Method of DNA Analysis Using the Estimation of the Algorithmic Complexity

Ioan OPREA[1,2], Sergiu PAŞCA[1], Vlad GAVRILĂ[1]

[1] *"Iuliu Haţieganu" University of Medicine & Pharmacy, Faculty of General Medicine*
[2] *"Babeş-Bolyai" University, Faculty of Physics*
Cluj-Napoca, Romania
ioan.op@gmail.com

Abstract

The computational approach has a major impact in current biomolecular research. The computation of the algorithmic complexity (Kolmogorov) is a recently introduced method of global analysis for DNA sequences. The complexity is not directly computable, but one can estimate it using the compressibility of the given sequence. Intuitively, the more compressible a sequence is, the less complex it is; a random sequence is virtually incompressible.

We have developed a set of programs which work together with *GenCompress* (Chen, Kwong, Li – 2001).We have used them for analyzing a variety of sequences: complete viral genome, chloroplastic genome, eukaryotic genes and fragments of genes, etc.

We have found that exons are virtually incompressible, as opposed to introns, which posess various compressibility ratios. This confirms the existence of long-range correlations within introns, but not in exons. The method can also be used for detecting low complexity sequences, as well as homogeneous domain-like regions. This new procedure has also enabled the construction of phylogenetic trees.

As a conclusion, the study of the complexity of DNA sequences is a powerful tool in genome analysis.

Keywords

algorithmic complexity, DNA sequences, long-range correlations, large-scale structure, phylogeny

Introduction

The computational approach has nowadays a major impact on biomolecular research. One of the great impediments is that we are still heavily dependent on powerful resources. However, a new class of small and highly specialized applications is developed, that work very well on personal computers. The aim of this paper is to develop and explain the power of such an approach on DNA sequences. The starting point is the suggestion made by Chen, Kwong & Li [1] regarding the possibility of a global analysis of nucleotide sequences. Similar ideas were developed, to a lesser extent, in [2,3].

Materials and Methods

Materials

We have used our method for analyzing *complete viral genomes* (lentiviruses: HIV, SIV, FIV, BIV, VMV, OMVV, CAEV, EIAV, HTLV1, HTLV2, STLV1, STLV2; other types of viruses: VACCG, bacteriophage lambda), *chloroplastic genomes* (CHMPXX), *mithocondrial genome* (African elephant), *exons* and *introns* of eukariotic genes (dystrophin gene - http://www.dmd.nl/dmdprobe.html).

Except for the dystrophin exons and introns, all the other sequences were obtained from *GenBank* (http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome).

The DNA is made of long sequences of four kinds of nitrogen containing bases $\{a,c,g,t\}$. These sequences are grouped in *coding regions* – exons (of the eukariotic genes) and *non-coding regions* – introns (various regulatory regions such as promoters, enhancers, silencers, long repeats with apparently no function, etc.). The coding regions are translated into proteins, while the vast majority of the non-coding regions seem to have no biological function whatsoever.

Our analysis concerns strictly the sequence of nucleotides (both coding and non-coding), regarded as a string made of four characters. The spatial structure of DNA (nucleosomes, chromatin, chromosomes) plays no role in computing the complexity.

Methods

We make use of the so called *algorithmic complexity* (Kolmogorov) [4]. Formally, the complexity of a sequence is defined by the length of the *shortest* computer program that outputs the given sequence:

$$K(x) = \min_{p}\{l(p) : U(p) = x\} \tag{1}$$

where x is a binary sequence, $l(p)$ is the length of the program p and U is a universal computer (universal Turing machine)[4]. Intuitively, the simpler the sequence, the shorter the program needed to describe it is. For instance the first n decimal points of π can be generated using a relatively short program. At the opposite end, we have the completely random sequences, for instance the sequence we get by tossing a coin n times. This sequence is, with an overwhelming probability its own shortest program (it shows no regularity whatsoever).

A further extension of the concept is the conditional complexity of a sequence x with respect to a sequence with respect to a sequence y. It is defined by the length of the shortest program that outputs x when given y as input:

$$K(x|y) = \min_{p}\{l(p) : U(p, y) = x\} \tag{2}$$

This helpes us define an algorithmic distance (*similarity metric*) [5] between two sequences x and y:

$$d(x, y) = \frac{\max\{K(x|y^*), K(y|x^*)\}}{\max\{K(x), K(y)\}} = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} \tag{3}$$

where x^*, y^* are the shortest programs that output x and y respectively, and for the second part of the equality we have used the symmetry of the complexity [4]:

$$K(x, y) = K(x) + K(y|x^*) = K(y) + K(x|y^*) \tag{4}$$

Finally, we shall also use the algorithmic mutual information, which is given by:

$$I(x : y) = K(x) - K(x|y^*) = K(x) + K(y) - K(x, y) \tag{5}$$

Unfortunately, the Kolmogorov complexity is not computable. Therefore, in order to estimate it we'll have to use an compression method. *The complexity can be approximated by the length of the compressed sequence.*

The efficiency of the compression is estimated by the so-called *compression ratio*, measured in bits per symbol:

$$CR = \frac{l(x^*)}{N(x)} \tag{6}$$

where $l(x^*)$ is the length of the compressed sequence, x^*, in bits, and $N(x)$ is the number of symbols in the original sequence x.

According to a famous theorem by Shannon, in order to store a message that contains M types of symbols, we need at most $\log_2 M$ bits. In the case of DNA sequences we have 4 kinds of nitrogen containing bases $\{a,c,g,t\}$, so, a priori, we need 2 bits to code each base. We shall see in a moment that this is not entirely true for natural DNA sequences.

It is known that DNA sequences are difficult to compress. Usual compression programs such as *zip* or *rar* give compression ratios higher than 2 bits per base (*bpb*). In other words, they expand the original sequence. In order to get a good estimation of the algorithmic complexity we need a compression method that is able to detect the subtle regularities in gene sequences (i.e., *approximate repeats* and *complemented palindromes* – specific elements in natural DNA sequences).

We have used *GenCompress* developed by Chen, Kwong şi Li [1] (available for free download at http://www.cs.cityu.edu.hk/~cssamk/gencomp/GenCompress1.htm). It currently achieves one of the best compresstion ratios of DNA sequences. This program compresses an input file which contains only the characters $\{a,c,g,t\}$. The result is a compressed file, the size of which is of most importance in our subsequent calculations (*Fig.1*).

We have also used a set of programs developed by us, that facilitate the work with *GenCompress*. They perform various operations of the raw data produced by *GenCompress*:

- compute the compression ratio – according to equation (6)
- compute the distance between two sequences – according to equation (3). We will use this distance in order to generate the phylogenetic tree of the lentiviruses.
- compute a specific length scale for an array of sequences – we generate this array by splitting an original sequence. Afterwards we compute the compression ratios for the fragments and then use the *self-correlator method*: let $\{(x_1,y_1)\ldots(x_n,y_n)\}$ be a set of points so that the x's are equally distanced (δ-the distance between x_i and x_{i+1}). At step k, we evaluate

$$s_k = \sum_{i=k}^{n} y_i y_{i-k+1}$$

When $s_k \approx 0$ we have reached the length scale $L=(k-1)*\delta$ that is specific to the sequence. To determine the point $s_k \approx 0$ we plot s_k versus k.

- compute the mutual information between subsequences of a given sequence – we split the original sequence at various points and compute the mutual information between the resulting subsequences according to equation (5). Thus, the algorithmic mutual information becomes a function of the position of the splitting point.
- generate random sequences for comparison with natural ones – using the internal random number generator of the computer
- compute the compression ratios of increasing length fragments, thus determinig the mimimun length at which compression appears
- split a given sequence into constant length fragments, extract a specific subsequence for further analyzing, automatically generate the file format specific to *GenCompress*, automatically run *GenCompress* for long lists of files, etc.

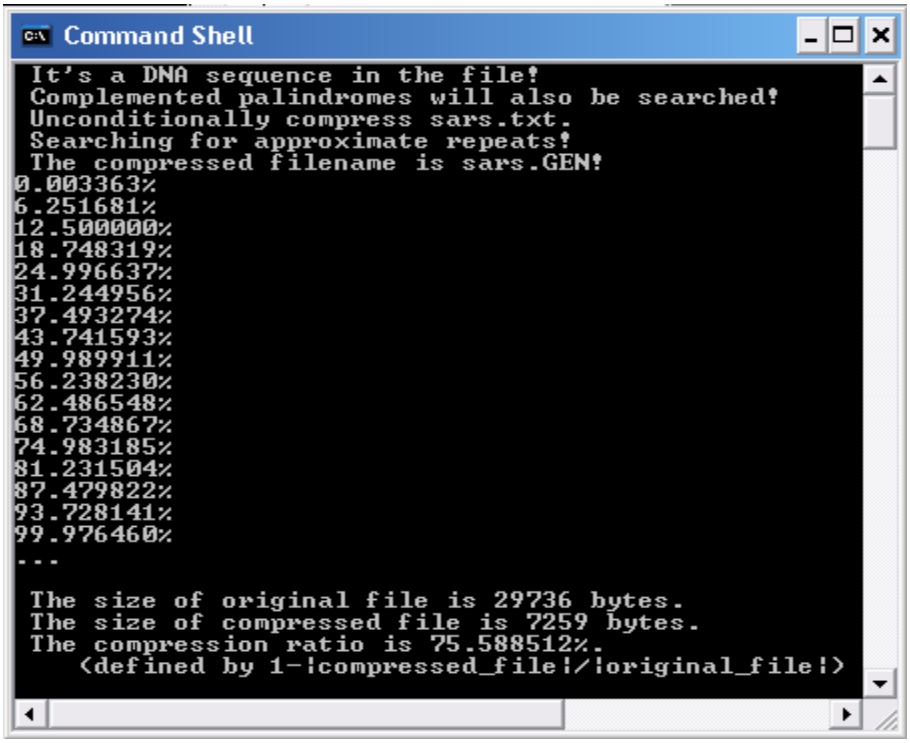

Fig.1. GenCompress output window

The investigations fall into four categories:

- the analysis of the exons compared to the introns with the purpose of identifying potential differences of compressibility (dystrophin gene)
- the construction of a complexity profile of a given sequence as well as the computation of a length scale by splitting the sequence into constant length fragments (VACCG, CHMPXX)
- the computation of the mutual information as a function of the distance by partitioning a sequence into two fragments with the purpose of identifying possible domains
- building phylogenetic trees through the complete genome analysis (lentiviruses) – the trees was built using the Fitch-Margoliash method as implemented in the PHYLIP package (version 3.62 - http://evolution.genetics.washington.edu/phylip.html)

Results

The analysis of natural versus random sequences of the same length shows clearly the existence of correlation in natural, real-world DNA sequences (*Fig. 2*).

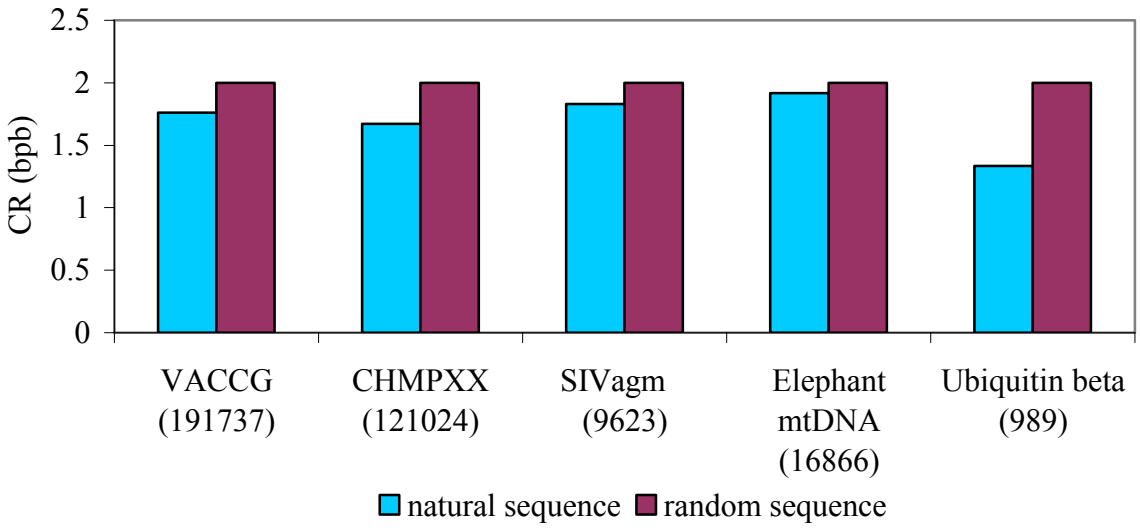

Figure 2. Comparison between natural and random sequences of the same length (the numbers represent the lengths in bases)

The effect is even more obvious if we compare the exons and introns of the dystrophin gene (*Fig. 3*).

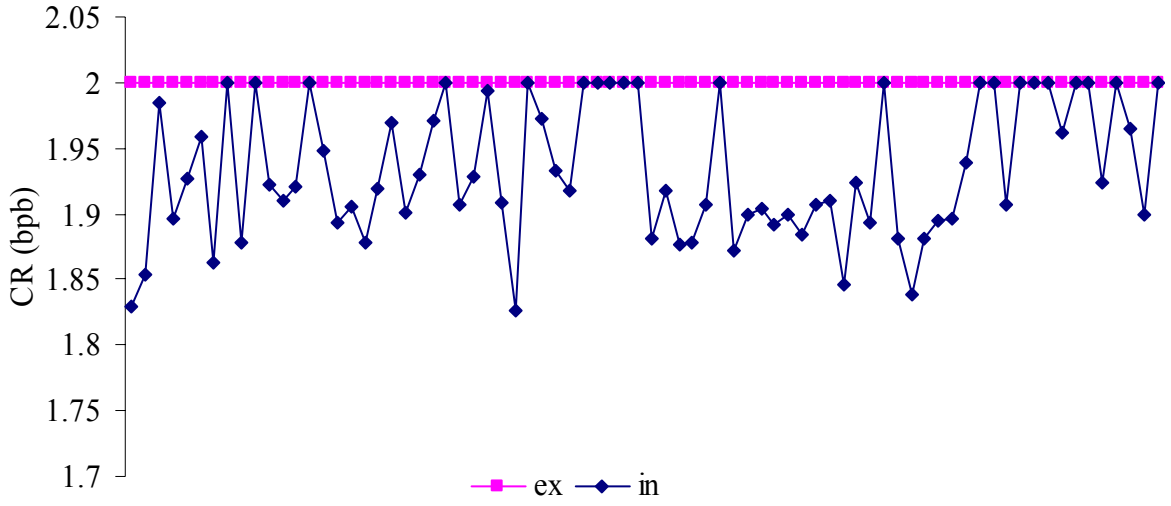

Figure 3. Dystrophin – Compression ratios of exons compared to introns

The difference between the complexity of the exons and that of introns is statistically significant (*t-test*: $p = 1.35*10^{-17}$). The complexity of the introns is lower than that of the exons: while the latter are virtually incompressible (CR = 2 bpb), the former exhibit correlations along the sequence, therefeore being more compressible ($CR_{average} = 1.933$ bpb).

The analysis of sequence fragments reveals theree aspects:

- the correlations appear to be stronger as the fragment length increases (*Fig. 4-5*; *Fig. 6*) – this can be explained by the spreading of the correlations at larger distances than the fragment length (splitting the sequence arbitrarily separates correlated regions); one can also determine the minimum length at which correlations become manifest (and so, fragments become compressible – 1600 bases for CHMPXX; *Fig. 7*)
- the compression ratios of the fragments are bounded from above by the compression ratio of random fragments of the same length ($CR_{random} = 2$ bpb), and from below by the overall compression ratio of the original sequence ($CR_{VACCG} = 1.76$ bpb; $CR_{CHMPXX} = 1.67$ bpb)
- the self-correlator gives a length scale specific for each sequence, of approximately 14.000 bases for VACCG and 6000 bases for CHMPXX – i.e., all the compressibility features of the sequence can, statistically, be found within a sequence of length equal to the length scale (*Fig. 8*; *Fig. 9*)

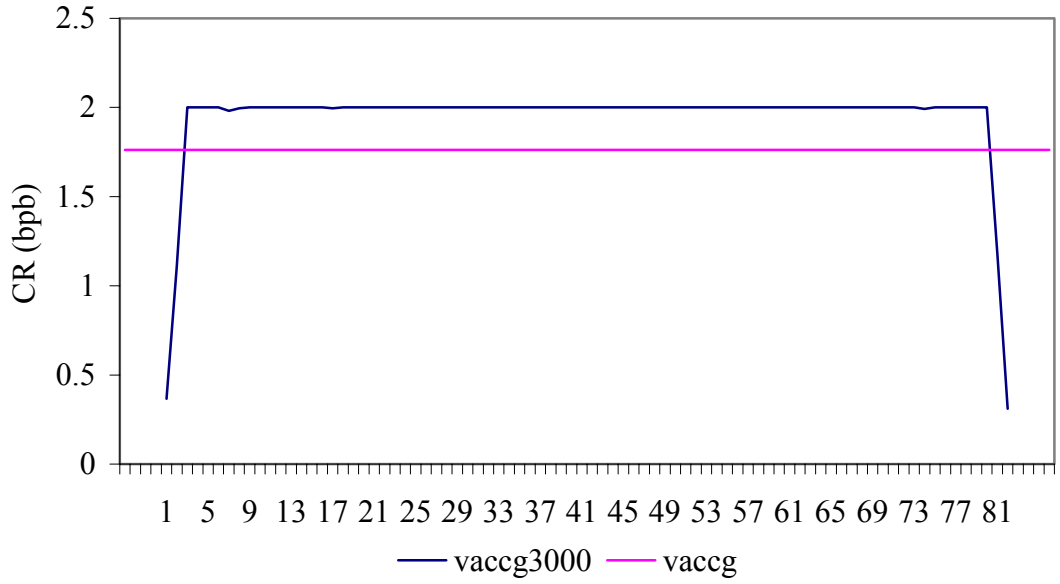

Figure 4. VACCG - Compression ratios of 3000 bases fragments. The fuchsia horizontal line represents the overall compressibility ratio of VACCG
(x-axis shows the number of the fragment)

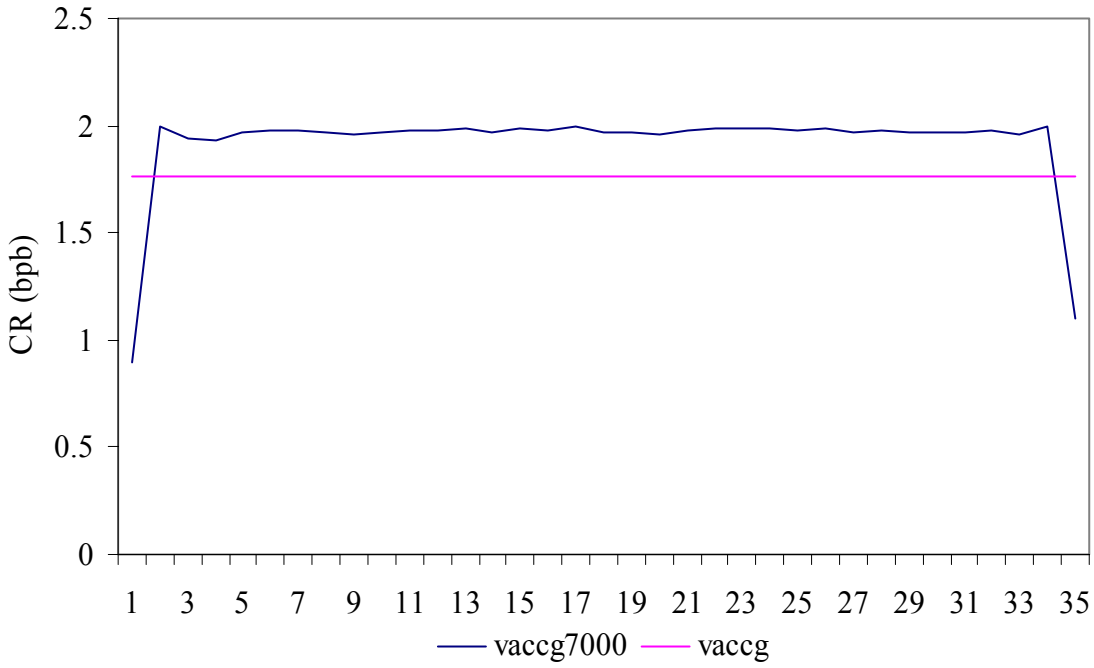

Figure 5. VACCG – Compression ratios of 7000 bases fragments. The fuchsia horizontal line represents the overall compressibility ratio of VACCG
(x-axis shows the number of the fragment)

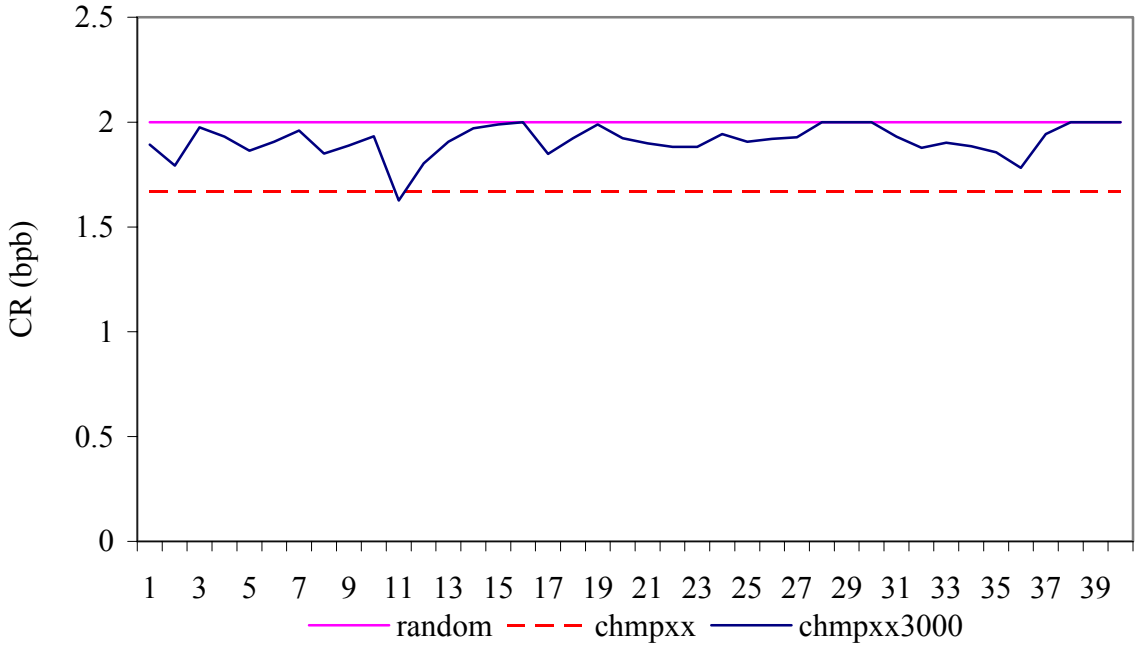

Figure 6. CHMPXX – Compression ratios of constant length fragments (3000 bases) versus random (but still 3000 bases) fragments. The dashed line represents the overall compression ratio of CHMPXX (x-axis shows the number of the fragment)

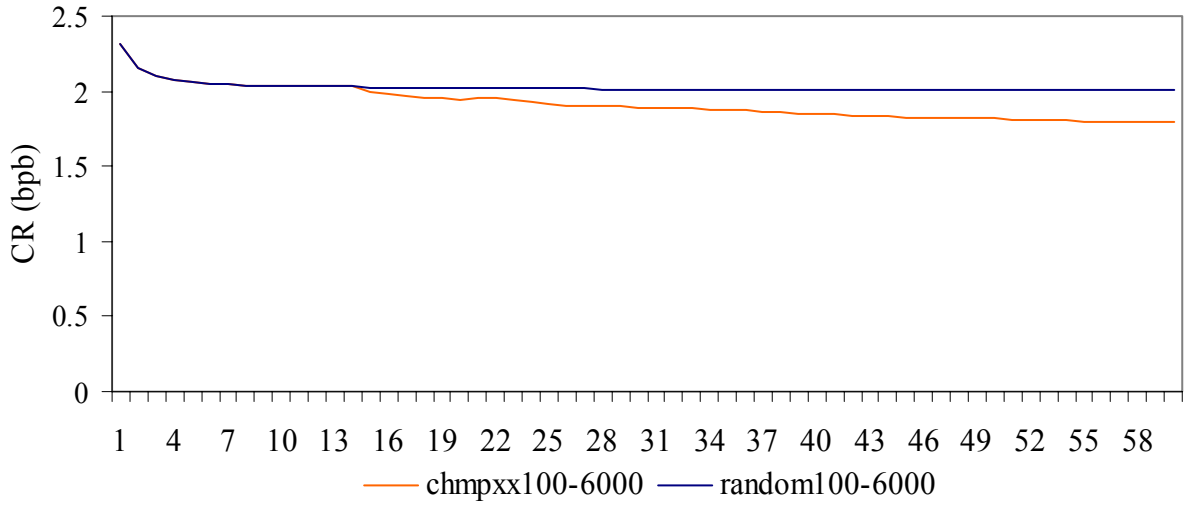

Figure 7. CHMPXX – Compression ratios of increasing length fragments (100 to 3000 bases) compared to fragments of the same length but coming form a random sequence (x-axis shows the number of the fragment)

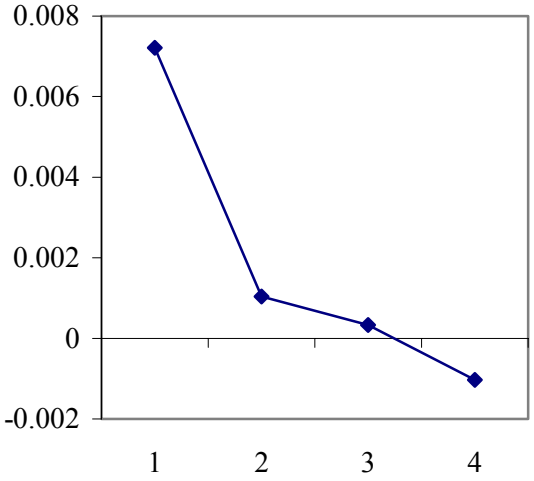

Figure 8. Self-correlator VACCG

δ =7000 bases $\Rightarrow$ ***L=14000 bases***

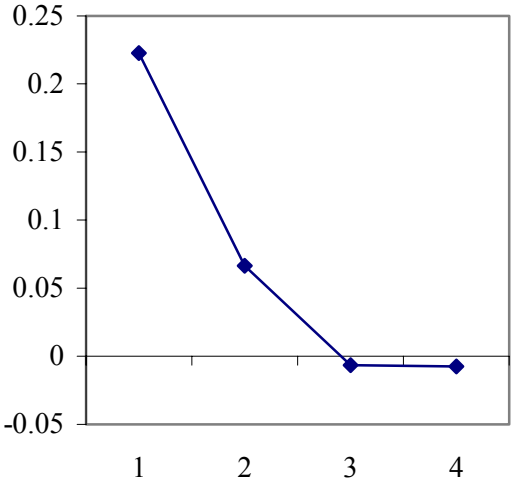

Figure 9. Self-correlator CHMPXX

δ =3000 bases $\Rightarrow$ ***L=6000 base***

By computing the mutual information between the two subsequences for different partition points, one can determine which is the best suited partition point, i.e., which splitting gives the most different subsequences (see [7] for a different approach). Thus, minima of the algorithmic mutual information give the boundaries of the relatively homogenous regions, which are a sort of 'informational domains'. Thus, the bacteriophage lambda (48502 bases) seems to be composed of two such domains, one of 22000 bases and the other one of 26502 bases (*Fig. 10*).

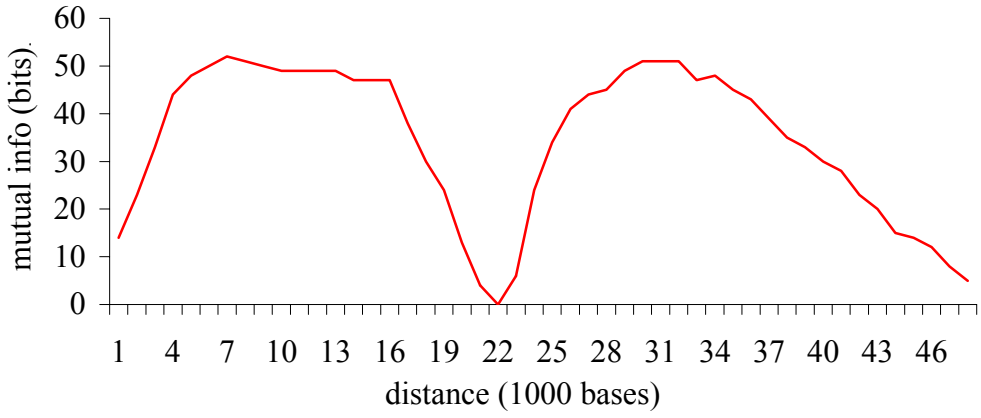

Figure 10. Algorithmic mutual information between the two fragments obtained by cutting the sequence of the bacteriophage lambda (the splitting point advances with 1000 bases at a step). - Notice the two distinct domains!

The fourth category of results is related to the possibility of constructing of phylogenetic trees by global genome analysis. We use the algorithmic distance (*similarity metric*) to calculate a matrix of distances (*Table 1*). This approach was originally suggested in [1,5]. We illustrate the method by generating a phylogenetic tree of lentiviruses (*Fig.10*).

	FIV	SIV	HIV1	HIV2	HTLV1	HTLV2	STLV1	STLV2	BIV	CAEV	EIAV	VMV	OMVV
FIV	0	0.983	0.982	0.986	1.013	1.014	1.013	1.016	0.986	0.982	0.980	0.983	0.982
SIV	0.983	0	0.981	0.978	1.003	1.002	1.002	1.004	0.981	0.982	0.982	0.982	0.981
HIV1	0.982	0.981	0	0.982	1.007	1.008	1.006	1.010	0.982	0.982	0.982	0.983	0.982
HIV2	0.986	0.978	0.982	0	0.998	1.000	0.999	1.001	0.982	0.983	0.985	0.985	0.984
HTLV1	1.013	1.003	1.007	0.998	0	0.971	0.332	0.971	0.995	1.012	1.008	1.014	1.014
HTLV2	1.014	1.002	1.008	1.000	0.971	0	0.975	0.918	0.994	1.013	1.009	1.014	1.015
STLV1	1.013	1.002	1.006	0.999	0.332	0.975	0	0.969	0.994	1.012	1.008	1.013	1.014
STLV2	1.016	1.004	1.010	1.001	0.971	0.918	0.969	0	0.995	1.015	1.010	1.015	1.016
BIV	0.986	0.981	0.982	0.982	0.995	0.994	0.994	0.995	0	0.986	0.983	0.985	0.985
CAEV	0.982	0.982	0.982	0.983	1.012	1.013	1.012	1.015	0.986	0	0.983	0.956	0.945
EIAV	0.980	0.982	0.982	0.985	1.008	1.009	1.008	1.010	0.983	0.983	0	0.983	0.983
VMV	0.983	0.982	0.983	0.985	1.014	1.014	1.013	1.015	0.985	0.956	0.983	0	0.830
OMVV	0.982	0.981	0.982	0.984	1.014	1.015	1.014	1.016	0.985	0.945	0.983	0.830	0

Table 1. Distance matrix for the construction of the phylogenetic tree of lentiviruses

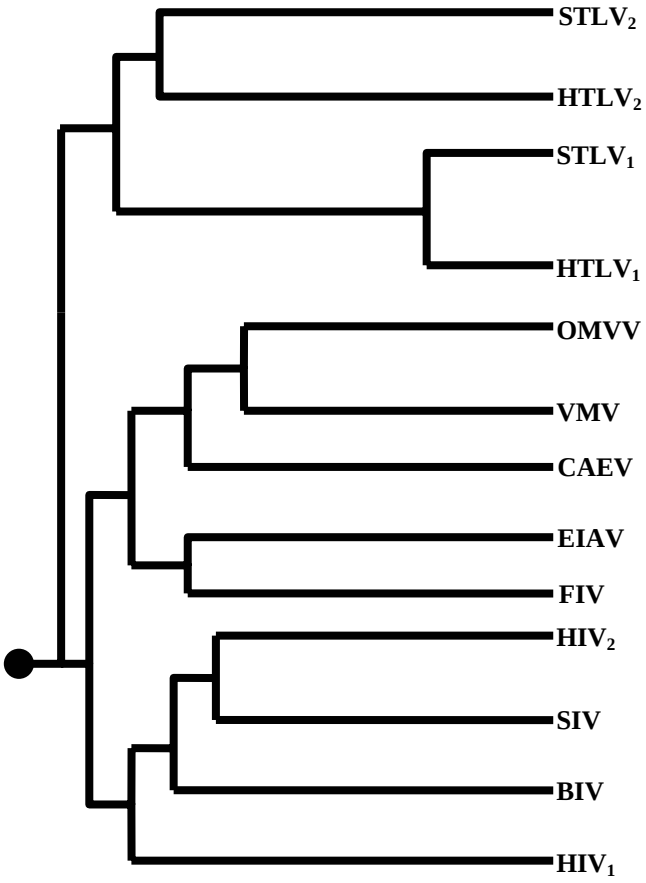

Figure 11. Phylogenetic tree of lentiviruses built using the similarity metric

Discussions

The analysis of the algorithmic complexity is a powerful tool for studying DNA sequences. As a global (and holistic) method it describes the informational content of a sequence. One can thus detect and characterize long range correlations existing in natural DNA sequences, as opposed to random ones. The different compression ratios of an exon compared to an intron are an example of maximization the information carried by the former. The best way of storing information is achieved when one uses sequences consisting of independent symbols (*white noise*) and that is exactly what we see in exons. It is also of great importance that exons be able to mutate more freely in order to allow for evolution, thus they tend to achieve random sequences. The introns, on the other hand, exhibit correlations between their bases, thus becoming compressible.

It would also be interesting to relate these correlations that induce compressibility to the *1/f spectra* discovered in the early nineties [10]. Theoretical models for generating these *1/f spectra* have already been developed, emphasizing the role of gene duplication [9]. One possible role for these correlations might be that of stabilizing the genome [6]. It has also been suggested that different complexity regions might have different mutation rates [1].

The analysis of the complexity profile shows the existence of a long range structure of natural DNA sequences. One can determine the minimum and maximum complexity regions, the minimum distance at which compressibility appears (hence, correlations) as well as the characteristic length scales. We have revealed the presence of certain ’compressibility units’ within natural genomes, that is, the compressibility appears only above a certain length (not to be confused with the natural length scale given by the self-correlator). Below this length, the sequences are virtually random. (Of course, these boundaries are determined by the sensibility of the method. One must keep in mind that compression is just an approximation for the complexity). The mutual information is yet another tool for detecting domain-like regions. The final purpose of such an analysis is to uncover some general patterns in the large scale structure of the genome (presumably it goes way beyond nucleosome level). Then, of course, the problem remains of how to link these ’informational structures’ to the biological function [8]. However, this is beyond the scope of this paper.

Another application of the method was proposed in [5] and it refers to the construction of phylogenetic trees. It has the great advantage of being completely automatic and allows the

analysis of a great number of species in a relatively short time. The results are consistent with the already known phylogenies, determined through homology studies; however, this method is far simpler.

To this end, we have developed an alternative method of genome investigation, based on the theory of algorihmic complexity. It remains to be determined by further studies, to what extent this approach is useful in DNA analysis on a large scale.

Conclusions

Our results confirm the existence of long range correlations in natural DNA sequences [7,8,9,10]. Many roles have been suggested for these correlations including the hope that some large scale structure of the genome might be discovered. This is not necessarily unrelated to self-organizing systems driven by evolution. Our approach suggests that there is such a structure at least at the symbolic level. It remains however to be established what is the biological meaning of this structure and what is the reason for which biological sequences exhibit this heterogeneity with respect to their complexity. The ultimate goal is, of course, to learn the "genome organization" principles, and explain this organization using our knowledge about evolution. Also, this might reveal some new aspects of evolution itself.

References

[1] Chen X, Kwong S, Li M. A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison. In: The Tenth Workshop on Genome Informatics, 14-15 December 1999, http://www.jsbi.org/journal/GIW99/GIW99F06.pdf

[2] Sato H, Yoshioka T, Konayaga A, Toyoda T. DNA Data Compression in the Post Genome Era. In: Genome Informatics 12: 512–514 (2001), http://www.jsbi.org/journal/GIW01/GIW01P130.pdf

[3] Matsumoto T, Sadakane K, Imai H, Okazaki T. Can General-Purpose Compression Schemes Really Compress DNA Sequences? In: The Fourth Annual International

Conference on Computational Molecular Biology, Tokyo, Japan, 8 – 11 April 2000, http://recomb2000.ims.u-tokyo.ac.jp/Posters/pdf/40.pdf

[4] Grünwald P, Vitányi P. Shannon Entropy and Kolmogorov Complexity, http://arxiv.org/PS_cache/cs/pdf/0410/0410002.pdf

[5] Li M, Chen X, Li X, Ma B, Vitányi P. The Similarity Metric, http://arxiv.org/PS_cache/cs/pdf/0111/0111054.pdf

[6] Huen YK.Brief Comments on Junk DNA:Is It Really Junk? In:Complexity International, http://www.complexity.org.au/ci/vol09/huen01/huen01.pdf

[7] Li W. The Complexity of DNA. In: Complexity, 3(2):33-37, http://www.nslij-genetics.org/wli/pub/complexity97.pdf

[8] Li W. The Study of Correlation Structures of DNA Sequences: A Critical Review, http://arxiv.org/PS_cache/adap-org/pdf/9704/9704003.pdf

[9] Li W, Marr TG, Kaneko K, Understanding Long-Range Correlations in DNA Sequences, http://citeseer.ist.psu.edu/cache/papers/cs/14616/http:zSzzSzlinkage.rockefeller.eduzSzwlizSzpubzSzlrc3.pdf/li94understanding.pdf

[10] Voss RF. Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences. In: Physical Review Letters 68(25):3805-3808, http://www.nslij-genetics.org/dnacorr/voss92.pdf