International Journal of
Molecular Sciences

MDPI

Article

Small-Angle Scattering and Multifractal Analysis of DNA Sequences

lEugen Mircea Anitas [1,2]

[1] Joint Institute for Nuclear Research, Dubna, Russia; anitas@theor.jinr.ru
[2] Horia Hulubei, National Institute of Physics and Nuclear Engineering, Bucharest-Magurele, Romania

Received: 06 June 2020; Accepted: 28 June 2020; Published: 30 June 2020

Abstract: The arrangement of A, C, G and T nucleotides in large DNA sequences of many prokaryotic and eukaryotic cells exhibit long-range correlations with fractal properties. Chaos game representation (CGR) of such DNA sequences, followed by a multifractal analysis, is a useful way to analyze the corresponding scaling properties. This approach provides a powerful visualization method to characterize their spatial inhomogeneity, and allows discrimination between mono- and multifractal distributions. However, in some cases, two different arbitrary point distributions, may generate indistinguishable multifractal spectra. By using a new model based on multiplicative deterministic cascades, here it is shown that small-angle scattering (SAS) formalism can be used to address such issue, and to extract additional structural information. It is shown that the box-counting dimension given by multifractal spectra can be recovered from the scattering exponent of SAS intensity in the fractal region. This approach is illustrated for point distributions of CGR data corresponding to *Escherichia coli*, *Phospholamban* and *Mouse mitochondrial* DNA, and it is shown that for the latter two cases, SAS allows extraction of the fractal iteration number and the scaling factor corresponding to "ACGT" square, or to recover the number of bases. The results are compared with a model based on multiplicative deterministic cascades, and respectively with one which takes into account the existence of forbidden sequences in DNA. This allows a classification of the DNA sequences in terms of random and deterministic fractals structures emerging in CGR.

Keywords: DNA sequences; multifractals; small-angle scattering; multiplicative cascades

1. Introduction

Understanding the correlations between DNA structure and its functions is one of the fundamental challenges in modern biology, with important implications in various biological processes, such as in replication or transcription [1]. Basically, eukaryotic DNA has a hierarchical organization in which the primary structure consists of a sequence of four nucleotides, i.e., adenine (A), thymine (T), guanine (G) and cytosine (C), arranged in two complimentary polynucleotide strands in the form of a double helix. In this configuration, each type of nucleotide on one strands, bonds with just one type of nucleotide on the other strand, according to base pairing rules: A with T, and C with G. At a higher level, DNA is tightly packed into nucleosomes (almost two turns of DNA wrapped around histone protein), separated by additional DNA fragments. Furthermore, the nucleosomes are arranged into a chromatin fiber, which forms loops, then chromatin domains, and finally chromosomes [2]. As such, the number of base pairs involved increases from about 10^2 bp (for short regions of DNA double helix), up to about 10^2 Mbp (for chromosomes).

It is argued that the primary structure contributes, through gene positions and transcriptional activity to the organization of chromatin at all scales, although the precise influence remains controversial [1]. Up to some extent, this may be linked to the existence of long-range correlations of nucleotide distributions [3–7]. The nature and origin of such correlations is intensively studied in the literature by performing various types of statistical analysis on DNA sequences [8–13]. One of the most accurate method in revealing the existence of power-law correlations with specific scale-invariance properties is the wavelet transform modulus maxima (WTMM) [13]. This method considers analyzing wavelets that make the wavelet transform microscope blind to low-frequency trends, and can reveal the hierarchy that governs the spatial distribution of multifractal measures [1].

More recently, with the advent of bioinformatics, numerical representation of DNA sequences has become an important approach in analyzing long-range correlations in big data sets as generated from high-throughput methods for sequencing. To this aim, a popular method, which is gaining an increased interest is the chaos game representation (CGR), introduced by Jeffrey in Ref. [14]. This is an iterative mapping technique which assigns to each nucleotide unique coordinates in 2D space, and it allows a visual representation of both local and global patterns in DNA sequences, revealing previously unknown structures [14]. One of the most important property, which makes CGR very useful in numerical encoding is that it gives a one-to-one representation, i.e., given a CGR point in the plane, one can trace back to the origin of the sequence, and therefore the original DNA sequence can always be reconstructed [15]. In addition, the 2D point distributions generated by CGR can be further analyzed using different methods, to extract additional structural information.

To this aim, one of the most common methods, used either standalone or in combination with other methods, is multifractal analysis, since it allows distinguishing between coding and non-coding sequences, or it can be used in phylogeny reconstruction and in investigating the clustering of protein structures. In particular, by using a CGR in which a DNA symbolic sequence is mapped onto a singular measure on the attractor of a particular iterated function system (IFS), the multifractal spectrum of the resulting measure is shown to be more sensitive for detecting dependence structures within DNA than the averaged contribution given by redundancy [16]. WTMM applied to multifractal analysis of CGR images show that the scale-invariance range of CGR edge can be extended to three orders of magnitude, and complete singularity spectra can be calculated [17]. By using CGR of protein sequences based on the detailed HP model, together with their multifractal and correlation analyses, a more precise phylogenetic tree of bacteria has been proposed [18]. CGR of randomly linked functional protein sequences used together with recurrent IFS helps to extract some biological functions of these proteins [19]. Multifractal detrended cross-correlation analysis of genome sequences using CGR shows the existence of multifractal nature and power-law correlation behavior between any pair of genome sequences [20]. A version of CGR, modified to visualize heteroplasmic mutations, used together with lacunarity analysis, reveals fractal properties of mitochondrial DNA sequences, and can quantitatively characterize Parkinson's disease [21].

However, as we shall see below, the multifractal spectra of different point distributions can be very similar, and additional analysis is required to discriminate between such arrangements, and which can provide new insights into their structure. To address this issue, a useful approach is to use the formalism underlying analysis of small-angle scattering (SAS) [22], largely used in molecular biology for investigating the structure of complex macromolecules. Physically, when X-rays are used, SAS is based on the electrostatic interaction of the electromagnetic wave with electrons, while in the case of neutrons, it is based on their interaction with the atomic nuclei (nuclear scattering), and in ferromagnetic materials, on interaction of neutron's magnetic moment with that of the atom (magnetic scattering). Theoretically, depending on the behavior of the SAS intensity, i.e., if we have a simple or a generalized power-law decay (i.e., a succession of maxima and minima superimposed on a simple power-law decay), one can use models specific to random (statistically self-similar) [23,24], and respectively to deterministic (exact self-similar) fractals [25].

A first step in this direction has been recently performed in Ref. [26], where SAS technique has been applied to study the structural properties of various fractal patterns generated by CGR, and an analytic expression of the corresponding SAS intensity has been derived for Sierpiński triangles. The SAS intensity from CGR has been calculated numerically by using a simplified version of the Debye formula [27]. The good agreement between numerical data and the theoretical model indicates that SAS technique can be successfully used to obtain the main structural characteristics, including the overall fractal size, scaling factor, fractal dimension, or the number of units composing the fractal, corresponding to CGRs of DNA sequences, and for which analytic expressions of intensity are not available or are hard to be derived. Such information may be useful to distinguish functional regions of DNA sequences or solving issues related to the classification of organisms. In particular, knowledge of the fractal dimension may provide information about the percentage content of coding regions, i.e., the higher the fractal dimension, the higher the coding percent [28]. In addition, the scaling factor may provide information about the relative sizes of coding regions, and which, in turn, describes the degree of self-similarity for different organisms, since it is related to the value of the fractal dimension [26].

Here, the results obtained in Ref. [26] are extended by performing a combined analysis of CGR using SAS and multifractal analysis of *Escherichia coli*, *Phospholamban* and *Mouse mitochondrial* DNA experimental data. The complementarity of the two methods is illustrated first on a model based on multiplicative deterministic cascades. A model in which both approaches can be hardly used to distinguish between various patterns is also provided, and is based on the existence of forbidden sequences in DNA. Furthermore, the two models are used as benchmarks against which the structures of the experimental data are compared in terms of the interplay between mass and surface-like fractals structures emerging in CGR. As such, it is shown that for *Escherichia coli* the structure resemble closer a random fractal, while in the case of *Phospholamban* and *Mouse mitochondrial* DNA, the corresponding structure is closer to deterministic ones.

2. Theoretical Background

2.1. Iterated Function Systems and Chaos Game Representation of DNA Sequences

The theoretical framework provided by iterated function systems (IFS) is very useful for classification and description of fractals. Mathematically, a (hyperbolic) IFS is given by a complete metric space $(\mathbf{X}, d)$ together with a finite set of contraction mappings $w_n : \mathbf{X} \to \mathbf{X}$, with respective contractivity factors $s_n, n = 1, 2, \cdots, N$ [29]. Using a shorthand notation, an IFS is $\{\mathbf{X};\, w_n, n = 1, 2, \cdots, N\}$ and $s = \max\{s_n, n = 1, 2, \cdots, N\}$. Please note that a transformation $f : \mathbf{X} \to \mathbf{X}$ on a metric space $(\mathbf{X}, d)$ is a contraction mapping if there is a constant (contractivity factor) $0 \leq s < 1$ such that $d\,(f(x), f(y)) \leq s \cdot d(x, y) \quad \forall x,\, y \in \mathbf{X}$.

If one considers an IFS with contractivity factor s, and $(\mathcal{H}(\mathbf{X}), h(d))$ the space of nonempty compact subsets with the Hausdorff metric $h(d)$, then the transformation $W : \mathcal{H}(\mathbf{X}) \to \mathcal{H}(\mathbf{X})$ defined by

$$W(B) = \cup_{n=1}^{N} w_n(B),\ \forall B \in \mathcal{H}(\mathbf{X}), \tag{1}$$

is a contraction mapping on the complete metric space $(\mathcal{H}(\mathbf{X}), h(d))$ with contractivity factor s [29], i.e.,

$$h(W(B), W(C)) \leq s \cdot h(B, C)\ \forall B,\ C \in \mathcal{H}(\mathbf{X}). \tag{2}$$

Its unique fixed point, $A \in \mathcal{H}(\mathbf{X})$ obeys $A = \cup_{n=1}^{N} w_n(A)$, is given by $A = \lim_{m\to\infty} W^{\circ m}(B)$ for any $B \in \mathcal{H}(\mathbf{X})$, and is called the *attractor* of the IFS [29].

Here, for rendering pictures of attractors we play chaos game using the random iteration algorithm. As such, one start by assigning the probability $p_n > 0$ to w_n for $n = 1, 2, \cdots, N$, where $\sum_{n=1}^{N} p_n = 1$. Then, a point $x_0 \in \mathbf{X}$ is chosen, and then recursively, next points are obtained according to:

$$x_k \in \{w_1(x_{k-1}), w_2(x_{k-1}) \cdots, w_N(x_{k-1})\}\,. \tag{3}$$

The probability of the event $x_k = w_n(x_{k-1})$ is p_n, and $k = 1, 2, \cdots$. This generates the sequence $\{x_k :\ k = 0, 1, \cdots\} \subset \mathbf{X}$ which converges to the attractor of IFS.

When the chaos game is played with 4 points (i.e., $N = 4$), the matrix representation of the IFS of affine maps is:

$$w_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \mathrm{a}_i & \mathrm{b}_i \\ \mathrm{c}_i & \mathrm{d}_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \mathrm{e}_i \\ \mathrm{f}_i \end{bmatrix} \tag{4}$$

where the coefficients $\mathrm{a}_i, \mathrm{b}_i, \mathrm{c}_i, \mathrm{d}_i, \mathrm{e}_i, \mathrm{f}_i$ with $i = 1, 2, 3, 4$. For a uniformly filled square, they are given in Table 1.

To display visually the underlying structure of the attractor resulting from a string of four letters, Jeffrey proposed [14] to construct a square with vertices "A", "C", "G" and "T", and to control the chaos game, not with the probabilities p_i (see Table 1), but with the DNA sequence. In particular, to choose the next point, one could use the next base in the DNA sequence. For a sequence of three letters "CGT", the algorithm translates as follows: first, plot "C" at half distance between the center of the ACGT square and "C" vertex, then, the next base "G" is plotted at half distance between the previous point and "G" vertex, and finally, the last bases "T" is plotted at half distance between the previous point and the "T" vertex.

Table 1. The coefficients $\mathrm{a}_i, \mathrm{b}_i, \mathrm{c}_i, \mathrm{d}_i, \mathrm{e}_i, \mathrm{f}_i$ of the affine transforms w_n (first column) in Equation (4), and the probabilities p_i (last column) associated with each affine map.

w	a	b	c	d	e	f	p
1	1/2	0	0	1/2	0	0	1/4
2	1/2	0	0	1/2	0	1/2	1/4
3	1/2	0	0	1/2	1/2	0	1/4
4	1/2	0	0	1/2	1/2	1/2	1/4

2.2. Fractals and Multifractals

Mathematically, description of fractals is based on concepts from measures theory [30], and a rigorous definition has been introduced by Hausdorff in Ref. [31]. This involves a subset S of the n-dimensional Euclidean space, and let us consider that $\{\mathrm{C}_i\}$ is a cover of S with $c_i = \mathrm{diam}\,(\mathrm{C}_i) \leq s$ and $s \in$ S. Then, the Hausdorff measure $m^\alpha\,(\mathrm{S})$ is given by taking the infimum over all possible coverings, i.e.,

$$m^\alpha\,(\mathrm{S}) = \lim_{s\to 0} \inf_{\{\mathrm{C}_i\}} \sum_i c_i^\alpha, \quad \text{with} \quad \alpha \in \mathbb{R}^+, \tag{5}$$

and thus, the fractal dimension D can be written as:

$$D \equiv \inf\{\alpha : m^\alpha\,(S) = 0\} = \sup\{\alpha : m^\alpha\,(S) = +\infty\}. \tag{6}$$

This corresponds to the value of α for which the Hausdorff measure changes from zero to infinity. When $\alpha = D$, $m^\alpha\,(S)$ can have arbitrary values within this range.

However, for most practical purposes, it is very difficult to use Equation (6) for determination of the fractal dimension. To avoid this issue, another approach is to determine the variation of the fractal measure M, inside a sphere of dimension n and radius r centered on the fractal. For fractal systems, one can write a mass–radius relation of the type [32]:

$$M\,(r) = A\,(r)\,r^D, \tag{7}$$

where $\lim_{r\to\infty} \log A\,(r)\,/\,\log r \to 0$. The above equation plays a central role in development of the concepts used thereafter, since it allows us to describe mass and surface fractals, and provide analytical expression for fractal dimensions for a large class of fractals.

Thus, let us consider further a fractal of size L composed of balls of size a. Then, the number of balls enclosed by the imaginary sphere of radius r with a ball in the center, is given by [32]:

$$N\left(r\right) \propto \left(r/a\right)^{D} \propto r^{D}, \tag{8}$$

with $l \lesssim r \lesssim L$. If the fractal is a line, one has $D = 1$, for a smooth surface, $D = 2$, while for a regular Euclidean 3D object, $D = 3$.

For fractals with a scaling factor β_{s}, one can use the property that at first iteration the fractal consists of k copies of itself, each of size $\beta_{\mathrm{s}} L$, and write that [32]:

$$M\left(L\right) = kM\left(\beta_{\mathrm{s}} L\right). \tag{9}$$

Then, by using Equation (7), one obtains:

$$k\beta_{\mathrm{s}}^{D_{\mathrm{m}}} = 1, \tag{10}$$

which can be used to obtain the fractal dimension D. For fractals with multiple scaling factors β_{si} and k_i copies with $i = 1, \cdots, n$, Equation (9) is rewritten as:

$$\sum_{i=1}^{n} k_i \beta_{si}^{D_{\mathrm{m}}} = 1. \tag{11}$$

Similarly, for surface fractals one can write:

$$S\left(r\right) = S_0 r^{2-D_{\mathrm{s}}}, \tag{12}$$

where $S(r)$ represents the area between the boundary of the (rough) surface and the envelope of all spheres of radius r centered on the boundary. Here, S_0 is a constant, which is the surface area itself for a smooth surface, i.e., when $D_{\mathrm{s}} = 2$.

The fractal dimensions D which appear in the above relations, are equivalent to the box-counting dimension and describe fractals with a single scaling factor. However, when several scaling factors are present, a more detailed description is required, and can be achieved by using the multifractal formalism [33,34], where one considers an object S covered by a grid of boxes $B_i(l)$ of size l. By considering that the measure determined by the probability of hitting the object in the box B_i is $\mu(B)$, the number of covered boxes N at resolution l is $N \propto 1/l^2$, and thus one can write [35]:

$$Z_s(l) = \sum_{i=1}^{N} p_i^s(l), \tag{13}$$

where Z is called the "partition function" and has a power-law behavior when $l \to 0$ and $N \to \infty$, so that $Z_s \propto l^{D_s(s-1)}$. Here, i denotes each individual box, and $p_i = \mu(B)$ with $\sum_{i=1}^{N} p_i = 1$, are the hitting probabilities. Then, the generalized dimension spectrum is given by [35]:

$$D_s \equiv \frac{1}{s-1} \lim_{l \to 0} \frac{\ln Z_s(l)}{\ln l}. \tag{14}$$

By considering the ratio $p_i \equiv N_i(l)/N$, which gives the relative weight of the i-th box, one can write:

$$D_s = \frac{1}{s-1} \lim_{l \to 0} \frac{\ln \sum_{i=1}^{N} p_i^s(l)}{\ln l}. \tag{15}$$

The generalized dimension spectrum is a monotonically decreasing function, with horizontal asymptotes at $\alpha_{max} = \lim_{q \to -\infty} D_s$ and $\alpha_{min} = \lim_{q \to \infty} D_s$. Their values can be used to describe the heterogeneity, i.e., if $\alpha_{max} \neq \alpha_{min}$ the fractal is heterogeneous (multifractal), and homogeneous otherwise.

An equivalent representation of generalized dimension spectrum is provided by $f(\alpha)$ spectrum, which gives a mathematically precise and intuitive description of the multifractal measure in terms of interwoven sets, with singularity strength α, and Hausdorff dimension $f(\alpha)$. Without going into the details of derivation here, it can be shown that the generalized dimension and $f(\alpha)$ spectra are related through a Legendre transform [36], i.e.,

$$f(\alpha) = s\alpha(s) - \tau(s), \tag{16}$$

where $\alpha(s) = \mathrm{d}\tau(s)/\mathrm{d}s$ and $\tau(s) = (s-1)D_s$.

In the remaining of the paper, Equations (15) and (16) are used to characterize the multifractal properties of point distributions generated by CGR described in the previous section. To illustrate the method, analytically solvable models given by multiplicative deterministic cascades are first analyzed in Section 3.1.1.

2.3. Small-Angle Scattering

One considers here a two-phase approximation in which microscopic scattering objects with scattering length b_j have the scattering length density (SLD) $\rho_{\mathrm{s}}(\boldsymbol{r}) = \sum_j b_j \delta(\boldsymbol{r} - \boldsymbol{c}_j)$, where $\boldsymbol{c}_j$ are the position vectors of the objects. The differential elastic cross section is defined by $\mathrm{d}\sigma/\mathrm{d}\Omega = |A_t(\boldsymbol{q})|^2$, where $\boldsymbol{q}$ is the scattering vector. For a three-dimensional object $A_t(\boldsymbol{q}) = \int_{V'} \rho_{\mathrm{s}}(\boldsymbol{r}) \exp(i\boldsymbol{q}\cdot\boldsymbol{r})\mathrm{d}\boldsymbol{r}$ is the total scattering amplitude and $V^{'}$ is the irradiated volume. When the objects are embedded in a solid matrix of SLD ρ_0, then the scattering contrast is defined by $\Delta\rho = \rho - \rho_0$, and the total scattering intensity can be written as [22]:

$$I(q) \equiv \frac{1}{V'}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = c\,|\Delta\rho|^2\,V^2\left\langle|F(\boldsymbol{q})|^2\right\rangle, \tag{17}$$

where c is the concentration of objects, V is their volume, and $F(\boldsymbol{q}) \equiv (1/V)\int_V \exp(-i\boldsymbol{q}\cdot\boldsymbol{r})\mathrm{d}\boldsymbol{r}$ is the normalized form factor, with $F(0) = 1$, and the symbol $\langle\cdots\rangle$ denotes the ensemble averaging over all orientations. In what follows, since the object are two-dimensional ACGT squares, the volume V in Equation (17) shall be replaced by the corresponding surface area. Please note that the above averaging procedure allows the rotation of the squares in three-dimensional space, with equal probability.

Since the positions of the points are generated by CGR, we can start with the Debye formula to calculate the scattering intensity, and thus Equation (17) can be written as [27]:

$$I(q) = NI_{\mathrm{s}}(q) + 2F_{\mathrm{s}}(q)^2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\sin qr_{ij}}{qr_{ij}}, \tag{18}$$

where $I_{\mathrm{s}}(q) = 1$ is the intensity scattered by each point, q is the magnitude of the scattering vector $\boldsymbol{q}$, and r_{ij} is the distance between arbitrary points i and j. When the number of points exceeds a few thousand, the computation of the term $\sin(qr_{ij})/(qr_{ij})$ is very time consuming, and thus it is handled via a pair-distance histogram $g(r)$, with a bin-width commensurate with the experimental resolution [37]. Therefore Equation (18) becomes:

$$I^D(q) = N + 2\sum_{i=1}^{M_{\mathrm{bins}}} g(r_i)\frac{\sin qr_i}{qr_i}, \tag{19}$$

where M_{bins} is the number of bins, and $g(r_i)$ is the pair-distance histogram at pair-distance r_i. The latter quantity is calculated from the positions of points inside the ACGT square.

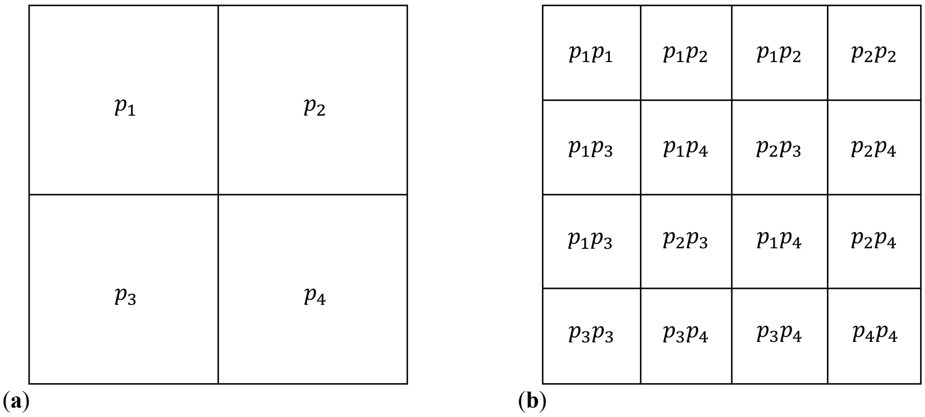
3. Results and Discussion

3.1. Analysis of Theoretical Models

To illustrate the complementarity between SAS and multifractal techniques in analyzing point distributions, one considers a model of multiplicative deterministic cascades. This is followed by a second model, based on forbidden sequences in DNA, and which shows that both the SAS and multifractal spectra can hardly be used to differentiate between various structures.

3.1.1. Multiplicative Deterministic Cascades

The multiplicative deterministic cascade model is constructed by dividing a square into for equal squares. To each of the subsquare, one assigns the probabilities $p_i \in [0,1]$, with $i = 1,2,3,4$. This is called the first iteration, and it is denoted $n = 1$ (Figure 1a). Then, at second iteration ($n = 2$), each of the four subsquares is further divided in four squares, and to each of them are assigned probabilities given by the Kronecker product of the matrices with the elements p_i placed as in Figure 1a. The results are shown in Figure 1b. At third iteration, one perform a similar division of squares, and to each of them one assigns the probabilities given by the Kronecker product of the matrices with as elements the probabilities from iteration $n = 1$, and respectively from $n = 2$. The multiplicative cascade model is obtained in the limit of high number of iteration, and the distribution of square values therefore depends on the initial choices of probabilities p_i. This model is similar to the model for the displacement of a viscous fluid by a nonviscous fluid in a porous medium, developed in Ref. [38]. However, in this reference, at $n = 2$ the probabilities p_i associated with each such division are multiplied in random order by probabilities p_i. The same random assignment of probabilities is kept for subsequent iterations, thus leading to a multiplicative random cascade model.



p_1	p_2
p_3	p_4

(**a**)

p_1p_1	p_1p_2	p_1p_2	p_2p_2
p_1p_3	p_1p_4	p_2p_3	p_2p_4
p_1p_3	p_2p_3	p_1p_4	p_2p_4
p_3p_3	p_3p_4	p_3p_4	p_4p_4

(**b**)

Figure 1. The initiator (**a**) and generator (**b**) of the multifractal lattice. The quantities p_1, p_2, p_3 and p_4 represent the probabilities associated with their corresponding cells in the lattice.

Figure 2 shows several realizations of the multiplicative deterministic cascades model, in which the probabilities p_i are given in Table 2. The common feature is the presence of a deterministic pattern, in which exact copies of the fractal appear at various scales. The structure of the pattern is very rich, and includes stripes of lines (model M2), single scale Sierpiński gaskets (model M4), or a clear superposition of Sierpiński gaskets (model M4).

Table 2. The probabilities p_i used to generate the models M1, M2, M3, M4, M5 and M6 in Figure 2.

Model	p_1	p_2	p_3	p_4
M1	1	1	1	0.5
M2	1	1	0.5	0.5
M3	1	0.75	0.75	0.75
M4	1	1	1	0
M5	1	1	0.5	0.25
M6	0.5	1	1	0.25
M7	1	1	1	1

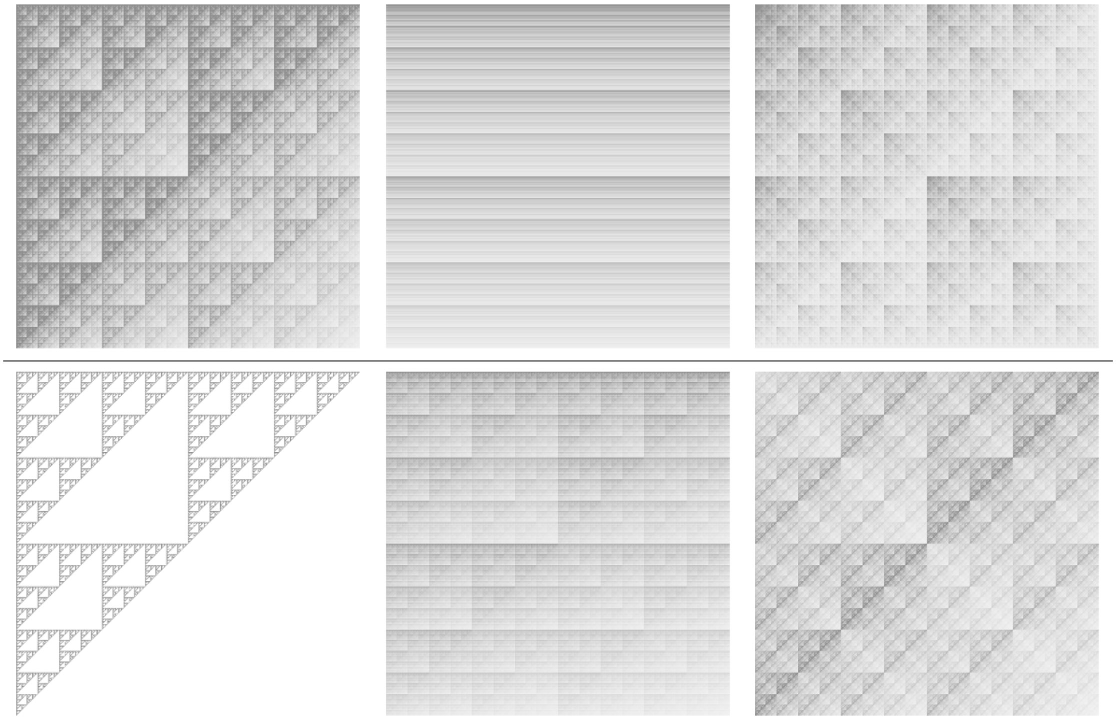


Figure 2. Six configurations M1, M2, M3, M4, M5, and M6 of the multiplicative deterministic cascade model, at iteration number $n = 11$. Upper part: M1, M2, M3 (from left to right). Lower part: M4, M5, M6 (from left to right). See Table 1 for the corresponding probabilities p_1, p_2, p_3 and p_4. The model M7 is not shown here, since it corresponds to a uniformly filled square.

For these models of multiplicative cascades, it can be shown that the generalized dimension spectrum has an analytic expression given by [39]:

$$D_s = \frac{1}{1-s} \log_2 \left(f_1^q + f_2^q + f_3^q + f_4^q \right), \tag{20}$$

where $f_j = p_j / \sum_{i=1}^{4} p_i$, and $j = 1, 2, 3, 4$.

The corresponding dimension spectra are presented in Figure 3a. The results clearly show that the spectra of models M1, M2, M3, M5 and M6 have a decreasing behavior, indicating that the corresponding structures are multifractals. However, for model M4 the spectrum is constant and indicates a simple fractal structure (Sierpiński gasket) with fractal dimension about 1.58, as expected. Also, in the case of model M7 (uniform square) the fractal dimension 2 is recovered. Please note that for models M5 and M6 the dimension spectra coincide for all s. They are also the most heterogeneous structures, since the difference $\alpha_{\text{min}} - \alpha_{\text{max}} \simeq 3.45 - 1.48 = 1.97$ is maximum, as compared to the other models.

The $f(\alpha)$ spectra are calculated by using Equation (16), and the results are presented in Figure 3b. Generally, these spectra are convex functions with a single maximum at $\alpha = \alpha_0$. This gives the most frequent value of the scaling indices α, and appears at $s = 0$, where $f(\alpha) = D_0$ [39]. Also, the scaling indices take vales in the finite range $[\alpha_{\min}, \alpha_{\max}]$. The smaller the length of this interval, the more homogeneous the structure. In particular, for models M4 and M7, the spectra degenerate into a single point at $f(\alpha) \simeq 1.58$ and respectively $f(\alpha) = 2$. Please note that for models M1, M3, M5 and M6, the $f(\alpha)$ spectra are asymmetric with respect to their maxima. Lower values of $f(\alpha)$ for $s < 0$ (models M1, M5 and M6) indicate a higher influence of the lowest values of the fractal measure in the spectral complexity. Similarly, the appearance of lower values of $f(\alpha)$ for $q > 0$ (model M3) indicate a higher influence of the lowest values of the fractal measure [40].

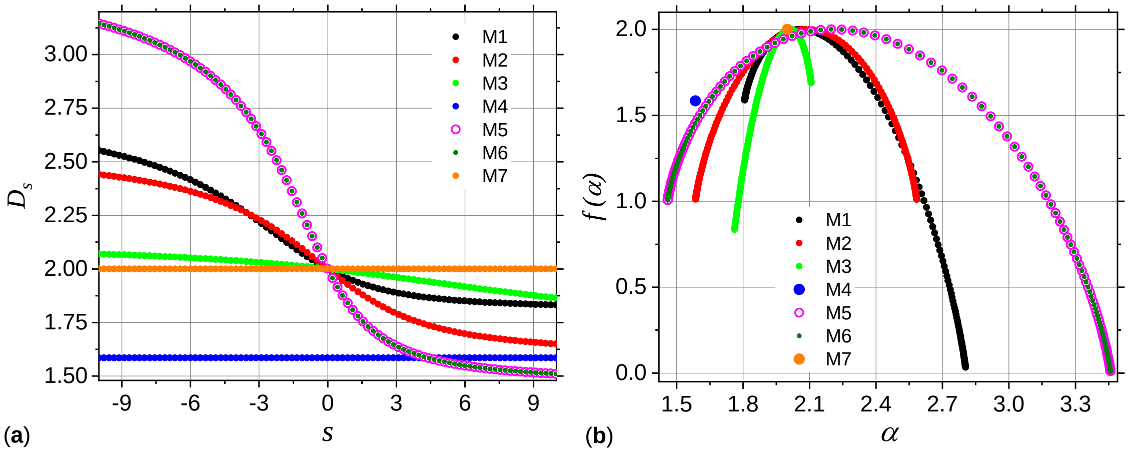


Figure 3. Generalized dimensions D_s (**a**) and $f(\alpha)$ spectra (**b**) for the configurations M1, M2, M3, M4, M5, M6 and M7. See Table 1 for the corresponding probabilities p_1, p_2, p_3 and p_4.

In the following one obtains the SAS from models M5 and M6. To this aim, the coordinates of the points needs to be extracted from Figure 2 and used as input in Debye equation (Equation (19)). However, a general property of such images is that the distribution of grey levels is concentrated within a short range, and thus image binarization at various thresholds becomes impracticable. Figure 4 shows this situation for models M5 and M6, where the number of pixels are mostly concentrated in the 0.7 ÷ 0.95 range. Despite this local pixel concentration, their distribution is different, and this indicate that SAS technique can distinguish between the two structures.

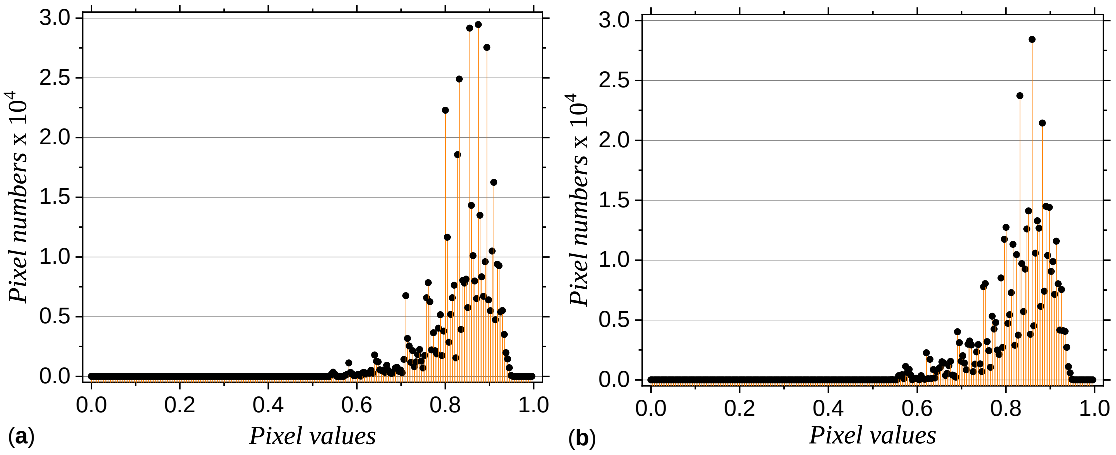


Figure 4. Histograms of the pixel levels in the images corresponding to models M5 (**a**) and M6 (**b**), shown in Figure 2 lower part.

Therefore, an equalization of histograms such that they span the entire range from 0 (black) to 1 (white) needs first to be performed. This is done in Figure 5, which shows both the transformed image and the corresponding pixel distribution (upper part—model M5, and lower part—model M6).

The histogram distribution of the models show a periodicity specific to multiplicative deterministic cascade structures, since the grey levels alternate in a regular fashion. The heights of the single maxima (for model M5) and of the groups of maxima (for model M6) follow a parabola-like behavior over the entire range of pixel values. These features now allows the analysis of the transformed images at various thresholds t. Figure 6 shows the corresponding binarized images for models M5 (upper part) and M6 (lower part) at thresholds 0.2, 0.4, 0.6 and 0.8 (from left to right).

The structure factor of the point distributions in Figure 6 is calculated by using Equation (19) and the results are shown in Figure 7. The main feature of the scattering curves in all cases is the presence of three structural regimes: $S(q) \propto q^0$ at $ql \lesssim 2\pi$ (Guinier region), $S(q) \propto q^{-D}$, with $D \simeq 2$ at $2\pi \lesssim ql \lesssim 2\pi l_0$ (fractal region), and $S(q) \simeq 1/k$ at $2\pi l_0 \lesssim ql$ (asymptotic region). Here, l is the overall fractal size (in pixels), k is the number of points composing the fractal, and l_0 is the pixel size. The length of the Guinier region provides information about the overall fractal size and they are very similar at each threshold, since the dimensions of squares are the same. In the fractal region one can see a generalized power-law decay (GPLD), i.e., the presence of a succession of maxima and minima superimposed on a simple power-law decay. A GPLD is characteristic to deterministic fractals [25] and can be used, as we shall see below, in extracting the fractal iteration number and the scaling factor.

The differences between scattering curves for models M5 and M6 are clearly visible at smaller values of the thresholds, and less visible at higher values, in particular, at $t = 0.8$ the differences are practically indistinguishable. This is because by increasing t the distribution of points becomes more uniform (see last column in Figure 6), and the deterministic character cease to dominate. Consequently, the maxima and minima are smoothed to the extent that the resulting curves coincide. Also, the scattering exponent $D \simeq 2$ is kept in all cases ant it coincides with the numerical values of D_0 given by multifractal spectra (see Figure 3). Finally, in the asymptotic region one can see that the total number of pixels k increases with t (since the ratio $1/k$ decreases), and this can be used to obtain the total number of points present in the fractal.

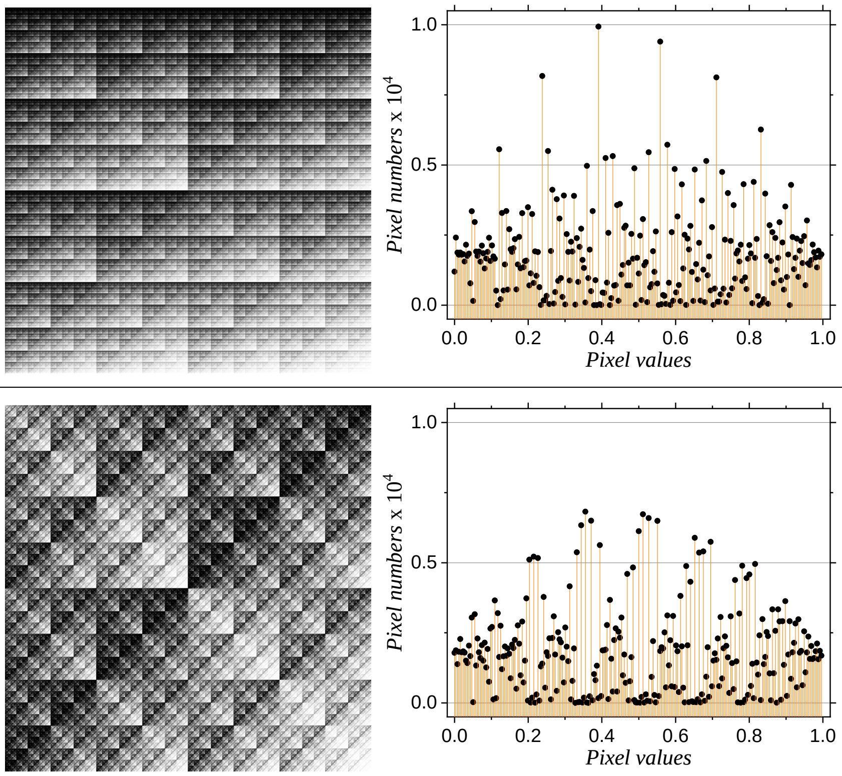


Figure 5. Images with equalized histograms for configurations M5 (upper-left corner) and M6 (low-left corner). Upper- and lower-right corers: the corresponding histograms.

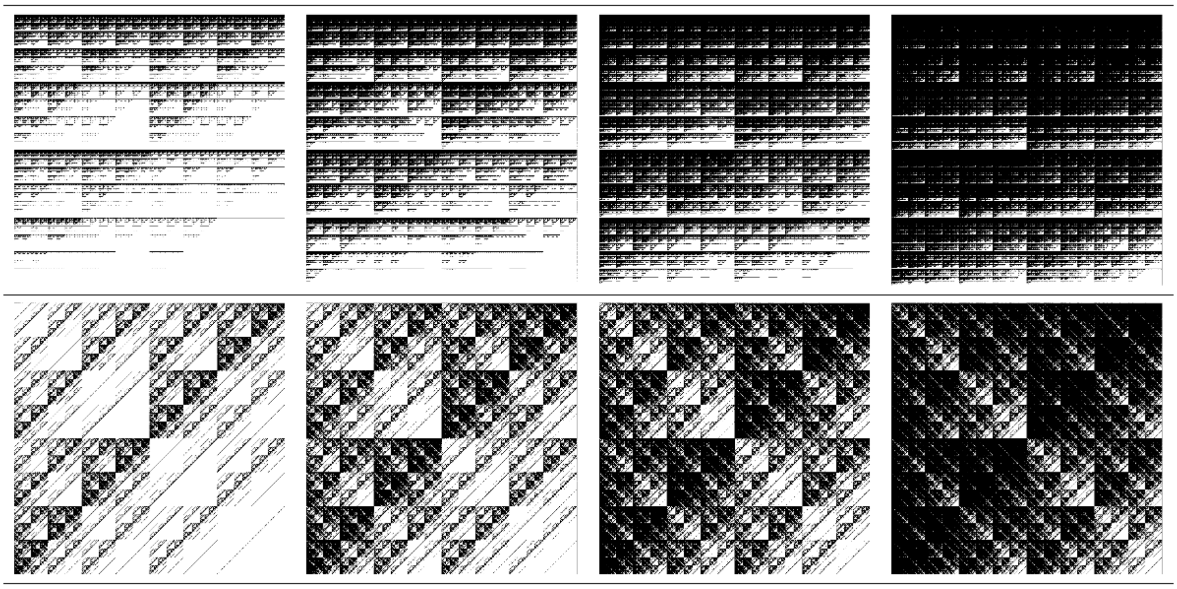


Figure 6. Binarized images of configurations M5 (upper-row) and M6 (lower-row) at different thresholds t. Columns from left to right: $t = 0.2$, $t = 0.4$, $t = 0.6$, $t = 0.8$.

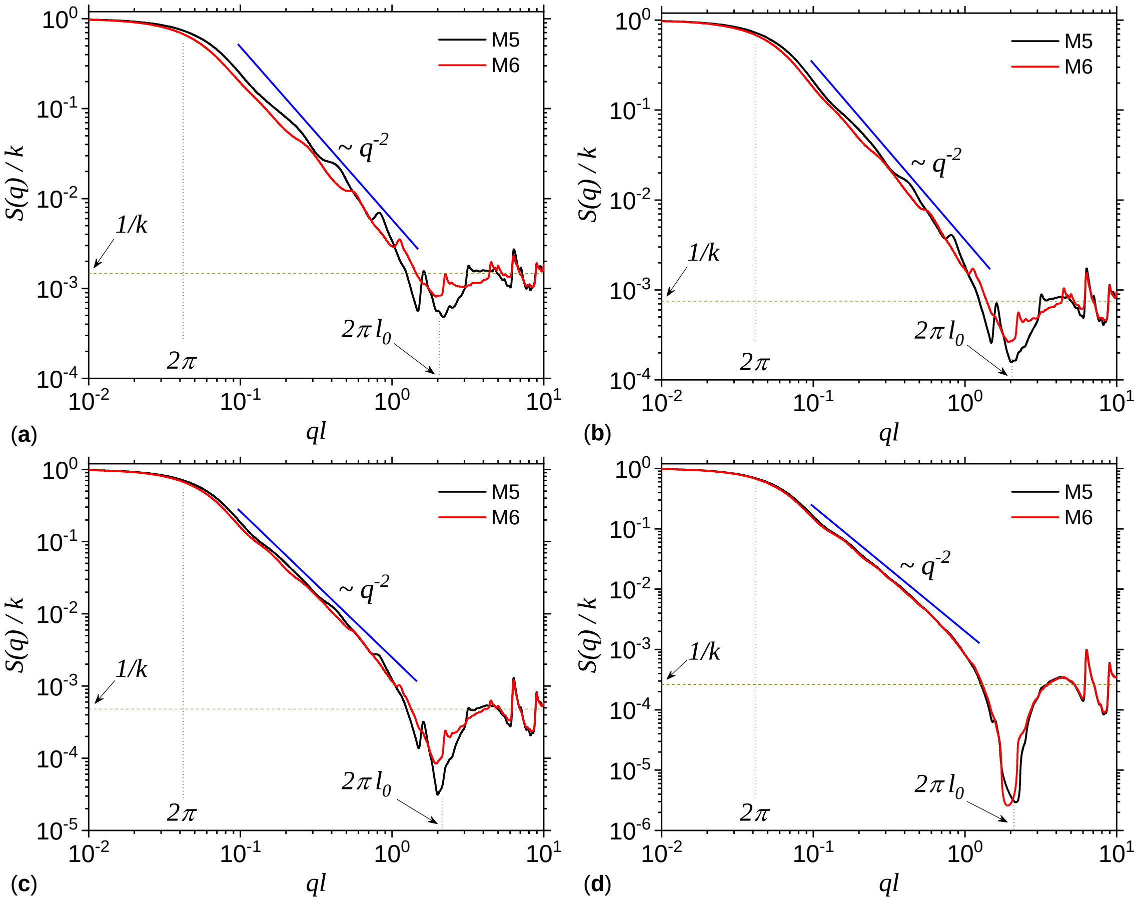


Figure 7. Structure factor (Equation (19)) corresponding to configurations M5 and M6, at various thresholds t. (**a**) $t = 0.2$. (**b**) $t = 0.4$. (**c**) $t = 0.6$. (**d**) $t = 0.8$. Here, l is the overall size of the fractal in pixels, k is the number of basic units composing the fractal, and l_0 is the pixel size.

3.1.2. Missing Sequences Models

Once a complete genome is available, a natural question which arise is whether there are subsequences of short strings of letters “A”, “C”, “G” or “T” that are missing, since this may reveal some biological meanings such as evolutionary relatedness of species [41]. To this aim, important steps have been performed in Ref. [42], where sequences are transformed into portraits, and thus the missing sequences can be visualized. Although this method has also the advantage of providing analytic expressions of the fractal dimensions of the patterns emerging from this visualization scheme [43], here CGR shall be used since in the case of missing of certain types of string composed on 2-contiguous letters, the resulting pattern has an obvious fractal structure (see Figure 8), while the fractal dimension can be extracted from either SAS or multifractal spectra, as discussed in the previous section. Moreover, by establishing relations between DNA sequences with missing subsequences, and the generalized Cantor sets, such as those presented in Ref. [44], the possibility of using SAS technique in analyzing structures with missing subsequences can be greatly extended, since theoretical expressions for SAS intensities are already available for some classes of generalized Cantor fractals [45].

In the following, one considers a random sequence of length 8×10^4 consisting from the four letters “A”, “C”, “G” or “T”, but that has never “A” followed by “T” (model TR12), “A” followed by “G” (model TR13), or “G” followed by “G” (model TR33). The corresponding structures obtained from CGR are shown in Figure 8. The resulting patterns has fractal properties in which the complementary structure, i.e., white regions, have various geometrical shapes (model dependent) and with sizes following a power-law distribution. The simplest complementary structure is shown in Figure 8 (middle part), and which consists from one square of edge length $1/8$, 4^1 squares of edge lengths $1/16$, 4^2 squares of edge lengths $1/32$ and so on. Such a distribution is specific to surface-like fractals and has important implications in behavior of the corresponding scattering curves. In particular, when the associated measure is multifractal, then the SAS intensity shows a not only a single mass fractal region, as in Figure 7, but a *succession* of mass fractal regime followed by a surface fractal one [46].

Since for all three models, both the number of complementary regions forming the surface fractal, as well as their scaling factor are unchanged, one may expect that the corresponding SAS curves from these models to be quite similar. This is confirmed numerically in Figure 9a, which shows that the SAS curves decay $\propto q^{-D}$, with $D \lesssim 2$ in the fractal regime, and they are hardly distinguishable. For comparison, the SAS intensity of a square, i.e., $D = 2$ is also included (black curve). The absence of maxima and minima in the fractal regime reflect a nearly uniform arrangement of the points, resulting from the random process used in generating their positions. The $f(\alpha)$-spectra for these models are shown in Figure 9b. They are confined in a narrow range of α values, i.e., $1.62 \lesssim \alpha \lesssim 1.95$ for models TR12 and TR13, and $1.72 \lesssim \alpha \lesssim 1.95$ for model TR33, thus indicating that the multifractal nature is not so pronounced as in the case of multiplicative cascade models (see also Figure 3b). Also, for comparison, Figure 9b includes the $f(\alpha)$-spectrum of a uniform distribution of points in the “ACGT” square, which is an Euclidean object. This is represented as a single point (black) of coordinates $(2, 2)$.

The results in Figure 9 illustrate that generally, SAS and $f(\alpha)$ can hardly be used to perform a clear differentiation between such structures. This issue may be eventually addressed by establishing first the relationship with the corresponding generalized Cantor set, as performed in Ref. [44], followed then by calculating the corresponding SAS intensities.

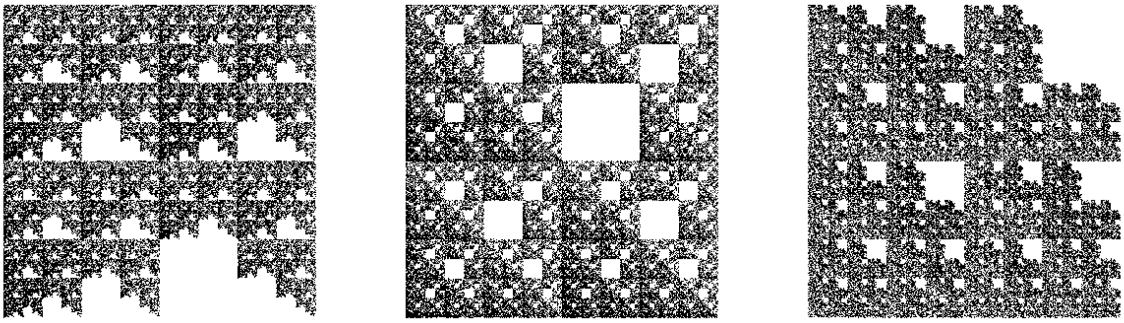


Figure 8. CGR in the "ACGT" square, with $A = (0,0)$, $T = (1/2,0)$, $G = (1/2,1/2)$, $C = (0,1/2)$ played with 8×10^4 points. Left part: "AT" moves are forbidden (model TR12). Middle part: "AG" moves are forbidden (model TR13). Right part: "GG" moves are forbidden (model TR33).

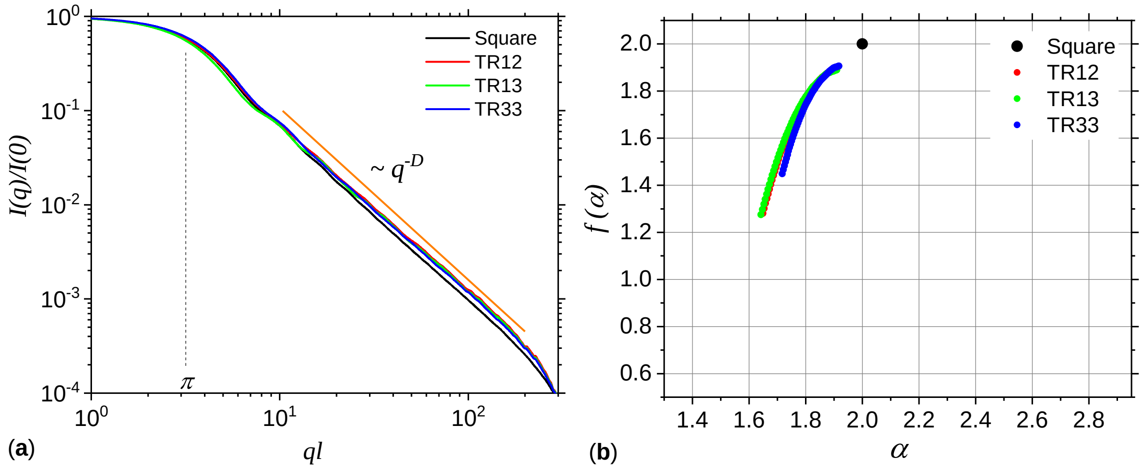


Figure 9. Scattering intensities (**a**) and $f(\alpha)$ spectra (**b**) for models TR12, TR13 and TR33. l is the overall size of the "ACGT" square. The data for a uniform square are included for comparison.

3.2. Application to DNA Sequences: Phospholamban, Mouse mitochondrion and Escherichia coli

The CGR followed by SAS and multifractal analysis is illustrated on *Phospholamban* [homo sapiens (human); GenBank ID: 5350] with 12,116 bp (Figure 10 left part), *Mouse mitochondria*, complete genome (GenBank ID: 342520) with 16,295 bp (Figure 10 middle part), and *Escherichia coli* O145:H28 strain 162,405 sequence161, whole genome shotgun sequence (NCBI Accession Version NZ_BJSS01000161) with 15,000 bp (Figure 10 right part).

The protein encoded by *Phospholamban* is found as a pentamer, is a major substrate for the cAMP-dependent protein kinase in cardiac muscle, and is an inhibitor of cardiac muscle sarcoplasmic reticulum Ca(2^+)-ATPase in the unphosphorylated state [47]. However, inhibition is relieved upon phosphorylation of the protein, and subsequent activation of the Ca(2^+) pump leads to enhanced muscle relaxation rates, thereby contributing to the inotropic response elicited in heart by beta-agonists. The encoded protein is a key regulator of cardiac diastolic function. Mutations in this gene are a cause of inherited human dilated cardiomyopathy with refractory congestive heart failure, and familial hypertrophic cardiomyopathy [47].

In the case of *Mouse mitochondrion*, the genome displays exceptional economy of organization, with tRNA genes interspersed between rRNA and protein-coding genes with zero or few non-coding nucleotides between coding sequences [48]. The genome is highly homologous in overall sequence as well as the organization to human mitochondrial DNA, and an important feature in that the translational start codon is AUN, with any of the four nucleotides in the third position, whereas the only translational stop codon is the orthodox UAA [48].

O145:H28 is one of the major non-O157 Shiga toxin (Stx)-producing *Escherichia coli* (STEC) lineages that causes severe diseases and is frequently isolated from humans, animals and foods [49]. Highly dynamic features of prophages and plasmids in the diversification of the STEC lineage, and the differential dynamics and impacts of these mobile genetic elements on the pangenome and virulence factor repertoire among O145:H28 strains [50].

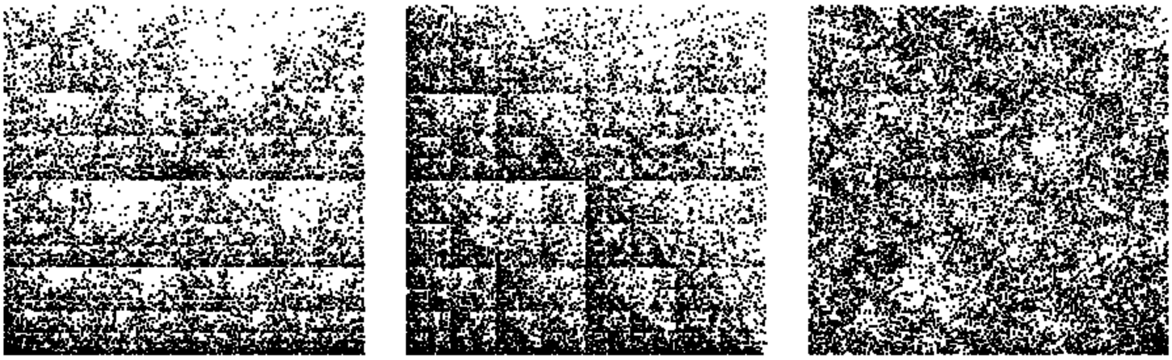


Figure 10. CGR in the "ACGT" square, with $A = (0,0)$, $T = (1/2,0)$, $G = (1/2,1/2)$, $C = (0,1/2)$ for *Phospholamban*, GenBank ID: 5350 (Left part), *Mouse mitochondrion*, GenBank ID: 342520 (Middle part), and *Escherichia coli*, NCBI Accession Version NZ_BJSS01000161.

The CGR of *Phospholmban* and *Mouse mitochondrion* in Figure 10 resemble more closer a deterministic-like fractal structure, in which a pattern repeats itself at increasingly smaller scales. In particular, the structure of *Phospholamban* resemble, in a first approximation, the structure of model TR12 shown in Figure 8.

Please note that in "CG" sub-quadrant of CGR from *Phospholamban*, i.e., in the upper-left region of the G-quadrant in Figure 10, the number of points are relatively much smaller as compared to other sub-quadrants. This indicates a paucity of subsequences ending in "CG" [14]. Similarly, the high number of points in "TT", "AT", "TA" and "AA" sub-quadrants, indicates a higher number of subsequences ending with these suffixes.

However, for *Mouse mitochondrion*, the white regions specific to model TR12, are partially populated with points, while their overall density is higher at the bottom. The paucity of subsequences ending in "CG" is still preserved, but however, the "AA" sub-sequence seems to be predominant. As it shall be seen below, this has important consequences on the behavior of SAS and $f(\alpha)$-spectra. A somehow similar arrangement can be seen also for *Mouse mitochondrion*, where a higher density of points is formed within the "ATC" triangle. The structure of *Escherichia coli* is characterized by the coexistence of regions with different densities, distributed more uniformly, as compared to *Phospholamban* and *Mouse mitochondrion*.

Figure 11a shows the SAS intensities from *Phospholamban*, *Mouse mitochondrion* and *Escherichia coli* calculated with Equation (19). For comparison, the scattering curve of a uniformly filled square of the same size as the "ACGT" square is included. The results show the presence of the fractal region at $\pi \lesssim ql \lesssim \pi/r_{\min}$, where l is the overall size of the square (360 × 360 pixels^2), and $r_{\min}$ is related to the smallest size of the pixels in the image. The common feature is the power-law decay with scattering exponent $D \lesssim 2$, revealing a mass fractal structure in all three cases. However, for *Phospholamban* and *Mouse mitochondrion*, the scattering curve has a generalized power-law decay (see **Introduction** section). To investigate this behavior in more details, Figure 11b shows the curves $q^D I(q)/I(0)$, which reveal more clearly the periodicity of *Phospholamban* and *Mouse mitochondrion* in the fractal regime. This is a signature of their deterministic nature, as indicated also by their structures shown in Figure 10a,b.

Then, the periodicity and the number of minima can be related to the scaling factor β and to the number of iterations [25]. In particular, the curve $q^D I(q)/I(0)$ is approximately log-periodic with the period equal to the inverse scaling factor, and thus $\beta \simeq 1/2$, while the number of iterations is equal to three. Please note that the value of the scaling factor is related to the CGR algorithm, and thus if one plot a given point, not at half distance between the previous point and the corresponding vertex, but at another fraction, or even using a 3D CGR as described in Ref. [51], one expects to recover also the corresponding scaling factor. Therefore, a SAS analysis could be employed to check the nature of the algorithm used to generate the points distribution.

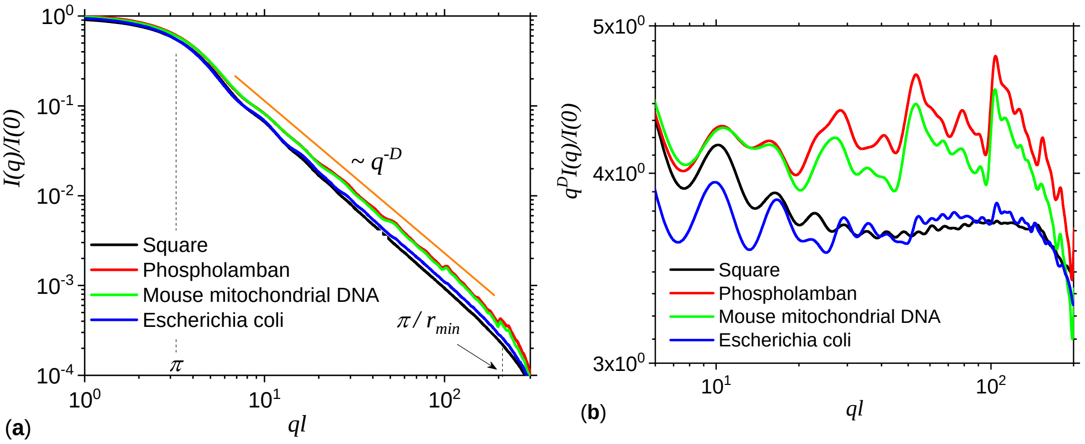


Figure 11. Scattering intensities of *Phospholamban* (red), *Mouse mitochondrion* DNA (green), and *Escherichia coli* (blue). For comparison, the scattering intensity of a uniform distribution of points within a square is shown. l is the overall size of the "ACGT" square, r_{min} is the minimum scale on which the samples have a fractal behavior, and D is the fractal dimension (see text for details).

Symmetry properties occurring in the "ACGT" square can also be revealed by performing a structural analysis in real space. Thus, by taking the Fourier transform of Equation (19), one obtains the corresponding pair-distance distribution function (pddf) $p(r)$, which gives the probability density of finding the distance r between two arbitrary points. For fractal structures, pddf can be used to reveal the fractal scaling factors, since it is characterized by a succession of groups of maxima and minima distributed periodically on a double logarithmic scale [45,52].

Figure 12a shows the pddf of *Phospholamban*, *Mouse mitochondrion* and *Escherichia coli*. For comparison, the pddf of a uniform distribution of points is included. The common feature is that all curves follow the same trend as that given by the uniform distribution, thus reflecting the symmetry properties of the square. In addition, the most common distances occur at $r/l \simeq 0.5$, i.e., at half the edge length of "ACGT" square, while the maximum distance between two arbitrary points occurs at $r/l \simeq 1.4$. However, the pddf of *Phospholamban* and *Mouse mitochondrion* are characterized by small maxima and minima in the form of wriggles, in the range $0.15 \lesssim r/r \lesssim 0.8$. This is the result of their deterministic nature and correspond to the generalized power-law decays occurring in Figure 11a,b in the fractal region. Finally, the pddf of *Escherichia coli* shows slight variations in the range $0.2 \lesssim r/l \lesssim 1$. In particular, for $0.2 \lesssim r/l \lesssim 0.7$ the pddf has smaller values as compared with the pddf of the uniform distribution of points, thus indicating the presence of more rarefied regions. However, for $0.7 \lesssim r/l \lesssim 1$, the pddf is slightly higher as compared to the pddf of the uniform distribution, thus indicating the presence of more dense regions (see also Figure 10c).

The non-uniform points density distribution in the CGR for these genomes is also reflected in Figure 12b, which shows that all structures are characterized by a multifractal structure. In particular, the *Phospholamban* appears to have the most heterogeneous distributions, reflected by the highest range of α values spanned by $f(\alpha)$ curve. Please note that the maxima of all spectra occurs at $f(\alpha) \lesssim 2$, in agreement with the values of fractal dimensions obtained from the scattering exponents in the fractal regime, shown in Figure 11a.

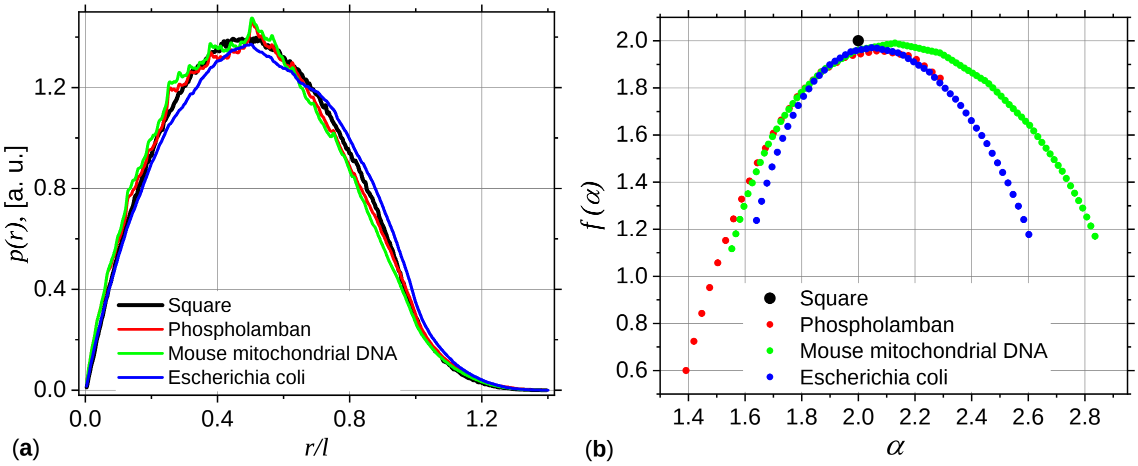


Figure 12. Pddf (**a**) and $f(\alpha)$ spectra (**b**) for *Phospholamban Mouse mitochondrion* DNA and *Escherichia coli*. The data for a uniform square are included for comparison.

4. Conclusions

In this work SAS and multifractal techniques are employed to perform a structural analysis of CGR of DNA sequences. To address the complementarity of the two techniques, they are applied first to two theoretical models, i.e., to a deterministic multiplicative cascades and to missing sequences, and then to genome data of *Phospholamban*, *Mouse mitochondrion* and *Escherichia coli*.

The deterministic multiplicative cascades model may serve as a benchmark against which other structures can be compared, since it provides an analytical expression for the generalized dimension spectra. It has been shown that within this model, multifractal analysis may not always differentiate between the generalized dimension, and consequently between $f(\alpha)$ spectra (Figure 3), but however, when an appropriate threshold is used, SAS may distinguish between such structures. The differentiation is based on the behavior of the scattering curves in the fractal regimes and on the specific way, i.e., amplitude, periodicity and number of most pronounced maxima and minima, change with the magnitude of the scattering vector (Figure 7).

The missing sequences model addresses the question of whether there are subsequences of short strings of letters ("A", "C", "G" and "T") that are missing. Here, a detailed analysis is performed on structures in which "A" is never followed by "T" or "G", and when "G" is never followed by "G". It is shown that SAS and multifractal techniques can hardly differentiate between such structures, since their corresponding spectra coincide (Figure 9). This may be attributed to the interplay between mass and surface-like fractals appearing in the "ACGT" square. While the mass fractals consist from the points resulted from CGR, surface-like fractals are formed by the power-law distribution of unoccupied regions (Figure 8).

Analysis of DNA sequences of *Phospholamban*, *Mouse mitochondrion* and *Escherichia coli* reveals that both SAS and multifractal techniques can be used to distinguish between structures (Figures 11 and 12), as well as to extract additional structural information. SAS shows that *Phospholamban* and *Mouse mitochondrion* DNA generate a more ordered arrangement in the CGR, as compared to *Escherichia coli*. Figure 10 shows the corresponding arrangements. In the former case, these resemble closer deterministic structures, and in addition to the fractal dimension, it allows us to obtain the scaling factor (from the log-periodicity of minima in the curve $q^D I(q)/I(0)$) and the fractal iteration number (from the number of these minima; Figure 9b). *Escherichia coli* generates a more uniform distribution but still with some density fluctuations (Figure 10 right part). For such a structure, SAS is characterized by the absence of clear periodicity of maxima and minima in the fractal regions (Figure 11b). The heterogeneity of such structures is clearly revealed also in the $f(\alpha)$ spectra (Figure 12b), and which show that *Phospholamban* and *Mouse mitochondrion* have the highest structural heterogeneity.

The analysis performed in this work provides good insights on the discriminating power of both SAS and multifractal formalism, and gives useful structural information which may help to distinguish between coding and non-coding sequences in DNA, phylogeny reconstruction or clustering of protein structures. In particular, the fractal dimension D of coding segments has smaller values as compared to non-coding segments [53]. This can be used to quantify the excess of short runs and the deficit of long runs of weak and of strong hydrogen bases in coding sequences. The dimensions D_{-2}, D_{-1} and D_1 from the multifractal spectra (Equation (15)) can be used to perform a classification of bacteria by assigning to each sequence a point in 2D and 3D spaces (D_{-1}, D_1), and respectively (D_1, D_1, D_{-2}). In this representation, bacteria that are close phylogenetically, are almost close in these spaces [54].

Conflicts of Interest: The author declares no conflict of interest.

References

1. Arneodo, A.; Vaillant, C.; Audit, B.; Argoul, F.; d'Aubenton Carafa, Y.; Thermes, C. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Phys. Rep.* **2011**, *498*, 45–188. doi:10.1016/j.physrep.2010.10.001.
2. Felsenfeld, G.; Groudine, M. Controlling the double helix. *Nature* **2003**, *421*, 448–453. doi:10.1038/nature01411.
3. Albrecht-Buehler, G. Fractal genome sequences. *Gene* **2012**, *498*, 20–27. doi:10.1016/j.gene.2012.01.090.
4. Albuquerque, E.L.; Fulco, U.L.; Freire, V.N.; Caetano, E.W.S.; Lyra, M.L.; de Moura, F.A.B.F. DNA-based nanobiostructured devices: The role of quasiperiodicity and correlation effects. *Phys. Rep.* **2014**, *535*, 139–209. doi:10.1016/j.physrep.2013.10.004.
5. Niu, X.H.; Hu, X.H.; Shi, F.; Xia, J.B. Predicting DNA binding proteins using support vector machine with hybrid fractal features. *J. Theor. Biol.* **2014**, *343*, 186–192. doi:10.1016/j.jtbi.2013.10.009.
6. Lennon, F.E.; Cianci, G.C.; Cipriai, N.A.; Hensing, T.A.; Zhang, H.J.; Chen, C.T.; Murgu, S.T.; Vokes, E.E.; Vannier, M.W.; Salgia, R. Lung cancer—A fractal viewpoint. *Nat. Rev. Clin. Oncol.* **2015**, *12*, 664–675. doi:10.1038/nrclinonc.2015.108.
7. Babic, M.; Mihelic, J.; Calì, M. Complex Network Characterization Using Graph Theory and Fractal Geometry: The Case Study of Lung Cancer DNA Sequences. *Appl. Sci.* **2020**, *10*, 3037. doi:10.3390/app10093037.
8. Li, W. Mutual information functions versus correlation functions. *J. Stat. Phys.* **1990**, *60*, 823–837. doi:10.1007/BF01025996.
9. Voss, R.F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **1992**, *68*, 3805–3808. doi:10.1103/PhysRevLett.68.3805.
10. Havlin, S.; Buldyrev, S.V.; Goldberger, A.L.; Mantegna, R.N.; Peng, C.K.; Simons, M.; Stanley, H.E. Statistical and linguistic features of DNA sequences. *Fractals* **1995**, *3*, 269–284. doi:10.1142/S0218348X95000229.
11. Herzel, H.; Große, I. Measuring correlations in symbol sequences. *Phys. A Stat. Mech. Appl.* **1995**, *216*, 518–542. doi:10.1016/0378-4371(95)00104-F.
12. Bernaola-Galván, P.; Román-Roldán, R.; Oliver, J.L. Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E* **1996**, *53*, 5181–5189. doi:10.1103/PhysRevE.53.5181.
13. Arneodo, A.; d'Aubenton Carafa, Y.; Bacry, E.; Graves, P.V.; Muzy, J.F.; Thermes, C. Wavelet based fractal analysis of DNA sequences. *Phys. D Nonlinear Phenom.* **1996**, *96*, 291–320. doi:10.1016/0167-2789(96)00029-2.
14. Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.* **1990**, *18*, 2163–2170. doi:10.1093/nar/18.8.2163.
15. Hoang, T.; Yin, C.; Yau, S.S.T. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics* **2016**, *108*, 134–142. doi:10.1016/j.ygeno.2016.08.002.
16. Gutiérrez, J.M.; Rodriguez, M.A.; Abramson, G. Multifractal analysis of DNA sequences using a novel chaos-game representation. *Phys. A Stat. Mech. Appl.* **2001**, *300*, 271–284. doi:10.1016/S0378-4371(01)00333-8.
17. Han, J.-J.; Fu, W.-J. Wavelet-based multifractal analysis of DNA sequences by using chaos-game representation. *Chin. Phys. B* **2010**, *19*, 010205. doi:10.1088/1674-1056/19/1/010205.

18. Yu, Z.G.; Anh, V.; Lau, K.S. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J. Theor. Biol.* **2004**, *226*, 341–348. doi:10.1016/j.jtbi.2003.09.009.
19. Zu-Guo, Y.; Qian-Jun, X.; Long, S.; Jun-Wu, Y.; Anh, V. Chaos game representation of functional protein sequences, and simulation and multifractal analysis of induced measures. *Chin. Phys. B* **2010**, *19*, 068701. doi:10.1088/1674-1056/19/6/068701.
20. Pal, M.; Kiran, V.S.; Rao, P.M.; Manimaran, P. Multifractal detrended cross-correlation analysis of genome sequences using chaos-game representation. *Phys. A Stat. Mech. Appl.* **2016**, *456*, 288–293. doi:10.1016/j.physa.2016.03.074.
21. Zaia, A.; Maponi, P.; Zannotti, M.; Casoli, T. Biocomplexity and Fractality in the Search of Biomarkers of Aging and Pathology: Mitochondrial DNA Profiling of Parkinson's Disease. *Int. J. Mol. Sci.* **2020**, *21*, 1758. doi:10.3390/ijms21051758.
22. Feigin, L.A.; Svergun, D.I. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*; Springer: Boston, MA, USA, 1987; p. 335. doi:10.1007/978-1-4757-6624-0.
23. Martin, J.E.; Hurd, A.J. Scattering from fractals. *J. Appl. Cryst.* **1987**, *20*, 61–78. doi:10.1107/S0021889887087107.
24. Schmidt, P.W. Small-angle scattering studies of disordered, porous and fractal systems. *J. Appl. Cryst.* **1991**, *24*, 414–435. doi:10.1107/S0021889891003400.
25. Cherny, A.Y.; Anitas, E.M.; Osipov, V.A.; Kuklin, A.I. Deterministic fractals: Extracting additional information from small-angle scattering data. *Phys. Rev. E* **2011**, *84*, 036203. doi:10.1103/PhysRevE.84.036203.
26. Anitas, E.M.; Slyamov, A. Structural characterization of chaos game fractals using small-angle scattering analysis. *PLoS ONE* **2017**, *12*, 1–16. doi:10.1371/journal.pone.0181385.
27. Debye, P. Zerstreuung von Röntgenstrahlen. *Ann. Phys.* **1915**, *351*, 809–823. doi:10.1002/andp.19153510606.
28. Provata, A.; Almirantis, Y. Fractal Cantor Patterns in the Sequence Structure of DNA. *Fractals* **2000**, *8*, 15–27. doi:10.1142/S0218348X00000044.
29. Barnsley, M.F. *Fractals Everywhere*, 2nd ed.; Morgan Kaufmann: Burlington, MA, USA, 2000.
30. Rogers, C.A. *Hausdorff Measures*; Cambridge University Press: Cambridge, UK, 1970; p. 179.
31. Dimension und äußeres Maß.
32. Gouyet, J.F. *Physics and Fractal Structures*; Masson: Paris, France, 1996; p. 234.
33. Arneodo, A.; Decoster, N.; Roux, S. A wavelet-based method for multifractal image analysis. I. Methodology and test applications on isotropic and anisotropic random rough surfaces. *Eur. Phys. J. B* **2000**, *15*, 567–600. doi:10.1007/s100510051161.
34. Decoster, N.; Roux, S.; Arnéodo, A. A wavelet-based method for multifractal image analysis. II. Applications to synthetic multifractal rough surfaces. *Eur. Phys. J. B* **2000**, *15*, 739–764. doi:10.1007/s100510051179.
35. Muzy, J.F.; Bacry, E.; Arneodo, A. Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. *Phys. Rev. E* **1993**, *47*, 875–884. doi:10.1103/PhysRevE.47.875.
36. Chhabra, A.; Jensen, R.V. Direct Determination of the f (alpha) Singularity Spectrum. *Phys. Rev. Lett.* **1989**, *62*, 1327–1130. doi:10.1103/PhysRevLett.62.1327.
37. Pantos, E.; van Garderen, H.F.; Hilbers, P.A.J.; Beelen, T.P.M.; van Santen, R.A. Simulation of small-angle scattering from large assemblies of multi-type scatterer particle. *J. Mol. Struct.* **1996**, *383*, 303. doi:10.1016/S0022-2860(96)09302-7.
38. Meakin, P. Diffusion-limited aggregation on multifractal lattices: A model for fluid-fluid displacement in porous media. *Phys. Rev. A* **1987**, *36*, 2833–2837. doi:10.1103/PhysRevA.36.2833.
39. Martinez, V.J.; Jones, B.J.T.; Dominguez-Tenreiro, R.; van de Weygaert, R. Clustering Paradigms and Multifractal Measures. *Astrophys. J.* **1990**, *357*. doi:10.1086/168890.
40. Tarquis, A.M.; Losada, J.C.; Benito, R.M.; Borondo, F. Multifractal analysis of tori destruction in a molecular Hamiltonian system. *Phys. Rev. E* **2001**, *65*, 016213. doi:10.1103/PhysRevE.65.016213.
41. Hao, B.; Xie, H.; Yu, Z.; Chen, G. Avoided Strings in Bacterial Complete Genomes and a Related Combinatorial Problem. Annals of Combinatorics. *Ann. Comb.* **2000**, *4*, 247–255. doi:10.1007/PL00001279.
42. Hao, B.L.; Lee, H.; Zhang, S.Y. Fractals related to long DNA sequences and complete genomes. *Chaos Solitons Fract.* **2000**, *11*, 825–836. doi:10.1016/S0960-0779(98)00182-9.

43. Hao, B.; Xie, H.; Yu, Z.; Chen, G.Y. Factorizable language: from dynamics to bacterial complete genomes. *Physica A* **2000**, *288*, 10–20. doi:10.1016/S0378-4371(00)00411-8.
44. Yang, Z.; Wang, P. DNA Sequences with Forbidden Words and the Generalized Cantor Set. *J. Appl. Math. Phys.* **2019**, *7*, 1687–1696. doi:10.1016/S0378-4371(00)00411-8.
45. Cherny, A.Y.; Anitas, E.M.; Kuklin, A.I.; Balasoiu, M.; Osipov, V.A. Scattering from generalized Cantor fractals. *J. Appl. Cryst.* **2010**, *43*, 790–797. doi:10.1107/S0021889810014184.
46. Anitas, E.M. Small-Angle Scattering from Fractals: Differentiating between Various Types of Structures. *Symmetry* **2020**, *12*, 65. doi:10.3390/sym12010065.
47. NCBI. PLN Phospholamban [Homo Sapiens (Human)]. Available online: https://www.ncbi.nlm.nih.gov/gene/5350 (accessed on 29 June 2020).
48. Bibb, M.J.; Van Etten, R.A.; Wright, C.T.; Walberg, M.W.; Clayton, D.A. Sequence and gene organization of mouse mitochondrial DNA. *Cell* **1981**, *26*, 167–180. doi:10.1016/0092-8674(81)90300-7.
49. Brooks, J.T.; Sowers, E.G.; Wells, J.G.; Green, K.D.; Griffin, P.M.; Hoekstra, P.M.; Strockbine, N.A. Non-O157 Shiga toxin-producing Escherichia coli infections in the United States. *J. Infect. Dis.* **2005**, *192*, 1422–1429. doi:10.1086/466536.
50. Nakamura, K.; Murase, K.; Sato, M.P.; Toyoda, A.; Itoh, T.; Mainil, J.G.; Piérard, D.; Yoshino, S.; Kimata, K.; Isobe, J.; et al. Differential dynamics and impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga toxin-producing Escherichia coli O145:H28. *Microb. Genom.* **2020**, *6*, 1–13. doi:10.1099/mgen.0.000323.
51. Feng, J.; Wang, T.-M. A 3D graphical representation of RNA secondary structures based on chaos game representation. *Chem. Phys. Lett.* **2008**, *454*, 355–361. doi:10.1016/j.cplett.2008.01.041.
52. Anitas, E.M.; Marcelli, G.; Szakacs, Z.; Todoran, R.; Todoran, D. Structural Properties of Vicsek-like Deterministic Multifractals. *Symmetry* **2019**, *11*, 806. doi:10.3390/sym11060806.
53. Berthelsen, C.L.; Glazier, J.A.; Skolnick, M.H. Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys. Rev. A* **1992**, *45*, 8902–8913. doi:10.1103/PhysRevA.45.8902.
54. Yu, Z.G.; Anh, V.; Lau, K.S. Measure representation and multifractal analysis of complete genomes. *Phys. Rev. E* **2001**, *64*, 031903. doi:10.1103/PhysRevE.64.031903.