Average Value of Sum of Exponents of Runs in Strings

Kazuhiko Kusano, Wataru Matsubara, Akira Ishino, and Ayumi Shinohara

Graduate School of Information Science, Tohoku University,
Aramaki aza Aoba 6-6-05, Aoba-ku, Sendai 980-8579, Japan
{kusano@shino., matsubara@shino., ishino@, ayumi@}ecei.tohoku.ac.jp

Abstract. A substring $w[i..j]$ in w is called a repetition of period p if $s[k] = s[k+p]$ for any $i \leq k \leq j-p$. Especially, a maximal repetition, which cannot be extended neither to left nor to right, is called a run. The ratio of the length of the run to its period, i.e. $\frac{j-i+1}{p}$, is called an exponent. The sum of exponents of runs in a string is of interest. The maximal value of the sum is still unknown, and the current upper bound is $2.9n$ given by Crochemore and Ilie, where n is the length of a string. In this paper we show a closed formula which exactly expresses the average value of it for any n and any alphabet size, and the limit of this value per unit length as n approaches infinity. For binary strings, the limit value is approximately 1.13103.

1 Introduction

Repetitions in strings are an important element in the analysis and processing of strings. We especially focus on the runs, which are non-extendable repetitions. Kolpakov and Kucherov showed that the maximal number of runs $\rho(n)$ in any strings of length n is at most cn for some constant c [4]. Although they gave no value for c, recently there have been several results lowering the value [1,3,9,11]. The currently known best upper bound is $c = 1.048$ [2,3]. It is conjectured that $c < 1$.

A repetition count of run is called an exponent, and the maximal sum of exponents is also well studied [1,5,10]. It is proved that the maximal sum of exponents is linear and the current best upper bound is $2.9n$ [1]. It is conjectured that the sum of exponents is less than $2n$.

Although the exact estimation of the maximal number $\rho(n)$ of runs is still unknown, Puglisi and Simpson [8] presented a formula that gives the number of runs in a string of length n on average as follows:

$$r(n) = \sum_{p=1}^{\frac{n}{2}} \sigma^{n-2p-1} \left((n-2p+1)\sigma - (n-2p)\right) \sum_{d|p} \mu(d)\sigma^{\frac{p}{d}},$$

where σ is the alphabet size and $\mu(n)$ is the Möbius function.

In this paper, we consider the average value $e(n)$ of sum of exponents of runs in strings of length n and prove that

$$e(n) = \sum_{p=1}^{\frac{n}{2}} L(p) \left(2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}\right).$$

Moreover, we show that

$$\lim_{n\to\infty} \frac{e(n)}{n} = \sum_{d=1}^{\infty} \mu(d) \left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma} + \frac{1}{d\sigma} \ln\left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma}\right)\right),$$

where $L(n)$ is the number of Lyndon words of length n.

Kazuhiko Kusano, Wataru Matsubara, Akira Ishino, Ayumi Shinohara : Average Value of Sum of Exponents of Runs in Strings, pp. 185–192.
Proceedings of PSC 2008, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04145-1

2 Definitions

Let $\Sigma = \{0, 1, 2, \ldots, \sigma - 1\}$ be an *alphabet* of size σ, that is, $|\Sigma| = \sigma$. The set of all the strings on Σ is denoted by Σ^*, and the set of all the strings of length n by Σ^n. For a string w, we denote its length by $|w|$. We index w from 0 to $|w| - 1$ and denote its ith letter by w, i.e. $w = w[0]w[1] \ldots w[|w| - 1]$. We denote by $w[i..j]$ the *substring* $w[i]w[i + 1] \ldots w[j]$ of w. Let ε be the empty string. We say that p is a *period* of w if $w[i] = w[i + p]$ holds for any $i \geq 0$. For a string $w = xyz$, x, y and z are called a *prefix*, *substring* and *suffix* of w, respectively.

A substring $w[i..j]$ of w is called a *run* if it is *periodic*, i.e., it has the shortest period p satisfying $p \leq \frac{j-i+1}{2}$ and it is *non-extendable*, i.e., it satisfies the following two conditions:

$$\begin{array}{ll} i = 0 & \text{or } w[i-1] \neq w[i+p-1], \\ j = n-1 & \text{or } w[j+1] \neq w[j-p+1]. \end{array}$$

We denote the run $w[i..j]$ by a triple (i, j, p). The *exponent* is the ratio $\frac{j-i+1}{p}$, and the *root* is the prefix $w[i..i + p - 1]$ of length p. We denote by $Runs(w)$ the number of runs contained in string w, and by $Exp(w)$ the sum of exponents of all runs in string w.

We say that a string w is *primitive* if w cannot be written as $w = u^k$ by any string u and any integer $k \geq 2$. A string w is called a *Lyndon word* if w is minimal in the lexicographical ordering of all its non-empty suffixes [6]. We denote by $L(n)$ the number of Lyndon words of length n over the given alphabet. By these definitions, Lyndon words must be primitive and the number of primitive strings of length n is equal to $nL(n)$. The Möbius function $\mu(n)$ is defined by

$$\mu(n) = \begin{cases} 0 & \text{if } n \text{ has one or more repeated prime factors,} \\ 1 & \text{if } n \text{ has an even number of distinct prime factors,} \\ -1 & \text{if } n \text{ has an odd number of distinct prime factors.} \end{cases}$$

It is known that $L(n)$ can be expressed as follows [7]:

$$L(n) = \frac{1}{n} \sum_{d|n} \mu\left(\frac{n}{d}\right) \sigma^d.$$

The notation $d|p$ means that d is a divisor of p.

3 Main Result

We are interested in the average number $r(n)$ of runs and the average value $e(n)$ of sum of exponents of runs in strings of length n, defined as follows:

$$\begin{aligned} r(n) &= average\{Runs(w) : w \in \Sigma^n\}, \\ e(n) &= average\{Exp(w) : w \in \Sigma^n\}. \end{aligned}$$

Puglisi and Simpson showed that the following equation holds.

Theorem 1 (Puglisi and Simpson [8])**.**

$$r(n) = \sum_{p=1}^{\frac{n}{2}} \sigma^{n-2p-1} \left((n-2p+1)\sigma - (n-2p)\right) \sum_{d|p} \mu(d)\sigma^{\frac{p}{d}}.$$

We prove the following equation in the sequel.

Theorem 2.

$$e(n) = \sum_{p=1}^{\frac{n}{2}} L(p) \left(2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}\right).$$

For a string w of length n and a positive integer p, we define a string $d(w,p)$ of length $n-p$ as follows:

$$d(w,p)[i] = w[i+p] - w[i] \pmod{\sigma} \qquad \text{for } 0 \le i < n-p,$$

where the operators $-$ and mod are applied to symbols as if these symbols are numbers. For example, for a string $w = \texttt{21010}$ on $\Sigma = \{\texttt{0},\texttt{1},\texttt{2}\}$, we have $d(w,1) = \texttt{2212}$ and $d(w,2) = \texttt{100}$.

A substring $w[i..j]$ of $w \in \Sigma^n$ is called a 0*-segment* if $w[i..j]$ is a maximal block of 0's, that is, $w[t] = \texttt{0}$ for every t $(i \le t \le j)$ and it satisfies the following two conditions:

$$\begin{array}{ll} i = 0 & \text{or } w[i-1] \neq \texttt{0}, \\ j = n-1 & \text{or } w[j+1] \neq \texttt{0}. \end{array}$$

We denote the 0-segment by a pair (i,j).

Example 3. For string 0012000102, 0-segments are $(0,1)$, $(4,6)$, $(8,8)$.

Lemma 4. *For any string w, a substring $w[i..j+p]$ is a run with period p if and only if $d(w,p)[i..j]$ is a 0-segment of length $\ge p$.*

Proof. When there exists 0-segment (i,j) in $d(w,p)$, it holds that $w[t] = w[t+p]$ $(i \le t \le j)$, i.e., $w[i..j+p]$ has the period p. $|w[i..j+p]| = j+p-i+1 \ge 2p$ if and only if $|d(w,p)[i..j]| = j-i+1 \ge p$. Moreover, $w[i..j+p]$ satisfies the non-extendable condition. Therefore, $w[i..j+p]$ is a run. The "only if" part is clear. □

We denote by $c(n,p)$ the number of 0-segments of length p in all strings from Σ^n, and by $C(n,p)$ the number of 0-segments of length $\ge p$ in all strings from Σ^n. By definition, $C(n,p) = \sum_{i=p}^{n} c(n,i)$. For 0-segments of length $\ge p$ in Σ^n, we denote the sum of $\frac{l}{p}$ by $C_e(n,p)$, where l is the length of each 0-segments, i.e., $C_e(n,p) = \sum_{i=p}^{n} c(n,i)\frac{i}{p}$.

Example 5. For $\sigma = 2$, $c(5,2)$ is 12 because among all strings of length 5, all 0-segments of length 2 are underlined as follows:

00000	00100	01000	01100	10000	10100	11000	11100
00001	00101	01001	01101	10001	10101	11001	11101
00010	00110	01010	01110	10010	10110	11010	11110
00011	00111	01011	01111	10011	10111	11011	11111

Similarly, $c(5,3) = 5$, $c(5,4) = 2$ Macro 3 $c(5,5) = 1$. Then $C(5,2) = 12+5+2+1 = 20$. $C_e(5,2) = 12 \cdot \frac{2}{2} + 5 \cdot \frac{3}{2} + 2 \cdot \frac{4}{2} + \frac{5}{2} = 26$.

Lemma 6. *For any positive integer n and $p \leq n$, it holds that*

$$C(n,p) = (n-p+1)\sigma^{n-p} - (n-p)\sigma^{n-p-1}, \text{ and}$$
$$C_e(n,p) = \frac{1}{p}\left(p(n-p+1)\sigma^{n-p} - (p-1)(n-p)\sigma^{n-p-1}\right).$$

Proof. Let $Q_{n,p}$ be the set of pairs of strings separated by 0-segments of length p giving in concatenation strings of length n:

$$Q_{n,p} = \{(\alpha,\beta) : \alpha 0^p \beta \in \Sigma^n \wedge \alpha,\beta \in \Sigma^* \wedge \alpha[|\alpha|-1] \neq 0 \wedge \beta[0] \neq 0\}.$$

Then $c(n,p) = |Q_{n,p}|$ because a one-to-one correspondence exists between 0-segments of length p and $Q_{n,p}$.

Example 7. For $\sigma = 2, n = 3$ and $p = 1$, there are 0-segments of length p in Σ^n as follows:

$$\Sigma^3 = \{000, 001, \underline{0}1\underline{0}, \underline{0}11, 100, 1\underline{0}1, 11\underline{0}, 111\}.$$

The number of 0-segments $c(3,1)$ is 5. Then $Q_{3,1}$ becomes as follows:

$$Q_{3,1} = \{(\varepsilon, 10), (01, \varepsilon), (\varepsilon, 11), (1,1), (11, \varepsilon)\}.$$

We get expression for $c(n,p)$ by considering the number of elements of $Q_{n,p}$ in two cases.

(1) For $p \leq n-1$,
When $\alpha = \varepsilon$, since $|\beta| = n-p$ and $\beta[0] \neq 0$, there are $(\sigma-1)$ choices for $\beta[0]$ and σ^{n-p-1} choices for $\beta[1..n-p-1]$. $|Q_{n,p}| = (\sigma-1)\sigma^{n-p-1}$. Similarly, when $\beta = \varepsilon$, $|Q_{n,p}| = (\sigma-1)\sigma^{n-p-1}$. In the case of $\alpha \neq \varepsilon$ and $\beta \neq \varepsilon$, there are $(n-p-1)$ choices for the position of 0^p, $(\sigma-1)$ choices for $\alpha[|\alpha|-1]$ and $\beta[0]$, and σ^{n-p-2} choices for the other characters since $|\alpha| + |\beta| = n-p$, $\alpha[|\alpha|-1] \neq 0$, and $\beta[0] \neq 0$. $|Q_{n,p}| = (n-p-1)(\sigma-1)^2\sigma^{n-p-2}$. For $p = n-1$, α or β is ε, and $(n-p-1)(\sigma-1)^2\sigma^{n-p-2}$ equal to 0. Therefore,

$$\begin{aligned} c(n,p) &= |Q_{n,p}| \\ &= 2(\sigma-1)\sigma^{n-p-1} + (n-p-1)(\sigma-1)^2\sigma^{n-p-2} \\ &= (n-p+1)\sigma^{n-p} - 2(n-p)\sigma^{n-p-1} + (n-p-1)\sigma^{n-p-2}. \end{aligned}$$

(2) For $p = n$,
Since $\alpha = \beta = \varepsilon$,

$$c(n,p) = |Q_{n,p}| = 1.$$

For $p \leq n-1$,

$$\begin{aligned} C(n,p) &= \sum_{i=p}^{n} c(n,i) \\ &= \sum_{i=p}^{n-2} \left((n-i+1)\sigma^{n-i} - 2(n-i)\sigma^{n-i-1} + (n-i-1)\sigma^{n-i-2}\right) + 2(\sigma-1) + 1 \\ &= (n-p+1)\sigma^{n-p} - (n-p)\sigma^{n-p-1}. \end{aligned}$$

This equation holds for $C(n,n)=1$.

For $p \leq n-1$,

$$\begin{aligned} C_e(n,p) &= \sum_{i=p}^{n} c(n,i)\frac{i}{p} \\ &= \sum_{i=p}^{n-2} \left((n-i+1)\sigma^{n-i} - 2(n-i)\sigma^{n-i-1} + (n-i-1)\sigma^{n-i-2}\right)\frac{i}{p} \\ &\quad +2(\sigma-1)\frac{n-1}{p} + \frac{n}{p} \\ &= \frac{1}{p}\left(p(n-p+1)\sigma^{n-p} - (p-1)(n-p)\sigma^{n-p-1}\right). \end{aligned}$$

This equation holds for $C_e(n,n)=1$. □

Lemma 8. *For any integer p and strings w and v of length n such that $d(w,p) = d(v,p)$, $w[i..i+p-1] = v[i..i+p-1]$ for some i if and only if $w = v$.*

Proof. ($=:$) We prove this by induction. Let $i \leq j < i+p$ and k are integers. For $k=0$, $w[j+kp] = v[j+kp]$ is hold. For $k \geq 1$, if $w[j+kp] = v[j+kp]$ is hold, $w[j+(k+1)p] = w[j+kp] + d(w,p)[j+kp] \pmod{\sigma} = v[j+kp] + d(v,p)[j+kp] \pmod{\sigma} = v[j+(k+1)p]$. Then, $w[i..n-1] = v[i..n-1]$. Similarly, $w[0..i+p-1] = v[0..i+p-1]$. Therefore, $w = v$.
($\Longleftarrow$) It is clear. □

When $w \in \Sigma^n$, the length of $d(w,p)$ is $n-p$ and the number of 0-segments in Σ^{n-p} is $C(n-p,p)$. By Lemma 4 and 8, there are $\sigma^p C(n-p,p)$ runs which have period p in Σ^n. However, runs may have different periods. For example, 0101010101 has periods both 2 and 4. To prevent counting these runs more than once, we should consider counting runs with the minimum period.

Lemma 9. *The ratio of the number of runs whose shortest period is p to the number of runs which have period p is $\frac{pL(p)}{\sigma^p}$.*

Proof. The number of runs which have period p is $\sigma^p C(n-p,p)$. On the other hand, if a run of period p has different period $q < p$, the run also has period $\gcd(p,q)$ by Periodicity Lemma [6]. So $w[i..j]$ has no period $q < p$ if its root is primitive. The number of primitive strings of length p is $pL(p)$. Therefore, the number of runs whose shortest period is p is $pL(p)C(n-p,p)$. □

By Lemma 4, 0-segments of length l ($l \geq p$) in $d(w,p)$ correspond to runs of length $l+p$ in w. The exponents of these runs are $\frac{l}{p}+1$. This and Lemma 6, 8, and 9 derive $\sigma^n e(n)$, the sum of exponents of all runs in Σ^n, as follows:

$$\begin{aligned} \sigma^n e(n) &= \sum_{p=1}^{\frac{n}{2}} pL(p)\left(C_e(n-p,p) + C(n-p,p)\right) \\ &= \sum_{p=1}^{\frac{n}{2}} L(p)\left(2p(n-2p+1)\sigma^{n-2p} - (2p-1)(n-2p)\sigma^{n-2p-1}\right). \end{aligned}$$

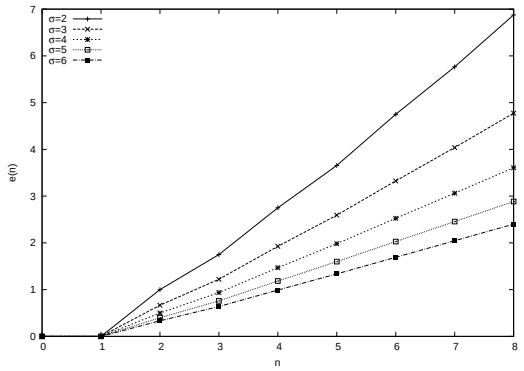

Figure 1. The average values $e(n)$ of sum of exponents of runs in strings on various sized alphabets.

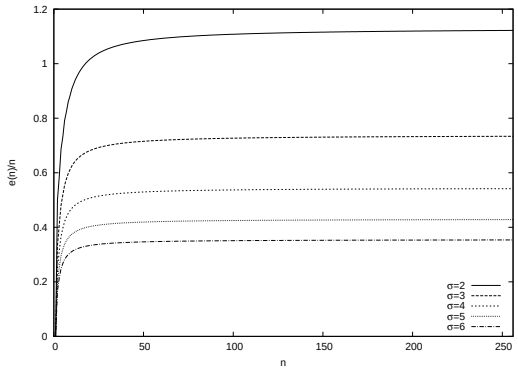

Figure 2. The average values $\frac{e(n)}{n}$ per unit length of sum of exponents of runs in strings on various sized alphabets.

We now get the Theorem 2. In Figure 1 it is shown that the average value $e(n)$ grows almost linearly, as n increases. The convergence of $\frac{e(n)}{n}$ is illustrated in Figure 2. For the limit of $e(n)$, we get the following theorem.

Theorem 10. *The limit of* $\frac{e(n)}{n}$*, as* $n \to \infty$*, is*

$$\sum_{d=1}^{\infty} \mu(d) \left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma} + \frac{1}{d\sigma} \ln \left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma} \right) \right).$$

To prove the theorem we deform $\frac{e(n)}{n}$.

Proposition 11.

$$\frac{e(n)}{n} = \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^{-2pd+p-1} \left(2(\sigma-1) - \frac{4pd}{n}(\sigma-1) + \frac{1}{pd} + \frac{2}{n}(\sigma-1) \right)$$

Proof. Let $2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}$ be $f(p)$.

$$\begin{aligned}
e(n) &= \sum_{p=1}^{\frac{n}{2}} L(p) f(p) \\
&= \sum_{p=1}^{\frac{n}{2}} \sum_{d|p} \mu\left(\frac{p}{d}\right) \sigma^d \frac{f(p)}{p} \\
&= \quad \mu(1)\sigma^1 \tfrac{f(1)}{1} \\
&\quad +\mu(2)\sigma^1 \tfrac{f(2)}{2} + \mu(1)\sigma^2 \tfrac{f(2)}{2} \\
&\quad +\mu(3)\sigma^1 \tfrac{f(3)}{3} \qquad\qquad + \mu(1)\sigma^3 \tfrac{f(3)}{3} \\
&\quad +\mu(4)\sigma^1 \tfrac{f(4)}{4} + \mu(2)\sigma^2 \tfrac{f(4)}{4} \qquad\qquad + \mu(1)\sigma^4 \tfrac{f(4)}{4} \\
&\quad +\mu(5)\sigma^1 \tfrac{f(5)}{5} \qquad\qquad\qquad\qquad + \mu(1)\sigma^5 \tfrac{f(5)}{5} \\
&\quad +\mu(6)\sigma^1 \tfrac{f(6)}{6} + \mu(3)\sigma^2 \tfrac{f(6)}{6} + \mu(2)\sigma^3 \tfrac{f(6)}{6} \qquad\qquad + \mu(1)\sigma^6 \tfrac{f(6)}{6} \\
&\qquad\qquad \vdots
\end{aligned}$$

$$= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^p \frac{f(pd)}{pd} \qquad \text{(Factor } \mu(d) \text{ out)}$$

$$= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \frac{1}{pd} \left(2pd(n-2pd+1)\sigma^{-2pd+p} - (2pd-1)(n-2pd)\sigma^{-2pd+p-1}\right)$$

$$\frac{e(n)}{n} = \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \frac{1}{npd} \left(2pd(n-2pd+1)\sigma^{-2pd+p} - (2pd-1)(n-2pd)\sigma^{-2pd+p-1}\right)$$

$$= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^{-2pd+p-1} \left(2(\sigma-1) - \frac{4pd}{n}(\sigma-1) + \frac{1}{pd} + \frac{2}{n}(\sigma-1)\right)$$

□

Now we prove Theorem 10.

Proof. When $n \to \infty$, $\frac{1}{n} \to 0$ and $\sigma^{-2pd+p-1}\frac{4pd}{n}$ is also negligible because $\sigma^{-2pd+p-1}$ is small enough when $\frac{pd}{n}$ is considerable.

$$\lim_{n\to\infty} \frac{e(n)}{n} = \lim_{n\to\infty} \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^{-2pd+p-1} \left(2(\sigma-1) + \frac{1}{pd}\right)$$

$$= \lim_{n\to\infty} \sum_{d=1}^{\frac{n}{2}} \mu(d) \left(\frac{2\sigma^{-2d}(\sigma-1)}{1-\sigma^{1-2d}} - \frac{1}{d\sigma}\ln(1-\sigma^{1-2d})\right)$$

$$= \sum_{d=1}^{\infty} \mu(d) \left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma} + \frac{1}{d\sigma}\ln\left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma}\right)\right)$$

□

Table 1 shows the limit values of $\frac{e(n)}{n}$ and $\frac{r(n)}{n}$.

σ	2	3	4	5	6
$\lim_{n\to\infty} \frac{e(n)}{n}$	1.13103	0.73822	0.54459	0.43039	0.35536
$\lim_{n\to\infty} \frac{r(n)}{n}$	0.41165	0.30491	0.23736	0.19329	0.16268

Table 1. The limit values of $\frac{e(n)}{n}$ and $\frac{r(n)}{n}$ for various sized alphabets.

4 Conclusion

We showed a formula which expresses the average value of sum of exponents of runs in strings exactly, although the upper bound of it is still open. The situation is similar to the numbers of runs, as Puglisi and Simpson showed in [8]. Moreover we gave the limit of the value per unit length as the length of strings approaches infinity. For the alphabet size $\sigma = 2$, the value is approximately 1.13103.

References

1. M. Crochemore and L. Ilie: *Analysis of Maximal Repetitions in Strings*, in Proc. 32nd International Symposium on Mathematical Foundations of Computer Science (MFCS 2007), vol. 4708 of LNCS, 2007, pp. 465–476.
2. M. Crochemore, L. Ilie, and L. Tinta: *The "runs" conjecture.* http://www.csd.uwo.ca/~ilie/runs.html.
3. M. Crochemore, L. Ilie, and L. Tinta: *Towards a solution to the "runs" conjecture*, in Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM 2008), vol. 5029 of LNCS, 2008, pp. 290–302.
4. R. Kolpakov and G. Kucherov: *Finding maximal repetitions in a word in linear time*, in Proc. 40th Annual Symposium on Foundations of Computer Science (FOCS'99), 1999, pp. 596–604.
5. R. Kolpakov and G. Kucherov: *On the sum of exponents of maximal repetitions in a word*, Tech. Rep. 99-R-034, LORIA, France, 1999.
6. M. Lothaire: *Algebraic combinatorics on words*, Cambridge University Press New York, 2002.
7. M. Lothaire: *Applied Combinatorics on Words*, Cambridge University Press, 2005.
8. S. J. Puglisi and J. Simpson: *The expected number of runs in a string.* Australasian Journal of Combinatorics, 2008, in press.
9. S. J. Puglisi, J. Simpson, and W. F. Smyth: *How many runs can a string contain?* Theoretical Computer Science, 2007, in press.
10. W. Rytter: *The number of runs in a string: Improved analysis of the linear upper bound*, in Proc. 23rd Annual Symposium on Theoretical Aspects of Computer Science (STACS 2006), vol. 3884 of LNCS, 2006, pp. 184–195.
11. W. Rytter: *The number of runs in a string.* Information and Computation, 205(9) 2007, pp. 1459–1469.