RAIRO-Theor. Inf. Appl. 55 (2021) 3
https://doi.org/10.1051/ita/2021001

RAIRO - Theoretical Informatics and Applications
www.rairo-ita.org

THE SIMPLEST BINARY WORD WITH ONLY THREE SQUARES[*]

Daniel Gabric and Jeffrey Shallit[**]

Abstract. We re-examine previous constructions of infinite binary words containing few distinct squares with the goal of finding the "simplest", in a certain sense. We exhibit several new constructions. Rather than using tedious case-based arguments to prove that the constructions have the desired property, we rely instead on theorem-proving software for their correctness.

Mathematics Subject Classification. 68R15.

Received July 27, 2020. Accepted February 19, 2021.

1. Introduction

One of the earliest results in combinatorics on words is that squares are unavoidable over a two-letter alphabet, but are avoidable over a three-letter alphabet [4, 15, 16]. Here a "square" is a nonempty word of the form xx, "unavoidable" means that every sufficiently long word contains a square subword, and "avoidable" means there exists an infinite word containing no squares.

Although squares are unavoidable over a two-letter alphabet, Entringer, Jackson, and Schatz [8] proved that there exist infinite binary words containing no squares of order ≥ 3. (The order of a square xx is $|x|$, the length of x.) This was later improved by Fraenkel and Simpson; they showed the existence of binary words having only three distinct squares.

The main tool for creating such words is the *morphism*: a map $h : \Sigma^* \to \Delta^*$ for alphabets Σ, Δ obeying the rule $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma^*$. A morphism is k-*uniform* if $|h(a)| = k$ for all $a \in \Sigma$. If it is k-uniform for some k, then we say it is *uniform*. A 1-uniform morphism is called a *coding*. If $\Delta \subseteq \Sigma$ we can iterate h, writing $h^2(x)$ for $h(h(x))$, and so forth. If further $h(a) = ax$ for some $a \in \Sigma$, $x \in \Sigma^*$, and $h^i(x) \neq \epsilon$ for all i, then iterating h infinitely produces an infinite word $h^\omega(a) = axh(x)h^2(x)\cdots$ called a *fixed point* of h. If an infinite word is the image, under a coding, of a fixed point of a k-uniform morphism, it is called k-*automatic*. The *weight* of a morphism $h : \Sigma^* \to \Sigma^*$ is defined to be $\sum_{a\in\Sigma} |h(a)|$, and the weight of a k-automatic infinite word is defined to be the weight of its defining morphism.

In this note we find the "simplest" infinite binary word having at most three distinct squares. Our criterion for simplicity is as follows:

(a) the word should be generated by a finite automaton of s states taking the base-k representation of n as input (*i.e.*, a k-automaton), most significant digit first; and
(b) the product $k \cdot s$ should be as small as possible.

[*] *Research of the second author is supported by NSERC grant 2018-04118.*
School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1 Canada.

[**] Corresponding author: shallit@uwaterloo.ca

Article published by EDP Sciences

By Cobham's theorem [7], this is same as saying the word is generated as the image, under a coding, of a fixed point of a k-uniform morphism over an alphabet of s letters.

One practical advantage to restricting our attention to k-automatic words is that the property of having exactly three distinct square factors can be stated in first-order logic, thus reducing the verification to a completely routine calculation using a decision procedure [6].

2. The Entringer-Jackson-Schatz contruction

We begin with a description of the construction of Entringer-Jackson-Schatz. Here very slightly modified from the original, it starts with an arbitrary squarefree word $\mathbf{z}$ over $\{0,1,2\}$ and applies the uniform morphism

$$\begin{aligned} h(0) &= 1100 \\ h(1) &= 0111 \\ h(2) &= 1010, \end{aligned}$$

to it. They proved that the resulting word $h(\mathbf{z})$ has no squares of order ≥ 3; in fact, the only squares that appear are $0^2, 1^2, (01)^2, (10)^2$, and $(11)^2$.

Although this is indeed a simple construction, in terms of automatic sequences, it can be improved. The minimum automaton size for $h(\mathbf{z})$, over all 2-automatic squarefree words $\mathbf{z}$, is 10, as can be verified by breadth-first search, with pruning if the prefix constructed so far requires 11 or more states.

This minimum number of states is achieved, for example, by applying h to the famous squarefree word $\mathbf{vtm} := \tau(g^\omega(0)) = 210201210120210201202101210201 2\cdots$, where

$$\begin{aligned} g(0) &= 01 & \tau(0) &= 2 \\ g(1) &= 20 & \tau(1) &= 1 \\ g(2) &= 23 & \tau(2) &= 0 \\ g(3) &= 02 & \tau(3) &= 1. \end{aligned}$$

Remark 2.1. The word **vtm** is (up to renaming) the classical squarefree word of Thue [16]. It can be defined in many different ways [3], including as the fixed point of the morphism defined by 2 → 210, 1 → 20, 0 → 1. The name **vtm** for this word comes from [5].

A novel alternative construction (not necessarily an image of **vtm**) needs only six states. This is the minimum possible number of states for a 2-automatic word containing no squares of order ≥ 3 and only 5 distinct squares.

Theorem 2.2. *Consider the infinite word $\rho(f^\omega(0))$, where*

$$\begin{aligned} f(0) &= 01 & \rho(0) &= 0 \\ f(1) &= 23 & \rho(1) &= 0 \\ f(2) &= 45 & \rho(2) &= 0 \\ f(3) &= 02 & \rho(3) &= 0 \\ f(4) &= 05 & \rho(4) &= 1 \\ f(5) &= 25 & \rho(5) &= 1. \end{aligned}$$

This is the lexicographically least word generated by a 2-automaton of ≤ 6 states, containing no squares of order ≥ 3, and only 5 distinct squares.

3. Only three distinct squares

The Entringer-Jackson-Schatz construction was optimally improved by Fraenkel and Simpson [9], as follows: they constructed an infinite binary word containing only 3 squares: 0^2, 1^2, and $(10)^2$.

Their construction is rather complicated, and also has a complicated proof. It starts with an infinite squarefree word $\mathbf{w}$ over $\{0, 1, 2\}$ avoiding the subwords 020 and 121. (Although they do not say so, an example of such a word is given by renaming the letters in $\mathbf{vtm} := \tau(g^\omega(0))$ above.) Then replace every occurrence of 12 with 132. Next, replace every remaining occurrence of 21 with 241. Finally, apply the morphism α defined as follows:

$$\begin{aligned}
\alpha(0) &= 011000111001\\
\alpha(1) &= 011100011001\\
\alpha(2) &= 011001110001\\
\alpha(3) &= 01100010111001\\
\alpha(4) &= 01110010110001.
\end{aligned}$$

The resulting word avoids all squares except 0^2, 1^2, and $(01)^2$.

Because of the inherent complexity of this construction, it seems desirable to find simpler ones. An example using 24-uniform morphisms was given by Rampersad *et al.* [14]. Define

$$\begin{aligned}
p(0) &= 012321012340121012321234\\
p(1) &= 012101234323401234321234\\
p(2) &= 012101232123401232101234\\
p(3) &= 012321234323401232101234\\
p(4) &= 012321234012101234321234,
\end{aligned}$$

and

$$\begin{aligned}
\beta(0) &= 011100\\
\beta(1) &= 101100\\
\beta(2) &= 111000\\
\beta(3) &= 110010\\
\beta(4) &= 110001.
\end{aligned}$$

Then $\beta(p^\omega(0))$ is an infinite word containing only the squares 0^2, 1^2, and $(01)^2$. This construction gives a 24-automatic sequence generated by an automaton of 18 states, so its weight is $24 \cdot 18 = 432$.

3.1. Ochem's word

Ochem [13] provided a different construction in 2006:

$$\begin{aligned}
\sigma(0) &= 00011001011000111001011001110001011100101100010111\\
\sigma(1) &= 00011001011000101110010110011100010110001110010111\\
\sigma(2) &= 00011001011000101110010110001110010111000101100111,
\end{aligned}$$

He showed that if $\mathbf{x}$ is a $(7/4 + \epsilon)$-free word, then $\sigma(\mathbf{x})$ contains only three squares.

In fact, we can also successfully apply σ to the word $\mathbf{vtm}$ above, even though it is not $(7/4 + \epsilon)$-free. Since σ is a uniform map, we know that $\sigma(\mathbf{vtm})$ is 2-automatic.

Theorem 3.1. *This word $\sigma(\mathbf{vtm})$ is a 2-automatic word containing only three distinct squares. It is generated by an automaton with 109 states (and has weight $2 \cdot 109 = 218$).*

3.2. The Harju-Nowotka construction

Harju and Nowotka [10] generated an infinite binary word with three squares by defining the map

$$\begin{aligned}\zeta(0) &= 111000110010110001110010\\ \zeta(1) &= 111000101100011100101100010\\ \zeta(2) &= 111000110010110001011100101100\ .\end{aligned}$$

and then applying it to **vtm**.

The morphism ζ is clearly not uniform. However, the lengths of the images of $0, 1, 2$ are (respectively) $24, 27, 30$ and form an arithmetic progression. This is enough to show that $\zeta(\mathbf{vtm})$ is 2-automatic, as the following result shows.

Theorem 3.2. *Let $\mathbf{vtm} = \tau(g^\omega(0))$ where g and τ are defined in Section 2. Let $h : \{0,1,2\}^* \to \Delta^*$ be a morphism. If the three lengths $|h(0)|$, $|h(1)|$, and $|h(2)|$ form an arithmetic progression, then $h(\mathbf{vtm})$ is 2-automatic.*

Proof. Suppose a, b are integers, with $a \geq 1$ and $a + 2b \geq 1$, such that $|h(i)| = a + ib$ for $i \in \{0, 1, 2\}$. Write $\mathbf{vtm} = c(0)c(1)c(2)\cdots$. An easy induction now shows that

$$|h(c(0)c(1)\cdots c(n-1))| = (a+b)n + bt_n,$$

for $n \geq 0$, where $\mathbf{t} = t_0t_1\cdots$ is the Thue-Morse word. To compute the n'th symbol of $h(\mathbf{vtm})$, divide n by $a+b$ to determine which block $h(c(i))$ it corresponds to; then adjust based on whether $t_i = 0$ or not. More precisely, define $n' := \lfloor n/(a+b) \rfloor$ and $m := n \bmod (a+b)$. Then

$$(h(\mathbf{vtm}))[n] := \begin{cases} (h(c(n')))[m], & \text{if } t_{n'} = 0;\\ (h(c(n'-1)))[m+a+b], & \text{if } t_{n'} = 1 \text{ and } t_{n'-1} = 0 \text{ and } m < b;\\ (h(c(n')))[m-b], & \text{if } t_{n'} = 1 \text{ and } t_{n'-1} = 0 \text{ and } m \geq b;\\ (h(c(n'-1)))[m+a], & \text{if } t_{n'} = 1 \text{ and } t_{n'-1} = 1 \text{ and } m < b;\\ (h(c(n')))[m-b], & \text{if } t_{n'} = 1 \text{ and } t_{n'-1} = 1 \text{ and } m \geq b.\end{cases}$$

For fixed a and b, an automaton on input n in base 2 can compute n' and m on the fly and do the required lookup. □

Theorem 3.3. *The infinite word $\zeta(\mathbf{vtm})$ contains only three distinct squares: $0^2, 1^2$, and $(01)^2$. It is generated by an automaton with 88 states, and has weight is $2 \cdot 88 = 176$.*

3.3. The Badkobeh-Crochemore construction

Yet another construction was given by Badkobeh and Crochemore [1, 2]. They defined the morphism

$$\begin{aligned}\xi(0) &= 000111\\ \xi(1) &= 0011\\ \xi(2) &= 01001110001101\ .\end{aligned}$$

of weight 24. Although ξ applied to a squarefree word can produce a word with more than three squares (consider 0102), it turns out that $\xi(\mathbf{vtm})$ is squarefree. Furthermore, although they do not mention it, ξ is a morphism of lowest total weight with this property.

Incidentally, we found another morphism with the same properties, of the same weight; it is

$$\begin{aligned}\kappa(0) &= 110100111000110100\\ \kappa(1) &= 1100\\ \kappa(2) &= 01\ .\end{aligned}$$

However, the lengths of the images of both of these morphisms are not in arithmetic progression, and so Theorem 3.2 does not apply. Indeed, we suspect (but did not prove) that neither $\xi(\mathbf{vtm})$ nor $\kappa(\mathbf{vtm})$ is a 2-automatic sequence. If they are 2-automatic, then more than 200 states are needed to generate them.

3.4. Our first construction

The previous section suggests looking for a morphism η of lowest total weight, where the lengths of the images of $0, 1, 2$ are in arithmetic progression, such that $\eta(\mathbf{vtm})$ has only 3 distinct squares. We found the following morphism, which is the smallest such, of weight 36.

$$\begin{aligned}\eta(0) &= 00011101\\ \eta(1) &= 001110001101\\ \eta(2) &= 0011000111001101\ .\end{aligned}$$

Theorem 3.4. *The infinite word $\eta(\mathbf{vtm})$ contains only three distinct squares: $0^2, 1^2$, and $(10)^2$. It is 2-automatic, and can be generated by an automaton of 27 states, so its weight is $2 \cdot 27 = 54$.*

3.5. The last construction

Finally, instead of using the strategy of applying a morphism to **vtm**, we can search directly for a k-automatic word of minimum total weight. It turns out that this minimum weight is 44, corresponding to a 2-automaton with 22 states; see Figure 1.

The corresponding representation is as the image, under the coding γ, of the fixed point of the morphism q defined below over the alphabet $\{0, 1, \ldots, 21\}$. We use commas to separate letters in the image of q, because of the large alphabet size.

$q(0) = 0, 1$	$\gamma(0) = 1$	$q(1) = 2, 3$	$\gamma(1) = 1$
$q(2) = 4, 5$	$\gamma(2) = 0$	$q(3) = 6, 7$	$\gamma(3) = 1$
$q(4) = 8, 9$	$\gamma(4) = 0$	$q(5) = 10, 11$	$\gamma(5) = 0$
$q(6) = 12, 7$	$\gamma(6) = 1$	$q(7) = 8, 9$	$\gamma(7) = 1$
$q(8) = 13, 14$	$\gamma(8) = 0$	$q(9) = 15, 1$	$\gamma(9) = 0$
$q(10) = 2, 16$	$\gamma(10) = 0$	$q(11) = 7, 17$	$\gamma(11) = 1$
$q(12) = 6, 4$	$\gamma(12) = 1$	$q(13) = 2, 16$	$\gamma(13) = 1$
$q(14) = 18, 19$	$\gamma(14) = 1$	$q(15) = 0, 1$	$\gamma(15) = 0$
$q(16) = 6, 7$	$\gamma(16) = 0$	$q(17) = 10, 11$	$\gamma(17) = 1$
$q(18) = 20, 9$	$\gamma(18) = 0$	$q(19) = 10, 21$	$\gamma(19) = 1$
$q(20) = 13, 14$	$\gamma(20) = 1$	$q(21) = 18, 19$	$\gamma(21) = 0$

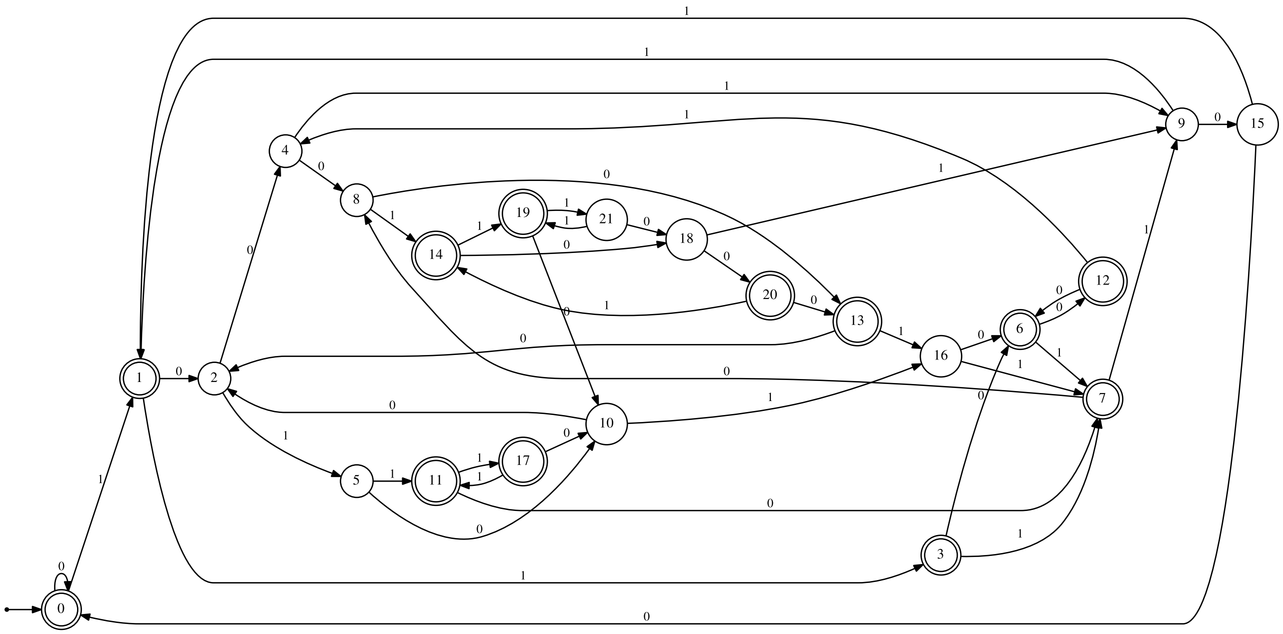

Figure 1. DFAO where accepting states have output 1 and all other states have output 0.

Theorem 3.5. *The infinite word*

$$\gamma(q^\omega(0)) = 110100110001110011010011100011010001110100110000\cdots$$

contains only 3 *distinct squares:* 0^2, 1^2, *and* $(10)^2$. *It has total weight* 44.

By exhaustive search we find that there are no k-automatic words containing only three distinct squares, with s states, for $3 \leq k \leq 44$ and $ks \leq 44$.

We propose this word as the simplest of all binary words with three squares.

4. Verifying the claims

We used breadth-first search to find candidates for the minimal examples presented here. The number of states in the minimal automaton were determined using the Myhill-Nerode theorem (see, *e.g.*, [11]). We used the theorem-proving software `Walnut` [12] to verify assertions about the squares contained in each word. For example, the claim about the 22-state automaton in the previous section can be proved as follows: create the automaton, and call it `Q` in `Walnut`, and then evaluate the following three statements:

```
eval qtest1 "Ei,n (n>=3) & At (t<n) => Q[i+t]=Q[i+t+n]":
eval qtest2 "Ei (Q[i]=Q[i+1])&(Q[i]=Q[i+2])&(Q[i]=Q[i+3])":
eval qtest3 "Ei (Q[i]=@0)&(Q[i+1]=@1)&(Q[i+2]=@0)&(Q[i+3]=@1)":
```

The first predicate asserts that there is a square of order ≥ 3 in the word. The second asserts that there is a square of the form $(00)^2$ or $(11)^2$. The third asserts that there is a square of the form $(01)^2$. Since all three queries return `false`, the word has the desired properties. The total computation time for this query is a few seconds on a laptop.

Each of Theorems 1, 2, 4, 5, 6 can be proved similarly, although some require significant memory resources and time. The `Walnut` code can be found on the website of the second author:

https://cs.uwaterloo.ca/~shallit/papers.html.

References

[1] G. Badkobeh, Infinite words containing the minimal number of repetitions. *J. Discrete Algorithms* **20** (2013) 38–42.

[2] G. Badkobeh and M. Crochemore, Fewest repetitions in infinite binary words. *RAIRO: ITA* **46** (2012) 17–31.

[3] J. Berstel, Sur la construction de mots sans carré. *Séminaire de Théorie des Nombres* (1978–1979) 18.01–18.15.

[4] J. Berstel, Axel Thue's Papers on Repetitions in Words: a Translation. Number 20 in Publications du Laboratoire de Combinatoire et d'Informatique Mathématique. Université du Québec à Montréal (1995).

[5] F. Blanchet-Sadri, J. Currie, N. Rampersad and N. Fox, Abelian complexity of fixed point of morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$. *INTEGERS: Elect. J. Combin. Number Theory* **14** (2014) #A11 (electronic).

[6] É. Charlier, N. Rampersad and J. Shallit, Enumeration and decidable properties of automatic sequences. *Internat. J. Found. Comp. Sci.* **23** (2012) 1035–1066.

[7] A. Cobham, Uniform tag sequences. *Math. Systems Theory* **6** (1972) 164–192.

[8] R.C. Entringer, D.E. Jackson and J.A. Schatz, On nonrepetitive sequences. *J. Combin. Theory Ser. A* **16** (1974) 159–164.

[9] A.S. Fraenkel and J. Simpson, How many squares must a binary sequence contain? *Electronic J. Combinatorics* **2** (1994) #R2.

[10] T. Harju and D. Nowotka, Binary words with few squares. *Bull. European Assoc. Theor. Comput. Sci.* **89** (2006) 164–166.

[11] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley (1979).

[12] H. Mousavi, Automatic theorem proving in `Walnut` (2016). http://arxiv.org/abs/1603.06017.

[13] P. Ochem, A generator of morphisms for infinite words. *RAIRO: ITA* **40** (2006) 427–441.

[14] N. Rampersad, J. Shallit and M.-w. Wang, Avoiding large squares in infinite binary words. *Theoret. Comput. Sci.* **339** (2005) 19–34.

[15] A. Thue, Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906) 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, edited by T. Nagell, Universitetsforlaget, Oslo (1977) 139–158.

[16] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912) 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, edited by T. Nagell, Universitetsforlaget, Oslo (1977) 413–478.