The Number of Positions Starting a Square in Binary Words

Tero Harju
Department of Mathematics
University of Turku, Finland
harju@utu.fi

Tomi Kärki
Department of Mathematics
University of Turku, Finland
topeka@utu.fi

Dirk Nowotka
Institute for Formal Methods in Computer Science (FMI)
Universität Stuttgart, Germany
nowotka@fmi.uni-stuttgart.de

Submitted: Sep 3, 2010; Accepted: Dec 14, 2010; Published: Jan 5, 2011
Mathematics Subject Classification: 68R15

Abstract

We consider the number $\sigma(w)$ of positions that do not start a square in binary words w. Letting $\sigma(n)$ denote the maximum of $\sigma(w)$ for length $|w| = n$, we show that $\lim \sigma(n)/n = 15/31$.

1 Square-free positions and strong words

Every binary word with at least 4 letters contains a square. A.S. Fraenkel and J. Simpson [2, 1] studied the number of distinct squares in binary word; see also Ilie [4], where it was shown that a binary word can contain at most $2n - \Theta(\log n)$ distinct squares. It has been conjectured that n is an upper bound in this case.

On the other hand, in an impressive paper [5] G. Kucherov, P. Ochem and M. Rao proved that the minimum number of occurrences of squares in binary words is asymptotically equal to $0.55080\ldots$ times the length of the word. Later Ochem and Rao [7] showed that this constant is exactly $103/187$.

In the present paper we count the minimum number of positions in binary words that starts a square, and we show that asymptotically this is $16/31 = 0.516\ldots$. For our convenience, we state the result in the dual case, i.e., we count the maximum number of positions that are square-free. Related question for borders of cyclic words was considered by T. Harju and D. Nowotka [3].

Several parts of the proofs are computer aided, both for searching the strong words (the main concept in the proofs) as well as for checking their compatibilities. We have included the Mathematica code for the search of strong words.

We refer to Lothaire [6] for elementary definitions in combinatorics on words. Let $A = \{a, b, c\}$ be a ternary alphabet, and $B = \{0, 1\}$ a binary alphabet. For a binary word $w = a_1a_2 \cdots a_n \in B^*$ with $a_i \in B$, we say that a position $i \in \{1, 2, \ldots, n\}$ *starts a square*, if $a_i \cdots a_{i+j-1} = a_{i+j} \cdots a_{i+2j-1}$ for some j such that $i + 2j - 1 \leq n$. Otherwise, the position i is *square-free* in w.

For $r, s \geq 1$, let $\sigma_w(r, s)$ denote the number of square-free positions i with $r < i \leq r+s$ in the word w. In order to simplify the treatment, we shall write $\sigma_w(u)$ instead of $\sigma_w(r, s)$ where $w = xuv$ such that $|x| = r$ and $|u| = s$. Hence while talking about $\sigma_w(u)$ the occurrence of the factor u in w will be implicitly, and without risk of confusion, assumed. Also, let $\sigma(w) = \sigma_w(w)$. For an integer $n \geq 1$, let

$$\sigma(n) = \max\{\sigma(w) : w \in B^*, |w| = n\} .$$

A word w is said to be *strong* if for all nonempty prefixes u of w,

$$\sigma_w(u) \geq |u|/2 .$$

We notice that if w is a strong word, then so is its *complement* $\bar{w}$ obtained from w by interchanging the letters 0 and 1.

Example 1. The short strong words, beginning with 0, are listed in Table 1. As an example consider the word $w = 0100110001001$ with $|w| = 13$. We have $\sigma(w) = 8$, and the square-free positions are marked by dots in the following copy $w = .0.10.01.100.0.10.0.1$. The ratio $8/13$ is much bigger than the asymptotic bound $15/31$ that will be proved in the sequel. One can easily check that w is a strong word. $\square$

0	0110	010001	0100110	01001100	010011000
01	01000	010011	0100111	01001101	010011010
010	01001	011001	0110010	01001110	010011100
011	01100	0100010	0110011	010001100	010011101
0100	01101	0100011	01000110	010001101	0100011001

Table 1: The first 30 short strong words.

Using *Mathematica* (version 7.01.0), one can calculate $\sigma(w)$ and the ratio $\sigma(w)/|w|$ using functions `Sigma` and `SigmaRatio` defined as

```
Sigma[Str_]:= StringLength[Str]-
Length[StringPosition[Str,x__ ~~x__,Overlaps -> True]] ,
```

```
SigmaRatio[Str_,j_]:= (j - Length[Select[StringPosition[Str,
x__ ~~x__, Overlaps -> True], #[[1]] < j + 1 &]])/j .
```

For checking whether a word is strong, one can use

```
Strong[Str_] :=Module[{strong, i}, strong = True; i = 0;
While[strong && i < StringLength[Str], i = i + 1;
strong = (SigmaRatio[Str, i] >= 1/2)]; strong].
```

A list of all strong words can be generated by the command

```
StrongList = {"0", "1"}; For[i = 1, i < Length[StrongList],
i++, If [Strong[StrongList[[i]] <> "0"], StrongList =
Append[StrongList, StrongList[[i]] <> "0"]];
If [Strong[StrongList[[i]] <> "1"], StrongList =
Append[StrongList, StrongList[[i]] <> "1"]]];
StrongList.
```

After a computer check, we have that there are only finitely many strong words, the longest of which have length 37. More precisely, we have the following lemma.

Lemma 1. *(1) There are 382 strong words the longest of which has length 37.*

(2) If w is a strong word with $|w| \geq 8$, then w begins with 0100 *or its complement* 1011*.*

The long strong words of length at least 27, starting with the letter 0, are in Table 2.

2 Decompositions

A *min-factor* $m(w)$ of a binary word w is the shortest prefix u of w such that $\sigma_w(u) < |u|/2$, if it exists. By the above observation, each binary word w with $|w| \geq 38$ does have a (unique) min-factor. The *min-decomposition* of w is the factorization $w = w_1 w_2 \cdots w_r w_{r+1}$, where $w_i = m(w_i \cdots w_{r+1})$ for $i = 1, 2, \ldots, r$ and the suffix w_{r+1} does not possess a min-factor. In particular, w_{r+1} is strong.

The following lemma will be crucial in the sequel.

Lemma 2. *Assume that $w = m(w)w'$ for a suffix w' with* 010 *or* 101 *a prefix of w'. Then the min-factor $m(w)$ is a strong word.*

Proof. In order to show that $m(w)$ is strong, consider the prefix p of length $|m(w)| - 1$. Then

$$\sigma_w(p) = \sigma_w(m(w)), \tag{1}$$

since w' begins with 010 or 101, and thus the last letter of $m(w)$ starts a square in w. By the definition of $m(w)$, we have $\sigma_w(m(w)) < |m(w)|/2$ and $\sigma_w(p) \geq |p|/2$. Hence, combining these with (1), we obtain

$$(|m(w)| - 1)/2 \leq \sigma_w(m(w)) < |m(w)|/2,$$

length	strong word
27	010011000100111011000100110
	010011000100111011001011100
	010011000100111011001011101
	010011000100111011001110010
	010011101100010011010001100
	010011101100010011010001101
28	0100110001001110110001001100
	0100110001001110110001001101
	0100110001001110110010111001
	0100111011000100110100011001
29	01001100010011101100010011000
	01001100010011101100010011010
	01001100010011101100101110010
	01001100010011101100101110011
	01001110110001001101000110010
	01001110110001001101000110011
30	010011000100111011000100110001
	010011000100111011000100110100
	010011000100111011001011100110
31	0100110001001110110001001100011
	0100110001001110110001001101000
	0100110001001110110001001101001
	0100110001001110110010111001100
	0100110001001110110010111001101
32	01001100010011101100010011000110
	01001100010011101100010011010001
33	010011000100111011000100110001101
	010011000100111011000100110100010
	010011000100111011000100110100011
34	0100110001001110110001001101000110
35	01001100010011101100010011010001100
	01001100010011101100010011010001101
36	010011000100111011000100110100011001
37	0100110001001110110001001101000110010
	0100110001001110110001001101000110011

Table 2: The long strong words.

which implies that $|m(w)|$ is odd and $\sigma_w(m(w)) = (|m(w)| - 1)/2$. Hence, since the last letter of $m(w)$ does not start a square in $m(w)$, we have

$$\sigma(m(w)) \geq \sigma_w(m(w)) + 1 = (|m(w)| + 1)/2 \,.$$

This completes the proof that $m(w)$ is strong. ☐

3 Asymptotic behaviour

In this section we consider the asymptotic behaviour of $\sigma(n)/n$, and prove the following result as a consequence of Theorems 7 and 9.

Theorem 3. *We have*

$$\lim \frac{\sigma(n)}{n} = \frac{15}{31} \,.$$

3.1 Upper bound

In the next lemmas, let

$$w = w_1 w_2 \cdots w_r w_{r+1} \tag{2}$$

be a min-decomposition of w for $r \geq 2$.

Lemma 4. *Each min-factor w_i, for $i = 1, 2, \ldots, r$, is of odd length.*

Proof. Assume that w_i is a min-factor of even length n. Let v be the prefix of w_i of length $n - 1$. Then

$$\sigma_w(v) \leq \sigma_w(w_i) \leq \frac{n}{2} - 1 = \frac{n-2}{2} < \frac{n-1}{2} \,,$$

which contradicts with the definition of a min-factor. ☐

Lemma 5. *Let $i < r$. If $|w_{i+1}| \geq 9$ then w_i is strong.*

Proof. Since w_{i+1} is a min-factor, by the definitions, its prefix of length $|w_{i+1}| - 1$ is a strong word. Each strong word of length at least eight begins with 010 or 101, and thus the claim follows from Lemma 2. ☐

The next lemma relies on computations.

Lemma 6. *If $|w_i| = 27$ and $|w_{i+1}| \geq 31$ for $i < r$, then w_i is one of the following two strong words,*

$$010011000100111011000100110 \quad \textit{or} \quad 101100111011000100111011001 \,.$$

Theorem 7. *We have*

$$\limsup \frac{\sigma(n)}{n} \leq \frac{15}{31} \,.$$

Proof. Let $w = w_1 w_2 \cdots w_r w_{r+1}$ be the min-decomposition of w. Recall that, for $i \leq r$, we have $\sigma_w(w_i) < |w_i|/2$, and that the prefix of length $|w_i|-1$ is strong whenever $|w_i| > 1$. Also, by Lemma 4, $|w_i|$ is odd for each $i \leq r$. We consider the factors

$$w_{i,i+k} = w_i w_{i+1} \ldots w_{i+k}\,,$$

where $i + k \leq r$. By symmetry, we can assume that in these considerations w_i begins with the letter 0. The other case is obtained by complementing the words in the following considerations.

Claim. For all $i \leq r-3$, we have $\sigma_w(w_{i,i+k})/|w_{i,i+k}| \leq 15/31$ for some $0 \leq k \leq 2$.

The claim leaves (some of the) suffixes $w_{r-2}w_{r-1}w_r w_{r+1}$ unconsidered. However, since these suffixes are always bounded by length, the claim of the theorem follows.

For the present claim , we obtain the following facts aided by computer checks.

For each index $j < r$, if $|w_{j+1}| > 29$, then the word $p = 01001100010011$ (or, in the symmetric case, its complement $\bar{p}$) is a prefix of w_{j+1}. Indeed, if $|w_{j+1}| > 29$, then $w_{j+1} \geq 31$ by Lemma 4, and its prefix of length 30 is strong. By Table 2, every strong word of length 30 has the prefix p or $\bar{p}$. By Lemma 2, w_j is strong, and after a computer check, we find that if $|w_j| \geq 25$ then w_j must be one of the words in Table 3, where the lengths of the words are at most 31. Therefore

$$\text{if } |w_{j+1}| > 29, \text{ then } |w_j| \leq 31\,. \tag{3}$$

Hence, by the definition of a min-factor, we have

$$\sigma_w(w_{j,j})/|w_{j,j}| \leq 15/31.$$

We also find by checking through the strong words of length 29, with the condition that w_j is a min-factor, that

$$\text{if } |w_j| = 29 \text{ with } j < r \text{ and } \sigma_{w_{j,j+1}}(w_j) \geq 14, \text{ then } |w_{j+1}| \leq 29\,. \tag{4}$$

Suppose then that $|w_i| > 31$ for $i \leq r-3$, and that, for all $k = 1, \ldots, r-i$,

$$\frac{\sigma_w(w_{i,i+k})}{|w_{i,i+k}|} > \frac{15}{31}. \tag{A}$$

In particular, by (A) and Lemma 5, the factor w_i is strong. Moreover, by (3), we have $|w_{i+1}| \leq 29$. If $|w_i| = 33$, then $\sigma_w(w_{i,i+1})/|w_{i,i+1}| \leq (16+14)/(33+29) = 15/31$, which contradicts with the assumption (A). Hence, we have $|w_i| = 35$ or 37.

First, let $|w_i| = 35$. By the assumption (A), we have to have $|w_{i+1}| = 29$ and $\sigma_w(w_{i+1}) = 14$. By (4), since $i \leq r-2$, also $|w_{i+2}| \leq 29$. But now,

$$\frac{\sigma_w(w_{i,i+2})}{|w_{i,i+2}|} \leq \frac{17+14+14}{35+29+29} = \frac{15}{31}\,.$$

Second, let $|w_i| = 37$. Then, by (A), we have $|w_{i+1}| = 27$ or 29. Since $i \leq r-3$, the case $|w_{i+1}| = 29$ leads to a contradiction. Namely, by (A) and (4), we must have $|w_{i+2}| \leq 29$. If $|w_{i+2}| \leq 27$, then

$$\frac{\sigma_w(w_{i,i+2})}{|w_{i,i+2}|} \leq \frac{18+14+13}{37+29+27} = \frac{15}{31}$$

contradicts with (A). On the other hand, if $|w_{i+2}| = 29$, then as above $|w_{i+3}| \leq 29$ and

$$\frac{\sigma_w(w_{i,i+3})}{|w_{i,i+3}|} \leq \frac{18+14+14+14}{37+29+29+29} = \frac{15}{31}\,.$$

This is again a contradiction.

Hence, it follows that we have the factor $w_i w_{i+1}$ with $|w_i| = 37$ and $|w_{i+1}| = 27$. In this case, the computer search finds that there is a unique solution for w_i,

$$w_i = 0100110001001110110001001101000110010$$

starting with 0, and w_{i+1} is one of the following two words of length 27,

$$w_{i+1} = 101100010011101100101110011\,, \tag{i1}$$
$$w_{i+1} = 101100010011101100101110010\,. \tag{i2}$$

These words differ from those in Lemma 6 which means $|w_{i+2}| \leq 29$, and

$$\frac{\sigma_w(w_{i,i+2})}{|w_{i,i+2}|} \leq \frac{18+13+14}{37+27+29} = \frac{15}{31}\,.$$

Again, this is a contradiction, and the claim follows. $\square$

length	strong word
25	0100110001001110110010111
25	1011001110110001001110110
25	1011001110110001001101000
25	1011001110110001001100011
27	101100111011000100111011001
31	0100110001001110110001001100011
31	0100110001001110110001001101000
31	1011001110110001001110110010111

Table 3: The set of strong words of length at least 25 preceding the word p = 01001100010011. Notice that as starting letters 0 and 1 are not symmetric, because of the chosen p. Also, there are no words in this list of length 29.

Example 2. In the previous proof for the unique min-factor w_i with $|w_i| = 37$ where $i = r - 2$, the computer search states that w_{i+1} is equal to either of the following words

$$1011000100111011001011100110 1\,,$$
$$1011000100111011001011100110 0\,.$$

The first one has no continuation, but for the second one, we have two candidates for w_{i+2} to be a min-factor. These are

$$0100111011000100110100011001 0\,,$$
$$0100111011000100110100011001 1\,.$$

$\square$

3.2 Lower bound

For the lower bound we construct good words from square-free ternary words using the following morphism. Let $h\colon \{\alpha, \beta, \bar\alpha, \bar\beta\}^* \to \{0,1\}^*$ be the 31-uniform morphism defined by

$$\begin{aligned} h(\alpha) &= 0100110001001110110001001101000\,,\\ h(\beta) &= 0100110001001110110001001100011\,,\\ h(\bar\alpha) &= 1011001110110001001110110010111\,,\\ h(\bar\beta) &= 1011001110110001001110110011100\,. \end{aligned}$$

We have $\sigma_{h(xy)}(h(x)) = 15 = \sigma(h(x)) - 1$ for all different $x, y \in \{\alpha, \beta, \bar\alpha\}$ except for $xy = \beta\bar\alpha$. Taking the complements, we have $\sigma_{h(xy)}(h(x)) = 15 = \sigma(h(x)) - 1$ for all $x, y \in \{\alpha, \bar\beta, \bar\alpha\}$ except for $xy = \bar\beta\alpha$.

Take then a square-free ternary word w on the alphabet $\{\alpha, \beta, \bar\alpha\}$ and change every occurrence of $\beta\bar\alpha$ by $\bar\beta\bar\alpha$. Denote the new square-free word on the alphabet $\{\alpha, \beta, \bar\alpha, \bar\beta\}$ by $\hat w$. We show that the words $h(\hat w)$ satisfy $\sigma(h(\hat w))/|h(\hat w)| > 15/31$. Let us first prove the following lemma.

Lemma 8. *There are no squares u^2 in $h(\hat w)$ such that $|u| \geq 31$.*

Proof. Suppose on the contrary that there is a square u^2 in $h(\hat w)$ where $|u| \geq 31$. Since $h(\hat w)$ consists of *blocks* $h(\alpha), h(\beta), h(\bar\alpha), h(\bar\beta)$ of length 31, we can write

$$u = xvy = x'v'y'\,, \tag{5}$$

where $x \neq \varepsilon$ is the prefix of the first u up to the beginning of a new block, $v = h(r)$ consists of full blocks, y is a prefix of the block following v such that $|y| < 31$ and $x'v'y'$ is the corresponding block decomposition for the second occurrence of u, denoted by u' in the sequel. Note that x and x' may be full blocks, and some or all of v, y, v', y' may be empty, and the corresponding elements in the two decompositions can be of different length. Moreover,

$$h(z) = yx' \tag{6}$$

for some letter $z \in \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}$.

(1) Assume $|x| \geq 5$. We notice that the word 01000 (resp. $00011, 10111, 11100$) occurs in $h(\hat{w})$ only as a suffix of $h(\alpha)$ (resp., $h(\beta), h(\bar{\alpha}), h(\bar{\beta})$). Since x is a prefix of $u = u'$ and also a suffix of some block, we conclude that $x' = x$, $v' = v$ and $y' = y$. Hence, $x' = x$ determines y and z uniquely, and the word $xv(yx')v$ is preceded by y. In other words, $(yx)v(yx')v = h(zrzr)$ must occur in $h(\hat{w})$. By the block decomposition (5), this implies that $zrzr$ is a factor of $\hat{w}$, which contradicts with the square-freeness of $\hat{w}$.

(2) Assume $|x| < 5$. Since $|u| \geq 31$, we have $|vy| \geq 27$. Hence, v contains a prefix 01001100010 or its complement. We notice that 01001100010 (resp. 10110011101) occurs in $h(\hat{w})$ only as a prefix of the block $h(\alpha)$ or $h(\beta)$ (resp. $h(\bar{\alpha})$ or $h(\bar{\beta})$). Hence, we conclude that in u' we must have $x' = x$, $v' = v$ and $y' = y$.

If $|y| \geq 28$, then $y = y'$ determines x' and z uniquely and $v(yx')v(y'x') = h(rzrz)$ is a factor of $h(\hat{w})$. We obtain a contradiction as above.

On the other hand, if $|y| < 28$, then $|x'| \geq 4$ by (6). A suffix $x' = x$ of any block with length at least four determines the block uniquely. Hence, the word $(yx)v(yx')v = h(zrzr)$ is a factor of $\hat{w}$. Again, this is a contradiction. □

Now we are ready to prove the lower bound.

Theorem 9. *We have*

$$\liminf \frac{\sigma(n)}{n} \geq \frac{15}{31}.$$

Proof. Let $\hat{w}$ be as in the previous proof obtained from a square-free ternary word w. Each square u^2 in $h(\hat{w})$ satisfies $|u| < 31$, and thus u^2 must occur inside $h(xyz)$ for some factor $xyz \in \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}^3$ in $\hat{w}$. However, we verify by a computer check that

$$\sigma_{h(xyz)}(h(x)) = 15 \tag{7}$$

for all factors xyz of $\hat{w}$. Hence, combining (7) with Lemma 8, we conclude that $\sigma_{h(\hat{w})}(h(x)) = \sigma(h(x)) - 1 = 15$ for every $x \in \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}$, which proves the claim. □

Acknowledgement. Tomi Kärki acknowledges the support of Magnus Ehrnrooth Foundation.

References

[1] A. S. Fraenkel and J. Simpson. How many squares can a string contain? *J. Combin. Theory Ser. A*, 82(1):112–120, 1998.

[2] A. S. Fraenkel and R. J. Simpson. How many squares must a binary sequence contain? *Electron. J. Combin.*, 2:R2, 1995.

[3] T. Harju and D. Nowotka. Border correlation of binary words. *J. Combin. Theory Ser. A*, 108(2):331–341, 2004.

[4] L. Ilie. A note on the number of squares in a word. *Theoret. Comput. Sci.*, 380(3):373–376, 2007.

[5] G. Kucherov, P. Ochem, and M. Rao. How many square occurrences must a binary sequence contain? *Electron. J. Combin.*, 10:R12, 2003.

[6] M. Lothaire. *Combinatorics on words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997.

[7] P. Ochem and M. Rao. Minimum frequencies of occurrences of squares and letters in infinite words. In *Mons Days of Theoretical Computer Science*, Mons, August 2008.