Problem
The Maximum Number of Runs in a String

Bill Smyth

1 Algorithms Research Group, Department of Computing & Software
McMaster University, Hamilton, Ontario, Canada L8S 4K1
smyth@mcmaster.ca
www.cas.mcmaster.ca/cas/research/algorithms.htm

2 Digital Ecosystems & Business Intelligence Institute
and Department of Computing, Curtin University, GPO Box U1987
Perth WA 6845, Australia
smyth@computing.edu.au

Given a nonempty string \( u \) and an integer \( e \geq 2 \), we call \( u^e \) a repetition; if \( u \) itself is not a repetition, then \( u^e \) is a proper repetition. Given a string \( x \), a repetition in \( x \) is a substring \( x[i..i+e|u|-1] = u^e \), where \( u^e \) is a proper repetition and neither \( x[i+e|u|..i+(e+1)|u|] \) nor \( x[i-|u|..i-1] \) equals \( u \). We say the repetition has period \(|u|\) and exponent \( e \); it can be specified by the integer triple \((i, |u|, e)\). It is well known [2] that the maximum number of repetitions in a string \( x = x[1..n] \) is \( \Theta(n \log n) \), and that the number of repetitions in \( x \) can be computed in \( \Theta(n \log n) \) time [2, 1, 10].

A string \( u \) is a run iff it is periodic of (minimum) period \( p \leq |u|/2 \). Thus \( x = abaabaabaabaab = (aba)^4ab \) is a run of period \(|aba| = 3 \). A substring \( u = x[i..j] \) of \( x \) is a run or maximal periodicity in \( x \) iff it is a run of period \( p \) and neither \( x[i-1..j] \) nor \( x[i..j+1] \) is a run of period \( p \). The run \( u \) has exponent \( e = \lfloor |u|/p \rfloor \) and possibly empty tail \( t = x[i+ep..j] \) (proper prefix of \( x[i..i+p-1] \)). Thus

\[
\begin{array}{c}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\
\end{array}
\]

\[
x = b a a a b a a b a a b a b a
\]

contains a run \( x[3..12] \) of period \( p = 3 \) and exponent \( e = 3 \) with tail \( t = a \) of length \( t = |t| = 1 \). It can be specified by a 4-tuple \((i, p, e, t) = (3, 3, 3, 1)\).

and it includes the repetitions \((aab)^3\), \((aba)^3\) and \((baa)^2\) of period \( p = 3 \).

In general it is easy to see that for \( e = 2 \) a run encodes \( t+1 \) repetitions;
for $c > 2$, $p$ repetitions. Clearly, computing all the runs in $x$ specifies all the repetitions in $x$. The idea of a run was introduced in [9].

Let $r_x$ denote the number of runs that actually occur in a given string $x$, and let $\rho(n)$ denote the maximum number of runs that can possibly occur in any string $x$ of given length $n$. A string $x = x[1..n]$ such that $r_x = \rho(n)$ is said to be run-maximal.

In [7, 8] it was shown that there exist universal positive constants $k_1$ and $k_2$ such that

$$\frac{\rho(n)}{n} < k_1 - k_2 \log_2 n / \sqrt{n},$$

but the proof was nonconstructive and provided no way of estimating the magnitude of $k_1$ and $k_2$. In [7], using a brute force algorithm, a table of $\rho(n)$ was computed for $n = 5, 6, \ldots, 31$, giving also for each $n$ an example of a run-maximal string; for every $n$ in this range, $\rho(n)/n < 1$ and $\rho(n) \leq \rho(n-1) + 2$. In [5] an infinite sequence $X = \{x_1, x_2, \ldots\}$ of strings was described, with $|x_{i+1}| > |x_i|$ for every $i \geq 1$, such that

$$\lim_{i \to \infty} \frac{r_{x_i}}{|x_i|} = \frac{3}{2\phi},$$

where $\phi = \frac{1 + \sqrt{5}}{2}$ is the golden mean. Moreover, it was conjectured that in fact

$$\lim_{n \to \infty} \frac{\rho(n)}{n} = \frac{3}{2\phi}.$$  \hspace{1cm} (1)

Recently a different and simpler construction was found [6] to yield another infinite sequence $X$ of strings for which the ratio $r_{x_i}/|x_i|$ approached the same limit; in addition, it was shown that for every $\epsilon > 0$ and for every sufficiently large $n = n(\epsilon)$, $\frac{3}{2\phi} - \epsilon$ provides an asymptotic lower bound on $\rho(n)/n$.

In 2006 considerable progress was made on the estimation of an upper bound on $\rho(n)/n$:

* $\rho(n)/n \leq 5.0$ [12];
* $\rho(n)/n \leq 3.48$ [11];
* $\rho(n)/n \leq 3.44$ [13];
* $\rho(n)/n \leq 1.6$ [3].

Thus the problem may be stated as follows:

**Is conjecture (1) true?**

If not, then characterize the function $\rho(n)/n$.

Help may be found in recent work studying the limitations imposed on the existence and length of runs in neighbourhoods of positions where two runs are known to exist [4, 14].
References