Random generation of Finite Sturmian Words *

Jean Berstel [(1)] (berstel@litp.ibp.fr)

Michel Pocchiola [(2)] (pocchiol@dmi.ens.fr)

(1) Laboratoire d'Informatique Théorique et de Programmation (LITP)
Institut Blaise Pascal
4, place Jussieu, 75252 Paris Cédex 05

(2) Laboratoire d'Informatique de l'Ecole normale supérieure (LIENS)
URA 1327, CNRS, 45 rue d'Ulm, 75230 Paris Cédex 05

June 16, 1999

Abstract

We present a bijection between the set $\mathcal{S}$ of factors of length m of Sturmian words and some set $\mathcal{T}$ of triples of nonnegative integers bounded by m. This bijection and its inverse are both computable in linear time. Its applications are : a bijective proof of Mignosi's formula for counting Sturmian words, a linear probabilistic algorithm for generating finite sturmian word at random, a linear on-line algorithm for computing the longest Sturmian prefix of a given word.

The construction of the bijection relies on concepts from combinatorial geometry.

1 Introduction

The Sturmian word with real parameters $\alpha, \beta \in [0, 1[$ is defined as the 0-1 bi-infinite sequence

$$(\lfloor \alpha(n+1) + \beta \rfloor - \lfloor \alpha n + \beta \rfloor)_{n \in \mathbb{Z}}. \tag{1}$$

Sturmian words have a long history. A clear exposition of early work by J. Bernoulli, Christoffel, and A. A. Markov is given in the book by

*Partially supported by PRC "Mathématiques et Informatique"

Venkov [21]. The term "sturmian" has been used by Hedlund and Morse in their development of symbolic dynamics [8, 9, 10]. The same objects are known as "characteristic" sequences, "cutting" sequences, "beatty" sequences or "billiard" trajectories. There is a large literature about properties of these sequences (see for example Series [19], Fraenkel *et al.* [7], Stolarsky [20]). From a combinatorial point of view, they have been considered by Rauzy [14, 15, 16], Brown [3], Ito, Yasutomi [11] in particular in relation with iterated morphisms, and by Séébold [18], Mignosi [12]. Sturmian words appear in ergodic theory [13], in computer graphics [2], and in crystallography. Dulucq and Gouyou-Beauchamps [5] considered the set of all finite words that are factors of some Sturmian word. They proved that the complement, say C, of this set is a context-free language, and they conjectured that C is inherently ambiguous. To show this, they in fact conjectured a formula for the number of factors of length m of Sturmian words. Since the generating series of these numbers is transcendental, the Chomsky-Schützenberger theorem would prove inherent ambiguity (see Flajolet [6] for a systematic exposition). This formula was proved later by Mignosi [12] and will be obtained in the present paper as a easy corollary of our main result.

The aim of this paper is to present a bijection between the set $\mathcal{S}$ of factors of length m of Sturmian words and some set $\mathcal{T}$ of triples (p, q, a) of nonnegative integers bounded by m. This bijection and its inverse are both computable in linear time. The derivation of the bijection involves the analysis of the partition of lines induced by the lattice points and some elementary concepts of combinatorial geometry.

The bijection has several interesting applications. First we obtain a straightforward proof of Mignosi's enumeration formula for factors of Sturmian words. A second consequence is a algorithm for random generation of Sturmian words of bounded given length which may have some application in Computer Graphics. More precisely it is a probabilistic algorithm in the sense that it successfully terminates only in the average. A third application is a linear time algorithm to test whether a given finite word is Sturmian. More precisely we give a linear algorithm to compute the longest prefix of a word that is Sturmian; this algorithm is on-line and test in constant time whether the given prefix can be extented by the next input symbol into a Sturmian word.

The present version of the paper contains only sketches of the proofs.

2 Results

We call a factor of length m of a Sturmian word, a *Sturmian m-factor.* It is easily verified that the sequence

$$(\lfloor \alpha(n+1)+\beta \rfloor - \lfloor \alpha n+\beta \rfloor)_{0\le n<m} \tag{2}$$

ranges over all Sturmian m-factors when α and β range over $[0,1[$. We denote by $\mathcal{S}$ the set of Sturmian m-factors.

The main result of the paper is a description of a natural bijection between the set $\mathcal{S}$ and a subset $\mathcal{T}$ of $\{0,1,\ldots,m\}^3$. The set $\mathcal{T}$ is the set of triples (a,p,q) such that either $a=m$ and $p=q=1$, or $0\le a\le m-p$ and $1\le q\le p$ with p and q coprime. Let B be the mapping which associates to a triple $(a,p,q)\in\mathcal{T}$ the sequence

$$B(a,p,q)=u_1-u_0, u_2-u_1,\ldots,u_m-u_{m-1} \tag{3}$$

where

$$u_n=\begin{cases}\lfloor \alpha n+\beta\rfloor, & \text{for } n=0,\ldots,a;\\ \lceil \alpha n+\beta\rceil-1, & \text{for } n=a+1,\ldots,m,\end{cases} \tag{4}$$

with $\alpha=q/p$ and $\beta=\lceil \alpha a\rceil-\alpha a$.

The sequence $B(a,p,q)$ is a $\{0,1\}$-sequence because α is less than 1.

Theorem 1 *The mapping B is a bijection from $\mathcal{T}$ onto the set $\mathcal{S}$ of Sturmian m-factors.*

Observe that B is computable in linear time. The fact that the inverse mapping B^{-1} is also computable in linear time will be a consequence of Proposition 2 below.

The proof is based on a geometric encoding of Sturmian words. Sturmian m-factors are defined by straight lines but many different straight lines may defined the same Sturmian m-factor. Each equivalence class is a convex polygon in the space of lines, and thus may be characterized by choosing a suitable edge. In some sense this is equivalent to represent a class by a pair of lines. This representation is specially convenient for computation because it allows to derive in a simply manner the associate triple in $\mathcal{T}$. This construction will be detailed in the next section.

We now come to the applications.

Mignosi [12] proved the following result.

Proposition 1 *The number of factors of length m of Sturmian words is given by the sum*

$$1 + \sum_{1}^{m} (m - i + 1)\phi(i), \tag{5}$$

where ϕ is the Euler function, i.e., $\phi(n)$ is the number of natural integers less than n and coprime to n.

The authors of the present paper gave an other proof in [1]. In view of Theorem above this proposition becomes straightforward. Indeed the formula (5) counts precisely the number of elements in set $\mathcal{T}$.

The next application concerns random generation of Sturmian m-factors. For this it suffices in view of Theorem above to generate triples (a, p, q) in the set $\mathcal{T}$ with uniform distribution. We implement this by the usual rejection algorithm : we generate triple (a, p, q) in $\{0, 1, \ldots, m\}^3$ uniformaly and reject whose which are not in $\mathcal{T}$. Since the size of $\mathcal{T}$ is asymptotically equal to m^3/π^2 the expected number of sampling is asymptotically π^2.

Finally we consider the problem of testing if a given word is Sturmian. More precisely we shall prove the following result.

Proposition 2 *Given a word of $\{0, 1\}^*$ its maximal Sturmian prefix can be computed on-line in time proportional to its size.*

Briefly the algorithm runs as follow. For each Sturmian prefix we maintain the polygon (in the space of lines) of all lines defining this Sturmian factor. We incrementaly transform this polygon by adding the geometric constraints defined by a new letter. The main point is that this can be done in constant time because the current polygon always has at most 4 edges. Details will be given in the last section.

3 The bijection

Recall that a *Sturmian m-factor* is a factor of length m of a Sturmian word, and that the sequence

$$(\lfloor \alpha(n + 1) + \beta \rfloor - \lfloor \alpha n + \beta \rfloor)_{0 \le n < m} \tag{6}$$

ranges over all Sturmian m-factors when α and β range over $[0, 1[$. We denote by $\mathcal{S}$ the set of Sturmian m-factors.

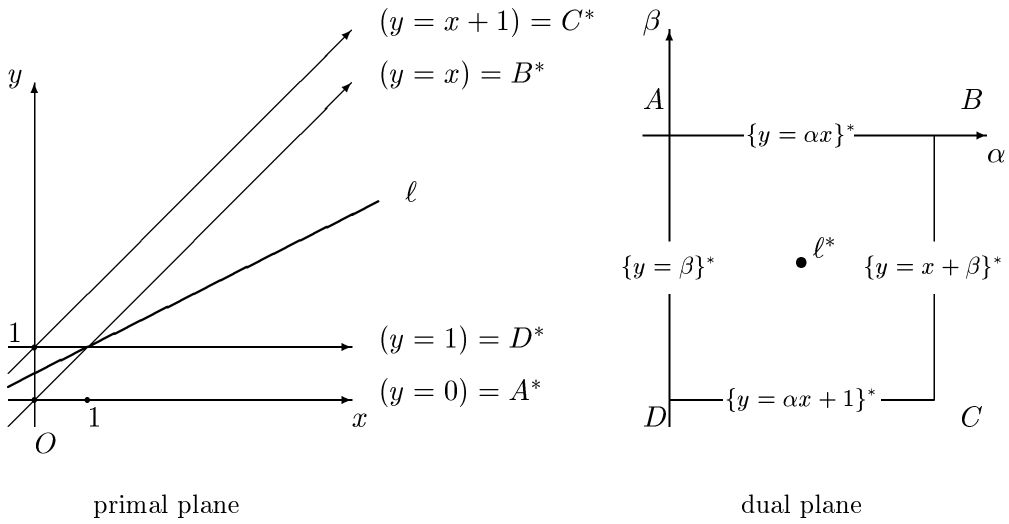

Figure 1: The duality transform.

In the sequel we consider the Sturmian m-factor (6) as a function of the straight line represented in Cartesian coordinates by the linear equation

$$y = \alpha x + \beta. \tag{7}$$

The real α is the *slope* of the straight line and the real β may be called the *intercept*. For a straight line ℓ with equation (7), we denote by $S(\ell)$ the Sturmian m-factor defined by (6) and by ℓ^+ the closed half-plane $y \leq \alpha x + \beta$.

Let $\mathcal{H}$ be the set of straight lines with finite slope and let $\mathcal{L}$ be the set of straight lines with equation $y = \alpha x + \beta$ with α and β in $[0, 1[$. For ℓ and ℓ' in $\mathcal{H}$ the *segment* $[\ell, \ell']$ is the set of straight lines with equation

$$y = (u\alpha + (1-u)\alpha')x + (u\beta + (1-u)\beta')$$

where u ranges over $[0, 1]$ and where α, α' and β, β' are the slopes and intercepts of ℓ and ℓ', respectively. Related notions, as *open segment, polygon, convexity*, etc. are defined as usual and freely used in the sequel.

The mapping S induces a natural partition of $\mathcal{L}$ whose parts are the $S^{-1}(s)$ when s ranges over $\mathcal{S}$. The following proposition relates this partition to the lattice set H defined by

$$H = \{\, (x, y) \in \mathbb{N} \mid 0 \leq x \leq m \,\}. \tag{8}$$

Proposition 3 *Two lines ℓ and ℓ' define the same Sturmian m-factor, i.e. $S(\ell) = S(\ell')$, if and only if*

$$\ell^+ \cap H = \ell'^+ \cap H.$$

Proof. The set $\ell^+ \cap H$ is clearly defined by the lattice points $(n, \lfloor \alpha n + \beta \rfloor)$ for $n = 0, \ldots, m$, and consequently by the Sturmian m-factor defined by ℓ since $\beta \in [0, 1[$. □

As shown by the previous proposition the Sturmians m-factors are closely related to the way lines separate the lattice set H or equivalently the way lattice points of H separate the set of lines $\mathcal{L}$. We briefly make an account of this classical subject of combinatorial geometry.

For any line $\ell \in \mathcal{H}$, let $c(\ell)$ denotes the number of lattice points in H lying on the line ℓ, i.e., $c(\ell) = \mathrm{Card}(\ell \cap H)$. The function $c : \mathcal{L} \to \mathbb{N}$ induces a partition of $\mathcal{L}$ whose parts or cells are the maximal connected components of $\mathcal{L}$ where the function c remains constant; we denote by $\mathcal{A}$ this partition. One can easily verify that the cells are relatively open convex polygons of $\mathcal{L}$ of dimension 2, 1 or 0 according to the value 0, 1 or ≥ 2 of the function c on the cell. The cells of dimension $0, 1$ and 2 are respectively called the vertices, edges and faces of the partition $\mathcal{A}$; theirs sets are denoted $\mathcal{V}$, $\mathcal{E}$ and $\mathcal{F}$, respectively.

To represent sets of lines by sets of points we use the duality transform $x \in \mathcal{H} \cup \mathcal{P} \mapsto x^* \in \mathcal{H} \cup \mathcal{P}$. Duality maps the line $\ell \in \mathcal{H}$ with equation $y = \alpha x + \beta$ to the point $\ell^* \in \mathcal{P}$ with coordinates $(\alpha, -\beta)$ and the point p with coordinates (α, β) to the line p^* with equation $y = \alpha x - \beta$. It can be easily verify that the duality transform is an involution, and that it preserves the incidence relation i.e.,

$$p \in \ell \Leftrightarrow \ell^* \in p^*, \qquad p \in \ell^+ \Leftrightarrow \ell^* \in (p^*)^+. \tag{9}$$

The (x, y)-plane and (α, β)-plane are called the *primal* and *dual* plane, respectively. The segment $[\ell, \ell']$ is then represented in the dual plane by the segment $[\ell^*, \ell'^*]$ (see Figure 1).

By the duality transform the partition $\mathcal{A}$ is represented by the arrangement of the square $\mathcal{L}^* = \{(\alpha, \beta) \mid 0 \leq \alpha, -\beta < 1\}$ induced by the dual lines in H^* which intersect the square $\mathcal{L}^*$, namely the $m(m+1)/2$ lines duals of the lattice points (x, y) such that $1 \leq y \leq x \leq n$. The partition $\mathcal{A}$, for $m = 5$, is represented in the dual plane in Figure 2: there are 17 vertices, 41 edges and 24 faces. Observe that we don't count the vertices and edges

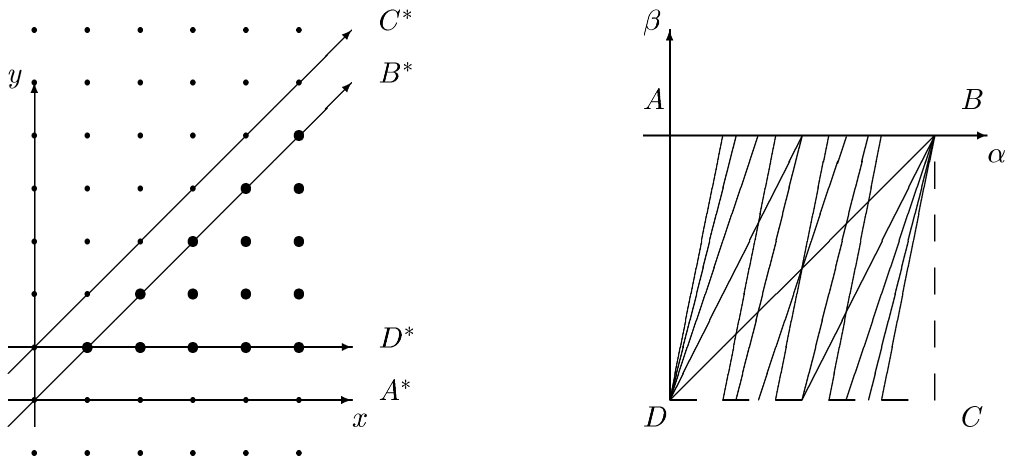

Figure 2: The partition $\mathcal{A}$ for $m = 5$

lying on the boundary $[D, C] \cup [B, C]$ of the square $\mathcal{L}^*$ and the external face since they don't belong to $\mathcal{V} \cup \mathcal{E} \cup \mathcal{F}$.

The following proposition asserts that the partition $\mathcal{A}$ is a refinement of the partition $\{S^{-1}(s)\}_{s \in \mathcal{S}}$.

Proposition 4 *The function S is constant on each cell of the partition $\mathcal{A}$.*

We denoted still by S the extension of S to $\mathcal{A}$. At this point the set $\mathcal{V}, \mathcal{E}, \mathcal{F}$ are natural candidates for an encoding of the Sturmian m-factors; in fact the following proposition discards the set $\mathcal{V}$ of vertices.

Proposition 5 *The restrictions of the function S to the sets $\mathcal{V}, \mathcal{E}, \mathcal{F}$ are injective (but not surjective), surjective (but not injective) and bijective, respectively.*

The encoding of the Sturmian words by the faces of the partition $\mathcal{A}$ is somehow difficult to handle. On the other side the edges are much easier to describe. This is done as follow.

For $E \in \mathcal{E}$ we denote by $\widehat{E}$ the common lattice point of the straight line in E, i.e. $\{\widehat{E}\} = \bigcap_{\ell \in E} \ell = \bigcap E$, and we denote by $\sup E$ ($\inf E$) the straight line through $\widehat{E}$ whose slope is the upper bound (lower bound) of the slopes of the lines in E. Clearly $E =]\inf E, \sup E[$.

Proposition 6 *Let $E \in \mathcal{E}$ and let (a, b) be the coordinates of the lattice point E^* and let $y = \alpha x + \beta$ the equation of the line* $\sup E$. *One has*

1. *α is a rational and $0 < \alpha \leq 1$*

2. *$0 \leq a \leq m$*

3. *$\beta = \lceil \alpha a \rceil - \alpha a$, $b = \lceil \alpha a \rceil$.*

4. *The Sturmian m-factor $S(E)$ is*

$$u_1 - u_0, u_2 - u_1, \ldots, u_m - u_{m-1} \tag{10}$$

where

$$u_n = \begin{cases} \lfloor \alpha n + \beta \rfloor, & \textit{for } n = 0, \ldots, a; \\ \lceil \alpha n + \beta \rceil - 1, & \textit{for } n = a + 1, \ldots, m. \end{cases} \tag{11}$$

Observe that the word $S(E)$ depends only of the triple (a, q, p) where $q/p = \alpha$.

The mapping S is not injective on the set of edges $\mathcal{E}$. We therefore shall distinguish a special edge among all edges associated to some Sturmian m-factor. This will be done by an ordering relation. For $s \in \mathcal{S}$ we put

$$\mathcal{E}(s) = \{\, E \in \mathcal{E} \mid S(E) = s \,\}, \tag{12}$$

and we define an ordering $\prec$ of the set $\mathcal{E}(s)$ by $E \prec E'$ if the first coordinate of $\widehat{E}$ is less than or equal to the first coordinate of $\widehat{E}'$.

Proposition 7 *The relation $\prec$ is a total order relation on $\mathcal{E}(s)$ for any $s \in \mathcal{S}$.*

Let $\mathcal{E}\!\downarrow$ be the set of minimal elements of $\mathcal{E}$ for the relation $\prec$. And let R be the application which maps each Sturmian m-factors s on the minimal element of the set of edges E such that $S(E) = s$.

Proposition 8 *The application R is a bijection from $\mathcal{S}$ to $\mathcal{E}\!\downarrow$ whose inverse map is the restriction of S to $\mathcal{E}\!\downarrow$.*

It remains to identify the set of triples (a, p, q) associated to the edges of $\mathcal{E}\!\downarrow$.

Proposition 9 *The application which associate to each edge $E \in \mathcal{E}$ the triple (a, p, q) such that a is the first coordinate of the lattice point $\widehat{E}$ and where q/p (with p and q coprime) is the slope of* $\sup E$ *is a bijection from $\mathcal{E}\!\downarrow$ to the set of triples (a, p, q) such that either $a = m$ and $p = q = 1$, either $0 \leq a \leq m - p$ and $1 \leq q \leq p$ with p and q coprime.*

The affiliation of Propositions 9, 8 and 6 gives Theorem 1.

4 Recognition of Sturmian m-factors

We consider the problem of testing if a given word is Sturmian.

Let $s = v_1 v_2 \ldots v_m$ be a $\{0,1\}$-word of length m and let $u_0, \ldots, u_m$ be the integer sequence defined by

$$u_n = v_1 + v_2 + \cdots + v_n.$$

The word s is a Sturmian m-factor if and only if exists a line $\ell \in \mathcal{L}$ such that $S(\ell) = s$. In terms of the slope α and the intercept β of the line ℓ the word s is a Sturmian m-factor if and only if the system of linear inequations

$$u_n \leq \alpha n + \beta < u_n + 1 \qquad n = 1, 2, \ldots, m \tag{13}$$

admits a solution. Observe that each letter in the word s adds two constraints to the above linear system.

Using the Megiddo's algorithm [17, chap 15] the existence of a solution to (13) can be decided in linear time in the number of constraints i.e., m in our case. However we can do much better in our case. Indeed it turns out that the size (number of vertices) of the polygon defined by (13) is at most 4.

Lemma 1 *Let s be a Sturmian word and let F be the face of the partition $\mathcal{A}$ such that $S(F) = s$. Then the size of the the convex polygon F is 3 or 4.*

Proposition 10 *Given a word of $\{0,1\}^*$ its maximal Sturmian prefix can be computed on line in time proportional to its size.*

Proof. For each Sturmian prefix we maintain the polygon (in the space of lines) of all lines defining this Sturmian factor. We incrementaly transform this polygon by adding the two geometric constraints defined by a new letter. Since the size of the current polygon is 3 or 4 the new constraints can be added in constant time. □

Acknowledgement

We would like to thank Dominique Gouyou-Beauchamps for helpful discussions.

References

[1] **J. Berstel, M. Pocchiola**, A geometric Proof of the Enumeration Formula for Sturmian Words, Technical Report LIENS-92-21

[2] **J. E. Bresenham**, Algorithm for computer control of a digital plotter, *IBM Systems J.* **4** (1965), 25–30.

[3] **Tom C. Brown**, A characterization of the quadratic irrationals, *Canad. Math. Bull.* **34** (1991), 36–41.

[4] **E. Coven, G. Hedlund**, Sequences with minimal block growth, *Math. Systems Theory* **7** (1973), 138–153.

[5] **S. Dulucq, D. Gouyou-Beauchamps**, Sur les facteurs des suites de Sturm, *Theoret. Comput. Sci.* **71** (1990), 381–400.

[6] **P. Flajolet**, Analytic models and ambiguity of context-free languages, *Theoret. Comput. Sci.* **49** (1987), 283–309.

[7] **A. S. Fraenkel, M. Mushkin, U. Tassa**, Determination of $\lfloor n\theta \rfloor$ by its sequence of differences, *Canad. Math. Bull.* **21** (1978), 441–446.

[8] **G.A. Hedlund**, Sturmian minimal sets, *Amer. J. Math* **66** (1944), 605–620.

[9] **G.A. Hedlund, M. Morse**, Symbolic dynamics, *Amer. J. Math* **60** (1938), 815–866.

[10] **G.A. Hedlund, M. Morse**, Sturmian sequences, *Amer. J. Math* **61** (1940), 1–42.

[11] **S. Ito, S. Yasutomi**, On continued fractions, substitutions and characteristic sequences, *Japan. J. Math.* **16** (1990), 287–306.

[12] **F. Mignosi**, On the number of factors of Sturmian words, *Theoret. Comput. Sci.* **82** (1991), 71–84.

[13] **M. Queffélec**, *Substitution Dynamical Systems – Spectral Analysis* Lecture Notes Math.,vol. 1294, Springer-Verlag, 1987.

[14] **G. Rauzy**, Suites à termes dans un alphabet fini, *Sémin. Théorie des Nombres* (1982–1983), 25-01,25-16, Bordeaux.

[15] **G. Rauzy**, Mots infinis en arithmétique, in :*Automata on infinite words* (D. Perrin ed.), *Lect. Notes Comp. Sci.* **192** (1985), 165–171.

[16] **G. Rauzy**, Sequences defined by iterated morphisms, in :*Workshop on Sequences* (R. Capocelli ed.), *Lecture Notes Comput. Sci.*, to appear.

[17] **A. Schrijver**, Theory of Linear and Integer Programming, *Wiley* (1986)

[18] **P. Séébold**, Fibonacci morphisms and Sturmian words, *Theoret. Comput. Sci.* **88** 1991, 367–384.

[19] **C. Series**, The geometry of Marhoff numbers, *The Mathematical Intelligencer* **7** (1985), 20–29.

[20] **K. B. Stolarsky**, Beatty sequences, continued fractions, and certain shift operators, *Cand. Math. Bull.* **19** (1976), 473–482.

[21] **B. A. Venkov**, *Elementary Number Theory*, Wolter-Noordhoff, Groningen, 1970.