On Abelian Circular Squares in Binary Words[1]

Aviezri S. Fraenkel[1], Jamie Simpson[2] and Mike Paterson[3]

[1] Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel
http://www.wisdom.weizmann.ac.il/~fraenkel fraenkel@wisdom.weizmann.ac.il
[2] School of Mathematics, Curtin University, Perth WA 6001, Australia;
http://www.cs.curtin.edu.au/~simpson simpson@cs.curtin.edu.au
[3] Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
http://www.dcs.warwick.ac.uk/ msp@dcs.warwick.ac.uk.

To the memory of Paul Erdös and to his living heritage

Abstract. An *abelian square* in a binary word is a pair of adjacent nonempty blocks of the same length, having the same number of 1s. An *abelian circular square* is an abelian square which is possibly wrapped around the word: the tail protruding from the right end of the word reappears at the left end. Two abelian circular squares are *equivalent* if they have the same length and contain the same number of ones. We prove that the longest word with only k inequivalent abelian circular squares contains $4k+2$ bits and has the form $(01)^{2k+1}$ or its complement. Possible connections to tandem repeats in the human genome are pointed out.

1 Introduction

A *binary word* or *binary sequence* is a finite or infinite concatenation of 0s and 1s. An *abelian square* in a binary word w is a pair of adjacent nonempty blocks in w, each having the same number of 1s and each having the same number of 0s. We say that two abelian squares are *equivalent* if they have the same length and contain the same number of 1s. Otherwise they are *inequivalent*. Equivalence is clearly an equivalence relation. The *conjugates* of a binary word are its cyclic permutations. Thus the conjugates of the word 0010 are 0010, 0100, 1000 and 0001.

An *abelian circular square* in a binary word is an abelian square in any of the word's conjugates. Equivalence and inequivalence of abelian circular squares are defined as for abelian squares. For example, the word w=0110110 contains four

[1] A first version of this paper appeared in Proc. 8th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science **1264**, Springer Verlag, Berlin, 1997, pp. 76–82, under the title "On Weak Circular Squares in Binary Words".

inequivalent abelian circular squares: 0110, 011011, 11 and 00. The last occurs in the conjugate 1001101 (and in some others). There are other abelian circular squares, such as 1001 which is equivalent to 0110; and 110110, 110011, both of which are equivalent to 011011. Unless otherwise specified, a *word* means a binary word, and its digits are *bits*. In this note we are interested in the longest word which contains only k inequivalent abelian circular squares. This is equivalent to finding a word of given length such that the number of inequivalent abelian squares appearing in its set of conjugates is minimal. We do not allow abelian circular squares whose total length is greater than that of the word, that is, we do not allow an abelian circular square to overlap itself.

We are interested in finding the longest word which contains only k inequivalent abelian circular squares. Our main result, Theorem 2, states that such words contain $4k+2$ bits, and have the form $(01)^{2k+1}$ or its complement, which is also its reverse.

We will prove Theorem 2 by showing, in Theorem 1, that if a word w contains n 1s, then for every $k \in \mathbb{Z}^+$, $2k \leq n$, one of the conjugates of w must contain an abelian square with $2k$ 1s, that is, w contains an abelian circular square, each half of which contains k 1s. Theorem 1 is based on Lemma 1, whose proof is the only place where we use circularity in an essential way.

For the non-circular case we formulate the following

Conjecture. A longest word containing only k inequivalent abelian squares contains $4k+3$ bits and has one of the forms: $(01)^{2k+1}0$, $0^{2k+1}10^{2k+1}$, or their complements.

2 Background

A *square* is a pair of adjacent nonempty blocks which are identical. Two squares are *equivalent* if one can be translated onto the other. Otherwise they are *inequivalent.* It has been shown many times that there exist infinite *ternary* sequences which are *squarefree*, i.e., they contain no square. See, e.g., Thue [1912], Morse and Hedlund [1944], Hawkins and Mientka [1956], Leech [1957], Novikov and Adjan [1968], Pleasants [1970], Burris and Nelson [1971/72], del Junco [1977], Ehrenfeucht and Rozenberg [1983]. (Currie [1993] wrote: "One reason for this sequence of rediscoveries is that nonrepetitive sequences have been used to construct counterexamples in many areas of mathematics: ergodic theory, formal language theory, universal algebra and group theory, for example…".) Actually, Thue [1912] showed more: there exists a doubly infinite squarefree ternary sequence which also avoids the two triples $a_1a_3a_1$ and $a_2a_3a_2$. See Berstel [1992, §4.2] for an exposition of the full result, and Berstel [1995] for an English translation of Thue's papers.

Every binary word of length 4, however, must already contain a square. Let $g(k)$ be the length of a longest binary word containing at most k inequivalent squares. Then clearly $g(0) = 3$; in fact, the only squarefree words of length 3 are 010 and its 1-complement 101. A computer disclosed that $g(1) = 7$; the only binary words of length 7 with only 1 square are

$$0001000, \qquad 0100010, \qquad 0111011$$

and their 1-complements and the reverse of 0111011 and its 1-complement. Further, $g(2) = 18$; the only binary words of length 18 which contain only 2 distinct squares are 010011000111001101 and its 1-complement (which is also its reverse).

We have seen that $g(0) = 3$, $g(1) = 7$, $g(2) = 18$. This data raises the following natural questions.

1. Is the set of values of the sequence $S = \{g(k) : k = 0, 1, \ldots\}$ infinite or finite?
2. What's the value of $g(3)$?

Erdös [1961, P. 240] wrote: "As far as I know, R. PELTESOHN and SUTHERLAND (unpublished) were the first to construct an infinite sequence formed from the symbols $0, 1, 2$ where no two consecutive blocks were identical. It is easy to see that in a sequence of length four formed from the symbols 0 and 1 two consecutive blocks will be identical. I understand that EUWE proved that in an infinite sequence formed from 0 and 1 there will be arbitrarily large identical consecutive blocks, but that there do not have to be three consecutive identical blocks".

Regarding the first of the above questions, Entringer, Jackson and Schatz [1974] considered the conjecture that S is infinite, citing a reference "which... seems to say that [this] conjecture... is true". This reference is Erdös [1961, P. 240], based, apparently, on the last sentence of the previous paragraph. They then go on to show that S is finite, by proving that $g(5) = \infty$, i.e., there exists an infinite binary sequence with only 5 squares! Fraenkel and Simpson [1995] showed that already $g(3) = \infty$, by constructing an arbitrarily long binary sequence whose only squares are 0^2, 1^2 and $(01)^2$. This is best possible, since $g(2)$ is finite.

Incidentally, it is curious that Erdös [1961] didn't mention Thue [1912] in the first sentence of the above quoted paragraph, nor any of the other early sources of this result. In this context we mention that George Szekeres told one of us (Simpson) that he studied together with Erdös, and that while they were undergraduates, someone asked them if they could produce an infinite squarefree ternary sequence. Szekeres couldn't remember whether they succeeded.

A word is *strongly squarefree* if it has no abelian square. Erdös [1961] has introduced this notion and raised the question of the maximum length of a word on a symbols which is strongly squarefree. Specifically, he wrote, in the paragraph immediately following the above, "Let us now call two consecutive blocks 'identical' if each symbol occurs the same number of times in both of them (i.e. we disregard order). I conjectured that in a sequence of length $2^k - 1$ formed from k symbols there must be two 'identical' blocks. This is true for $k \leq 3$, but for $k = 4$ de BRUIJN and I disproved it and perhaps an infinite sequence of four symbols can be formed without consecutive 'identical' blocks".

Pleasants [1970] constructed an arbitrarily long word on $a = 5$ symbols which is strongly squarefree, and showed that for $a \leq 3$, every sufficiently long word contains an abelian square. Keränen [1992] has shown that Pleasants' $a = 5$ can be replaced by $a = 4$, thus proving the conjecture of Erdös and de Bruijn.

Every binary word of length at least four has an abelian square because it has a square; so it is natural to ask for the longest word having only k abelian squares. In the present paper we take a first step in this direction, by solving this problem for the case of abelian circular squares.

We also remark that questions regarding squares in sequences arise in molecular biology, where they are known as *repeats*, or *tandem repeats*. In fact, the most frequent repeat in the human genome seems to be the word GT, with high *copy number* (the number of times GT is repeated). Only about 10% of the human genome codes for proteins; the rest, called by biologists *junk DNA*, contains the tandem repeats. Biologists are slowly modifying their view as to the appropriateness of the terminology "junk DNA". For example, Trifonov [1989] argues that the copy number influences the functions of DNA chains adjacent to the repeated word, such as their binding power and gene expression; it can even cause certain diseases if too high or too low; and it also influences the unwinding capability of the DNA helix. Algorithms for identifying repeats and databases of repeats in the human genome are maintained by Milosavljević [1998].

Since the copy number at a given site changes from one individual to another, the copy number has also been used in DNA-*fingerprinting*. This application appears to have been originated by Alec Jeffreys' group in Leicester. See, e.g., Jeffreys, Wilson and Thein [1985] and Jeffreys, Turner and Debenham [1991]. Further elaborations on applications of DNA-fingerprinting to medicine and forensic medicine are given in Raskó and Downes [1995 (ch. 6, especially p. 156, and ch. 12, especially pp. 379–380)], where it is also stated that the human genome contains some 500,000 repeated words. (Keywords for human genome applications are VNTR (Variable Number Tandem Repeats) and mini- and microsatellite sequences for the basic subwords that are repeated.)

3 A Lemma

We may represent any word w of length m containing n 1s by the sequence

$$(a_0, a_1, \ldots, a_{n-1}; m), \tag{1}$$

where the j-th digit of w is 1 if and only if $j = a_i$ for some i ($a_i, j \in \{0, \ldots, m-1\}$). Thus the word $w = 100100011$ corresponds to the sequence $(0, 3, 7, 8; 9)$. Since we are concerned with *circular* squares, arithmetic on the a_is is done over the ring $\mathbb{Z}_m$ of integers modulo m, i.e., it is to be interpreted modulo m, with the result in the interval $[0, m-1]$, while arithmetic on the subscripts of the a_is is over $\mathbb{Z}_n$, i.e., it is to be interpreted modulo n, with the result in the interval $[0, n-1]$. The inequalities below involving the a_is are over $\mathbb{Z}_m$, for which we use $\prec$, $\preceq$, $\succ$, $\succeq$. Arithmetic on the indexes is over $\mathbb{Z}_n$.

As a warm-up, we consider the ways in which an abelian circular square with 1s at the *specified locations* $a_{i+1}, \ldots, a_{i+2k}$ $(2k \leq n)$ and in the *specified order* does not exist in a given word w of the form (1). First, it does not exist when the left-most set of k 1s is separated by more 0s than lie in between and around the k 1s to their right. Example: $w = 0100001010101$, where $m = 13$, $n = 5$, $(a_0, a_1, a_2, a_3, a_4) = (1, 6, 8, 10, 12)$. Let $k = 2$, $i = 4$. Then $i + 1 = 0$, and there is no abelian circular square consisting of a_0, a_1, a_2, a_3. However, w clearly contains an abelian circular square with four 1s, e.g., 0101 0101, which consists of a *different* set of a_is. The same situation occurs when "left" and "right" are interchanged. Secondly, it does not exist when there is no 0 to the left of the left-most 1 nor to the right of the right-most 1, and when the total number of 0s between them is odd. Example: w=010011010 with $m = 9$, $n = 4$, $(a_0, a_1, a_2, a_3) = (1, 4, 5, 7)$. Let $k = 2$, $i = 1$. There is no abelian circular square consisting of a_2, a_3, a_0, a_1 in this order. Again we see, however, that w does contain an abelian circular square, e.g., 0110 1001, which consists of the *same* set of a_i but in a different order. We can in fact show that the conditions

(a) $a_{i+k} - a_{i+1} + 1 \preceq a_{i+2k+1} - a_{i+k} - 1$,
(b) $a_{i+k+1} - a_i - 1 \succeq a_{i+2k} - a_{i+k+1} + 1$,
(c) if $a_{i+1} - a_i = 1$, $a_{i+2k+1} - a_{i+2k} = 1$ then $a_{i+2k} - a_{i+1} - 2k + 1$ is even,

are necessary and sufficient for the existence of an abelian circular square with $2k$ 1s at *specified* locations and in a *given* order, but we won't need this fact in the sequel.

Lemma 1. *Given a word w of the form* (1). *Let k be a fixed positive integer in $\mathbb{Z}_n$. Then the inequality*

$$a_{i+k} - a_i \succ a_{i+2k} - a_{i+k} \tag{2}$$

cannot hold for all i in $\mathbb{Z}_n$. Similarly, (2) *with the inequality sign reversed, namely*

$$a_{i+k} - a_i \prec a_{i+2k} - a_{i+k}, \tag{3}$$

cannot hold for all i in $\mathbb{Z}_n$.

Proof. Suppose that (2) holds for all i. Iterating this inequality with i replaced by $i + k$, we get the infinite chain,

$$a_{i+k} - a_i \succ a_{i+2k} - a_{i+k} \succ a_{i+3k} - a_{i+2k} \succ \ldots .$$

In any block of length $n+1$ of this chain there must be two identical terms, which is a contradiction. This proves the first statement of the lemma. The second part is proved in the same way with the inequalities reversed. $\square$

For the noncircular case Lemma 1 fails — the chain we used in the proof is no longer infinite, which precipitated the contradiction.

4 Two Theorems

We use Lemma 1 to prove our first theorem.

Theorem 1. *Let $k \geq 1$ and $n \geq 2k$. Then there exists some $i \in \mathbb{Z}_n$ such that the word* (1) *contains an abelian circular square using the* $2k$ 1*s at positions* $a_{i+1}, \ldots, a_{i+2k}$ *in this order.*

Proof. If there is some i such that (2) becomes an equality, i.e., $a_{i+k} - a_i = a_{i+2k} - a_{i+k}$, we are done, since the left-hand side is the number of bits in the semi-closed interval $I_1 = (a_i, a_{i+k}]$, and the right-hand side is the number of bits in $I_2 = (a_{i+k}, a_{i+2k}]$, so we have the abelian circular square consisting of the $2k$ 1s at $a_{i+1}, \ldots, a_{i+2k}$, with k of them in I_1 and the other k in I_2. So we assume that there is no such i. Then for every i, either (2) holds or (3) holds. By Lemma 1 we may thus assume, without loss of generality, that there is some i for which (2) holds, but when we replace i by $i+1$, then the reverse inequality, namely,

$$a_{i+k+1} - a_{i+1} \prec a_{i+2k+1} - a_{i+k+1} \tag{4}$$

holds. Let $I_3 = [a_{i+1}, a_{i+k+1})$ and $I_4 = [a_{i+k+1}, a_{i+2k+1})$. (See Figure 1.)

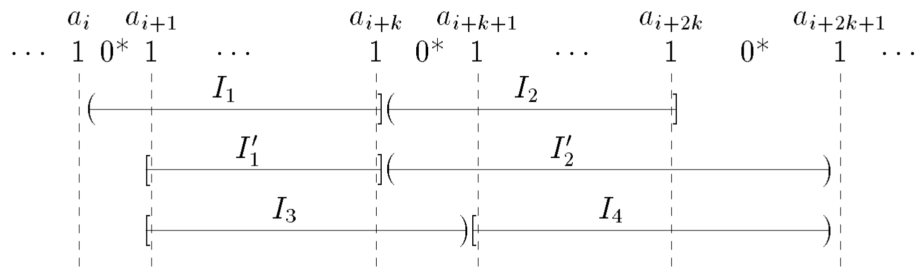

Figure 1.

By (2), $|I_1| > |I_2|$. We can slide the left (open) boundary of I_1 to the right, bit by bit, up to the 0 to the left of a_{i+1}, thus reducing the size of I_1, without changing the number of 1s it contains. If at any stage equality holds, we are done. Similarly, we may slide the right (closed) boundary of I_2 to the right, without changing the number of 1s, thus increasing the size of I_2. If there is never equality, we have even $|I_1'| > |I_2'|$, where $I_1' = [a_{i+1}, a_{i+k}]$ is closed, and $I_2' = (a_{i+k}, a_{i+2k+1})$ is open. Since $|I_3| \geq |I_1'|$ and $|I_2'| \geq |I_4|$, we therefore have $|I_3| > |I_4|$, which contradicts (4). Thus equality must hold at some stage, so w contains an abelian circular square. $\square$

Theorem 2. *For any $k \geq 0$, a longest word containing only k inequivalent abelian circular squares has length $4k+2$ and has the form $(01)^{2k+1}$ or its complement.*

Proof. It is clear that the word $(01)^{2k+1}$ and its complement each have length $4k+2$ and contain precisely k abelian circular squares. We must show that there does not exist a longer word containing only k abelian circular squares, and that $(01)^{2k+1}$ and its complement are the only words of length $4k+2$ containing k abelian circular squares.

Given any word of length at least $4k+3$, without loss of generality suppose it contains at least as many 1s as 0s. Then it contains at least $2k+2$ 1s. By Theorem 1 it then contains abelian circular squares with $2i$ 1s for each value of i from 1 to $k+1$. Thus it contains at least $k+1$ abelian circular squares. It follows that the longest word containing only k abelian circular squares has length $4k+2$.

Now suppose that we have a word of length $4k+2$ containing k abelian circular squares. Again we may assume that it contains at least as many 1s as 0s, and by an argument similar to the above conclude that it has exactly $2k+1$ 1s, $2k+1$ 0s and abelian circular squares with $2i$ 1s for each value of i from 1 to k. If the word contains 2 adjacent 0s then it will also contain the abelian square 00, so this cannot occur. Interchanging 0s and 1s does not alter the number of abelian squares, so we conclude that the word does not contain adjacent 1s or adjacent 0s. It must therefore be $(01)^{2k+1}$ or its complement as required. $\square$

References

1. J. Berstel [1992], Axel Thue's work on repetitions in words, in: *Séries Formelles et Combinatoire Algébrique* (P. Leroux and C. Reutenauer, eds.), Publ. du LaCIM, Vol. 11, Université de Québec à Montréal, Canada, pp. 65-80.
2. J. Berstel [1995], Axel Thue's papers on repetition in words: an English translation, Publ. du LACIM, Vol. 20, Université de Québec à Montréal, Canada.
3. S. Burris and E. Nelson [1971/72], Embedding the dual of π_∞ in the lattice of equational classes of semigroups, *Algebra Universalis* **1**, 248–153.
4. J.D. Currie [1993], Open problems in pattern avoidance, *Amer. Math. Monthly* **100**, 790–793.
5. A. del Junco [1977], A transformation with simple spectrum which is not rank one, *Canad. J. Math.* **29**, 655–663.
6. A. Ehrenfeucht and G. Rozenberg [1983], On the separating power of EOL systems, *RAIRO Inform. Théor.* **17**, 13–22.
7. R. Entringer, D. Jackson and J. Schatz [1974], On nonrepetitive sequences, *J. Combin. Theory* (Ser. A) **16**, 159–164.
8. P. Erdös [1961], Some unsolved problems, *Magyar Tud. Akad. Mat. Kutato. Int. Kozl.* **6**, 221–254.
9. A.S. Fraenkel and R.J. Simpson [1995], How many squares must a binary sequence contain? *Electronic J. Combinatorics* **2** (1995) R2, 9pp. http://ejc.math.gatech.edu:8080/Journal/journalhome.html
10. D. Hawkins and W.E. Mientka [1956], On sequences which contain no repetitions, *Math. Student* **24**, 185–187.
11. A.J. Jeffreys, M. Turner and P. Debenham [1991], The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework, *Am. J. Hum. Genet.* **48**, 824–840.

12. A.J. Jeffreys, V. Wilson and S.L. Thein [1985], Hypervariable 'minisatellite' regions in human DNA, *Nature* **314**, 67–73.
13. A.J. Jeffreys, V. Wilson and S.L. Thein [1985], Individual-specific 'fingerprints' of human DNA, *Nature* **316**, 76–79.
14. V. Keränen [1992], Abelian squares are avoidable on 4 letters, Automata, Languages and Programming: Lecture Notes in Computer Science 623, 41–52, Springer Verlag.
15. J.A. Leech [1957], A problem on strings of beads, *Math. Gaz.* **41**, 277–278.
16. A. Milosavljević [1998], Repeat Analysis, Ch. 13, Sect. 4, Imperial Cancer Research Fund Handbook of Genome Analysis, Blackwell Scientific Publications.
17. M. Morse and G.A. Hedlund [1944], Unending chess, symbolic dynamics and a problem in semigroups, *Duke Math. J.* **11**, 1–7.
18. P.S. Novikov and S.I. Adjan [1968], Infinite periodic groups I, II, III, *Izv. Akad. Nauk. SSSR Ser. Mat.* **32**, 212–244; 251–524; 709–731.
19. P.A.B. Pleasants [1970], Non-repetitive sequences, *Proc. Cambridge Phil. Soc.* **68**, 267–274.
20. I. Raskó and C.S. Downes [1995], *Genes in Medicine: Molecular Biology and Human Genetic Disorders*, Chapman & Hall, London.
21. A. Thue [1912], Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr., I. Mat. Nat. Kl. Christiania* **I**, 1–67.
22. E.N. Trifonov [1989], The multiple codes of nucleotide sequences, *Bull. Math. Biology* **51**, 417-432.

This article was processed using the LaTeX macro package with LLNCS style