7th International Workshop
for Young Mathematicians
Applied Mathematics
Kraków, 20-26 September 2004
pp. 57-68

ITERATED FUNCTION SYSTEMS AND MULTIFRACTAL ANALYSIS OF DNA SEQUENCES

GRZEGORZ HARAŃCZYK

It is human nature to find patterns in things
- whether in the shape of clouds,
the arrangement of sand, a chain of events, or the digits of π.

ABSTRACT. The recent progress in experimental techniques of molecular genetics has made available a large amount of genome data. The availability induces many questions and opens the possibility to establish some global properties of the DNA sequences. This leads to a requirement for specialized tools to view and analyze the data.

This is a review of a few recent papers, in which a new approach based on dynamical systems techniques was proposed. We use an iterated function system (IFS) model to simulate the multifractal structure of the DNA sequence.

1. Introduction

DNA (deoxyribonucleic acid) stores genetic information for long term use. It is a linear sequence of four bases (adenine, cytosine, thymine, guanine) which provides information for protein synthesis as well as information for signals to regulate the inner workings of the organism. Since the mid-1970s, molecular biologists have been able to obtain the sequences of longer and longer stretches of DNA, culminating in 1995 with completion of the first complete sequences of entire genomes.

Presently, only the function of a few percent of the DNA is known, the rest has been believed to be "junk". There are a total of approximately 3,000,000,000 bases in each human genome and about 97% of it has been designated "junk" (the ratio of functional and junk DNA differs widely per species). The most exhaustive knowledge is about the genes responsible for the bodily structures, the structural genes, which are the simplest part of the system. But the knowledge about the most important part of this system, the regulator genes, is incomplete. The genetic code language of these genes is only partially known.

In this paper we apply dynamical systems techniques to develop a mathematical tool for representing DNA sequences and revealing some underlying structure in those sequences.

We use the idea of iterated functions systems to introduce a Chaos Game Representation Algorithm that produces characteristic patterns for symbolic sequences. This

Key words and phrases. multifractal analysis, chaos game, IFS, DNA sequences.

technique converts a sequence into a two-dimensional representation that preserves subsequence structure and provides a visual representation.

Based on this idea, we present an initial analysis of the genetic data, describing some of the insights that can be gleaned from the sequence.

The paper is organized as follows. The next section contains some notions, definitions and basic properties concerning iterated function systems. This leads to the definition of Chaos Game Representation. Next, in Section 3, we apply our model to DNA sequences.

In Section 4 we introduce the concept of generalized dimensions and multifractal formalism. We also show relation between generalized dimensions and multifractal spectrum.

Finally, in Section 5, we present some results and practical applications of this approach.

2. IFS REPRESENTATION FOR SYMBOLIC SEQUENCES

We begin by recalling the definitions and properties concerned with iterated function systems that will be needed throughout this paper. An iterated function system (IFS) is a special case of regular stochastic dynamical systems, which is specified by N maps transforming a metric space into itself and N probabilities which characterize the likelihood of choosing a particular map at each step of the evolution of the system.

Under certain conditions using the Banach Contraction Principle one can prove the existence of a unique attractive invariant measure for an IFS. The support of this measure is called the attractor of the IFS and has fractal structure for a wide class of IFS models. The basic properties of iterated function systems can be found in various books, for example, Lasota and Mackey [1], Barnsley [2] or Jürgens, Peitgen and Saupe [3].

Let $X \subset \mathbb{R}^d$ be a compact set. Assume that the $S_i : X \rightarrow X$, are strict contractions, i.e., Lipschitz functions with the Lipschitz constants $L_i < 1$ for $i = 1, \ldots, N$. If, in addition, there are given $p_i > 0$, satisfying $\sum_{i=1}^{N} p_i = 1$, then the family $(S, p) = (S_i, p_i)_{i=1}^{N}$ is called an iterated function system (with constant probabilities).

The IFS under consideration satisfies sufficient conditions for the existence of the attractor. Let us denote this attractor by $\mathcal{A}$.

From a computational viewpoint, an attractor can be generated according to two techniques: deterministic and stochastic.

Using the deterministic procedure we build the sequence of sets X_n:

$$\begin{cases} X_0 = X \\ X_{n+1} = \bigcup_{i=1}^{N} S_i(X_n). \end{cases}$$

When n is large, X_n is an approximation of the real attractor $\mathcal{A}$.

According to the stochastic principle we choose a point $x_0 \in X$ and then successively define the sequence $\{x_n\}$ by choosing

$$x_{n+1} \in \{S_1(x_n), \ldots, S_N(x_n)\},$$

for $n = 0, 1, \ldots$ in such a way that $x_{n+1} = S_k(x_n)$ with probability p_k.

Then $\bigcup_n \{x_n\}$ is an approximation of the real attractor of (S, p).

The initial point x_0 can be arbitrary chosen in X, because all the maps S_i are strongly contracting. The following fact is a basic for the computational construction of attractors. If the IFS satisfies above conditions, then for every $\varepsilon > 0$ there exist n_0 and m_0 such that $\text{dist}(\{x_n, \ldots, x_{n+m}\}, \mathcal{A}) < \varepsilon$ for every $n > n_0$ and $m > m_0$ (here dist stands for the Hausdorff distance). The larger n and m are, the more precise the approximation is.

The second approach provides a convenient framework for the representation, description and analysis of symbolic sequences from an alphabet $\{a_1, \ldots, a_N\}$, since any sequence $\tau = a_{i_1} \ldots a_{i_L}$ of length L corresponds to a set

$$CGR(\tau) = \{x \in X : x = S_{i_k}(S_{i_{k-1}}(\ldots S_{i_1}(x_0))), \quad k = 1, \ldots, L\}.$$

This method is called Chaos Game Representation (CGR). The main advantage of using CGR is that it represents both statistical properties of frequencies of symbols as well as sequentiality properties – i.e., which symbols follow others.

We can also use the deterministic method for representing symbolic sequences. In such a way the sequence $\tau = a_{i_1} \ldots a_{i_L}$ corresponds to a region of the attractor $S_{i_L}(S_{i_{L-1}}(\ldots(S_{i_1}(X))))$ called an order-L iterator of the attractor [17].

3. Chaos game representation of gene structure

A DNA sequence can be treated as a string composed from four letters A, C, T and G, representing the nucleotides adenine, cytosine, thymine and guanine, respectively. We use the chaos game representation introduced in previous section to represent the DNA sequences.

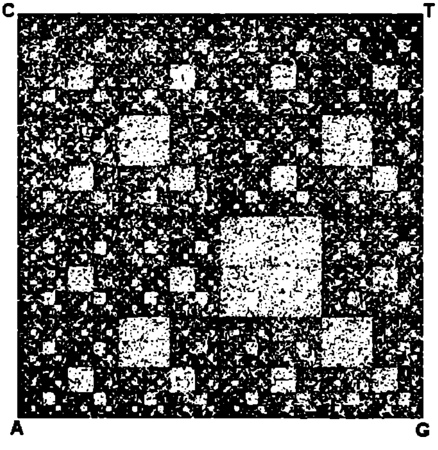

Figure 1. CGR for the DNA sequence (HUMHBB).

To this aim, we identify the DNA bases A, C, T and G with four maps S_A, S_C, S_T and S_G transforming the unit square $X = [0, 1] \times [0, 1] \subset \mathbb{R}^2$ into itself.

Let $x_0 = (\frac{1}{2}, \frac{1}{2})$ and

$$\begin{aligned} S_A(x,y) &= \Big(\frac{1}{2}x, \frac{1}{2}y\Big), \\ S_C(x,y) &= \Big(\frac{1}{2}x, \frac{1}{2}y + \frac{1}{2}\Big), \\ S_T(x,y) &= \Big(\frac{1}{2}x + \frac{1}{2}, \frac{1}{2}y + \frac{1}{2}\Big), \\ S_G(x,y) &= \Big(\frac{1}{2}x + \frac{1}{2}, \frac{1}{2}y\Big). \end{aligned}$$

For a given DNA sequence τ, e.g., $\tau = GAATTCTAATCTCC\ldots$ [1] , we obtain

$$CGR(\tau) = \{x_0, S_G(x_0), S_A(S_G(x_0)), S_A(S_A(S_G(x_0))), S_T(S_A(S_A(S_G(x_0)))), \ldots\}.$$

The genetic sequence is represented by points within the square X, where the four vertices of the square correspond to the four DNA bases. We can label the corners by A at $(0,0)$, C at (0,1), T at $(1,1)$ and G at $(1,0)$.

Then roughly speaking, we perform the following steps: First, we pick an arbitrary starting point x_0, e.g., x_0 at the center of the square $(\frac{1}{2}, \frac{1}{2})$. Then, each letter in the sequence tell us how to move. The first point x_1, representing the first base in the DNA sequence, is plotted half way between x_0 and the corner representing that base. The second point x_2 is plotted half way between previous point, x_1, and the corner representing the second base etc.... This procedure is continued until the sequence is completed.

We analyzed a few different DNA sequences [2] and we obtained examples of distinctive patterns (see Figure 2.).

The pictures uncover a complex structure, which varies depending on the sequence. There are slight differences within the same species – compare patterns for *Mus musculus*: AC099415, AC108947, AC131721, AC141647, or for *Zebrafish*: BX000363, BX088654, BX255951, BX890565, or for *Human*: BX664615, M94081, NT007819.

This approach to representation of gene structure was also proposed by Jeffrey [4].

We propose the model according to which purines (A,G) and pyrimidines (C,T) are connected with opposite corners of the square, whereas Jeffrey used the natural dictionary ordering of genetic alphabet, i.e., A at $(0,0)$, C at (0,1), G at $(1,1)$ and T at $(1,0)$.

The different ordering do not change statistical properties of the obtained patterns, but in some cases the features of obtained patterns can be easier explained. We can observe that mysterious characteristic "double-scoops" for Jeffrey's patterns are built from "forbidden squares" corresponding to a fact that guanine almost never immediately follows cytosine (forbidden word "CG"). But in some cases explanation of obtained patterns is not so simply (*Escherichia coli* – L10328 shown in Figure 2.).

To analyze the chaos game we need a suitable formal set of instruments which allows us to specify precisely characteristics which distinguish some patterns from others. We propose to use generalized dimensions and multifractal spectrum.

[1] HUMHBB - Human beta globin region on chromosome 11; DNA linear 73308 bp

[2] available in the GenBank database at www.ncbi.nlm.nih.gov

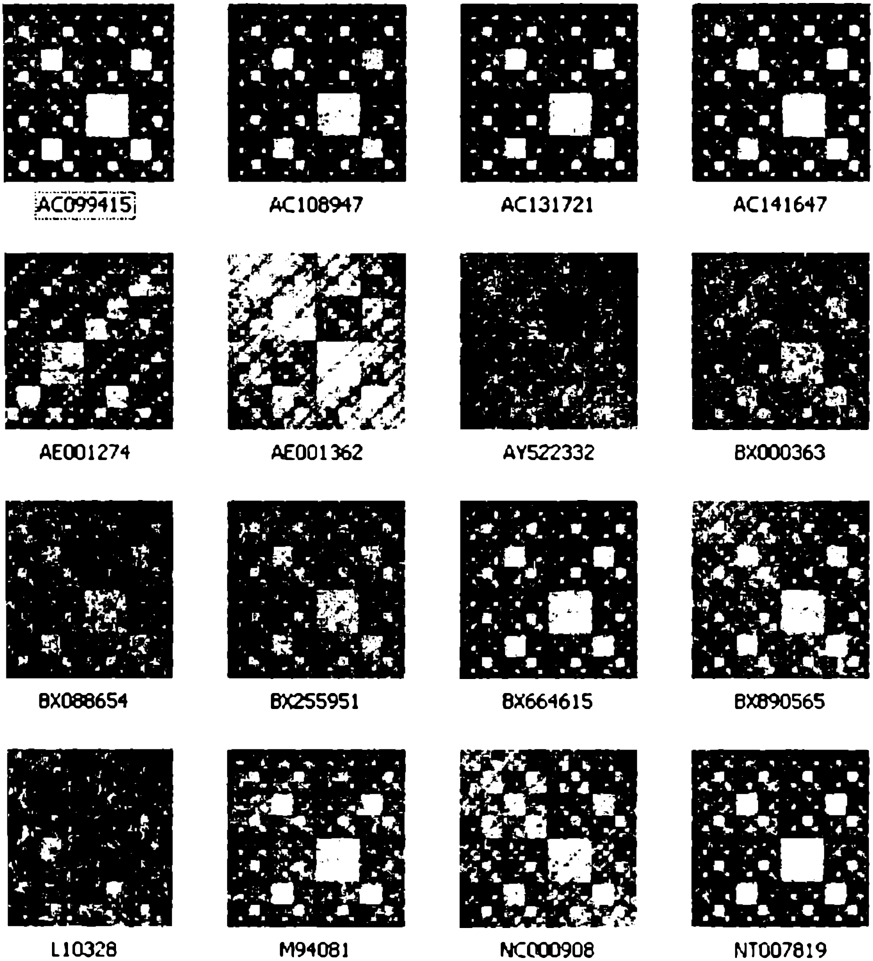

FIGURE 2. A few possible gene-patterns.

ℓemark. This approach, i.e., using chaos game representation, can be applied ›resenting a large class of symbolic sequences, not only to the genetic ones.

th random symbolic sequence (equal probabilities for each symbol) the result is ›nt. There is no pattern at all, the CGR-algorithm produces a square uniformly with dots. If the probabilities are not equal, the shape of the attractor is .nged, but the shading may be visible.

:an be also applied to the trajectories of a dynamical system. This requires lucing a finite coverings of the phase space, the corresponding encoding of traies into symbolic sequences, e.g., the symbolic dynamics of the logistic equation re 4.).

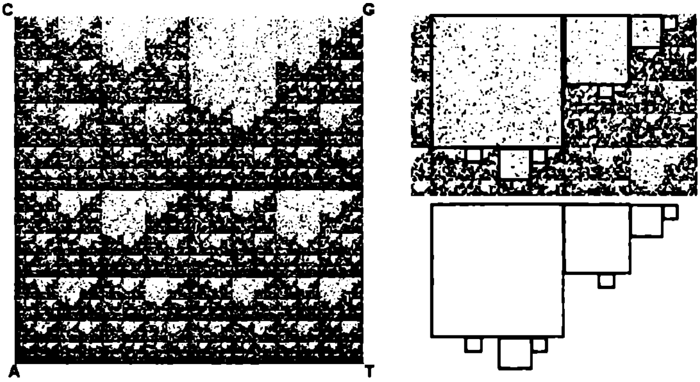

FIGURE 3. CGR for the DNA sequence (HUMHBB) according to Jeffrey's approach (on the left) and origin of 'scoops'.

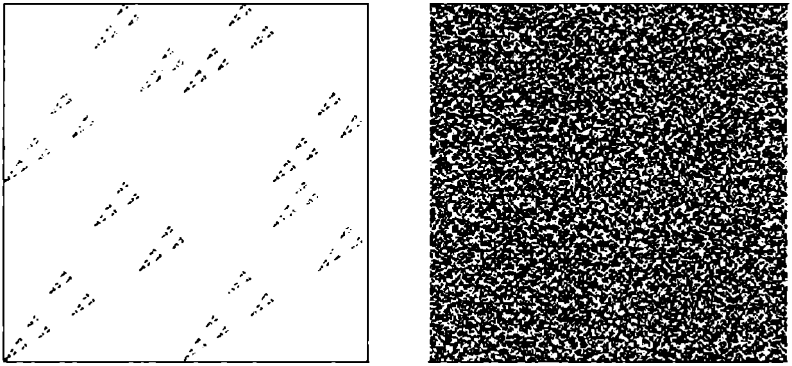

FIGURE 4. CGR for the symbolic dynamics generated by the logistic map (on the left) and for a random sequence.

We can consider as an example some financial data as well, e.g., closing prices of some stocks.

Not all the data can be represented as a string over an alphabet of four symbols, but we can convert it into such an alphabet using for example coarse-graining procedure.

4. TOOLS

When we consider the presence of complex self-similar geometrical structure of the CGR-gene patterns, the technique that immediately comes to mind is fractal analysis – analysis of self-similar sets.

The term self-similarity hardly needs an explanation. Self-similarity means that each piece of a set (however small) is identical to the whole after some rescaling. Self-similar structures appear in a variety of natural phenomena, but the most natural objects do not display this precise form. The range of magnification within which we see similar forms in nature is finite and a magnified view of one part is not precisely reproduce the whole object, but do not have the same qualitative appearance. Therefore, fractals can only be used as models for natural shapes.

When we think about fractals we usually perceive them as static objects. But this point of view tells us little about the evolution or origin of a given structure, because there are many phenomena in nature which can not be illustrated by sets. In other words, to talk about fractals while ignoring the dynamic processes which created them would be inadequate. The fractal box-counting dimension is the basic notion for describing structures that have a scaling symmetry, but it does not consider the distribution of points on the attractor (see Figure 5.). It was the main reason to extend the idea of self-similarity from sets to measures.

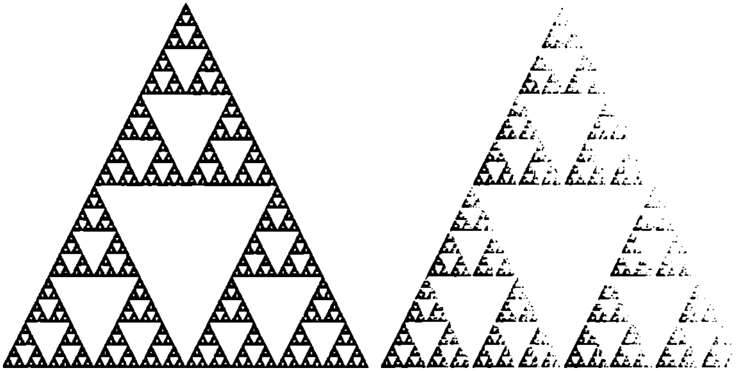

FIGURE 5. The fractal dimension is not sensitive enough to distinguish between the uniform and non-uniform Sierpiński triangle.

The concept of multifractals, self-similar measures, was introduced and described for the first time in 1970s by Mandelbrot in the context of fully developed turbulence [5]-[6]. Systems with multifractal structure are very important as mathematical models and highly diverse. For example, in dynamical systems we are interested not only in the shape of attractors but also how often a given region of the attractor is visited. Self-similar measures have been used to describe the turbulent flow of fluids, percolations, diffusion-limited aggregation systems (DLA systems), finance, and cosmic string theory. It is worth to emphasize that the multifractals are strongly related to thermodynamics and theory of probability. In this paper we show application of this theory to the analysis of DNA sequences.

In this section we define generalized dimensions and establish its principal properties. We introduce also multifractal formalism, following Evertsz and Mandelbrot [7]. In particular, in Subsection 4.2 we give a formula for multifractal spectrum.

4.1. **Generalized dimensions D_q.** Let X be a subset of $\mathbb{R}^d$ which is divided into d-dimensional cubes X_i of size l, $i = 1, \ldots, N$. We assume that μ is probabilistic measure on X, i.e., $\mu(X) = 1$. Let C be a non-empty bounded subset of X. For

$i = 1, \dots, N$ we define $C_i = C \cap X_i$, and we denote $\mu(C_i)$ by μ_i. Let N_l be the minimal number of boxes with length l that are required to cover C.

The generalized dimension D_q of C for the parameter q ($q \in \mathbb{R}$) is defined as

$$D_q = \frac{1}{q-1} \lim_{l \to 0} \frac{\ln I(q,l)}{\ln l}, \tag{1}$$

where

$$I(q,l) = \sum_{i=1}^{N_l} \mu_i^q, \quad q \neq 1.$$

Additionally, applying l'Hospital's rule, we define $D_1 = \lim_{q \to 1} D_q$.

There are two main reasons for the importance of the generalized dimensions. Firstly, D_q is designed to reflect not only the fractal geometry of the underlying objects, but also the dynamics which takes place in them. Secondly, using this definition we can readily find the well-known fractal dimensions for integer q as special cases, i.e.,

- D_0 the box-counting dimension (fractal, capacity dimension), $I(0,l) = N_l$,
- D_1 the information dimension,
- D_2 the correlation dimension.

Let us present a few of the main properties of the generalized dimensions: The generalized dimension D_q is defined for all real q and is a monotone decreasing function of q. As q in (1) varies, different subsets, which are associated with different scaling indices, become dominant.

What do the D_q tell us?

q can be treated as a "filter", namely:

- $q > 0$ highlights dense portions of the pattern,
- $q < 0$ highlights sparse portions of the set,
- when $q \to \infty$, D_q shows strongest clustering regions,
- when $q \to -\infty$, D_q shows least dense regions.

For uniform measure μ ($\mu_i = \frac{1}{N_l}$, $i = 1, \dots, N_l$), D_q does not depend on q, and simply equals box-counting dimension, as $\ln(I(q,l)) = (1-q) \ln N_l$.

A measure for which the D_q dimension varies with q is called multifractal measure. Only in the case of the well-known simple fractals, monofractals, a single dimension suffices, $D_q = const$ for all q.

4.2. **Multifractal spectrum.** The symbol $\propto$ indicates an asymptotic relation (scaling law):

$$\gamma \propto \beta^\eta \quad \equiv \quad \eta = \lim_{\beta \to 0} \frac{\ln \gamma}{\ln \beta};$$

The proportionality constant $c(\eta)$ ($\gamma = c(\eta)\beta^\eta$) can be weakly dependent on β:

$$\lim_{\beta \to 0} \frac{\ln c(\eta)}{\ln \beta} = 0.$$

In the notation of Subsection 4.1, we define the coarse Hölder exponent by the formula:

$$\alpha = \frac{\ln \mu_i}{\ln l}.$$

Note that for a multifractal α will be restricted to an interval $\alpha_{min} < \alpha < \alpha_{max}$ while for a fractal there will be an unique α (because for all $i = 1, \ldots, N_l$, $\mu_i = const$ and $N_l \propto l^{-D_0}$). To obtain the frequency distribution $f(\alpha)$, one has to evaluate for each value α the number $N_l(\alpha)$ of boxes of size l having a coarse Hölder exponent equal to α.

Let

$$f_l(\alpha) = -\frac{\ln N_l(\alpha)}{\ln l}.$$

As $l \to 0$, then $f_l(\alpha)$ tends to well-defined limit $f(\alpha)$. This definition means that, for each α, the number of boxes increases for decreasing l as $N_l(\alpha) \propto l^{-f(\alpha)}$. So multifractals are objects whose structure cannot be described by a single scaling behavior but we need all the spectrum of values. The exponent $f(\alpha)$, called the multifractal or singularity spectrum, is a continuous function of α. In the simplest cases, the graph of $f(\alpha)$ is an upsidedown bell shaped curve, which values could be interpreted as a fractal dimension of the subsets of boxes of size l having coarse Hölder exponent α in the limit $l \to 0$.

The function $f(\alpha)$ was firstly defined in 1986 by a group of physicists in their seminal paper [8].

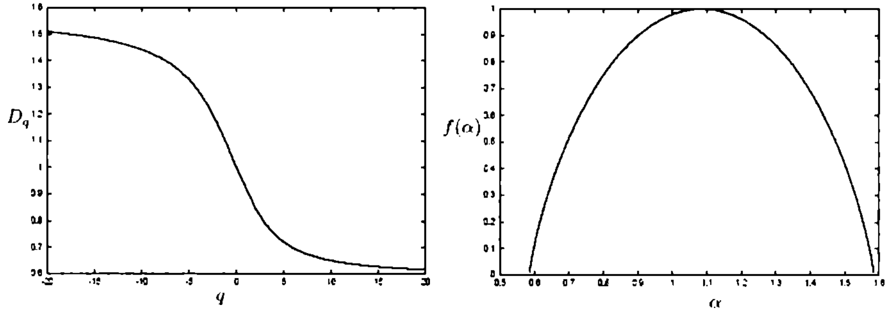

FIGURE 6. The typical shape of the graph of D_q and $f(\alpha)$.

4.3. The relation between D_q and $f(\alpha)$. We assume, that $f(\alpha)$ is differentiable. For a given value of q, let $\alpha(q)$ be such that

$$\frac{d}{d\alpha}[q\alpha - f(\alpha)]\Big|_{\alpha=\alpha(q)} = 0,$$

$$\frac{d^2}{d\alpha^2}[q\alpha - f(\alpha)]\Big|_{\alpha=\alpha(q)} > 0,$$

and, as a consequence

$$\frac{d}{d\alpha}f(\alpha)\Big|_{\alpha=\alpha(q)}=0,\quad \frac{d^2}{d\alpha^2}f(\alpha)\Big|_{\alpha=\alpha(q)}<0.$$

Introducing

$$\tau(q)=q\alpha(q)-f(\alpha(q)),$$

from (1) we obtain

$$D_q=\frac{\tau(q)}{q-1}.$$

Thus we may derive an explicit formula for the relation between D_q and $f(\alpha)$:

$$f(\alpha)=q(\alpha)-[q(\alpha)-1]D_{q(\alpha)}$$

Scaling function $\tau(q)$ is called correlation exponent or mass exponent of the qth order. So for the purpose of multifractal description we may use either $(f(\alpha(q)),\alpha(q))$ or $(\tau(q),q)$.

This means that $f(\alpha)$ can be computed from $\tau(q)$ and vice versa. The relation between $f(\alpha)$ and $\tau(q)$ is called a negative Legendre transform.

Thanks to its sensitivity for distributions of points on the attractor the multifractal analysis can be successfully applied in image analysis and object classification. There is a lot of algorithms and computer programs for the calculation of the $f(\alpha)$-spectrum of multifractal structures (e.g., [9]).

5. Practical application

Chaos Game Representation of DNA sequence provides a visual representation for initial analysis. The next step of research in this area is to study the differences and similarities between genomes using more sophisticated analysis of genetic data. Multifractal analysis presented above helps us to compare precisely different sequences. Hence it can be used for their classification.

The study of inter-species sequence comparisons is important for identifying elements in the genome because determining the sequence differences between species can provide insight into the distinct features of different organisms, help to define the generic basis for speciation and facilitate the characterization of mutational processes.

Many interesting results were obtained using similar methods, but in the case of a one dimensional model. It was applied to the problem of recognition of an organism based on fragments of their DNA sequences [10].

Based on this idea Yu et al. [11] developed a fast algorithm for deriving species phylogeny based on the measure representation of DNA and protein sequences. It helps to determine relationships, and reconstructing changes that must have occurred to create biologically relevant differences.

6. Conclusions

In recent years, more and more mathematicians have started working on the analysis and developing of various mathematical models for representing and describing DNA sequences. A variety of statistical modeling, numerical simulations and theoretical approaches were used, e.g.:

- data-driven pseudo-random walk in two- or four-dimensional space [12],
- entropic profiles [13],
- statistical analysis of time series (Lévy statistics) [14],
- statistical long-range correlation analysis [15]-[16],
- Hao's frame representation [17],
- linguistic analysis [18]-[20],
- symbolic dynamics and dynamical entropies [21]-[22].

In this paper we presented another approach. This is an interesting starting point of investigation, complementing more traditional approaches in analysis of DNA sequences. One of the next steps in this research is to examine the usefulness of these techniques for investigating DNA sequence structure.

There is no doubt that DNA has a significant role in organizing the development of an organism. The more we learn about genes, the more evident becomes the need for a good understanding of dynamic effects in biology – in growth, in development, in regulation of genetic networks, in ecosystems, and in the evolution.

References

[1] A. Lasota, M.C. Mackey, *Chaos, Fractals, and Noise. Stochastic Aspects of Dynamics*, Springer-Verlag, New York,1994.

[2] M. Barnsley, *Fractals Everywhere*, Academic Press, San Diego, 1988.

[3] H. Jürgens, H.O. Peitgen, D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer, New York, 1992.

[4] H. Joel Jeffrey, *Chaos game representation of gene structure*, Nucleic Acids Research 18, 2163, (1990).

[5] B.B. Mandelbrot, *Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier*, J. Fluid Mech. 62 (1974) 331

[6] B.B. Mandelbrot, *Multiplications aléatories itérées et distributions invariance par moyenne pondérée aléatorie I,II*, Comptes Rendus (Paris): 278A (1974) 289-292 and 355-358

[7] B.B. Mandelbrot, C.J.G. Evertsz *Multifractal Measures*, appendix B in [3], 921-953

[8] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, B.I. Shraiman: *Fractal measures and their singularities: The characterization of strange sets*, Phys. Rev. A 33, 1141-1151 (1986).

[9] A. Chhabra, R.V. Jensen, *Direct Determination of the $f(\alpha)$ Singularity Spectrum*, Phys. Rev. Lett. 62, 1327, (1989)

[10] V.V. Anh, K.S. Lau, Z.G. Yu, *Recognition of an organism from fragments of its complete genome*, Phys. Rev. E 66, 031910, (2002).

[11] Z.G. Yu, V.V. Anh, *Phylogenetic tree of prokaryotes based on complete genomes using fractal and correlation analyses*, Second Asia-Pacific Bioinformatics Conference (APBC 2004), Dunedin, New Zealand, (2004).

[12] C.L. Berthelsen, J.A. Glazier, M.H. Skolnick, *Global fractal dimension of human DNA sequences treated as pseudorandom walks*,Phys. Rev. A 45, 8902-8913 (1992).

[13] J.L. Oliver, P. Bernaola-Galván, J. Guerrero-Garcia, R. Román-Roldán *Entropic profiles of DNA Sequences Trough Chaos-game-derived Images*, J. Theor. Biol. 160, 457-470, (1993).

[14] N. Scafetta, V. Latora, P. Grigolini, *Lévy scaling: The diffusion entropy analysis applied to DNA sequences*, Phys. Rev. E 66, 031906, (2002).

[15] W. Li, T.G. Marr, K. Kaneko, *Understanding long-range correlation in DNA sequences*, Physica D 75, 392-416, (1994).

[16] H. Herzel, I. Große, *Measuring correlations in symbolic sequences*, Physica A 216, 518-542, (1995).

[17] P. Tino, *Multifractal properties of Hao's geometric repersentation of DNA sequences*, Physica A 480, 304 (3-4), (2002).

[18] R.N. Mantegna et al., *Linguistic Features of Noncoding DNA Sequences*, Phys. Rev. Lett. 73, 3169-3172, (1994).

[19] N.E. Israeloff, M. Kagalenko, K. Chan,*Can Zipf Distinguish Language from Noise in Noncoding DNA?* , Phys. Rev. Lett. 76, 1976, (1996).

[20] S. Bonhoeffer et al., *No Signs of Hidden Language in Noncoding DNA*, Phys. Rev. Lett. 76, 1977, (1996).

[21] H. Herzel, W. Ebeling, A.O. Schmitt, *Entropies of biosequences: The role of repeats*, Phys. Rev. E 50, 50615071, (1994).

[22] J. Freund, W. Ebeling, K. Rateitschak, *Self-similar sequences and universal scaling of dynamical entropies*, Phys. Rev. E 54, 5561-5566, (1996).