Hierarchical multifractal representation of symbolic sequences and application to human chromosomes

A. Provata and P. Katsaloulis
Institute of Physical Chemistry, National Center for Scientific Research "Demokritos," 15310 Athens, Greece

(Received 23 September 2009; published 8 February 2010)

The two-dimensional density correlation matrix is constructed for symbolic sequences using contiguous segments of arbitrary size. The multifractal spectrum obtained from this matrix motif is shown to characterize the correlations in the symbolic sequences. This method is applied to entire human chromosomes, shuffled human chromosomes, reconstructed human genomic sequences and to artificial random sequences. It is shown that all human chromosomes have common characteristics in their multifractal spectrum and deviate substantially from random and uncorrelated sequences of the same size. Small deviations are observed between the longer and the shorter chromosomes, especially for the higher (in absolute values) statistical moments. The correlations are crucial for the form of the multifractal spectrum; surrogate shuffled chromosomes present randomlike spectrum, distinctly different from the actual chromosomes. Analytical approaches based on hierarchical superposition of tensor products show that retaining pair correlations in the sequences leads to a closer representation of the genomic multifractal spectra, especially in the region of negative exponents, due to the underrepresentation of various functional units (such as the cytosine-guanine *CG* combination and its complementary *GC* complex). Retaining higher-order correlations in the construction of the tensor products is a way to approach closer the structure of the multifractal spectra of the actual genomic sequences. This hierarchical approach is generic and is applicable to other correlated symbolic sequences.

DOI: 10.1103/PhysRevE.81.026102 PACS number(s): 89.75.Fb, 05.45.Df, 87.14.gk, 87.15.A−

I. INTRODUCTION

The presence of nontrivial statistical correlations in the genome of many organisms has been the subject of intense investigations in the past three decades, starting in 1992 with the seminal papers by Li *et al.* [1], Peng *et al.* [2], and Voss [3]. Since then a great number of studies have been devoted to the exploration of correlations in the genomes of many organisms and to the understanding of the mechanisms producing these correlations. More specifically, scaling laws were detected in the primary structure of DNA and in particular in the noncoding parts of higher eucaryotic DNA [1–14]. Later, the size distributions of coding and noncoding DNA segments were shown to follow distinctly different statistics, short range in the former case and long range in the latter [8]. The CG/GC content of the genome (C=cytosine, G=guanine) was also shown to present long-range features in its distribution within genomic sequences [15–23].

The identification of long-range distributions in the size of noncoding DNA sequences interrupted by coding short-range distributed sequences has induced the hypothesis of stochastic fractal, Cantor-like structures in the genome. Investigations in this direction have demonstrated such tendencies and fractal dimensions were obtained for various categories of organisms [24–26]. In addition to this, the chaos game representation of genomic sequences has also demonstrated fractal features [27].

The complexity of the genomic sequences, due to the many superposed and often counteracting evolutionary processes, leads to the hypothesis that the DNA cannot be characterized by a simple fractal dimension but it could involve many scales in its structure. This hypothesis points directly to the notion of multifractality, where many scales are involved in the construction of the system. It is well known that usual fractal and multifractal systems emerge from multiple iterations of the same process or multiple superpositions of the same initial unit. Such hierarchical processes are known to take place during the evolution of organisms, where DNA duplications and repetitions are abundantly observed [28]. In particular, one repetitive element alone, the ALu sequence, is shown to cover 10.7% of the human genome. In the current study the presence of multiple scaling in the human genome will be discussed through the search for multifractality. In addition, the origin of this multifractality will be investigated through comparisons with elementary iterative processes (tensorial products of density correlation matrices) producing specific types of multifractal spectra.

In the next section the probability of finite-size blocks containing a specific configuration of base pairs (bps) will be defined. It will be shown that the frequency of occurrence of a certain DNA block is different from the product of frequencies of individual bps, which is a direct indication of correlations in DNA. The construction of the probability block-density matrix will further demonstrate the nonuniform distribution of blocks and the presence of correlations. In the same section the calculation of the multifractal spectrum based on the probability density matrix will be discussed. In Sec. III the multifractal spectrum of all human chromosomes will be calculated and direct comparison will be undertaken with the spectrum of various surrogate data. In Sec. IV it will be shown how to obtain iteratively the probability distribution of blocks from multiple superpositions of the single bps distributions. This iterative process yields an analytical expression of the multifractal spectrum, which can be compared with the one calculated directly from the genomic sequences (chromosomes). When higher-order blocks are used for the construction of the probability block-density matrix, this hierarchical approach gradually approaches the results of the real chromosomes. In the final section the main conclusions of this study will be recapitulated and open problems will be discussed.

II. METHOD: 2D DENSITY CORRELATION MATRIX

A. Construction of the 2D density correlation matrix

Consider a symbolic sequence $S=S_1S_2S_3\cdots S_N$, where N is the length of the sequence and S_i takes values from an m-letter alphabet. As a working example genomic sequences will be used, in which case $m=4$ and S_i can take values from the set $\{A,C,G,T\}$ (A stands for adenine, C for cytosine, G for guanine, and T for thymine). Define as $P_{\{S_l\}}$ the probability to find a given block of size l within the sequence S. For random uncorrelated sequences,

$$P_{\{S_l\}}=\prod_{i=1}^{m} p_i^{m_i}, \quad \sum_{i=1}^{m} m_i = l \tag{1}$$

where p_i denotes the single symbol probability (of symbol S_i) and m_i denotes the number of times the symbol S_i is found within the block of size l. In the general case correlations oblige the block probability to deviate substantially from Eq. (1). This is usually the case for sequences originating from digitalized natural data, or conversion of continuous sequences into finite alphabets, or partition of the state space of continuous dynamical systems into a finite number of elements. Also, for symbol sequences originating from natural processes, such as genomic and protein sequences, natural languages, music etc, correlations are built in as a result of evolution.

To obtain information about the presence of a given block of size n $\{B_n\}=[B_1B_2\cdots B_n]$ following in juxtaposition a block $\{A_n\}=[A_1A_2\cdots A_n]$, we construct the frequency matrix M_n^m of all possible combinations of blocks. For an m-letter alphabet and blocks of size n, the number of all possible combinations are m^n and the 2-D frequency matrix M_m^n contains $m^n \times m^n$ elements. If A and B denote specific combinations, $\{A\}\equiv[A_1A_2\cdots A_n]$ and $\{B\}\equiv[B_1B_2\cdots B_n]$, the elements of the frequency matrix correspond to

$$M_{\{AB\}} \equiv p_{[A_1A_2\cdots A_nB_1B_2\cdots B_n]} \tag{2}$$

$$\neq p_{[A_1A_2\cdots A_n]}\cdot p_{[B_1B_2\cdots B_n]} \equiv M_{\{A\}\{B\}} \tag{3}$$

and their precise value needs to be computed directly from the sequence S. In particular, for the case of DNA sequences ($m=4$) and for two-letter words ($n=2$) the matrix M_4^2 has dimensions (16×16) and takes the following form:

$$M_4^2=\begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots & p_{AATT} \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots & p_{ACTT} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ p_{TTAA} & p_{TTAC} & p_{TTAG} & p_{TTAT} & p_{TTCA} & \cdots & p_{TTTT} \end{pmatrix}. \tag{4}$$

From its construction, the matrix M contains information on the frequency of occurrence of all words of length $2n$ as well as on the probability of a given block of size n to follow a specific block of the same size n. Although in this work the size of the two blocks is equal, the results can be generalized for blocks of different sizes. Due to the construction of M the sum of all its elements should be equal to unity. This is true for all sequences, whether or not they are uniform or correlated. Note, that for a uniform, random distribution of the letters and very long (infinite length) sequences, the sum of each column and row should be equal to $1/(m^n)$. This does not hold for nonuniform or correlated sequences.

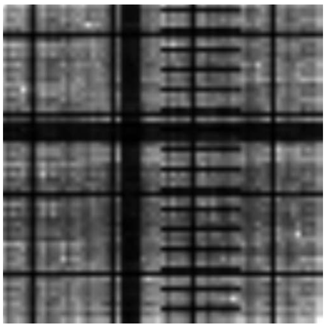

FIG. 1. The density correlation matrix for human chromosome 10, with $n=3$.

The matrix M_m^n can be represented with a rough surface, the local height of which is defined by the block probabilities. If the various blocks have similar frequencies, the surface is almost flat. If there are variations in the block frequencies these are mirrored in the variety of surface heights and consequently in the complexity of the multifractal spectrum. In DNA sequences it is known that certain combinations, such as the CG one, are reserved for specific functions (promoter sequences, regulatory elements, etc.). These combinations are not abundant and consequently their frequency is relatively small [28]. As a result, blocks of any size which include the CG (or GC) complex are also infrequent. The superposition of these blocks with others which have various frequencies create a variety of scales, which give nontrivial structure to the multifractal spectrum of the density correlation matrix.

As an example in Fig. 1 the density correlation matrix of the human chromosome 10, for $n=3$, is depicted. The variation in the gray scales represents the local value p_{ij}. Low p_{ij} values are represented by black colors, while high p_{ij}'s are represented by white colors. One can observe different scalings in the coloring, ranging from black, to dark gray, to light gray and to white. Low density layers which correspond to the rare CG and GC content (large black cross) and larger blocks containing multiple CG combinations (smaller black crosses) clearly appear. Maximum values such as p_{AAAA}, p_{TTTT}, and single point mutations of those, can be seen with white and light gray colors in the upper left and lower right part of Fig. 1.

B. Block correlations in DNA

To obtain some intuition about the type of correlations we search, it is useful to construct the simple frequency matrix for blocks of size $2n=2$. Together the frequency values found in the human chromosome 10 are reported,

$$M=\begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix},$$

$$M_{ch:10}=\begin{pmatrix} 0.0958 & 0.0505 & 0.0705 & 0.0752 \\ 0.0734 & 0.0536 & 0.0103 & 0.0706 \\ 0.0597 & 0.0440 & 0.0537 & 0.0505 \\ 0.0631 & 0.0597 & 0.0736 & 0.0959 \end{pmatrix}. \tag{5}$$

In the above data the accuracy is of the order of ± 0.0005. Curiously enough, the values of the frequencies $p_{ij}, i, j \in \{A, C, G, T\}$ are close in all human chromosomes for all combinations of i and j. In other words, the numerical values of the matrix M are almost the same as in Eq. (5) for all human chromosomes. One first observation is that this matrix is not symmetric around the diagonal, $p_{ij} \neq p_{ji}$, while symmetry is always apparent in the case of random, uncorrelated sequences. Note also the small frequency of appearance of the doublet CG. This is directly related to the specificity of the CG complex, which is not abundant in the genome but is retained for specific functions related to promoting the gene transcription [28], as discussed also in Sec. II A. The same is true for the complimentary GC sequence.

To demonstrate pedagogically the presence of correlations it is enough to calculate the same matrix as products of frequencies of finding a single letter. The individual bps frequencies computed for chromosome 10 are

$$p_A = 0.291\,921, \quad p_C = 0.207\,966,$$

$$p_G = 0.207\,859, \quad p_T = 0.292\,219. \tag{6}$$

In this uncorrelated case the binary frequencies are:

$$M_{uncor}=\begin{pmatrix} p_A p_A & p_A p_C & p_A p_G & p_A p_T \\ p_C p_A & p_C p_C & p_C p_G & p_C p_T \\ p_G p_A & p_G p_C & p_G p_G & p_G p_T \\ p_T p_A & p_T p_C & p_T p_G & p_T p_T \end{pmatrix} = \begin{pmatrix} 0.0852 & 0.0607 & 0.0607 & 0.0853 \\ 0.0607 & 0.0432 & 0.0432 & 0.0608 \\ 0.0607 & 0.0432 & 0.0432 & 0.0608 \\ 0.0853 & 0.0609 & 0.0608 & 0.0854 \end{pmatrix}. \tag{7}$$

As expected this matrix is symmetric around the diagonal. Note also, the difference in the numerical values between frequency matrices (5) and (7).

Having obtained a first impression on the type of correlations that we seek, we will next characterize the correlations through the multifractal spectra of higher-order frequency matrices.

C. Calculation of the multifractal spectrum

The distribution of words of size n in the above manner leads to the construction of a two-dimensional (2D) mesh with different density on each site, constituting a surface above this mesh. The complexity of the sequence is now mirrored in the complexity of the relief of this surface.

Natural surfaces, or surfaces obtained from natural sequences as above, tend to display characteristic heights over a variety of scales and often they present self-similar features [29–31]. For single-scale self-similar surfaces, one single exponent, the fractal dimension D is enough to characterize the complexity of the structure. In the case of more complex systems, such as the DNA, where a great number of evolutionary processes have been involved in their formation, it is hard to imagine that a single scaling exponent can adequately describe their structure. A spectrum of exponents is more appropriate for the quantitative description of a complex sequence and the corresponding constructed surface. The continuous spectrum of exponents, each of which describes the local distribution of specific heights, is obtained through the multifractal description of the surface and is also called the singularity spectrum. The singularity spectrum is obtained through the generalized dimensions and the exponents D_q of order q are obtained using the box-counting technique. Namely, the surface is divided in squares (or in general "boxes") of linear size ϵ. In each "box" in position $\{i, j\}$ the surface height/intensity is denoted as p_{ij}. The qth order exponent is calculated as

$$D_q = \frac{1}{q-1} \lim_{\epsilon \to 0} \frac{\log \sum_{i,j=1}^{m^n} p_{ij}^q}{\log \epsilon}. \tag{8}$$

In the case of the density correlation matrix [Eq. (5) in Sec. II A], each matrix element corresponds to the local "height" of the surface and formula (8) can be directly applied. The limit $\epsilon \to 0$ corresponds to the single-site calculation. The case $q=0$ corresponds to the "capacity dimension," which is best estimated with the usual box-counting technique. The case $q=1$ presents a singularity and is calculated through de l' Hospital's rule. It is also called the "information dimension" and is given as

$$D_1 = \lim_{\epsilon \to 0} \frac{\sum_{i,j=1}^{m^{2n}} p_{ij} \log p_{ij}}{\log \epsilon} \tag{9}$$

The case $q=2$ is also known as the "correlation dimension." In the cases of nonfractal, random systems or simple (one-scale) fractals, all exponents converge a) to the capacity dimensions for simple, one-scale fractals or b) to the embedding dimensions for random, nonfractal surfaces. In the case of systems with multiple scales, the expression for D_q is a decreasing function of q.

The computation of D_q using Eqs. (8) and (9) can proceed directly using the block probabilities obtained from the DNA sequences, as in Sec. II B. Next we will calculate the multifractal spectrum of all human chromosomes and compare them with random surrogate data, while in Sec. IV we will show how to construct the 2D density correlation matrix M using a hierarchical superposition of blocks (tensor product)

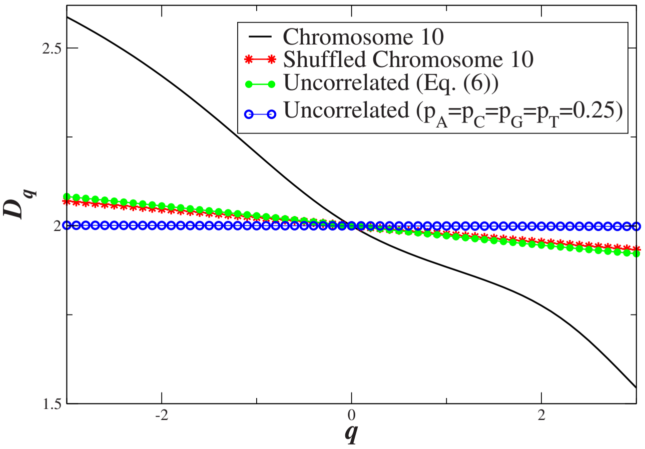

FIG. 2. (Color online) The multifractal exponents D_q as a function of q for human chromosome 10 and various surrogate data.

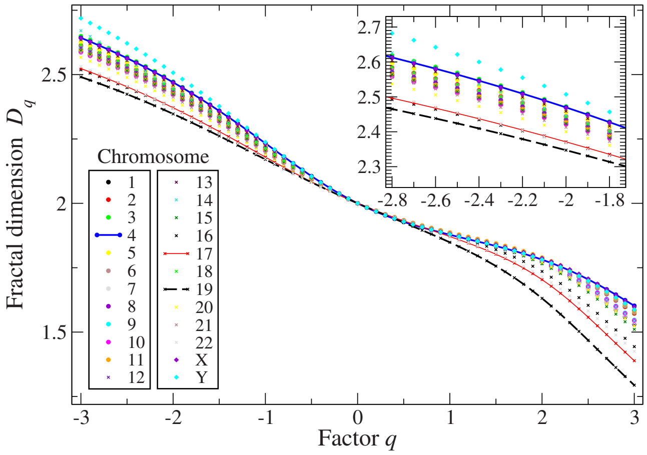

FIG. 3. (Color online) The multifractal exponent D_q as a function of q for all human chromosomes.

and obtain an analytical approximation to the multifractal spectrum by retaining various degrees of correlations.

III. APPLICATION TO ARTIFICIAL RANDOM SEQUENCES AND GENOMIC SEQUENCES

Using the method described in Sec. II A we first construct the density correlation matrix for all 24 human chromosomes (22 autosomes +2 sex chromosomes). We use blocks of size $n=8$ which produce the matrices M_4^8. For each one of these matrices we calculate the multifractal dimensions D_q using Eqs. (8) and (9). Figure 2 presents the multifractal exponents D_q as a function of q from real chromosomal data and various surrogates. In this figure are shown: (a) the multifractal exponents obtained from the chromosome 10 sequence (thick solid black line). (b) The red line (with symbols stars) corresponds to shuffled data from the same chromosome. The original sequence has been shuffled a number of times equal to 10× (chromosomal size in bps). (c) The green line (with symbols filled circles) represents an artificial random sequence of size equal to chromosome 10 and single symbol probabilities given by Eq. (6). (d) The blue line (with symbols open circles) corresponds to a sequence of size equal to chromosome 10 and equal probabilities for all four symbols. Note that there are a number of unknown bps in chromosome 10, which are found with frequency $p_N=1.7\times 10^{-5}$. This introduces an error of the order of 10^{-5} in our calculations which is of the order of the accuracy used in this work.

As expected for case (d) (the completely random sequence, with equiprobable distribution of all 4 bps), the frequency covers equiprobably the 2D space and thus all the exponents collapse to the value 2 (this will be shown later in Sec. II A). The case of random data with different bps frequency (case c) deviates slightly from the equal frequency distribution, as seen in the same figure. The same is true for the case of the shuffled data (case b), where all correlations are destroyed due to the shuffling process. When all correlations break the resulting sequence becomes statistically equivalent to a random one [see case (c)] and thus presents identical multifractal exponents.

In Fig. 3 the multifractal exponents D_q are computed for all human chromosomes. The inset depicts details of the same figure. Three lines were added as eyeguides. The blue (thick solid) line which corresponds to chromosome 4, the red (thin solid) line for chromosome 17 and the black (dashed) line for chromosome 19. All sequences show $D_q=2$ for $q=0$. This is expected, from Eq. (8), since the exponent for $q=0$ represents the fractal dimension (capacity dimension) of the substrate which is equal to 2. Note that longer chromosomes (numbers 1, 2…11) follow the blue (thick) line and are depicted at the top in Fig. 3, while shorter ones follow the red (thin) and black (dashed) lines and are depicted lower on the same diagram.

Infrequent events (extreme events) which take place with low probabilities p_{ij} become evident and give larger contributions for negative values of q, while frequent events contribute more in the positive q values. As discussed earlier the complexes *CG* and *GC* are very infrequent [see also matrix (5)] and they are responsible for the large deviations in Fig. 3 for negative q. At the other end, for positive q, groups such as poly-A and poly-T contribute to the divergence from the homogeneous value $D_q=2$.

From the above discussion the existence of correlations in the base sequence of DNA becomes evident with the use of the density correlation matrix, while the comparison of the multifractal spectra of genomic sequences with random and uncorrelated ones reveals the presence of specific structures (sequences) with functional roles in the genome.

IV. ANALYTICAL HIERARCHICAL APPROACH USING TENSOR PRODUCTS

A. Higher-order tensor products of symbol sequences

Many physical systems emerge as superpositions of themselves from finer to larger scales and this way correlations are generated and propagate. In particular, in the case of DNA sequences it is widely accepted that the genome suffers extensive duplications and many copies of the same segment maybe found within the same chromosome in adjacent positions [28]. Assuming in this case that the genome has emerged from multiple superpositions of itself, the probability $p_{\{A_{2n}\}}$ of a certain string $\{A_{2n}\}=[A_1A_2\ldots A_nA_{n+1}\ldots A_{2n}]$ to result from the superpositions of strings $\{A_n\}=[A_1A_2\ldots A_n]$ and $\{A'_n\}=[A_{n+1}A_{n+2}\ldots A_{2n}]$ will be computed. This includes the case of duplication when $A_n=A'_n$.

To take into account all possible configurations of strings of size $2n$ we define the superposition through the tensor product of matrices (or outer product in case of vectors),

$$M_m^{2n}=M_m^n\otimes M_m^n \tag{10}$$

where the symbol $\otimes$ denotes the tensor multiplication. For the simple case of a 2-letter alphabet (letters: A,C) the superposition of all possible strings containing 2 letters results in the probability density matrix of all strings containing $2\times2=4$ letters,

$$\begin{aligned}M_2^2\otimes M_2^2&=\begin{bmatrix}p_{AA}&p_{AC}\\p_{CA}&p_{CC}\end{bmatrix}\otimes\begin{bmatrix}p_{AA}&p_{AC}\\p_{CA}&p_{CC}\end{bmatrix}\\&=\begin{bmatrix}p_{AA}\begin{bmatrix}p_{AA}&p_{AC}\\p_{CA}&p_{CC}\end{bmatrix}&p_{AC}\begin{bmatrix}p_{AA}&p_{AC}\\p_{CA}&p_{CC}\end{bmatrix}\\p_{CA}\begin{bmatrix}p_{AA}&p_{AC}\\p_{CA}&p_{CC}\end{bmatrix}&p_{CC}\begin{bmatrix}p_{AA}&p_{AC}\\p_{CA}&p_{CC}\end{bmatrix}\end{bmatrix}\\&=\begin{bmatrix}p_{AA}p_{AA}&p_{AA}p_{AC}&p_{AC}p_{AA}&p_{AC}p_{AC}\\p_{AA}p_{CA}&p_{AA}p_{CC}&p_{AC}p_{CA}&p_{AC}p_{CC}\\p_{CA}p_{AA}&p_{CA}p_{AC}&p_{CC}p_{AA}&p_{CC}p_{AC}\\p_{CA}p_{CA}&p_{CA}p_{CC}&p_{CC}p_{CA}&p_{CC}p_{CC}\end{bmatrix}.\end{aligned} \tag{11}$$

Note that using the treatment of Eq. (11) the correlations of doublets are retained but above two letters the probabilities are treated as decorrelated. This might be true in some cases (for some combinations) but is not true in the general case. If one wishes to retain correlations of higher orders, one may start the superposition procedure from longer words $n>2$. Using the superposition of a matrix with itself we quickly produce matrices of large sizes. Consider for example the matrix M_4^2 of form (5). The tensor product of M_4^2 (size 16×16) yields a matrix M_4^4 of size $4^4\times4^4=256\times256$. Further superposing these matrices yields a matrix of size M_4^8 of size $4^8\times4^8=65\,536\times65\,536$.

Certainly, using tensor calculus one can find tensor products of matrices with different dimensionality and thus obtain the probability of finding a word of arbitrary size n from superpositions of words of sizes n_1 and n_2, provided that $n_1+n_2=n$. Still, it is interesting to investigate the case of self-tensor product of matrices, because it is easy to calculate analytically the multifractal spectrum of the kth tensor product (k superpositions) and to compare it with the multifractal spectrum obtained from the real data.

It is now possible to obtain a spectrum of scaling exponents even from simple superpositions of frequency matrices which retain low-order genomic correlations. This will demonstrate that even a simple superposition of DNA strings, in the form of repetitions or duplications is enough to create long-range correlations. We will start with the pedagogical example presented in Eq. (7). This matrix representing uncorrelated data, with $p_A=p_T\sim0.29=\alpha$, $p_C=p_G\sim0.21=\beta$, becomes

$$M_{uncor}=\begin{pmatrix}\alpha^2&\alpha\beta&\alpha\beta&\alpha^2\\\alpha\beta&\beta^2&\beta^2&\alpha\beta\\\alpha\beta&\beta^2&\beta^2&\alpha\beta\\\alpha^2&\alpha\beta&\alpha\beta&\alpha^2\end{pmatrix}. \tag{12}$$

In matrix (12) the combination $\alpha\beta$ appears with frequency 8, the combination α^2 with frequency 4 and the combination β^2 also with frequency 4. In the case of n successive superpositions of matrix M, i.e., the tensor product of order n, it can be shown that the number of superpositions are given by

$$\begin{aligned}(2\alpha+2\beta)^{2n}&=\sum_{k=0}^{2n}\frac{(2n)!}{k!(2n-k)!}(2\alpha)^k(2\beta)^{2n-k}\\&=\sum_{k=0}^{2n}\frac{(2n)!2^{2n}}{k!(2n-k)!}\alpha^k\beta^{2n-k},\end{aligned} \tag{13}$$

where the variable k can take both even and odd values. From expression (13) it can be directly inferred that in the matrix of size $4^n\times4^n$ probabilities of strength

$$P_k=\alpha^k\beta^{2n-k} \tag{14}$$

are met as often as

$$R_k=\frac{(2n)!2^{2n}}{k!(2n-k)!}. \tag{15}$$

Using Eq. (8) it becomes now easy to calculate the exponents D_q as follows:

$$D_q=\frac{1}{q-1}\lim_{\epsilon\to0}\frac{\log\sum_{k=1}^{2n}R_kP_k^q}{\log\,\epsilon} \tag{16}$$

where the various p_{ij} in representation (8) have been grouped together in representation P_k of Eq. (16) and each one of them is found R_k times. The value of ϵ is calculated as

$$\epsilon=(1/4)^n \tag{17}$$

because in every iteration the size of the matrix increase 4 times in each direction and after n iterations the linear increase is $L1=4^n$. Thus each cell (site) has linear size proportional to the inverse of $L1$ and this corresponds to the smallest value of cell size ($\epsilon\to0$). Taking into account Eqs. (14)–(17) the values of D_q are calculated as

$$D_q=\frac{1}{q-1}\frac{\log[2^{2n}(\alpha^q+\beta^q)^{2n}]}{\log(1/4)^n}, \tag{18}$$

which ultimately simplifies to

$$D_q=\frac{1}{q-1}\left[-1-\frac{\log(\alpha^q+\beta^q)}{\log 2}\right]. \tag{19}$$

Note that

$$D_q = 2 \quad \text{for} \quad \alpha = \beta = 1/4 \tag{20}$$

which corresponds to the homogeneous and uncorrelated distribution of the 4 symbols (bps). Expression (19) can be directly computed for different values, e.g., α=0.29 and β =0.21 which approximate the human chromosome 10 and can be compared with the D_q exponents directly computed from the block distributions in the corresponding sequences.

It is also interesting to repeat the calculations Eqs. (13)–(19), using different values of single base concentrations, namely $p_A = a$, $p_C = c$, $p_G = g$ and $p_T = t$. The corresponding matrix takes then form (7) and after n successive superpositions of this matrix M the complexity of the 2D surface reads as

$$(a + c + g + t)^{2n} = \sum_{k_1,k_2,k_3=0}^{2n} \frac{(2n)!}{k_1!k_2!k_3!(2n - k_1 - k_2 - k_3)!} \times a^{k_1}c^{k_2}g^{k_3}t^{(2n-k_1-k_2-k_3)}. \tag{21}$$

From Eq. (16), with

$$P_{\{k_1k_2k_3\}} = a^{k_1}c^{k_2}g^{k_3}t^{2n-k_1-k_2-k_3} \tag{22}$$

and with

$$R_{\{k_1k_2k_3\}} = \frac{(2n)!}{k_1!k_2!k_3!(2n - k_1 - k_2 - k_3)!} \tag{23}$$

it becomes easy to calculate the exponents D_q. The resulting expression is

$$D_q = \frac{1}{q-1}\left[- \frac{\log(a^q + c^q + g^q + t^q)}{\log 2} \right], \tag{24}$$

which reduces to Eq. (19) when $a = t$ and $c = g$. Note that the calculation presented here corresponds to multiple (n) superpositions of the uncorrelated frequencies of single bps, while the block distributions used in Sec. III carry information about block correlations. We thus expect specific deviations of the spectrum calculated in Eqs. (19) and (24) with the one directly obtained using the block frequencies (Figs. 2 and 3).

B. Tensor product with nearest-neighbor correlations

In a second approximation, it is possible to calculate tensor products of the matrix $M_{ch:10}$, which retains nearest-neighbor correlations in the DNA sequences. In this case one needs to calculate the complexity of the tensor product of order n of the matrix $M_{ch:10}$. Following closely the analysis in Sec. II A one can write the complexity of the tensor product of order n as

$$(p_{aa} + p_{ac} + p_{ag} + p_{at} + p_{ca} + \cdots + p_{gg})^n = \sum_{k_1,k_2,\cdots k_{15}=0}^{n} \frac{n!}{k_1!k_2!\cdots k_{15}!(n - k_1 - k_2 - \cdots - k_{15})!} \times p_{aa}^{k_1}p_{ac}^{k_2}p_{ag}^{k_3}\cdots p_{tt}^{(n-k_1-k_2-\cdots-k_{15})}. \tag{25}$$

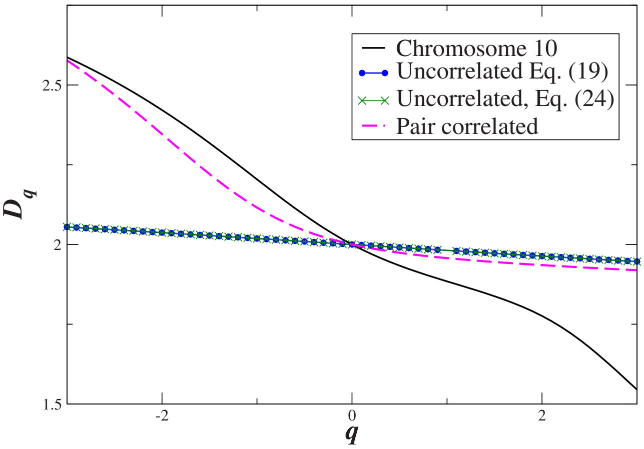

FIG. 4. (Color online) Comparison of multifractal spectrum of chromosome 10 with analytical hierarchical approaches: (a) the solid black line represents the generalized exponents D_q for human chromosome 10. (b) The blue line (with symbol circles) corresponds to a sequence of 2 symbols with probabilities α=0.29 and β=0.21, as calculated from chromosome 10 [see Eq. (19)]. c) The green line (with symbols triangles) is almost identical to the green one and is obtained from Eq. (24) with four single symbol probabilities given by Eq. (6). (d) The purple (dashed) line is obtained from Eq. (26) where pair correlations in chromosome 10 are taken into account.

The resulting spectrum of exponents is then calculated as

$$D_q = - \frac{\log(p_{aa}^q + p_{ac}^q + p_{ag}^q + p_{at}^q + p_{ca}^q + \cdots + p_{gg}^q)}{2(q-1)\log 2}. \tag{26}$$

Expression (26) retains also correlations of nearest neighbors in the DNA sequence. In Fig. 4 the three approximations to the multifractal spectrum [Eqs. (19), (24), and (26)] are presented together with the original spectrum calculated from the genomic data (chromosome 10). In particular D_q is depicted for (a) the chromosome 10 DNA sequence, (b) the approximation of single bps with equal $p_A = p_T$=0.29 and $p_C = p_G$=0.21, (c) the approximation of single bps with different values of p_A, p_T, p_C and p_G as obtained from the sequence [Eq. (6)], and (d) keeping correlations in blocks of size 2 (nearest neighbors, Eq. (26)). As depicted in Fig. 4 the case (d) which includes pair correlations describes closer the real data, in terms of the multifractal exponents D_q which account for the bps frequencies in this chromosome. Although the approximation obtained through the pair correlations resembles the DNA spectrum, it is still considered as unsatisfactory, due to ignoring correlations of higher order.

Continuing this iteration process and keeping correlations in larger and larger blocks the spectrum D_q can gradually improve, approaching better the experimental (genomic) data. Generalization of Eqs. (19), (24), and (26) to blocks of size s yields the following approximation for the multifractal spectrum:

$$D_q = - \frac{\log C_q}{2s(q-1)\log 2}, \tag{27}$$

where C_q is

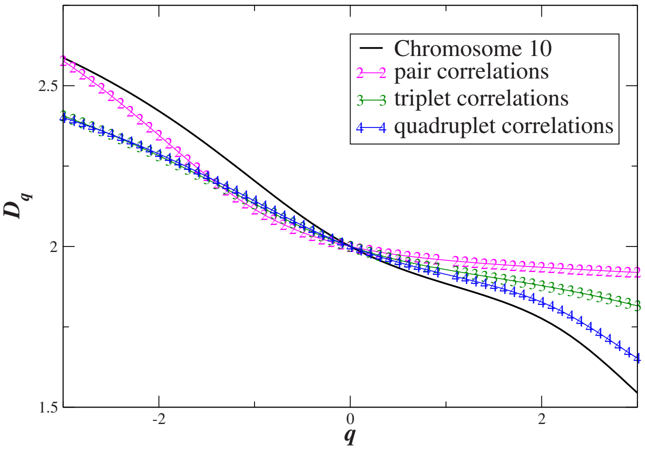

FIG. 5. (Color online) Comparison of multifractal spectrum of chromosome 10 with analytical hierarchical approaches: (a) the solid black line represents the generalized exponents D_q for human chromosome 10. (b) The purple line with symbol "2" is obtained from Eq. (26) where pair correlations in chromosome 10 are taken into account. (c) The green line with symbol "3" is obtained from Eq. (28) where third-order correlations in chromosome 10 are retained. (d) The blue line with symbol "4" is obtained from Eq. (28) where 4th order correlations in chromosome 10 are retained.

$$\begin{aligned} C_q &= p^q_{aa\cdots aa} + p^q_{aa\cdots ac} + p^q_{aa\cdots ag} + \cdots + p^q_{gg\cdots gg} \\ &= \sum_{\{i1,i2,i3,i4\}} p^q_{\{i1i2i3i4\}} \end{aligned} \tag{28}$$

$$\text{with} \quad \{i1,i2,i3,i4\} \in \{A,C,G,T\}$$

and the sum runs over all 4^s combinations of blocks. Figure 5 compares the multifractal spectrum calculated from chromosome 10 with the analytical superposition method, when retaining correlations in blocks of sizes $s=2$ (pair correlations), $s=3$ (triplet correlations), $s=4$ (quadruplet correlations), and $s=5$ (quintuplet correlations). As the size of the block (which equals the order of correlations) increases, the corresponding spectrum estimates better the one obtained directly from the data. Higher-order correlations induce numerical errors due to the small values of the various frequencies which exceed the computer precisions. The same procedure can be followed for all other chromosomes and gives qualitatively similar results. The hierarchical approach presented is based on superpositions of blocks of identical length but different composition. In principle, one could complexify further the hierarchical process by varying the length and the composition of blocks at the same time.

From the comparison of the multifractal spectra obtained through the analytical approach when retaining various orders of correlations with the actual chromosomal spectra, we understand that: (a) multiple superpositions of single bps data are not enough to reproduce the correlations of the human genomic sequences. (b) Structural correlations in the genome have been created by various evolutionary processes which cannot be accounted for using simple superpositions. (c) By taking into account higher-order genomic correlations during the analytical hierarchical approach, a closer representation of the chromosomal multifractal spectra is achieved and (d) If superpositions of multiple blocks of different lengths are taken into account then the DNA structure can, gradually, be recovered.

V. CONCLUSIONS

The 2D block correlation matrix is constructed as tensor product of block frequencies in symbol sequences. Considering this matrix as a 2D surface where the local height represents the corresponding block probability, its multifractal spectrum is calculated. The multifractal spectrum demonstrates the existence of different length scales participating in the construction of the symbolic sequence.

As an application we consider the 22 human autosomes plus the two sex chromosomes. The calculated spectra demonstrate the presence of nontrivial correlations and give substantial deviations from random uncorrelated data. These deviations can be attributed (a) to the many length scales which are involved in the evolutionary construction of the primary structure of DNA and (b) to the fact that specific sequences are not abundantly distributed but are retained for specific functional or structural tasks and thus appear with relatively low frequencies. The demonstrated multifractality may stem from the apparent superposition of different functional units carried by each genomic sequence. In addition, the superposition of segments is responsible for creating many length scales in the system and thus this method may prove valuable in shedding light to the still open problem of anomalous scaling of intron (noncoding strings) versus exon (coding) segments that was first discovered in the early 1990s.

From the deterministic theory of fractals it is known that fractal structures are mathematically created through superpositions of seed sequences. Retaining block probabilities of low order (second to fifth) we have created correlation density matrices from superpositions of these block probabilities. Due to the hierarchical nature of the process we were able to calculate exactly the multifractal spectra obtained using tensor products of matrices. It was demonstrated that this process approaches closer the actual chromosomal multifractal spectrum as the order of the retained correlations is increased. However, numerical errors prohibit the continuation of this process for very large blocks because the block probabilities decrease to even smaller values, beyond the computer precision. Using blocks of higher orders and different lengths to construct the tensor product is a way to obtain an increasingly better approximation to the genomic multifractal spectra.

This hierarchical tensor product superposition approach is generic. It can be used for the detection and the comparison of correlations in many natural and artificial symbolic sequences, such as natural languages, music, binary sequences generated by computer, protein sequences and other continuous data sequences converted to binary ones. It can provide a way to determine whether the sequence correlations are created as a result of superposition processes of shorter (lower-order) segments.

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[2] C. K. Peng, S. V. Buyldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. I. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[3] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[4] W. Ebeling and G. Nicolis, Chaos, Solitons Fractals **2**, 635 (1992).
[5] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).
[6] S. Karlin and V. Brendel Science **257**, 39 (1992); **259**, 677 (1993); S. Karlin, B. Blaisdell, R. Sapolsky, L. Cardon, and C. Burge, Nucleic Acids Res. **21**, 703 (1993).
[7] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).
[8] A. Provata and Y. Almirantis, Physica A **247**, 482 (1997); Y. Almirantis and A. Provata, J. Stat. Phys. **97**, 233 (1999).
[9] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, and H. E. Stanley, Phys. Rev. E **65**, 041905 (2002).
[10] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg, and J. L. Oliver, Phys. Rev. E **75**, 032903 (2007).
[11] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy, and C. Thermes Eur. Phys. J. B **1**, 259 (1998); Physica A **249**, 439 (1998).
[12] O. V. Usatenko, V. A. Yampol'skii, K. E. Kechedzhy, and S. S. Mel'nyk, Phys. Rev. E **68**, 061107 (2003).
[13] P. Bernaola-Galvan, J. L. Oliver, and R. Roman-Roldan, Phys. Rev. Lett. **83**, 3336 (1999).
[14] V. Afreixo, P. J. S. G. Ferreira, and D. Santos, Phys. Rev. E **70**, 031910 (2004).
[15] W. T. Li, Gene **300**, 129 (2002).
[16] P. Bernaola-Galvan, J. L. Oliver, P. Carpena, O. Clay, and G. Bernardi, Gene **333**, 121 (2004).
[17] W. T. Li and D. Holste, Fluct. Noise Lett. **4**, L453 (2004).
[18] J. Cheng and L. X. Zhang, Chaos, Solitons Fractals **25**, 339 (2005).
[19] P. W. Messer and P. F. Arndt, Nucleic Acids Res. **34**, W692 (2006).
[20] P. Katsaloulis, T. Theoharis, and A. Provata, Physica A **316**, 380 (2002); J. Theor. Biol. **258**, 18 (2009).
[21] P. Katsaloulis, T. Theoharis, W. M. Zheng, B. L. Hao, A. Bountis, Y. Almirantis, and A. Provata, Physica A **366**, 308 (2006).
[22] L. Han, B. Su, W. H. Li, and Z. M. Zhao, Genome Biol. **9**, R79 (2008).
[23] J. Freudenberg, M. Wang, Y. Yang, and W. T. Li, BMC Bioinf. **10**, S66 (2009).
[24] A. Provata and Y. Almirantis, Fractals **8**, 15 (2000).
[25] S. Garte, J. Theor. Biol. **230**, 251 (2004).
[26] M. A. Zaks, Phys. Rev. E **65**, 011111 (2001).
[27] B. Hao, Physica A **282**, 225 (2000).
[28] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed. (Garland, New York, 1994).
[29] B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982).
[30] J. Feder, *Fractals*, (Plenum Press, New York, 1988); T. Vicsek, *Fractal Growth Phenomena* (World Scientific, Singapore, 1989).
[31] H. Takayasu, *Fractals in the Physical Sciences* (Manchester University Press, Manchester, 1990).