RAIRO-Inf. Theor. Appl. **40** (2006) 473-484
DOI: 10.1051/ita:2006030

SQUARES AND OVERLAPS IN THE THUE-MORSE SEQUENCE AND SOME VARIANTS

Shandy Brown[1], Narad Rampersad[2], Jeffrey Shallit[2] and Troy Vasiga[2]

Abstract. We consider the position and number of occurrences of squares in the Thue-Morse sequence, and show that the corresponding sequences are 2-regular. We also prove that changing any finite but nonzero number of bits in the Thue-Morse sequence creates an overlap, and any linear subsequence of the Thue-Morse sequence (except those corresponding to decimation by a power of 2) contains an overlap.

Mathematics Subject Classification. 68Q45, 68R15.

1. Introduction

Let $\mathbf{t} = \mathtt{01101001}\cdots = t_0t_1t_2\cdots$ be the Thue-Morse sequence defined by $t_i =$ the sum of the bits, modulo 2, in the binary expansion of i. Alternately, $\mathbf{t}$ can be described as the fixed point of the morphism μ that sends $\mathtt{0} \to \mathtt{01}$, $\mathtt{1} \to \mathtt{10}$. Thue proved [5, 9] that $\mathbf{t}$ contains no overlaps, that is, no subwords of the form $axaxa$ where $a \in \{\mathtt{0}, \mathtt{1}\}$ and $x \in \{\mathtt{0}, \mathtt{1}\}^*$.

Of course, $\mathbf{t}$ contains squares, that is, nonempty subwords of the form xx. In this paper, we define sequences based on the size and number of squares beginning at a given position of $\mathbf{t}$, and show that these sequences are easy to compute; more precisely, they are 2-regular in the sense of Allouche and Shallit [2, 3].

Next, we consider the overlap-freeness of some variants of $\mathbf{t}$. We show that changing any finite but positive number of bits of $\mathbf{t}$ yields a word with overlaps.

Keywords and phrases. Thue-Morse word, overlap-free word, automatic sequence.

[1] Digital Thinkery, 199 Carter Avenue, Waterloo, Ontario N2J 3K5, Canada; shandy@geeky.net

[2] School of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada; nrampersad@math.uwaterloo.ca; shallit@graceland.uwaterloo.ca; tmjvasiga@cs.uwaterloo.ca

We also show that any linear subsequence of **t** (other than those corresponding to decimation by a power of 2) contains overlaps.

In this paper we use the concepts of k-automatic and k-regular sequences. Roughly speaking, a sequence $(a_n)_{n\geq 0}$ is k-automatic if there exists a deterministic finite automaton that, on input n expressed in base k, reaches a state q with an associated output of a_n [4]. Alternatively, $(a_n)_{n\geq 0}$ is k-automatic if the set

$$\{(a_{k^i n+c})_{n\geq 0} \ : \ i \geq 0 \text{ and } 0 \leq c < k^i\}$$

is finite. A set of non-negative integers is k-automatic if its associated characteristic sequence is k-automatic. We will use the following basic facts about k-automatic sets [4, Th. 5.6.3]:

Lemma 1.1. *The class of k-automatic sets is closed under intersection and set addition (i.e., the operation $R + S = \{r + s \ : \ r \in R, s \in S\}$).*

We also discuss a generalization of k-automatic sequences, called k-regular sequences. A sequence is k-regular if the set of sequences generated by subsequences of the form

$$(a_{k^i n+c})_{n\geq 0}$$

for $i \geq 0$, $0 \leq c < k^i$, is finitely generated. For more details, see [4]. It follows that every k-automatic sequence is also k-regular.

2. Occurrences of subwords in automatic sequences

Given any infinite word $\mathbf{a} = a_0 a_1 a_2 \cdots$ we say that a subword w of length k *begins at position p* if $\mathbf{a} = a_0 \cdots a_{p-1} w a_{p+k} a_{p+k+1} \cdots$ Our first result shows that the set of positions of occurrences of any subword in an automatic sequence is automatic.

Theorem 2.1. *Let $\mathbf{a} = a_0 a_1 a_2 \cdots$ be a k-automatic sequence over the alphabet Δ, and let $w \in \Delta^*$. Then the set of positions p such that w occurs beginning at position p is k-automatic.*

Proof. Write $w = b_0 b_1 \cdots b_{n-1}$. Then each set $S_j = \{i \ : \ a_i = b_j\}$ is k-automatic. By Lemma 1.1, the sets $S_j - j$ are k-automatic. Also by Lemma 1.1, the intersection of all sets of the form $S_j - j$ is k-automatic. But this is precisely the set of positions p such that w occurs beginning at p. □

Example 2.2. Let us find an automaton for the starting positions of occurrences of 00 in **t**, the Thue-Morse word. We start with the well-known automaton for the sequence **t** in Figure 1. This automaton implements the fact that t_i is just the sum of the bits (taken mod 2) of i expressed in base 2.

The automaton for the shifted version of **t**, *i.e.*, $\mathbf{s} = s_0 s_1 s_2 \cdots$ where $s_i = t_{i+1}$, is given in Figure 2. To see that this works, note that the automaton must, on input n expressed in base 2, compute the parity of the sum of the bits of $n + 1$ expressed in base 2. It therefore suffices to keep track of both the sum of the bits

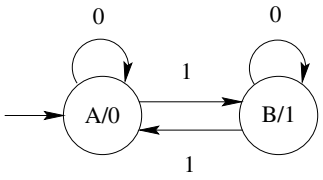

FIGURE 1. The Thue-Morse automaton.

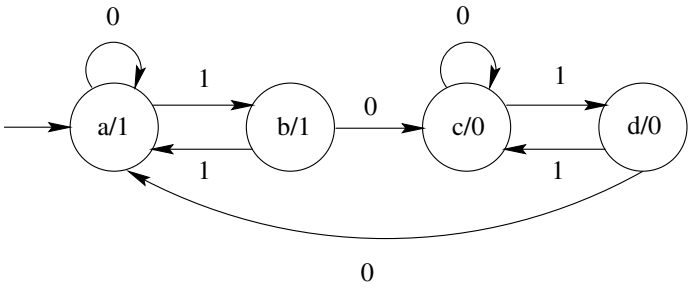

FIGURE 2. Automaton for the shifted version of Thue-Morse.

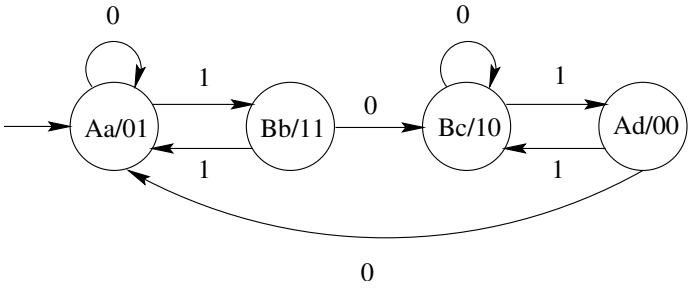

FIGURE 3. Automaton for occurrences of two-bit subwords of Thue-Morse.

seen so far (mod 2), together with the parity of the number of trailing 1's in the string seen so far.

To intersect positions we take the cross product, and after deleting states unreachable from the start state, we obtain the automaton in Figure 3. Note that reaching state Ad corresponds to occurrences of `00` and reaching state Bb corresponds to occurrences of `11`.

We can see how this construction relates to the proof of Theorem 2.1 in the following way. Write $w = \texttt{00}$ and define sets S_j with respect to w and $\mathbf{t}$ as in the proof of Theorem 2.1. Then $\mathbf{t}$ is the characteristic sequence of S_0 and $\mathbf{s}$ is the characteristic sequence of $S_1 - 1$. The intersection of S_0 and $S_1 - 1$ gives the starting positions of occurrences of `00`.

3. Square-counting sequences

Let $\mathbf{a} = a_0a_1a_2\cdots$ be a sequence. Based on $\mathbf{a}$ we can define two sequences that concern the squares in $\mathbf{a}$, as follows:

$$\begin{aligned} A(i) &:= \text{the number of squares beginning at a position} \leq i \text{ of } \mathbf{a} \\ B(i) &:= \text{the number of positions } p \leq i \text{ that mark the beginning of} \\ &\quad \text{a square of } \mathbf{a}. \end{aligned}$$

We can also consider a variant of the sequence A. If we use the subscript d, then only *distinct* squares are counted (as opposed to all occurrences with multiplicity, which is the default).

Another variant is to consider the corresponding first difference sequences. We have

$$\begin{aligned} \Delta A(i) &:= \text{the number of squares beginning at position } i \text{ of } \mathbf{a} \\ \Delta B(i) &:= 1, \text{ if there is a square beginning at position } i \text{ in } \mathbf{a} \text{ and 0 otherwise.} \end{aligned}$$

Finally, we can also consider the sequence

$$\begin{aligned} C(i) &:= \text{the length of } x, \text{ where } xx \text{ is the largest square beginning at} \\ &\quad \text{position } i \text{ of } \mathbf{a}. \end{aligned}$$

Of course, the domain of $A, \Delta A$, and C is the natural numbers together with ∞, since, for example, there may be arbitrarily many or arbitrarily large squares beginning at a given position; for example, consider the periodic sequence $\mathtt{000}\cdots$. It is perhaps surprising these sequences can be infinite everywhere even for non-periodic sequences, as the next result shows. Recall that an infinite word $\mathbf{x}$ is said to be *ultimately periodic* if it can be written in the form $yz^\omega = yzzz\cdots$ for some words y, z with z nonempty.

Theorem 3.1. *There exists a non-ultimately-periodic infinite word* $\mathbf{x}$ *over* $\{\mathtt{0},\mathtt{1}\}$ *such that for all integers* $m, n \geq 1$ *there exist finite words* u, v *and an infinite word* $\mathbf{w}$ *such that* $\mathbf{x} = uv^2\mathbf{w}$ *and* $|u| = m$, $|v| \geq n$.

Proof. The following procedure generates longer and longer prefixes x_i of an infinite word $\mathbf{x} = a_0a_1a_2\cdots$ with the desired properties:

$x_0 \leftarrow \epsilon$
for $i := 1$ to ∞ do
    $x_i \leftarrow x_{i-1}$ concatenated with $\mathtt{0}^i\mathtt{1}$
    for $j := i$ downto 1 do
        $x_i \leftarrow x_i$ concatenated with the string formed by deleting the first $j-1$ characters of x_i.

For example, here are the first few values of x_i:

$x_0 = \epsilon$
$x_1 = \mathtt{0101}$
$x_2 = \mathtt{01010011010010101001101001}$

By construction, each x_i is a prefix of x_j for $i < j$, so there is a unique infinite word $\mathbf{x}$ of which all the x_i are prefixes. Since each x_i contains $\mathtt{10^i1}$ as a substring, $\mathbf{x}$ cannot be ultimately periodic. And the loop on j ensures that arbitrarily large squares begin at each position. □

A simple modification of this construction generates a word with arbitrarily large and arbitrarily long powers beginning at every position.

In the next section we show that if $\mathbf{a} = \mathbf{t}$, the Thue-Morse sequence, then each of these sequences A and B (with or without subscripts), ΔA and ΔB, and C, is 2-automatic.

4. The squares in the Thue-Morse word

Pansiot [7] and Brlek [6] described the squares in the Thue-Morse word:

Theorem 4.1. *All squares in* $\mathbf{t}$ *are of the form* $\mu^k(\mathtt{00})$, $\mu^k(\mathtt{11})$, $\mu^k(\mathtt{010010})$, *or* $\mu^k(\mathtt{101101})$ *for some* $k \geq 0$, *and all these squares actually occur.*

We now describe all the positions where these squares can occur. We use the following notation: $\overline{\mathtt{0}} = \mathtt{1}$ and $\overline{\mathtt{1}} = \mathtt{0}$.

Lemma 4.2. *Let* $a \in \{\mathtt{0}, \mathtt{1}\}$.

(a) *The only occurrences of* $\mu^k(aa)$ *in* $\mathbf{t}$ *begin at positions of the form* $2^k \cdot p$, *where* aa *occurs at position* p.
(b) *The only occurrences of* $\mu^k(a\overline{a}aa\overline{a}a)$ *in* $\mathbf{t}$ *begin at positions of the form* $2^k \cdot p$, *where* $a\overline{a}aa\overline{a}a$ *occurs at position* p.

Proof. We prove the two claims by induction on k. The results are trivially true for $k = 0$. For the induction step, assume the claims are true for k; we prove them for $k + 1$.

For (a), we claim that $\mu^{k+1}(aa)$ must begin at an even position of $\mathbf{t}$. For if it appears at an odd position, and $k \geq 1$, then, since $\mu^{k+1}(a)$ begins with $a\overline{a}\,\overline{a}a$, $\overline{a}\,\overline{a}$ occurs at an even position and would thus be the image under μ of a letter, which is not the case. If $k = 0$, then we would have $a|\overline{a}a|\overline{a}$ as a subword of $\mathbf{t}$, where the straight bars | separate pairs of letters that are images under μ of a single letter. It therefore follows that the letter following the last $\overline{a}$ must be a, and so $\mathbf{t}$ contains the overlap $a\overline{a}a\overline{a}a$, a contradiction. Now, since $\mu^{k+1}(aa)$ appears at an even position q, it is the image of $\mu^k(aa)$ at position $q/2$. But by induction $q/2 = 2^k \cdot p$, so $q = 2^{k+1} \cdot p$, as desired.

A similar argument proves (b). □

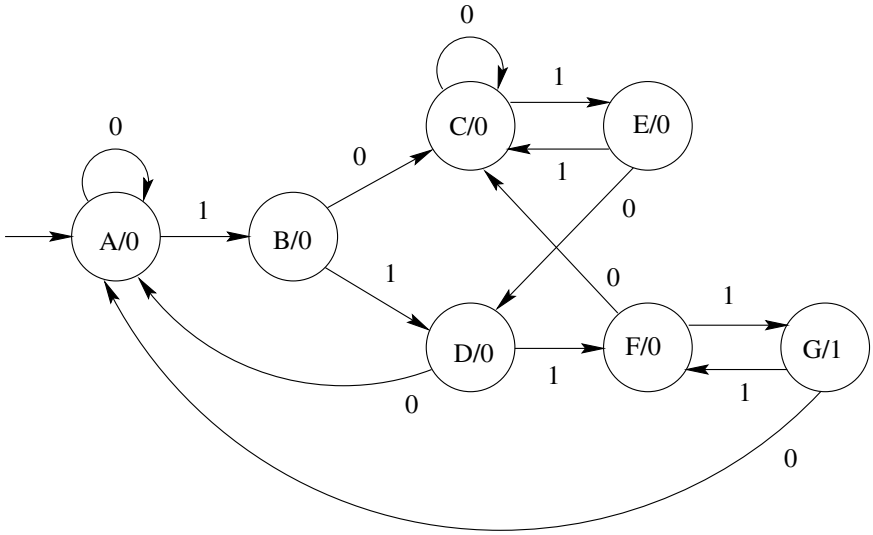

FIGURE 4. Automaton for occurrences of `010010` of Thue-Morse.

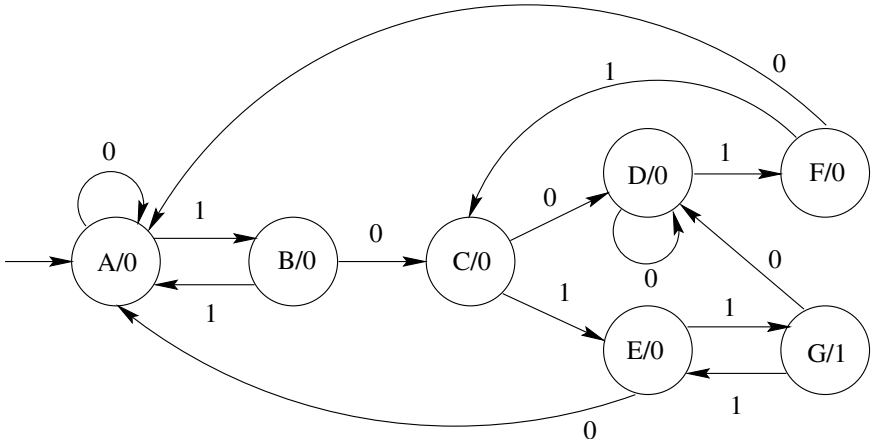

FIGURE 5. Automaton for occurrences of `101101` of Thue-Morse.

We note also that for any position i, there is at most one square in $\mathbf{t}$ beginning at position i.

Previously we showed that the set of positions of `00` (resp., `11`) in $\mathbf{t}$ is 2-automatic; see Figure 3. A similar technique can be carried out for `010010` and `101101`. After minimization, we obtain the automata in Figures 4 and 5.

Theorem 4.3. *The sequence ΔB is 2-automatic. The sequence B is 2-regular.*

Proof. As we have seen in Theorem 2.1, for each subword w of $\mathbf{t}$, there is a regular language L_w consisting of the base-2 representations of those indices i

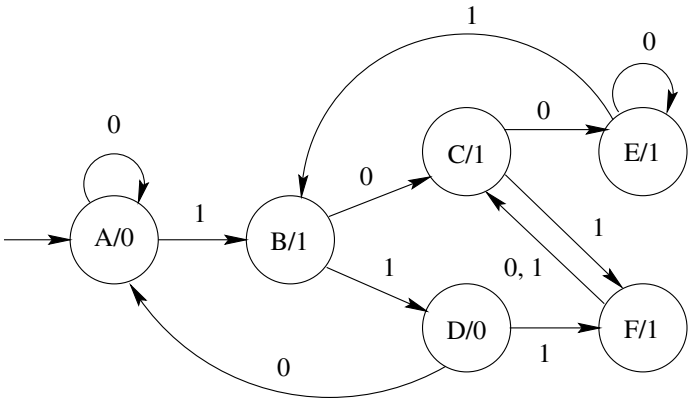

FIGURE 6. Automaton for starting positions of squares in $\mathbf{t}$.

where w occurs. (The corresponding automata were given explicitly for $w \in \{\mathtt{00}, \mathtt{11}, \mathtt{010010}, \mathtt{101101}\}$ in Figs. 3–5.) Theorem 4.1 shows that the set of base-2 representations of starting indices of all squares in $\mathbf{t}$ is given by

$$(L_{\mathtt{00}} \cup L_{\mathtt{11}} \cup L_{\mathtt{010010}} \cup L_{\mathtt{101101}})\mathtt{0}^*, \tag{1}$$

which is evidently a regular language. Thus ΔB is 2-automatic.

To show that the sequence B is 2-regular, we first define the convolution of two sequences. If $\mathbf{a} = (a_i)_{i\geq 0}$ and $\mathbf{b} = (b_i)_{i\geq 0}$ are two sequences of integers, then the convolution $\mathbf{c} = \mathbf{a} \star \mathbf{b}$ is defined as follows: if $\mathbf{c} = (c_i)_{i\geq 0}$, then

$$c_i = \sum_{j+k=i} a_j b_k.$$

If $\mathbf{a}$ and $\mathbf{b}$ are k-regular sequences, then by [4], Theorem 16.4.1, $\mathbf{c} = \mathbf{a} \star \mathbf{b}$ is k-regular. To conclude that B is 2-regular, we note that

$$B(i) = \sum_{j\leq i} \Delta B(j) = \mathbf{a} \star \Delta B,$$

where $\mathbf{a}$ is the constant sequence $1, 1, 1, \ldots$ □

Figure 6 gives the 2-automaton for the set of starting positions of squares in $\mathbf{t}$, that is, the 2-automaton generating the sequence ΔB for $\mathbf{t}$. It can be obtained through the usual construction combining the automata for $L_{\mathtt{00}}$, $L_{\mathtt{11}}$, $L_{\mathtt{010010}}$, and $L_{\mathtt{101101}}$, and equation (1), and then minimizing the result.

We now turn to the sequence C. First, we prove a lemma.

Lemma 4.4. *Let* $\mathbf{a} = a_0a_1a_2\cdots$ *be a* k*-regular sequence of integers with* $a_0 = 0$. *For each integer* $j \geq 1$ *define a new sequence* $\mathbf{a}/j = b_0b_1b_2\cdots$ *as follows:*

$$b_n = \begin{cases} a_{n/j}, & \text{if } n \equiv 0 \bmod j; \\ 0, & \text{otherwise.} \end{cases}$$

Let c *be an integer. Then the sequence*

$$\mathbf{a}(c,k) := \mathbf{a} + c\cdot(\mathbf{a}/k) + c^2\cdot(\mathbf{a}/k^2) + \cdots$$

is also k*-regular.*

Proof. Write $\mathbf{a}(c,k) = d_0d_1d_2\cdots$. It suffices to show that each subsequence d_{k^in+e}, $0 \leq e < k^i$, is a linear combination of elements of the k-kernel of $\mathbf{a}$. Let $j = \nu_k(\gcd(k^i,e))$, where $\nu_k(n)$ is the exponent of the highest power of k dividing n. Then

$$d_{k^in+e} = a_{k^in+e} + c\cdot a_{k^{i-1}n+e/k} + c^2\cdot a_{k^{i-2}n+e/k^2} + \cdots + c^j\cdot a_{k^{i-j}n+e/k^j}.$$

□

Theorem 4.5. *The sequence* C *is* 2*-regular.*

Proof. For each subword w of $\mathbf{t}$, let $\mathbf{s}_w$ be the characteristic sequence for the occurrences of w in $\mathbf{t}$. Then

$$C = \mathbf{s}_{00}(2,2) + \mathbf{s}_{11}(2,2) + 3\cdot\mathbf{s}_{010010}(2,2) + 3\cdot\mathbf{s}_{101101}(2,2).$$

By Lemma 4.4, C is 2-regular. (Here we have used the fact that the sum of k-regular sequences is k-regular; see [4], Th. 16.2.1.) □

We can also obtain a defining set of relations for the sequence C. For $n \geq 0$, they are as follows (we omit the proof):

$$\begin{aligned} C(2n) &= 2C(n) \\ C(4n+1) &= 1 \\ C(8n+7) &= 1 + \frac{2}{3}C(8n+3) \\ C(16n+3) &= 0 \\ C(16n+11) &= 3 - C(8n+3). \end{aligned}$$

Moreover, the sequence $C(8n+3)$ is the fixed point of the map $0 \to 03$, $3 \to 00$, which is just a recoding of the period-doubling sequence [4], Example 6.3.4.

Now we turn to counting the distinct squares.

Theorem 4.6. *For $j \geq 0$ we have*

$$A_d(i) = \begin{cases} 4j+5, & \text{if } 2^{j+3} \leq i < 10 \cdot 2^j; \\ 4j+6, & \text{if } 10 \cdot 2^j \leq i < 11 \cdot 2^j; \\ 4j+7, & \text{if } 11 \cdot 2^j \leq i < 15 \cdot 2^j; \\ 4j+8, & \text{if } 15 \cdot 2^j \leq i < 2^{j+4}. \end{cases}$$

Proof. It is easy to check that 00 occurs for the first time in $\mathbf{t}$ at position 5, 11 occurs for the first time at position 1, 010010 occurs for the first time at position 15, and 101101 occurs for the first time at position 11. From Lemma 4.2, it follows that the square

$\mu^k(00)$ occurs for the first time at position $5 \cdot 2^k$
$\mu^k(11)$ occurs for the first time at position 2^k
$\mu^k(010010)$ occurs for the first time at position $15 \cdot 2^k$
$\mu^k(101101)$ occurs for the first time at position $11 \cdot 2^k$. □

Corollary 4.7. *For $\mathbf{t}$, the sequence A_d is 2-regular. The sequence ΔA_d is 2-automatic.*

We note that the same sorts of results hold for the paperfolding sequence $\mathbf{p}$ [4], p. 155. It is known [8] that any square xx in the paperfolding sequence satisfies $|x| \leq 5$. It follows that there are only finitely many distinct squares in $\mathbf{p}$, and the desired results then follow by Theorem 2.1.

5. Overlap-freeness of some variants of the Thue-Morse sequence

As is well-known, the Thue-Morse sequence is overlap-free, that is, it contains no subwords of the form $axaxa$ where $a \in \{0, 1\}$ and $x \in \{0, 1\}^*$. In this section we change focus somewhat and consider some variants of the Thue-Morse sequence, showing that these variants do not preserve overlapfreeness.

In our first theorem we consider taking the Thue-Morse sequence and changing some finite, but nonzero, number of bits.

Theorem 5.1. *Let $\mathbf{t}'$ be an infinite word obtained from $\mathbf{t}$ by changing k bits, $0 < k < \infty$. Then $\mathbf{t}'$ contains an overlap.*

First, we state a very useful lemma [1]:

Lemma 5.2. *If $\mathbf{u}$ is an infinite overlap-free word, then there exist a finite word $x \in \{\epsilon, 0, 1, 00, 11\}$ and an infinite overlap-free word $\mathbf{z}$ such that $\mathbf{u} = x\mu(\mathbf{z})$.*

Now we can prove Theorem 5.1.

Proof. The proof is by contradiction. Let k be minimal such that $\mathbf{t}'$ is overlap-free. Suppose $k = 1$. Changing the bits with index 0, 1, or 2 in $\mathbf{t}$ creates the overlaps 111, 01010, and 1001001 respectively. For each of the remaining bits,

TABLE 1. Changing a bit in the Thue-Morse word.

Original subword	Modified subword	Original subword	Modified subword
00101100	00**1**1**1**100	10010110	100**0**0110
00101101	00**1**1**1**101	10011001	100**0**1001
00110010	001**0**0010	10011010	100**0**1010
00110100	001**0**0100	10100101	101**1**0101
01001011	010**1**1011	10100110	101**1**0110
01001100	010**1**1100	10110011	101**0**0011
01011001	010**0**1001	10110100	101**0**0100
01011010	010**0**1010	11001011	110**1**1011
01100101	011**1**0101	11001101	110**1**1101
01100110	011**1**0110	11010010	110**0**0010
01101001	011**1**1001	11010011	110**0**0011

having index ≥ 3, consider the subword $b_1b_2\cdots b_8$ of length 8 in $\mathbf{t}$, where b_4 is the bit to be changed. There are 22 such subwords of length 8 and changing b_4 in any of these words creates an overlap, as shown in Table 1.

We assume then that $k > 1$. By Lemma 5.2 we can write $\mathbf{t}' = x\mu(\mathbf{y})$, where $x \in \{\epsilon, 0, 1, 00, 11\}$ and $\mathbf{y}$ is overlap-free. We have three cases, $x \in \{\epsilon, 0, 00\}$. (The cases where $x \in \{1, 11\}$ are similar to those where $x \in \{0, 00\}$.)

Case 1. $x = \epsilon$, $\mathbf{t}' = \mu(\mathbf{y})$. We write $\mathbf{y} = u\mathbf{v}$, where $\mu(u)$ is the minimal word containing the k changed bits of $\mathbf{t}$. Note that the factorization $\mathbf{t}' = \mu(\mathbf{y})$ implies that if the bit in position $2i$ ($2i+1$ resp.), $i < |u|$, of $\mu(u)$ differs from its corresponding bit in $\mathbf{t}$, then so does the bit in position $2i+1$ ($2i$ resp.). Since each two bit subword of $\mathbf{v}$ beginning at position $2i$, $i \geq |u|$, in $\mathbf{t}'$ is the image under μ of the i-th bit of $\mathbf{t}$, we see that the only bits of $u\mathbf{v}$ that differ from $\mathbf{t}$ occur in u, where $|u| = |\mu(u)|/2$. Thus $\mathbf{y}$ differs from $\mathbf{t}$ in $k/2$ bits and is overlap-free, contradicting the minimality of k.

Case 2. $x = 0$, $\mathbf{t}' = 0\mu(\mathbf{y})$. Inspection of Figure 3 shows that all occurrences of 00 in $\mathbf{t}$ begin at an odd position. Thus, somewhere after the last bit changed in $\mathbf{t}'$, there must be an occurrence of 00 that begins in an odd position. But then 00 must be the image under μ of either 0 or 1, which is impossible.

Case 3. $x = 00$, $\mathbf{t}' = 00\mu(\mathbf{y})$. By [1], Lemma 2d, if $00\mu(\mathbf{y})$ is overlap-free, then $1\mathbf{y}$ is overlap-free. But by an argument similar to that of case 1, $1\mathbf{y}$ differs from $\mathbf{t}$ in fewer than k bits, contradicting the minimality of k. □

We remark that a similar result holds for $\mathcal{S}(\mathbf{t}) = t_1t_2t_3\cdots$, the Thue-Morse sequence shifted by one symbol. However, the same result does not hold for $\mathcal{S}^2(\mathbf{t}) = t_2t_3\cdots$, as the string $\overline{t_2}t_3t_4\cdots$ is easily seen to be overlap-free.

Pansiot [7], Corollary 2, showed that any word formed by making a finite but positive number of insertions and deletions to $\mathbf{t}$ cannot be generated by an iterated morphism. Our results can be viewed as a complement to this result.

Now we consider linear subsequences of the Thue-Morse sequence, for example $(t_{24n+7})_{n\geq 0}$.

Theorem 5.3. *Let i, a be integers with $i \geq 1$, $a \geq 0$. Then $(t_{in+a})_{n\geq 0}$ is overlap-free if and only if i is a power of 2.*

Proof. Suppose i is a power of 2. Then, repeatedly using the identities $t_{2k} = t_k$, $t_{2k+1} = \overline{t_{2k}}$ we find that there exists a' such that either $t_{in+a} = t_{n+a'}$ or $t_{in+a} = \overline{t_{n+a'}}$ for all $n \geq 0$. Thus $(t_{in+a})_{n\geq 0}$ equals a shift of either $\mathbf{t}$ or $\overline{\mathbf{t}}$, and hence is overlap-free.

Now suppose i is not a power of 2. Write $i = 2^u \cdot i'$ where i' is odd. Again, repeatedly using the identities $t_{2k} = t_k$, $t_{2k+1} = \overline{t_{2k}}$ we find that there exists a' such that either $t_{in+a} = t_{i'n+a'}$ or $t_{in+a} = \overline{t_{i'n+a'}}$ for all $n \geq 0$. Thus, without loss of generality, we may assume i is odd.

We will now show that the subsequence $(t_{in+a})_{n\geq 0}$ contains either the overlap $\mathtt{000}$ or $\mathtt{111}$. We introduce some notation: by $[x]_2$ we mean the integer that x represents in base 2. By $s_2(x)$ we mean the sum of the bits of the base-2 representation of x.

There are several cases to consider.

Case 1. The base-2 expansion of i ends in at least two 1's. Let $x \in \{\mathtt{0},\mathtt{1}\}^*$ be a string such that $[x\mathtt{01}^j]_2 = i$ for some $j \geq 2$. Choose n such that $in + a \equiv$ mod $2^{u+1} + 12^{u+2}$ for some u chosen much larger than $\log_2 i$. Now the base-2 expansion of $in + a$ looks like $y\mathtt{10}^u\mathtt{1}$. On the other hand, the base-2 expansion of $i(n-1)+a$ is $y\mathtt{01}^{u-|x|-j}\overline{x}\mathtt{10}^{j-2}\mathtt{10}$, and the base-2 expansion of $i(n-2)+a$ is $y\mathtt{01}^{u-|x|-j-1}\overline{x}\mathtt{10}^{j-1}\mathtt{11}$. By choosing the parity of u appropriately, we can force the parity of the sum of the bits of all three strings to be the same. Since $t_j = s_2(j)$ mod 2, the subsequence contains either the overlap $\mathtt{000}$ or $\mathtt{111}$.

Case 2. The base-2 expansion of i ends in $\mathtt{01}$. Let $x \in \{\mathtt{0},\mathtt{1}\}^*$ be a string such that $[x\mathtt{01}]_2 = i$.

Case 2a. $s_2(x)$ is odd.

Choose n such that $in + a \equiv 2^u \mod 2^{u+1}$ for some u chosen much larger than $\log_2 i$. Now the base-2 expansion of $in + a$ looks like $y\mathtt{10}^u$. On the other hand, the base-2 expansion of $i(n+1)+a$ is $y\mathtt{10}^{u-|x|-2}x\mathtt{01}$, and the base-2 expansion of $i(n+2)+a$ is $y\mathtt{10}^{u-|x|-3}x\mathtt{010}$. Since $s_2(x)$ is odd, the parity of the sum of the bits of all three strings is the same. Thus the subsequence contains either the overlap $\mathtt{000}$ or $\mathtt{111}$.

Case 2b. $s_2(x)$ is even.

Choose n such that $in+a \equiv 2^{u+1}+1 \mod 2^{u+2}$ for some u chosen much larger than $\log_2 i$. Then the base-2 expansion of $in+a$ is $y\mathtt{10}^u\mathtt{1}$. There are two subcases to consider.

Case 2bi. $x = z\mathtt{10}^l$, $l \geq 1$. Then the base-2 expansion of $i(n-1)+a$ is $y\mathtt{01}^{u-|x|-1}\overline{z}\mathtt{10}^{l+2}$, while the base-2 expansion of $i(n+1)+a$ is $y\mathtt{10}^{u-|x|-1}z\mathtt{10}^l\mathtt{10}$.

Case 2bii. $x = z\mathtt{01}^l$, $l \geq 1$. Then the base-2 expansion of $i(n-1)+a$ is $y\mathtt{01}^{u-|x|-1}\overline{z}\mathtt{10}^{l-1}\mathtt{100}$, while the base-2 expansion of $i(n+1)a$ is $y\mathtt{10}^{u-|x|-1}z\mathtt{01}^{l+1}\mathtt{0}$.

Since $s_2(x)$ is even, in both case 2bi and 2bii we can choose u to force the parity of the sum of the bits of all three strings to be the same.

Thus the subsequence contains either the overlap $\mathtt{000}$ or $\mathtt{111}$. □

Noted added in proof. Gwénaël Richomme points out (2nd Sep. 2005) that any Sturmian word is an example of an non-ultimately-periodic word containing arbitrarily large squares beginning at every position. For instance, the Fibonacci word is an example of a 4 power-free word with this property. Richemomme also notes that one can construct a cube free binary word with this property.

Acknowledgements. We thank the referees for their helpful suggestions.

References

[1] J.-P. Allouche, J. Currie, and J. Shallit, Extremal infinite overlap-free binary words. *Electronic J. Combinatorics* **5** (1998), #R27 (electronic). `http://www.combinatorics.org/Volume_5/Abstracts/v5i1r27.html`

[2] J.-P. Allouche and J.O. Shallit, The ring of k-regular sequences. *Theoret. Comput. Sci.* **98** (1992) 163–197.

[3] J.-P. Allouche and J.O. Shallit, The ring of k-regular sequences, II. *Theoret. Comput. Sci.* **307** (2003) 3–29.

[4] J.-P. Allouche and J.O. Shallit, *Automatic Sequences: Theory, Applications, Generalizations.* Cambridge University Press (2003).

[5] 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 (February 1995).

[6] S. Brlek, Enumeration of factors in the Thue-Morse word. *Disc. Appl. Math.* **24** (1989) 83–96.

[7] J.J. Pansiot, The Morse sequence and iterated morphisms. *Inform. Process. Lett.* **12** (1981) 68–70.

[8] H. Prodinger and F.J. Urbanek, Infinite 0–1-sequences without long adjacent identical blocks. *Discrete Math.* **28** (1979) 277–289.

[9] 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*, T. Nagell, editor, Universitetsforlaget, Oslo (1977) 413–478.