Entropy analysis of substitutive sequences revisited

To cite this article: K Karamanos 2001 *J. Phys. A: Math. Gen.* **34** 9231

J. Phys. A: Math. Gen. **34** (2001) 9231–9241 PII: S0305-4470(01)17972-2

Entropy analysis of substitutive sequences revisited

K Karamanos

Centre for Nonlinear Phenomena and Complex Systems, Université Libre de Bruxelles, CP 231, Campus Plaine, Boulevard du Triomphe, B-1050, Brussels, Belgium

E-mail: kkaramam@ulb.ac.be

Received 3 October 2000, in final form 17 July 2001
Published 19 October 2001
Online at stacks.iop.org/JPhysA/34/9231

Abstract

A given finite sequence of letters over a finite alphabet can always be algorithmically generated, in particular by a Turing machine. This fact is at the heart of complexity theory in the sense of Kolmogorov and Chaitin. A relevant question in this context is whether, given a statistically 'sufficiently long' sequence, there exists a deterministic finite automaton that generates it. In this paper we propose a simple criterion, based on measuring block entropies by lumping, which is satisfied by all automatic sequences. On the basis of this, one can determine that a given sequence is not automatic and obtain interesting information when the sequence is automatic. Following previous work on the Feigenbaum sequence, we give a necessary entropy-based condition valid for all automatic sequences read by lumping. Applications of these ideas to representative examples are discussed. In particular, we establish new entropic decimation schemes for the Thue–Morse, the Rudin–Shapiro and the paperfolding sequences read by lumping.

PACS numbers: 02.10.−v, 02.30.Lt, 05.45.−a, 05.70.−a, 65.20.+w

1. Introduction

Nature provides us with a wide variety of symbolic strings ranging from the sequences generated by the symbolic dynamics of nonlinear systems to the RNA and DNA sequences or the DLA patterns [4, 23, 27].

Entropy-like quantities are a very useful tool for the analysis of such sequences. Of special interest are the *block entropies*, extending Shannon's classical definition of the entropy of a single state to the entropy of a succession of states [23]. In particular, it has been shown that the *scaling* of the block entropies with length sometimes gives interesting information on the structure of the sequence [13, 14].

In [16], we have shown that the estimation of the block entropies actually depends on the way of reading, that is, on the *observer*. This has an immediate bearing on the 'decoding' procedure, as different values of the block entropies mean different kinds and amounts of information extracted by the symbolic sequence. By using *lumping*, we have established a new decimation scheme for the symbolic dynamics of the Feigenbaum attractors of unimodal maps [12, 22]. The coarse-grained statistical properties of the attractors have been subsequently derived, with emphasis on the behaviour of the block entropies.

Lumping is the reading of the symbolic sequence by 'taking portions' (see equation (1)), as opposed to *gliding* where one has essentially a 'moving frame'. Note that gliding is the standard convention in the literature. Reading the symbolic sequence in a specific way is also called *decimation* of the sequence.

The importance of the distinction between *gliding* and *lumping* codes in genetics has already been recognized long ago in [10], see also [20] (called at that time *overlapping* and *non-overlapping* codes). As mentioned in [16], the realization that the kind and amount of information of a given symbolic sequence may depend on the way that reading brings symbolic dynamics closer to natural languages, in which the existence of distinct privileged words conveying a precise 'meaning' is crucial. Moreover, in [6] it has been shown that the estimation of the (conditional) block entropies with the usual prescription of gliding cannot help us to distinguish between sequences with different spectral properties and different levels of complexity.

A similar situation arises in the quite different context of supramolecular chemistry, where certain inorganic molecules become capable of pattern recognition [19]. The inorganic skeleton of the macromolecule is then 'read' by the molecules of its environment due to stereochemical interactions. This type of 'reading' corresponds essentially to 'lumping' as dealt with in the present paper.

The important question which arises in the light of these results is whether one can invent some criteria which could illuminate the structure of a symbolic sequence and give us some more specific information beyond that afforded by block entropies computed by gliding. The objective of the present paper is to derive an entropy criterion for the particular, yet quite important property of *automaticity* of the sequence.

We recall that a sequence is called *automatic* if it is the image of a letter-to-letter projection of the fixed point of a set of substitutions of constant length. A substitution is called *uniform* or *of constant length* if all the images of the letters have the same length. The term 'automatic' comes from the fact that an automatic sequence is generated by a finite automaton. For instance, the Feigenbaum symbolic sequence can in an equivalent manner be generated by the Metropolis–Stein–Stein algorithm [16, 22], or as the fixed point $(\sigma^F)^\infty(R)$ of the set of substitutions of length 2: $\sigma^F(R) = RL, \sigma^F(L) = RR$ starting with R, or by the finite automaton of figure 1. For more details about automatic sequences the reader is referred to [9] and for their role in physics to [1].

In this paper we show how the procedure of reading the symbolic sequences by lumping is useful and helps us to decimate some important automatic sequences from the mathematical literature in a different way. In section 2 we present our main results in the form of two propositions. In section 3 we re-examine some automatic sequences from this standpoint and in section 4 we investigate the automaticity of some other substitutive sequences by lumping and with an example from biology. In section 5 we draw the principal conclusions.

2. Entropy analysis by lumping

Consider a subsequence of length N selected out of a very long (theoretically infinite) symbolic sequence. We stipulate that this subsequence is to be read in terms of distinct 'blocks' of length n,

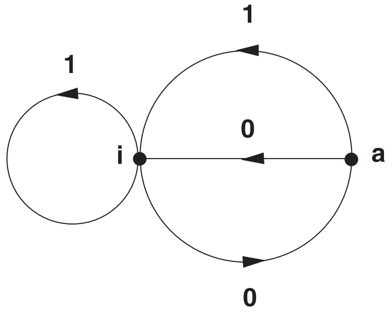

Figure 1. Deterministic finite automaton predicted by Cobham's algorithmic procedure. This automaton contains two states: i and a and the function of exit F corresponds to each state by a symbol; either $F(i) = R = 1$ or $F(a) = L = 0$. To calculate the nth term of the 2^∞ sequence we first express the number n in its binary form and then we start running the automaton from its initial state, according to the binary digits of n. In this trip we read the symbols contained in the binary expansion of n from the left to the right following the targets indicated by the letters. For instance, $n = 3 = (11_2)$ gives the run $i \to i \to i$ so that $u(3) = R = 1$, while $n = 9 = (1001_2)$ gives the run $i \to i \to a \to i \to i$ so that $u(9) = R = 1$.

$$\ldots \underbrace{A_1 \ldots A_n}_{B_1} \underbrace{A_{n+1} \ldots A_{2n}}_{B_2} \ldots \underbrace{A_{jn+1} \ldots A_{(j+1)n}}_{B_{j+1}} \ldots \tag{1}$$

We call this reading procedure *lumping*. We shall follow lumping in this paper.

The following quantities characterize the information content of the sequence [13, 18]:

(i) The dynamical (Shannon-like) block entropy for blocks of length n

$$H(n) = - \sum_{(A_1,\ldots,A_n)} p^{(n)}(A_1, \ldots, A_n) \cdot \ln p^{(n)}(A_1, \ldots, A_n) \tag{2}$$

where the probability of occurrence of a block $A_1, \ldots, A_n$, denoted by $p^{(n)}(A_1, \ldots, A_n)$, is defined (when it exists) in the statistical limit as

$$p^{(n)}(A_1, \ldots, A_n) = \frac{\text{No of blocks of the form } A_1, \ldots, A_n \text{ encountered when lumping}}{\text{total No of blocks encountered when lumping}} \tag{3}$$

starting from the beginning of the sequence and the associated entropy per letter

$$h^{(n)} = \frac{H(n)}{n}. \tag{4}$$

(ii) The conditional entropy or entropy excess associated with the addition of a symbol to the right of an n-block

$$h_{(n)} = H(n+1) - H(n). \tag{5}$$

(iii) The entropy of the source (a topological invariant), defined as the limit (if it exists)

$$h = \lim_{n\to\infty} h_{(n)} = \lim_{n\to\infty} h^{(n)} \tag{6}$$

which is the discrete analogue of the metric or Kolmogorov entropy.

We now turn to the selection problem, that is, to the possibility of the emergence of some preferred configurations (blocks) out of the complete set of different possibilities. The number of all possible symbolic sequences of length n (complexions in the sense of Boltzmann) in a K-letter alphabet is

$$N_K = K^n. \tag{7}$$

Yet not all of these configurations are necessarily realized by the dynamics, nor are they equiprobable. A remarkable theorem by McMillan [18] gives a partial answer to the selection problem asserting that for stationary and ergodic sources the probability of occurrence of a block $(A_1, \ldots, A_n)$ is

$$p^{(n)}(A_1, \ldots, A_n) \sim \mathrm{e}^{-H(n)} \tag{8}$$

for almost all blocks $(A_1, \ldots, A_n)$. In order to determine the abundance of long blocks one is thus led to examine the scaling properties of $H(n)$ as a function of n.

We are now in a position to state our main result, see also [17]. Let m^k be the length of a block encountered when lumping and $H(m^k)$ the associated block entropy. The following property then holds.

Proposition 1. *If the symbolic sequence $(u_n)_{n\in\mathcal{N}}$ is m-automatic, then*

$$\exists k_o \in \{0, 1\} \quad m \in \mathcal{N}^* \quad \forall \quad k \geqslant k_o : \quad H(m^{k_o}) = H(m^k) \tag{9}$$

when lumpingstarts from the beginning of the sequence.

Proof. Suppose that the infinite sequence $(u_n)_{n\in\mathcal{N}}$ taking values from a finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ is m-automatic. Then, according to theorem 3 of [9], the sequence $(u_n)_{n\in\mathcal{N}}$ can in an equivalent manner be generated as the fixed point of a set of substitutions of length m, plus a letter-to-letter projection. (The role of the automaton is not essential in the sequel.) Let us call this substitution σ and write down explicitly

$$\sigma(a_1) = b_{11}b_{12}\cdots b_{1m} \qquad \sigma(a_2) = b_{21}b_{22}\cdots b_{2m} \qquad \cdots \qquad \sigma(a_n) = b_{n1}b_{n2}\cdots b_{nm}$$

where all *b's* belong to the alphabet. Automaticity will then entail, for instance, that

$$u_n = p(\sigma^\infty(a_1))$$

where p is a letter-to-letter projection and

$$\begin{aligned} p(\sigma(a_1)) &= p(b_{11}b_{12}\cdots b_{1m}) = p(b_{11})p(b_{12})\cdots p(b_{1m}) = c_{11}c_{12}\cdots c_{1m} \\ p(\sigma(a_2)) &= p(b_{21}b_{22}\cdots b_{2m}) = p(b_{21})p(b_{22})\cdots p(b_{2m}) = c_{21}c_{22}\cdots c_{2m} \\ &\cdots \\ p(\sigma(a_n)) &= p(b_{n1}b_{n2}\cdots b_{nm}) = p(b_{n1})p(b_{n2})\cdots p(b_{nm}) = c_{n1}c_{n2}\cdots c_{nm}. \end{aligned}$$

It is now a theorem that the only blocks appearing in the expression of $\sigma^\infty(a_1)$ when lumping by blocks of length m, are the blocks $b_{11}b_{12}\cdots b_{1m}, b_{21}b_{22}\cdots b_{2m}, \cdots b_{n1}b_{n2}\cdots b_{nm}$. Furthermore, one can act on $\sigma^\infty(a_1)$ by σ itself

$$\sigma^\infty(a_1) = \sigma(\sigma^\infty(a_1))$$

which leaves $\sigma^\infty(a_1)$ invariant.

In the same way, the only blocks appearing in the expression of u_n when lumping by blocks of length m, are the blocks $c_{11}c_{12}\cdots c_{1m}, c_{21}c_{22}\cdots c_{2m}, \cdots c_{n1}c_{n2}\cdots c_{nm}$.

Furthermore, one can write

$$u_n = p(\sigma^\infty(a_1)) = p(\sigma(\sigma^\infty(a_1))).$$

Simply by counting the blocks when first lumping on $p(\sigma^\infty(a_1))$ by blocks of length m and then on $p(\sigma(\sigma^\infty(a_1)))$ by blocks of length m^2, (note that it is the same sequence), we find that

$$H(m) = H(m^2).$$

Following inductively the same argument and observing the expansion factor m each time that σ appears, this last relation implies in a straightforward manner that

$$\forall k: \quad H(m) = H(m^2) = \cdots = H(m^k).$$

In particular, we also have

$$H(1) = H(m)$$

when p is a bijection and this completes the proof. □

The meaning of proposition 1 is that for m-automatic sequences there is always an envelope in the diagram $H(n)/n$ versus n, falling off exponentially as $\sim m^{-k}$ for blocks of length m^k, $k = 1, 2, \ldots$. For infinite ergodic strings, the conclusion does not depend on the starting point. Similar conclusions hold if instead of a one-to-one letter projection we have a one-to-many letters projection of constant length. In particular, we have the following result.

Proposition 2. *If the symbolic sequence $(u_n)_{n\in\mathcal{N}}$ is the image of the fixed point of a set of substitutions of length m by a projection of constant length μ, then*

$$\exists \quad k_o \in \{0, 1\} \quad m \in \mathcal{N}^* \quad \forall \quad k \geqslant k_o: \quad H(\mu \cdot m^{k_o}) = H(\mu \cdot m^k) \tag{10}$$

when lumping starts from the beginning of the sequence.

Typical examples of this kind of sequences are the $\mu \cdot 2^\infty$ Feigenbaum sequences [16]. As a concrete example one can indeed consider the $3 \cdot 2^\infty$ Feigenbaum sequence, which is generated by the 2^∞ Feigenbaum sequence with $\sigma^F(R) = RL$, $\sigma^F(L) = RR$ starting with R, after the projection of constant length $\mu = 3$, $p_3(R) = RLL$, $p_3(L) = RLR$, so that

$$u_n^3 = p_3(\sigma^{F\infty}(R)) = RLLRLRRLLRLLRLLRLRRLLRL\ldots.$$

For this sequence we have shown in [16] that the following decimation scheme holds for the entropies calculated by lumping:

$$H(3 \cdot 2r) = H(3 \cdot r).$$

Other examples (with $\mu = 1$) are the Rudin–Shapiro and the paperfolding sequences which are studied below.

3. Substitutions of constant length

To illustrate the propositions derived in section 2, we shall first re-examine some substitutions common in the mathematical literature using lumping. Note that the conditional block entropies or entropy excess using gliding have been analysed exhaustively and computed for every n for the Thue–Morse, the Rudin–Shapiro and the paperfolding sequences in [6], see also [2, 7].

Note also that the Thue–Morse sequence has a continuous singular spectrum, the Rudin–Shapiro sequence has Lebesgue spectrum and the paperfolding sequence has discrete spectrum [25, 26].

3.1. Decimation of the Thue–Morse sequence

The Thue–Morse sequence is defined as the fixed point (i.e. the infinite iteration $(\sigma^T)^\infty(0)$) of the substitution σ^T, defined on the alphabet $\{0, 1\}$ by

$$\sigma^T(0) = 01 \qquad \sigma^T(1) = 10. \tag{11}$$

Its first terms are

$$0110100110010110\ldots$$

This sequence appears in many contexts ranging from combinatorics to chess. For a survey see [3].

We introduce the *decimation operator* $\hat{M}$, whose action amounts on the replacements

$$\hat{M}(01) = 0 \qquad \hat{M}(10) = 1 \tag{12}$$

when lumping starts from the beginning of the sequence (essentially the inverse of the substitution σ^T, defined in (11)).

It is then evident that

$$\hat{M}^n((\sigma^T)^\infty(0)) = (\sigma^T)^\infty(0) \tag{13}$$

which implies the following invariance property of the block entropies

$$H^T(2^k) = H^T(2) = H^T(1) = \ln 2 \tag{14}$$

in accordance with proposition 1.

Simply by counting words and observing the reduction factor 2 due to (12), equation (13) enables us to introduce the following additional entropic decimation scheme for subsequences of the full (infinite) Thue–Morse sequence:

$$H^T(2 \cdot r) = H^T(r). \tag{15}$$

3.2. Decimation of the Rudin—Shapiro sequence

The Rudin–Shapiro sequence is the image of the fixed point $(\sigma^R)^\infty(a)$ of the substitution σ^R

$$\sigma^R(a) = ab \qquad \sigma^R(b) = ac \qquad \sigma^R(c) = db \qquad \sigma^R(d) = dc \tag{16}$$

by the projection

$$p^R(a) = p^R(b) = 0 \qquad p^R(c) = p^R(d) = 1. \tag{17}$$

Its first terms are

$$0001001000011101\ldots$$

Following the same procedure as above, one sees that the following invariance property holds:

$$H^R(2^k) = H^R(2) = 2\ln 2. \tag{18}$$

Unlike the case of the Feigenbaum and the Thue–Morse sequences, we cannot introduce a stronger entropic decimation scheme for the Rudin–Shapiro sequence as in (15). The reason is that when we count the blocks of an odd length, many different blocks of the $(\sigma^R)^\infty(a)$ sequence may correspond to the same block of the $p^R((\sigma^R)^\infty(a))$ Rudin–Shapiro sequence. For instance, the blocks *cab* and *dba* which one encounters when reading the $(\sigma^R)^\infty(a)$ by lumping, are both projected to the block 100 of the Rudin–Shapiro sequence and the blocks *aca* and *bdb* are projected to the block 010.

3.3. Decimation of the paperfolding sequence

The paperfolding sequence is the image of the fixed point $(\sigma^P)^\infty(a)$ of the substitution σ^P

$$\sigma^P(a) = ab \qquad \sigma^P(b) = cb \qquad \sigma^P(c) = ad \qquad \sigma^P(d) = cd \tag{19}$$

by the projection

$$p^P(a) = p^P(b) = 1 \qquad p^P(c) = p^P(d) = 0. \tag{20}$$

Its first terms are

$$1101100111001001\ldots$$

We now have the following invariance property:

$$H^P(2^k) = H^P(2) = 2\ln 2. \qquad (21)$$

Again, unlike the case of the Feigenbaum and the Thue–Morse sequences, we cannot establish a stronger entropic decimation scheme for the paperfolding sequence. The reason here is that when we count the blocks of an odd length, many different blocks of the $(\sigma^P)^\infty(a)$ sequence may correspond to the same block of the $p^P((\sigma^P)^\infty(a))$ paperfolding sequence. For instance, the blocks *abc* and *bad* which one encounters when reading the $(\sigma^P)^\infty(a)$ by lumping, are both projected to the block 110 of the paperfolding sequence, and the blocks *bcd* and *adc* are both projected to the block 100 of the paperfolding sequence.

3.4. Discussion of the results

Comparing equations (14), (18) and (21) we conclude that the new decimation scheme allows one to establish some new ordering relations between block entropies, reflecting differences in the complexity of the sequences, contrary to the use of gliding. In particular, if we compare the invariant block entropies for blocks of length 2^k we observe that

$$H^T(2^k) < H^R(2^k) = H^P(2^k). \qquad (22)$$

Retrospectively, this could be expected, because the self-similar tree of the Thue–Morse sequence is certainly simpler than the corresponding structures of the Rudin–Shapiro and the paperfolding sequences. Moreover, the self-similar tree of the Thue–Morse sequence does not contain any projection at the end.

4. Further examples: substitutions of variable length

4.1. A first example

Consider first the fixed point of the substitution

$$\sigma^a(0) = 012 \qquad \sigma^a(1) = 1212 \qquad \sigma^a(2) = 00. \qquad (23)$$

Application of the entropy analysis by lumping shows numerically that

$$H^a(3) = H^a(9) = H^a(27) \qquad (24)$$

indicating the possibility that the sequence is isomorphic to a substitution of constant length 3. A careful examination of the blocks appearing and of their frequencies of occurrence, shows that as 1 is always followed by 2, the sequence can, in an equivalent manner, be considered as the fixed point of the substitution

$$\sigma^b(0) = 012 \qquad \sigma^b(1) = 121 \qquad \sigma^b(2) = 200 \qquad (25)$$

so that it is 3-automatic.

4.2. A second example

Consider now the fixed point of the substitution

$$\sigma^c(0) = 12 \qquad \sigma^c(1) = 102 \qquad \sigma^c(2) = 0 \qquad (26)$$

which in view of the theory developed in [11], is isomorphic to a substitution of constant length, although this is not evident at first sight.

Application of the entropy analysis by lumping shows numerically that

$$H^c(2) = H^c(4) = H^c(8) = H^c(16) = H^c(32) \qquad (27)$$

indicating the possibility that the sequence is isomorphic to a substitution of constant length 2.

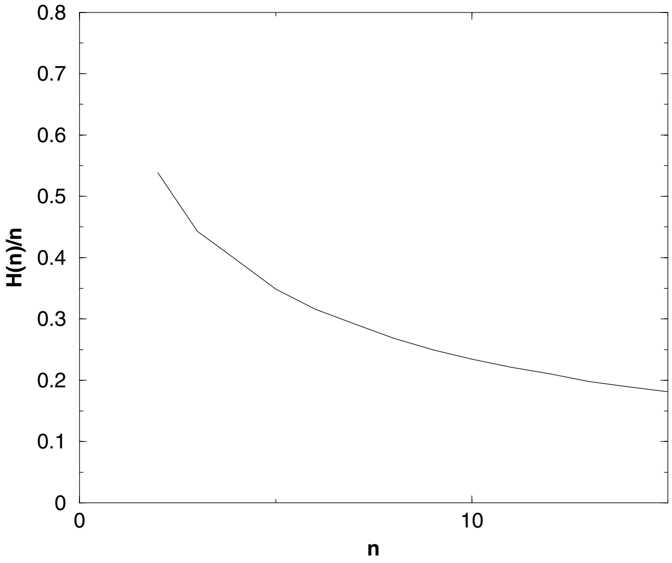

Figure 2. Entropy per letter $h^{(n)}$ as a function of n, measured by lumping, obtained numerically by the first 2584 terms of the Fibonacci sequence. The same result is obtained with gliding. We observe a monotonic decay, which is the signal of non-automaticity in view of proposition 1.

A careful examination of the blocks appearing and of their frequencies of occurrence, shows that the sequence can, in an equivalent manner, be considered as the fixed point of the substitution

$$\sigma^d(a) = ab \qquad \sigma^d(b) = ca \qquad \sigma^d(c) = cd \qquad \sigma^d(d) = ac \tag{28}$$

projected by

$$h^d(a) = 10 \qquad h^d(b) = 21 \qquad h^d(c) = 20 \qquad h^d(d) = 12. \tag{29}$$

Thus application of entropy analysis by lumping reveals the hidden structure of the sequence.

4.3. Decimation of the Fibonacci sequence

An example of a sequence which is substitutive but does not satisfy the entropy condition (9) (so that it is not automatic), is the Fibonacci sequence, defined as the fixed point of the substitution

$$\sigma^{Fi}(0) = 01 \qquad \sigma^{Fi}(1) = 0. \tag{30}$$

For this sequence we have found numerically that the entropies calculated by lumping are equal to the corresponding entropies calculated by gliding, which are known [5], see also figure 2.

4.4. Decimation of the Chacon sequence

Another interesting substitutive sequence has been introduced by Chacon in [8], see also [15], by juxtaposition of blocks B_n with the following rule

$$\begin{aligned} B_0 &= 0 \\ B_{n+1} &= B_n B_n 1 B_n. \end{aligned} \tag{31}$$

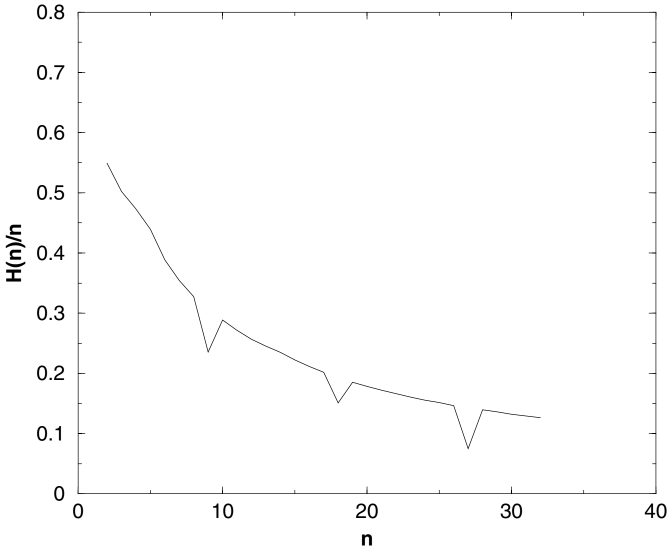

Figure 3. Entropy per letter $h^{(n)}$ as a function of n, measured by lumping, obtained numerically by the first 9840 terms of the Chacon sequence. We observe a non-monotonic decay for values of n being in arithmetic progression of 9, $n = 9k, k \geqslant 2$.

The resulting sequence reading

$$0010001010010\ldots$$

is invariant under the replacements $0 \to 0010$, $1 \to 1$ and it is a typical example of a system which is *weakly mixing* but not *strongly mixing*. It has thus a special place in the ergodic hierarchy.

Application of entropy analysis by lumping to this sequence shows numerically that the block entropies present a non-monotonic behaviour for $n = 9k, k \geqslant 2$, that is, for an arithmetic progression of 9 (see figure 3). This quite surprising fact is probably due to the special property of the Chacon sequence to have a letter 1 in all positions $n = 3^k, k \geqslant 1$. One can indeed show that the unity following the two B_n's in the definition of the Chacon sequence is in position 3^k from the beginning of the sequence.

Proof. For $k = 1$, it holds. Let us suppose that it holds for $k = t$. Then for $k = t + 1$, the unity following the two B_t's is in position

$$2\left(3^t + \tfrac{1}{2}(3^t - 1)\right) + 1 = 3^{t+1}. \qquad \square$$

This might be a good starting point for further studies of the entropic behaviour of this sequence.

5. Perspectives

Finally, the analysis of DNA and RNA sequences has attracted considerable interest these last years; see, for instance, [21]. To show how our diagnostics are applied to such cases, we have considered a part of the complete genome of the virus *lambda fage* (a coding sequence) and of the *human beta globin region of chromosome 11* (a non-coding sequence with coding only

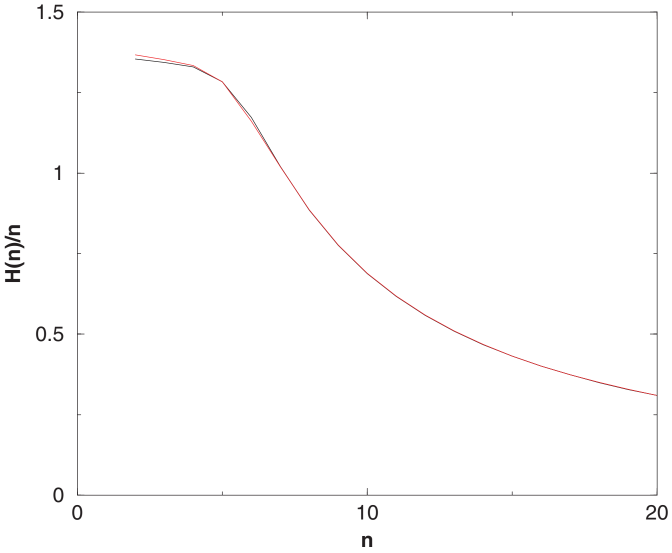

Figure 4. Entropy per letter $h^{(n)}$ as a function of n, measured by lumping, obtained numerically by the first 9780 basis proteins of the virus *lambda fage* (upper curve), and the *human beta globin region of chromosome 11* (lower curve). We observe a monotonic decay, which is the signal of non-automaticity in view of proposition 1. Note that a sequence of the same length, which is a part of one of the above-considered automatic sequences, gives non-monotonic behaviour.

3%). Figure 4 depicts the result of preliminary investigations, which strongly suggests that there is no small length automaticity in both these sequences.

Another interesting perspective is opened by the use of different *automaticity measures*, as in [24, 28, 29]. These measures could characterize the 'distance from being automatic', in some sense, and they will soon be the subject of further studies.

6. Conclusions

In this paper we derived a new *diagnostic* for automaticity. When one disposes of an unknown symbolic sequence and applies the entropy analysis by lumping, then if the sequence does not obey the invariance property predicted by the propositions of section 2, it is certainly non-automatic. Conversely, if one observes the adequate invariance property, then the sequence is a candidate to be automatic, or to be the image of the fixed point of a set of substitutions of constant length by a projection of constant length.

We have analysed the block entropies of some well-known automatic sequences from this standpoint and found that, under the convention that the sequence is to be read in terms of hypersymbols, new relations show up and the entropies satisfy some well-defined invariance properties.

Although the results are relatively straightforward to obtain, we believe that they deserve attention because of their very broad range of applicability, from theoretical information science to telecommunications and biology.

To the author's knowledge, the question of a functional relation between the block entropies when gliding and when lumping for an arbitrary sequence has not yet been addressed in the literature. A plausible conjecture supported by numerical work could be that the block entropies calculated by gliding form an *upper bound* for the block entropies calculated by lumping. It is also interesting to mention that for the special case of the Feigenbaum sequence, we have calculated exhaustively in [16] the block entropies by lumping and related them to the block entropies by gliding.

Acknowledgments

I am indebted to J-P Allouche and G Nicolis for useful discussions and support. I also thank V Berthé for interesting discussions, V Basios for collaboration and for the preparation of a special software for the implementation of some of the ideas put forward in this paper, as well as I Almirantis for making available biological data to me. I also like thank G Didier for useful exchange of software. The author has been a NATO Fellow for the last three academic years and is currently a Van Buuren Foundation Fellow and a Petsalys-Lepage Foundation Fellow. This work has been supported in part by the Pôles d'Attraction Interuniversitaires program of the Belgian Federal Office of Scientific, Technical and Cultural Affairs.

References

[1] Allouche J P 1987 *Pour la Science* **114** 94

[2] Allouche J P 1991 *Bull. Belg. Math. Soc.* **1** 133

[3] Allouche J P and Shallit J 1999 The ubiquitous Prouhet–Thue–Morse sequence *Sequences and their applications Proc. SETA'98* ed C Ding, T Helleseth and H Niederreiter (Berlin: Springer) pp 1–16

[4] Bai-Lin H 1994 *Chaos* (Singapore: World Scientific)

[5] Berthé V 1995 *Beyond Quasicrystals* (Les Houches: Les Éditions de Physique, Springer) p 441 and references therein

[6] Berthé V 1994 *J. Phys. A: Math. Gen.* **27** 7993 and references therein

[7] Cassaigne J 1997 *Bull. Belg. Math. Soc.* **4** 67

[8] Chacon R V 1969 *Proc. Am. Math. Soc.* **22** 559

[9] Cobham A 1972 *Math. Syst. Theory* **6** 164

[10] Crick F H C, Barnett L, Brenner S and Watts-Tobin R J 1961 *Nature* **192** 1227

[11] Dekking F M 1978 *Zeit. Wahr.* **41** 221

[12] Derrida B, Gervois A and Pomeau Y 1978 *Ann. Inst. Henri Poincaré , Section A: Physique Théorique* **29** 305

[13] Ebeling W and Nicolis G 1991 *Europhys. Lett.* **14** 191

[14] Ebeling W and Nicolis G 1992 *Chaos Solitons Fractals* **2** 635

[15] Ferenczi S 1995 *Bull. S.M.F.* **123**

[16] Karamanos K and Nicolis G 1999 *Chaos Solitons Fractals* **10** 1135

[17] Karamanos K 2001 AIP Conference Proceedings **573** p 278

[18] Khinchin A I 1957 *Mathematical Foundations of Information Theory* (New York: Dover)

[19] Lehn J M 1995 *Supramolecular Chemistry* (Weinheim: VCH)

[20] Lewin B 1997 *Genes VI* (New York: Oxford University Press)

[21] Mantegna R N, Buldyrev S V, Goldberger A L, Havlin S, Peng C K, Simons M and Stanley H E 1995 *Phys. Rev.* E **52** 2939

[22] Metropolis N, Stein M L and Stein P R 1973 *J. Comb. Theory* A **15** 25

[23] Nicolis G and Gaspard P 1994 *Chaos Solitons Fractals* **4** 41

[24] Pomerance C, Robson J M and Shallit J 1997 *Theor. Comp. Sci.* **180** 181

[25] Queffélec M 1987 *Substitution Dynamical Systems. Spectral Analysis (Lecture Notes in Mathematics)* vol 1294 (Berlin: Springer)

[26] Queffélec M 1995 *Beyond Quasicrystals* (Les Houches: Les Éditions de Physique, Springer) vol 3 p 369 and references therein

[27] Schröder M 1991 *Fractals, Chaos, Power Laws* (New York: Freeman)

[28] Shallit J 1996 *Journal de Théorie des Nombres de Bordeaux* **8** 347

[29] Shallit J and Breitbart Y 1996 *J. Comput. Syst. Sci.* **53** 10