HOW MANY RUNS CAN A STRING CONTAIN?

SIMON J. PUGLISI, JAMIE SIMPSON, AND BILL SMYTH

Abstract. Given a string $\boldsymbol{x} = \boldsymbol{x}[1..n]$, a ***repetition*** of period p in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r = \boldsymbol{x}[i+1..i+rp]$, $p = |\boldsymbol{u}|$, $r \geq 2$, where neither $\boldsymbol{u} = \boldsymbol{x}[i+1..i+p]$ nor $\boldsymbol{x}[i+1..i+(r+1)p+1]$ 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 $\boldsymbol{u}^r\boldsymbol{t} = \boldsymbol{x}[i+1..i+rp+|\boldsymbol{t}|]$ of $\boldsymbol{x}$, where $\boldsymbol{u}^r$ is a repetition, $\boldsymbol{t}$ a proper prefix of $\boldsymbol{u}$, and no repetition of period p begins at position i of $\boldsymbol{x}$ or ends at position $i+rp+|\boldsymbol{t}|+1$.

In 2000 Kolpakov and Kucherov showed that the maximum number $\rho(n)$ of runs in any string $\boldsymbol{x}[1..n]$ is $O(n)$, but their proof was nonconstructive and provided no specific constant of proportionality. At the same time, they presented experimental data to prompt the conjecture: $\rho(n) < n$. Recently, Rytter [10] made a significant step toward proving this conjecture by showing that $\rho(n) < 5n$. In this paper we improve Rytter's approach and press the bound on $\rho(n)$ further, proving $\rho(n) \leq 3.48n$.

1. Introduction

Repetitions and other forms of periodicity have long been considered fundamental characteristics of strings. In fact, the work often cited as having founded stringology [13] is an investigation of the periodicity properties of infinite strings. Today, the detection of repetitions has become of practical interest; for instance, in the field of bioinformatics. Algorithms for this task are now a standard part of any software for whole genome analysis.

A run is a series of overlapping repetitions that all have the same period (we give a formal definition shortly). The idea of computing the repetitions in $\boldsymbol{x} = \boldsymbol{x}[1..n]$ by computing the runs is attractive because the number of runs is linear in string length [7], while the number of repetitions can be $\Theta(n \log n)$ [1]. The only known linear-time algorithm for computing all runs (hence all repetitions) is due to Kolpakov and Kucherov [7]. Unfortunately this algorithm requires significant algorithmic machinery and working memory, and is thus not suitable for very long (for instance, genome-sized) strings. These inadequacies motivate us to improve our theoretical understanding of the nature of runs. We expect that, with a more precise understanding of the way in which these structures occur, it will become possible to design simpler algorithms that will compute runs in a more direct and efficient manner.

The $\Theta(n)$ complexity of Kolpakov and Kucherov's algorithm hinges on a lengthy and technical proof [7] that the maximum number $\rho(n)$ of runs that could exist in any string $\boldsymbol{x}$ is at most

$$k_1 n - k_2 \log_2 n\sqrt{n}, \tag{1.1}$$

where k_1 and k_2 are positive constants.

The proof of (1.1) provides no information about the magnitude of the constants k_1 and k_2. Nevertheless Kolpakov & Kucherov provide experimental evidence to prompt the conjecture [12] that $\rho(n) < n$. Progress toward proving this conjecture has been scant. Franek et al. [4] proved a lower bound $\rho(n) > 0.927n$ over an infinite set of string lengths n corresponding to "run-rich" strings; more recently, Franek and Yang [5] showed that this bound holds for all sufficiently large n. Fan et al. [2] and also Simpson [11] have proved several intricate results that place restrictions on the nature and extent of repetitions that occur in areas of high periodicity within the string. While these results do apply to runs, it is not yet obvious how they can be used to improve the upper bound on $\rho(n)$. The most significant step to date was made recently by Rytter [10], who showed that $\rho(n) \leq 5n$. In our paper we rely heavily on Rytter's ideas and improve the upper bound to $3.48n$.

Throughout this paper we use boldface to denote strings and think of a string as an array; thus $\boldsymbol{x} = \boldsymbol{x}[1..n]$ is a string of length $n = |\boldsymbol{x}|$. Terminology and notation generally follow [12].

A ***repetition*** in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r = \boldsymbol{x}[i{+}1..i{+}rp]$, where $r \geq 2$, $p = |\boldsymbol{u}|$, and neither $\boldsymbol{u} = \boldsymbol{x}[i{+}1..i{+}p]$ nor $\boldsymbol{x}[i{+}1..i{+}(r{+}1)p]$ is a repetition. We call $\boldsymbol{u}$ the ***generator***, p the ***period***, and r the ***exponent*** of the repetition. A repetition can thus be encoded as an integer triple (i, p, r). In order to compute all repetitions efficiently, Main [9] defined a ***run*** or ***maximal periodicity*** in $\boldsymbol{x}$ as a substring $\boldsymbol{u}^r\boldsymbol{t} = \boldsymbol{x}[i{+}1..i{+}rp{+}t]$, where $\boldsymbol{u}^r$ is a repetition, $\boldsymbol{t}$ a proper prefix of $\boldsymbol{u}$, $t = |\boldsymbol{t}|$, and no repetition of period p begins at position i of $\boldsymbol{x}$ or ends at position $i{+}rp{+}t{+}1$. The ***generator***, ***period*** and ***exponent*** of a run are defined as for a repetition, and $\boldsymbol{t}$ is called the ***tail***. Thus a run is economically represented by a 4-tuple (i, p, r, t). Since a run includes $t{+}1$ or p repetitions, respectively, according as $r = 2$ or $r > 2$, it follows that there are at most as many runs as repetitions. Further, by computing all runs we implicitly compute all repetitions.

For a real number $\theta \geq 2$, a θ ***highly periodic run***, henceforth a ***θ-hp run***, is a run in which the generator is itself periodic and has length at least θ times the length of its (minimum) period. We call the period s of the generator the ***subperiod*** of the θ-hp run, and the prefix of the run of length s its ***subgenerator***. Thus a θ-hp run of period p and subperiod s satisfies $\theta \leq p/s$. Rytter uses θ-hp runs with $\theta = 4$ which he calls simply hp-runs. Towards the end of this paprt we will set $\theta = 8$ but for the initial results keep it as an unevaluated parameter.

2. Some Lemmas

A central result about periodicity in strings is the Periodicity Lemma of Fine and Wilf [3].

Lemma 2.1. (The Periodicity Lemma) *Let $\boldsymbol{x}$ be a string having two periods p and q. If $|\boldsymbol{x}| \geq p + q - \gcd(p, q)$ then $\boldsymbol{x}$ also has period $\gcd(p, q)$.*

The next result also applies to strings having two periods, but with string length less than the Fine-Wilf bound.

Lemma 2.2. *Let $\boldsymbol{x}$ be a string having two periods p and q with $q > p$. Then the string's suffix and prefix of length $|\boldsymbol{x}| - p$ both have period $q - p$.*

This is Lemma 8.1.1 of [8] and Lemma 2.1 of [6].

Suppose $\boldsymbol{x} = \boldsymbol{uv}$ for nonempty $\boldsymbol{u}$ and $\boldsymbol{v}$; then $\boldsymbol{vu}$ is called a ***rotation*** of $\boldsymbol{x}$. The following lemma is also required [12, p. 26].

Lemma 2.3. *If $\boldsymbol{x}$ and a rotation of $\boldsymbol{x}$ are equal, then $\boldsymbol{x}$ is a repetition.*

To count the number of runs in a string $\boldsymbol{x}$ we bound separately the number of θ-hp runs and the other runs. Lemma 2.4 shows that θ-hp runs with similarly sized subperiods starting close together must have the same subperiod. Lemma 2.7 bounds the number of θ-hp runs in $\boldsymbol{x}$ which have the same subperiod. Lemmas 2.4 and 2.7 are used in Lemma 2.8 which gives an upper bound on the number of θ-hp runs having subperiod in a certain interval. These results are combined with Lemma 2.9, taken straight from Rytter's paper, to give our main result.

Lemma 2.4. *Suppose that θ-hp runs begin at positions $k_1 + 1$ and $k_2 + 1$ respectively of a string $\boldsymbol{x}$, with periods p_1 and p_2 respectively and subperiods s_1 and s_2 respectively. If $L \le s_i \le U$ for $i = 1, 2$,*

$$(2\theta - 1)L - U \ge k_2 - k_1 \ge 0 \tag{2.1}$$

and

$$(\theta/2 - 1)L \ge U, \tag{2.2}$$

then $s_1 = s_2$.

Proof: We consider three cases.

Case 1. $k_2 - k_1 \le p_1 - s_1 - s_2$. Since $\boldsymbol{x}[k_1 + 1..k_1 + p_1]$ has period s_1 and $\boldsymbol{x}[k_2 + 1..k + p_2]$ has period s_2, their intersection $\boldsymbol{x}[k_2 + 1.. \min(k_1 + p_1, k_2 + p_2)]$ has both periods. We show that this intersection is sufficiently long to apply the Periodicity Lemma. Its length is $\min(p_1 + k_1 - k_2, p_2)$. By the assumption for this case

$$p_1 + k_1 - k_2 \quad \ge s_1 + s_2.$$

Also,

$$\begin{aligned} p_2 &\ge \theta s_2 \\ &\ge (\theta - 1)L + s_2 \\ &> U + s_2, \text{ by (2.2)} \\ &\ge s_1 + s_2. \end{aligned}$$

Thus the length of the intersection is greater than the sum of the subperiods. By the Periodicity Lemma both $\boldsymbol{x}[k_1 + 1..k_1 + p_1]$ and $\boldsymbol{x}[k_2 + 1..k_2 + p_2]$ have period $\gcd(s_1, s_2)$. However we assumed their minimum periods were s_1 and s_2 respectively. To avoid a contradiction we must have $s_1 = s_2$.

Case 2. $p_1 - s_1 - s_2 < k_2 - k_1 \le p_1$. This time we consider the intersection of $\boldsymbol{x}[k_1 + p_1 + 1..k_1 + 2p_1]$ and $\boldsymbol{x}[k_2 + 1..k_2 + p_2]$, which have periods s_1 and s_2 respectively. Their intersection is $\boldsymbol{x}[k_1 + p_1 + 1.. \min(k_1 + 2p_1, k_2 + p_2)]$, which has length $\min(p_1, k_2 - k_1 + p_2 - p_1)$. By similar reasoning to that used in Case 1 we

see that $p_1 > s_1 + s_2$. Also, using (2.1) and the assumption for this case,

$$\begin{aligned} k_2 - k_1 + p_2 - p_1 &> \theta s_2 - s_1 - s_2 \\ &\geq (\theta - 2)s_2 + s_2 - s_1 \\ &\geq (\theta - 2)L + s_2 - s_1 \\ &\geq 2U + s_2 - s_1, \text{ by (2.2)} \\ &\geq s_1 + s_2. \end{aligned}$$

In each case the intersection has sufficient length for the Periodicity Lemma to apply and we get $s_1 = s_2$ as in Case 1.

Case 3. $p_1 \leq k_2 - k_1 \leq (2\theta - 1)L - U$. We again consider the intersection of $\boldsymbol{x}[k_1 + p_1 + 1..k_1 + 2p_1]$ and $\boldsymbol{x}[k_2 + 1..k_2 + p_2]$, which is now $\boldsymbol{x}[k_2 + 1..\min(k_1 + 2p_1, k_2 + p_2)]$ with length $\min(2p_1 - k_2 + k_1, p_2)$. If the minimum is $2p_1 - k_2 + k_1$ then the length is at least

$$\begin{aligned} & 2p_1 - k_2 + k_1 \\ \geq\ & (2\theta - 1)s_1 + s_1 - (2\theta - 1)L + U \\ \geq\ & s_1 + U \\ \geq\ & s_1 + s_2. \end{aligned}$$

If it is p_2 then the length is at least θL which by (2.2) is at least $2L + 2U \geq s_1 + s_2$. As it the other cases we get $s_1 = s_2$. □

Observe that condition (2.2) implies $\theta \geq 4$.

We now tighten Lemma 10 of [10]. Our modifications remove the requirement that the runs α and β discussed in that lemma are *neighbours* (in Rytter's sense) and instead relate their offset from one another to the subperiod they share. To formulate these results, we need the following definition [10]: a θ-hp run starting at position k in $\boldsymbol{x}$ with subperiod s is said to be ***left-periodic*** iff $\boldsymbol{x}[k-1] = \boldsymbol{x}[k-1+s]$.

Lemma 2.5. *Let α and β be left-periodic θ-hp runs beginning at positions $k_\alpha + 1$ and $k_\beta + 1$ respectively of a string $\boldsymbol{x}$, with periods p_α and p_β respectively, both with subperiod s. If $k_\alpha < k_\beta < k_\alpha + 2s$, then $k_\alpha + p_\alpha = k_\beta + p_\beta$.*

Proof: Since α is a run, $\boldsymbol{x}[k_\alpha] \neq \boldsymbol{x}[k_\alpha + p_\alpha]$. However, since α is left-periodic and has period p_α, $\boldsymbol{x}[k_\alpha + p_\alpha + s] = \boldsymbol{x}[k_\alpha + s] = \boldsymbol{x}[k_\alpha]$. We conclude that

$$\boldsymbol{x}[k_\alpha + p_\alpha] \neq \boldsymbol{x}[k_\alpha + p_\alpha + s]. \tag{2.3}$$

Similarly,

$$\boldsymbol{x}[k_\beta + p_\beta] \neq \boldsymbol{x}[k_\beta + p_\beta + s]. \tag{2.4}$$

Let $\boldsymbol{y} = \boldsymbol{x}[k_\alpha + p_\alpha..k_\alpha + p_\alpha + s]$ and $\boldsymbol{z} = \boldsymbol{x}[k_\beta + p_\beta..k_\beta + p_\beta + s]$. Because of (2.3) and (2.4) neither $\boldsymbol{y}$ nor $\boldsymbol{z}$ has period s. We consider four cases.

Suppose $k_\beta + p_\beta \geq k_\alpha + p_\alpha + s$. By hypothesis $k_\beta < k_\alpha + 2s < k_\alpha + p_\alpha$, so that $\boldsymbol{y}$ is a factor of $\boldsymbol{x}[k_\beta + 1..k_\beta + p_\beta]$ of period s. But this is impossible as $\boldsymbol{y}$ does not have period s. We conclude that

$$k_\alpha + p_\alpha + s > k_\beta + p_\beta. \tag{2.5}$$

Now suppose that $k_\alpha+p_\alpha \geq k_\beta+p_\beta+s$. Since $k_\beta+p_\beta \geq k_\alpha+1$, $\boldsymbol{z}$ is a factor of $\boldsymbol{x}[k_\alpha+1..k_\alpha+p_\alpha]$ of period s. This also is impossible and we conclude that

$$k_\beta+p_\beta+s > k_\alpha+p_\alpha. \tag{2.6}$$

Next suppose $k_\beta+p_\beta < k_\alpha+p_\alpha$. By (2.6) $k_\alpha+p_\alpha+s \leq k_\beta+p_\beta+2s < k_\beta+2p_\beta$, so $\boldsymbol{y}$ is a factor of $\boldsymbol{x}[k_\beta+p_\beta+1..k_\beta+2p_\beta]$ of period s. Again this is impossible and we conclude that

$$k_\beta+p_\beta \geq k_\alpha+p_\alpha. \tag{2.7}$$

Finally suppose that $k_\alpha+p_\alpha < k_\beta+p_\beta$. By (2.5) $k_\beta+p_\beta+s < k_\alpha+p_\alpha+2s \leq k_\alpha+2p_\alpha$, so $\boldsymbol{z}$ is a factor of $\boldsymbol{x}[k_\alpha+p_\alpha+1..k_\alpha+2p_\alpha]$, again impossible. It follows that $k_\alpha+p_\alpha = k_\beta+p_\beta$, as required. □

The next result follows easily from Lemma 2.5 above.

Lemma 2.6. *Let α and β be left-periodic θ-hp runs starting at $k_\alpha+1$ and $k_\beta+1$, respectively, both having subperiod s. If $k_\alpha < k_\beta < k_\alpha + 2s$, then $k_\beta = k_\alpha + s$.*

Proof: By Lemma 2.5 the generator of β is a prefix of α and so the two runs have the same subgenerator $\boldsymbol{s} = \boldsymbol{x}[k_\alpha + 1..k_\alpha + s]$. If $k_\beta \neq k_\alpha + s$ then $\boldsymbol{s}$ equals a rotation of itself and so by Lemma 2.3 the subperiod of the runs is smaller than s, a contradiction. Therefore $k_\beta = k_\alpha + s$. □

Remark 1: Let α and β be 2-hp runs of subperiod s starting at positions $k_\alpha+1$ and $k_\beta+1 > k_\alpha+1$ respectively in $\boldsymbol{x}$. Observe that if $k_\beta-k_\alpha \leq s$, β is necessarily left-periodic. At the same time, Lemma 2.6 tells us that if α and β are left-periodic, then $k_\beta–k_\alpha \geq s$. We conclude that while two 2-hp runs of subperiod s may possibly begin at positions less than s apart (an example is given in [10]) if the leftmost of the two is not left-periodic, nevertheless due to left-periodicity a third such run can only begin at distance s or more from the start of the second.

Lemma 2.4 concerned θ-hp runs with different subperiods. The next lemma uses the results on left-periodic runs to consider those with the same subperiod.

Lemma 2.7. *For $\theta \geq 2$, the number of θ-hp runs with subperiod s in a string of length n is less than n/s.*

Proof: In view of Remark 1, we may suppose that $\boldsymbol{x}$ contains $m+1$ θ-hp runs with subperiod s beginning at $\boldsymbol{x}[k+1]$, $\boldsymbol{x}[k+t+1]$, $\boldsymbol{x}[k+t+s+1]$, $\boldsymbol{x}[k+t+2s+1]$,..., $\boldsymbol{x}[k+t+ms+1]$, $0 \leq t \leq s$, and that m is maximal; that is, there is no such run beginning at $\boldsymbol{x}[k+t+(m+1)s+1]$. Suppose the run beginning at $\boldsymbol{x}[k+t+ms+1]$ has period p. Since it is a θ-hp run we have $p \geq \theta s$. If p were greater than or equal to $(\theta+1)s$ we would have a θ-hp run beginning at $\boldsymbol{x}[k+t+(m+1)s+1]$, which we have denied. Therefore $p < (\theta+1)s$.

We show that no such run can begin at any position from $\boldsymbol{x}[k+t+ms+2]$ to $\boldsymbol{x}[k+t+ms+p-s+1]$. Any θ-hp run beginning in $\boldsymbol{x}[k+t+ms+2..k+t+(m+2)s]$ with subperiod s must have a sub-generator which is a rotation of the sub-generator of the other θ-hp runs. This implies the run is left-periodic and we can apply Lemma 2.6. But this lemma would imply that such a run would begin at $\boldsymbol{x}[k+t+(m+1)s+1]$ which is forbidden. If a θ-hp run started in $\boldsymbol{x}[k+t+(m+2)s+1..k+t+p-s+1]$ then its generator would extend at least to $\boldsymbol{x}[k+t+(m+2+\theta)s]$ which is beyond

$\boldsymbol{x}[k+t+(m+1)s+p]$, so the s periodicity would extend to $\boldsymbol{x}[k+t+ms+2\theta s]$ and $\boldsymbol{x}[k+t+ms+1..k+t+ms+2\theta s]$ would have period s, and not be a highly periodic run. This is a contradiction and we conclude that no θ-hp runs with subperiod s begin in this interval. Thus any sequence of θ-hp runs each with subperiod s beginning at positions $\boldsymbol{x}[k+1]$, $\boldsymbol{x}[k+t+1]$, $\boldsymbol{x}[k+t+s+1]$, $\boldsymbol{x}[k+t+2s+1]$,..., $\boldsymbol{x}[k+t+ms+1]$, must be followed by an interval of length at least $(\theta-1)s$ in which no such run begins. That is, we have an interval of length $(m+\theta-1)s$ in which $m+1$ runs begin. It follows that the whole string contains less than n/s θ-hp runs with subperiod s. □

Remark 2: The first two runs described in this proof may start close together, but then the starts of the later pairs are s positions apart, and the final starting position is followed by an interval in which no such run can begin. One might think that the low density at the end would outweigh the high density at the start, and that the lemma could be strengthened. The following example shows that asymptotically this is not the case. The string $\boldsymbol{x} = ((ab)^m a)^l$ has length $l(2m+1)$. For $l \geq 2$, $m \geq 4$, it contains 4-hp runs with subperiod 2 beginning at positions 1 and $\{1+2i+j(2m+1) : 1 \leq i \leq m-4, 0 \leq j \leq l-2\}$. Thus it contains $(l-1)(m-4)+1$ 4-hp runs with subperiod 2. The number of runs per unit length of $\boldsymbol{x}$ is therefore

$$\frac{(l-1)(m-4)+1}{l(2m+1)}.$$

This approaches $1/2$ as l and m become large.

Lemma 2.8. *Let L and U satisfy*

$$(\frac{\theta}{2}-1)L \geq U > L > 0. \tag{2.8}$$

Then the number of θ-hp runs with subperiod in the interval $[L,U]$ is less than n/L.

Proof: If all θ-hp runs with subperiod in the interval $[L,U]$ have the same subperiod s then by Lemma 2.7 we have less than $n/s \leq n/L$ such runs altogether, with average separation between their starting positions greater than s. If the string contains two such runs with unequal subperiods then, by Lemma 2.4 their starting positions are separated by at least $(2\theta-1)L-U$. Using (2.8)

$$\begin{aligned}(2\theta-1)L-U &= 4(\theta/2-1)L+3L-U \\ &\geq 3U+3L.\end{aligned}$$

We conclude that the number of θ-hp runs with subperiod in the interval $[L,U]$ is maximised when they all have the same subperiod, and this is less than n/L. □

Lemma 2.8 will enable us to bound the number of θ-hp runs in a string. We bound the number of other runs using the next two lemmas. Lemma 2.9 will be used to bound the number with smaller periods, and Lemma 2.10 to bound the others.

Let Φ be the set of positive integers exactly divisible by an even power of 2, possibly 2^0. That is, integers of the form $2^i m$ where i is even and m is odd. Thus $\Phi = \{1,3,4,5,7,9,\dots.\}$.) Let

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

The following is Lemma 7 in Rytter's paper [10].

Lemma 2.9. *The number of runs with period p or less in a string of length n is at most $H(p)n$.*

The next lemma, which strengthens Rytter's "Three Neighbours Lemma", shows that three runs with similarly sized periods must include a θ-hp run if they have starting positions sufficiently close together. This will allow us to bound the number of non-θ-hp runs.

Lemma 2.10. *Suppose that a string $\boldsymbol{x}$ contains runs beginning at positions k_1+1, k_2+1 and k_3+1 with periods p_1, p_2 and p_3 respectively and that $k_1 < k_2 < k_3$. Suppose also that L and U are positive numbers such that*

$$L \le p_i \le U < 2L \tag{2.9}$$

for $i \in \{1,2,3\}$. If

$$3L - 2U \ge k_3 - k_1, \tag{2.10}$$

then either the run beginning at k_2+1 or the run beginning at k_3+1 is a θ-hp run with subperiod at most $\gcd(|p_1-p_2|, |p_2-p_3|)$ and

$$\theta \ge \frac{2L}{U-L}. \tag{2.11}$$

Proof: For the sake of contradiction suppose that $p_1 = p_3$. The intersection of the first and third runs is $\boldsymbol{x}[k_3+1..\min(k_1+2p_1, k_3+2p_3)]$ and its length is $\min(2p_1-(k_3-k_1), 2p_3)$. Now,

$$\begin{aligned} 2p_1-(k_3-k_1) &\ge p_1 + L - (3L-2U) \\ &\ge p_1 + 2U - 2L \\ &> p_1. \end{aligned}$$

So the length of the overlap is at least the common period, implying that the whole of $\boldsymbol{x}[k_1+1..k_3+2p_3]$ has period $p_1 = p_3$, contradicting the hypothesis that they are distinct runs. We conclude that $p_1 \ne p_3$. A similar analysis shows that the three periods are pairwise distinct. Note that this requires that $U > L$.

The intersection of the first two runs is $\boldsymbol{x}[k_2+1..\min(k_1+2p_1, k_2+2p_2)]$. This has periods p_1 and p_2. Using Lemma 2.2 we see that

$$\boldsymbol{x}[k_2+1..\min(k_1+2p_1, k_2+2p_2) - \min(p_1,p_2)]$$

has period $|p_1-p_2|$. Since $\min(k_1+2p_1, k_2+2p_2) - \min(p_1,p_2) \ge \min(k_1+p_1, k_2+p_2)$,

$$\boldsymbol{x}[k_2+1..\min(k_1+p_1, k_2+p_2)] \tag{2.12}$$

has period $|p_1-p_2|$. By considering the second and third runs in the same way we find that

$$\boldsymbol{x}[k_3+1..\min(k_2+p_2, k_3+p_3)] \tag{2.13}$$

has period $|p_2-p_3|$. The intersection of the factors in displays (2.12) and (2.13) has both period $|p_1-p_2|$ and $|p_2-p_3|$. We show that the length of this intersection is sufficient to apply the Periodicity Lemma. The length is

$$\min(k_1+p_1, k_2+p_2, k_3+p_3) - k_3.$$

Applying (2.9) and (2.10) we get

$$\begin{aligned}\min(k_1+p_1, k_2+p_2, k_3+p_3) - k_3 &\geq \min(p_1,p_2,p_3) - (k_3-k_1)\\ &\geq L - (3L-2U)\\ &= 2U-2L\\ &\geq |p_1-p_2| + |p_2-p_3|.\end{aligned}$$

Thus the Periodicity Lemma applies and the intersection has period

$$g = \gcd(|p_1-p_2|, |p_2-p_3|).$$

The period g clearly extends to the union of the factors in displays (2.12) and (2.13), and so

$$\boldsymbol{x}[k_2+1..\min(k_2+p_2, k_3+p_3)]$$

has period g. If the minimum is k_2+p_2 then the run beginning at k_2+1 is a θ-hp run with subperiod at most g. If the minimum is k_3+p_3 then, since $k_3 > k_2$, the run beginning at k_3+1 is a θ-hp run with subperiod at most g.

This establishes the first part of the Lemma. For the second part we bound the size of g. It is easy to see that $g \leq U-L$ but we will show that $g \leq (U-L)/2$. Recall that if $a > b$, then $\gcd(a,b) \leq a/2$; thus $g \leq (U-L)/2$ unless $|p_1-p_2| = |p_2-p_3|$. That is, unless $p_1-p_2 = p_2-p_3$ or $p_1-p_2 = p_3-p_2$. The second alternative would imply $p_1 = p_3$ which we noted earlier was impossible. The first alternative would mean $2g = |(p_1-p_2)+(p_2-p_3)| = |p_1-p_3| \leq U-L$ and so again we have $g \leq (U-L)/2$.

Thus the run beginning at $\boldsymbol{x}[k_2+1]$ or the run beginning at $\boldsymbol{x}[k_3+1]$has subperiod g; since by hypothesis such a run has a period of length at least L, it is therefore a θ-hp run with $\theta \geq L/g \geq 2L/(U-L)$, as required. □

Note that in this lemma we have made no assumptions about the relative sizes of k_1+2p_1, k_2+2p_2 and k_3+2p_3, or about the relative sizes of p_1, p_2 and p_3.

3. The Main Result

Now we prove our main theorem. To do this we use θ-hp runs with $\theta = 8$. It will be seen that this and the values used in place of L and U are sufficient for the lemmas to apply.

Theorem 3.1. *The number of runs in a string $\boldsymbol{x}$ of length n is less than* $3.48n$.

Proof: We count separately those runs in $\boldsymbol{x}$ which are 8-hp runs and those which are not.

For those which are, let $L(k)$, integer $k \geq 0$, be the set of 8-hp runs with subperiod in the interval $[2\times 3^k, 2\times 3^{k+1})$. Note that we cannot have an any hp-run with subperiod 1, so every possible 8-hp run is counted in one of these intervals. By Lemma 2.8 $|L(k)| < n/(2\times 3^k)$. The total number of 8-hp runs in the string is then less than

$$\sum_{k=0}^{\infty} |L(k)| < \sum_{k=0}^{\infty} \frac{n}{2\times 3^k} = 0.75n. \tag{3.1}$$

We now consider the other runs. We partition these into two sets. N_1 is the set with periods in the interval $[1, \lfloor (5/4)^{16} \rfloor] = [1, 35]$ and N_2 those with period greater than 35. By Lemma 2.9,

$$|N_1| \leq H(35)n = 2.16540n. \tag{3.2}$$

To bound $|N_2|$ we let $M(k)$ be the set of runs which are not 8-hp and have periods in the interval $[(5/4)^k, (5/4)^{k+1})$. To count such runs we apply Lemma 2.10 with $L = (5/4)^k$ and $U = (5/4)^{k+1}$. If three such runs were to begin in an interval of length $\frac{1}{2}(5/4)^k = 3L - 2U$, one of the runs would be θ-hp with

$$\theta \geq \frac{2(5/4)^k}{(5/4)^{k+1} - (5/4)^k} = 8,$$

a contradiction. Therefore we have at most two such runs in such an interval and

$$|M(k)| \leq \frac{2n}{\frac{1}{2}(5/4)^k} = 4n(\frac{4}{5})^k.$$

Thus

$$|N_2| = \sum_{k=16}^{\infty} |M(k)| < 4n \sum_{k=16}^{\infty} (\frac{4}{5})^k = 0.56295n. \tag{3.3}$$

We obtain the bound on the total number of runs by summing the bounds in (3.1), (3.2) and (3.3). □

Remark 3: The experiments of Kolpakov and Kucherov [7] suggest that in fact there are *no* θ-hp runs in run-maximal strings. If indeed the conjecture $\rho(n) < n$ is correct, it appears therefore that quite different methods will be required in order to prove it.

Acknowledgements: We thank Professor Rytter for providing us with an early preprint of [10].

References

[1] Maxime Crochemore, An optimal algorithm for computing all the repetitions in a word, *Inform. Process. Letters* 12–5 (1981) 244–248.

[2] Kangmin Fan, Simon J. Puglisi, W. F. Smyth, and Andrew Turpin, A new periodicity lemma, *SIAM Journal on Discrete Mathematics* (2006) to appear.

[3] N.J. Fine, H.S. Wilf, Uniqueness theorem for periodic functions, *Proc. Amer. Math. Soc.*, 16 (1965) 109–114.

[4] Frantisek Franek, Jamie Simpson, and W. F. Smyth, The maximum number of runs in a string, In Mirka Miller and Kunsoo Park, editors, *Proc. 14th Australasian Workshop on Combinatorial Algorithms*, Seoul (2003) 36–45.

[5] Frantisek Franek and Qian Yang, An asymptotic lower bound for the maximal-number-of-runs function, In Jan Holub and Jan Žďárek, editors, *Proc. Prague Stringology Conf. '06*, Prague (2006) 3–8.

[6] M.Gabriella Castelli, Filippo Mignosi, Antonio Restivo, Fine and Wilf's theorem for three periods and a generalization of Sturmian Words, *Theoretical Comput. Sci.* 218 (1999) 83–94.

[7] Roman Kolpakov and Gregory Kucherov, On maximal repetitions in words, *Journal of Discrete Algorithms*, 1(1) 159–186, 2000.

[8] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge (2002).

[9] M. G. Main, Detecting leftmost maximal periodicities, *Discrete Applied Mathematics*, 25:145–153, 1989.

[10] Wojciech Rytter, The number of runs in a string: Improved analysis of the linear upper bound, In B. Durand and W. Thomas, editors, *STACS 2006*, number 3884 in Lecture Notes in Computer Science, pages 184–195. Springer-Verlag, Berlin, 2006.
[11] Jamie Simpson, Intersecting periodic words, *Theoretical Comput. Sci.*, to appear.
[12] Bill Smyth, *Computing Patterns in Strings*, Addison-Wesley-Pearson Education Limited, Essex, England (2003).
[13] Axel Thue, Über unendliche zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana*, 7:1–22, 1906.

Department of Computing
Curtin University of Technology
GPO Box U1987
Perth, Western Australia 6845
E-mail address: puglissj@cs.curtin.edu.au

Department of Mathematics and Statistics
Curtin University of Technology
GPO Box U1987
Perth, Western Australia 6845
E-mail address: simpson@maths.curtin.edu.au

Department of Computing
Curtin University of Technology
GPO Box U1987
Perth, Western Australia 6845
E-mail address: smyth@cs.curtin.edu.au

Department of Computing & Software
McMaster University
Hamilton, Ontario, Canada L8S 4K1
E-mail address: smyth@mcmaster.ca