

Enumeration and Decidable Properties of Automatic Sequences

Narad Rampersad
Department of Mathematics
University of Liège
4000 Liège
Belgium

Emilie Charlier & Jeffrey Shallit
School of Computer Science, University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
shallit@cs.uwaterloo.ca
<http://www.cs.uwaterloo.ca/~shallit>

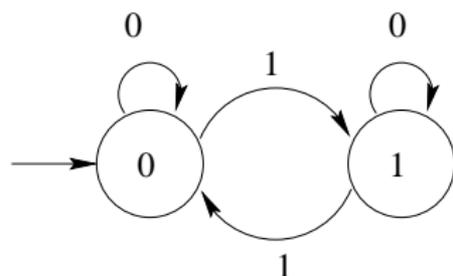
Automatic sequences

A sequence $(a_n)_{n \geq 0}$ is said to be **k -automatic** if there exists

- ▶ a deterministic finite automaton
- ▶ (with an output $\tau(q)$ associated with each state q)
- ▶ that, on input the base- k representation of n , reaches some state p
- ▶ and $\tau(p) = a_n$.

Example: the Thue-Morse sequence

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
t_n	0	1	1	0	1	0	0	1	1	0	0	1	0	1	1	0



Here t_n counts the **number of 1's (mod 2)** in the base-2 representation of n .

Fixed points of uniform morphisms

The Thue-Morse sequence can also be viewed in another way: as the fixed point of the uniform morphism $0 \rightarrow 01, 1 \rightarrow 10$.

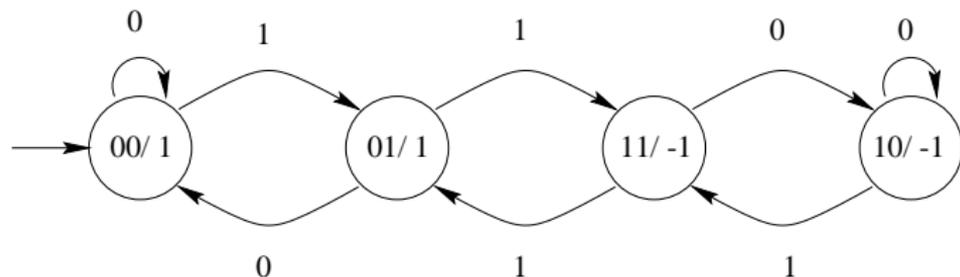
(A morphism is uniform if each letter gets mapped to a word of the same length.)

The class of fixed points of uniform morphisms has been widely studied.

However, a larger class is the class of k -automatic sequences.

The Rudin-Shapiro sequence

The *Rudin-Shapiro* sequence $\mathbf{r} = (r_n)_{n \geq 0}$, is defined by $r_n = 1$ (resp. -1) according to whether the number of (possibly overlapping) occurrences of “11” in the binary expansion of n is even (resp. odd). Then $(r_n)_{n \geq 0}$ is 2-automatic, since it is generated by the automaton below.



Here the meaning of a state labeled ab/c is that the running sum of the number of occurrences of “11” so far is congruent to a modulo 2, the last digit input was b , and the output is c .

The Rudin-Shapiro sequence

The Rudin-Shapiro sequence \mathbf{r} is **not** a fixed point of a uniform morphism.

Proof. If it were the fixed point of a k -uniform morphism for k not a power of 2, then by a theorem of Cobham, it would be ultimately periodic.

But it is known that the largest power in Rudin-Shapiro is 4, contradiction.

So it must be the fixed point of a 2^j -uniform morphism h for some $j \geq 1$.

Now \mathbf{r} starts 00; if $\mathbf{r} = h(\mathbf{r})$ then \mathbf{r} starts with $h(0)h(0)$. This means that $\mathbf{r}[2^j - 1] = \mathbf{r}[2^{j+1} - 1]$.

But the number of occurrences of 11 in $2^{j+1} - 1$ is of opposite parity of the number of occurrences of 11 in $2^j - 1$, a contradiction.

The Rudin-Shapiro sequence

However, \mathbf{r} is the **image** (under a coding – a letter-to-letter morphism) of a fixed point of a uniform morphism:

$$a \rightarrow ab; b \rightarrow ac; c \rightarrow db; d \rightarrow dc$$

followed by the coding $a, b \rightarrow 1; c, d \rightarrow -1$.

The class of k -automatic sequences coincides with the class of codings of fixed points of uniform morphisms (Cobham's theorem).

Some decidable properties of automatic sequences

Ultimate periodicity is decidable (Honkala, 1986; Leroux, 2005).

Squarefreeness and **overlapfreeness** is decidable (Allouche, Rampersad, Shallit, 2009).

More generally, the property of avoiding α -powers (α rational) is decidable (Allouche, Rampersad, Shallit, 2009), as is **containing infinitely many α -powers**, or **containing infinitely many distinct α -powers**.

Recurrence and **uniform recurrence** are decidable (Nicolas and Pritykin, 2009).

All of these are subsumed by the following result, which follows from the work of Büchi, Bruyère, Michaux, Villemaire, and others:

Theorem.

If we can express a property $P(n)$ of an integer n using quantifiers, logical operations, integer variables, the operations of addition, subtraction, indexing of a k -automatic sequence \mathbf{x} , and comparison of integers or elements of \mathbf{x} , then $\exists n P(n)$ and $\exists_{\infty} n P(n)$ are decidable.

A decidability theorem

Proof.

Given an automaton M generating \mathbf{x} , we transform it into an automaton M' accepting the base- k representations of those n for which the property holds.

The existential quantifier is implemented using nondeterminism, and the universal quantifier is implemented using suitable negations.

Addition and subtraction are performed digit-by-digit, keeping track of carries; comparison of integers is done similarly.

Checking $\exists nP(n)$ is then done by seeing if M' accepts anything (has a path from initial state to an accepting state).

Checking $\exists_{\infty} nP(n)$ is done by seeing if M' accepts infinitely many strings (has a path from initial state to an accepting state of length $\geq r$, the number of states).

An example

Consider the property of having an overlap.

A sequence \mathbf{x} has an overlap beginning at position i if and only if there exists an index $\ell \geq 1$, such that $\mathbf{x}[i + j] = \mathbf{x}[i + j + \ell]$ for $0 \leq j \leq \ell$.

Given a DFA M_1 generating \mathbf{x} , we first create an NFA M_2 that on input (i, ℓ) accepts if there exists j , $0 \leq j \leq \ell$, such that $\mathbf{x}[i + j] \neq \mathbf{x}[i + j + \ell]$. To do this, M_2 guesses the base- k representation of j , digit-by-digit, verifies that $j \leq \ell$, computes $i + j$ and $i + j + \ell$ on the fly, and accepts if $\mathbf{x}[i + j] \neq \mathbf{x}[i + j + \ell]$.

We now convert M_2 to a DFA using the subset construction, and change the “finality” of each state, obtaining a DFA M_3 . Then M_3 accepts those pairs (i, ℓ) such that $\mathbf{x}[i + j] = \mathbf{x}[i + j + \ell]$ for all j with $0 \leq j \leq \ell$. Now we create an NFA M_4 that on input i guesses ℓ and accepts if M_3 accepts (i, ℓ) .

When we talk about “accepting pairs (i, ℓ) ”, we really mean their base- k expansion.

The base- k expansion of a pair is defined by taking the canonical base- k expansion of both integers, and then padding the shorter with leading zeroes (or trailing zeroes, if we are using the reversed representation).

Problem: if “on input i we guess ℓ and accept if M accepts (i, ℓ) ”, then it could be that ℓ is much longer than i .

To handle this, we also allow non-canonical expansions where both elements of a pair have leading zeroes.

Additional decidability properties

A word w is *bordered* if it begins and ends with the same word x with $0 < |x| \leq |w|/2$.

An example in English is **ing**oing — it begins and ends with **ing**.

Otherwise it is *unbordered*.

Theorem.

Let $\mathbf{x} = a(0)a(1)a(2)\cdots$ be a k -automatic sequence. Then the associated infinite sequence $\mathbf{b} = b(0)b(1)b(2)\cdots$ defined by

$$b(n) = \begin{cases} 1, & \text{if } \mathbf{x} \text{ has an unbordered factor of length } n; \\ 0, & \text{otherwise;} \end{cases}$$

is k -automatic.

Proof. The sequence \mathbf{x} has an unbordered factor of length n

iff

$\exists j \geq 0$ such that the factor of length n beginning at position j of \mathbf{x} is unbordered

iff

there exists an integer $j \geq 0$ such that for all possible lengths l with $1 \leq l \leq n/2$, there is an integer i with $0 \leq i < l$ such that the supposed border of length l beginning and ending the factor of length n beginning at position j of \mathbf{x} actually differs in the i 'th position

iff

there exists an integer $j \geq 0$ such that for all integers l with $1 \leq l \leq n/2$ there exists an integer i with $0 \leq i < l$ such that $\mathbf{x}[j+i] \neq \mathbf{x}[j+n-l+i]$. ■

Unbordered factors

Example. Consider the problem of determining for which lengths the Thue-Morse sequence has an unbordered factor. Currie & Saari (2009) proved that if $n \not\equiv 1 \pmod{6}$, then there is an unbordered factor of length n .

However, this is not a necessary condition, as

$$\mathbf{t}[39..69] = 0011010010110100110010110100101,$$

which is an unbordered factor of length 31. They left it as an open problem to give a complete characterization of the lengths for which \mathbf{t} has an unbordered factor. Our method shows the characteristic sequence of such lengths is 2-automatic.

Further, we conjecture that there is an unbordered factor of length n in \mathbf{t} if and only if the base-2 expansion of n (starting with the most significant digit) is not of the form $1(01^*0)^*10^*1$.

In principle this could be verified, purely mechanically, by our method, but we have not yet done so.

Additional decidability properties

The following question is decidable, but does not seem to follow from the theorem mentioned previously: given a k -automatic sequence, does it contain powers of arbitrarily large exponent?

Proof. \mathbf{x} has powers of arbitrarily high exponent iff the set of pairs

$$S := \{(n, j) : \exists i \geq 0 \text{ such that for all } t \text{ with } 0 \leq t < n \text{ we have} \\ \mathbf{x}[i + t] = \mathbf{x}[i + j + t]\}$$

contains pairs (n, j) with n/j arbitrarily large iff

for all $i \geq 0$ S contains a pair (n, j) with $n > j \cdot 2^i$ iff

L , the set of base- k encodings of pairs in S , contains, for each i , strings ending in

$$\overbrace{[* , 0][* , 0] \cdots [* , 0]}^i [b , 0]$$

for some $b \neq 0$, where $*$ means any digit. This is decidable. 

Theorem. It is decidable if a k -automatic sequence is recurrent.

Proof. $(a_n)_{n \geq 0}$ is recurrent if every factor that occurs, occurs infinitely often.

It suffices to test whether each factor that occurs, occurs again in a later position.

In other words, $(a_n)_{n \geq 0}$ is recurrent iff for all $i \geq 0$, $\ell \geq 1$, there exists $j > i$ such that $a[i..i + \ell - 1] = a[j..j + \ell - 1]$.

This is decidable, by our Theorem.

The class of k -automatic sequences is an interesting one but

- it must be over a finite alphabet.

Hence they cannot be used to enumerate unbounded quantities.

We would like a generalization over an infinite alphabet.

k -regular sequences

To do so, we start from the following characterization of automatic sequences.

A sequence $(a_n)_{n \geq 0}$ is k -automatic if and only if the set

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

is finite.

A sequence $(a_n)_{n \geq 0}$ over \mathbb{Z} is k -regular if and only if there exists a finite set S of sequences such that each sequence in

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

can be expressed as a linear combination of sequences in S .

Examples of k -regular sequences

- ▶ $s_k(n)$, the sum of the base- k digits of n
- ▶ the Mallows-Propp sequence: the unique monotone sequence $(a(n))_{n \geq 0}$ of non-negative integers such that $a(a(n)) = 2n$ for $n \neq 1$
- ▶ the number of overlap-free binary words of length n (Carpi; Cassaigne)

Theorem.

Let S be a set of pairs of non-negative integers such that the language of base- k representations

$$L = \{(m, n)_k : (m, n) \in S\}$$

is regular (accepted by a finite automaton).

Then the sequence $(a_m)_{m \geq 0}$ defined by

$$a_m = |\{n : (m, n) \in S\}|$$

is k -regular.

With this theorem we can recover or improve many results from the literature.

Improved results

If $\mathbf{a} = (a_n)_{n \geq 0}$ is a k -automatic sequence, then the following associated sequences are k -regular.

- ▶ its **subword complexity function**, $n \rightarrow$ number of distinct factors of length n
 - ▶ Previously known for fixed points of k -uniform morphisms (Mossé, 1996)
- ▶ its **palindrome complexity function**, $n \rightarrow$ number of distinct factors of length n that are palindromes
 - ▶ Previously known for fixed points of primitive k -uniform morphisms (Allouche, Baake, Cassaigne, Damanik, 2003)
- ▶ its **sequence of separator lengths** (length of smallest factor that begins at position n and does not occur previously)
 - ▶ Previously known for fixed points of k -uniform circular morphisms (Garel, 1997)

If $\mathbf{a} = (a_n)_{n \geq 0}$ is a k -automatic sequence, then the following associated sequences are k -regular sequences:

- ▶ the number of distinct square factors of length n ; the number of squares beginning at (centered at, ending at) position n ; the length of the longest square beginning at (centered at, ending at) position n ; the number of palindromes beginning at (centered at, ending at) position n ; the number of distinct recurrent factors of length n ; etc.,
 - ▶ Previously known for the Thue-Morse sequence (Brown, Rampersad, Shallit, Vasiga, 2006)

If $(a_n)_{n \geq 0}$ is a k -automatic sequence, then the following associated sequences are k -regular sequences:

- ▶ The **number of unbordered factors of length n**
 - ▶ For the Thue-Morse sequence we have a conjectured recursive expression for the number $f(n)$ of unbordered factors of length n . This could, in principle, be verified purely mechanically, by our method.

If $(a_n)_{n \geq 0}$ is a k -automatic sequence, then the following associated sequences are k -regular sequences:

- ▶ The **recurrence function** of \mathbf{a} , $n \rightarrow$ the smallest integer t such that every factor of length t of \mathbf{a} contains every factor of length n
- ▶ The **appearance function** of \mathbf{a} , $n \rightarrow$ the smallest integer t such that the prefix of length t of \mathbf{a} contains every factor of length n

We illustrate the idea behind all these results with an example: subword complexity (number of distinct factors).

The number of distinct factors of length n in \mathbf{a} equals the number of **first occurrences of each factor**.

This equals the number of indices i such that there is no index $j < i$ with the **factor of length n beginning at position i** equal to the **factor of length n beginning at position j** .

So given a k -automatic sequence $\mathbf{a} = (a_n)_{n \geq 0}$ consider the set

$$S = \{(n, i) : \text{for all } j \text{ with } 0 \leq j < i \text{ there exists } t$$
$$\text{with } 0 \leq t < n \text{ such that } a_{i+t} \neq a_{j+t}\}.$$

From our first theorem, the set of base- k encodings of S is a regular language and hence by our second theorem, the subword complexity function is k -regular.

In a paper to be presented at WORDS 2011, I have shown that

- ▶ Given a regular language L encoding a set S of pairs of integers, the quantity $\sup_{(p,q) \in S} \frac{p}{q}$ is either infinite or rational, and it is computable
- ▶ The **critical exponent** of an automatic sequence (exponent of the largest power of any factor) is a rational number and is computable.
- ▶ The **optimal constant for linear recurrence** for an automatic sequence is rational and computable.

Linear recurrence

A sequence $\mathbf{a} = (a_n)_{n \geq 0}$ is **linearly recurrent** if there is a constant C such that for all $\ell \geq 0$, and all factors x of length ℓ occurring in \mathbf{a} , any two consecutive occurrences of x are separated by at most $C\ell$ positions.

Given \mathbf{a} , can we determine the smallest value of C that works?

The idea is, given the automaton for \mathbf{a} , to construct an automaton accepting the language of pairs (d, ℓ) such that

- (a) there is some factor of length ℓ for which there is another occurrence at distance d and
- (b) this occurrence is actually the very next occurrence.

Then $\sup_{(d, \ell) \in S} \frac{d}{\ell}$ gives the optimal C .

Is the property of abelian-squarefreeness (abelian k -power-freeness) decidable for automatic sequences?

For fixed points of some morphisms, this can be done (Currie & Rampersad, 2011).

Can you find a decision procedure for similar questions about **morphic** sequences (images of fixed points of morphisms, not necessarily uniform)?

We can do this for **some** morphic sequences, such as the Fibonacci word, where addition can be performed in some associated base.

But the general case is still open.