THE EXPECTED NUMBER OF RUNS IN A WORD

SIMON PUGLISI AND JAMIE SIMPSON

Introduction A word is a finite or infinite sequence of symbols taken from a (usually finite) alphabet. Examples are sections of texts, sequences of DNA codons and computer files. The study of their combinatorial properties is usually considered to have begun with the work of Thue in 1906 ([9]) but has accelerated in the last decade with publication of three books ([4],[5],[6]) on the subject by M. Lothaire (who, incidentally, is not a human being but a gang of French mathematicians). Applications exist in many areas of mathematics and computer science although computer scientists usually refer to “strings” rather than “words”. The usual notation for a word of \( n \) elements is \( x = x[1..n] \), with \( x[i] \) being the \( i \)th element and \( x[i..j] \) block of elements from position \( i \) to position \( j \). The length of the word is its number of elements. Thus a word of length 8 on the alphabet \( \{a, b, c\} \) is

\[ x = aabaabac \]

and \( x[5..7] = aba \). Some of the notation used in the area is borrowed from semigroup theory so that \( x[i..j] \) is a factor of \( x \), and a set of 2 or more identical adjacent factors form a power. Thus the word \( x \) above contains the squares \( aa \) which we can write as \( a^2 \), \((aab)^2\) and \((aba)^2\). A word which is not a power is primitive. A word \( x \) or factor \( x \) is periodic with period \( p \) if \( x[i] = x[i+p] \) for all \( i \) such that \( i \) and \( i + p \) are in the word. In the example above \( x[1..2] \) and \( x[4..5] \) have period 1 and \( x[1..7] \) has period 3. Note that a power is necessarily periodic, but a periodic word is a power only if its period is a proper divisor of its length.

A run of period \( p \) in a word is a factor \( x[i..j] \) with period \( p \), length at least \( 2p \) and such that neither \( x[i-1..j] \) nor \( x[i..j+1] \) have period \( p \). Runs are also called maximal repetitions. In the example above \( x[1..2] \) and \( x[4..5] \) are runs of period 1 and \( x[1..7] \) is a run of period 3. Runs have importance in data compression ([?]) a run \( x[i..j] \) with period \( p \) can be fully described by reporting \( i, j, p \) and \( x[i..i+p-1] \). In recent years there has been great interest in the maximum number of runs that can occur in a word of length \( n \). We call this number \( \rho(n) \). This takes larger values than one might expect, for example \( \rho(n) = \).

We use bold face letters for words and factors, and use ordinary face for letters and numbers. Recall that the Mobius function is \( \mu(n) \) defined by \( \mu(1) = 1 \) and if \( n \) has prime factorisation

\[ n = \prod_{i=1}^{l} p_i^{\alpha_i} \]

then \( \mu(n) = 0 \) if \( \alpha_i > 1 \) for any \( i \) and \( (-1)^l \) otherwise. A word \( x \) is primitive if it there do not exist a word \( y \) and integer \( n \) greater than 1 such that \( x = y^n \). We write \( P_q(n) \) for the number of primitive words of length \( n \) on an alphabet of size \( q \). The following result appears in Lothaire [4] as Equation (1.3.7) and is easily proved using Mobius inversion.

Lemma 0.1. \( P_q(n) = \sum_{d|n} \mu(d) q^{n/d} \)
Theorem 0.2. The number of runs of period $p$ in the set of all words of length $n$ on an alphabet of size $q$ is

$$q^{n-2p-1}((n - 2p + 1)q - (n - 2p))P_q(p)$$

for $n \geq 2p$.

Proof: We prove this by induction on $n$. Let $N(n,p,q)$ be the number of runs of period $p$ in the set of all words of length $n$ on an alphabet of size $q$. If $n = 2p$ each primitive prefix of length $p$ gives rise to exactly one run. Thus there are exactly $P_q(p)$ runs among these words. Thus $N(2p,p,q) = P_q(2p)$ and the formula holds for $n = 2p$.

Now suppose the statement holds for $n = k$. Each word of length $k+1$ will contain the same set of runs as its length $k$ prefix. Each length $k$ word is the prefix of $q$ length $k+1$ words. This gives us $q$ runs among the length $k+1$ words. The only way a word of length $k+1$ can contain one more run of period $p$ than does its length $k$ prefix is if it has the form $uavvb$ where $a \neq b$, $|vb| = p$ and $vb$ is primitive. It is easily seen that $u$ has length $k - 2p$. There are thus $q^{k-2p}$ choices for $u$, $q-1$ choices for $a$ and $P_q(k)$ choices for $vb$. Thus the number of extra runs equals $q^{k-2p}(q-1)P_q(p)$. So the total number of runs in the $q^{k+1}$ words of length $k+1$ is

$$qN(k,p,q) + q^{k-2p}(q-1)P_q(p)$$

$$= q(q^{k-2p-1}((k - 2p + 1)q - (k - 2p))P_q(p)) + q^{k-2p}(q-1)P_q(p)$$

$$= q^{k-2p}((k - 2p + 2)q - (k + 1 - 2p))P_q(p),$$

as required. □

We now obtain a formula for the total number of runs in the set of all words of length $n$ on an alphabet of size $q$. We obtain this by summing the formula from the last theorem over all $p$ satisfying $p \leq n/2$.

Theorem 0.3. The number of runs in the set of all words of length $n$ on an alphabet of size $q$ is

$$\sum_{p=1}^{[n/2]} q^{n-2p-1}((n - 2p + 1)q - (n - 2p)) \sum_{d|p} \mu(d)q^{p/d}.$$ 

Proof: Let the required number be $M(n,q)$. Using the notation in the proof of the last theorem we have

$$M(n,q) = \sum_{p=1}^{[n/2]} N(n,p,q)$$

$$= \sum_{p=1}^{[n/2]} q^{n-2p-1}((n - 2p + 1)q - (n - 2p))P_q(p)$$

$$= \sum_{p=1}^{[n/2]} q^{n-2p-1}((n - 2p + 1)q - (n - 2p)) \sum_{d|p} \mu(d)q^{p/d}.$$ 

□
Remark. After changing the order of summation and some simplification the expression above becomes

\[ q^{n-1} \sum_{d=1}^{\lfloor n/2 \rfloor} \mu(d) \left\{ (nq+q-n) \frac{1-q^{(1-2d)f}}{q^{2d-1} - 1} - 2d(q-1) \frac{q^{(1-2d)f} (f - (f+1)q^{2d-1}) + q^{2d-1}}{(q^{2d-1})^2} \right\} \]

where \( f = \lfloor n/2d \rfloor \). This isn’t much of an improvement. We now look at some values of these functions. The first values shows the expected number of runs in words of length \( n \) using an alphabet of size \( q \). Since there are \( q^n \) such words the expected number of runs is just \( M(n, q)/q^n \).

<table>
<thead>
<tr>
<th>Alphabet size:</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word length: 2</td>
<td>0.5</td>
<td>0.3333</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.5556</td>
<td>0.36</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>1.4375</td>
<td>1.1235</td>
<td>0.7376</td>
<td>0.3871</td>
</tr>
<tr>
<td></td>
<td>3.4043</td>
<td>2.6318</td>
<td>1.7022</td>
<td>0.8824</td>
</tr>
<tr>
<td></td>
<td>7.4914</td>
<td>5.6789</td>
<td>3.6350</td>
<td>1.8733</td>
</tr>
</tbody>
</table>

Table 1. Expected number of runs in words of various lengths using various sized alphabets.

The expected number of runs decreases with the alphabet size. If we divide the figures in the table by the word lengths we get the expected number of runs per unit length. We will show below that this approaches a limit depending only on the alphabet size. Some values of the expected number of runs per unit length are shown in the next table.

<table>
<thead>
<tr>
<th>Alphabet size:</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word length: 2</td>
<td>0.2500</td>
<td>0.1667</td>
<td>0.1000</td>
<td>0.0500</td>
</tr>
<tr>
<td></td>
<td>0.2500</td>
<td>0.1852</td>
<td>0.1200</td>
<td>0.0633</td>
</tr>
<tr>
<td></td>
<td>0.2875</td>
<td>0.2247</td>
<td>0.1475</td>
<td>0.0774</td>
</tr>
<tr>
<td></td>
<td>0.3404</td>
<td>0.2632</td>
<td>0.1702</td>
<td>0.0882</td>
</tr>
<tr>
<td></td>
<td>0.3746</td>
<td>0.2839</td>
<td>0.1818</td>
<td>0.0937</td>
</tr>
<tr>
<td></td>
<td>0.3968</td>
<td>0.2965</td>
<td>0.1887</td>
<td>0.0969</td>
</tr>
<tr>
<td></td>
<td>0.4042</td>
<td>0.3007</td>
<td>0.1910</td>
<td>0.0980</td>
</tr>
<tr>
<td></td>
<td>0.4079</td>
<td>0.3028</td>
<td>0.1921</td>
<td>0.0986</td>
</tr>
<tr>
<td>Limits</td>
<td>0.4116</td>
<td>0.3049</td>
<td>0.1933</td>
<td>0.0991</td>
</tr>
</tbody>
</table>

Table 2. Expected number of runs per unit length in words of various lengths using various sized alphabets. The last row gives the limits obtained from Corollary 0.5.

To obtain a formula for the limits of the columns of the table we first obtain a formula for \( M(n+1, q)/q^n n + 1 - M(n, q)/q^n \).

**Theorem 0.4.** The limit of \( M(n+1, q)/q^{n+1} - M(n, q)/q^n \) as \( n \) goes to infinity equals

\[ \frac{q - 1}{q} \sum_{n=1}^{\infty} \frac{\mu(d)}{(q^{2d-1} - 1)} \]
**Proof:** We show that the limit approaches this sum as \( n \) goes through the even integers. The case for odd integers is similar. Writing \( 2N \) for \( n \) and substituting in the formula from Theorem 0.3 we get,

\[
\sum_{p=1}^{N} \frac{q^{2N-2p}((2N - 2p + 2)q - (2N + 1 - 2p))}{q^{2N+1}} = q^{2N-2p-1}((2N - 2p + 1)q - (2N - 2p)) \sum_{d|p} \mu(d) q^{p/d}
\]

\[
= \sum_{p=1}^{N} q^{1-p} \sum_{d|p} \mu(d) q^{p/d}
\]

\[
= \frac{q-1}{q} \sum_{p=1}^{N} q^{-2p} \sum_{d|p} \mu(d) q^{p/d}.
\]

Changing the order of summation, and replacing \( p \) with \( \delta d \) this becomes,

\[
\frac{q-1}{q} \sum_{d=1}^{N} \mu(d) \sum_{\delta=1}^{[N/d]} q^{-(2d-1)\delta}
\]

\[
= \frac{q-1}{q} \sum_{d=1}^{N} \mu(d) \frac{1-q^{-(N/d)(2d-1)}}{1-q^{-(2d-1)}} q^{-(2d-1)}
\]

\[
= \frac{q-1}{q} \sum_{d=1}^{N} \frac{\mu(d)}{q^{2d-1}-1} (1 - q^{-[N/d](2d-1)}).
\]

We claim that as \( N \) goes to infinity the sum of the terms involving \( q^{-[N/d](2d-1)} \) goes to zero. The absolute value of this sum is,

\[
\sum_{d=1}^{N} \frac{|\mu(d)|}{q^{2d-1}-1} q^{-[N/d](2d-1)}.
\]

It is easily checked that \( [N/d](2d-1) \geq N \) so the sum is at most,

\[
\sum_{d=1}^{N} \frac{|\mu(d)|}{q^{2d-1}-1} q^{-N} < q^{-N} \sum_{d=1}^{\infty} \frac{1}{q^{2d-1}-1} < 2q^{-N}
\]

which clearly approaches 0. Thus the infinite sum approaches

\[
\frac{q-1}{q} \sum_{d=1}^{\infty} \frac{\mu(d)}{q^{2d-1}-1}.
\]

as required. □

**Corollary 0.5.** The expected number of runs per unit length in a word on an alphabet of size \( q \) tends to

\[
\frac{q-1}{q} \sum_{d=1}^{\infty} \frac{\mu(d)}{q^{2d-1}-1}
\]

as the word length approaches infinity.
Proof: For fixed $q$ and using earlier notation, the expected number of runs per unit length in a word of length $n$ is $M(n, q)/nq^n$. We write $L$ for $\frac{q-1}{q} \sum_{n=1}^{\infty} \frac{\mu(d)}{q^{n-1}}$ and let $\epsilon$ be a positive number. We will define a number $N$ below and show that for $n > N$ we have

$$|\frac{M(n, q)}{nq^n} - L| < \epsilon.$$ 

Note that

$$M(n, q) = \frac{1}{n} \sum_{m=1}^{n-1} \left( \frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} \right).$$

Since $M(1, q) = 0$ so we don’t need to allow for the beginning of the sum. By Theorem 0.4 there exists $N_1$ such that $m > N_1$ implies that

$$(0.1) \quad |\frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} - L| < \epsilon/2.$$ 

Let

$$B = \frac{1}{n} \sum_{m=1}^{n-1} \left( \frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} \right)$$

and choose $N$ so that

$$\frac{B - (N_1 + 1)L}{N} < \epsilon/2.$$ 

Thus,

$$|\frac{M(n, q)}{nq^n} - L| = \frac{1}{n} \sum_{m=1}^{n-1} \left( \frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} \right) - L| = \frac{1}{n} \sum_{m=1}^{N_1} \left( \frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} \right) + \frac{1}{n} \sum_{m=N_1+1}^{n-1} \left( \frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} \right) - L| = \frac{1}{n} \{B + \sum_{m=N_1+1}^{n-1} \left( \frac{M(m+1, q)}{q^{m+1}} - \frac{M(m, q)}{q^m} \right)\} - L|. $$

Thus, using (0.1),

$$\frac{B}{n} + \frac{n-1-N_1}{n} (L - \frac{\epsilon}{2}) - L < \frac{M(n, q)}{nq^n} - L < \frac{B}{n} + \frac{n-1-N_1}{n} (L + \frac{\epsilon}{2}) - L \Rightarrow \frac{B - (N_1 + 1)L}{n} = \frac{(n-1-N_1)\epsilon}{2n} < \frac{M(n, q)}{nq^n} - L < \frac{B - (N_1 + 1)L}{n} + \frac{(n-1-N_1)\epsilon}{2n} \Rightarrow -\epsilon < \frac{M(n, q)}{nq^n} - L < \epsilon$$

as required. □

Some values of the limit obtained in this corollary are shown in Table 2.
References


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