Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

Statistical compressibility analysis of DNA sequences by generalized entropy-like quantities: Towards algorithmic laws for Biology ?

K. Karamanos
Centre for Nonlinear Phenomena and Complex Systems,Université Libre de Bruxelles, Campus Plaine, C.P. 231, Boulevard du Triomphe, B-1050, Brussels

I. Kotsireas
Department of Physics and Computing, Wilfrid Laurier University, 75 University Avenue, Waterloo, Ontario N2L 3C5, Canada

A. Peratzakis and K. Eftaxias
Department of Physics, Section of Solid State Physics, University of Athens, Panepistimiopolis, GR 15784 Zografos, Athens, Greece

A detailed entropy analysis by the recent novelty of 'lumping' is performed in some DNA sequences. On the basis of this, we first report here a negative answer to the question 'can the DNA sequences at the level of nucleotides be generated by a deterministic finite automaton of essentially a small number of states, in the statistical limit?'. What is observed in all cases is an almost linear scaling of the block entropies - up to the numerical precision - close to the one of a mixing ergodic system with a very high topological entropy. The basic result that we report here is that the all the examined biological sequences appear to be very little compressible (they lie near to the incompressible limit). The topological entropy of coding regions appears to be even higher than that of non-coding regions.

I. 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 [1, 14, 15].

Recently, there has been a growing interest in unraveling the mysteries of DNA strings and a significant number of works has been reported [19– 41]. DNA strings are written in an alphabet of four letters $\{a, c, t, g\}$ where a stands for *adenine,* c for *cytosine,* g for *guanine,* and t for *thymine.* a and g are *purines,* and c and t *purimidines*.

As has already been observed in [50, 51],
'...The expected main product of the Human Genome Project is a large collection of 10.000 to 1.000.000 bases long sequences of human DNA fragments, characterized originally only to one or another hybridization marker. The expected first user of this product is the sequence analyst for whom the characterization of the sequences, i.e. identification and mapping of biologically meaningfull and/or structurally distinct sequence regions, will be the main task.

A significant part of the sequence characterization problem is the abundance of heretofore functionally 'silent' and structurally ill-defined intervening sequences and spacers. Together with dispersed and tandem repeats which are structurally recognizable but functionally silent as well, this makes over 90 % of the human genome currently meaningless. These sequences, as nonsensical as they appear, also have to be located and described in order to clearly distinguish them from those scarce sequences that are believed to be ubderstood.

One could think of several levels of the sequence characterization, from preliminary screening for few major features all the way down to search for numerous particular functional sites. To begin with, it would de desirable to simply tell whether this or another sequence fragment is promising at all. For example, even an apparently featureless sequence would seem to be *a priori* more interesting or at least more complex than, say, a tandemly repeating sequence motif. Thus it would be desirable to be able to somehow evaluate the complexity of the sequences...'

One of the earlier, if not the earliest, and most inspiring work on the possible relation of nucleotide sequences and aspects of their undrelying dynamics, termed at that time as "dynamic linguistics", was investigated by J.S. Nicolis and co-workers in the end of 80's and summarized in his book [41]. In this work the transinfornation (or *mutual information*) was chosen as the indicator of " ...the 'bulk' information about the 'grammar' and 'syntax' of the 'language'..." (c.f [41]) for symbolic strings of biological origin, namely the sequences of RNA-polymerase-III and human embryonic cDNA.

Except from the transinformation, 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 successsion of states [14]. In particular, it has been shown in the literature that the *scaling* of the block entropies with length gives sometimes

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

interesting information on the structure of the sequence [12, 13].

In particular, we have derived an entropy criterion for the specific, yet quite important algorithmic property of automaticity of a sequence. We recall that, a sequence is called *automatic* if it is generated by a finite automaton (The lowest level Turing machine.) For more details about automatic sequences the reader is referred to [4], and for their role in Physics to [5].

Our criterion is based on the entropy analysis by *lumping*. *Lumping* is the reading of the symbolic sequence by 'taking portions' (see eq.(1)), as opposed to *gliding* where one has essentially a 'moving frame'. Notice that gliding is the standard convention in the literature. Reading the symbolic sequence in a specific way is also called *decimation* of the sequence.

An important aspect of recent research activities about DNA strings concerns the idendification of the *coding* and *non-coding* regions in the chromosome. The coding regions encode for the production of a given protein, while the non-coding regions do not. In the light of this research many interesting algorithms and methods - often heuristic - have been developed. Noawadays, in the basis of these algorithms and of experimental data there is a general agreement for a given chromosome region to be considered as coding region or not. Still, as already pointed out, most of these results are heuristic and have been invented for this particular situation (they are not of general purpose in mathematical physics).

In this framework our group have proposed a systematic series of numerical experimentation aiming at revealing some algorithmic aspects of genomic sequences. Some of these results are presented here. Note that entropy analysis of some chromosomes by different techniques have been reported in [34, 36].

A second interesting point in this context is that, according to different independent methods, coding regions appear to have higher topological entropy than non-coding regions. There are some exceptions to that rule, reported at [34].

In the present paper we shall re-examine from an entropic point of view by lumping some chromosomes which are already known to be (mostly) coding or non-coding.

This treatement has the additional merit to pose the problem of possible automaticity and other algorithmic properties in a firm and well-understood mathematical basis, although there always is the serious drawback of *patchiness* [33] means, of important fluctuations around the mean statistical values. To apply the standard arsenal of Non-equilibrium Statistical Mechanics one have to suppose stationarity, and this is what it has been done explicitly or implicitly in the statistical treatments of the problem [34–37]. There are however some works in the literature where non-stationarity is postulated and checked in an *a posteriori* basis. Nevertheless, the present length of the analysed chromosomes guarantees - at least numerical - stability.

The basic conclusions in the literature are retrieved from our entropy analysis by lumping. None of the examined sequences appears to be automatic. Block entropies follow a monotonic decay. The scaling of the block entropies appears to be (almost) linear, close to the one of a (strongly) mixing system in the regime of developed chaos.

Moreover, the Kolmogorov-Sinai (K-S) entropy seems to be larger for viruses that for human beings. Thic could be a first signal that differences at the level of functionality and structure (or even of anatomy) are reflected at the chromosome level of organisation of life. Also, the K-S entropy seems to be larger for coding chromosomes than for non-coding ones, thus supporting a similar, previous conclusion existing at the literature [34, 35].

Concerning the compressibility issue now, one could ask oneself if the amount of information stored in the present important biological data bases, is possible to be replaced by some other means, occupying less place in computer memory, as for instance, deterministic or stochastic automata or some clearcut 'syntactic rules'. In the last case, the big data bases could be replaced by smallest ones. This is a question of obvious practical importance not only for the biologist but also for the bioinformatician. As we shall see, we will give a negative answer to this question.

The paper is articulated as follows. In Sec 2 we present the mathematical formulation of the entropy analysis by lumping. In Sec 3 we present the example of an automatic sequence, taken from the world on Nonlinear Science, namely the Feigenbaum sequence. In Sec 4 we present the computational platform, and in Sec 5 our main results concerning DNA sequences. Finally, in Sec 6 we draw the main conclusions and discuss future work.

II. ENTROPY ANALYSIS BY LUMPING

For reasons of completeness and for later use, we compile here the basic points of the method of the 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,

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

We call this reading procedure *lumping*. We shall follow lumping in the sequel.

The following quantities characterize the information content of the sequence [11, 12]

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

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

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

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

$$\frac{No.\ of\ blocks,\ A_1...A_n,\ encountered\ when\ lumping}{total\ No.\ of\ blocks} \tag{3}$$

starting from the beginning of the sequence, and the associate 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 analog of metric or Kolmogorov entropy.

We now turn to the selection problem, that is to the possibility of 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 due to McMillan [11], gives a partial answer to the selection problem asserting that for stationary and ergodic sources the probability of occurrence of a block $(A_1, ..., A_n)$ is

$$p_n(A_1, ..., A_n) \sim e^{-H(n)} \tag{8}$$

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

Another advandage of considering the quantity of block entropy – apart from its faster search requirements and easier implementation – is that it is equiped by its very construction with a quite powerful diagnostic for the onset of finite size effects. It is well known that block entropy is underestimated where even in some cases (gaussian assumptions of errors in estimation) specific formulas have been derived for correction, [39].

This underestimation of $H(m)$ for large values of m is due to the simple fact that not all words will be represented adequately if one loocks long enough samples. The situation becomes more and more prominent for calculating $H(m)$ by 'lumping' instead of 'gliding'. Indeed in the case of 'lumping' an exponentially fast decaying tail towards value zero follows after the plateau.

Since the probabilities of the words of length m are calculated by their frequencies, i.e. $p_n = |B_n| / N_{[sample]}$ where $N_{[sample]}$ is the size of the avalable data-sample i.e. the length of the 'text' under considereation in 'letters' then as $B_n \to 0$ for long words (since $m \to N_{[sample]}$) the block entropy calculated will reach a maximum value, its plateau, at

$$H_{MAX} = \log_{[b]}(N_{[sample]})$$

where b the lenght of the alphabet. In our case, of course, $b = 4$, see for example Fig. 4. This way, the value of H_{MAX} can determine a safe border for finite size effects.

After this small digression, we recall here the main result of the entropy analysis by lumping, see also [17, 18]. Let m^k be the length of a block encountered when lumping, $H(m^k)$ the associated block entropy. The following property then holds.

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

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

  when lumping, starting from the beginning of the sequence.

The meaning of the previous proposition 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 a length m^k, $k = 1, 2, ...$ 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.

- 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

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

  when lumping, starting from the beginning of the sequence.

Our propositions give an interesting *diagnostic* for automaticity. When one disposes of an unknown symbolic sequence and applies numerically the entropy analysis by lumping, then if the sequence does not obey such an

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

invariance property predicted by the propositions, it is certainly non-automatic. In the opposite, if one observes the adequate invariance property, then the sequence is a candidate to be automatic.

III. THE EXAMPLE OF THE FEIGENBAUM SEQUENCE

Before proceeding to the analysis of DNA strings (which as we shall see presently turn out not to be automatic) we first give an example of the entropy analysis by lumping of a 2-automatic sequence: the period-doubling or Feigenbaum sequence, much studied in the literature [8, 13, 16].

The Feigenbaum symbolic sequence can in an equivalent manner be generated by the Metropolis, Stein and Stein algorithm [6, 16], 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 Fig. 1.

According to our first proposition, for this sequence it holds

$$H(1) = H(2) = ... = H(2^k) \tag{11}$$

when lumping, and for any integer r,

$$H(2 \cdot r) = H(r) \tag{12}$$

as we have shown in [16]. The diagram $H(n)/n$ versus n for this sequence is shown in Fig. 2.

Thus, the Feigenbaum sequence appears to be extremely compressible from the viewpoint of algorithmic information theory, as one memorizing the finite automaton of Fig. 1 (instead of memorizing the full sequence), can reproduce every term ans so, the complete sequence. We say that the information carried by the Feigenbaum sequence is 'algorithmically compressible'.

The period-doubling sequence, is the only for which an exact functional relation between the block-entropies when *lumping* and when *gliding* exists in the literature, so that it is an instructive example.

IV. THE COMPUTATIONAL PLATFORM

In the following Section we shall present the results of some medium-scale computations on DNA sequences. The computations have been performed with a new, user friendly Maple package called Tools for Symbolic Dynamics.

This package is intended to provide a set of tools to facilitate the analysis of the information content of symbolic sequences in order to

- test theoretical predictions up to a very high degree of accuracy,
- perform efficiently different kinds of analyses to unkown sequences with the aim to discover potential patterns
- produce graphics to help interpret the results.

The initial version of Tools for Symbolic Dynamics has been written in Maple 5. Ports for both Maple 6 and Maple 7 are currently avilable. The package may be used with both the text and the graphical inerfaces of Maple.

The package Tools for Symbolic Dynamics has been tested in the analysis of various *substitutive* sequences (sequences generated by a set of substitution rules for each letter, over a finite alphabet) and the correcteness of the relevant routines has been verified on large-scale computations [19]. Using Tools for Symbolic Dynamics we can perform entropy computations by gliding or lumping on known or unknown symbolic sequences over a finite alphabet with millions of terms, thanks to a suitable search algorithm developed by one of us (IK), thus bypassing partially the well-known problem of *combinatorial explosion* [13].

In the next Sec we shall present the results obtained with the help of Tools for Symbolic Dynamics on real DNA sequences.

V. RESULTS

Essentially we have examined the DNA sequences shown in the following Table.

Real name	% of coding	Length
lambda virus	99	48502
human beta-globin region	3	73326
human vitamine D gene	3	55136
adenovirus AD2	78	35937
human interferone	6	5961
human heparan HSPG2 gene	92	7272

where for *lambda virus* and for the *adenovirus AD2* we have considered the *complete genome*.

For the first sequence (labda virus) the first few block entropies (where we have good enough statistics) are listed below

Block length n	$H(n)$
1	1.39
2	2.76
3	4.12
4	5.47
5	6.77

For the other sequences from the table, we have found similar results.

Based on the numerical regression, we have found that the block entropies by lumping fit very well with a linear regression of the form

$$H(n) = H_o + nh \tag{13}$$

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

where, according to the theory developed in [14], H_o expresses the statistical error around the regression and should be almost equal to zero, and h is the K-S entropy of the process. Moreover, H_o can be related to the effects of 'bad statistics'. In fact, it turn out to be larger for (very) small sequences.

After $n > 4$, we do not have enough statistical precision for the entropies and we do not include the points for $n > 4$ in the regression. The effects of bad statistics are further explained in Fig.3.

The parameters of the regression H_o and h are presented in the table below.

Real name	h	% h_{max}	H_o
lambda virus	1.361	0.982	0.032
human beta-globin region	1.325	0.956	0.042
human vitamine D gene	1.314	0.948	0.040
adenovirus AD2	1.360	0.981	0.026
human interferone	1.292	0.932	0.072
human heparan HSPG2 gene	1.329	0.958	0.071

In fact the ratio $h/h_{max} = h/\ln 4$, could be used in our case as an 'automaticity measure', expressing somehow the distance from being automatic.

For the m-automaticity one could indeed consider the ratio

$$d = max(k) \quad \| \frac{H(m^k) - H(m^{k-1})}{H_{max}(m^k)} \| \tag{14}$$

of the heights of the 'deviations from automatic behavior' (in view of our first proposition) as an % measure of non m-automaticity.

In fact, for linear scaling of the block entropies this leads just to

$$d = \frac{h}{h_{max}} = \frac{h}{\ln 4} \tag{15}$$

in our case. This indicator of complexity, coincides formally in our case with the 'compression coefficient' defined in Information Theory [11]. The meaninng of the compression coefficient is that 'it measures in a natural way the compression of the given uncoded text by the given means of coding...'.

In the light of this discussion, the basic conclusion arising from the above Table is that the information carried by this biological strings is very little compressible. This is also in agreement with the work of the Boston [34, 35] and the Berlin groups [36], where for the human chromosome 22 the compressibility in the region $n < 5$ is around 97.5 %.

To summarize, from this Table, we crystalize the following conclusions.

- None of the six above examined sequences presents a small automaticity.
- The entropy per letter $h^{(n)}$ follow a monotonic decay, clearly indicating the mixing character of the underlying dynamics. No traces of non-monotonicity of the $h^{(n)}$ can be found, even for very large n.
- All DNA sequences examined, present a large K-S entropy h_{KS}, indicating that life is working in very large 'distance from automaticity'.
- Viruses present in general largest h_{KS} entropy, possibly indicating that differences in the funtionality and the structure of living organisms are reflected in the entropies of their DNA and conversely.
- Coding regions present in general largest h_{KS} entropies than the non-coding ones.

We have also done a calculation of the block entropies inside a coding region of the lambda fage from 191 to 735.

Block length n	$H(n)$
1	1.4
2	2.7
3	3.7
4	4.5
5	4.6

This results are similar to that of the Berlin group [36], where they observe first an 'almost random' region which after $n > 5$ becomes clearly differentiated from randomness and presents a sublinear scaling. (In our case, we do not have enough precision (due to the small length of the sequence) to observe this region.)

In no case we observe an exponential envelope predicted by the propositions about automaticity. (In any case the positivity of the K-S entropy alone, suffices to guarantee the non existence of a deterministic finite automaton wich generates the sequence, as determinism implies that the K-S entropy is zero.)

VI. CONCLUSIONS AND OUTLOOK

In this work we have performed an analysis of some DNA sequences by lumping. The basic novelty of this method is that, unlike the Fourier transform or the conventional entropy analysis by gliding, it gives results that can be related with algorithmic characteristics of the sequences and, in particular, with the property of automaticity.

The basic conclusions of this analysis is that the DNA sequences are not automatic and the block entropies present an almost linear scaling with the word length, characteristic of a mixing ergodic system. Moreover, non-coding regions have in general smaller h_{KS} entropies than coding ones, a fact that could indicate some 'local order' in these parts of the sequences. Once again we have avoided to concatenate text.

This conclusion comes in support of a conjecture given in [34, 35]. However, the techniques followed in the two

works are quite different. In [34] the authors have performed entropy analysis by gliding on sequences coming from concatenated coding or non-coding parts of real DNA sequences. In this work we have performed entropy analysis by lumping on some suitable pre-selected sequences which are known to be mostly coding or non-coding. Clearly, more numerical work is needed in order to support further or reject these conjectures.

Recently also, E. N. Trifonov [50, 51] has proposed his own theory of linguistic complexity and applied it to biologically relevant problems [52]. The link between this theory and the present calculations should be exploited. Yet another perspective is opened by a recent 'parity code interpretation of nucleotide alphabet composition' introduced in [54], where the nucleotide alphabet is considered as an error-correcting code.

About the compressibility issue now, it seems that life works in very big distance from being automatic. In particular, the biological sequences examined in this work appear to be very little compressible (they lie near to the incompressible limit). Thus it seems that, the present big biological data bases cannot be replaced in the near future by smallest data bases, except if new, strong, compressibility algorithms show up.

Acknowledgments. The authors would like to thank Profs. G. Nicolis, R. Thomas, Y. Almirantis, A. Provata and J.S. Nicolis, for helpful discussions. We also thank Prof. A. Carbone and M. Gromov (IHES) for useful critiscism. Financial support from the Van Buuren Foundation and a Petsalys-Lepage Foundation are gratefuly acknowledged. IK would like to thank the *Center of Nonlinear Phenomena and Complex Systems (CENOLI)* of ULB for the kind invitation to Brussels, as well as the *Ontario Research Centre for Computer Algebra (ORCCA)* for supporting this collaboration. The first author (KK) would like to thank IHES for a travel grant. The authors would also like to thank the *Center of Symbolic Calculation (CSO)* of the ULB, where a first version of the computational platform has been developed. 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.

[1] H. Bai-Lin, *Chaos,* World Scientific, Singapore (1994).
[2] V. Berthé, in *Beyond quasicrystals, Les Houches,* F. Axel and D. Gratias (Eds.), Les Éditions de Physique, Springer 441–463 (1995); and references therein.
[3] V. Berthé, *J. Phys. A: Math. Gen.* **27,** 7993–8006 (1994); and references therein.
[4] A. Cobham, *Uniform tag sequences, Math. Systems Theory* **6,** 164–192 (1972).
[5] J.-P. Allouche,*Pour la Science,* 94 (April 1987).
[6] N. Metropolis, M. L. Stein and P. R. Stein, *J. Comb. Th.* **A 15(1),** 25–44 (1973)
[7] B. Derrida, A. Gervois and Y. Pomeau, *Ann. Inst. Henri Poincaré, Section A: Physique Théorique* **XXIX(3),** 305–356 (1978).
[8] P. Grassberger, *Int. J. Theor. Phys.* **25(9),** 907 (1986).
[9] F.H.C. Crick, L. Barnett, S. Brenner and R. J. Watts-Tobin, *Nature* **192(4809)** 1227 (1961).
[10] J. M. Lehn,*Supramolecular chemistry,* VCH, Weinheim (1995).
[11] A. I. Khinchin,*Mathematical Foundations of Information Theory*, Dover, New York (1957).
[12] W. Ebeling and G. Nicolis, *Europhys. Lett.* **14(3),** 191–196 (1991)
[13] W. Ebeling and G. Nicolis, *Chaos, Solitons & Fractals* **2,** 635 (1992).
[14] Nicolis G., and Gaspard P., *Chaos, Solitons & Fractals* **4(1)**, 41 (1994).
[15] M. Schröder, *Fractals, Chaos, Power Laws* Freeman, New York (1991).
[16] K. Karamanos and G. Nicolis, *Symbolic dynamics and entropy analysis of Feigenbaum limit sets, Chaos, Solitons & Fractals* **10(7),** 1135–1150 (1999).
[17] K. Karamanos, *Entropy analysis of automatic sequences revisited: an entropy diagnostic for automaticity, Proceedings of Computing Anticipatory Systems 2000, CASYS2000, AIP Conference Proceedings* **573,***D. Dubois (Ed.),* 278–284 (2001).
[18] K. Karamanos, *Entropy analysis of substitutive sequences revisited,J. Phys. A: Math. Gen.* **34,** 9231–9241 (2001).
[19] K. Karamanos and I. Kotsireas, *Thorough numerical entropy analysis of some substitutive sequences by lumping, Kybernetes* **31(9/10),** 1409–1417 (2002).
[20] L. L. Gatlin, *The information content of DNA I, II, J. Theor. Biol.* **10,** 281–300 (1966).
[21] B. E. Blaisdell, *A prevalent persistent global nonrandomness that distinguishes coding and non-coding eukaryotic nuclear DNA sequences, J. Mol Evol.* **19,** 122–133 (1983).
[22] R. Staden, *Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes, Nucleic Acids Res.* **12,** 551–567 (1984).
[23] R. Staden, *Graphic methods to determine the function of nucleic acid sequences, Nucleic Acids Res.* **12,** 521–538 (1984).
[24] M. I. Granero-Porati and A. Porati, *Information parameters and randomness in mithochondrial DNA, J. Mol. Evol.* **27,** 109–113 (1988).
[25] R. F. Voss, *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, Phys. Rev. Lett.* **68,** 3805–3808 (1992).
[26] W. Li and K. Kaneko, *Long-range correlations and partial* $1/f^a$ *spectrum in a noncoding DNA sequence, Europhys. Lett.* **17,** 655–660 (1992).
[27] G. S. Mani, *Correlations between the coding and noncoding regions in DNA, J. Theor. Biol.* **158,** 429–445 (1992).
[28] G. S. Mani, *Long-range doublet correlations in DNA and*

the coding regions, J. Theor. Biol. **158,** 447–464 (1992).

[29] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, *Long-range correlations in nucleotide sequences, Nature (London)* **356,** 168–170 (1992).

[30] S. Nee, *Uncorrelated DNA walks, Nature (London)* **357,** 450 (1992).

[31] V. V. Prabhu and J. M.Claverie, *Correlations in intronless DNA, Nature (London)* **359,** 782 (1992).

[32] P. J. Munson and R. C. Taylor, *DNA correlations, Nature (London)* **360** 636 (1992).

[33] S. Karlin and V. Brendel, *Patchiness and correlations in DNA sequences, Science* **259,** 677–680 (1993).

[34] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons M. and H. E. Stanley, *Phys. Rev.* **E 52(3)**, 2939 (1995).

[35] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng and M. Simons *Scaling features of noncoding DNA, Physica A* **273,** 1–18 (1999).

[36] D. Holste, I. Grosse and H. Herzel, *Statistical analysis of the DNA sequence of human chromosome 22, Phys. Rev. E* **64,** 0419171–9 (2001).

[37] D. Holste, I. Grosse, S. Beirer, P. Schieg and H. Herzel, *Repetitive sequences and correlations in human DNA sequences,* preprint (2002).

[38] W. Ebeling, T. Pöschel and K.F. Albrecht, *Entropy, Transinformation and Word Distribution of Information - Carrying Sequences, Int. J. Bif. and Chaos* **5,** 51–61 (1995).

[39] H. Herzel and I. Grosse, *Measuring correlations in symbol sequences, Physica A* **216,** 518–542 (1995).

[40] T. Pöschel, W. Ebeling and H. Rosé, *Guessing Probability Distributions from small Samples, J. Stat. Phys.* **80,** 1443–1451 (1995).

[41] J. S. Nicolis, *Chaos and Information Processing,* Word Scientific, Singapore (1991).

[42] Y. Almirantis and A. Provata, *The "clustered structure" of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences, Bull. Math. Biol.* **59,** 975–992 (1997).

[43] A. Provata and Y. Almirantis, *Scaling properties of coding and non-coding DNA sequences, Physica A* **247,** 482–496 (1997).

[44] Y. Almirantis and A. Provata, *Long and short-range correlations in genome organisation, J. Stat Phys.* **97,** 233-262 (1999).

[45] A. Provata and Y. Almirantis, *Fractal Cantor patterns in the sequence structure of DNA, Fractals* **8,** 15–27 (2000).

[46] Y. Almirantis and A. Provata, *An evolutionary model for the origin of non-randomness, long-range order and fractality in the genome, BioEssays* **23(7),** 647–656 (2001).

[47] Z.-G. Yu and B. Wang, *A time series model of CDS sequences in complete genome, Chaos, Solitons & Fractals* **12,** 519–526 (2001).

[48] Z.-G. Yu and V. Anh, *Time series model based on global structure of complete genome, Chaos, Solitons & Fractals* **12,** 1827–1834 (2001).

[49] B. Lewin, *Genes VI,* Oxford University Press, New York (1997).

[50] E. N. Trifonov, *Making Sense of the Human Genome,Structure & Methods, Volume 1: Human Genome Initiative & DNA Recombination, Ramaswamy H. Sarma & Makti H. Sarma, (Eds.)* Adenine Press (1990).

[51] O. Popov, D. M. Segal and E. N. Trifonov, *Linguistic complexity of protein sequences as compared to texts of human languages, BioSystems* **38,** 65–74 (1996).

[52] A. Bolshoy, K. Shapiro, E. N. Trifonov and I. Ioshikhes, *Enhancement of the nucleosomal pattern in sequences of lower complexity, Nucl. Ac. Res.* **25(16),** 3248–3254 (1997).

[53] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa and C. Thermes, *Long range correlations between DNA bending sites: Relation to the structure and dynamics of Nucleosomes, J. Mol. Biol.* in press (2002).

[54] D. A. Mac Donaill, *A parity code interpretation of nucleotide alphabet composition, Chem. Comm.,* 2062–2063 (2002).

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

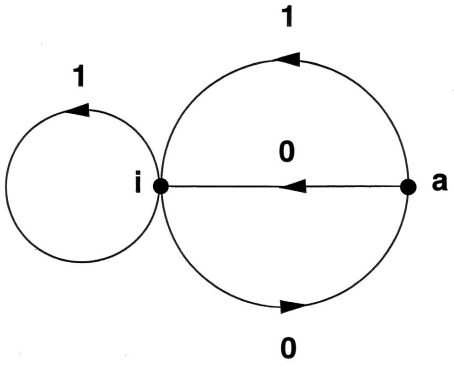

FIG. 1: Deterministic finite automaton predicted by Cobham's algorithmic procedure. This automaton contains two states: i and a and to each state corresponds by the function of exit F a symbol; either $F(i) = R = 1$ or $F(a) = L = 0$. To calculate the n^{th} term of the 2^{∞} sequence we first express 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 $n = 3 = (11$ base $2)$ gives the run $i \to i \to i$ so that $u(3) = R = 1$, while $n = 9 = (1001$ base $2)$ gives the run $i \to i \to a \to i \to i$ so that $u(9) = R = 1$.

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

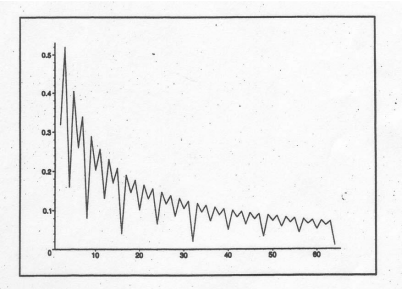

FIG. 2: Entropy analysis by lumping of the Feigenbaum symbolic sequence. Plot of $H(n)/n$ versus n.

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

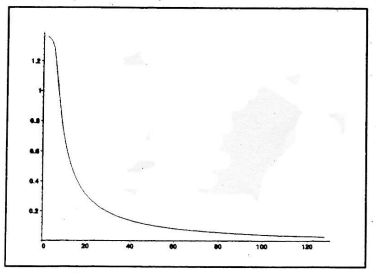

FIG. 3: Entropy analysis by lumping of the first 9840 terms of the DNA sequence of lamba virus.

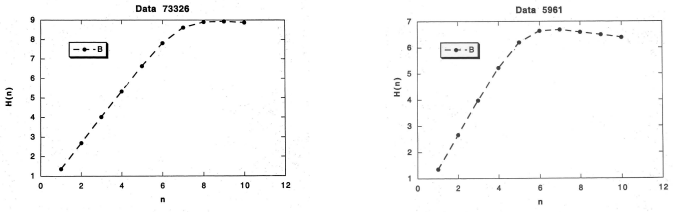

FIG. 4: Effect of bad statistics in the diagram $H(n)/n$ versus n of the human beta-globin region (a) and of the human interferone (b)After $n > 4$ the statistics becomes poor. Apparently, this result is more pronnounced for a short sequence (b).

Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications, Elounda, Greece, August 18-20, 2006 (pp481-491)

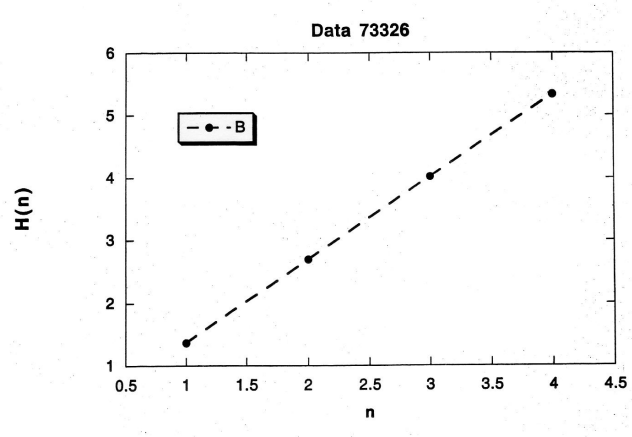

FIG. 5: Entropy analysis by lumping of the *human beta-globin region*. A linear scaling of the block entropies is observed where we have good statistics.