Repetitions of Words and the Thue-Morse sequence

Carl D. Offner

1 Introduction

In his essay *Notes on Structured Programming*, Edsger Dijkstra (1972) uses as an example the problem of constructing a program that generates successively longer

> sequences of 0's, 1's, and 2's without non-empty, element-wise equal, adjoining subsequences, generating these sequences in alphabetical order until a sequence of length 100 (i.e., of 100 digits) has been generated.

Dijkstra creates a program that generates these strings, using backtracking where necessary to get out of situations like the following: suppose we have arrived at the string 0102010, which has no two identical adjacent substrings. This string cannot be extended to maintain that constraint, however—appending 0, 1, or 2 leads in each case to a string that does contain two identical adjacent substrings. So at this point Dijkstra's program backtracks—it deletes the last element of the string and tries another. (And sometimes it has to delete more than 1 element.)

The obvious question that this problem addresses is that of determining if there is an upper limit to the length of such sequences.

Dijkstra says he got this example from Niklaus Wirth. What Wirth evidently did not tell him was that this problem was completely solved around 1906 by the Norwegian mathematician Axel Thue, who showed that there is an infinitely long sequence having this property, and also showed how to generate it.

A nice exposition of this result is contained in the first chapter of Salomaa's little book, *Jewels of Formal Language Theory* (1981). I am going to follow Salomaa's treatment here, except that the proof I give of Theorem 1 is considerably simpler than the one he gives, and the proof of Theorem 4 is reworded to make the main idea stand out.

2 Square-free and cube-free sequences

A sequence (or string) over an alphabet is said to be *square-free* if no two non-empty adjoining substrings are equal. Another way to say this is to say that the sequence does not contain a subsequence of the form xx, where x is a finite non-empty word.

It is easy to see that over an alphabet of two symbols ($\{a, b\}$, say) there is no square-free sequence of length greater than 3. This is because any sequence of length 4 either contains aa or bb or is one of the two square sequences *abab* or *baba*.

For this reason, any square-free word of length greater than 3 must be a word over an alphabet containing at least three symbols. Thue proved that there is an infinite such word (also called an ω-word); i.e., a sequence $\{a_i : 1 \leq i < \infty\}$ containing no squares.

In order to prove this result, we first have to prove an auxiliary result. We need two definitions:

- A word is *cube-free* if it contains no subsequence of the form xxx, where x is a finite non-empty word.
- A word is *strongly cube-free* if it contains no subsequence of the form xxa, where x is a finite non-empty word and a is the first symbol in x.

Clearly any square-free word is strongly cube-free, and any strongly cube-free word is cube-free.

Thue first proved that there is a strongly cube-free sequence over any alphabet containing at least two symbols[1]. This is what we will show now.

3 The Thue-Morse sequence

Define the Thue-Morse sequence $\{s_n : n \geq 0\}$ over the alphabet $\{+1, -1\}$ (equivalently, over the alphabet $\{+, -\}$) by

$$s_n = \begin{cases} -1 & \text{if } n \text{ has an odd number of 1's in its binary representation} \\ +1 & \text{if } n \text{ has an even number of 1's in its binary representation} \end{cases}$$

The sequence starts like this:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
s_n	+	−	−	+	−	+	+	−	−	+	+	−	+	−	−	+	−	+	+	−	+	−	−	+	+

The following properties are then easy to see, just by looking at the binary representation of n:

1. $s_{2n} = s_n$
2. $s_{2n+1} = -s_n$
3. $s_{2^n+j} = -s_j$ for $0 \leq j < 2^n$

Now from properties 1 and 2 it immediately follows that

4. If $s_j = s_{j+1}$ then j must be odd.

As an immediate corollary, we have

[1] Of course by what we have just shown, we know that this sequence cannot be square-free. But we will use it to construct a square-free sequence.

5. There are no 3 consecutive terms of the sequence which are equal.

From property 1, we see that the even-numbered elements of the sequence are just the sequence again, and property 2 shows that the odd-numbered elements of the sequence are just "minus the sequence".

Let us call the even-numbered elements of the sequence the "even subsequence", and denote this subsequence by $\mathcal{E}$. Similarly, let us call the odd-numbered elements the "odd subsequence", and denote this subsequence by $\mathcal{O}$. So "$\mathcal{O} = -\mathcal{E}$".

6. In any 5 consecutive elements of the sequence, there must be 2 consecutive elements which are identical.

This is because, if it were not so, the 5 elements would have to be $+-+-+$ or $-+-+-$. The 1st, 3rd, and 5th elements of these 5 elements then would be $+++$ or $---$ and would belong to either $\mathcal{O}$(if the first of the 5 elements was an odd-numbered element) or $\mathcal{E}$(if it was an even-numbered element). That is, $\mathcal{O}$ or $\mathcal{E}$ would have 3 consecutive equal terms, which is impossible.

1 Theorem *The Thue-Morse sequence is strongly cube-free. That is, there are no values $a \geq 0, b > 0$ such that*

1. $s_a = s_{a+b} = s_{a+2b}$

2. $s_{a+j} = s_{a+b+j} \quad (0 < j < b)$

Proof. If there were values a and b as stated, then

1. b could not equal 1 (this is property 5).

2. b could not be odd and greater than 1. For then there would be at least 7 elements between a and $a + 2b$, so by property 6, there would have to be 2 consecutive equal elements. In fact, since the elements from a to $a + b$ are the same as the elements from $a + b$ to $a + 2b$, there would have to be 2 distinct pairs of consecutive equal elements, separated by the distance b. Since b is odd, one of those pairs of equal elements would have to start at an even position, violating property 4.

3. b could not be even. For if so, let b be minimal. Then as in the proof of property 6 above, either $\mathcal{E}$ or $\mathcal{O}$ would satisfy the equations a) and b) in the statement of this theorem, with $b/2$ in place of b. But $b/2$ cannot be even, since b was minimal, and it cannot be odd, as was shown in parts 1 and 2 of this proof. □

4 Square-free sequences

Now, following Salomaa, we finish up Thue's construction of a square-free sequence. First, we show how to construct a square-free sequence on an alphabet of four letters:

2 Lemma *There exists a square-free sequence β over an alphabet of four elements.*

Proof. Consider the alphabet of four elements constructed of pairs of symbols of our two-element alphabet considered above. That is, define the alphabet Σ by

$$\Sigma = \{[++], [+-], [-+], [--]\}$$

Starting with the Thue-Morse sequence $\alpha = \{c_i : 1 \le i < \infty\}$ (each c_i being + or $-$), we define a new sequence $\beta = \{d_i : 1 \le i < \infty\}$ by the rule

$$d_j = [c_j c_{j+1}] \quad \text{for all } j \ge 1$$

We can see that the sequence β is square-free. For suppose yy occurs as a subword in β, where

$$y = d_{j+1} \dots d_{j+t} = d_{j+t+1} \dots d_{j+2t}$$

for some $t \ge 1$. Then we have

$$[c_{j+1}c_{j+2}] \dots [c_{j+t}c_{j+t+1}] = [c_{j+t+1}c_{j+t+2}] \dots [c_{j+2t}c_{j+2t+1}]$$

And this in turn shows that $c_{j+1} = c_{j+t+1} = c_{j+2t+1}$ and also that $c_{j+i} = c_{j+t+i}$ for $1 \le i \le t$. Hence

$$(c_{j+1} \dots c_{j+t})^2 c_{j+1}$$

occurs as a subword in the Thue-Morse sequence α, contradicting the fact that this sequence is strongly cube-free. □

Finally, we show how this square-free sequence β can be modified to form a square-free sequence γ on an alphabet consisting of only three letters. As we have already seen this result is best possible.

The construction hinges on the following observation: If we denote the letters of Σ by

$$[{+}{+}] = 1, [{+}{-}] = 2, [{-}{+}] = 3, [{-}{-}] = 4$$

then we can easily see that in the word β, the letter 1 must be followed by either 1 or 2 (since $[{+}{+}]$ and whatever follows it in β have to overlap in α). And in fact, 1 cannot be followed by 1, since that would mean that α contained the string $+ + +$, which we know is impossible. Hence 1 can only be followed by 2. Exactly the same kind of reasoning proves all the statements of the following lemma:

3 Lemma *In the word β,*

- *1 is always preceded by 3 and followed by 2.*
- *4 is always preceded by 2 and followed by 3.*

This has the following consequence: Suppose we replace all the occurrences of 1 and 4 in the sequence β by a letter a. Call the resulting string γ. Then we can reconstruct β from γ by simply looking at what precedes or follows each a.

4 Theorem *The sequence γ is square-free. That is, there is a square-free sequence on an alphabet of three letters.*

Proof. If it were not, there would be a non-empty string y such that γ contained y^2.

Now first of all, y must have length at least 2. This is because if it had length 1, yy would have to be of the form aa (it certainly could not be 22 or 33). This would mean that β contained one of the substrings 11, 14, 41, or 44—and we have already seen that this cannot happen.

But then every occurrence of a in y has a letter either preceding it or following it in y. As we have seen, this means that we can reconstruct the unique string (call it x) in β that y must have come from. This shows that β had to contain the sequence x^2, contradicting the fact that β is square-free. □

5 And of course there is more...

The Thue-Morse sequence has other interesting properties as well. Here are two:

1. Mahler (1929) proved that the binary number

   $$.0110100110010110\ldots$$

   whose successive binary digits form the Thue-Morse sequence (with 0 substituted for -1) is transcendental.

2. The Thue-Morse sequence plays a role in the theory of chaotic dynamical systems. In fact, Morse rediscovered this sequence in the late 1930's in an investigation of dynamical systems. More recently, it has been found that the binary number constructed above from the Thue-Morse sequence is the coordinate (in a certain natural sense) of the so-called Feigenbaum-Myrberg point in the Mandelbrojt set. Some interesting surveys of this general area are contained in the wonderful book *The Beauty of Fractals* (Peitgen and Richter, 1986).

References

Dijkstra, Edsger W. 1972. *Notes on structured programming*, Structured programming (C. A. R. Hoare, ed.), Academic Press, London and New York, 1972. [QA76.6.D33].

Mahler, Kurt. 1929. *Arithmetische Eigenschaften der Lösungen einer Klasse von Funktionalgleichungen*, Mathematische Annalen **101**, 342–366. Errata, **103** (1930), 532.

Peitgen, Heinz-Otto and Peter H. Richter. 1986. *The beauty of fractals : images of complex dynamical systems*, Springer, Berlin, New York. [QA447.P45].

Salomaa, Arto. 1981. *Jewels of formal language theory*, Computer Science Press, Rockville, Maryland. [QA267.3.S25].