FRACTAL DIMENSION OF HUMAN CHROMOSOME 22

Paul CRISTEA*, George Alex POPESCU**

*University "Politehnica" of Bucharest, Spl. Independentei 313, 77206 Bucharest, Romania, pcristea@dsp.pub.ro
** SUPELEC - École Supérieure d'Électricité, Plateau de Moulon - 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France, georgealex.popescu@supelec.fr

Abstract: *Based on complex genomic signal analysis, it has recently been reported that DNA sequences show large-scale regularities, at the scale of whole chromosomes. The paper shows that these long range correlations can also be found in a fractal-like structure of DNA genomic signals, as well as in their 1/f noise. Specifically, the paper studies the fractal dimension of DNA segments along homo sapiens chromosome 22. The results reveal that the probabilities of occurrence of nucleotides and groups of nucleotides in a DNA sequence depend on the distribution of nucleotides along the entire sequence and that this correlation is stronger in the extra-genic, non-coding, regions.*

Key words: *Fractal Analysis, Fractal Dimension, Genomic Signals*

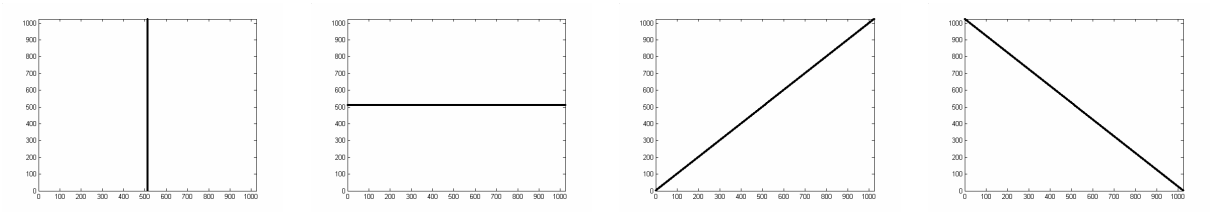
1. INTRODUCTION

Recently it has been found that seemingly random phenomena from very different domains -- like commuter traffic, earthquakes, electric circuits, flood records and market time series -- display some common behavior resulting from correlations between distant elementary events. These long range correlations show up in a fractal-like structure of the data describing such systems. DNA sequences share to some extent these features and belong to the same family of fractal-like objects. The fractal patterns emerge because individual events in an apparently random system are actually correlated with previous occurrences of other events. The analysis of genomic signals corresponding to DNA sequences [2-4] has revealed large scales statistical properties of at the scale of whole chromosomes. Such long range correlations are expressed both in a fractal-like structure of DNA genomic signals and in 1/f noise [7, 11]. The position of nucleotides – adenine, guanine, cytosine and thymine – in a DNA sequence depends on the distribution of nucleotides on the entire chromosome. The patterns of nucleotide occurrence in DNA sequences bear similarities to the 1/f noise, typical for fluctuations, but ubiquitous in nature and techniques. Fluctuations are the time analogues of fractal shapes, such as snowflakes and coastlines, which have the property of self-similarity over several scales of magnitude: the parts resemble the system as a whole. Base pairs in DNA do not occur in a completely random fashion, especially outside the coding regions, in the so called "junk DNA" area that does not encode directly information about protein synthesis. Exons – the encoding regions of DNA -- lack long-range correlation and resemble to white noise. This is so primarily because the exons encode proteins for which the functionality is given by the structure – an essentially qualitative feature that is improperly described by quantitative parameters that could be correlated. Long-range correlations - which extend over distances of hundreds of thousands to tens of millions of base pairs, i.e., up to the scale of whole chromosomes, have a functional role in the control of crossing-over and species separation [4], and also could represent a trade-off between efficient information storage and protection against error in the genetic code by adding some redundancy to the encoding.
Also very recently, it has been claimed that the prokaryote DNA sequences were the least correlated and that correlations would increase as organisms moved up the evolutionary ladder. These claims are not confirmed by the genomic signal analysis that reveals long range correlations for all the studied taxa.
The paper presents some preliminary results in the study of fractal structures in DNA sequences of *homo sapiens*. Specifically, the fractal dimension of the cumulated phase of the complex genomic signal corresponding to segments of the *homo sapiens* chromosome 22 has been measured in a sliding window approach.

2. FRACTAL DIMENSION OF SIGNALS

The dimension of a fractal-like structure can be measured using a multiresolution technique as for instance the doubling (merging) boxes method described in [5]. Gradually increasing the size of the box used to explore the structure of a distribution of points in the background array, the number of the boxes containing points that belong to the given set – which provides a quantitative measure of the extension of the set of points at the given resolution – decreases according to a power law, the exponent being the fractal dimension of the set of points.

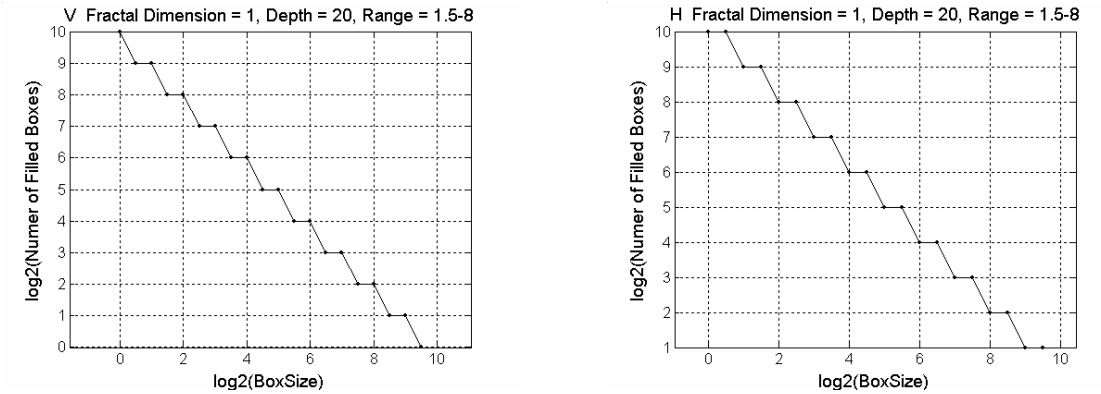
The self-similarity of the structure is revealed by the linearity of the plot log(*N*) *vs* log (*B*), where *N* is the number of field boxes and *B* the size of the box, while the slope gives the (fractal) dimension. The range of linearity reveals the range of scales for which the fractal self-similarity property holds. Figure 1 exemplifies the application of the method for several linear signals embedded in 2D arrays comprising 1024 pixels with value one (black points) out of the total of 1024 x 1024 pixels of the background array, initially zero (white).



(a) (b) (c) (d)

Fig. 1. Examples of linear signals embedded in a 2D background array of size 1024 x 1024 pixels.

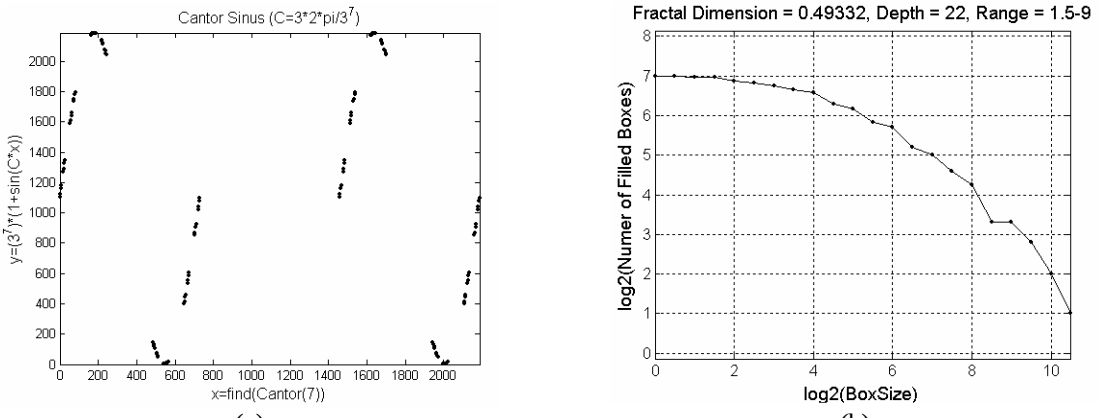
The corresponding dependencies between the logarithms of the number of filled boxes (*N*) and the size of the boxes (*B*) is given in figures 2 a and 2 b. To the vertical line in Fig 1 a corresponds the diagram in Fig. 2a, while to the horizontal and oblique lines in Fig 1 b, c, d corresponds the same diagram in Fig. 2 b. In all these instances the slope of the regression line gives the exact fractal dimension one.



(a) (b)

Fig. 2. Log-log plot of the dependence between the number of filled boxes and the box size for the linear signals shown in Fig 1. The vertical line in Fig. 1a generates the plot in Fig. 2 a, while all other lines give the diagram in Fig. 2 b. For all the instances the fractal dimension results one as expected.

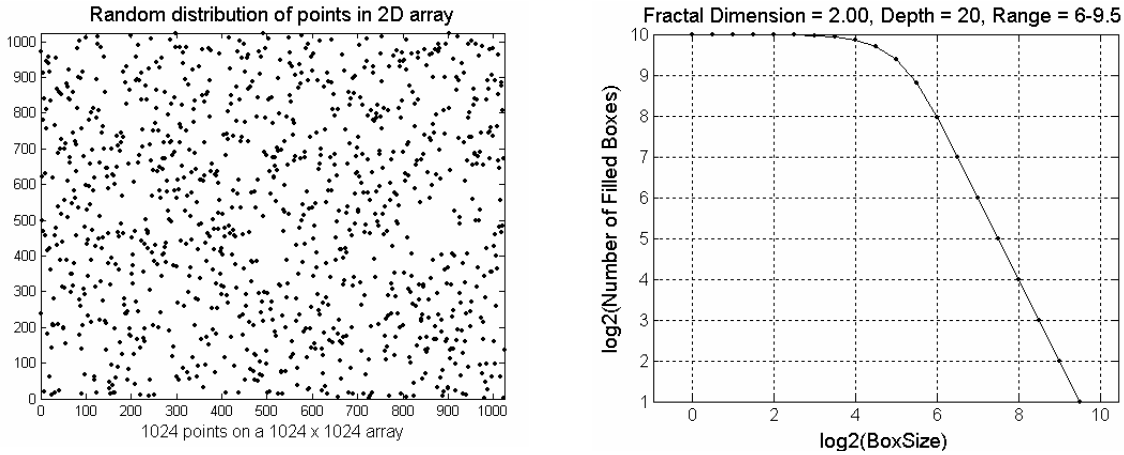
A more complex case is shown in Fig. 3. The Cantor sinus signal is obtained by selecting a Cantor subset of points (of order 7) from the 2187 points of the discrete curve of a sinus extended over three periods. The set of points comprises 2^7 = 128 points (ones), out of the total of 2187 x 2187 = 3^{14} empty initial positions of the background (zeros).



(a) (b)

Fig. 3. (a) Cantor sinus signal obtained by selecting a Cantor subset of points from the discrete curve of a sinus extended over three periods. (b) Log-log plot of the dependence between the number of filled boxes and the box size for the signal shown in Fig. 3 a.

The log-log plot of the dependence between the number of filled boxes and the box size reveals in this case a different fractal-like behavior for various ranges of scales. The set of points can be approximated for each such range of scales with a different fractal, but does not exhibit the self-similarity at all scales typical for fractals.



(a) (b)

Fig. 4. (a) Random signal comprising 1024 points embedded in an 1024 x 1024 background array. (b) Log-log plot of the dependence between the number of filled boxes and the box size for the signal shown in Fig. 4 a.

Figure 4 presents the case of a random signal comprising 1024 points distributed one in each horizontal location and randomly over the vertical locations. The diagram in Fig. 4 b reveals that the self-similarity holds only for large enough boxes (with approximately $\log_2(B) \geq 5.5$, i.e., with an area 2^{11} times the area of an initial pixel). For smaller boxes, the structures are no longer similar and the fractal behavior degenerates in point-like features.
The properties of the investigation method shown by these tests must be taken into account when analyzing the fractal structure and the fractal dimension of genomic signals.

3. RESULTS FOR GENOMIC SIGNALS

DNA sequences from contig NT_011519 of *Homo Sapiens* chromosome 22 have been downloaded from the GeneBank [10] and converted into complex genomic signals as described in previous work [2-4]. The cumulated phase of the genomic signal has been investigated using a sliding window moving along the sequence. The local fractal dimension of the signal inside the window has been determined in an attempt to identify the local properties of the various regions.

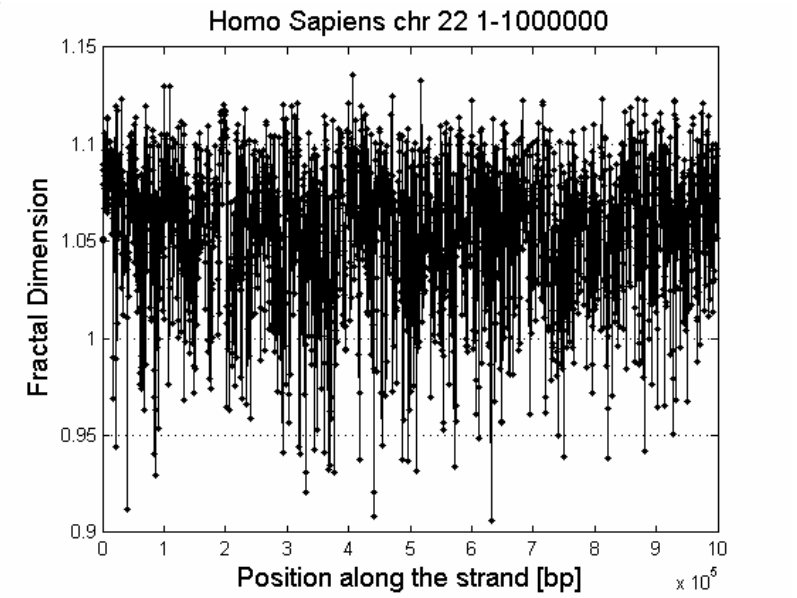


Fig.4. Fractal dimension along a segment of 1Mbp at the begining of contig NT_011519 [10] of Homo Sapiens chromosome 22.

The windows have been chosen of the size 1024 bp, and the advance of the sliding window has been 512 pixels along the strand. Figure 4 gives the results for a 1Mbp segment at the beginning of the contig NT_011519. The

average fractal dimension for the whole segment is of about 1.05, slightly over the dimension of a 'simple' line, but there are significant fluctuations of the dimensions from various regions of the contig. An attentive analyses of the results is necessary to correlate the measured fractal dimension to the annotation of the chromosome and the attach biological meaning to the numerical values.

3. CONCLUSIONS

The method for measuring the fractal dimension of a set of points, the corresponding algorithm and its MATLAB® implementation allow the analysis of the structure of large scale signals to reveal long range correlations with obvious biological significance. More research is necessary to clarify the meaning of the fractal dimension and of the range of scale over which the genomic signal is self similar. Other types of genomic signals, primarily the cumulated sums of the complex signals that generate the genomic complex path will be investigated.

REFERENCES

[1] M. F. Barnsley, A. Jacquin, L. Reuter and A. D. Sloan, "Harnessing chaos for image synthesis", *Computer Graphics*, **22**, 1988.

[2] M. F. Barnsley, A. D. Sloan, "A better way to compress images", *Byte*, **13**, 1988, 215-223

[3] P. Cristea, "Large Scale Features in DNA Genomic Signals", ELSEVIER, Signal Processing, Special Issue on Genomic Signal Processing, **83,** 2003, 871-888.

[4] P. Cristea, "Conversion of Nitrogenous Base Sequences into Genomic Signals", *Journal of Cellular and Molecular Medicine*, **6**, 2, April – June 2002, 279-303.

[5] P. Cristea, "Genomic Signals of Re-Oriented ORFs", *EURASIP Journal on Applied Signal Processing, Special Issue on Genomic Signal Processing*, under print, 2003.

[6] P. Cristea, "An Efficient Algorithm for Measuring Fractal Dimension of Complex Sequences", Proceedings of IAFA'2003, Bucharest, Romania, 7-10 May 2003.

[7] N.V. Dokholyan, S.V. Buldyrev, *et al.*, Distribution of Base Pair Repeats in Coding and Noncoding DNA Sequences, *Physical Review Letters*, **79**, 25, pp. 5182-5185, 1997.

[8] J. Hutchinson, "Fractals and self-similarity", *Indiana U. J. Math.*, **30**, 1981, 713-747.

[9] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome", *Nature*, **409**, 2001, 860-911.

[10] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, National Center for Biotehnology Information, GenBank, *http://www.ncbi.nlm.nih.gov/genoms/*.

[11] R. F. Voss, "Fractals in Nature: From characterization to simulation", in *The Science of Fractal Images*, Heinz-Otto Peitgen, Dietmar Saupe, editors (Springer), 1988.

[12] J. D. Watson, F. H. C. Crick, "A Structure for Deoxyribose Nucleic Acid", *Nature*, **171** (737), 1953.