The Number of Runs in a Ternary Word

Hideo Bannai, Mathieu Giraud, Kazuhiko Kusano, Wataru Matsubara, Ayumi Shinohara, Jamie Simpson

The Number of Runs in a Ternary Word

Hideo Bannai[1], Mathieu Giraud[2], Kazuhiko Kusano[3], Wataru Matsubara[3], Ayumi Shinohara[3], and Jamie Simpson[4]

[1] Kyushu University, Japan
[2] Université Lille, France
[3] Tohoku University, Japan
[4] Curtin University, Australia

Abstract. It is not known that the maximum number of runs in a word of length n is attained by a binary word though it seems likely that this is the case. In this note, we report observations on runs in *ternary words*, in which every small factor contains all three letters.

1 Introduction

A *run* of period p in a word x is a factor $\mathbf{x}[i..j]$ with least period p, length at least $2p$ and such that neither $\mathbf{x}[i-1..j]$ nor $\mathbf{x}[i..j+1]$ has period p. Runs are also called *maximal repetitions*. Let $r(x)$ be the number of runs in x. In recent years there has been great interest in the maximum number of runs that can occur in a word of length n which we call $\rho(n)$. In 2000 Kolpakov and Kucherov [5] showed that $\rho(n) = O(n)$ but their method gave no information about the size of the implied constant. Since then a number of authors have obtained upper and lower bounds on $\rho(n)$, the best to date being that

$$0.944 < \lim_{n\to\infty} \frac{\rho(n)}{n} < 1.029 \tag{1}$$

The upper bound here is due to Crochemore, Ilie and Tinta [1,2], and the lower bound to Kusano et al [6] and Simpson [8]. It is known [4] that that $\lim_{n\to\infty} \rho(n)/n$ exists. It is also known that the expected number of runs in binary words of length n tends to about $0.412n$ as n goes to infinity and for ternary words to about $0.305n$ [7].

It is not known that the maximum number of runs in a word of length n is attained by a binary word though it seems likely that this is the case. Between lengths 17 and 35, all the words with the maximum number of runs are binary [3]. More generally, the high density words we know about are binary and [7] showed the expected density decreases with alphabet size.

Part of the problem is that whatever you can do with a two letter alphabet you can do with a three letter alphabet by just not using the third letter. Even if we insist that each letter be used we can achieve the same asymptotic run density with three letters as with two by taking a good binary word on the alphabet $\{a, b\}$ and attaching c to the front. Insisting that the frequency of each letter is greater than some bound doesn't help. We could take a binary word of length $n/3$, make three copies of it using the alphabets $\{a, b\}$, $\{b, c\}$ and $\{c, a\}$ then concatenate them. This will give a word of length n with the same run density as the original word and with each letter having frequency $n/3$. In both these cases we are dealing with binary words in disguise.

Hideo Bannai, Mathieu Giraud, Kazuhiko Kusano, Wataru Matsubara, Ayumi Shinohara, Jamie Simpson: The Number of Runs in a Ternary Word, pp. 178–181 .
Proceedings of PSC 2010, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04597-8

To make a word really ternary, we use the following definition. Let k be an integer with $k \geq 3$. A *ternary word of order* k is a word on the alphabet $\{a, b, c\}$ in which each factor of length k contains at least one occurrence of each letter.

Say that the maximum number of runs in such a word of length n is $\rho(k, n)$. We have

$$\rho(n) \geq \rho(k+1, n) \geq \rho(k, n)$$

and, when the order goes to the infinity, $\rho(k, n)$ will approach $\rho(n)$. Showing that

$$\lim_{n\to\infty} \frac{\rho(n)}{n} > \lim_{n\to\infty} \frac{\rho(k, n)}{n}$$

for all k would be equivalent to showing that, for bounded order, two letters are better than three. Towards this aim we have obtained upper and lower bounds for $\rho(k, n)$ for various values of k.

2 Lower bound

In [6] some of us used a search technique to find long run-rich words. The basic idea of this was to start with n run-rich words, augment each by adding a or b to its end giving $2n$ words. Then select the n most run-rich words from this collection and repeat the process. Using these techniques, we constructed a ternary word of order 9 and length 9686 containing 7728 runs. Concatenating this with itself infinitely often (which gives an extra 11 runs for each copy) produces an infinite order 9 word having run density $7739/9686 \approx 0.798988$. The word begins:

$aabbccaabbccbbccaabbccaabbccbbccaabbccaacaabbccaabbccbbccaabbc$
$caabbccbbccaabbccaaccaabbccaabbccbbccaabbccaabbccbbccaabbccaac$
$aabbccaabbccbbccaabbccaabbccbbccaabbccaaccaabbccaabbccbbccaabb$
$ccaacaabbccaabbccbbccaabbccaabbccbbccaabbccaacaabbccaabbccbbcc$
$aabbccaabbccbbccaabbccaaccaabbccaabbccbbccaabbccaabbccbbccaabb$
$ccaacaabbccaabbccbbccaabbccaabbccbbccaabbccaaccaabbccaabbccbbc$
$caabbccaacaabbccaabbccbbccaabbccaabbccbbccaabbccaaccaabbccaabb$
$ccbbccaabbccaabbccbbccaabbccaacaabbccaabbccbbccaabbccaabbccbbc$
$caabbccaaccaabbccaabbccbbccaabbccaabbccbbccaabbccaacaabbccaabb$
$ccbbccaabbccaabbccbbccaabbccaaccaabbccaabbccbbccaabbccaacaabbc$
$caabbccbbccaabbccaabbccbbccaabbccaacaabbccaabbccbbccaabbccaabb$
$ccbbccaabbccaaccaabbccaabbccbbccaabbccaabbccbbccaabbccaacaabbc$

Thus we have:

$$\lim_{n\to\infty} \frac{\rho(9, n)}{n} \geq 0.798988$$

3 Upper bound

To get an upper bound on $\rho(k,n)$, we used the techniques of [4] for various values of k. These techniques give an exact bound on the number of runs with period less than some bound p and then use a result of Crochemore and Ilie [1] which states that the number of runs with period greater than or equal to p in a word of length n is at most $6n/p$. Results for different values of k are given in the following table, which used $p = 9$. Concerning runs with period less than 9, even ternary words of order $k = 19$ do not achieve the 11/13 maximal run density for binary words [4].

Order k	Exact bound (period ≤ 9)	Upper bound on $\rho(k,n)/n$	
3	0/1	0.6000	
4	7/13	1.1385	*cabcabccabcabbcabcabbcabcaabcabcaabcabc*
5	7/11	1.2364	*abbccaabbca*
6	7/11	1.2364	
7	2/3	1.2667	*aacaabbcaabbcbbcaabbc*
8	17/24	1.3083	*bbccaabaabbccaabbccbccaa*
9	17/24	1.3083	
10	17/24	1.3083	
11	17/24	1.3083	
12	17/24	1.3083	
13	17/24	1.3083	
14	17/24	1.3083	
15	3/4	1.3500	*acaaccaacaabaabbaabaabbcbbccbbcbbcca*
16	3/4	1.3500	
17	3/4	1.3500	
18	3/4	1.3500	
19	25/33	1.3576	
$+\infty$	11/13	1.4462	*aababbabaabab*

Table 1. Upper bounds on $\rho(k,n)$ for various orders k. The second column gives the exact bound on the number of such runs with period at most 9. The third column adds this to 6/(9+1), the bound on the run density of runs with period greater than 9, to give the required upper bound. The last column shows examples of strings, that, concatenated with themselves infinitely often, give the exact bound.

For example, we have:

$$\lim_{n\to\infty} \frac{\rho(9,n)}{n} \leq 1.3083$$

Our results are not strong enough to be more than suggestive. To get more convincing evidence for the superiority of binary words, we would need to get the upper bound on $\rho(k,n)$ to be less than the lower bound on $\rho(n)$ given in (1). Perhaps this can be done using the more powerful techniques of Crochemore and Ilie.

4 Other remarks

We mention another condition on run-rich words which is suggested by experimental evidence. This is that run-rich words do not need to contain cubes, in particular the cubes *aaa* and *bbb*. This is not necessary since, for example, the word *aaa* contains the maximum possible number of runs for a 3 letter word, but we could use instead the word *aab* which has the same number of runs but no cube.

We also note the following observation which is in the other direction to our earlier results. Let x be a string of size n on the binary alphabet $\{a, b\}$. Let $\bar{x}$ be the string obtained by rewriting x with the as replaced by bs and the bs by as. The words x and $\bar{x}$ have the same number of runs.

For a binary word y of length p, consider the word $\tau(y)$ of size $(n+1)p$ obtained by the rewriting $\tau(a) = cx$ and $\tau(b) = c\bar{x}$. The word $\tau(y)$ is of ternary order $n+2$. It can be shown that no run is lost with such rewriting, thus

$$r(\tau(y)) \;\geq\; r(y) + p \cdot r(x)$$

If we select x and y to be run-maximal, that is $r(x) = r(\bar{x}) = \rho(n)$ and $r(y) = \rho(p)$, then we have

$$\rho(n+2, (n+1)p) \;\geq\; r(\tau(y)) \;\geq\; \rho(p) + p \cdot \rho(n)$$

thus

$$\lim_{n\to\infty} \frac{\rho(n+2, (n+1)p)}{(n+1)p} \;\geq\; \lim_{n\to\infty} \frac{\rho(n)}{n+1} \;=\; \lim_{n\to\infty} \frac{\rho(n)}{n}.$$

References

1. M. Crochemore and L. Ilie: *Maximal repetitions in strings*. J. Comput. Systems Sci., 74(5) 2008, pp. 796–807.
2. 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), P. Ferragina and G. M. Landau, eds., vol. 5029 of LNCS, 2008, pp. 290–302, See also `http://www.csd.uwo.ca/~ilie/runs.html`.
3. F. Franek: `http://www.cas.mcmaster.ca/~franek/research.html`.
4. M. Giraud: *Asymptotic behavior of the numbers of runs and microruns*. Information and Computation, 207(11) 2009, pp. 1221–1228.
5. R. Kolpakov and G. Kucherov: *On maximal repetitions in words*. Journal of Discrete Algorithms, 1(1) 2000, pp. 159–186.
6. W. Matsubara, K. Kusano, A. Ishino, H. Bannai, and A. Shinohara: *New lower bounds for the maximum number of runs in a string*, in Proceedings of the Prague Stringology Conference 2008 (PSC 2008), J. Holub and J. Žďárek, eds., Czech Technical University in Prague, Czech Republic, 2008, pp. 140–145.
7. S. J. Puglisi and J. Simpson: *The expected number of runs in a word*. Australasian J. Combinatorics, 42 2008, pp. 45–54.
8. J. Simpson: *Modified Padovan words and the maximum number of runs in a word*. Australasian J. Combinatorics, 46 2010, pp. 129–146.