Coding with variable block maps

Vicente Acuña[a], Gilles Didier[b], Alejandro Maass[c,*]

[a] *Laboratory of Bioinformatics and Mathematics of Genome, Centro de Modelamiento Matemático UMR 2071 UCHILE-CNRS, Casilla 170/3 correo 3, Santiago, Chile*

[b] *Institut de Mathématiques de Luminy, 163 avenue de Luminy, Case 907, 13288 Marseille Cedex 9, France*

[c] *Departamento de Ingeniería Matemática and Laboratory of Bioinformatics and Mathematics of Genome, Centro de Modelamiento Matemático, Universidad de Chile, UMR 2071 UCHILE-CNRS, Casilla 170/3 correo 3, Santiago, Chile*

Abstract

In this article we study a special class of sliding block maps that we call variable block maps. We characterize the subsets of finite and infinite sequences that can be obtained as the image of another subset of symbolic sequences by a variable block map. On the other way, we show that the coding process induced by such kind of block maps can be reversed, even with partial knowledge about the variable block maps, and we give an explicit construction of a canonical antecedent.

Keywords: Symbolic sequence; Prefix code; Recoding

1. Introduction

One of the fundamental purposes in Symbolic Dynamics is to recode systems of finite or infinite words into others in order to be able to interpret them in a given framework. For instance, the theory of finite state codes aims to construct good representations of a given symbolic channel using the technique of splitting and amalgamation of leveled graphs (sofic systems in the language of symbolic dynamics). The literature in this subject and its applications is huge (for a complete list of references see [6]).

A basic coding operation in Symbolic Dynamics is the N-block presentation of a sequence (in fact, this is a sequence of splitting). That is, each symbol of the given sequence is replaced by the word consisting of itself and the next $N-1$ following symbols in the sequence, each time this is possible. The classical use of this operation is rather technical: to transform a given subshift of finite type into a topological Markov chain. The application of this operation which motivates this article was introduced in [3]. In this article the author considers the N-block presentation operation as a way to understand the role of a symbol in a sequence in relation to its context. The main result of [3] states that given a symbolic sequence w there exists a *maximal* sequence w^* written in a bigger alphabet, called the N-antecedent of w,

* Corresponding author.

E-mail addresses: viacuna@dim.uchile.cl (V. Acuña), didier@iml.univ-mrs.fr (G. Didier), amaass@dim.uchile.cl (A. Maass).

such that: (i) the N-block presentations of w^* and w coincide up to a bijection of the letters appearing in the respective N-block presentations and (ii) any other sequence verifying property (i) can be obtained from w^* by a letter-to-letter projection. The idea was that the maximal sequence found by this method can distinguish the context of a symbol in the starting sequence (here represented by the next $N-1$ symbols). This construction was applied in genetic sequences analysis to improve significantly the manual alignments between different HIV nucleotide sequences [5] and to measure similarity between sequences without alignment [4].

Although the notion of "the context" of a symbol inside a sequence or its role inside the sequence could be vague, a natural idea is to define it statistically as those words containing the symbol in the sequence such that the conditional frequency of the symbol given each of such words does not change "significantly" if we extend the word. In [3] it is assumed that such words or context are deterministically chosen of constant length N to the right of the symbol.

In this article we introduce the notion of *variable block maps*. They are defined from finite families of finite words where we distinguish the position of a center origin, such that, two words in such family once centered in their corresponding origins have at least a different touching letter. This concept is proposed to be used as the deterministic context of the symbols in a given sequence. One of their properties is that it considers both directions in a sequence and it admits words of variable length. Then it can be expected to recover more relations among symbols and it can be adapted to the statistical definition of context.

Given a variable block map $\mathcal{V}$ we define the $\mathcal{V}$-coding of a sequence u to be the one obtained by replacing each symbol of u by the element of $\mathcal{V}$ that is centered around it, each time this procedure is possible. Then we study the question whether there is a *maximal sequence* (in the same sense as before) and a variable block map $\tilde{\mathcal{V}}$ such that its $\tilde{\mathcal{V}}$-coding coincides with the previous one once we rename letters. As in [3], we prove this is possible under some combinatorial conditions and we give explicit relations among symbols that allow to construct it. We also characterize the sequences that are $\mathcal{V}$-codings for some variable block map $\mathcal{V}$.

This article is organized as follows. In Section 2, we define variable block maps and introduce origin and length functions associated to them. Section 3 is devoted to the study of the length and origin functions. Characterization results are proved in Section 4, where Section 4.2 is devoted to prove maximality of the variable block map. Finally, in Section 5 we prove the following curiosity for symbolic dynamics: any mixing sofic system verifying a straightforward necessary condition can be obtained as a factor of a fullshift by a map defined by the length function of a variable block map.

2. Definitions and background

2.1. Words and subshifts

We give some basic concepts and notations from symbolic dynamics and language theory. More details can be obtained in [6] and references therein.

An alphabet $\mathcal{A}$ is a finite set of elements called symbols. A finite or infinite (onesided or twosided) sequence of symbols in $\mathcal{A}$ is called a word; they are written $w = w_0 \dots w_{n-1}$, $w = (w_i)_{i\in\mathbb{N}}$ and $w = (w_i)_{i\in\mathbb{Z}}$, respectively, and indexes are referred as the positions of the word. The length of w is denoted by $|w|$; the empty word, of length 0, is denoted by ε. The set of all finite words of length l over $\mathcal{A}$ is $\mathcal{A}^l$, the set of all finite words is $\mathcal{A}^* = \bigcup_{l\in\mathbb{N}} \mathcal{A}^l$ and the set of finite words of positive length is $\mathcal{A}^+ = \bigcup_{l>0} \mathcal{A}^l$. The sets of onesided and twosided infinite sequences in $\mathcal{A}$ are $\mathcal{A}^{\mathbb{N}}$ and $\mathcal{A}^{\mathbb{Z}}$, respectively.

Let $u, v \in \mathcal{A}^*$ and $x \in \mathcal{A}^{\mathbb{N}}$. The concatenation of u and v is denoted by $uv = u_0 \dots u_{|u|-1} v_0 \dots v_{|v|-1} \in A^*$ and $ux = u_0 \dots u_{|u|-1} x_0 x_1 \dots \in A^{\mathbb{N}}$ is the concatenation of the finite sequence u with the (onesided) infinite sequence x. If $t \in \mathcal{A}^* \bigcup \mathcal{A}^{\mathbb{N}}$, $u \in \mathcal{A}^*$ and $v \in \mathcal{A}^* \bigcup \mathcal{A}^{\mathbb{N}}$ are such that $t = uv$, then u is said to be a prefix of t and v a suffix of t.

Let $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$, $v \in \mathcal{A}^+$ and k a position of u. The word v occurs in u at position k if $u_k \dots u_{k+|v|-1} = v$. A finite word $v \in \mathcal{A}^+$ is a subword of $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ if v occurs in u at some position.

Let Σ be a subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$. Denote by $\mathrm{L}_n(\Sigma)$ the set containing the subwords of length n of elements in Σ. The set $\mathrm{L}(\Sigma) = \bigcup_{n\geqslant 1} \mathrm{L}_n(\Sigma)$ is called the language of Σ. For u in $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ we set $\mathrm{L}_n(u) = \mathrm{L}_n(\{u\})$ and $\mathrm{L}(u) = \mathrm{L}(\{u\})$.

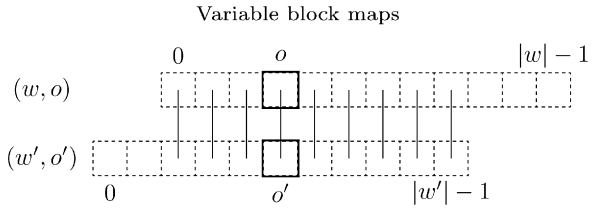

Fig. 1. Two pointed words. At least one of the coupled symbols in the positions joined by a vertical line has to be different if (w, o) and (w', o') belong to the same variable block map.

Let $\mathbb{K} = \mathbb{N}$ or $\mathbb{Z}$. The shift map $\sigma : \mathcal{A}^{\mathbb{K}} \to \mathcal{A}^{\mathbb{K}}$ is defined by $\sigma(x)_i = x_{i+1}$ for $x \in \mathcal{A}^{\mathbb{K}}$ and $i \in \mathbb{K}$. A set $\Sigma \subseteq \mathcal{A}^{\mathbb{K}}$ is said to be a subshift (onesided if $\mathbb{K} = \mathbb{N}$ and twosided if $\mathbb{K} = \mathbb{Z}$) if it is closed for the product topology of $\mathcal{A}^{\mathbb{K}}$ and invariant for the shift map: $\sigma(\Sigma) = \Sigma$. In this context $\mathcal{A}^{\mathbb{K}}$ is called a fullshift.

A subshift $\Sigma \subseteq \mathcal{A}^{\mathbb{K}}$ is of finite type (one abbreviates SFT) if there is an integer $L \geqslant 1$ such that given $x \in \mathcal{A}^{\mathbb{K}}$, $x \in \Sigma$ if and only if $x_i \dots x_{i+L-1} \in \mathrm{L}_L(\Sigma)$ for all $i \in \mathbb{K}$.

Let Σ and Γ be two subshifts (both one or twosided and not necessarily defined in the same fullshift). A block map between Σ and Γ is a map $\Pi : \Sigma \to \Gamma$ given by $\Pi(x)_i = \pi(x_{i-m} \dots x_i \dots x_{i+a})$ for any $x \in \Sigma$ and $i \in \mathbb{K}$, where $m, a \in \mathbb{N}$ (memory and anticipation, respectively) and $\pi : \mathrm{L}_{m+a+1}(\Sigma) \to \mathrm{L}_1(\Gamma)$. A block map is said to be a factor map if it is surjective and a conjugation if it is bijective.

A subshift Σ is called sofic if it is the image of a SFT by a block map (of course SFTs are sofic).

2.2. Variable block maps

2.2.1. Definition

Let $\mathcal{A}$ be an alphabet. A pointed word on $\mathcal{A}$ is a couple (w, o) where $w \in \mathcal{A}^+$ and $0 \leqslant o < |w|$ is an arbitrary position of w called its origin.

A finite set $\mathcal{V}$ of pointed words on $\mathcal{A}$ such that for all $(w, o), (w', o') \in \mathcal{V}$, $(w, o) \neq (w', o')$, there is an integer $k \in \{-\min\{o, o'\}, \dots, \min\{|w| - o, |w'| - o'\} - 1\}$ verifying $w_{o+k} \neq w'_{o'+k}$, is called a *variable block map* on $\mathcal{A}$ (see Fig. 1).

The following is an example of a variable block map on $\{a, b\}$:

$$\mathcal{V}_{\mathrm{ex}} = \{(\mathtt{aa}, 0), (\mathtt{aab}, 1), (\mathtt{ababa}, 2), (\mathtt{ba}, 0), (\mathtt{babb}, 2), (\mathtt{bb}, 0)\}.$$

Given a word u in $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ and a position k of u, there is at most one element $(w, o) \in \mathcal{V}$ such that $u_{k-o} \dots u_{k-o+|w|-1} = w$. It is worth mentioning that such a word may not exist. For instance, there is no pointed word in $\mathcal{V}_{\mathrm{ex}}$ for the underlined position of $u = \mathtt{bb\underline{a}b} \dots$.

2.2.2. Admissible positions

Let $\mathcal{V}$ be a variable block map over $\mathcal{A}$. Let $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}}$. If there exists a greatest position k of u and an element (w, k) in $\mathcal{V}$ such that w occurs in u at position 0 we say u has a $\mathcal{V}$-start position k and denote it by $\mathrm{start}_{\mathcal{V}}(u)$. Let $u \in A^+$. Symmetrically, if there exists a smallest position k of u and an element (w, o) in $\mathcal{V}$ such that w occurs in u at position $k - o$ and $|u| - k + o = |w|$ we say u has a $\mathcal{V}$-end position k and denote it by $\mathrm{end}_{\mathcal{V}}(u)$.

The set of $\mathcal{V}$-admissible positions of a word u is:

- $\mathbb{Z}$ if $u \in \mathcal{A}^{\mathbb{Z}}$;
- $\{n + \mathrm{start}_{\mathcal{V}}(u)\}_{n \in \mathbb{N}}$ if u has a start position and $\emptyset$ otherwise, if $u \in \mathcal{A}^{\mathbb{N}}$;
- $\{\mathrm{start}_{\mathcal{V}}(u), \dots, \mathrm{end}_{\mathcal{V}}(u)\}$ if u has both a start and an end position and $\emptyset$ otherwise, if $u \in \mathcal{A}^+$.

A word $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ is $\mathcal{V}$-compatible if its set of $\mathcal{V}$-admissible positions is not empty and if for all $\mathcal{V}$-admissible positions k there is an element $(w, o) \in \mathcal{V}$ such that w occurs in u at position $k - o$. A subset $S \subseteq \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ is $\mathcal{V}$-compatible if all of its elements are $\mathcal{V}$-compatible.

A finite set $\mathcal{P} \subseteq \mathcal{A}^+$ is a prefix code if no word in $\mathcal{P}$ is a prefix of another word in $\mathcal{P}$. A prefix code can be seen as a variable block map with all elements centered at zero. For more details about the "Theory of Codes" see [1].

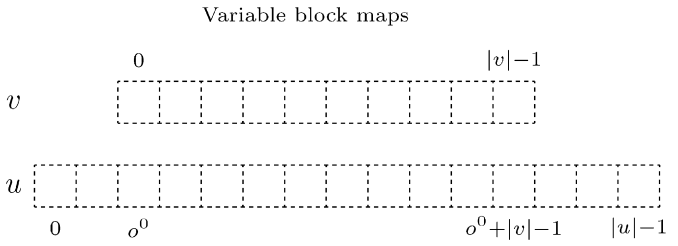

Fig. 2. Top: the coding. Bottom: the decoding.

2.2.3. $\mathcal{V}$-codings

Let $\mathcal{V}$ be a variable block map over $\mathcal{A}$ and $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ be a $\mathcal{V}$-compatible sequence. The $\mathcal{V}$-coding of u denoted $\Phi_{\mathcal{V}}(u)$ is the sequence v such that for all $\mathcal{V}$-admissible positions k of u:

- $v_{k-\text{start}_{\mathcal{V}}(u)}$ is the unique element $(w, o) \in \mathcal{V}$ such that w occurs in u at position $k - o$ if $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}}$;
- v_k is the unique element $(w, o) \in \mathcal{V}$ such that w occurs in u at position $k - o$ if $u \in \mathcal{A}^{\mathbb{Z}}$.

The $\mathcal{V}$-coding of a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ is the set of the $\mathcal{V}$-codings of its elements. From the definition of the $\mathcal{V}$-start position, if $v \in \mathcal{V}^{\mathbb{N}} \bigcup \mathcal{V}^+$ is a $\mathcal{V}$-coding of a $\mathcal{V}$-compatible sequence then, setting $v_l = (w[l], o^l)$ for all positions l of v, one has $o^0 > o^i - i$ for all $i > 0$. Symmetrically, from the definition of $\text{end}_{\mathcal{V}}(u)$, for $v \in \mathcal{V}^+$ one has $|w[|v| - 1]| - o^{|v|-1} > |w[i]| - o^i + i - |v| + 1$ for all $0 \leqslant i < |v| - 1$.

If all the elements of a variable block map $\mathcal{V}$ are centered at 0 and have the same length N, then the $\mathcal{V}$-coding of a sequence is nothing but its classical N-block presentation (see [6]).

Since elements of $\mathcal{V}$ overlap in the coding transformation, then not all sequences of elements of $\mathcal{V}$ are allowed in a $\mathcal{V}$-coding. For instance in any $\mathcal{V}_{\text{ex}}$-coding, element (aab, 1) can be followed only by (ba, 0) or (bb, 0). However, since all the "overlapping constraints" are local and a finite number, the subset of sequences of $\mathcal{V}^{\mathbb{Z}}$ which are $\mathcal{V}$-codings is a SFT. It comes that the subset of $\mathcal{V}$-compatible sequences of $\mathcal{A}^{\mathbb{Z}}$ is a SFT too.

The equivalent subsets of $\mathcal{V}^{\mathbb{N}}$, $\mathcal{V}^+$, $\mathcal{A}^{\mathbb{N}}$ and $\mathcal{A}^+$ are SFT with left and eventually right boundary conditions, so in the $\mathcal{V}^+$ and $\mathcal{A}^+$ cases, particular rational languages.

In the following "$\mathcal{V}$-coding" stands for "$\mathcal{V}$-coding of a $\mathcal{V}$-compatible sequence".

2.2.4. $\mathcal{V}$-decodings

The $\mathcal{V}$-decoding of a $\mathcal{V}$-coding $v \in \mathcal{V}^{\mathbb{Z}}$ is the sequence u defined for all $k \in \mathbb{Z}$ by $u_k = w_o$ where $(w, o) = v_k$.

Let $v \in \mathcal{V}^{\mathbb{N}} \bigcup \mathcal{V}^+$ be a $\mathcal{V}$-coding and set $v_l = (w[l], o^l)$ in all positions l of v. The $\mathcal{V}$-decoding of v (see Fig. 2) is the sequence u defined in each position $k \in \mathbb{N}$ if $v \in \mathcal{V}^{\mathbb{N}}$ or $0 \leqslant k < |v| + o^0 + |w[|v| - 1]| - o^{|v|-1}$ if $v \in \mathcal{V}^+$ by:

- $u_k = w[0]_k$ for $0 \leqslant k < o^0$;
- $u_{k+o^0} = w[k]_{o^k}$ for $k \in \mathbb{N}$ and $v \in \mathcal{V}^{\mathbb{N}}$;
- $u_{k+o^0} = w[k]_{o^k}$ for $0 \leqslant k < |v|$ and $v \in \mathcal{V}^+$;
- $u_{k+o^0} = w[|v| - 1]_{k-|v|+1+o^{|v|-1}}$ for $|v| \leqslant k < |v| + |w[|v| - 1]| - o^{|v|-1}$ and $v \in \mathcal{V}^+$.

The $\mathcal{V}$-decoding of a subset of $\mathcal{V}^+ \bigcup \mathcal{V}^{\mathbb{N}} \bigcup \mathcal{V}^{\mathbb{Z}}$ is the set of the $\mathcal{V}$-decodings of its elements.

2.2.5. Remarks

One can find the rules defining the $\mathcal{V}$-compatible sequences somehow restrictive. Nevertheless, they make the $\mathcal{V}$-coding and the $\mathcal{V}$-decoding transformations one to one and onto from the set of $\mathcal{V}$-compatible sequences of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ to its image. This fact is particularly important for finite or onesided infinite sequences.

For instance, the $\mathcal{V}_{\text{ex}}$-coding of aabaa is (aab, 1)(ba, 0)(aa, 0) and not
(aa, 0)(aab, 1)(ba, 0)(aa, 0) which, with our definition, is not even a $\mathcal{V}_{\text{ex}}$-coding.

Let $\mathcal{B}$ be a finite alphabet of the same cardinality as $\mathcal{V}$ and let $\theta : \mathcal{V} \to \mathcal{B}$ be a bijection. The (block) map $\Pi_{\mathcal{V},\theta}$ is defined on $\mathcal{V}$-compatible sequences $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ by $\Pi_{\mathcal{V},\theta}(u) = \theta(\Phi_{\mathcal{V}}(u)) \in \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$, where θ is applied componentwise. Given a $\mathcal{V}$-compatible sequence $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$, $\Pi_{\mathcal{V},\theta}(u)$ is called the $(\mathcal{V}, \theta)$-coding

of u. A sequence $v \in \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ is said to be a $\mathcal{V}$-coding if there exists a bijection $\theta : \mathcal{V} \to \mathcal{B}$ such that it is the $(\mathcal{V}, \theta)$-coding of some $\mathcal{V}$-compatible sequence $u \in \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$.

3. Length and origin relations

We have defined the notion of $\mathcal{V}$-coding for finite and infinite $\mathcal{V}$-compatible sequences. Our purpose (in next section) is to give a characterization of symbolic sequences that are $\mathcal{V}$-codings for some variable block map $\mathcal{V}$ as it was done in [3] for constant length prefix codes. In this purpose one has to consider the length and origin position of the words in the code as new variables.

A length–origin pair over a finite alphabet $\mathcal{B}$ is a couple of functions (L, O) with $\mathrm{L} : \mathcal{B} \to \mathbb{N}$ (length function) and $\mathrm{O} : \mathcal{B} \to \mathbb{N}$ (origin function) such that $0 \leqslant \mathrm{O}(b) < \mathrm{L}(b)$ for all $b \in \mathcal{B}$. Given a centered code $\mathcal{V}$ on the alphabet $\mathcal{A}$, there is a natural length–origin pair $(\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})$ on $\mathcal{V}$ defined, respectively, by $\mathrm{O}_{\mathcal{V}}((w, o)) = o$ and $\mathrm{L}_{\mathcal{V}}((w, o)) = |w|$ in any $(w, o) \in \mathcal{V}$. In this case, we also define the "word" function $\mathrm{W}_{\mathcal{V}}((w, o)) = w$.

Definition 1. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$. One says that $[X, (\mathrm{L}, \mathrm{O})]$ is boundary-standard if:

- for all $u \in X \bigcap (\mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}})$ one has $\mathrm{O}(u_0) > \mathrm{O}(u_i) - i$ for all positions $i > 0$ of u;
- for all $u \in X \bigcap \mathcal{B}^+$ one has $\mathrm{L}(u_{|u|-1}) - \mathrm{O}(u_{|u|-1}) > \mathrm{L}(u_i) - \mathrm{O}(u_i) - |u| + 1 + i$ for all positions $0 \leqslant i < |u| - 1$ of u.

The following remark is plain.

Remark 2. Let $\mathcal{V}$ be a variable block map over $\mathcal{A}$. If a subset $X \subseteq \mathcal{V}^+ \bigcup \mathcal{V}^{\mathbb{N}} \bigcup \mathcal{V}^{\mathbb{Z}}$ is the $\mathcal{V}$-coding of a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ then $[X, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]$ is boundary-standard.

Definition 3. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$ such that $[X, (\mathrm{L}, \mathrm{O})]$ is boundary-standard. We call (L, O)-antecedent of X any symbolic subset $Y \subseteq \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$, where $\mathcal{A}$ is a finite alphabet, such that there exist a variable block map $\mathcal{V}$ over $\mathcal{A}$ with Y compatible with $\mathcal{V}$ and a bijection $\theta : \mathcal{V} \to \mathcal{B}$ verifying $\mathrm{L} = \mathrm{L}_{\mathcal{V}} \circ \theta^{-1}$, $\mathrm{O} = \mathrm{O}_{\mathcal{V}} \circ \theta^{-1}$ and X is the $(\mathcal{V}, \theta)$-coding of Y.

We will see in this section that, given $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and a length–origin pair (L, O) over $\mathcal{B}$, it is not always possible to find an (L, O)-antecedent of X. We will need two additional concepts.

Definition 4. Let $\mathcal{A}$ be a finite alphabet, $S \subseteq \mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$, $\mathtt{a} \in \mathcal{A}$ and $k \in \mathbb{Z}$. Define by $\Delta_k^S(\mathtt{a})$ the subset:

- $\{\mathtt{a}\}$ if $k = 0$;
- $\{\mathtt{b} \in \mathcal{A} | \mathtt{a}w\mathtt{b} \in \mathrm{L}(S)$ for a word $w \in \mathcal{A}^{k-1}\}$ if $k > 0$;
- $\{\mathtt{b} \in \mathcal{A} | \mathtt{b}w\mathtt{a} \in \mathrm{L}(S)$ for a word $w \in \mathcal{A}^{-k-1}\}$ if $k < 0$.

Definition 5. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over the finite alphabet $\mathcal{B}$. For $(i, j) \in \mathbb{N}^2$, the relation $\overset{i,j}{\smile}$ over symbols of $\mathcal{B}$ is recursively defined by applying the rules: $\mathtt{a} \overset{i,j}{\smile} \mathtt{b}$ if

(1) $0 \leqslant i < \mathrm{L}(\mathtt{a})$, $0 \leqslant j < \mathrm{L}(\mathtt{b})$ and $\mathtt{a} \in \Delta^X_{j-\mathrm{O}(\mathtt{b})-i+\mathrm{O}(\mathtt{a})}(\mathtt{b})$;

(2) there exist $\mathtt{c} \in \mathcal{B}$ and $k \in \mathbb{N}$ such that $\mathtt{a} \overset{i,k}{\smile} \mathtt{c}$ and $\mathtt{c} \overset{k,j}{\smile} \mathtt{b}$.

Observe that if $\mathtt{a} \overset{i,j}{\smile} \mathtt{b}$ then $0 \leqslant i < \mathrm{L}(\mathtt{a})$ and $0 \leqslant j < \mathrm{L}(\mathtt{b})$. Nevertheless, this fact does not imply that $\mathtt{a} \in \Delta^X_{j-\mathrm{O}(\mathtt{b})-(i-\mathrm{O}(\mathtt{a}))}(\mathtt{b})$. From definition remark that $\mathtt{a} \overset{i,j}{\smile} \mathtt{b}$ if and only if $\mathtt{b} \overset{j,i}{\smile} \mathtt{a}$, and that, $\mathtt{a} \overset{i,k}{\smile} \mathtt{b}$ and $\mathtt{b} \overset{k,j}{\smile} \mathtt{c}$ implies $\mathtt{a} \overset{i,j}{\smile} \mathtt{c}$. This relation depends upon X and (L, O) but we do not use them as index to avoid extra notations. An intuitive interest of the previous defined relation is given by the following lemma. Its easy proof is left to the reader.

Lemma 6. *Let $\mathcal{V}$ be a variable block map over the alphabet $\mathcal{A}$, S a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$, T the $\mathcal{V}$-coding of S and $\mathtt{a}, \mathtt{b} \in \mathcal{V}$.*

(1) *For $k \in \mathbb{Z}$, $\mathtt{b} \in \Delta^T_k(\mathtt{a})$ if and only if there exists $w \in S$ and $0 \leqslant p < |w|$ such that $\mathrm{W}_{\mathcal{V}}(\mathtt{a})$ appears in w with its center $\mathrm{O}_{\mathcal{V}}(\mathtt{a})$ in position p and $\mathrm{W}_{\mathcal{V}}(\mathtt{b})$ appears in w with its center $\mathrm{O}_{\mathcal{V}}(\mathtt{b})$ in position $p + k$. Moreover,*

if $\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - 1 - \mathrm{O}_{\mathcal{V}}(\mathtt{a}) \geqslant k - \mathrm{O}_{\mathcal{V}}(\mathtt{b})$, *then*

$$\mathtt{a} \overset{\mathrm{O}_{\mathcal{V}}(\mathtt{a})+\ell, \mathrm{O}_{\mathcal{V}}(\mathtt{b})-k+\ell}{\smile} \mathtt{b}$$

for $\min\{\mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{O}_{\mathcal{V}}(\mathtt{b}) - k\} \leqslant \ell < \min\{\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b}) + k\}$.

(2) *For* $i, j \in \mathbb{N}$, *if* $\mathtt{a} \overset{i,j}{\smile} \mathtt{b}$ *then* $\mathrm{W}_{\mathcal{V}}(\mathtt{a})_i = \mathrm{W}_{\mathcal{V}}(\mathtt{b})_j$.

In what follows we investigate some relations between the language of a $\mathcal{V}$-coding and the corresponding length and origin functions associated to $\mathcal{V}$. An interesting fact is that the length function of symbols appearing more than once in the $\mathcal{V}$-coding and not only in a periodic way are bounded with respect to their first return times. The results obtained here are still partial in the sense that much more relations appear if we consider repetitions of words instead of symbols.

Theorem 7. *Let $\mathcal{V}$ be a variable block map over the alphabet $\mathcal{A}$, S a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ and T the $\mathcal{V}$-coding of S. For $\mathtt{a} \in L_1(T)$ and $k \in \mathbb{Z}$, if $\#(\Delta_k^T(\mathtt{a})) > 1$ then for all symbols $\mathtt{b} \in \Delta_k^T(\mathtt{a})$ at least one of the two following inequalities holds*:

- $\mathrm{O}_{\mathcal{V}}(\mathtt{a}) < \mathrm{O}_{\mathcal{V}}(\mathtt{b}) - k$;
- $\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}) < k + \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})$.

Proof. Assume there is a symbol $\mathtt{b} \in \Delta_k^T(\mathtt{a})$ such that $\mathrm{O}_{\mathcal{V}}(\mathtt{a}) \geqslant \mathrm{O}_{\mathcal{V}}(\mathtt{b}) - k$ and $\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}) \geqslant k + \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})$. In S this condition means that $\mathrm{W}_{\mathcal{V}}(b)$ is a subword of $\mathrm{W}_{\mathcal{V}}(a)$ that appears in position $\mathrm{O}_{\mathcal{V}}(a) + k - \mathrm{O}_{\mathcal{V}}(b)$.

Since $\#(\Delta_k^T(\mathtt{a})) > 1$ there is a symbol $\mathtt{c} \neq \mathtt{b}$ in $\Delta_k^T(\mathtt{a})$. By Lemma 6 there exists $w \in S$ and $0 \leqslant p < |w|$ such that $\mathrm{W}_{\mathcal{V}}(\mathtt{a})$ appears in w with its center in position p and $\mathrm{W}_{\mathcal{V}}(\mathtt{c})$ appears in w with its center in position $p + k$. But $\mathrm{W}_{\mathcal{V}}(\mathtt{b})$ is a subword of $\mathrm{W}_{\mathcal{V}}(\mathtt{a})$ that appears in w with its center in position $p + k$. This is in contradiction with the fact that T is a $\mathcal{V}$-coding of S. □

Lemma 8. *Let $\mathcal{V}$ be a variable block map over the alphabet $\mathcal{A}$, S a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ and T the $\mathcal{V}$-coding of S. For $\mathtt{a} \in \mathcal{V}$, if there exists a word $w \in \mathcal{V}^*$ such that $\mathrm{L}_{\mathcal{V}}(\mathtt{a}) \geqslant 2(|w| + 1)$ and $\mathtt{a}w\mathtt{a}$ occurs in T then $\mathtt{a} \overset{i \bmod |w|+1, i}{\smile} \mathtt{a}$ for all $0 \leqslant i < \mathrm{L}_{\mathcal{V}}(\mathtt{a})$.*

Proof. Since $\mathtt{a} \in \Delta_{|w|+1}^T(\mathtt{a})$ ($\mathtt{a}w\mathtt{a}$ occurs in T) with $|w| \leqslant \mathrm{L}_{\mathcal{V}}(\mathtt{a})$, one has $\mathtt{a} \overset{k, k+|w|+1}{\smile} \mathtt{a}$ for all $k \in \{0, \ldots, \mathrm{L}_{\mathcal{V}}(\mathtt{a}) - |w| - 2\}$. If $\mathrm{L}_{\mathcal{V}}(\mathtt{a}) \geqslant 2(|w|+1)$ the preceding relation holds in particular for all $k \in \{0, \ldots, \lfloor(\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - 1)/2\rfloor\}$. It comes that $\mathtt{a} \overset{i \bmod |w|+1, i}{\smile} \mathtt{a}$ for all $0 \leqslant i < \mathrm{L}_{\mathcal{V}}(\mathtt{a})$. □

Theorem 9. *Let $\mathcal{V}$ be a variable block map over the alphabet $\mathcal{A}$, S a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ and T the $\mathcal{V}$-coding of S. Let $\mathtt{a} \in \mathcal{V}$. If there exist $v \in (\mathcal{V} \setminus \{\mathtt{a}\})^*$ and $w \in (\mathcal{V} \setminus \{\mathtt{a}\})^*$ with $|v| \neq |w|$ such that $\mathtt{a}v\mathtt{a}$ and $\mathtt{a}w\mathtt{a}$ occur in T then*:

$$\mathrm{L}_{\mathcal{V}}(\mathtt{a}) < 2(\max\{|v|, |w|\} + 1).$$

Proof. Assume $|v| < |w|$ and $\mathrm{L}_{\mathcal{V}}(\mathtt{a}) \geqslant 2(|w| + 1)$. Put $g = \gcd(|v| + 1, |w| + 1)$.

Applying two times Lemma 8, one gets both $\mathtt{a} \overset{i \bmod |w|+1, i}{\smile} \mathtt{a}$ and $\mathtt{a} \overset{i \bmod |v|+1, i}{\smile} \mathtt{a}$ for all $0 \leqslant i < \mathrm{L}_{\mathcal{V}}(\mathtt{a})$. It comes $\mathtt{a} \overset{i \bmod g, i}{\smile} \mathtt{a}$ for all $0 \leqslant i < \mathrm{L}_{\mathcal{V}}(\mathtt{a})$.

Let $\mathtt{b}$ be the symbol occurring at position $g - 1$ of w. Observe that $|w| > g$ and thus $\mathtt{b} \neq \mathtt{a}$. Since $\mathtt{b} \in \Delta_g^S(\mathtt{a})$ one has $\mathtt{a} \overset{k+\mathrm{O}_{\mathcal{V}}(\mathtt{a})+g, k+\mathrm{O}_{\mathcal{V}}(\mathtt{b})}{\smile} \mathtt{b}$ for all $-\min\{\mathrm{O}_{\mathcal{V}}(\mathtt{a}) + g, \mathrm{O}_{\mathcal{V}}(\mathtt{b})\} \leqslant k < \min\{\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}) - g, \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})\}$. With the preceding relations, one gets $\mathtt{a} \overset{j+\mathrm{O}_{\mathcal{V}}(\mathtt{a}), j+\mathrm{O}_{\mathcal{V}}(\mathtt{b})}{\smile} \mathtt{b}$ for all $-\min\{\mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{O}_{\mathcal{V}}(\mathtt{b})\} \leqslant j < \min\{\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})\}$ which, with the second item of Lemma 6, leads to a contradiction with the fact that $\mathcal{V}$ is a variable block map. □

4. Characterization of $\mathcal{V}$-codings

In this section we will characterize the sets of sequences that are the $\mathcal{V}$-coding of a compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$, where $\mathcal{V}$ is a variable block map over the alphabet $\mathcal{A}$. To simplify statements we introduce a relation which is a special case of Definition 5.

Definition 10. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$, (L, O) be a length–origin pair over $\mathcal{B}$ and $k \in \mathbb{Z}$. The relation $\overset{k}{\sim}$ over symbols of $\mathcal{B}$ is defined by $\mathtt{a} \overset{k}{\sim} \mathtt{b}$ if one of the two following conditions holds:
- $\mathtt{a} = \mathtt{b}$ and $k \in \{-\mathrm{O}(\mathtt{a}), \ldots, \mathrm{L}(\mathtt{a}) - \mathrm{O}(\mathtt{a}) - 1\}$;
- $\mathtt{a} \overset{k+\mathrm{O}(\mathtt{a}),k+\mathrm{O}(\mathtt{b})}{\smile} \mathtt{b}$.

Again this relation depends upon X and (L, O) which are not used as index to avoid extra notations. With hypotheses and notations of Lemma 6 and by defining $\overset{k}{\sim}$ from T and $(\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})$ one has $\mathtt{a} \overset{k}{\sim} \mathtt{b}$ implies $\mathrm{w}_{\mathcal{V}}(\mathtt{a})_{k+\mathrm{O}(\mathtt{a})} = \mathrm{w}_{\mathcal{V}}(\mathtt{b})_{k+\mathrm{O}(\mathtt{b})}$ for all $\mathtt{a}, \mathtt{b} \in \mathcal{V}$ (Lemma 6 part (2)).

Let $k \in \mathbb{Z}$. We distinguish the subset of $\mathcal{B}$, $I_k = \{\mathtt{a} \in \mathcal{B} | -\mathrm{O}(\mathtt{a}) \leqslant k < \mathrm{L}(\mathtt{a}) - \mathrm{O}(\mathtt{a})\}$. First, observe that symbols in $\mathcal{B} \setminus I_k$ are not related by $\overset{k}{\sim}$ with any other symbol in $\mathcal{B}$ (not even themselves). Thus, for all $k \in \mathbb{Z}$ the relation $\overset{k}{\sim}$ is an equivalence relation over I_k.

Let $\mathtt{a} \in \mathcal{B}$ and $k \in \mathbb{Z}$. The notation $[\mathtt{a}]_k$ stands for:
- the equivalence class of $\mathtt{a}$ for $\overset{k}{\sim}$ if $\mathtt{a} \in I_k$;
- $\emptyset$ otherwise.

Before giving the construction of a *canonical antecedent* (in next subsection) we introduce the identifiability property and show it is a necessary condition for the existence of an $(\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})$-antecedent.

Definition 11. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$. We say that $[X, (\mathrm{L}, \mathrm{O})]$ verifies the identifiability property if for all $\mathtt{a}, \mathtt{b} \in \mathcal{B}$ with $\mathtt{a} \neq \mathtt{b}$ there exists an integer $k \in \{-\min\{\mathrm{O}(\mathtt{a}), \mathrm{O}(\mathtt{b})\}, \ldots, \min\{\mathrm{L}(\mathtt{a}) - \mathrm{O}(\mathtt{a}), \mathrm{L}(\mathtt{b}) - \mathrm{O}(\mathtt{b})\} - 1\}$ such that $\mathtt{a} \notin [\mathtt{b}]_k$.

Theorem 12. *Let $\mathcal{V}$ be a variable block map over the alphabet $\mathcal{A}$, S a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ and T the $\mathcal{V}$-coding of S. Then $[T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]$ verifies the identifiability property.*

Proof. Assume $[T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]$ does not verify the identifiability property. That is, there are two symbols $\mathtt{a}$ and $\mathtt{b}$ in $\mathcal{V}$ with $\mathtt{a} \neq \mathtt{b}$ and $\mathtt{a} \overset{k}{\sim} \mathtt{b}$ for all
$k \in \{-\min\{\mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{O}_{\mathcal{V}}(\mathtt{b})\}, \ldots, \min\{\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})\} - 1\}$. Then, by part (2) of Lemma 6,

$$\mathrm{w}_{\mathcal{V}}(\mathtt{a})_{\mathrm{O}_{\mathcal{V}}(\mathtt{a})+k} = \mathrm{w}_{\mathcal{V}}(\mathtt{b})_{\mathrm{O}_{\mathcal{V}}(\mathtt{b})+k}$$

for all $k \in \{-\min\{\mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{O}_{\mathcal{V}}(\mathtt{b})\}, \ldots, \min\{\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})\} - 1\}$. This contradicts the fact that $\mathcal{V}$ is a variable block map. □

4.1. Construction

Definition 13. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$, (L, O) be a length–origin pair over $\mathcal{B}$ and $k, l \in \mathbb{Z}$. The *transition function* $\mathrm{F}_{k,l}$ is defined for all $D \subseteq \mathcal{B}$ by

$$\mathrm{F}_{k,l}(D) = \{\mathtt{b} \in \mathcal{B} | \exists \mathtt{a} \in D \text{ with } \mathtt{a} \overset{k+\mathrm{O}(\mathtt{a}),l+\mathrm{O}(\mathtt{b})}{\smile} \mathtt{b}\}.$$

If $D = \{\mathtt{a}\}$ one writes $\mathrm{F}_{k,l}(\mathtt{a})$ instead of $\mathrm{F}_{k,l}(D)$.

Lemma 14. *Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$, (L, O) be a length–origin pair over $\mathcal{B}$, $k, l, m \in \mathbb{Z}$, $D \subseteq \mathcal{B}$ and $\mathtt{a}, \mathtt{b} \in \mathcal{B}$. The following properties hold*:
(1) $\mathtt{b} \in \mathrm{F}_{k,l}(\mathtt{a})$ *if and only if* $\mathtt{a} \in \mathrm{F}_{l,k}(\mathtt{b})$;
(2) *if* $\mathrm{F}_{k,l}(D) \neq \emptyset$ *then* $\mathrm{F}_{l,m}(\mathrm{F}_{k,l}(D)) \subseteq \mathrm{F}_{k,m}(D)$;

(3) *if* $\mathtt{b} \in \mathrm{F}_{k,l}(\mathtt{a})$ *then* $\mathrm{F}_{k,l}(\mathtt{a}) = [\mathtt{b}]_l$;
(4) $\mathrm{F}_{k,l}(\mathtt{a}) = \mathrm{F}_{k,l}([\mathtt{a}]_k)$.

Proof. (1) and (2) are straightforward.

(3) Let $\mathtt{b}, \mathtt{b}' \in \mathrm{F}_{k,l}(\mathtt{a})$, then $\mathtt{b} \overset{\mathrm{O}(\mathtt{b})+l,\mathrm{O}(a)+k}{\smile} \mathtt{a}$ and $\mathtt{b}' \overset{\mathrm{O}(\mathtt{b}')+l,\mathrm{O}(a)+k}{\smile} \mathtt{a}$, which implies $\mathtt{b} \overset{\mathrm{O}(\mathtt{b})+l,\mathrm{O}(\mathtt{b}')+l}{\smile} \mathtt{b}'$ and thus $\mathtt{b} \overset{l}{\sim} \mathtt{b}'$. One concludes that $\mathrm{F}_{k,l}(\mathtt{a})$ is contained in $[\mathtt{b}]_l$. Let $\mathtt{c} \overset{l}{\sim} \mathtt{b}$. Since $\mathtt{b} \overset{\mathrm{O}(b)+l,\mathrm{O}(a)+k}{\smile} \mathtt{a}$ one gets $\mathtt{c} \overset{\mathrm{O}(c)+l,\mathrm{O}(a)+k}{\smile} \mathtt{a}$ and thus $\mathtt{c} \in \mathrm{F}_{k,l}(\mathtt{a})$. This implies $\mathrm{F}_{k,l}(\mathtt{a}) = [\mathtt{b}]_l$.

(4) Clearly, $\mathrm{F}_{k,l}(\mathtt{a}) \subseteq \mathrm{F}_{k,l}([\mathtt{a}]_k)$. Let $\mathtt{b} \in \mathrm{F}_{k,l}([\mathtt{a}]_k)$. Then there is $\mathtt{a}' \in [\mathtt{a}]_k$ such that $\mathtt{b} \overset{\mathrm{O}_{\mathcal{V}}(\mathtt{b})+l,\mathrm{O}_{\mathcal{V}}(\mathtt{a}')+k}{\smile} \mathtt{a}'$. Since $\mathtt{a} \overset{k}{\sim} \mathtt{a}'$, one concludes that $\mathtt{b} \overset{\mathrm{O}_{\mathcal{V}}(\mathtt{b})+l,\mathrm{O}_{\mathcal{V}}(\mathtt{a})+k}{\smile} \mathtt{a}$, which implies the desired result. □

Definition 15. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$. The set $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]} \subseteq \mathcal{P}(\mathcal{B})^{\mathbb{Z}}$ is defined in the following way, where $\mathcal{P}(\mathcal{B})$ is the set of subsets of $\mathcal{B}$. An element $\tilde{\mathtt{u}} \in \mathcal{P}(\mathcal{B})^{\mathbb{Z}}$ belongs to $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ if:
- there is $k \in \mathbb{Z}$ such that $\tilde{\mathtt{u}}_k \neq \emptyset$;
- for all $k \in \mathbb{Z}$ if $\tilde{\mathtt{u}}_k \neq \emptyset$ there is a symbol $\mathtt{a} \in \mathcal{B}$ such that $\tilde{\mathtt{u}}_k = [\mathtt{a}]_k$;
- for all $k, l \in \mathbb{Z}$, $\tilde{\mathtt{u}}_l = \mathrm{F}_{k,l}(\tilde{\mathtt{u}}_k)$.

Since $\mathcal{B}$ is finite and for all $\mathtt{a} \in \mathcal{B}$, $[\mathtt{a}]_k$ is empty for all but a finite set of integers, then $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ is finite.

Let us notice that conditions in Definition 15 are consistent. In fact, assume that $\tilde{\mathtt{u}}_k = [\mathtt{a}]_k$ and $\tilde{\mathtt{u}}_l = \mathrm{F}_{k,l}(\tilde{\mathtt{u}}_k) \neq \emptyset$, then by Lemma 14, parts (3) and (4), there is $\mathtt{b} \in \mathcal{B}$ such that $\tilde{\mathtt{u}}_l = [\mathtt{b}]_l$.

Lemma 16. *Let* $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$, (L, O) *be a length–origin pair over* $\mathcal{B}$, $k \in \mathbb{Z}$ *and* $\mathtt{a} \in \mathcal{B}$. *At most one of the following properties holds*:
- *there is a unique element* $\tilde{\mathtt{v}} \in \mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ *such that* $\mathtt{a} \in \tilde{\mathtt{v}}_k$ *if* $-\mathrm{O}(\mathtt{a}) \leqslant k < \mathrm{L}(\mathtt{a}) - \mathrm{O}(\mathtt{a})$;
- *there is no element* $\tilde{\mathtt{v}} \in \mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ *such that* $\mathtt{a} \in \tilde{\mathtt{v}}_k$.

Proof. Let $\mathtt{a} \in \mathcal{B}$ such that $-\mathrm{O}(\mathtt{a}) \leqslant k < \mathrm{L}(\mathtt{a}) - \mathrm{O}(\mathtt{a})$ (i.e. $\mathtt{a} \in I_k$). One has that $[\mathtt{a}]_k \neq \emptyset$. Define $\tilde{\mathtt{v}} \in \mathcal{P}(\mathcal{B})^{\mathbb{Z}}$ by $\tilde{\mathtt{v}}_k = [\mathtt{a}]_k$ and $\tilde{\mathtt{v}}_l = \mathrm{F}_{k,l}([\mathtt{a}]_k)$ for all $l \in \mathbb{Z} \setminus \{k\}$. Then $\tilde{\mathtt{v}}_l$ is either empty or equal to $[\mathtt{b}]_l$ for some $\mathtt{b} \in \mathcal{B}$ and $\tilde{\mathtt{v}}$ satisfies all the conditions of Definition 15. Thus, for all $\mathtt{a} \in I_k$ there is an element $\tilde{\mathtt{v}} \in \mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ such that $\tilde{\mathtt{v}}_k = [\mathtt{a}]_k$. This element is the unique one in $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ granting this equality because of the third condition of Definition 15. □

Lemma 16 allows us to define for all symbols $\mathtt{a} \in \mathcal{B}$ and all $k \in \{-\mathrm{O}(\mathtt{a}), \ldots, \mathrm{L}(\mathtt{a}) - \mathrm{O}(\mathtt{a}) - 1\}$ the unique element $\tilde{\mathtt{v}}$ of $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ such that $\mathtt{a} \in \tilde{\mathtt{v}}_k$ (equivalently $[\mathtt{a}]_k = \tilde{\mathtt{v}}_k$). We call this element $\phi_k(\mathtt{a})$. The following lemma is straightforward.

Lemma 17. *Let* $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$, (L, O) *be a length–origin pair over* $\mathcal{B}$, $k \in \mathbb{Z}$ *and* $\mathtt{a}, \mathtt{b} \in \mathcal{B}$. *If* $\mathtt{b} \in \mathrm{F}_{k,l}(\mathtt{a})$ *then* $\phi_k(\mathtt{a}) = \phi_l(\mathtt{b})$.

To each $\mathtt{a} \in \mathcal{B}$ one associates the pointed word over $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$

$$p(\mathtt{a}) = (\phi_{-\mathrm{O}(\mathtt{a})}(\mathtt{a}) \ldots \phi_{\mathrm{L}(\mathtt{a})-\mathrm{O}(\mathtt{a})-1}(\mathtt{a}), \mathrm{O}(\mathtt{a})).$$

We denote by $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]} = \{p(\mathtt{a}) | \mathtt{a} \in \mathcal{B}\}$. If $[X, (\mathrm{L}, \mathrm{O})]$ verifies the identifiability property then $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$ is a variable block map and p is a bijection from $\mathcal{B}$ to $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$. We denote by $\theta_{[X,(\mathrm{L},\mathrm{O})]} = p^{-1} : \mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]} \to \mathcal{B}$.

Definition 18. Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$ such that $[X, (\mathrm{L}, \mathrm{O})]$ is boundary-standard and verifies the identifiability property. The (L, O)-antecedent of X, denoted by $A([X, (\mathrm{L}, \mathrm{O})])$, is equal to $\bigcup_{v \in X} a_{[X,(\mathrm{L},\mathrm{O})]}(v)$, where $a_{[X,(\mathrm{L},\mathrm{O})]}(v)$ is defined for all $v \in X$ by:
- $a_{[X,(\mathrm{L},\mathrm{O})]}(v) = \phi_0(v) = (\phi_0(v_i))_{i \in \mathbb{Z}}$ if $v \in \mathcal{B}^{\mathbb{Z}}$;
- $a_{[X,(\mathrm{L},\mathrm{O})]}(v) = \phi_{-\mathrm{O}(v_0)}(v_0) \ldots \phi_{-1}(v_0)\phi_0(v)$ if $v \in \mathcal{B}^{\mathbb{N}}$;
- $a_{[X,(\mathrm{L},\mathrm{O})]}(v) = \phi_{-\mathrm{O}(v_0)}(v_0) \ldots \phi_{-1}(v_0)\phi_0(v)\phi_1(v_{|v|-1}) \ldots \phi_{\mathrm{L}(v_{|v|-1})-\mathrm{O}(v_{|v|-1})-1}(v_{|v|-1})$ if $v \in \mathcal{B}^+$.

Lemma 19. *Let $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$ such that $[X, (\mathrm{L}, \mathrm{O})]$ is boundary-standard and verifies the identifiability property. The $(\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}, \theta_{[X,(\mathrm{L},\mathrm{O})]})$-coding of $A([X, (\mathrm{L}, \mathrm{O})])$ is X.*

Proof. Let $v \in X \bigcap \mathcal{B}^{\mathbb{Z}}$. Consider v' to be the $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$-coding of $u = a_{[X,(\mathrm{L},\mathrm{O})]}(v)$. For $j \in \mathbb{Z}$, v'_j is the unique element (w, o) of $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$ such that $w = u_{[j-o, j+|w|-o-1]} = \phi_0(v_{[j-o, j+|w|-o-1]})$. For all $j - \mathrm{O}_{\mathcal{V}}(\mathtt{a}) \leqslant l < j + \mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a})$ one has that $v_l \in \Delta^X_{l-j}(v_j)$, which implies that $v_l \in \mathrm{F}_{l-j,0}(v_j)$ and thus by Lemma 17 $\phi_0(v_l) = \phi_{l-j}(v_k)$. Thus, $v'_j = (\phi_{-\mathrm{O}(v_j)}(v_j) \dots \phi_{\mathrm{L}(v_j)-\mathrm{O}(v_j)-1}(v_j), \mathrm{O}(v_j))$ and $\theta_{[X,(\mathrm{L},\mathrm{O})]}(v'_j) = v_j$.

The same property holds if $v \in X \bigcap (\mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^+)$. One just has to remark that if $[X, (\mathrm{L}, \mathrm{O})]$ is boundary-standard then $a_{[X,(\mathrm{L},\mathrm{O})]}(v)$ is $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$-admissible and the first symbol occurring in its $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$-coding is v_0 and the last one, when it is finite, is $v_{|v|-1}$. □

Now we are in conditions to state the main result of the section.

Theorem 20. *Let $\mathcal{B}$ be a finite alphabet, $X \subseteq \mathcal{B}^+ \bigcup \mathcal{B}^{\mathbb{N}} \bigcup \mathcal{B}^{\mathbb{Z}}$ and (L, O) be a length–origin pair over $\mathcal{B}$ such that $[X, (\mathrm{L}, \mathrm{O})]$ is boundary-standard. The following properties are equivalent*:

(1) *There exist a finite alphabet $\mathcal{A}$, a variable block map $\mathcal{V}$ over $\mathcal{A}$, a $\mathcal{V}$-compatible subset Y of $\mathcal{A}^{\mathbb{Z}}$ and a bijection $\theta : \mathcal{V} \to \mathcal{B}$ verifying $\mathrm{L} = \mathrm{L}_{\mathcal{V}} \circ \theta^{-1}$, $\mathrm{O} = \mathrm{O}_{\mathcal{V}} \circ \theta^{-1}$ and $\theta^{-1}(X)$ is the $\mathcal{V}$-coding of Y.*

(2) *$[X, (\mathrm{L}, \mathrm{O})]$ verifies the identifiability property.*

Proof. (1) ⇒ (2) Follows from Theorem 12.

(2) ⇒ (1) If condition (2) holds, then the construction in this subsection shows how to build a new alphabet $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$, a variable block map $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$, a bijection $\theta_{[X,(\mathrm{L},\mathrm{O})]}$ from $\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}$ to $\mathcal{B}$ and a set of sequences $A([X, (\mathrm{L}, \mathrm{O})])$ over $\mathcal{N}_{[X,(\mathrm{L},\mathrm{O})]}$ such that X is the $(\mathcal{V}_{[X,(\mathrm{L},\mathrm{O})]}, \theta_{[X,(\mathrm{L},\mathrm{O})]})$-coding of $A([X, (\mathrm{L}, \mathrm{O})])$ (Lemma 19). □

4.2. Maximal recoding

In this subsection we analyze a canonical property of the (L, O)-antecedent constructed in previous subsection.

Theorem 21. *Let $\mathcal{V}$ be a variable block map over the alphabet $\mathcal{A}$, S a $\mathcal{V}$-compatible subset of $\mathcal{A}^+ \bigcup \mathcal{A}^{\mathbb{N}} \bigcup \mathcal{A}^{\mathbb{Z}}$ and T the $\mathcal{V}$-coding of S. There is a surjective map $\psi : \mathcal{N}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]} \to \mathcal{A}$ such that $S = \psi(A([T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]))$, where in the last equality ψ is applied coordinatewise.*

Proof. From Remark 2 $[T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]$ is boundary-standard and from Theorem 12 it verifies the identifiability property. Then elements described in previous subsection: $\mathcal{N}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]}$, $\mathcal{V}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]}$ and $A([T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})])$, are well defined.

Consider relations $(\overset{i,j}{\smile}: \ i, j \in \mathbb{N})$ associated to $[T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]$. Let $\mathtt{a}, \mathtt{b} \in \mathcal{V}$ and $k \in \{-\min\{\mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{O}_{\mathcal{V}}(\mathtt{b})\}, \dots, \min\{\mathrm{L}_{\mathcal{V}}(\mathtt{a}) - \mathrm{O}_{\mathcal{V}}(\mathtt{a}), \mathrm{L}_{\mathcal{V}}(\mathtt{b}) - \mathrm{O}_{\mathcal{V}}(\mathtt{b})\} - 1\}$.

From Lemma 6, one has that $\mathtt{a} \overset{k}{\sim} \mathtt{b}$ implies $\mathrm{w}_{\mathcal{V}}(\mathtt{a})_{k+\mathrm{O}_{\mathcal{V}}(\mathtt{a})} = \mathrm{w}_{\mathcal{V}}(\mathtt{b})_{k+\mathrm{O}_{\mathcal{V}}(\mathtt{b})}$. Now, from the definition of $\mathcal{V}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]}$ one gets

$$\mathtt{a} \overset{k}{\sim} \mathtt{b} \ \Leftrightarrow \ \phi_k(a) = \phi_k(b) \ \Leftrightarrow \ \mathrm{w}_{\mathcal{V}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]}}(p(\mathtt{a}))_{k+\mathrm{O}_{\mathcal{V}}(\mathtt{a})} = \mathrm{w}_{\mathcal{V}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]}}(p(\mathtt{b}))_{k+\mathrm{O}_{\mathcal{V}}(\mathtt{b})}.$$

Thus, there exists a surjective map $\psi : \mathcal{N}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]} \to \mathcal{A}$ such that for all elements $\mathtt{a} \in \mathcal{V}$ and all $0 \leqslant i < \mathrm{L}_{\mathcal{V}}(\mathtt{a})$ one has $\mathrm{w}_{\mathcal{V}}(\mathtt{a})_i = \psi(\mathrm{w}_{\mathcal{V}_{[T,(\mathrm{L}_{\mathcal{V}},\mathrm{O}_{\mathcal{V}})]}}(p(\mathtt{a}))_i)$. One concludes that $S = \psi(A([T, (\mathrm{L}_{\mathcal{V}}, \mathrm{O}_{\mathcal{V}})]))$. □

5. Length subshifts

In this section we provide another point of view to understand the complexity of the coding–decoding process induced by a given variable block map $\mathcal{V}$ over an alphabet $\mathcal{A}$. To simplify the discussion we assume that $\mathcal{A}^{\mathbb{Z}}$ only contains $\mathcal{V}$-compatible sequences.

In $\mathcal{A}^{\mathbb{Z}}$ one can define two natural block maps: the coding operation $\Phi_{\mathcal{V}}$, defined previously, and the block map associated with the length function $\mathrm{L}_{\mathcal{V}}$. To be more precise, put $\mathcal{L} = \{\mathrm{L}_{\mathcal{V}}(w) : (w, o) \in \mathcal{V}\}$ and define $\ell_{\mathcal{V}} : \mathcal{A}^{\mathbb{Z}} \to \mathcal{L}^{\mathbb{Z}}$ by $(\ell_C(x))_i = \mathrm{L}_{\mathcal{V}}(w_i(x))$ where $(w_i(x), o_i(x))$ is the unique element of $\mathcal{V}$ such that $w_i(x)$ appears in x centered at position i. The length subshift associated to $\mathcal{V}$ is $X_{\mathrm{L}_{\mathcal{V}}} = \ell_{\mathcal{V}}(\mathcal{A}^{\mathbb{Z}})$. It is clear that $X_{\mathrm{L}_{\mathcal{V}}}$ is a mixing sofic subshift of $\mathcal{L}^{\mathbb{Z}}$ (it is the image of a mixing subshift by a block map). Moreover, it has a *receptive fixed* point $\mathbf{u} = (u)_{i\in\mathbb{Z}}$. That is, there exist synchronized words w_1 and w_2 in $L(X_{\mathrm{L}_{\mathcal{V}}})$ and a positive integer i_0 such that for all $i \geqslant i_0$ one has $w_1 u^i w_2 \in L(X_{\mathrm{L}_{\mathcal{V}}})$. Such kind of fixed point plays an important role in the construction of factor maps between sofic subshifts in symbolic dynamics (for more details see [2]). In the following theorem we prove that any sofic subshift verifying such conditions can be the length subshift associated to a variable block map. This illustrates the complexity of the decoding process given a variable block map. One needs the following result that is a direct consequence of Theorem 3.3 of [2].

Lemma 22. *Let $X \subseteq \mathcal{D}^{\mathbb{Z}}$ be a mixing sofic subshift having a receptive fixed point. Then there exists a block map $F : \mathcal{D}^{\mathbb{Z}} \to \mathcal{D}^{\mathbb{Z}}$ such that $F(\mathcal{D}^{\mathbb{Z}}) = X$.*

Theorem 23. *Let $X \subseteq \mathcal{D}^{\mathbb{Z}}$ be a mixing sofic subshift having a receptive fixed point. There exists a variable block map $\mathcal{V}$ over $\mathcal{D}$ such that $X_{\mathrm{L}_{\mathcal{V}}}$ and X are conjugate by a letter-to-letter block map.*

Proof. We can assume $\mathcal{D} = \{0, \ldots, |\mathcal{D}| - 1\}$. By hypothesis, from Lemma 22, there exists a block map $F : \mathcal{D}^{\mathbb{Z}} \to \mathcal{D}^{\mathbb{Z}}$ such that $F(\mathcal{D}^{\mathbb{Z}}) = X$. Moreover, we can assume there is $f : \mathcal{D}^r \to \mathcal{D}$ such that $F(x)_i = f(x_i \ldots x_{i+r-1})$ for $x \in \mathcal{D}^{\mathbb{Z}}$ and $i \in \mathbb{Z}$. Let $\mathcal{V} = \{(uv, 0) \ : \ u \in \mathcal{D}^r, \ v \in \mathcal{D}^{f(u)+1}\}$. It is straightforward that $\mathcal{V}$ is a variable block map over $\mathcal{D}$. In fact, it is a prefix code. Observe that for any $(w, o) \in \mathcal{V}$ one has $\mathrm{L}_{\mathcal{V}}(w) = r + f(u) + 1$ where $w = uv$ with $u \in \mathcal{D}^r$ and $v \in \mathcal{D}^{f(u)+1}$. Therefore $X_{\mathrm{L}_{\mathcal{V}}} \subseteq \mathcal{A} = \{r + 1, \ldots, r + |\mathcal{D}|\}$ is conjugate with X up to the identification of a letter $d \in \mathcal{D}$ with $r + d + 1 \in \mathcal{A}$. □

Acknowledgments

The second author thanks the CMM-CNRS who made this collaboration possible. The support and hospitality of both institutions are very much appreciated.
The first and third authors acknowledge financial support from Programa Iniciativa Científica Milenio P01-005 and P04-069-F. Both authors thank the anonymous referee for helpful comments and suggestions.

References

[1] J. Berstel, D. Perrin, Theory of Codes, Academic Press, New York, 1985.
[2] M. Boyle, Lower entropy factors of sofic systems, Ergodic Theory Dynamical Systems 3 (4) (1983) 541–557.
[3] G. Didier, Caractérisation des *N*-écritures et application à l'étude des suites de complexité ultimement $n + cte$, Theoret. Comput. Sci. 215 (1999) 31–49.
[4] G. Didier, I. Laprevotte, M. Pupin, A. Hénaut, Local decoding of sequences and alignment-free comparison, J. Comput. Biol., to appear.
[5] I. Laprevotte, M. Pupin, E. Coward, G. Didier, C. Terzian, C. Devauchelle, A. Hénaut, HIV-1 and HIV-2 LTR nucleotide sequences: assessment of the alignment by *N*-block presentation, "retroviral signatures" of overrepeated oligonucleotides, and a probable important role of scrambled stepwise duplications/deletions in molecular evolution, Mol. Biol. Evol. 18 (7) (2001) 1231–1245.
[6] D. Lind, B. Marcus, An Introduction to Symbolic Dynamics and Coding, Cambridge University Press, Cambridge, 1995.