Notes on the Thue-Morse Sequence

Terry R. McConnell
Syracuse University

September 17, 2014

Abstract

Exercise solutions and other miscellaneous results about the Thue-Morse sequence and related matters.

1 Introduction

The Thue-Morse sequence is an infinite sequence of binary digits that contains no overlapping substrings. For example, in the sequence that begins 0110110111... the 4th digit is shared between two consecutive instances of 0110 - it is an example of an overlap. In general, a sequence contains an overlap if and only if it contains a substring of the form avava, where a is either a 0 or 1 digit, and v is a finite string of digits. See, e.g., Lemma 2.1.1 of [1].

A sequence with these properties was first constructed by Axel Thue in 1906. (Thue published a related paper with further results in 1912.)

Thue's main use for his sequence was as a tool for constructing an example of an infinite word on 3 letters (a, b, and c, for example) that contains no square factors, i.e. substrings of the form XX, where X is a finite string of letters. It is easily seen to be impossible to avoid squares in long words of 2 letters. The only square-free words on the two letters a and b are a, ab, aba, b, ba, and bab.

While it necessarily cannot avoid squares, Thue's infinite word on 2 letters does successfully avoid overlaps, which can be viewed as more intimate forms of contact between substrings than mere adjacency (as in squares.) It should be noted that Thue's word contains no cubes - substrings of the form XXX - since these entail overlaps.

There are by now many known methods for constructing the Thue-Morse sequence $\mathbf{t}$. For example, consider the *substitution* μ, defined on the set of all finite non-empty sequences formed using the letters a and b, that replaces each a by ab, and each b by ba. Thus, e.g., $\mu(abba) = \mu^3(a) = abbabaab$. The successive finite strings obtained by iterating μ applied to the letter a give ever longer finite prefixes of $\mathbf{t}$. The resulting infinite word starts out

$$\mathbf{t} = abbabaabbaababbabaababbaabbabaabbaababbaabbabaababbabaabbaababbabaababbaabbab\ldots$$

The terms of the sequence have a strange, whipsaw-like regularity that never quite settles into a truly repetitive pattern. Rather, it has a fractal quality, repeating blocks and their mirror images on ever larger scales of length. Thus, it provides a simple model of chaotic behavior, and in this context it was rediscovered by Marston Morse in a 1921 study of chaotic behavior in certain dynamical systems.

Thue's square free word $\mathbf{m}$ on 3 letters starts out

$$\mathbf{m} = abcacbabcbacabcacbacabcbabcacbabcbacabcbabcacbacabcacbabcbacabcacbacabcbab\ldots$$

A construction of this sequence based on $\mathbf{t}$ is given on p. 26 of [1]. See the solution of Exercise 2.3.2 below for an alternative construction.

(Unless otherwise stated, all exercises and results are from [1].)

2 Exercise 2.2.1 Solution

The purpose of this exercise is to prove that μ is essentially the only morphism that fixes **t.** More precisely, if $\nu : \{a, b\}^* \to \{a, b\}^*$ is a morphism such that $\nu(\mathbf{t}) = \mathbf{t}$, then $\nu = \mu^n$ for some natural number n

The exercise suggests first proving the following result, of some independent interest, which characterizes the form of square factors of **t**: If wuu is a nonempty prefix of **t**, then either $|u| = 2^k$ or $|u| = 3(2^k)$, and, in either case, $|w|$ is a multiple of 2^k.

Let $S = \{ab, ba\}^*$. We shall use the result of Lemma 2.2.5, which says that if $x, y \in \{a, b\}$, and both v and xvy belong to S, then we cannot have x = y. In this connection, it is useful to note that if wu is a prefix of **t** and $|w|$ is even, then $w \in S$. If also u is even, then $u \in S$.

Suppose wuu is a prefix of **t** and both $|w|$ and $|u|$ are even. Then we have wuu = $\mu(rss)$ for some $r, s \in \{a, b\}^*$. By repeated application of this observation we may suppose that at least one of $|w|$ and $|u|$ is odd. Let us suppose that $|w|$ is odd, say $w = vb$ with $v \in S$. Then u must begin with a, since $bb \notin S$. We distinguish two subcases depending on whether $|u|$ is even or odd.

Case $|u|$ even: Suppose u begins with ab, say u = abz. If z is empty then **t** has a prefix of the form vbabab which contains an overlap. (Note that both sequences baba and abab do occur in **t**: baba starting at the 3rd place, and abab at the 11th place of **t**.) On the other hand if z is nonempty, then z must end in b since that character plus the following first a of the second u must belong to S. Thus, let z = yb. Then **t** has the prefix vbabybabyb. But this contains the overlap b(aby)b(aby)b. If, on the other hand, u = aaz with z nonempty, then z must still have the form yb (for an even simpler reason - we cannot have a cube aaa.) But then the prefix vbaaybaayb also contains an overlap: b(aay)b(aay)b.

Case $|u|$ odd: We may assume $|u| > 1$. Then u has the form abz. (aaz is impossible because the second u begins at an even position and must start with an element of S.) Since $|z| \geq 1$ it has the form z = ya or z = yb. If y is empty we have the example abaaba, occurring at the 16th place of **t**. (The example babbab occurs at the 12th place. These are the only possibilities with $|u| = 3$. For example, aabaab cannot occur because the preceding letter would either produce the cube aaa, or the overlap baabaab.) The case of nonempty y cannot occur. First suppose z = yb. Then we have a prefix vbabybabyb which has the overlap b(aby)b(aby)b. On the other hand, if z = ya, then we have a prefix vbabyaabya. But y must start with an a since by begins with an element of S. If y starts ab then we have the overlapping factor, babab. If y starts aax then we have a prefix vbabaaxaabaaxa. Note that both $|y|$ and $|x|$ are even. It follows that both axa and baaxaab belong to S. But this contradicts Lemma 2.2.5. In sum, when both $|w|$ and $|u|$ are odd, only $|u| = 1$ and $|u| = 3$ can occur.

Finally, we must consider the case $|w|$ even and $|u|$ odd. Clearly we cannot have $|u| = 1$ since neither aa nor bb belongs to S, So suppose $|u| \geq 3$. In this case uu belongs to S, so we way assume without loss of generality that u = abz. z must end in b since, combined with the first a of the second u, the two must be an element of S. Thus, u = abyb, and we have a prefix wabybabyb. But then both y and byb belong to S, and this again contradicts lemma 2.2.5.

Suppose ν is a morphism that fixes **t**. It follows from the foregoing that for some $n \geq 0$ and prefix $\mathbf{t}_k$ of **t** having odd length, we must have one of the following: $\nu(a)\nu(b)\nu(b) = \mu^n(\mathbf{t}_k uu)$ with $u = a, b, aba$, or bab.

We shall use the following simple observation: if $\mu^n(\mathbf{t}_k) = \mu^n(s)$, then $s = \mathbf{t}_k$. (This follows since $\mu^n(a) \neq \mu^n(b)$, but the two have equal length.)

Suppose $u = b$ so that $\nu(abba) = \mu^n(\mathbf{t}_k bb\mathbf{t}_k)$. Then $\nu(abba)$ is a prefix of **t** having length $2^n(2k + 2)$, so that $\nu(abba) = \mu^n(\mathbf{t}_{2k+2})$. It follows that $\mathbf{t}_{2k+2} = \mathbf{t}_k bb\mathbf{t}_k$. If $k \geq 3$ then $\mathbf{t}_k bbabb$ is a prefix of **t**, hence $\mathbf{t}_k bba$ belongs to S since it has even length. But then bb also belongs to S, and this is a contradiction. It follows that $k = 1$, i.e. $\nu(a) = \mu^n(a)$ and $\nu(b) = \mu^n(b)$, hence $\nu = \mu^n$.

The other cases are similar: If $u = a$ we conclude that $\mathbf{t}_{2k+2} = \mathbf{t}_k aa\mathbf{t}_k$. This is impossible even for $k = 1$, since it produces the cube factor aaa. If $u = bab$ we have $\mathbf{t}_{2k+6} = \mathbf{t}_k babbab\mathbf{t}_k$. If $k \geq 3$ then we would have an overlapping factor $babab$. Finally, if $u = aba$ then $\mathbf{t}_{2k+6} = \mathbf{t}_k abaaba\mathbf{t}_k$. If $k \geq 3$ then we have the overlapping factor $baabaab$.

3 Exercise 2.2.3 Solution

The purpose of the exercise is to show that there exist uncountably many infinite binary sequences that have no overlapping factors. (It is easy to show that there are at least countably many such sequences by removing ever longer prefixes from $\mathbf{t}$.) If $\mathbf{s}$ has no overlapping factors then we might try to produce a new example by adding a letter at the beginning. This will succeed unless it happens that $\mathbf{s}$ has some prefix that is a square. If $\mathbf{s}$ has prefix vava then $a\mathbf{s}$ begins with the overlap avava. In this connection, the following lemma will be useful:

Lemma 3.1. *If* $\mathbf{s}$ *has no overlapping factors and* $\mu(\mathbf{s})$ *has a square prefix, then* $\mathbf{s}$ *itself has a square prefix.*

Proof. Suppose $\mu(\mathbf{s}) = xaxay$. Then $\mu(b\mathbf{s}) = baxaxay$ has an overlapping factor. By Lemma 2.2.6, $b\mathbf{s}$ must have an overlapping factor, and by hypothesis this factor must be a prefix. Thus we may suppose $b\mathbf{s} = bvbvbw$. Cancelling b we see that $\mathbf{s}$ has the square prefix $vbvb$.

Next, we shall prove the following pair of statements by induction on n:

(3.1) $a\mu^n(b)$ has no overlaps and no prefix vbvb

(3.2) $b\mu^n(b)$ has no overlaps and no prefix vava

Both statements are obvious for $n = 0$. Suppose for a positive n we had $a\mu^n(b) = aubauby$. Then $ba\mu^n(b) = \mu(b)\mu^n(b) = \mu(b\mu^{n-1}(b))$ has an overlap at the beginning, but this is impossible by Lemma 2.2.6 and the inductive hypothesis. A similar argument shows that $b\mu^n b$ can have no prefix of the form $vava$. Repeated use of Lemmas 3.1 and 2.2.6 shows that $\mu^n(b)$ can have no square prefix at all, and thus both sequences $a\mu^n(b)$ and $b\mu^n(b)$ have no overlapping factors.

Let n_1 be an odd natural number and $n_j - n_{j-1}, j \geq 2$ be odd numbers. We shall next prove that

(3.3) $a\mu^{n_1}(b)\mu^{n_2}(b)\dots\mu^{n_k}(b)$ has no overlaps and no prefix vbvb,

and,

(3.4) $b\mu^{n_1}(b)\mu^{n_2}(b)\dots\mu^{n_k}(b)$ has no overlaps and no prefix vava.

The proof is by induction on k, the case $k = 1$ having been subsumed in the proof of (3.1) and (3.2).

If $b\mu^{n_1}(b)\mu^{n_2}(b)\dots\mu^{n_k}(b)$ had a prefix $vava$ then $ab\mu^{n_1}(b)\mu^{n_2}(b)\dots\mu^{n_k}(b) = \mu(a\mu^{n_1-1}(b)\mu^{n_2-1}(b)\dots\mu^{n_k-1}(b))$ begins with the overlap $avava$. Thus $a\mu^{n_1-1}(b)\mu^{n_2-1}(b)\dots\mu^{n_k-1}(b)$ has an overlap, and it follows from the inductive hypothesis that this must begin at the initial a. (If it began later in the sequence, we could factor at least μ^{n_1-1} and apply the inductive hypothesis to its argument.)

Thus $\mu^{n_1-1}(b)\mu^{n_2-1}(b)\dots\mu^{n_k-1}(b)$ has a prefix $uaua$. But

$$\mu^{n_1-1}(b)\mu^{n_2-1}(b)\dots\mu^{n_k-1}(b) = \mu^{n_1-1}(b\mu^{n_2-n_1}(b)\dots\mu^{n_k-n_1}(b)).$$

Since $s = b\mu^{n_2-n_1}(b)\dots\mu^{n_k-n_1}(b)$ has no overlap, by inductive hypothesis and Lemma 3.1, s must have a square prefix vv with $\mu^{n_1-1}(v) = ua$. Also, by inductive hypothesis, v must end in b. But this is impossible when $n_1 - 1$ is even. This completes the inductive step for (3.4). The argument is similar for obtaining the inductive step of (3.3).

Finally, we shall show that the strings in (3.4) are distinct for distinct sequences $n_1, n_2, \dots$. This suffices to complete the proof, since it is easy to show that the set of all such sequences is uncountable. By considering the first position at which two such sequences differ, it suffices to show that $\mu^i(b)b$ cannot be a prefix of $\mu^j(b)$ if $j > i$. (Note that the factor that would follow $\mu^i(b)$ begins with the letter b.) Let $k = j - i - 1 \geq 0$. We have $\mu^j(b) = \mu^i(ba\mu^k(b)) = \mu^i(b)\mu^i(a\mu^k(b))$, showing that $\mu^i(b)$ is followed in $\mu^j(b)$ by the letter a.

We remark that the set $ab\mu^{n_1}(b)\mu^{n_2}(b)\dots$ is also an uncountable set of words without overlapping factors. Write $ab\mu^{n_1}(b)\cdots = \mu(a\mu^{n_1-1}(b)\dots)$ with $n_1 - 1$ odd, and use the result of (3.3) and Lemma 2.2.6. This formulation is not necessary for the current problem, but is useful in Problem 2.3.6.

4 Exercise 2.3.6 Solution

The problem is to show that the set of all infinite square free words on three letters (a,b,c) is uncountable. It suffices to use Theorem 2.3.1 with the sequences constructed at the end of the solution 2.2.3. Any such sequence **a** starts with *abb* and can be uniquely parsed into tokens a, ab, or abb, using the obvious greedy algorithm. Moreover, exactly as in the text, there is a unique word **b** such that $\delta(\mathbf{b}) = \mathbf{a}$, where δ is the substitution $\delta(c) = a, \delta(b) = ab, \delta(a) = abb$. Then **b** is square free by Theorem 2.3.1. The cardinality of the set of **b** so obtained is the same as the cardinality of the sequences **a**, i.e., uncountable.

5 Exercise 2.3.2 Solution

This exercise tasks the reader to provide the omitted proof of Proposition 2.3.2. That proposition indicates a somewhat more satisfying construction of the square-free word **m** than the one based on **t**: Define a substitution ϕ on the alphabet $\{a, b, c\}$ by $\phi(a) = abc, \phi(b) = ac$, and $\phi(c) = b$. The infinite word **m** is then obtained by iterating ϕ on a, just as **t** was obtained by iterating μ on a.

For the proof, let δ be the substitution introduced above in the solution of Exercise 2.3.6. One checks that $\delta \circ \phi = \mu \circ \delta$ on $\{a, b, c\}$. Therefore $\delta \circ \phi^\omega(a) = \mu^\omega(\delta(a)) = \mu^\omega(abb) = \mathbf{t}$. Thus $\mathbf{b} = \phi^\omega(a)$ satisfies $\delta(\mathbf{b}) = \mathbf{t}$. But, as shown in the middle of page 26, the word with this property is unique, and since **m** is such a word, we have $\mathbf{b} = \mathbf{m}$.

References

[1] M.Lothaire,*Combinatorics on Words*, Cambridge Mathematical Library, Cambridge, UK, 1997.