined in linear time.

However, there is a problem with (1): the proof is nonconstructive and gives no information about the magnitude of the constants k_1 and k_2. Nevertheless Kolpakov & Kucherov provide convincing experimental evidence that

* $\rho(x) < x$;

* $\rho(x)$ is achieved by a cube-free string $\boldsymbol{x}$ on alphabet $\{a, b\}$;
* $\rho(x+1) \leq \rho(x)+2$.

As far as we know, the only published work that addresses these fundamental questions of periodicity is [5], where an infinite family of strings $\boldsymbol{x}$ is constructed that is conjectured for sufficiently large x to achieve $\rho(x) < x$.

In order to show that in general $\rho(x) < x$, it seems to be necessary to establish restrictions on the number of runs that can occur near a position in $\boldsymbol{x}$ at which one or two runs are already known to occur. Perhaps the most famous theoretical result available for such a purpose is the "periodicity lemma":

Lemma 1 [4] *Let p and q be two periods of $\boldsymbol{x}$, and let $d = \gcd(p, q)$. If $p+q \leq x+d$, then d is also a period of $\boldsymbol{x}$.* □

Unfortunately this lemma provides no special information about runs, or the squares with which runs must begin, and it places no restrictions on the positions at which periodic substrings may occur. To our knowledge the only result that provides such information is the following:

Lemma 2 [10, Lemma 8.1.14] *Let $\boldsymbol{u}^2$ be a repetition, and suppose $\boldsymbol{w} \neq \boldsymbol{u}^k$ for any $k \geq 1$. If $\boldsymbol{u}^2$ is a prefix of $\boldsymbol{w}^2$, in turn a proper prefix of $\boldsymbol{v}^2$, then $u+w \leq v$.*

Our main result in this paper is essentially a generalization of this result, that we call a "new periodicity lemma".

2 New Periodicity Lemma

In this section we prove results that establish restrictions on the squares that can occur in the neighbourhood of positions in a string at which one or two squares already appear. We begin with three simple definitions:

Definition 3 *A square $\boldsymbol{u}^2$ is said to be **irreducible** if $\boldsymbol{u}$ is not a repetition.*

Definition 4 *A square $\boldsymbol{u}^2$ is said to be **minimal** if no proper prefix of $\boldsymbol{u}^2$ is a square.*

Lemma 5 *If $\boldsymbol{u}^2$ is minimal, then $\boldsymbol{u}^2$ is irreducible.*

Proof. By Definition 4, no proper prefix of $\boldsymbol{u}^2$ is a square; thus, in particular, no prefix of $\boldsymbol{u}$ is a square. Therefore $\boldsymbol{u}$ cannot be a repetition, and so by Definition 3 $\boldsymbol{u}^2$ is irreducible. □

The existence of a minimal square already imposes significant limitations on the nature of other squares that can exist, as the following result shows:

Lemma 6 *If $\boldsymbol{x} = \boldsymbol{u}^2$ is minimal, then for all integers $k \geq 0$ and $w \in u/2..u-1$:*

(a) *if*
$$k + w \leq u, \quad k+3w \geq 2u, \tag{2}$$
$\boldsymbol{x}[k+1..k+2w]$ *is not a square;*

(b) *if*
$$k + w > u, \quad k+2w \leq 2u, \tag{3}$$
either $\boldsymbol{x}[k+1..k+2w]$ *is not a square* **or** $\boldsymbol{x}[w'+1..w'+u]$ *has period $u-w$, where*
$$w' = (k+w)-u. \quad \square$$

Proof. Suppose that for some pair of integers k and w satisfying either (2) or (3), $\boldsymbol{x}[k+1..k+2w] = \boldsymbol{w}^2$.

First assume that $k = 0$. Then if (2) holds, either $w = u$, a contradiction, or else $w < u$, contradicting the minimality of $\boldsymbol{u}^2$. On the other hand, if (3) holds, then both $w > u$ and $w \leq u$ must hold, again a contradiction. Thus we can assume that $k \geq 1$.

(a) Suppose that (2) holds, let $w' = u-(k+w)$, and consider
$$\widehat{\boldsymbol{w}} = \boldsymbol{x}[1..w-w'] = \boldsymbol{x}[k+w'+1..k+w].$$
Since by (2)
$$(w-w') - (k+w') = k+3w-2u \geq 0,$$
the substring $\boldsymbol{x}[1..k+w]$ has period $k+w'$. Again by (2),
$$(k+w) - 2(k+w') = (k+w) - 2(u-w) \geq 0,$$
so that $\boldsymbol{x}[1..k+w]$ has prefix $\left(\boldsymbol{x}[1..k+w']\right)^2$, contradicting the minimality of $\boldsymbol{u}^2$. Thus in case (a) no such k and w can exist.

(b) Next we suppose that (3) holds, so that $w < u$, hence that $k-w' = u-w > 0$.
Consider
$$\boldsymbol{w} = \boldsymbol{x}[w'+1..w'+w] = \boldsymbol{x}[k+1..u+w'].$$
Since by (3) $w'+w = k+(2w-u) \geq k$, the substring $\boldsymbol{x}[w'+1..w'+u]$ of length u has period $k-w' = u-w$, as required. $\square$

To show that in case (a) of Lemma 6 the assumption that $k+3w \geq 2u$ (as well as the weaker condition $w \geq u/2$) is necessary, consider the example $u = 14, k = 6, w = 5$:

$$\boldsymbol{x} = \boldsymbol{u}^2 = abbaba(babab)(bab\|ab)(babab)ababbab.$$

Here $\boldsymbol{w} = babab$ and $\boldsymbol{w}^3$ is a substring of $\boldsymbol{x}$.

To show that in case (b) of Lemma 6 the substring $\boldsymbol{w}^2$ can in fact exist, consider the example $u = 11, k = 4, w = 8$ with $w' = 1$:

$$\boldsymbol{x} = \boldsymbol{u}^2 = babc(abcabca\|b)(abcabcab)ca.$$

The substring $\boldsymbol{x}[2..12] = (abc)^3ab$ has period $u - w = 3$.

We turn now to the situation in which a minimal square and an irreducible square occur at the same position. We first prove two basic lemmas that describe the relationship between minimality and irreducibility, then go on to prove our main result.

Lemma 7 *If $\boldsymbol{v}^2$ is irreducible with minimal proper prefix $\boldsymbol{u}^2$, then*

$$v > \max\{u+1, 3u/2\}.$$

Proof. Observe that $1 \leq u < v$, and observe further that $u+1 \geq 3u/2$ if and only if $u \leq 2$.

For $u = 1$, $\boldsymbol{u}^2 = \lambda^2$ for some letter λ and the shortest irreducible square $\boldsymbol{v}^2 = (\lambda^2\mu)^2$ for some letter $\mu \neq \lambda$. Thus for $u = 1$, $v \geq 3 > u+1$, as required.

For $u = 2$, since $\boldsymbol{u}^2$ is minimal, $\boldsymbol{u}^2 = (\lambda\mu)^2$ and the shortest irreducible square $\boldsymbol{v}^2 = (\lambda\mu\lambda\mu\nu)^2$ for some letter ν. Thus for $u = 2$, $v \geq 5 > u+1$, as required.

Suppose therefore that $u \geq 3$, and suppose further, without loss of generality, that $v < 2u$. Then

$$\boldsymbol{v} = \boldsymbol{u}\boldsymbol{u}[1..v-u] = \boldsymbol{u}[v-u+1..u]\boldsymbol{v}[2u-v+1..v],$$

where $\boldsymbol{y} = \boldsymbol{u}[1..v-u]$ of length $v-u$ is a prefix of $\boldsymbol{u}$, hence of $\boldsymbol{v}$, and $\boldsymbol{z} = \boldsymbol{u}[v-u+1..u]$ of length $2u-v$ is a prefix of $\boldsymbol{v}$, hence of $\boldsymbol{u}$. If now we assume $2v \leq 3u$, it follows that $v-u \leq 2u-v$, so that $\boldsymbol{y}$ is also a prefix of $\boldsymbol{z}$. Thus $\boldsymbol{u}$ has prefix $\boldsymbol{y}^2$ and so $\boldsymbol{u}^2$ cannot be minimal, a contradiction. We conclude that $2v > 3u$, as required. $\Box$

Lemma 8 *If $\boldsymbol{x} = \boldsymbol{v}^2$ is irreducible with minimal proper prefix $\boldsymbol{u}^2$, $v < 2u$, then*

$$\boldsymbol{x} = \boldsymbol{u_1}\boldsymbol{u_2}\boldsymbol{u_1}\boldsymbol{u_1}\boldsymbol{u_2}\boldsymbol{u_1}\boldsymbol{u_2}\boldsymbol{u_1}\boldsymbol{u_1}\boldsymbol{u_2},$$

where $u_1 = 2u-v, u_2 = 2v-3u$.

Proof. Since $v < 2u$, $u \geq 3$ by Lemma 7. Since $u_1+u_2 = v-u < u$, $\boldsymbol{u}$ has proper prefix $\boldsymbol{u_1u_2}$. Since $u-(v-u) = 2u-v = u_1$, $\boldsymbol{u} = \boldsymbol{u_1u_2u_1}$. Since $\boldsymbol{v} = \boldsymbol{uu}[1..v-u] = \boldsymbol{uu_1u_2}$, the result follows. $\Box$

For the proof of our main result, the following definitions are helpful. A substring of a given string $\boldsymbol{x}$ is said to be ***internal*** if and only if it is neither a prefix nor a suffix of $\boldsymbol{x}$. And if $\boldsymbol{x} = \boldsymbol{x_1x_2}$, $\boldsymbol{x_2}$ nonempty, then $\boldsymbol{x_2x_1} = R_{x_1}(\boldsymbol{x})$ is said to be the x_1^{th} ***rotation*** of $\boldsymbol{x}$.

We also make use of the following two well-known results::

Lemma 9 [14, p. 76] *Let $\boldsymbol{x}$ be a string of length n and minimum period p, and let $j \in 1..n-1$ be an integer. Then $R_j(\boldsymbol{x}) = \boldsymbol{x}$ if and only if $\boldsymbol{x}$ is a repetition and p divides j.* $\Box$

Lemma 10 [14, p. 76] *Let $\boldsymbol{x}$ be a string of length at least 3. If $\boldsymbol{x}$ is a repetition, then so is every rotation of $\boldsymbol{x}$.* $\Box$

Lemma 11 (NPL) *If $\boldsymbol{x}$ has minimal prefix $\boldsymbol{u}^2$ and irreducible prefix $\boldsymbol{v}^2$, $u < v < 2u$, then for every $w \in u+1..v-1$ and for every $k \in 0..v-u-1$, $\boldsymbol{x}[k+1..k+2w]$ is not a square.*

Proof. Suppose that for $0 \leq k < v–u$ and $u < w < v$, $\boldsymbol{w}^2 = \boldsymbol{x}[k+1..k+2w]$. For $k = 0$, Lemma 2 requires that $u+w \leq v$, hence that $2u < v$, a contradiction. Thus the lemma holds for $k = 0$, and so we assume $k \geq 1$. Making use of the notation of Lemma 8, we consider two main cases:

I $1 \leq k < u_1$ (k small)

Since $k+1 < u_1+1$ and $k+w \geq u+2 = 2u_1+u_2+2$, therefore $\boldsymbol{u_2u_1} = \boldsymbol{x}[u_1+1..2u_1+u_2]$ is an internal substring of $\boldsymbol{w} = \boldsymbol{x}[k+1..k+w]$. Then

$$\boldsymbol{u_2u_1} = \boldsymbol{x}[u_1+1+w..2u_1+u_2+w]$$

is also an internal substring of $\boldsymbol{w} = \boldsymbol{x}[k+1+w..k+2w]$. Further, since

$$u_1+1+w > u_1+1+u = 3u_1+u_2+1,$$
$$2u_1+u_2+w < 2u_1+u_2+v = 5u_1+3u_2,$$

therefore $\boldsymbol{u_2u_1}$ is an internal substring of $\boldsymbol{x}[3u_1+u_2+1..5u_1+3u_2] = (\boldsymbol{u_2u_1})^2$. Thus $\boldsymbol{u_2u_1} = R_j(\boldsymbol{u_2u_1})$ for some $j > 0$, and so by Lemma 9 $\boldsymbol{u_2u_1}$ is a repetition. Then by Lemma 10 $\boldsymbol{u_1u_2}$ is a repetition, contradicting the minimality of $\boldsymbol{u}^2$. Hence $k \notin 1..u_1-1$.

II $u_1 \le k < v-u$ (k large)

This case divides into two subcases:

II.1 $k+w \ge v$ ($k+w$ large)

Observe that now $\boldsymbol{w_1} = \boldsymbol{x}[k{+}1..k{+}w]$ must have prefix $R_{k-u_1}(\boldsymbol{u_2}\boldsymbol{u_1}^2)$, while $\boldsymbol{w_2} = \boldsymbol{x}[k{+}1{+}w..k{+}2w]$ must have prefix $R_j(\boldsymbol{u_2}\boldsymbol{u_1}^2)$ for some j. In order that j should equal $k-u_1$, the beginning of $\boldsymbol{w_2}$ would have to be offset by v from the beginning of $\boldsymbol{w_1}$, implying that $w = v$, an impossibility. Therefore $j \ne k-u_1$.

We conclude that $\boldsymbol{w}$ has a prefix equal to two distinct rotations of $\boldsymbol{u_2}\boldsymbol{u_1}^2$, so that by Lemma 9, $R_j(\boldsymbol{u_2}\boldsymbol{u_1}^2)$ is a repetition. Hence by Lemma 10 so also is $\boldsymbol{u} = \boldsymbol{u_1}\boldsymbol{u_2}\boldsymbol{u_1}$, contradicting the minimality of $\boldsymbol{u}^2$. Thus this subcase is impossible.

II.2 $k+w \le v-1$ ($k+w$ small)

Since $k \ge u_1$ and $w \ge u+1$, we may assume from now on that

$$u+u_1+1 \le k+w \le v-1. \tag{4}$$

Again we must divide into subcases:

II.2.1 $k+2w \ge u+v$ ($k+2w$ large)

Since $w \le v-1$, II.2 implies that

$$k+2w \le 2v-u_1-2,$$

so that $\boldsymbol{w_2} = \boldsymbol{x}[k+1+w..k+2w]$ must have suffix $R_{j_2}(\boldsymbol{u_2}\boldsymbol{u_1}^2)$ for some $j_2 \in 0..u_1+u_2$. At the same time, (4) implies that $\boldsymbol{w_1} = \boldsymbol{x}[k+1..k+w]$ must have suffix $R_{j_1}(\boldsymbol{u_2}\boldsymbol{u_1}^2)$ for some $j_1 \in 1..u_2-1$.

Suppose $j_1 = j_2$. Then since $\boldsymbol{w_1}$ terminates at some position in the second occurrence of $\boldsymbol{u_2}$ in $\boldsymbol{x}$, so also must $\boldsymbol{w_2}$ terminate at the same position in the fourth occurrence of $\boldsymbol{u_2}$ in $\boldsymbol{x}$. But this again requires that $w = v$, contradicting the assumption that $w \le v-1$.

Hence $j_1 \neq j_2$ and so $\boldsymbol{w}$ has a suffix that is two distinct rotations of $\boldsymbol{u_2 u_1}^2$. Again using Lemma 9 and Lemma 10, we argue that $\boldsymbol{u} = \boldsymbol{u_1 u_2 u_1}$ is a repetition, contradicting the minimality of $\boldsymbol{u}^2$. Thus II.2.1 is impossible.

II.2.2 $k{+}2w < u{+}v$ ($k{+}2w$ small)

By (4) we can write

$$\boldsymbol{w_1} = \boldsymbol{u_2}[k{-}u_1{+}1..u_2]\boldsymbol{u_1 u_1 u_2}[1..K],$$

where $k{-}u_1{+}1 \leq K \leq u_2{-}1$. Setting

$$\boldsymbol{z} = \boldsymbol{w}[1..z] = \boldsymbol{u_2}[k{-}u_1{+}1..K],$$

we may write $w = u{+}z$, where

$$z = K{-}(k{-}u_1). \tag{5}$$

Observe that $\boldsymbol{w_1}$ has prefix $\boldsymbol{y} = \boldsymbol{u_2}[k{-}u_1{+}1..u_2]\boldsymbol{u_1}$ of length $y = u{-}k$ that must also be a prefix of

$$\boldsymbol{w_2} = \boldsymbol{x}[u{+}u_1{+}K{+}1..u{+}u_1{+}K{+}w].$$

Thus $\boldsymbol{y}$ has border $\boldsymbol{u_2}[K{+}1..u_2]\boldsymbol{u_1}$ of length $u_1{+}u_2{-}K$ and therefore period

$$(u{-}k) - (u_1{+}u_2{-}K) = z.$$

Hence for some integer $q \geq 1$ and some proper prefix $\boldsymbol{z}^*$ of $\boldsymbol{z}$,

$$\begin{aligned} \boldsymbol{w} = \boldsymbol{w_1} &= \boldsymbol{z}^q \boldsymbol{z}^* \boldsymbol{t} && (6) \\ = \boldsymbol{w_2} &= \boldsymbol{z}^{q-1} \boldsymbol{z}^* \boldsymbol{u_2}[1..z]\boldsymbol{t}, && (7) \end{aligned}$$

where $\boldsymbol{u_2}[1..z] = R_j(\boldsymbol{z})$ for some $j \in 0..z{-}1$, and because of (6),

$$\boldsymbol{t} = \boldsymbol{u_1 u_2}[1..K],$$

while as a consequence of (7),

$$\begin{aligned} \boldsymbol{t} &= \boldsymbol{u_2}[z{+}1..z{+}u_1{+}K], \quad z{+}u_1{+}K \leq u_2; \\ &= \boldsymbol{u_2}[z{+}1..u_2]\boldsymbol{u_1}[1..z{+}u_1{+}K{-}u_2], \quad \text{otherwise.} \end{aligned}$$

Thus $\boldsymbol{t}$ is a border of $\boldsymbol{x_1} = \boldsymbol{x}[v{+}1..k{+}2w] = \boldsymbol{x}[1..k{+}2w{-}v]$, so that $x_1{-}t$ is a period of $\boldsymbol{x_1}$. Since $x_1 = k{+}2w{-}v = u_1{+}z{+}t$, it follows that u_1+z is a period of $\boldsymbol{x_1}$. In other words, $\boldsymbol{x}$ has a prefix of length u_1+z+t of period u_1+z. Since

$$\begin{aligned}(u_1+z+t) - 2(u_1+z) &= (u_1+K) - (u_1+z) \\ &= K-z \\ &= k-u_1, \quad \text{by (5)} \\ &\geq 0, \quad \text{by II}\end{aligned}$$

we conclude that $\boldsymbol{x}$ has prefix $\boldsymbol{s}^2$, where $\boldsymbol{s} = \boldsymbol{u_1u_2}[1..z]$, in contradiction to the minimality of $\boldsymbol{u}^2$. Therefore II.2.2 also is impossible. $\square$

We can state an equivalent of Lemma 11 for runs. Observe first that by definition every run is irreducible. Observe also that if a run of period u and tail $\boldsymbol{t}$ occurs at position i in $\boldsymbol{x}$, no run of the same period can occur at any position $j \in i..i+u+t$. Thus, if we define a ***minimal run*** to be a run of generator $\boldsymbol{u}$ where $\boldsymbol{u}^2$ is a minimal square, we can state

Lemma 12 *Suppose $\boldsymbol{x}$ has a minimal run of period u as prefix and another run of period $v < 2u$ as prefix. Then for every positive integer $k < v-u$ and for every $w \in u..v$, no run of period w occurs at position $k+1$ of $\boldsymbol{x}$.* $\square$

3 Discussion

We have proved two main lemmas (6 and 11) that restrict the periods w of squares that can occur at positions $i+k$ in $\boldsymbol{x}$ when at position i either one (Lemma 6) or two (Lemma 11) squares are known to occur. It seems that, with the exception of [10, Lemma 8.1.14], such properties have not previously been studied. In particular, we hope that with the help of Lemma 11, it will be possible to establish, or at least make progress with, the three conjectures arising out of [8].

The Main/Kolpakov-Kucherov algorithm [11, 8] is the only known linear-time algorithm for computing all the runs in a given string $\boldsymbol{x}$. It is complex and, until recently, depended for its worst-case linear behaviour on the use of Farach's algorithm [3], also complex and not space-efficient, for linear-time computation of suffix trees. Since 2003 two worst-case linear-time suffix array construction algorithms [6, 7] have been available for use in the computation of the LZ factorization, but even after the substitution of suffix arrays for suffix trees in the all-runs algorithm, significant complications remain. We expect that, with a more precise understanding of the periodicity of runs, it will become possible to design simpler algorithms that will compute all the runs in a string in a more direct and more space-efficient manner.

References

1. Alberto Apostolico & Franco P. Preparata, **Optimal off-line detection of repetitions in a string**, *Theoret. Comput. Sci. 22* (1983) 297–315.
2. Maxime Crochemore, **An optimal algorithm for computing the repetitions in a word**, *Inform. Process. Lett. 12–5* (1981) 244–250.
3. Martin Farach, **Optimal suffix tree construction with large alphabets**, *Proc.* 38^{th} *IEEE Symp. Found. Computer Science* (1997) 137–143.
4. N. J. Fine & Herbert S. Wilf, **Uniqueness theorems for periodic functions**, *Proc. Amer. Math. Soc. 16* (1965) 109–114.
5. Frantisek Franek, R. J. Simpson & W. F. Smyth, **The maximum number of runs in a string**, *Proc.* 14^{th} *Australasian Workshop on Combinatorial Algorithms*, Mirka Miller & Kunsoo Park (eds.) (2003) 26–35.
6. Juha Kärkkäinen & Peter Sanders, **Simple linear work suffix array construction**, *Proc.* 30^{th} *Internat. Colloq. Automata, Languages & Programming* (2003) 943–955.
7. Pang Ko & Srinivas Aluru, **Space efficient linear time construction of suffix arrays**, *Proc.* 14^{th} *Annual Symp. Combinatorial Pattern Matching*, R. Baeza-Yates, E. Chávez & M. Crochemore (eds.), LNCS 2676, Springer-Verlag (2003) 200–210.
8. Roman Kolpakov & Gregory Kucherov, **On maximal repetitions in words**, *J. Discrete Algs. 1* (2000) 159–186.
9. Abraham Lempel & Jacob Ziv, **On the complexity of finite sequences**, *IEEE Trans. Information Theory 22* (1976) 75–81.
10. M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press (2002) 504 pp.
11. Michael G. Main, **Detecting leftmost maximal periodicities**, *Discrete Applied Maths. 25* (1989) 145–153.
12. Michael G. Main & Richard J. Lorentz, **An $O(n \log n)$ algorithm for finding all repetitions in a string**, *J. Algs. 5* (1984) 422–432.
13. Edward M. McCreight, **A space-economical suffix tree construction algorithm**, *J. Assoc. Comput. Mach. 32–2* (1976) 262–272.
14. Bill Smyth, *Computing Patterns in Strings*, Pearson Addison-Wesley (2003) 423 pp.
15. Axel Thue, **Über unendliche zeichenreihen**, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana 7* (1906) 1-22.
16. Peter Weiner, **Linear pattern matching algorithms**, *Proc. 14th Annual IEEE Symp. Switching & Automata Theory* (1973) 1–11.