The Thue-Morse sequence

V. Berthé

The Thue-Morse sequence appears in numerous fields and has been discovered and rediscovered in different contexts. Let us begin with an arithmetical definition. The Thue-Morse sequence is defined as the sequence $u = (u_n)_n$, which counts the sum modulo 2 of the digits of n in base 2 (u_n gives the parity of the number of 1's in the binary expansion of n). This sequence can also be generated by an iterative process called **substitution**. Let $\mathcal{A} = \{a, b\}$ and let $\mathcal{A}^*$ denote the set of words defined on the alphabet $\mathcal{A}$. Consider the map $\sigma : \mathcal{A} \to \mathcal{A}^*$ defined by $\sigma(a) = ab$ and $\sigma(b) = ba$. The map σ extends to a morphism of $\mathcal{A}^*$ by concatenation. We then can iterate σ and the nested words $\sigma^n(a)$ converge in the product topology to the infinite sequence which begins by $\sigma^n(a)$, for every n. This sequence is precisely the Thue-Morse sequence.

This sequence was first discovered by Prouhet in 1851 as a solution of the so called **Prouhet-Tarry-Escott problem** (Tarry and Escott re-introduced this problem after Prouhet in the years 1910-1920): consider a finite set of integers that can be partitioned into c classes with the same cardinality s such that the sums of the elements, the sums of squares, ..., the sums of the k-th powers in each class are independent of the class. If the sums of the $(k+1)$-th powers are not equal, then the solution is said to be an exact solution of order k. The Thue-Morse sequence provides a solution of degree k exactly (when $c = 2$) by considering the classes $\{0 \leq n \leq 2^{k+1} - 1,\ u_n = a\}$ and $\{0 \leq n \leq 2^{k+1} - 1,\ u_n = b\}$. Note that this partition is conjectured to be the unique partition of the set $\{0, 1, \ldots, 2^{k+1} - 1\}$ into 2 classes providing an exact solution of degree k. For more results on this subject, see the survey [3].

The sequence u was next rediscovered by Thue in 1912 [8]. Thue tried to construct arbitrarily long words on a two letter alphabet without cubes, i.e., without factors of the form www, where w is a non-empty word. It is easily seen that there are no infinite square-free sequences on two letters. It is then natural to ask whether there are infinite sequences free of powers $2 + \varepsilon$, for any $\varepsilon > 0$. Such a sequence is called overlap-free; in other words, none of its factors is of the form $xuxux$. In fact, the Thue-Morse sequence is an infinite overlap-free word on two letters. Moreover all overlap-free words are derived from this sequence. Furthermore, if one defines the Thue-Morse sequence on $\{0, 1\}$ (by mapping a to 1 and b to -1), then the sequence $(u_{n+1} - u_n)_n$ defined on the three-letter alphabet $\{-1, 0, 1\}$ is square-free. By applying the morphism

$\mu(-1) = a, \mu(0) = ab, \mu(1) = abb$, one also gets an infinite cube-free word on two letters. These results have been extended whithin the theory of avoidable and unavoidable patterns in strings. Note that such combinatorial properties were used to solve various algebraic problems, as to provide a negative answer to the Burnside problem. For more information on the subject, see the references and results quoted in [6].

The sequence u was next introduced by Morse [7] in 1921 in order to show the existence of non-periodic recurrent geodesics over simply connected surfaces with constant negative curvature, by coding geodesics by an infinite sequence of 0's and 1's, according to which boundary of the surface it meets. Indeed the Thue-Morse sequence is a **recurrent** sequence, i.e., every factor appears in an infinite number of places with bounded gaps. In other words, the **symbolic dynamical system** generated by the Thue-Morse sequence is **minimal** (see for instance [4]).

The Thue-Morse sequence is a typical example of a k-automatic sequence. Actually, like every fixed point of a substitution of constant length, it can be generated by a finite machine called a **finite automaton** as follows. A k-automaton is given by a finite set of states $\mathcal{S}$, one state being called the initial state, by k maps from $\mathcal{S}$ into itself (we denote them $0, 1, \ldots, k-1$) and by an output map φ from $\mathcal{S}$ into a given set Y. Such an automaton generates a sequence with values in Y as follows: feed the automaton with the digits of the base-k expansion of n, by starting with the initial state; then define u_n as the image under φ of the reached state. In the Thue-Morse case, the automaton has two states, say $\{a, b\}$, the map 0 maps each state to itself whereas the map 1 exchanges both states, the output map is the indentity map and the state a is the initial sate.

Automatic squences have many nice characterizations (see for instance the survey [1])- Automatic sequences are exactly the letter-to-letter images of fixed points of constant length substitutions. Furthermore, this is equivalent to the fact that the following subset of subsequences (called the k-kernel) $\{(u_{k^t n+r})_n,\ t \geq 0,\ 0 \leq r \leq k^t - 1\}$ is finite or to the fact that the series $\sum_n u_n X^n$ is algebraic over $\mathbb{F}_k(X)$, in the case k is a prime power. Note that on the other hand the real number with dyadic expansion the Thue-Morse sequence is transcendental. For more references and connections with physics, see [2].

Consider the following sequence $v = (v_n)_n$, which counts modulo 2 the number of 11's (possibly with overlap) in the base-2 expansion of n. The sequence v is easily seen to have a finite 2-kernel and hence to be be 2-automatic. This sequence was introduced independently by Rudin and Shapiro (see the references in [4]) in order to mimimize uniformly $|\sum_{n=0}^{N-1} a_n e^{int}|$, for a sequence (a_n) defined over $\{-1, 1\}$. The *Rudin-Shapiro* sequence hence provides $\sup_t |\sum_{n=0}^{N-1} v_n e^{int}| \leq (2 + \sqrt{2})\sqrt{N}$.

The dynamical system generated by each of the two sequences u and v is **strictly ergodic**, since both underlying substitutions are **primitive** (see for instance [4]). But although very similar in their definition, these two sequences share very distinct spectral properties. The Morse system has a singular simple spectrum whereas the dynamical system generated by the Rudin-Shapiro sequence provides an example of a system with finite spectral multiplicity and a Lebesgue component in the spectrum. For more references on the ergodic, spectral and harmonic properties of substitutive sequences, see [4].

If (almost) everything is known concerning the Thue-Morse and the Rudin-Shapiro sequences, then the situation is completely different for the fascinating **Kolakoski** sequence. The Kolakoski sequence is the self-determined sequence defined over the alphabet $\{1, 2\}$ as follows. The sequence begins with 2 and the sequence of lengths of the consecutive strings of 2's and 1's is the sequence itself. Hence this sequence is equal to $22112122122112\ldots$ For a survey of related properties and conjectures, see [5].

References

[1] J.-P. ALLOUCHE *Automates finis en théorie des nombres*, Exp. Math. **5** (1987), 239–266.

[2] *"Beyond quasicrystals"*, Actes de l'École de Physique Théorique des Houches: "Beyond quasicrystals", Les Éditions de Physique, Springer, (1995).

[3] P. BORWEIN, C. INGALLS *The Prouhet-Tarry-Escott problem revisited*, Enseign. Math. **40** (1994), 3–27.

[4] M. QUEFFÉLEC *Substitution dynamical systems. Spectral analysis*, Lecture Notes in Mathematics **1294**, Springer-Verlag, 1987.

[5] F. M. DEKKING *What is the long range order in the Kolakoski sequence*, The mathematics of long-range aperiodic order (Waterloo, ON, 1995), 115–125, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., 489, Kluwer acad. Publ., Dordrecht, 1997.

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

[7] M. MORSE *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84–100.

[8] A. THUE Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen (1912), Selected mathematical papers of Axel Thue, Universiteitsforlaget (1977).