Sturmian Words, Sturmian Trees and Sturmian Graphs

A Survey of Some Recent Results

Jean Berstel

Institut Gaspard-Monge, Université Paris-Est

CAI 2007, Thessaloniki

Outline

1. Sturmian words
   - Factors
   - Central words
2. Burrows-Wheeler transform
   - Burrows-Wheeler transform and Sturmian words
   - Gessel-Reutenauer transformation
3. Sturmian trees
   - Definition and examples
   - Slow automata
   - Rank and degree
   - Results
4. Sturmian graphs

Sturmian words

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = \boxed{a}\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ \boxed{b}\ a\ a\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = \boxed{a\ b}\ a\ a\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ \boxed{b\ a}\ a\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a b	aa ab ba	aab aba baa bab	$aaba$ $abaa$ $abab$ $baab$ $baba$	$aabaa$ $aabab$ $abaab$ $ababa$ $baaba$ $babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ b\ \boxed{a\ a}\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$x = a\ b\ a\ \boxed{a\ b}\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = \boxed{a\ b\ a}\ a\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ \boxed{b\ a\ a}\ b\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ b\ \boxed{a\ a\ b}\ a\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$x = a\ b\ a\ \boxed{a\ b\ a}\ b\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Factors of an infinite word

A factor of a word x is a finite word that occurs in x.

Example (Fibonacci word)

$$x = a\ b\ a\ a\ \boxed{b\ a\ b}\ a\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ a\ b\ a\ a\ b\ \ldots$$

Length: n	1	2	3	4	5
Factors	a	aa	aab	$aaba$	$aabaa$
	b	ab	aba	$abaa$	$aabab$
		ba	baa	$abab$	$abaab$
			bab	$baab$	$ababa$
				$baba$	$baaba$
					$babaa$
#	2	3	4	5	6

Sturmian words

Proposition (Hedlund & Morse)

An infinite word x is ultimately periodic iff there is an integer n such that x has at most n distinct factors of length n.

Definition

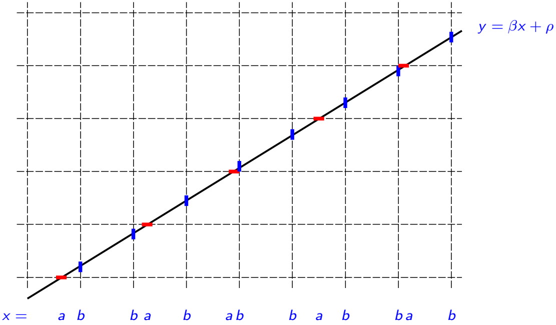
An infinite word x is Sturmian if the number of its factors of length n is $n+1$ for each n.

Sturmian words are non ultimately periodic words with the smallest complexity.

Example (Fibonacci word: $f_{n+2} = f_{n+1}f_n$)

$$
\begin{aligned}
f_0 &= a \\
f_1 &= ab \\
f_2 &= aba \\
f_3 &= abaab \\
f_4 &= abaababa \\
f_5 &= abaababaabaab \\
f_\omega &= abaababa\cdots
\end{aligned}
$$

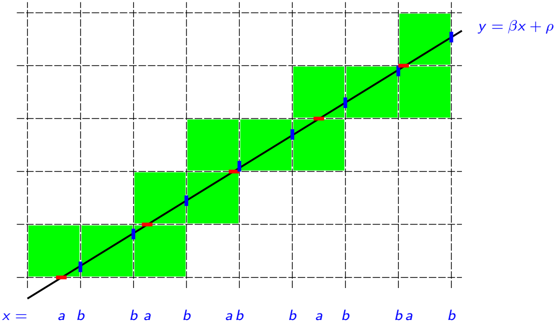
A characterization: cutting sequences



Theorem

A infinite word is Sturmian iff it is the cutting sequence of a straight line $y = \beta x + \rho$ with an irrational slope β.

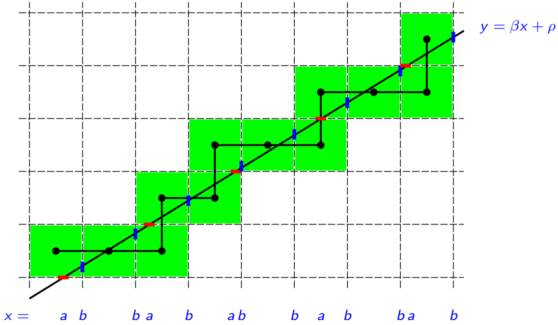
A characterization: cutting sequences



Theorem

A infinite word is Sturmian iff it is the cutting sequence of a straight line $y = \beta x + \rho$ with an irrational slope β.

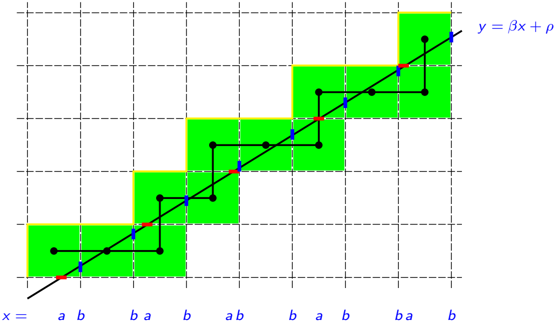
A characterization: cutting sequences



Theorem

A infinite word is Sturmian iff it is the cutting sequence of a straight line $y = \beta x + \rho$ with an irrational slope β.

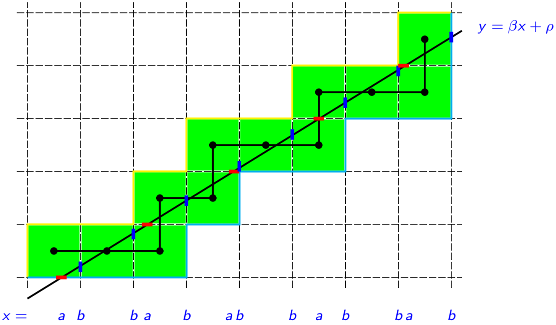
A characterization: cutting sequences



Theorem

A infinite word is Sturmian iff it is the cutting sequence of a straight line $y = \beta x + \rho$ with an irrational slope β.

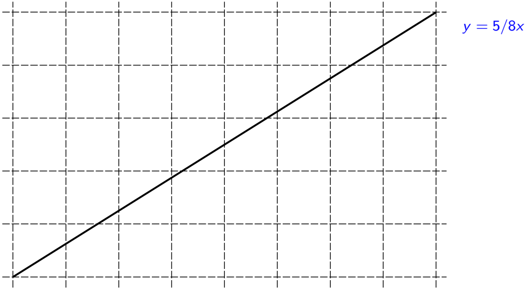
A characterization: cutting sequences



Theorem

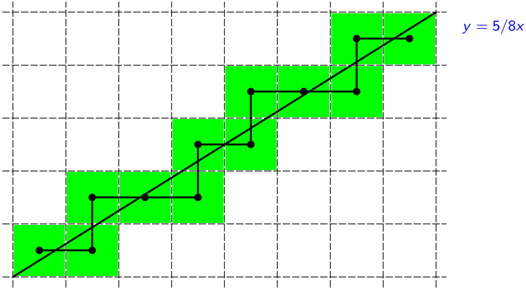
A infinite word is Sturmian iff it is the cutting sequence of a straight line $y = \beta x + \rho$ with an irrational slope β.

Central words, Christoffel words, standard words



$x = 01001010010$ is a central word.

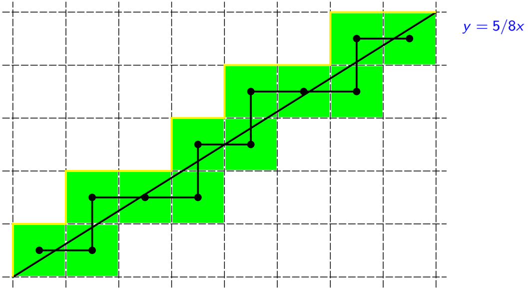
Central words, Christoffel words, standard words



$x = 01001010010$ is a central word.

- A central word x is a palindrome

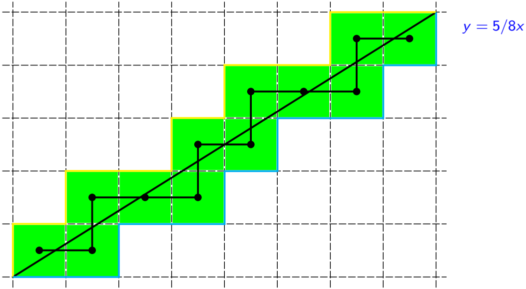
Central words, Christoffel words, standard words



$x = 01001010010$ is a central word.

- A central word x is a palindrome
- The upper Christoffel word is $1x0$

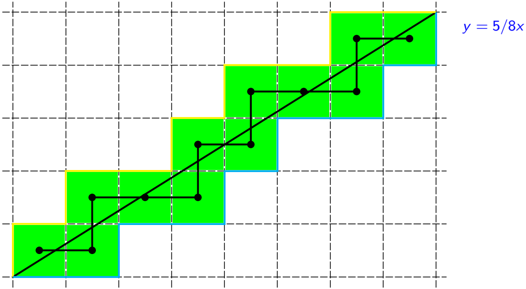
Central words, Christoffel words, standard words



$x = 01001010010$ is a central word.

- A central word x is a palindrome
- The upper Christoffel word is $1x0$
- The lower Christoffel word is $0x1$

Central words, Christoffel words, standard words



$x = 01001010010$ is a central word.

- A central word x is a palindrome
- The upper Christoffel word is $1x0$
- The lower Christoffel word is $0x1$
- The words is $x10$ and $x01$ are standard words

Construction of all standard words

Definition

A finite or infinite sequence $d = (d_0, d_1, \ldots)$ of integers with $d_0 \geq 0$, $d_n > 0$ for $n \geq 1$ is a directive sequence. Define

$$s_{-1} = b, \quad s_0 = a, \quad s_{n+1} = s_n^{d_n} s_{n-1} \quad (n \geq 0)$$

Each word s_n is a standard word produced by the directive sequence $(d_0, \ldots, d_{n-1})$. The infinite word $s = \lim s_n$ is the characteristic word produced by the directive sequence d.

Example

$d = (1, 1, 1, 1, 1)$ produces $s_5 = abaababaabaab$. Indeed

$$\begin{aligned}
s_0 &= a \\
s_1 &= ab \\
s_2 &= aba \\
s_3 &= abaab \\
s_4 &= abaababa \\
s_5 &= abaababaabaab
\end{aligned}$$

Recall that

$$s_{-1} = b, \quad s_0 = a, \quad s_{n+1} = s_n^{d_n} s_{n-1} \quad (n \geq 0)$$

Other examples

- For $d = (1, 1, 1, 2)$, one gets $s_{-1} = b$, $s_= a$, $s_1 = ab$, $s_2 = aba$, $s_3 = abaab$ and $s_4 = (abaab)(abaab)aba$.
- The sequence $d = (0, 1, 1, 1, 2)$ produces $s_{-1} = b, s_0 = a$, $s_1 = b$, $s_2 = ba$, $s_3 = bab$, $s_4 = babba$, $s_5 = (babba)(babba)bab$.
- $d = (1, 1, 1, 1, 1, 1, \ldots)$ produces the infinite Fibonacci word.

Theorem

The set of standard words is the set of all words s_n produced by all directive sequences. The set of characteristic words is the set of all limits of directive sequences.

Balanced words

Definition

- A set X of finite words is balanced if, for every letter a and every u, v in X of the same length,

$$\Big| |u|_a - |v|_a \Big| \leq 1$$

- A finite or infinite word is balanced if the set of its factors is balanced.

Example

- The Thue-Morse word $t = abbabaabbaababba \cdots$ is not balanced since aa and bb are factors.
- The Fibonacci word is balanced. The word $baab$ s a factor of the Fibonacci word but its square $baabbaab$ is not balanced.

Theorem

- *An infinite word is Sturmian if and only if it is balanced and aperiodic.*
- *An finite word is balanced if and only if it a factor of a Sturmian word.*

Strongly balanced words

O. Jenkinson, L. Zamboni, "Characterisations of balanced words via orderings", *Theoret. Comput. Sci.* **310** (2004), 247-271.

A word w is strongly balanced if it is primitive and if w^2 is balanced.

Example

- The word 01110 is balanced but not strongly balanced since $w^2 = 0111001110$.
- The word 1010010 is strongly balanced.

Proposition

The following properties are equivalent

- w *is strongly balanced;*
- w *is a conjugate of a standard word;*
- *every conjugate of* w *is balanced;*
- w *is conjugate to an upper Christoffel word;*
- w *is conjugate to a lower Christoffel word.*

Example

$w = 1010010$ is conjugate to the standard word 0101001 produced by $d = (1, 2, 1)$, and to the upper Christoffel word 1010100 and to the lower Christoffel word 0010101.

Burrows-Wheeler transform

Burrows-Wheeler Transform

- The Burrows-Wheeler Transform (BWT) is a reversible transformation that produces a permutation BWT(w) of an input sequence w.
- The transform is easier to compress.
- BWT is used in the BZIP2 algorithm.
- BWT has a strong relation to the Gessel-Reutenauer transform.

References

- S. Mantaci, A. Restivo, G. Rosone, M. Sciortino, "An extension of the Burrows Wheeler transform", to appear in the special issue of TCS devoted to the Burrows Wheeler Transform.
- S. Mantaci, A. Restivo, M. Sciortino, "Burrows Wheeler transform and Sturmian words", *Inform. Proc. Letters* **86** (2003), 241-246.
- M. Crochemore, J. Désarménien, D. Perrin, "A note on the Burrows-Wheeler transformation", *Theoret. Comput. Sci.* **332** (2005), 567-572.

How does Burrows-Wheeler Transform work?

INPUT: $w = abraca$
OUTPUT: $\mathrm{BWT}(w) = x = caraab$ and 1

$$\begin{bmatrix} a & a & b & r & a & c \\ a & b & r & a & c & a \\ a & c & a & a & b & r \\ b & r & a & c & a & a \\ c & a & a & b & r & a \\ r & a & c & a & a & b \end{bmatrix}$$

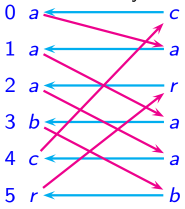
1. The conjugates of w are ordered lexicographically.
2. The output is the last column of the table: $caraab$, and the position of the input:1 (numbering starts at 0).

Two words u and v are conjugate if and only if $\mathrm{BWT}(u) = \mathrm{BWT}(v)$.

Reversibility of the Burrows-Wheeler Transform

The Burrows-Wheeler Transform is reversible: given $\mathrm{BWT}(w)$ and an index i, it is possible to recover w.
Given $\mathrm{BWT}(w) = x = caraab$ and $i = 1$, do the following

1. The first column f of the table is obtained by sorting the letters in x (the last column).



2. Define a permutation τ on the set $\{0, \ldots, n-1\}$ that maps a position in f to the corresponding position in x.

$$\tau = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 5 & 4 & 0 \end{pmatrix}$$

Thus

$$\begin{array}{ccccccc} & 1 & 3 & 5 & 2 & 4 & 0 \\ w = & a & b & r & a & c & a \end{array}$$

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0, 1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form 1^q0^p. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0,1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form 1^q0^p. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0,1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form 1^q0^p. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0,1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form 1^q0^p. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0,1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form 1^q0^p. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0, 1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form $1^q 0^p$. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Properties of the Burrows-Wheeler Transform

1. BWT is injective on conjugacy classes.
2. BWT is not surjective: there are words (e.g. *bccaaab*) that are not images of a word.
3. Why useful: it produces a clustering effect. In each row but one, the symbol in the last colum is the symbol preceeding the conjugate. So conjugates that are gouped together have also the same final symbol.

Theorem

A word w over $\{0,1\}$ is strongly balanced if and only if its Burrows-Wheeler Transform is of the form 1^q0^p. Moreover, in the table, each row is obtained from the preceeding by replacing a factor 01 by a factor 10, and all columns also are conjugates.

0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0

Gessel-Reutenauer bijection

I. Gessel, C. Reutenauer, “Counting permutations with given cycle structure and descent set”, *J. Comb. Theory A*, **64**, 1993, 189-215.

Definition

The standardization associates to $w = a_1 \cdots a_n$ over an ordered alphabet A a permutation σ defined by

$$\sigma(i) < \sigma(j) \quad \text{iff} \quad a_i < a_j \text{ or } (a_i = a_j \text{ and } i < j)$$

Example

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ c & c & b & b & b & c & a & c & a & a & a & b & b & a \\ 11 & 12 & 6 & 7 & 8 & 13 & 1 & 14 & 2 & 3 & 4 & 9 & 10 & 5 \end{pmatrix}$$

$$\begin{matrix} (1 & 11 & 4 & 7) & (2 & 12 & 9) & (3 & 6 & 13 & 10) & (5 & 8 & 14) \\ c & a & b & a & c & b & a & b & c & b & a & b & c & a \end{matrix}$$

Gessel-Reutenauer transform

Theorem (Gessel-Reutenauer)

The standardization σ induces a bijection between all words over A and the family of multisets of conjugacy classes of primitive words over A.

Definition

Define a new order on finite order on words by

$$u \preceq v \quad \text{iff} \quad u^\omega < v^\omega \text{ or } (u^\omega = v^\omega \text{ and } |u| < |v|)$$

Example

$aba \prec ab$ because $abaaba\cdots < ababab\cdots$.

Gessel-Reutenauer coding

INPUT: $S = \{caba, bcba, bca, cba\}$
OUTPUT: $ccbbbcacaaabba$ and $9, 10, 11, 14$

1. Sort the conjugates of words in S by $\prec$
2. The output word is the sequence of last letters
3. The output indices are the positions of the input words

1		a	b	a	c
2		a	b	c	
3		a	b	c	b
4		a	c	a	b
5		a	c	b	
6		b	a	b	c
7		b	a	c	a
8		b	a	c	
9	$\longrightarrow$	b	c	a	
10	$\longrightarrow$	b	c	b	a
11	$\longrightarrow$	c	a	b	a
12		c	a	b	
13		c	b	a	b
14	$\longrightarrow$	c	b	a	

Gessel-Reutenauer decoding

INPUT: $ccbbbcacaaabba$ and $9, 10, 11, 14$
OUPUT: $S = \{caba, bcba, bca, cba\}$

1. Sort the input alphabetically (with an order-preserving sorting algorithm)
2. Compute the letter-correspndence permutation
3. Output the permutation in cycle form, and compute the multiset

$$\begin{pmatrix} a & a & a & a & a & b & b & b & b & b & c & c & c & c \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 7 & 9 & 10 & 11 & 14 & 3 & 4 & 5 & 12 & 13 & 1 & 2 & 6 & 8 \\ c & c & b & b & b & c & a & c & a & a & a & b & b & a \end{pmatrix}$$

Cycle form

$$\begin{matrix} (& 1 & 7 & 4 & 11 &) & (& 2 & 9 & 12 &) & (& 3 & 10 & 13 & 6 &) & (& 5 & 14 & 8 &) \\ & a & b & a & c & & & a & b & c & & & a & b & c & b & & & a & c & b & \end{matrix}$$

Output

$caba \quad bca \quad bcba \quad cba$

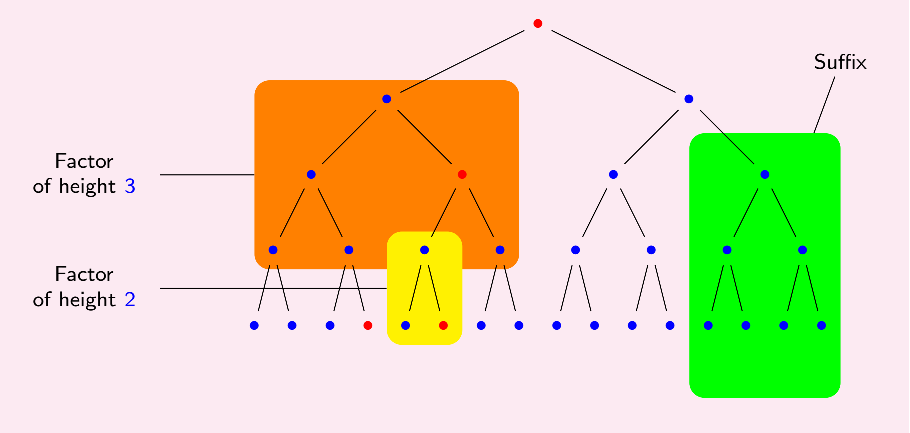
Sturmian trees

Factors in a tree

J. Berstel, L. Boasson, O. Carton, I. Fagnot, "A First Investigation of Sturmian Trees", *STACS'2007*, LNCS **4393**, 73-84.

Definition

A factor of height h of a tree t is a subtree of height h that occurs in t.
A suffix of tree t is an infinite subtree of t.



Rational tree

Definition

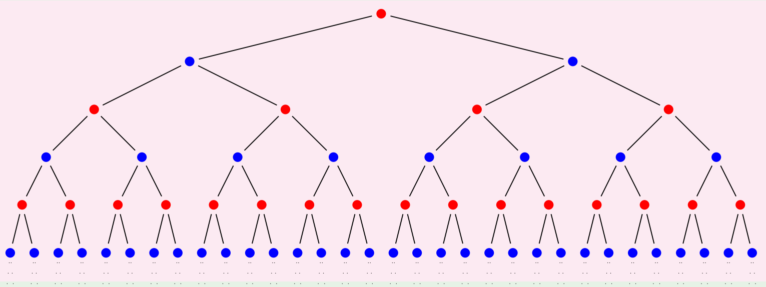
A tree is rational if it has a finite number of distinct suffixes.

Proposition (Carpi, De Luca, Varricchio)

A complete tree t is rational if there is some integer h such that t has at most h distinct factors of height h.

Example (Only two distinct suffixes)

Red nodes at even levels, blue nodes at odd levels.



Sturmian tree

Definition

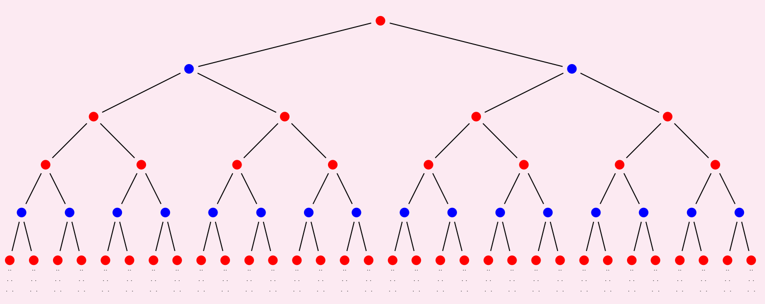
A tree is Sturmian if it has $h+1$ distinct factors of height h for each h.

Remark

Sturmian trees are irrational trees with the smallest complexity.

Example (Easy one: uniform tree)

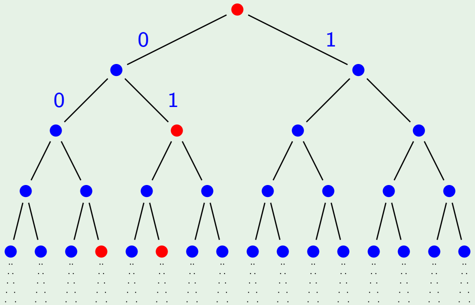
An Sturmian word $x = abaaba\ldots$ is repeated on each branch.



Example (Unexpected one: Dyck tree)

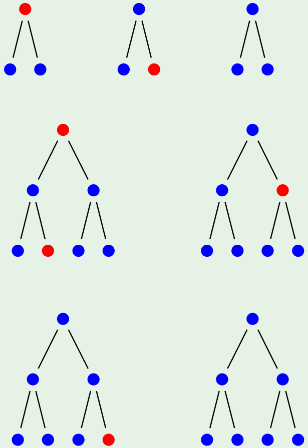
A node is • if it is a Dyck word over the alphabet $\{0,1\}$.

The Dyck tree



$D_2^* = \{\varepsilon, 01, 0101, 0011, \ldots\}$

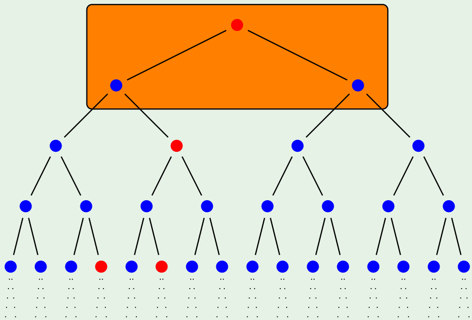
Its factors



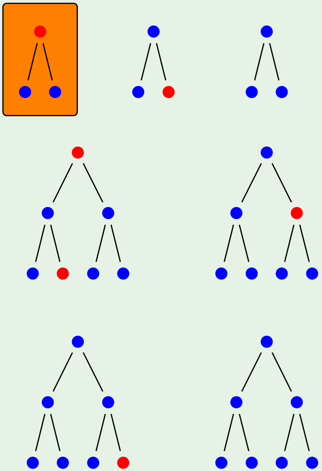
Example (Unexpected one: Dyck tree)

A node is ● if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



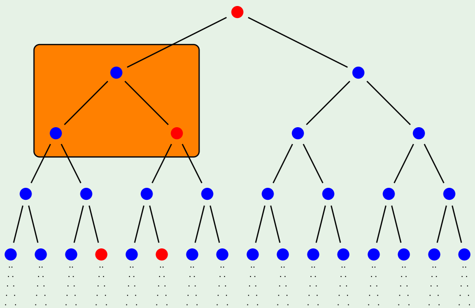
Its factors



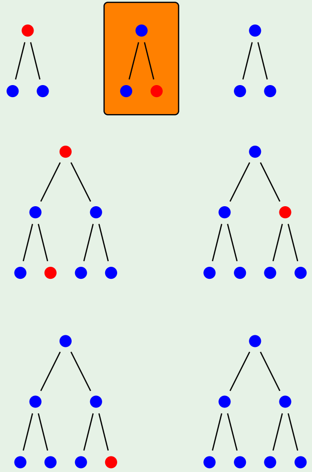
Example (Unexpected one: Dyck tree)

A node is • if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



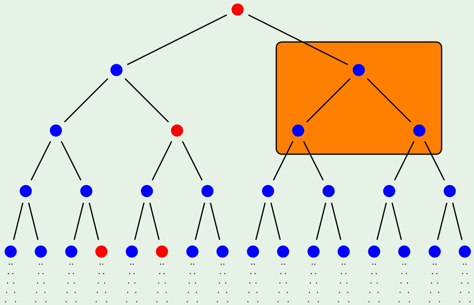
Its factors



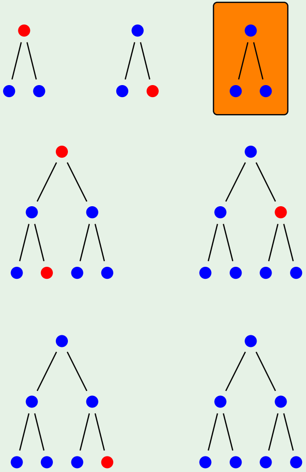
Example (Unexpected one: Dyck tree)

A node is • if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



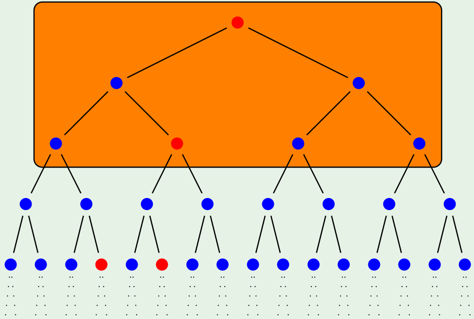
Its factors



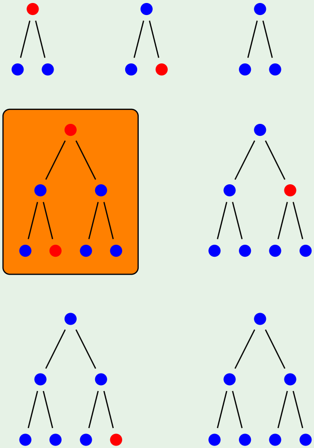
Example (Unexpected one: Dyck tree)

A node is ● if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



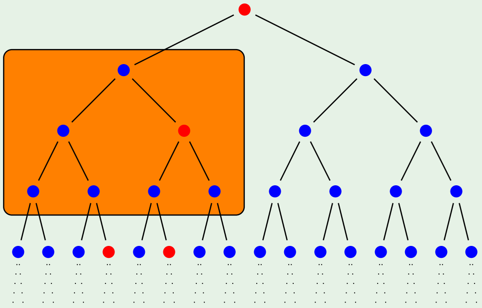
Its factors



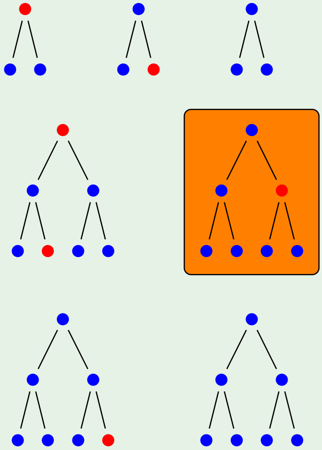
Example (Unexpected one: Dyck tree)

A node is $\bullet$ if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



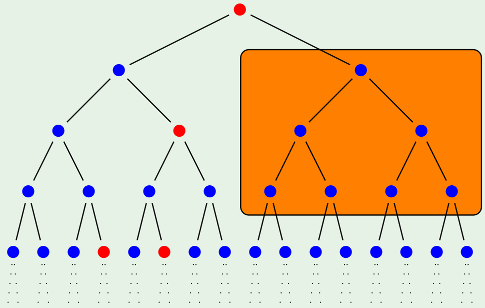
Its factors



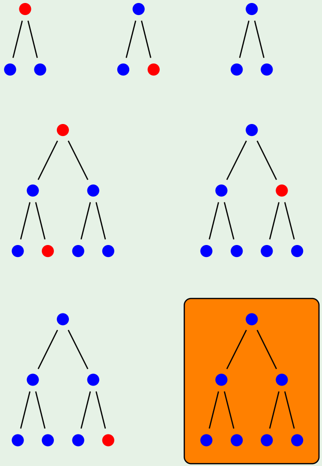
Example (Unexpected one: Dyck tree)

A node is $\bullet$ if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



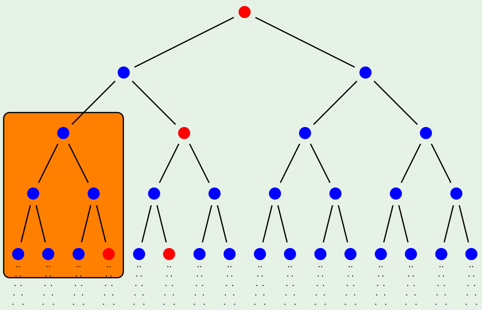
Its factors



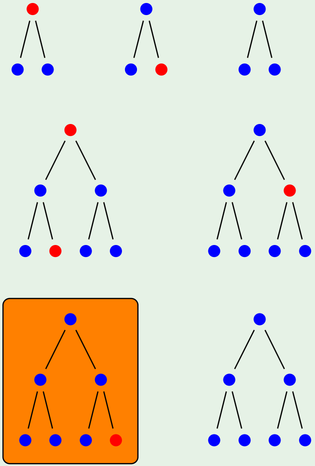
Example (Unexpected one: Dyck tree)

A node is • if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



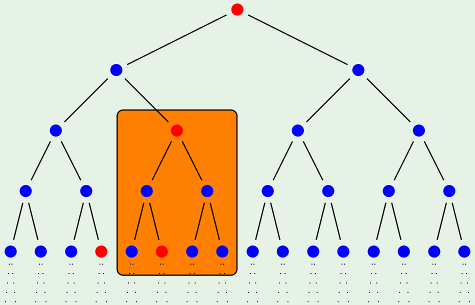
Its factors



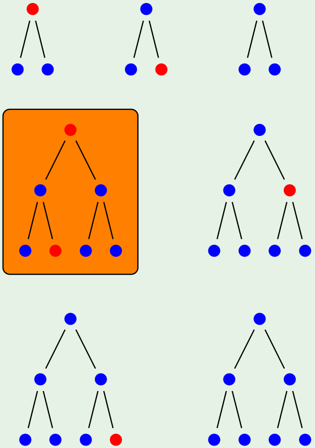
Example (Unexpected one: Dyck tree)

A node is • if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



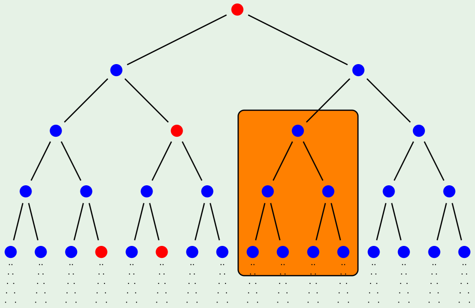
Its factors



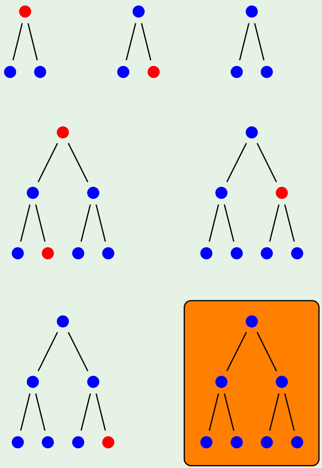
Example (Unexpected one: Dyck tree)

A node is $\bullet$ if it is a Dyck word over the alphabet $\{0, 1\}$.

The Dyck tree



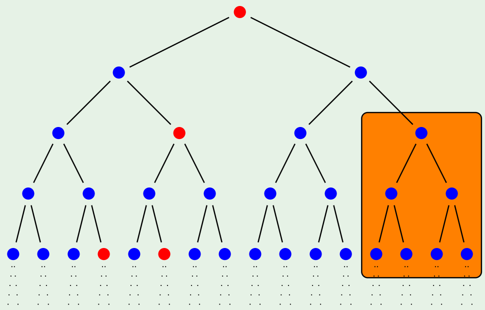
Its factors



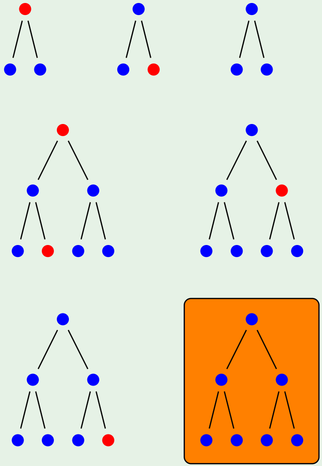
Example (Unexpected one: Dyck tree)

A node is ● if it is a Dyck word over the alphabet $\{0,1\}$.

The Dyck tree



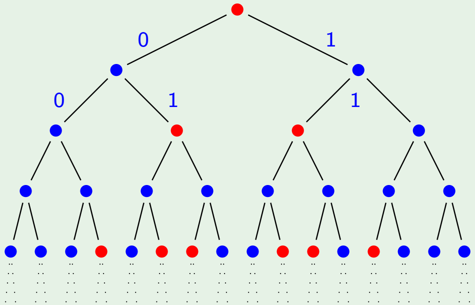
Its factors



Example (The two-sided Dyck tree is not Sturmian)

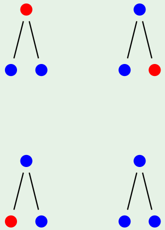
A node is • if it is a tow-sided Dyck word over the alphabet $\{0, 1\}$.

The two-sided Dyck tree



$D_2^* = \{\varepsilon, 01, 10, 0011, 0101, 0110, 1001, 1010, 1100, \ldots\}$

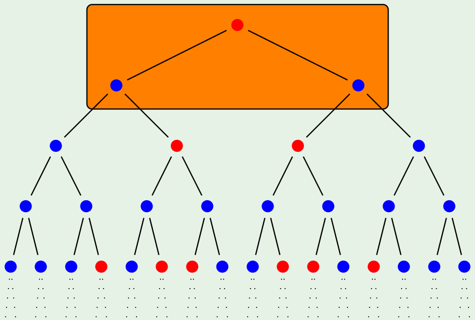
Its four factors of height 2



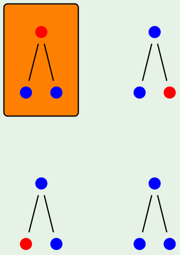
Example (The two-sided Dyck tree is not Sturmian)

A node is • if it is a tow-sided Dyck word over the alphabet $\{0, 1\}$.

The two-sided Dyck tree



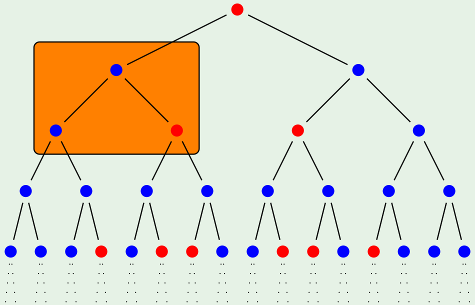
Its four factors of height 2



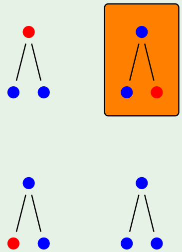
Example (The two-sided Dyck tree is not Sturmian)

A node is • if it is a tow-sided Dyck word over the alphabet $\{0, 1\}$.

The two-sided Dyck tree



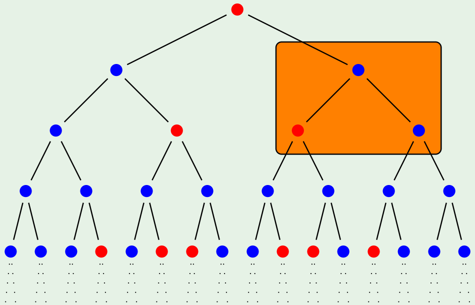
Its four factors of height 2



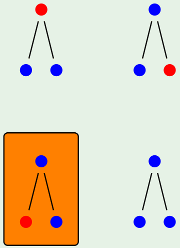
Example (The two-sided Dyck tree is not Sturmian)

A node is • if it is a tow-sided Dyck word over the alphabet $\{0, 1\}$.

The two-sided Dyck tree



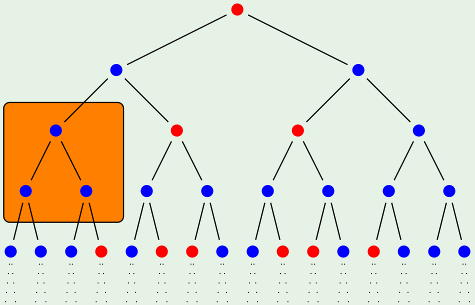
Its four factors of height 2



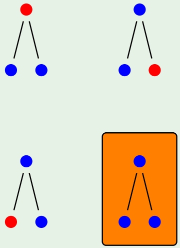
Example (The two-sided Dyck tree is not Sturmian)

A node is ● if it is a tow-sided Dyck word over the alphabet $\{0,1\}$.

The two-sided Dyck tree



Its four factors of height 2



Slow automata

Let $\mathcal{A}$ a (infinite) minimal deterministic automaton over $D = \{0, 1\}$ with states Q and final states F automaton.

Definition

The Moore equivalence $\sim_h$ of order h is

$$q \sim_1 q' \iff (q \in F \Leftrightarrow q' \in F)$$
$$q \sim_{h+1} q' \iff (q \sim_h q') \text{ and } (\forall a \in D \;\; q \cdot a \sim_h q' \cdot a)$$

Definition

An infinite automaton is slow iff the Moore equivalence $\sim_h$ of order h has $h + 1$ classes for each h.

Remark

In a slow automaton, exactly one equivalence class of $\sim_h$ is split into two classes of $\sim_{h+1}$.

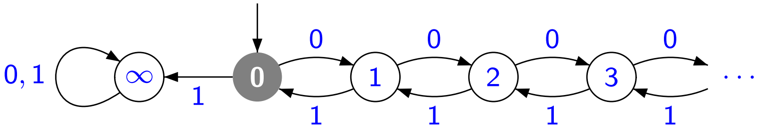
Proposition

A tree t is Sturmian iff the minimal automaton of its language is slow.

A first slow automaton

Automaton of the Dyck language.

State 0 is both the initial and the unique terminal state.



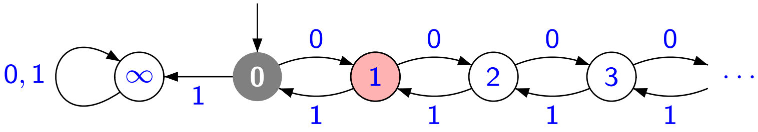
Moore equivalences:

$0 \mid 12345 \cdots \infty$

A first slow automaton

Automaton of the Dyck language.

State 0 is both the initial and the unique terminal state.



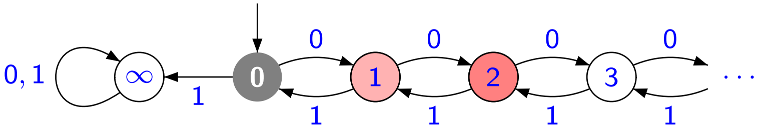
Moore equivalences:

$0 \mid 12345 \cdots \infty$
$0 \mid 1 \mid 2345 \cdots \infty$

A first slow automaton

Automaton of the Dyck language.

State 0 is both the initial and the unique terminal state.



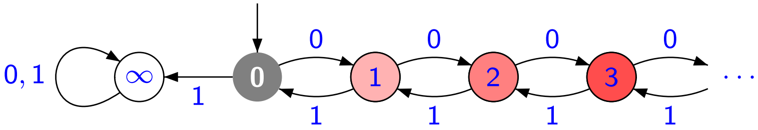
Moore equivalences:

$0 \mid 12345 \cdots \infty$
$0 \mid 1 \mid 2345 \cdots \infty$
$0 \mid 1 \mid 2 \mid 345 \cdots \infty$

A first slow automaton

Automaton of the Dyck language.

State 0 is both the initial and the unique terminal state.

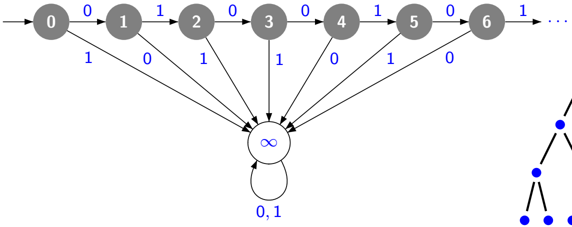


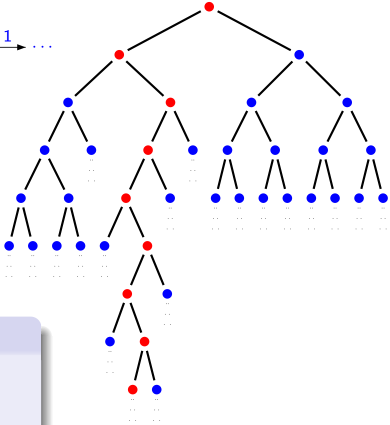
Moore equivalences:

$0 \mid 12345 \cdots \infty$
$0 \mid 1 \mid 2345 \cdots \infty$
$0 \mid 1 \mid 2 \mid 345 \cdots \infty$
$0 \mid 1 \mid 2 \mid 3 \mid 45 \cdots \infty$

Another slow automaton

Automaton accepting the prefixes of $01001010\cdots$.



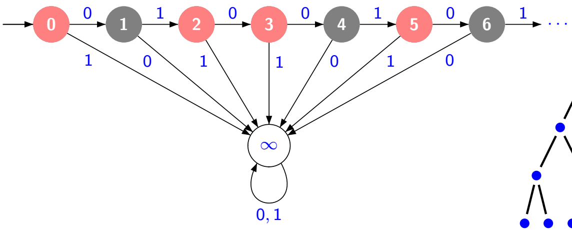


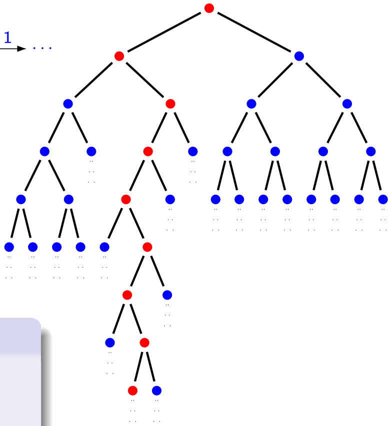
Equivalence classes:

$\infty \mid 012\cdots$

Another slow automaton

Automaton accepting the prefixes of $01001010\cdots$.



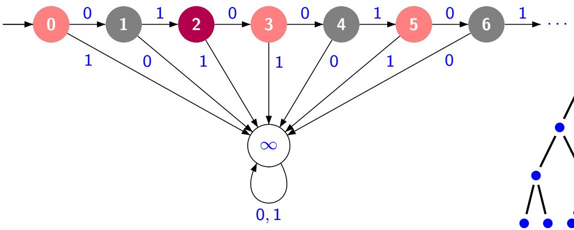


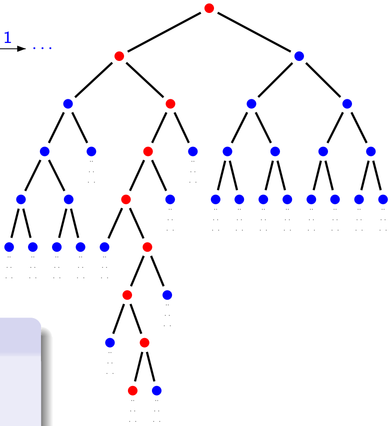
Equivalence classes:

$\infty \mid 012\cdots$
$\infty \mid 0235\cdots \mid 146\cdots$

Another slow automaton

Automaton accepting the prefixes of 01001010 $\cdots$.





Equivalence classes:

∞ | 012 $\cdots$

∞ | 0235 $\cdots$ | 146 $\cdots$

∞ | 035 $\cdots$ | 2 $\cdots$ | 146 $\cdots$

Rank and degree

Definition

- A node is called irrational if the infinite subtree rooted in this node is not rational.
- The rank is the number of distinct rational subtrees.
- The degree is the number of branches of irrational nodes.

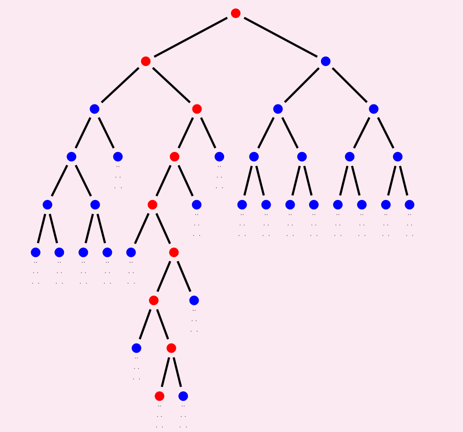
Examples

- The uniform tree has rank 0 and degree ∞.
- The Dyck tree has rank 1 and degree ∞.
- The indicator tree of a Sturmian word has rank 1 and degree 1.

Rank and degree of the indicator tree

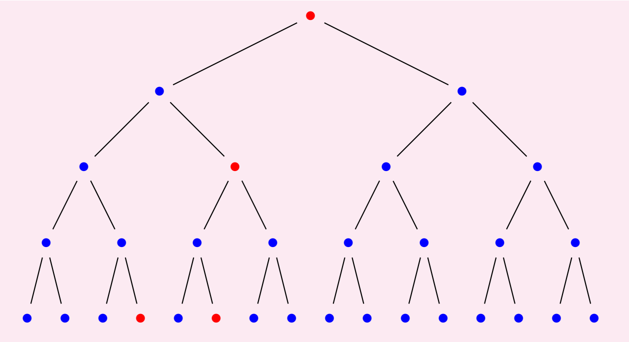
Take any Sturmian word (e.g. $01001010\cdots$) and distinguish the branch labeled by this word.

The only rational tree is the tree rooted in the blue node. The only irrational path is composed of the red nodes.



Rank and degree of the Dyck tree

The only rational subtree is composed of blue nodes only, so the degree is 1.



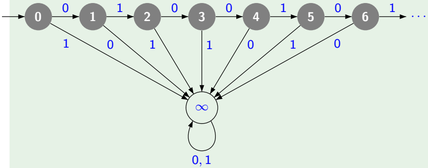
The degree is infinite because every (prefix of a) Dyck word extends to an infinite irrational path by concatenating some infinite product of distinct Dyck words.

Results

degree	rank	
	finite	infinite
1	characterized	example later
≥ 2, finite	proved to be empty	example in full paper
infinite	example of Dyck tree	example in full paper

Characterization: a generalization of the indicator tree

Example (Indicator tree)



General situation

- More than one rational subtrees
- The infinite path is interleaved with a fixed finite path.

Sturmian graphs

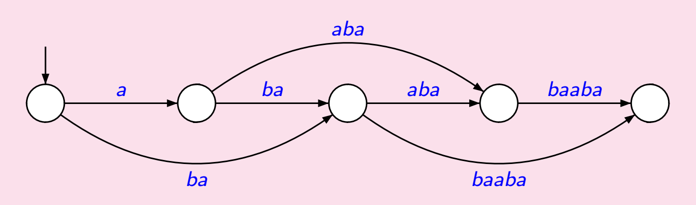
This name is given in the paper
Ch. Epifanio, F. Mignosi, J. Shallit, I. Venturini, "On Sturmian graphs", *Discrete Appl. Math* **155** (2007), 1014-1030, to graphs associated to central words.

Definition

The CDWAG (compact directed acyclic word graph) $G(w)$ of a word w is the minimal automaton recognizing the set of suffixes of w, after removing nonfinal states with outdegree 1.

Example

For $w = abaababaaba$, the automaton $G(w)$ is (all states are final)



Directive sequence of central words

Definition

The standard word s produced by $d = (d_0, d_1, \ldots, d_k)$ is $s = s_k$, where $s_{-1} = b$, $s_0 = a$, $s_{n+1} = s_n^{d_n} s_{n-1}$. The central word produced by d is the word $c = s^{=}$, that is s without its last two letters.

Example

- For $d = (1, 2, 2)$, the standard words are $s_1 = ab$, $s_2 = (ab)^2 a$, $s = s_3 = (ababa)^2 ab$. The central word is

  $$c = ababaababa$$

  Observe that $c = ababaababa = u_0 u_1^2 u_2$, where $u_i = \widetilde{s_i}$.
- For $d = (1, 2, 1, 1)$, the standard words are $s_1 = ab$, $s_2 = (ab)^2 a$, $s_3 = ababaab$, and $s = s_4 = ababaababa ba$ It is the same as the word s_3 up to the two last letters. So it defines the same central word.

Directive sequences $(d_0, \ldots, d_n, 1)$ and $(d_0, \ldots, d_n + 1)$ produce the same standard word up to the two last letters, so they produce the same central word.

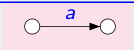
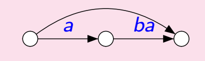
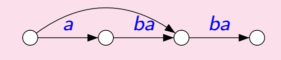
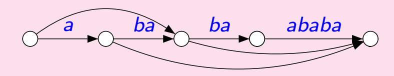
CDAWGs of central words

Construction

The CDAWG of a central word c with dirctive sequence $d = d'1$ is constructed by induction. Set $d' = d''\delta$.

1. if $\delta \neq 1$, repeat the last edge of the graph of d'.
2. otherwise, add a new state and $1 + \delta$ edges to this state. All these edges have the same label.

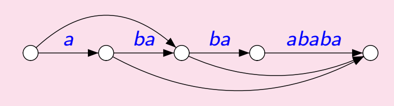
Example

d	s	c	G
$11 = 2$	*aba*	*a*	
$12 = 111$	*ababa*	*a|ba*	
$13 = 121$	*ababaab*	*a|ba|ba*	
$122 = 1211$	*ababaababaab*	*a|ba|ba|ababa*	

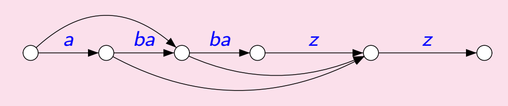
Example (continued)

Notation : s_d the word produced by the directive sequence d, $u_d = \widetilde{s}_d$. For instance $s_{122} = ((ab)(ab)a)(ababa)ab$, $s_{1211} = ((ab)(ab)a)(ab)(ababa)$.

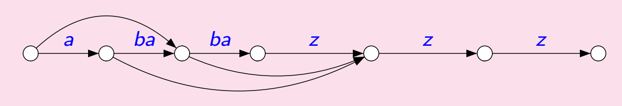
$d = 1211$, $s_{1211} = ababaabababa$, $c = a|ba|ba|z = u_0 u_1^2 u_{12}$ (with $z = ababa$)



$d = 123 = 1221$, $s_{1221} = ababaababaababababa$, $c = a|ba|ba|z|z = u_0 u_1^2 u_{12}^2$

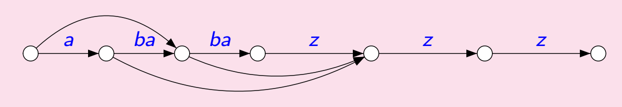


$d = 1231$, $s_{1231} = ababaababaababaababaab$, $c = a|ba|ba|z|z|z = u_0 u_1^2 u_{12}^3$

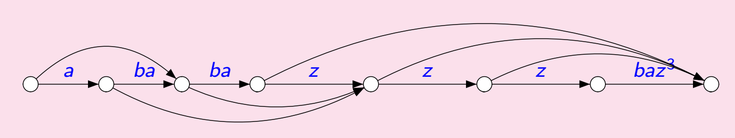


Example (end)

$d = 1231$, $s_{1231} = ababaababaababaababaab$, $c = a|ba|ba|z|z|z = u_0 u_1^2 u_{12}^3$



$d = 1232 = 12311$, $s_{12311} = ababaababaababaababaababaababaababaab$,
$c = a|ba|ba|z|z|z|baz^3 = u_0 u_1^2 u_{12}^3 u_{123}$



Size of the CDAWG

Length of the central word

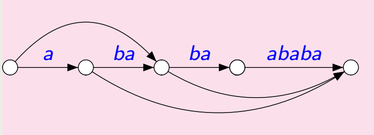
The length of the central word c defined by $d = (d_0, d_1, \ldots, d_k)$ is $|\ell_k| - 2$, where $\ell_n = |s_n| - 2$ and

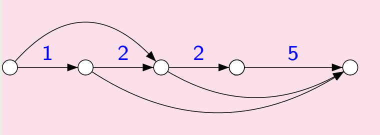
$$\ell_{-1} = \ell_0 = 1, \quad \ell_{n+1} = d_n \ell_n + \ell_{n-1} .$$

Observation

Let $H(c)$ be the graph obtained from the $G(c)$ by replacing each label by its length. $H(c)$ counts from 0 to $|c|$: each integer h with $0 \leq h \leq |c|$ is the sum of the weights of exactly one path in $H(c)$ starting at the initial state.

Example





Size of counting graphs

Problem

What is the minimal size of a graph with outdegree at most 2 counting from 0 to n?

If the size of the labels increase exponentially, like for the Fibonacci word, then the size is $O(\log n)$. It is conjectured that the bound $O(\log n)$ always holds. Related to the following conjecture.

Conjecture (Zaremba)

There exists an integer K such that for all positive m, there exists some $i \perp m$, $i < m$ such that all partial quotients in the continued fraction expansions of i/m are bounded by K.

Sturmian graphs