AUSTRALASIAN JOURNAL OF COMBINATORICS
Volume **46** (2010), Pages 129–145

Modified Padovan words and the maximum number of runs in a word

Jamie Simpson

Department of Mathematics and Statistics
Curtin University of Technology
G.P.O. Box U1987
Perth, WA 6845
Australia
simpson@maths.curtin.edu.au

Abstract

A *run* (or maximal periodicity) in a word is a periodic factor whose length is at least twice the period and which cannot be extended to the left or right without changing the period. Recently Kusano et al. [6] used a clever search technique to find run-rich words and were able to show that the number of runs in a word of length n can be greater than $0.94457564n$. In this paper we use a two-stage process to construct words with a (very slightly) higher run density than theirs. We first produce ternary words which we call Modified Padovan words, then apply a morphism to these to produce run-rich binary words. The Modified Padovan words have interesting and surprising properties.

1 Introduction

We use the usual notation for combinatorics on words. A word of n elements is $x = x[1\,..\,n]$, with $x[i]$ being the ith element and $x[i\,..\,j]$ the *factor* of elements from position i to position j. If $i = 1$ then the factor as a *prefix* and if $j = n$ then it is a *suffix*. The letters in x come from some *alphabet* A. The set of all finite words with letters from A is A^*. The *length* of x, written $|x|$, is the number of occurrences of letters in x and the number of occurrences of the letter a in x is $|x|_a$. Two or more adjacent identical factors form a *power*. 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 $x[i]$ and $x[i+p]$ are in x. Two words x and y are *conjugate* if there exist words u and v such that $x = uv$ and $y = vu$. If $x = a_1a_2\ldots a_n$ then the *reverse* of x, written $R(x)$, is $a_n\ldots a_2a_1$. If $x = uvu$ we say that x has *border* u, and we see that x has period $|x| - |u|$.

Finally a *run* (or *maximal periodicity*) in a word x is a factor $x[i\,..\,j]$ having minimum period p, length at least $2p$ and such that neither $x[i-1\,..\,j]$ nor $x[i\,..\,j+1]$

is a factor with period p. Runs are important because of their applications in data compression and computational biology (see, for example, [5]). In recent years a number of papers have appeared concerning the function $\rho(n)$ which is the maximum number of runs that can occur in a word of length n. In 2000 Kolpakov and Kucherov [5] showed that $\rho(n) = O(n)$ but their method did not give any information about the size of the implied constant. Then in [12] Rytter showed that $\rho(n) < 5n$. This bound was improved in [11] and [1] and most recently by Crochemore and Ilie[1] to $\rho(n) < 1.029n$. Their method is difficult and heavily computational. Recently Giraud [4] has produced weaker results using a much simpler technique. He has also shown that $\lim_{n\to\infty} \rho(n)/n$ exists. In the other direction Franek et al. [3] produced words with a run density (number of runs per unit length) greater than 0.927. Then in [6] Kusano et al. published a word with density $56733/60064 \approx 0.944542$. Using the ideas of the present paper, which come from an analysis of the Kusano et al. word, Simon Puglisi and the author constructed a word of length 29196442 containing 27578248 runs giving a density greater than 0.94457564. Hideo Bannai and his colleagues have kindly published this word on their web site [2]. From all this we see that

$$0.9445756 < \lim_{n\to\infty} \frac{\rho(n)}{n} < 1.029. \tag{1.1}$$

We also note that the mean number of runs in words has been studied by Puglisi and Simpson [10] who showed that for random words the expected number of runs per unit length decreases with alphabet size. With a binary alphabet its limit as word length goes to infinity is about 0.4116. In this paper we describe a method for constructing a sequence of run-rich words and show that their asymptotic density is

$$\frac{11\psi^2 + 7\psi - 6}{11\psi^2 + 8\psi - 6} = 0.944575712\ldots$$

where ψ is the real root of $z^3 - z - 1 = 0$. This is a very small improvement on (1.1). The words are constructed in a two stage process. We first produce ternary words called Modified Padovan words using an iterative process, then apply a morphism h to them to produce run-rich binary words. Modified Padovan words have some surprising and interesting properties. Further analysis of them may lead to more substantial improvements to (1.1).

The paper contains six sections. In the next we discuss circular words and prove some lemmas about them. In the third section we review the definition and properties of the Padovan sequence and define Modified Padovan words. In the fourth and longest section we investigate the number and structure of the runs in Modified Padovan words and in the fifth apply these results to show that a morphism applied to the Modified Padovan words produces words with the stated density. In the last section we compare our words with those constructed in [3], mention some more properties of Modified Padovan words and suggest avenues for further work.

[1] See `http://www.uwo.ca/faculty/ilie/runs.html`

[2] `http://www.shino.ecei.tohoku.ac.jp/runs/`

2 Circular words

We will often use circular words which are words with their ends joined. To indicate that x is the circular word abc we write

$$x = \langle abc \rangle.$$

Note that if x and y are conjugates then $\langle x \rangle$ and $\langle y \rangle$ are the same circular word. In some ways $\langle x \rangle$ is analogous to the set of conjugates of x or to the infinite word x^ω; for example $\langle x \rangle$, the set of conjugates of x and x^ω all have the same set of factors with length not exceeding $|x|$. A *run* in a circular word $\langle x \rangle$ coincides with a run in the infinite word x^ω beginning in its length x prefix, though we exclude an infinite run with period $|x|$. Such a run may have length greater than $|x|$ but we show below that it cannot have period greater than $|x|$. The number of runs in $\langle u \rangle$ will generally be different from the number in u. For example, aba contains no runs but $\langle aba \rangle$ contains the run aa and $aabaa$ contains two runs but $\langle aabaa \rangle$ contains only one. One of the advantages of using circular words is that for the Modified Padovan words, which we define below, the formula for the number of runs is much simpler than it would be for ordinary linear words. It is usual to say that if $x[i+1\,..\,i+n]$ is a run with period p then its *generator* is $x[i+1\,..\,i+p]$. We will vary from this convention by saying that the *circular generator* of the run is $\langle x[i+1\,..\,i+p] \rangle$. This means, for example, that $abcabca$ and $cabcabcab$ have the same circular generator since $\langle abc \rangle = \langle cab \rangle$.

We will need the following lemmas which concern periodicity in circular words. The first is the well-known Periodicity Lemma of Fine and Wilf [2] which we state without proof.

Lemma 2.1. (The Periodicity Lemma) *If x is a word having two periods p and q and $|x| \geq p + q - \gcd(p, q)$ then x also has period $\gcd(p, q)$.*

Lemma 2.2. *A circular word of length n has no runs with period greater than or equal to n.*

Proof. We specified above that a circular word of length n does not contain a run of period n. Suppose, for the sake of contradiction, that $\langle x[1\,..\,n] \rangle$ contains a run beginning at $x[i+1]$ with period p where $p > n$. Then the factor $y[i+1\,..\,i+2p]$ of the (linear) word $y = x^\omega$ has periods p and n. By the Periodicity Lemma it therefore has period $\gcd(p, n)$ and so is not a run with period p. □

Lemma 2.3. *If x and y are words of length p, m and n are positive integers less than p with $m + n > p$ and*

$$x[1\,..\,m] = y[p - m + 1\,..\,p]$$

$$x[p - n + 1\,..\,p] = y[1\,..\,n]$$

then x and y have borders of length $m + n - p$.

Proof.

$$\begin{aligned} x[1\,..\,m+n-p] &= y[p-m+1\,..\,b] \\ &= x[p-m+1+p-n\,..\,p] \\ &= x[2p-m-n+1\,..\,p] \\ &= x[p-(m+n-p)+1\,..\,p]. \end{aligned}$$

Similarly we can show $y[1\,..\,m+n-p] = y[p-(m+n-p)+1\,..\,p]$. □

In applications of this lemma x and y will be factors of a circular word which intersect each other at each end.

3 The Padovan Sequence and Modified Padovan words

The Padovan ([15], [13]) sequence is named after the architect Richard Padovan. It is analogous to the Fibonacci sequence but rather than using the Fibonacci recurrence formula $F_n = F_{n-1} + F_{n-2}$ it uses, for $n \geq 3$,

$$P_n = P_{n-2} + P_{n-3} \tag{3.1}$$

with $P_0 = 1 = P_1 = P_2$. The first few terms are

$$1, 1, 1, 2, 2, 3, 4, 5, 7, 9, 12, 16, \ldots$$

The ratio of consecutive terms in the Fibonacci sequence approaches the Golden Ratio $(1 + \sqrt{(}5))/2$ which is the larger solution of $x^2 = x + 1$. With the Padovan sequence P_{n+1}/P_n approaches the Plastic Number

$$\psi = 1.324717957\cdots \tag{3.2}$$

which is the real root of $x^3 = x + 1$. The Padovan numbers satisfy a number of identities including, for $n \geq 5$,

$$P_n = P_{n-1} + P_{n-5}, \tag{3.3}$$

$$P_n = P_{n-3} + P_{n-4} + P_{n-5}, \tag{3.4}$$

$$\sum_{i=0}^{n} P_i = P_{n+5} - 2 \tag{3.5}$$

and

$$P_k \sim \frac{\psi^{k+2}}{3\psi^2 - 1} \text{ as } k \to \infty. \tag{3.6}$$

We will construct words p_k on the alphabet $\{a, b, c\}$ using the mapping

$$\phi(x) = R(f(x)) \tag{3.7}$$

where $R(x)$ is the reverse of x and f is the morphism $f(a) = aacab$, $f(b) = acab$ and $f(c) = ac$.

The words $p_1, \ldots, p_5$ are defined in the following table. For larger indices we define them recursively using

$$p_{k+5} = \phi(p_k). \tag{3.8}$$

We call these words *Modified Padovan words*. The construction is similar to that of *DOL* words — see [9].

p_1	b
p_2	a
p_3	ac
p_4	ba
p_5	aca
p_6	baca
p_7	bacaa
p_8	cabacaa
p_9	bacaabaca
p_{10}	bacaacabacaa
p_{11}	bacaacabacaabaca
p_{12}	bacaabacaacabacaabaca
p_{13}	bacaabacaacabacaabacabacaaca
p_{14}	bacaacabacaabacabacaabacaacabacaabaca
p_{15}	bacaabacaacabacaabacabacaacabacaabacaacabacaabaca

Table 1: The first 15 Modified Padovan words. Note that in each case $|p_k| = P_k$.

Theorem 3.1. *For all $k \geq 6$,*

$$|p_k|_a = P_{k-2},\ |p_k|_b = P_{k-6},\ |p_k|_c = P_{k-5}$$

and for $k \geq 1$

$$|p_k| = P_k.$$

Proof. We prove this by induction on k. For values of k in the interval $[6, 10]$ we see that the first part of the statement holds by inspecting Table 1. Suppose it holds up to p_{k+4}. From (3.8) and the definition of f we see that

$$|p_{k+5}|_a = 3|p_k|_a + 2|p_k|_b + |p_k|_c.$$

Using (3.1), (3.4) and the induction hypothesis this equals

$$\begin{aligned} 3P_{k-2} + 2P_{k-6} + P_{k-5} &= 2P_{k-2} + P_{k-4} + 2P_{k-5} + 2P_{k-6} \\ &= 2P_{k-2} + 2P_{k-3} + P_{k-4} \\ &= P_{k-1} + 2P_{k-2} + P_{k-3} \\ &= P_k + P_{k-1} + P_{k-2} \\ &= P_{k+3}. \end{aligned}$$

This establishes the first part of the theorem. Similarly, using (3.3),

$$\begin{aligned} |p_{k+5}|_b &= |p_k|_a + |p_k|_b \\ &= P_{k-2} + P_{k-6} \\ &= P_{k-1} \end{aligned}$$

$$\begin{aligned} |p_{k+5}|_c &= |p_k|_a + |p_k|_b + |p_k|_c \\ &= P_{k-2} + P_{k-6} + P_{k-5} \\ &= P_{k-2} + P_{k-3} \\ &= P_k. \end{aligned}$$

So the first three parts of the theorem hold by induction. From Table 1 the last part holds for $1 \leq k \leq 5$. For larger values of k

$$\begin{aligned} |p_{k+5}| &= |p_{k+5}|_a + |p_{k+5}|_b + |p_{k+5}|_c \\ &= P_{k+3} + P_{k-1} + P_k \\ &= P_{k+5}. \end{aligned}$$

□

4 Runs in Modified Padovan Words

The number of runs of different periods in Modified Padovan words of low index are shown in Table 2. The results in this and other tables were obtained by constructing words and counting runs with a computer. The reader will be able to guess how the pattern in the table continues. The next few results prove that the guess is probably right. "Probably" because I have not been able to show that the periods of all runs in $\langle p_k \rangle$ are Padovan numbers, although this is so for all Modified Padovan words we have examined. Subject to this qualification we will see that the lengths of Modified Padovan words are Padovan numbers, that the periods of runs in Modified Padovan words are Padovan numbers, that the number of runs with a given period in a circular Modified Padovan word is a Padovan number and that the circular generators of these runs are circular Modified Padovan words.

Word	Length	1	3	4	5	7	9	12	16	21	28	37	49	65	86
p_5	3	1													
p_6	4	0													
p_7	5	1													
p_8	7	1	1												
p_9	9	1	0	1	1										
p_{10}	12	2	1	0	1										
p_{11}	16	2	1	1	1	1									
p_{12}	21	3	1	1	2	0	1	1							
p_{13}	28	4	2	1	2	1	0	1	1						
p_{14}	37	5	2	2	3	1	1	1	1	1					
p_{15}	49	7	3	2	4	1	1	2	1	1	1				
p_{16}	65	9	4	3	5	2	1	2	2	1	1	1			
p_{17}	86	12	5	4	7	2	2	3	2	2	1	1	1		
p_{18}	114	16	7	5	9	3	2	4	3	2	2	1	1	1	
p_{19}	151	21	9	7	16	4	3	5	4	3	2	2	1	1	1

Table 2: The number of runs with various periods in $\langle p_k \rangle$ for k from 5 to 19. The number at the top of each column is the period. Except in the first column, all positive numbers appearing in this table are Padovan numbers.

Lemma 4.1. *For words u and v in $\{a, b, c\}^*$*

(a) $\langle \phi(uv) \rangle = \langle \phi(u)\phi(v) \rangle$

(b) $\langle \phi(uvu) \rangle = \langle \phi(u)\phi(v)\phi(u) \rangle$

(c) $\langle \phi(uvuvu) \rangle = \langle \phi(u)\phi(v)\phi(u)\phi(v)\phi(u) \rangle$.

Proof.

$$\begin{aligned} \langle \phi(uv) \rangle &= \langle R(f(uv)) \rangle \\ &= \langle R(f(v))R(f(u)) \rangle \\ &= \langle \phi(v)\phi(u) \rangle \\ &= \langle \phi(u)\phi(v) \rangle \end{aligned}$$

which proves (a). The other parts can be proved in a similar way. □

This lemma could be extended to apply to $\phi(x)$ whenever x is a word on the alphabet $\{u, v\}$ for which $R(x)$ is a conjugate of x.

Lemma 4.2. *For $k \geq 12$ there exist words u and v such that*

$$\langle p_k \rangle = \langle uvuvu \rangle \tag{4.1}$$

$$\langle p_{k-2} \rangle = \langle uvu \rangle \tag{4.2}$$

$$\langle p_{k-3} \rangle = \langle uv \rangle \tag{4.3}$$

$$\langle p_{k-7} \rangle = \langle u \rangle. \tag{4.4}$$

Proof. We prove this by induction on k. For $12 \leq k \leq 16$ the appropriate partitions of p_k are given in Table 3, with the factors u underlined. By comparison with Table 1 we see that (4.2), (4.3) and (4.4) hold for these values of k.

Now suppose that the statement holds for words from p_{12} to p_{k-1}, where $k \geq 17$, and consider p_k. The induction hypothesis implies that there exist words u and v satisfying

$$\begin{aligned} \langle p_{k-5} \rangle &= \langle uvuvu \rangle \\ \langle p_{k-7} \rangle &= \langle uvu \rangle \\ \langle p_{k-8} \rangle &= \langle uv \rangle \\ \langle p_{k-12} \rangle &= \langle u \rangle. \end{aligned}$$

By Lemma 4.1 we have

$$\begin{aligned} \langle p_k \rangle &= \langle \phi(p_{k-5}) \rangle = \langle \phi(u)\phi(v)\phi(u)\phi(v)\phi(u) \rangle \\ \langle p_{k-2} \rangle &= \langle \phi(p_{k-7}) \rangle = \langle \phi(u)\phi(v)\phi(u) \rangle \\ \langle p_{k-3} \rangle &= \langle \phi(p_{k-8}) \rangle = \langle \phi(u)\phi(v) \rangle \\ \langle p_{k-7} \rangle &= \langle \phi(p_{k-12}) \rangle = \langle \phi(u) \rangle \end{aligned}$$

so that (4.1), (4.2), (4.3) and (4.4) hold with $\phi(u)$ and $\phi(v)$ in the roles of u and v. □

p_{12}	aca bacaab aca bacaab aca
p_{13}	abac aacabaca abac aacabaca abac
p_{14}	abaca acabacaabac abaca acabacaabac abaca
p_{15}	acabaca abacaacabacaab acabaca abacaacabacaab acabaca
p_{16}	acabacaab acaacabacaabacabaca acabacaab ...
	... acaacabacaabacabaca acabacaab

Table 3: Modified Padovan words partitioned in the form $uvuvu$.

The next few results will give exact values for the number of runs in $\langle p_k \rangle$ having period P_j for different values of j. For positive integers j other than 1 and 3 we write $N(k, j)$ for the number of runs in $\langle p_k \rangle$ with period P_j. The reason for the omission is that $P_1 = P_2 = 1$ and $P_3 = P_4 = 2$. We will avoid confusion by counting all runs with period 1 in $N(k, 2)$ and all with period 2 in $N(k, 4)$ and setting $N(k, 1) = N(k, 3) = 0$.

Theorem 4.3. *Using the notation from the last paragraph, if $k \geq 12$ and $j \in [4, k-5]$ then*

$$N(k, j) = N(k-3, j) + N(k-2, j). \tag{4.5}$$

Proof. By Lemma 4.2 $\langle p_k \rangle$ may be written $\langle uvuvu \rangle$ where $\langle uvu \rangle = \langle p_{k-2} \rangle$ and $\langle uv \rangle = \langle p_{k-3} \rangle$. To clarify our discussion we label the components of $\langle p_k \rangle$ as $\langle u_1 v_1 u_2 v_2 u_3 \rangle$. Consider a run with period P_j for $j \leq k-5$ which begins in $u_1 v_1$. Since $|u_2 v_2 u_3| = P_{k-2} > 2P_{k-5}$ such a run lies entirely in the linear word $u_1 v_1 u_2 v_2 u_3$ and will therefore

appear in $\langle u_1 v_1 \rangle$. In the other direction, any run in $\langle u_1 v_1 \rangle$ with period not exceeding P_{k-5} will lie entirely in $u_1 v_1 u_1 v_1 u_1$. Thus we have a bijection between the runs with period P_j beginning in $u_1 v_1$ and those in $\langle p_{k-3} \rangle$. The number of such runs is therefore $N(k-3, j)$.

Now consider such runs beginning in $u_2 v_2 u_3$. By similar arguments to those above these lie entirely in $u_2 v_2 u_3 u_1 v_1 u_2$ and will therefore appear in $\langle u_2 v_2 u_3 \rangle = \langle p_{k-2} \rangle$. As in the previous case we get a correspondence between such runs beginning in $u_2 v_2 u_3$ and those in $\langle p_{k-2} \rangle$, so there are $N(k-2, j)$ of them. The total number of runs with period P_j, $j \leq k-5$, in $\langle p_k \rangle$ is then $N(k-3, j) + N(k-2, j)$, as required. □

Corollary 4.4. *Using the notation of the last theorem,*

(a) $N(5,2) = 1$*; for* $k \geq 7$, $N(k,2) = P_{k-7}$*; and for all other values of* k, $N(k,2) = 0$.

(b) $N(k,4) = 0$ *for all values of* k.

(c) $N(8,5) = 1$*; for* $k \geq 10$, $N(k,5) = P_{k-10}$*; and for all other values of* k, $N(k,5) = 0$.

(d) $N(9,6) = 1$*; for* $k \geq 11$, $N(k,6) = P_{k-11}$*; and for all other values of* k, $N(k,6) = 0$.

(e) For $k \geq 9$, $N(k,7) = P_{k-9}$*; and for all other values of* k, $N(k,7) = 0$.

(f) $N(11,8) = 1$*; for* $k \geq 13$, $N(k,8) = P_{k-13}$*; and for all other values of* k, $N(k,8) = 0$.

(g) $N(12,9) = 1$*; for* $k \geq 14$, $N(k,9) = P_{k-14}$*; and for all other values of* k, $N(k,9) = 0$.

Proof. For $k \leq 11$ these statements come from Table 2. For larger values they follow inductively using Theorem 4.3 □

Lemma 4.5. *For* $k \geq 14$

$$\begin{aligned} &(a) \quad N(k, k-4) = 1, \\ &(b) \quad N(k, k-3) = 1, \\ &(c) \quad N(k, k-2) = 1, \\ &(d) \quad N(k, k-1) = 0. \end{aligned}$$

Proof. We prove the four parts of the lemma by induction on k. For values of k from 14 to 19 they are shown in Table 2. Suppose all parts of the lemma hold for indices greater than 13 and less than k where $k \geq 20$ and consider $\langle p_k \rangle$. By the induction hypothesis and Theorem 4.3 we see that $N(k, k-8)$ equals

$$\begin{aligned} & N(k-3, k-8) + N(k-2, k-8) \qquad (4.6) \\ = \; & N(k-6, k-8) + 2N(k-5, k-8) + N(k-4, k-8) \\ = \; & 4. \end{aligned}$$

(a) This is the most technical part of the proof. The reader may find it helpful to draw a small diagram in the margin. As in the proof of Theorem 4.3 we write

$$p_k = u_1v_1u_2v_2u_3.$$

By part (c) of the induction hypothesis $\langle p_{k-2}\rangle = \langle u_2v_2u_3\rangle$ contains one run with period P_{k-4}, so $\langle p_k\rangle$ also contains at least one run with this period. Suppose it contains two runs. The second run cannot be a factor of $uvuuvu$ for then it would also be contained in $\langle p_{k-2}\rangle$. Neither can it lie in $uvuvu$ for then it would lie in $\langle uv\rangle = \langle p_{k-3}\rangle$ which would contradict part (d) of the induction hypothesis. We conclude that either it lies in $u_3u_1v_1u_2v_2$ intersecting both u_3 and v_2 or it lies in $v_1u_2v_2u_3u_1$ intersecting both v_1 and u_1. Either way it follows that uvu has period P_{k-4}. That is,

$$uvu = sts \tag{4.7}$$

where $|st| = P_{k-4}$. Since $\langle uvu\rangle = \langle p_{k-2}\rangle$ this means $|s| + |t| = P_{k-4}$ and $2|s| + |t| = P_{k-2}$ which implies that $|s| = P_{k-2} - P_{k-4} = P_{k-5}$ and $|t| = P_{k-4} - P_{k-5} = P_{k-9}$. Since $|u| = P_{k-7}$ s has border u and therefore has period

$$|s| - |u| = P_{k-5} - P_{k-7} = P_{k-8}.$$

Thus, from (4.7), uv has prefix s with length P_{k-5} and period P_{k-8}, and vu has a suffix with the same properties. Since $|u| > |P_{k-8}|$ the word $v_1u_2v_2$ contains a central factor with period P_{k-8}. The factors v_2u_3 and u_1v_1 also contain copies of s. These cannot coalesce into a single run with period P_{k-8} as then $\langle u\rangle = \langle p_{k-7}\rangle$ would contain a run with this period, contradicting part (d) of the induction hypothesis. Thus $\langle p_k\rangle$ contains at least three runs with period P_{k-8}. If it contained others then by symmetry their number would be even, giving a total odd number of period P_{k-8} runs in $\langle p_k\rangle$. This contradicts (4.6). We conclude that $\langle p_k\rangle$ contains exactly one run with period P_{k-4}.

(b) As in part (a) let $\langle p_k\rangle = \langle uvuvu\rangle$ and $\langle p_{k-3}\rangle = \langle uv\rangle$. Therefore $\langle p_k\rangle$ contains the run $uvuv$ which has period P_{k-3} so $N(k, k-3) \geq 1$.

Suppose there are two runs in $\langle p_k\rangle$ with period P_{k-3}. Write their length $2P_{k-3}$ prefixes as x_1x_2 and y_1y_2 where $x_1 = x_2$ and $y_1 = y_2$. These two prefixes have an intersection of length $4P_{k-3} - P_k$ made up of a component of length m say, which is a prefix of x_1x_2 and a suffix of y_1y_2, and a component of length n say which is a suffix of x_1x_2 and a prefix of y_1y_2. Thus

$$m + n = 4P_{k-3} - P_k. \tag{4.8}$$

If either of these components had length P_{k-3} or more then the whole of the union of x_1x_2 and y_1y_2 would have period P_{k-3} which contradicts our assumption that they are separate runs. Thus m and n are each less than P_{k-3}. Since $x_1 = x_2$ and $y_1 = y_2$ we have $x_1[1\,..\,m] = y_1[P_{k-3} - m + 1\,..\,P_{k-3}]$ and $x_1[P_{k-3} - n + 1\,..\,P_{k-3}] = y_1[1\,..\,n]$. By

Lemma 2.3 each of x_1, x_2, y_1 and y_2 has a border of length $m+n-P_{k-3} = 3P_{k-3}-P_k$ and therefore period $P_k - 2P_{k-3}$ which, by (3.1) and (3.3), equals P_{k-7}. By (4.8)

$$\begin{aligned} m &= (m+n)-n \\ &> 4P_{k-3} - P_k - P_{k-3} \\ &= 2P_{k-3} - P_{k-2} \\ &= P_{k-3} - P_{k-7} \\ &> P_{k-7}. \end{aligned}$$

Thus y_2 and x_1 each have period P_{k-7} and their intersection has length greater than P_{k-7} so their union has period P_{k-7}. Similarly the union of x_2 and y_1 has this period. Since the two unions cover the whole of $\langle p_k \rangle$ we have $N(k, k-7) = 2$. But by Theorem 4.3 and the induction hypothesis

$$\begin{aligned} N(k,k-7) &= N(k-2,k-7) + N(k-3,k-7) \\ &= N(k-4,k-7) + 2N(k-5,k-7) + N(k-6,k-7) \\ &= 1+2+0. \end{aligned}$$

This contradiction completes the proof of part (b).

(c) As in part (b) we see that $\langle p_k \rangle$ contains the run $uvuuvu$ which has period P_{k-2} so $N(k, k-2) \geq 1$. Suppose there are two runs in p_k with period P_{k-2}. The total length of these is at least $4P_{k-2}$ so their overlap has length at least $4P_{k-2} - P_k$. This overlap is made up of two components - one which is a suffix of the first run and a prefix of the second and another which is a prefix of the first and a suffix of the second. The longer of these components has length at least $(4P_{k-2} - P_k)/2$ which is greater than P_{k-2}. Thus the two runs intersect in a factor longer than their common period, which means their union has period P_{k-2} and they are not separate runs.

(d) Let $\langle p_k \rangle = \langle x[1\,..\,P_k] \rangle$ and suppose, for the sake of contradiction, that $x[1\,..\,P_{k-1}] = x[1+P_{k-1}\,..\,2P_{k-1}]$. Then

$$x[1\,..\,2P_{k-1} - P_k] = x[P_k + 1\,..\,2P_{k-1}] = x[P_k - P_{k-1} + 1\,..\,P_{k-1}]$$

so that $x[1\,..\,P_{k-1}]$ has a border of length $2P_{k-1} - P_k$, and so is periodic with period $P_k - P_{k-1} = P_{k-5}$ by (3.3). Thus both $x[1\,..\,P_{k-1}]$ and $x[P_{k-1}+1\,..\,2P_{k-1}]$ have this period. Since the factor $x[1\,..\,2P_{k-1} - P_k]$ is common to these and has length greater than the period we find the whole of $x[P_{k-1}+1\,..\,P_{k-1}]$ has period P_{k-5}. This means there is only one run in p_k with this period. But by parts (b) and (c) and the induction hypothesis $N(k-2, k-5) = N(k-3, k-5) = 1$ so that, by Theorem 4.3, $N(k, k-5) = 2$. This contradiction completes the proof. □

	1	2	3	5	6	8	11	13	19	24	32	37	56	69	93	125	162	218	287	380	505	667
$h(p_1)$	3	3	3	1	1																	
$h(p_2)$	6	5	4	2	1	1	1	1														
$h(p_3)$	8	7	6	2	1	2	1	1														
$h(p_4)$	9	8	7	3	1	2	1	1	1													
$h(p_5)$	14	12	10	4	2	3	1	2	1	1	1											
$h(p_6)$	17	15	13	5	2	4	1	2	1	1	1	1										
$h(p_7)$	23	20	17	7	3	5	2	3	1	2	1	1	1									
$h(p_8)$	31	27	23	9	4	7	2	4	2	2	2	1	1	1								
$h(p_9)$	40	35	30	12	5	9	3	5	2	3	2	2	1	1	1							
$h(p_{10})$	54	47	40	16	7	12	4	7	3	4	3	2	2	1	1	1						
$h(p_{11})$	71	62	53	21	9	16	5	9	4	5	4	3	2	2	1	1	1					
$h(p_{12})$	94	82	70	28	12	21	7	12	5	7	5	4	3	2	2	1	1	1				
$h(p_{13})$	125	109	93	37	16	28	9	16	7	9	7	5	4	3	2	2	1	1	1			
$h(p_{14})$	165	144	123	49	21	37	12	21	9	12	9	7	5	4	3	2	2	1	1	1		
$h(p_{15})$	219	191	163	65	28	49	16	28	12	16	12	9	7	5	4	3	2	2	1	1	1	
$h(p_{16})$	290	253	216	86	37	65	21	37	16	21	16	12	9	7	5	4	3	2	2	1	1	1

Table 4: Numbers of runs of each period in $h(p_k)$ for different periods. The number at the top of each column is the period.

Theorem 4.6. *Using the notation of Theorem 4.3, if $j \geq 10$ then*
(a) for $k \leq j+1$ $N(k,j) = 0$, and
(b) for $k \geq j+2$ $N(k,j) = P_{k-j-1}$.

Proof. Fix a value of j greater than 9. By Lemma 2.2 we have $N(k,j) = 0$ for $k \leq j$. Now consider the case $j = k-1$. From Table 2, $N(11,10) = N(12,11) = N(13,12) = 0$ and by (d) of Lemma 4.5 $N(k,k-1) = 0$ for $k \geq 14$, so (a) holds. The cases $N(12,10)$, $N(13,10)$ and $N(13,11)$ also come from Table 2. Otherwise $k \geq 14$ and we may apply Lemma 4.5. By parts (a), (b) and (c) of that lemma part (b) holds for $k = j+2$, $j+3$ and $j+4$. We prove the rest by induction on k. Suppose the statement holds for all values less than some $k > j+4$. Then by Theorem 4.3

$$N(k,j) = N(k-3,j) + N(k-2,j)$$

which by the induction hypothesis equals

$$P_{k-j-4} + P_{k-j-3} = P_{k-j-1}$$

as required. □

5 Constructing run-rich words

We apply the following morphism to the words p_k.

$$\begin{aligned} h(a) &= 10100101100101001011010 0 \\ h(b) &= 1010010110100 \\ h(c) &= 10100101. \end{aligned}$$

The numbers of runs of each period in $h(p_k)$ for $k \leq 16$ are given in Table 4.

We write H_k for the length of $\langle h(p_k) \rangle$ and $M(k,q)$ for the number of runs in $\langle h(p_k) \rangle$ with period q. For $k \geq 6$ Theorem 3.1 implies that

$$\begin{aligned} H_k &= |h(a)||p_k|_a + |h(b)||p_k|_b + |h(c)||p_k|_c \\ &= 24|p_k|_a + 13|p_k|_b + 8|p_k|_c \\ &= 24P_{k-2} + 13P_{k-6} + 8P_{k-5}. \end{aligned} \tag{5.1}$$

The first few values of H_k are given below.

Lemma 5.1. *For $10 \leq j \leq k-2$*

$$M(k, H_j) \geq P_{k-j-1}.$$

Proof. Each run in p_k with period P_j will produce a run with period H_j in $h(p_k)$, so for all k and j we have $M(k,H_j) \geq N(k,j)$. In Theorem 4.6 we showed that if $j \geq 10$ and $k \geq j+2$ then $N(k,j) = P_{k-j-1}$. □

$$\begin{aligned} H_1 &= 8 \\ H_2 &= 24 \\ H_3 &= 32 \\ H_4 &= 37 \\ H_5 &= 56 \\ H_6 &= 69 \\ H_7 &= 93 \end{aligned}$$

Table 5: Values of H_k.

As well as runs with period H_j, $h(p_k)$ contains runs with other periods resulting from runs within the factors $h(a)$, $h(b)$ and $h(c)$ and concatenations of these factors. In Table 4 we see that there are runs with periods in the set

$$Q = \{1, 2, 3, 5, 6, 11, 13, 19\}.$$

Lemma 5.2. *For $k \geq 12$ and $q \leq H_{k-5}$*

$$M(k, q) = M(k-3, q) + M(k-2, q).$$

We omit the proof of this which is essentially the same as that of Theorem 4.3. The role of $uvuvu$ in that proof is taken by $h(u)h(v)h(u)h(v)h(u)$.

Lemma 5.3. *For $k \geq 14$*

$$\begin{aligned} M(k,1) &\geq 6P_{k-2} + 3P_{k-6} + 2P_{k-5} \\ M(k,2) &\geq 5P_{k-2} + 3P_{k-6} + 2P_{k-5} \\ M(k,3) &\geq 4P_{k-2} + 3P_{k-6} + 2P_{k-5} \\ M(k,5) &\geq P_{k+1} \\ M(k,6) &\geq P_{k-2} \\ M(k,8) &\geq P_k \\ M(k,11) &\geq P_{k-4} \\ M(k,13) &\geq P_{k-2} \\ M(k,19) &\geq P_{k-5} \end{aligned}$$

and for $3 \leq j \leq 9$

$$M(k, H_j) \geq P_{k-j-2}. \tag{5.2}$$

Proof. The cases $k = 14$, 15 and 16 are given in Table 4. For larger k we apply Lemma 5.2 and the theorem follows by induction. □

Theorem 5.4. *For $k \geq 14$ the number of runs in $\langle h(p_k) \rangle$ is at least*

$$5P_{k+3} + 2P_k + 6P_{k-2} - 2. \tag{5.3}$$

Proof. The number of runs in $h(p_k)$ is at least

$$\sum_{q\in Q} M(k,q) + M(k,H_1) + \sum_{j=2}^{k-2} M(k,H_j). \tag{5.4}$$

Note that $H_1 = 8$ is not a member of Q. The first term here, by Lemma 5.3, is at least

$$P_{k+1} + 17P_{k-2} + P_{k-4} + 7P_{k-5} + 9P_{k-6}. \tag{5.5}$$

By application of the identities (3.1), (3.2) and (3.3) we can reduce this to

$$4P_{k+3} + 2P_k + 6P_{k-2}. \tag{5.6}$$

For the second term in (5.4) Lemma 5.3 gives

$$M(k,H_1) = M(k,8) \geq P_k. \tag{5.7}$$

For the last term in (5.4) we use Lemma 5.1 and (5.2) of Lemma 5.3 to obtain

$$\begin{aligned} \sum_{j=2}^{k-2} M(k,H_j) &\geq \sum_{j=2}^{k-2} P_{k-j-2} \\ &= \sum_{j=0}^{k-4} P_j \\ &= P_{k+1} - 2. \end{aligned}$$

The total (5.3) is obtained by summing the bounds for the three terms in (5.4) and simplifying. □

Corollary 5.5.

$$\lim_{n\to\infty} \frac{\rho(n)}{n} \;\geq\; \frac{11\psi^2 + 7\psi - 6}{11\psi^2 + 8\psi - 6} \;>\; 0.94457571235.$$

Proof. Using $P_k \sim \psi^{k+3}/(3\psi^2 - 1)$ from (3.6), (5.1) and the bound in the theorem, the limit of the run density is at least

$$\begin{aligned} & \lim_{k\to\infty} \frac{5P_{k+3} + 2P_k + 6P_{k-2} - 2}{24P_{k-2} + 13P_{k-6} + 8P_{k-5}} \\ = & \lim_{k\to\infty} \frac{5\psi^{k+6} + 2\psi^{k+3} + 6\psi^{k+1} - 2}{24\psi^{k+1} + 13\psi^{k-3} + 8\psi^{k-2}} \\ = & \frac{5\psi^6 + 2\psi^3 + 6\psi}{24\psi + 13\psi^{-3} + 8\psi^{-2}} \\ = & \frac{11\psi^2 + 7\psi - 6}{11\psi^2 + 8\psi - 6}. \end{aligned}$$

The last step uses the definition of ψ : $\psi^3 = \psi + 1$. One notes that the run density in $\langle p_k \rangle$ is the same as that in p_k^ω, from which the statement of the corollary follows. □

6 Discussion

Hideo Bannai has sent us a preprint of [8] which presents a sequence of words which appear, like ours, to approach a run density of 0.9445757... Their construction technique is quite different from ours, and establishing the asymptotic density depends on a fascinating but unproven conjecture.

In this paper we have improved (in the seventh decimal place) the lower bound on $\lim_{n\to\infty} \rho(n)/n$: one wonders where it will all end. The authors of [3] foolishly conjectured that their bound was best possible. We will not be so reckless. They constructed words (henceforth called FSS words) with run density approaching $3/(1+\sqrt{5})$. To do this they devised an iterative technique by inspecting some run-rich words published in [5]. These FSS words have connections with the Fibonacci sequence: the periods of runs in the words are Fibonacci numbers and the asymptotic density may be written $3(1+\phi)/2$ where ϕ is the Golden Ratio. The periods of our words $h(p_k)$ are Padovan numbers and the formula for their asymptotic density (Corollary 5.5) involves the Plastic Number ([13], [15]). Perhaps even richer words exist whose run periods satisfy the recurrence $Q_{n+4} = Q_n + Q_{n+1}$ (following $F_{n+2} = F_n + F_{n+1}$ and $P_{n+3} = P_n + P_{n+1}$ for the Fibonacci and Padovan sequences). The FSS words seem to be optimal for short lengths; our words are better when the length reaches about 125.

Another difference between our words and the FSS words is in their factor complexities. Recall that this is a function $C(n)$ which counts the number of distinct factors with length n. Computational evidence suggests that for the FSS words we have $C(n) \leq 2n$ for all n and $C(n) = 2n$ for given n when the words are sufficiently long. Modified Padovan words have $C(n) \leq 2n+1$ for all n, again with equality with sufficiently long words, and our binary words satisfy $C(n) \leq 2n$ for n in $\{1, 2, 3\}$ and $C(n) \leq 2n + 1$ for larger n.

We have not discussed (unmodified) Padovan words p^*_k which are analogous to Fibonacci words. They are constructed with the morphism $f(a) = b$, $f(b) = c$ and $f(c) = ab$ and satisfy the recurrence $p^*_{k+3} = p^*_k p^*_{k+1}$. They give a run density around 0.53 and if we apply the morphism h to them we get binary words with a run density around 0.90. In comparison the Modified Padovan words have a run density of $\psi^{-4} + \psi^{-7} + \psi^{-11} \approx 0.5908$. This can be proved in the way we proved Theorem 5.4 and its corollary.

Acknowledgements

I thank Hideo Bannai for useful discussions, for publishing our original word on the internet and for providing us with the preprint [6], and Simon Puglisi and Bill Smyth for computational assistance and for useful and entertaining discussions. I also thank N.J.A. Sloane and his Encyclopedia of Integer Sequences [14]. Early in this work I needed to identify the sequence beginning $1, 5, 21, 86, 351, \ldots$ Quick as a flash the Encyclopedia revealed that this is every fifth term in the Padovan sequence. Finally I thank the referee for careful reading and insightful comments.

References

[1] M. Crochemore and L. Ilie, Maximal repetitions in strings, *J. Comput. Syst. Sci.* 74 (2008), 796–807.

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

[3] F. Franek, J. Simpson and W.F. Smyth, The maximum number of runs in a string, in *Proc. 14th Australas. Workshop Combin. Algorithms* (AWOCA), (2003), 26–35.

[4] M. Giraud, Not so many runs in strings, *Proc. LATA 2008* (2008), 245–252.

[5] R. Kolpakov and G. Kucherov, On maximal repetitions in words, *J. Discrete Algorithms* 1 (2000), 159–186.

[6] K. Kusano, W. Matsubara, A. Ishino, H. Bannai and A. Shinohara, New lower bounds for the maximum number of runs in a string, *Proc. Prague Stringology Conference 2008* (2008), 140–145.

[7] M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and its Applications 90, Cambridge, 2002.

[8] W. Matsubara, K. Kusano, H. Bannai and A. Shinohara, A series of run-rich strings, Preprint, 10 pp.

[9] G. Păun, G. Rozenberg and A. Salomaa, *DNA Computing: New Computing Paradigms*, Springer, 1988.

[10] S.J. Puglisi and J. Simpson, The expected number of runs in a word, *Australas. J. Combin.* 42 (2008), 45–54.

[11] S.J. Puglisi, J. Simpson and W.F. Smyth, How many runs can a string contain?, *Theor. Comp. Sc.* 401 (2008), 165–171.

[12] W. Rytter, The number of runs in a string: improved analysis of the linear upper bound, in B. Durand and W. Thomas, eds., *STACS 2006*, no. 3884 in Lec. Notes Comp. Sci., 184–195. Springer-Verlag, Berlin, 2006.

[13] I. Stewart, Mathematical Recreations: Tales of a neglected number, *Scientific American* June 1996, 92–93.

[14] N.J.A. Sloane, (2008), *The On-Line Encyclopedia of Integer Sequences*, `www.research.att.com/~njas/sequences/.`

[15] Wikipedia, *Padovan sequence*, `en.wikipedia.org/wiki/Padovan_sequence.`

(Received 14 Jan 2009)