

Combinatorics on words

Introduction to some problems

Jean Berstel

`http://www-igm.univ-mlv.fr/~berstel`

Institut Gaspard-Monge
Université de Marne-la-Vallée
France

Outline

- Subword complexity of finite words.
- Construction of infinite words.
- Powers and periodic words.
- Open problems.

Subword complexity of finite words

$p_x(n)$ = number of distinct factors of x of length n .

Example:

1) $x = 0011001$

n	0	1	2	3	4	5	6	7
$p_x(n)$	1	2	4	4	4	3	2	1

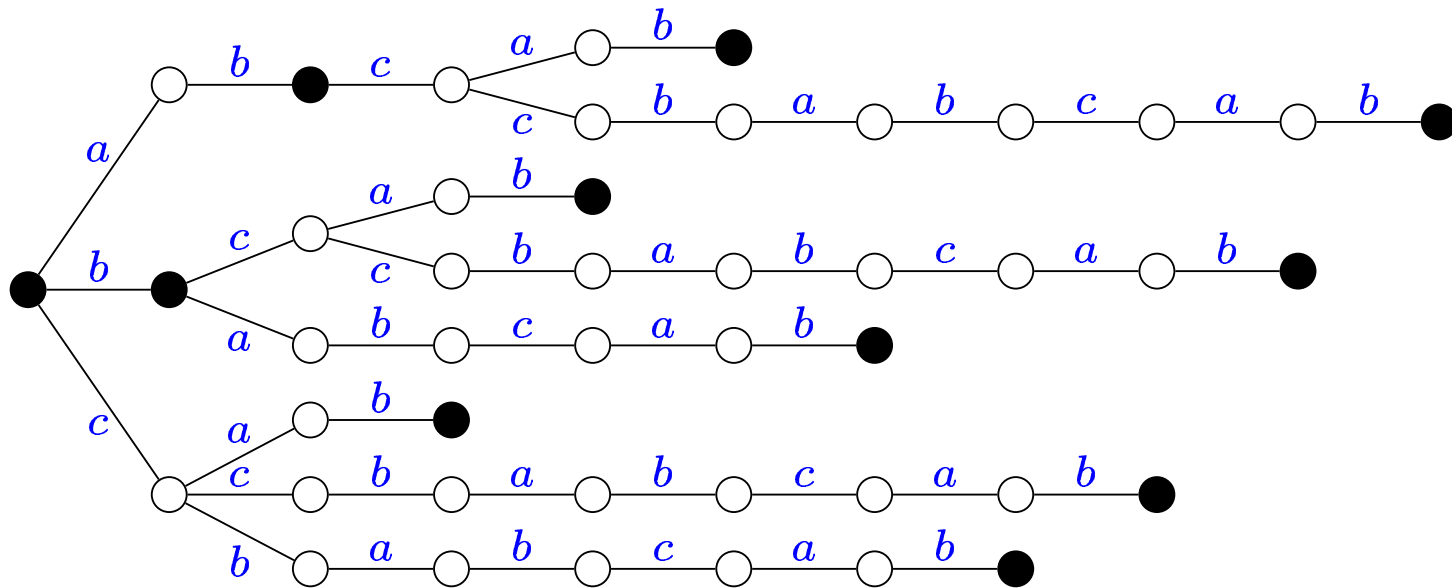
$$M_x = 2, H_x = 4$$

$$M_x = \min\{i \mid p_x(i) \text{ is maximal}\}$$

$$H_x = \max\{i \mid p_x(i) \text{ is maximal}\}$$

Subword complexity of finite words (cont'd)

The suffix tree of the word $w = abccbabcab$



n	0	1	2	3	4	5	6	7	8	9	10
$p_w(n)$	1	3	6	7	7	6	5	4	3	2	1

$$M_w = 3, H_w = 4$$

Subword complexity of finite words (3)

A factor is *unrepeated* if it appears only once in w .

Theorem (Carpi & de Luca) H_w is the smallest n such that any factor of length $\geq n$ is unrepeated.

Theorem (Carpi & de Luca, Levé & Séébold) For any word w , there is an integer M_w such that:

1. $p_w(i) < p_w(i + 1)$ for $0 \leq i < M_w$,
2. $p_w(i) = p_w(i + 1)$ for $M_w \leq i < H_w$,
3. $p_w(i) = p_w(i + 1) + 1$ for $H_w \leq i < |w|$.

Theorem (Carpi & de Luca) A word w is determined by its factors of length at most $H_w + 1$ (and even by the poset of these factors).

Subword complexity of infinite words

Theorem (Morse & Hedlund, Coven & Hedlund).

Let x be an infinite word over k letters. The following are equivalent:

1. x is ultimately periodic,
2. $p_x(n) = p_x(n + 1)$ for some n ,
3. $p_x(n) < n + k - 1$ for some $n \geq 1$,
4. $p_x(n)$ is bounded.

Thus, either $p_x(n)$ is ultimately constant or $p_x(n) \geq n + 1$ for all n . A word is **Sturmian** if $p_x(n) = n + 1$ for all $n \geq 0$. A Sturmian word is binary because $p_x(1) = 2$.

Construction of infinite words

Let (w_n) be a sequence of finite words. The *limit*

$$x = \lim_{n \rightarrow \infty} w_n$$

exists if, for each i , there exists N_i such that for all $n \geq N_i$,

$$|w_n| \geq i \quad \text{and} \quad x(i) = w_n(i) .$$

Example. For $w_n = a^n b^{n^2}$, one gets $\lim w_n = a^\omega$.

Explicit description

Characteristic word of a set of integers.

a) Squares $0, 1, 4, 9, \dots$

11001000010000001000...

b) The *spectrum* of $\frac{1+\sqrt{5}}{2}$ is the set $S_\tau = \{\lfloor n\tau \rfloor : n \geq 1\}$. The infinite binary word f is defined by

$$f_n = \begin{cases} a & \text{if } n+1 \in S_\tau \\ b & \text{otherwise} \end{cases}$$

$f = abaababaabaababaababa \dots$

Explicit description

c) Thue-Morse word

$$t = 01101001100101101001011001101001 \dots$$

defined by

$t_n =$ the number of 1's in the binary expansion $\text{bin}(n)$ of n modulo 2.

d) More generally, by a finite automaton working on binary expansion :

$t_n = 1$ iff $\text{bin}(n)$ is accepted by the automaton.

These are *automatic sequences*.

Infinite products

Any infinite product $x_0x_1 \cdots x_n \cdots$ of nonempty words has a limit.

$$c = 0110111001011101111000 \cdots$$

The *Champernowne* word is the product of the words $\text{bin}(n)$ (binary representation of n).

- Every word is factor of c : $p_c(n) = 2^n$.
- c is recurrent : every factor that appears in c appears infinitely many times.
- It is not uniformly recurrent : the gap between consecutive occurrences of a given factor is not bounded.

Words generated by iterating a morphism

A morphism $h : A^* \rightarrow A^*$ is *prolongable* in the letter a if

$$h(a) = ax$$

for some word x with $h^n(x) \neq \varepsilon$ for all $n \geq 0$. Then

$$h^2(a) = axh(x)$$

$$h^3(a) = axh(x)h^2(a)$$

and the sequence $(h^n(x))$ converges to

$$h^\omega(a) = axh(x)h^2(x) \cdots h^n(x) \cdots$$

Words generated by a morphism (cont'd)

$$h : \begin{array}{l} a \mapsto aba \\ b \mapsto abb \end{array}$$

Then

$$h^3(a) = a \, ba \, abbaba \, abaabbabbabaabbaba$$

Of course, $h^n(a)$ is always a prefix of $h^{n+1}(a)$. If $x = h^\omega(a)$, then

$$x = h(x)$$

that is x is a fixed point of h .

Words generated by a morphism (cont'd)

$$h : \begin{array}{l} a \mapsto aba \\ b \mapsto abb \end{array}$$

Then

$$\begin{array}{ll} u_1 = h(a) = aba & v_1 = h(b) = abb \\ u_2 = h^2(a) = abaabbaba & v_2 = h^2(b) = abaabbabb \\ & = u_1 v_1 u_1 \end{array}$$

and

$$\begin{aligned} u_3 = h^3(a) &= abaabbaba \, abaabbabb \, abaabbaba \\ &= h^2(a) h^2(b) h^2(a) = u_2 v_2 u_2 \end{aligned}$$

This gives a system of recurrence relations for the words u_n and v_n .

Substitution

$f : B^* \rightarrow B^*$ a morphism prolongable in the letter b .

$g : B^* \rightarrow A^*$ be a letter-to-letter morphism.

The pair (f, g) is a *substitution*. It generates the word $g(f^\omega(b))$.

The word of squares

$$s = 1100100001000000100 \dots$$

is generated by

$$\begin{array}{ll} a \mapsto a1 & a \mapsto 1 \\ f : 1 \mapsto 001 & g : 1 \mapsto 1 \\ 0 \mapsto 0 & 0 \mapsto 0 \end{array}$$

Indeed

$$f^\omega(a) = a100100001 \dots$$

Tag machine

A *Tag machine* is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921) : *is it decidable whether, for a word w on the tape, the reading head can reach the writing head ?*

Kolakoski sequence $\frac{22}{2} \frac{11}{2} \frac{21}{11} \frac{22}{2} \frac{12}{1} \frac{22}{2} \frac{11}{2} \dots$

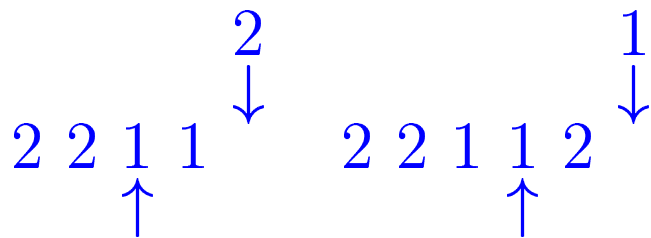
2 2 1 1 2
↑ ↓

Tag machine

A *Tag machine* is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921) : *is it decidable whether, for a word w on the tape, the reading head can reach the writing head ?*

Kolakoski sequence $\frac{22}{2} \frac{11}{2} \frac{21}{11} \frac{22}{2} \frac{12}{1} \frac{22}{2} \frac{11}{2} \dots$

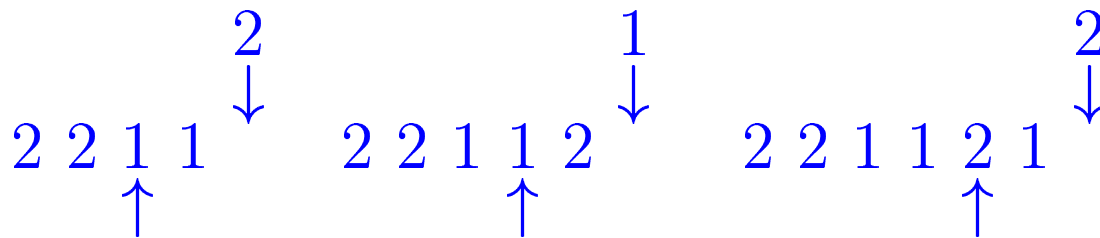


Tag machine

A *Tag machine* is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921) : *is it decidable whether, for a word w on the tape, the reading head can reach the writing head ?*

Kolakoski sequence $\frac{22}{2} \frac{11}{2} \frac{21}{11} \frac{22}{2} \frac{12}{1} \frac{22}{2} \frac{11}{2} \dots$

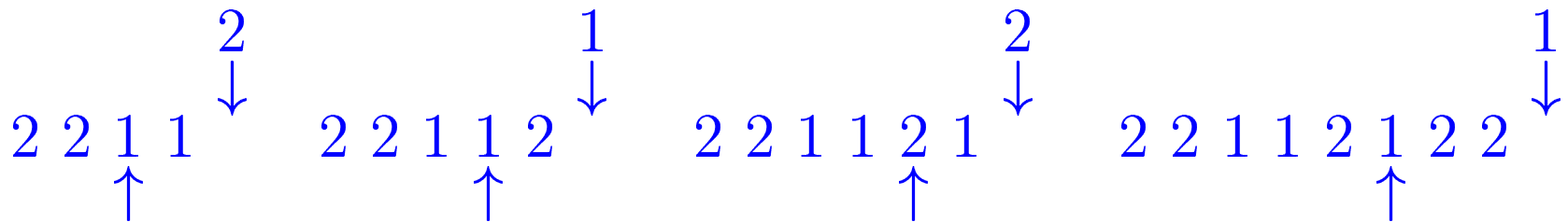


Tag machine

A *Tag machine* is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

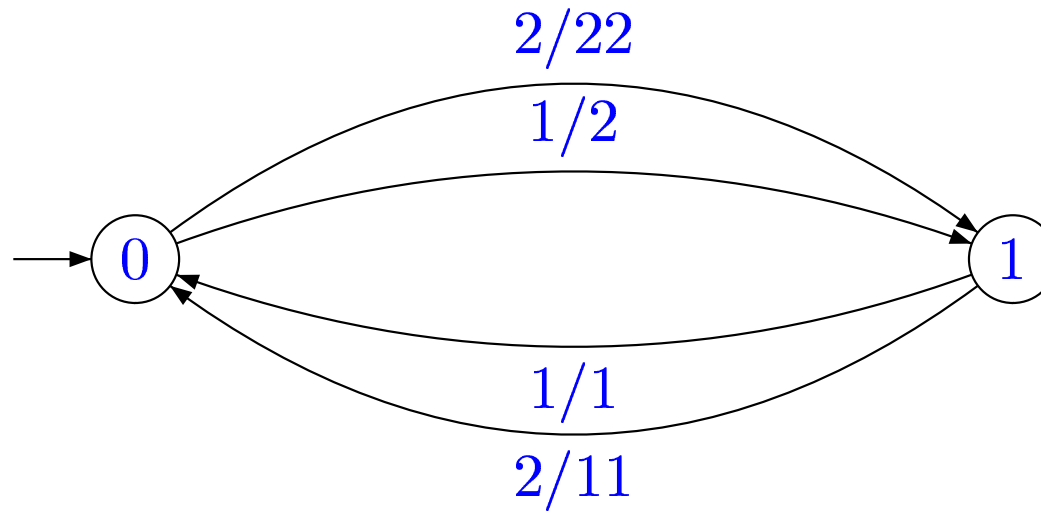
The question asked originally by Post (1921) : *is it decidable whether, for a word w on the tape, the reading head can reach the writing head ?*

Kolakoski sequence $\frac{22}{2} \frac{11}{2} \frac{21}{11} \frac{22}{2} \frac{12}{1} \frac{22}{2} \frac{11}{2} \dots$



Iterating sequential functions

A sequential function is a morphism with states. For the Kolakoski sequence:



2

22

2211

221121

221121221

...

The sequential machine can be viewed as a special case of a Tag machine, when reading and writing on the same tape.

Toeplitz words

$x = x_0?x_1?x_2?\dots$ with x_n words and $?$ a placeholder.

$y = a_0a_1a_2\dots$ with a_n letters.

The *Toeplitz* product is

$$x\tau y = x_0a_0x_1a_1x_2a_2\dots$$

Example Consider $x = ab?ab?ab?\dots = (ab?)^\omega$. Then

$$x\tau x = ab\mathbf{a}ab\mathbf{b}ab?ab\mathbf{a}ab\mathbf{b}ab?\dots = (abaabbab?)^\omega$$

$$x\tau x\tau x = abaabbab\mathbf{a}abaabbab\mathbf{b}abaabbab\mathbf{a}abaabbab\dots$$

The limit exists, and is of course a fixed point:

$$y = x\tau y$$

Toeplitz words and iterated morphisms

We consider words $x = w^\omega$ of type (p, q) , that is $|w| = p$ and w contains q placeholders. (E.g. $w = aa?b?$ has type $(5, 2)$.)

Theorem (Cassaigne & Karhumäki) *Let y be generated by a word of type (p, q) .*

- *if $q = 1$, then y can be obtained by iterating a morphism;*
- *if q divides p , then y can be obtained by a substitution;*
- *otherwise y can be obtained by iterating periodically q morphisms.*

The word obtained by $x = ab?ab?ab? \dots = (ab?)^\omega$ is generated by

$$a \mapsto aba$$

$$b \mapsto abb$$

Palindromic closure

The (right) *palindromic closure* w^π of a word w is the shortest palindrome word that starts with w .

$$(01001)^\pi = 01001\mathbf{0} \quad (01001010)^\pi = 01001010\mathbf{010}$$

Given a word $d = a_0a_1 \cdots a_n \cdots$, the word d^π *directed* by d is the limit of the sequence $u_0 = \varepsilon$ and

$$u_{n+1} = (u_n a_n)^\pi$$

For $d = 010101 \cdots$ one gets

$$\begin{array}{ll} 0 & \underline{0} \\ 1 & 0\underline{1}0 \\ 0 & 010\underline{0}10 \\ 1 & 010010\underline{1}0010 \end{array}$$

The limits of binary words are the Sturmian words. A word d^π is a fixed point of a morphism if and only if d is periodic (de Luca, Justin & Pirillo).

Repetitions

- A *repetition* is a non trivial power of a word.
- For example, *ababa* is a power of exponent $5/2$. The fractional power $u^{p/q}$ is defined when q divides the length of u .

$$\begin{aligned} |u| &= kq \\ p/q &= n + r/q \quad 0 \leq r < q \\ u^{p/q} &= u^n u' \quad |u'| = rk, u' \text{ prefix of } u \end{aligned}$$

- A *repetition-free* word is a word that contains no repetition.
- For example, if $u = ottr$, then

$$u^{7/4} = ottrott$$

Power-free words

Several types of power-free words

- A *square-free* word is a word that contains no factor that is a square.
- Let $r > 1$ be a real number. A word is r -free if it contains no factor of the form u^k for $k \geq r$, k rational.
- A word is k^+ -free if it is r -free for all $r > k$ (not necessarily for k).
- A word is k^- -free if it is k -free but not r -free for $r < k$.

Examples. Consider the morphisms

$$a \mapsto aba$$

$$b \mapsto abb$$

$$a \mapsto ab$$

$$b \mapsto ba$$

The word generated by the first morphism is 3^- -free, the word generated by the second (Thue-Morse) is 2^+ -free (= overlap-free).

Proof

The word generated by iterating the morphism $f : \begin{array}{l} a \mapsto aba \\ b \mapsto abb \end{array}$ is

$$z = abaabbabaabaabbabbabaabbabaabaabbabaabaabbabb \dots$$

The words aab , $f(aab) = aba\ aba\ abb$, and in fact all words $f^n(aab)$ are cubes except for their last letter.

Assume that $f(w)$ contains a cube uuu .

a) $|u|$ is a multiple of 3.

The initial letter of u appears at positions $i, i + |u|, i + 2|u|$. If $|u| \not\equiv 0 \pmod 3$, the initial letter of u appears in $f(a)$ or $f(b)$ at the first, the second and the third position. This is not the case.

b) f is injective, and the preimages are the last letters of the images.

c) One may assume $i \equiv 0 \pmod 3$. Thus w also contains a cube.

Powers and periodic words

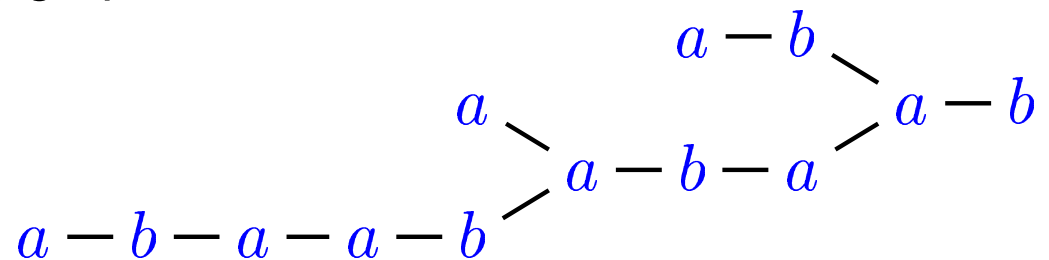
Let $\rho \geq 1$ be a real number. A word w is ρ -legal if it has a suffix which is a repetition of exponent at least ρ . For instance, $abaababa$ is $5/2$ -legal.

An infinite word x is ρ -legal if all its long enough prefixes are ρ -legal.

Example. The Fibonacci word $f = \varphi(f)$ is 2-legal, that is every long enough suffix of f ends with a square (f does not contain bb , aaa , $babab$).

$$f = abaababaabaababaabaababaabaab \dots$$

Indeed, consider the following graph.



Powers and periodic words (cont'd)

Theorem (Mignosi, Restivo & Salemi)

- Every τ^2 -legal infinite word is ultimately periodic;
- The Fibonacci word is $(\tau^2 - \varepsilon)$ -legal for every $\varepsilon > 0$.

$$(\tau = \frac{1 + \sqrt{5}}{2})$$

Every long enough prefix of the Fibonacci word ends with a repetition of exponent $\tau^2 - \varepsilon$.

Theorem (Mignosi, Pirillo) *The Fibonacci word is $(1 + \tau^2)^-$ -free, that is, it has repetitions of exponent $(1 + \tau^2) - \varepsilon$ for every $\varepsilon > 0$, but no repetition of exponent $1 + \tau^2$.*

Open problems

Problem *Prove or disprove that it is decidable whether a morphism is cube-free*

Special cases are known (Leconte, Keränen). Since no algorithm is known, perhaps it is undecidable ?

Open problems : repetition threshold

Every binary word of length 4 contains a square, and there exist infinite binary 2^+ -free words.

Every ternary word of length 39 contains a repetition of exponent $7/4$, and there exists (Dejean) an infinite ternary $(7/4)^+$ -free word.

The repetition-*threshold* is the smallest number $s(k)$ such that there exists an infinite word over k letters that has only repetitions of exponent less than or equal to k .

k	2	3	4	5	...	11
$s(k)$	2	$7/4$	$7/5$	$5/4$...	$11/10$

Problem *Is it true that the repetition threshold is always $k/(k-1)$?*

Open problems : avoidable pattern

- E is a *pattern* alphabet, A is a *target* alphabet.
- $\mathcal{M}(E, A)$ is the set of morphisms from E^+ to A^+ .
- For p over E , the *pattern language* of p over A is the set $H(p) = \{h(p) \mid h \in \mathcal{M}(E, A)\}$.
- A word w over A *avoids* p if no factor of w is in $H(p)$.

Examples

- A square-free word is a word that avoids the pattern $\alpha\alpha$.
- No word over n letters of length $n + 1$ avoids the pattern $\alpha\beta\alpha$.

A pattern p is *k -avoidable* if there exists an infinite word over k letters that avoids p .

Problem Is there a pattern that is 4- unavoidable and 5-avoidable ?