Repetitions in words

Recent results and open 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.
- Power-free words and power-free morphisms.
- Test sets and test words for power-free morphisms.
- 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

$M_x = 2, H_x = 4$

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

Subword complexity of finite words (cont'd)

2) de Bruijn words $=$ product of Lyndon words of length dividing n (Fredricksen Maiorana):

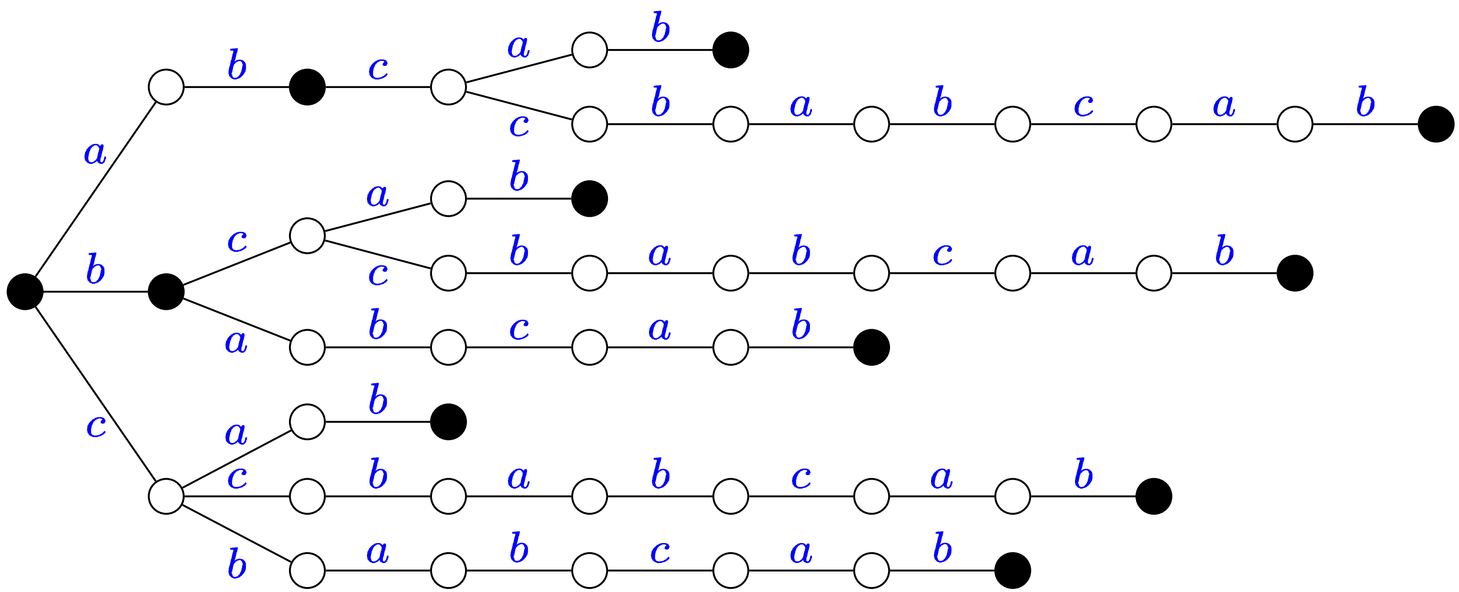
$$n = 3 \qquad \begin{array}{r} 0 \\ 001 \\ 011 \\ 1 \end{array} \qquad x = 00010111$$

n	0	1	2	3	4	5	6	7	8
$p_x(n)$	1	2	4	6	5	4	3	2	1

$$M_x = H_x = 3$$

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 : explicit description

Characteristic word of a set of integers.

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

$$1100100001000000100\ldots$$

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

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

$$f = abaababaabaababaababa\cdots$$

Explicit description

c) Thue-Morse word

$$t = 0110100110010110100101100110 1001 \cdots$$

defined by

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

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

$t_n = 1$ iff bin(n) is accepted by the automaton.

These are *automatic sequences*.

Infinite products

Any infinite product $x_0 x_1 \cdots x_n \cdots$ of nonempty words has a limit.

$$c = 0110111001011101111000 \cdots$$

The *Champernowne* word is the product of the words $\mathrm{bin}(n)$ (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 $h : A^* \to A^*$ 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

$$\begin{aligned} h^2(a) &= axh(x) \\ h^3(a) &= axh(x)h^2(a) \end{aligned}$$

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)

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

Then

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

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

$f : B^* \to B^*$ a morphism prolongable in the letter b.
$g : B^* \to A^*$ 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

$$s = 110010000100000010 0\cdots$$

is generated by

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

Indeed

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

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 $\underset{2}{\underline{2\,2}}\;\underset{2}{\underline{1\,1}}\;\underset{1}{\underline{2}}\;\underset{1}{\underline{1}}\;\underset{2}{\underline{2\,2}}\;\underset{1}{\underline{1}}\;\underset{2}{\underline{2\,2}}\;\underset{2}{\underline{1\,1}}\cdots$

$$\begin{array}{ccccc} &&&&2\\ &&&&\downarrow\\ 2&2&1&1&\\ &&\uparrow&& \end{array}$$

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 $\quad \underset{2}{\underline{2\,2}}\,\underset{2}{\underline{1\,1}}\,\underset{1}{\underline{2}}\,\underset{1}{\underline{1}}\,\underset{2}{\underline{2\,2}}\,\underset{1}{\underline{1}}\,\underset{2}{\underline{2\,2}}\,\underset{2}{\underline{1\,1}}\cdots$

$$\begin{array}{ccccc} & & & & 2\\ & & & & \downarrow\\ 2 & 2 & 1 & 1 & \\ & & \uparrow & & \end{array} \qquad \begin{array}{cccccc} & & & & & 1\\ & & & & & \downarrow\\ 2 & 2 & 1 & 1 & 2 & \\ & & & \uparrow & & \end{array}$$

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 $\quad \underset{2}{\underline{2\,2}}\,\underset{2}{\underline{1\,1}}\,\underset{1}{\underline{2}}\,\underset{1}{\underline{1}}\,\underset{2}{\underline{2\,2}}\,\underset{1}{\underline{1}}\,\underset{2}{\underline{2\,2}}\,\underset{2}{\underline{1\,1}}\cdots$

$$2\;\;2\;\;\underset{\uparrow}{1}\;\;1\;\;\overset{2}{\downarrow}\qquad 2\;\;2\;\;1\;\;\underset{\uparrow}{1}\;\;2\;\;\overset{1}{\downarrow}\qquad 2\;\;2\;\;1\;\;1\;\;\underset{\uparrow}{2}\;\;1\;\;\overset{2}{\downarrow}$$

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.

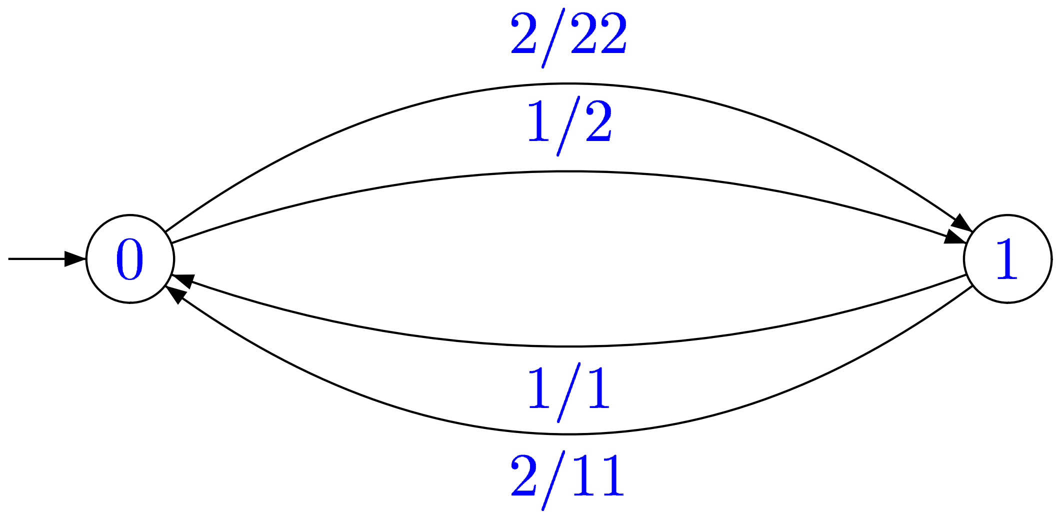
The question asked originally by Post (1921) : *is it decidable whether, for a word w on the tape, the reading head can reach the writing head ?*

Kolakoski sequence $\underset{2}{\underline{2\,2}}\,\underset{2}{\underline{1\,1}}\,\underset{1}{\underline{2}}\,\underset{1}{\underline{1}}\,\underset{2}{\underline{2\,2}}\,\underset{1}{\underline{1}}\,\underset{2}{\underline{2\,2}}\,\underset{2}{\underline{1\,1}}\cdots$

$$2\;\;2\;\;\underset{\uparrow}{1}\;\;1\;\;\overset{2}{\downarrow}\qquad 2\;\;2\;\;1\;\;\underset{\uparrow}{1}\;\;2\;\;\overset{1}{\downarrow}\qquad 2\;\;2\;\;1\;\;1\;\;\underset{\uparrow}{2}\;\;1\;\;\overset{2}{\downarrow}\qquad 2\;\;2\;\;1\;\;1\;\;2\;\;\underset{\uparrow}{1}\;\;2\;\;2\;\;\overset{1}{\downarrow}$$

Iterating sequential functions

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



2

22

2211

221121

221121221

$\ldots$

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

Toeplitz words

$x = x_0?x_1?x_2?\cdots$ with x_n words and $?$ a placeholder.
$y = a_0a_1a_2\cdots$ with a_n letters.
The *Toeplitz* product is

$$x\tau y = x_0a_0x_1a_1x_2a_2\cdots$$

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

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

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 $x = ab?ab?ab? \cdots = (ab?)^\omega$ is generated by

$$\begin{aligned} a &\mapsto aba \\ b &\mapsto abb \end{aligned}$$

Palindromic closure

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

$$(01001)^\pi = 010010 \qquad (01001010)^\pi = 01001010010$$

Given a word $d = a_0 a_1 \cdots a_n \cdots$, 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 morphims 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.

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

- 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

$$\begin{aligned} a &\mapsto aba \\ b &\mapsto abb \end{aligned} \qquad\qquad \begin{aligned} a &\mapsto ab \\ b &\mapsto ba \end{aligned}$$

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

$$z = abaabbabaabaabbabbabaabbabaabaabbabaabaabbabb\cdots$$

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 $|u| \not\equiv 0 \bmod 3$, 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 $i \equiv 0 \bmod 3$. Thus w also contains a cube.

Powers and periodic words

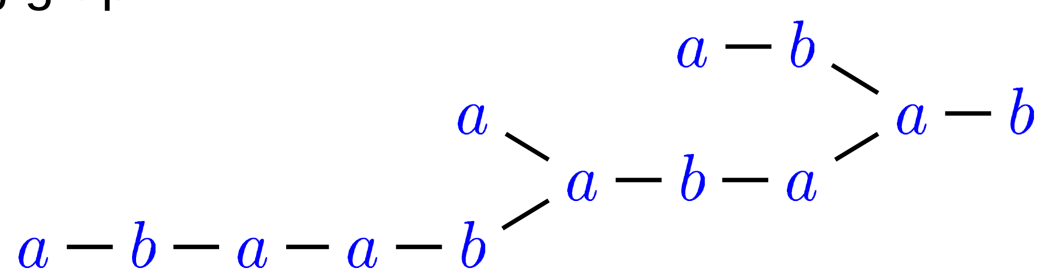
Let $\rho \geq 1$ be a real number. A word w is ρ-*legal* if it has a suffix which is a repetition of exponent at least ρ. For instance, $abaababa$ is $5/2$-legal.

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, $babab$).

$$f = abaab\,aba\,abaab\,abaababa\,abaababaabaab\cdots$$

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 $\varepsilon > 0$.*

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

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 $1+\tau^2$.*

Morphisms

A simple way to generate infinite power-free words : use a morphism.

Even better : use a power-free morphism: A morphism $f : A^* \to B^*$ is k-*free* if, for each k-free word w over A^*, the word $f(w)$ is k-free over B^*.

1)

$$\begin{aligned} a &\mapsto aba \\ b &\mapsto abb \end{aligned}$$

is a cube-free morphism.

2)

$$\begin{aligned} a &\mapsto abc \\ b &\mapsto ac \\ c &\mapsto b \end{aligned}$$

generates an infinite square-free word (used by M. Hall). It is not a square-free morphism ($aba \mapsto abcacabc$).

Morphisms

3)

$$\begin{aligned} a &\mapsto ab \\ b &\mapsto ba \end{aligned}$$

(Thue-Morse) is an overlap-free morphism.

4)

$$\begin{aligned} a &\mapsto abcab \\ b &\mapsto acabcb \\ c &\mapsto acbcacb \end{aligned}$$

is a square-free morphism (proved by Thue 1912). Its *length* is 18. There is no square-free endomorphism over 3 letters with smaller length (Carpi).

Morphisms

5)

$$\begin{aligned} a &\mapsto aababb \\ b &\mapsto aabbab \\ c &\mapsto abbaab \end{aligned}$$

is a cube-free morphism from 3 letters to 2 letters (Bean, Ehrenfeucht, McNulty).

6)

$$\begin{aligned} a &\mapsto abacabcacbabcbacbc \\ b &\mapsto abacabcacbacabacbc \\ c &\mapsto abacabcacbcabcbabc \\ d &\mapsto abacabcbacabacbabc \\ d &\mapsto abacabcbacbcacbabc \end{aligned}$$

is a square-free morphism from 5 letters to 3 letters (Brandenburg).

Proving a morphism to be k-free

The set of k-free endomorphisms is a monoid. If you are lucky, there is a characterization.

Theorem (Thue 1912) *An endomorphism f over a 2 letter alphabet is overlap-free if and only if it the product of the Thue-Morse morphism and the morphism that exchanges the two letters:*

$$f = E^{\epsilon}\mu^{k}$$

for some $\epsilon = 0, 1$ and $k \geq 0$.

Here $E(a) = b$, $E(b) = a$, and

$$\mu : \begin{array}{c} a \mapsto ab \\ b \mapsto ba \end{array}$$

Proving a morphism to be k-free

Usually, the monoid is not finitely generated.

Let $h : A^* \to B^*$ be a nonerasing morphism, and set

$$M(h) = \max_{a \in A} |h(a)|,\; m(h) = \min_{a \in A} |h(a)|\,.$$

Theorem (Crochemore) *The morphism h is square-free if and only if h preserves square-free words of length* $K(h) = \max(3, 1 + \lceil (M(h) - 3)/m(h) \rceil)$.

For Thue's morphism, check the images of square-free words of length 3.

$$\begin{aligned} a &\mapsto abcab \\ b &\mapsto acabcb \\ c &\mapsto acbcacb \end{aligned}$$

Test sets : cube-free morphisms

A set $T \subset A^*$ is a *test set* for k-free morphisms $A^* \to B^*$ if, for any $f : A^* \to B^*$,

$$f \text{ is } k\text{-free if and only if } f \text{ is } k\text{-free on } T.$$

A test set that is a singleton is a *test word*.

Theorem (Karhumäki, Leconte) *Let f be a binary morphism. Then f is cube-free if and only if f is cube-free on cube-free words of length at most 7.*

The set of cube-free binary words of length at most 7 is a test set for cube-free binary morphisms.

Test sets : cube-free morphisms

The test sets for cube-free binary morphisms are the following.

Theorem (Richomme & Wlazinski) *A set T of cube-free words is a test set for cube-free binary morphisms if and only if*

$$F(T) \supset X \cup \tilde{X} \cup E(X \cup \tilde{X})$$

where $F(T)$ is the set of factors of T and
$X = \{abbabba, ababba, aabba, ababa\}$.

Theorem (Richomme & Wlazinski) *Let f be a binary morphism. Then f is cube-free if and only if*

$$f(abaabbabaabaabbaababaabbabaabaabbaababaab)$$

is a cube-free word.

Thus, $abaabbabaabaabbaababaabbabaabaabbaababaab$ is a test word.

Test sets : square-free morphisms

Theorem (Crochemore) *Let* f *be a ternary morphism. Then* f *is square-free if and only if* f *is square-free on square-free words of length* 5..

Bounds are known for special families of morphisms (uniform, infix).

Test sets : ovelap-free morphisms

Theorem (Berstel & Séébold, Richomme & Séébold) *Let* f *be a binary endomorphism. Then the following are equivalent*

- f *is overlap-free,*
- *the four words* $f(aab)$, $f(aba)$, $f(bab)$,$f(abb)$ *are overlap-free,*
- *the word* $f(bbabaa)$ *is overlap-free.*

Richomme & Séébold have characterized *all* test sets for overlap-free binary morphisms.

Theorem (Richomme & Wlazinski) *A set* T *of overlap-free words is a test set for overlap morphisms from* $A^* \to B^*$ *with* $|A| = 2$, $|B| \geq 3$ *if and only if* $F(T)$ *contains the four words* $aba, bab, abba, baab$*. The word* $abbabaab$ *is a test word.*

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 and infinite word over k letters that has only repetitions of exponent less than or equal to k.

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

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 ?*