Scaling properties and fractality in the distribution of coding segments in eukaryotic genomes revealed through a block entropy approach

Labrini Athanasopoulou, Stavros Athanasopoulos, Kostas Karamanos, and Yannis Almirantis[*]
Institute of Biology, NRCPS "Demokritos," 15310 Athens, Greece

(Received 28 July 2010; revised manuscript received 19 September 2010; published 11 November 2010)

Statistical methods, including block entropy based approaches, have already been used in the study of long-range features of genomic sequences seen as symbol series, either considering the full alphabet of the four nucleotides or the binary purine or pyrimidine character set. Here we explore the alternation of short protein-coding segments with long noncoding spacers in entire chromosomes, focusing on the scaling properties of block entropy. In previous studies, it has been shown that the sizes of noncoding spacers follow power-law-like distributions in most chromosomes of eukaryotic organisms from distant taxa. We have developed a simple evolutionary model based on well-known molecular events (segmental duplications followed by elimination of most of the duplicated genes) which reproduces the observed linearity in log-log plots. The scaling properties of block entropy $H(n)$ have been studied in several works. Their findings suggest that linearity in semilogarithmic scale characterizes symbol sequences which exhibit fractal properties and long-range order, while this linearity has been shown in the case of the logistic map at the Feigenbaum accumulation point. The present work starts with the observation that the block entropy of the Cantor-like binary symbol series scales in a similar way. Then, we perform the same analysis for the full set of human chromosomes and for several chromosomes of other eukaryotes. A similar but less extended linearity in semilogarithmic scale, indicating fractality, is observed, while randomly formed surrogate sequences clearly lack this type of scaling. Genomic sequences always present entropy values much lower than their random surrogates. Symbol sequences produced by the aforementioned evolutionary model follow the scaling found in genomic sequences, thus corroborating the conjecture that "segmental duplication-gene elimination" dynamics may have contributed to the observed long rangeness in the coding or noncoding alternation in genomes.

DOI: 10.1103/PhysRevE.82.051917 PACS number(s): 87.18.Wd, 89.75.Da, 89.70.Cf, 87.23.Kg

I. INTRODUCTION

In the early 1990s, when long DNA sequences became available, the first studies of long-range features of genomes also appeared. Li and Kaneko [1] using mutual information function and spectral analysis, Peng *et al.* [2] using the concept of the DNA walk, Voss [3] using a method identifying the appearance of $1/f$ noise, Arneodo *et al.* [4] by means of wavelet analysis, and several other research groups in the following years have found long-range correlations in the nucleotide sequences of the long noncoding regions in eukaryotic genomes.

In a previous work we have studied the size distribution of distances between coding segments in several eukaryotic and microbial genomes [5]. Power laws have been found in most cases, despite the limited length of the annotated sequences (i.e., sequences for which the coordinates of protein-coding segments are known) available at the time. Using a box-counting method, fractality has been detected in the juxtaposition of coding and noncoding segments in these sequences [6]. Recently, with many eukaryotic genomes sequenced and annotated, the complete set of chromosomes of the human genome [7] and of the genomes of several model organisms [8] have been studied and the formation of power-law-like size distributions of the noncoding spacers is found to be the rule. Considering the extent of linearity and the power-law exponent in double-logarithmic scale for different genomes of distant organisms, in combination with known differences in their genomic evolution, we have proposed an evolutionary scenario [8] which, as simulations have shown, may reproduce the power laws observed in real genomes.

Entropy-based approaches have also been used in the study of DNA sequences (see, e.g., [9–11]). In these studies it is generally observed that genomes have entropies close to their maximal value (i.e., they are quasirandom in this respect; see [9]) at least when examined at the level of their nucleotide sequence. Here, we choose to study entire chromosomes at the level of alternation of regions of different functionality (coding and noncoding segments) by means of block entropy.

II. BLOCK ENTROPY AND FRACTAL STRUCTURE

A. Modes of block entropy scaling

Let us suppose a symbol sequence of length N, with symbols taken from a binary alphabet (0, 1) and let $p_n(A_1,\ldots,A_n)$ be the probability to find the block or word $(A_1,\ldots,A_n)$ of length n in this sequence. The Shannon-like entropy or block entropy for words of length n is defined as

$$H(n) = -\sum p_n(A_1,\ldots,A_n)\ln p_n(A_1,\ldots,A_n). \qquad (1)$$

A standard treatment and description of the essential properties of block entropy and of other related quantities may be found in [12–14]. Here we briefly summarize only the results of immediate relevance to the purposes of our study.

[*] yalmir@bio.demokritos.gr

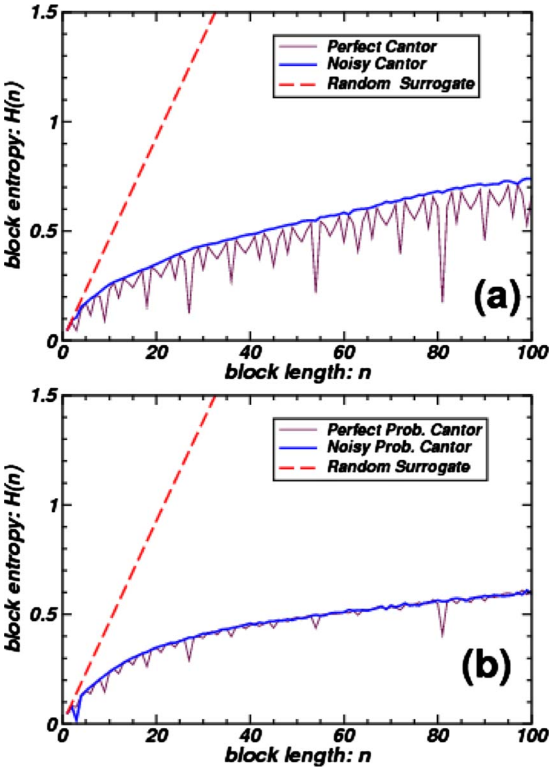

FIG. 1. (Color online) Block entropy $H(n)$ is plotted as a function of the word length n for (a) a deterministic and (b) a probabilistic ternary Cantor-like sequences of length $N=3^{13}$. In both cases, when a weak noise due to 5000 insertions and deletions is added (~0.3% of the sequence length), the nonmonotonicity of the initial curve disappears. Probabilities used for the construction of the probabilistic sequence are $p(101)=p(110)=p(011)=1/3$. $H(n)$-n curves (with marked linearity), which corresponds to random surrogate sequences of same length and number of ones are also included.

Word probabilities and related statistical quantities computed for a symbol sequence differ, depending on the choice of the way of reading. Several modes of reading may be considered. Gliding, which is the standard convention in the literature, goes by exhaustively reading all possible words of length n. This is achieved by moving the frame of length n one letter each time. Alternatively, reading by "lumping" means to take only words of length n sampled with a constant step k. In the present analysis we apply reading by the typical lumping ($k=n$). That means after reading the initial word of length n of the sequence, the next counted word is the one starting at $n+1$ and so on up to the end of the sequence. Thus, each letter of the sequence belongs only to one counted word. The terms overlapping widows and nonoverlapping windows have also been used in the literature. Lumping is chosen here because it is particularly suitable when the sequence under examination presents features such as periodicity and fractality or, more generally, exhibits an iterative structure [15,16]. When applying lumping on a symbol sequence generated by consecutive iterations, i.e., in the case of the logistic map at the accumulation point, plotting $H(n)$ against n produces a nonmonotonic plot with local minima determined by a suitable "decimation scheme" (see Eqs. 26, 27, and 31 in Ref. [15]). This nonmonotonicity is also proven for several other symbolic sequences produced algorithmically, such as the Cantor and Thue-Morse sequences [16,17]. It may be conjectured that this is a general feature of a wider class of fractal symbolic sequences, at least when the structure of the sequence obeys some form of invariance (expressed, e.g., by an exponent in a power-law size distribution).

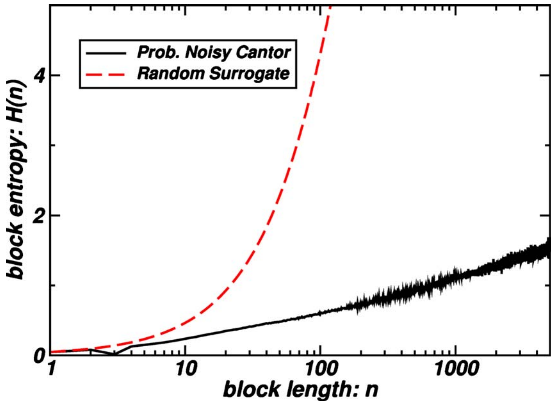

FIG. 2. (Color online) Block entropy $H(n)$ is plotted in semilogarithmic scale as a function of the word length n for a probabilistic ternary Cantor-like sequence with a weak noise added due to insertions and deletions. Sequence length, percentage of noise added, and triplet probabilities are taken as in Fig. 1(b). Linearity is clearly observed up to very high values of word length n. This feature reveals the existence of long-range correlations in the sequence structure. Also, the $H(n)$ curve for a random surrogate sequence of same length and number of ones is included.

Crucial scaling features of $H(n)$ have been investigated by several authors. Ebeling and Nicolis conjectured the following specific form for the scaling of $H(n)$,

$$H(n) = e + gn^{\mu_0}(\ln n)^{\mu_1} + nh, \tag{2}$$

for symbolic sequences generated by nonlinear dynamics including languagelike processes [13,14,18]. More specifically,

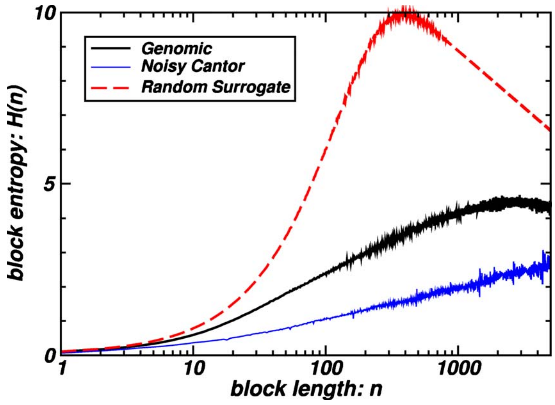

FIG. 3. (Color online) Block entropy $H(n)$ is plotted in semilogarithmic scale as a function of the word length n for human chromosome 21 with s.f.=100 (see in the text), alongside with a deterministic noisy Cantor-like sequence, with 1% indel occurrences, of (almost) equal length and number of "coding segments." Also, a common random surrogate is included.

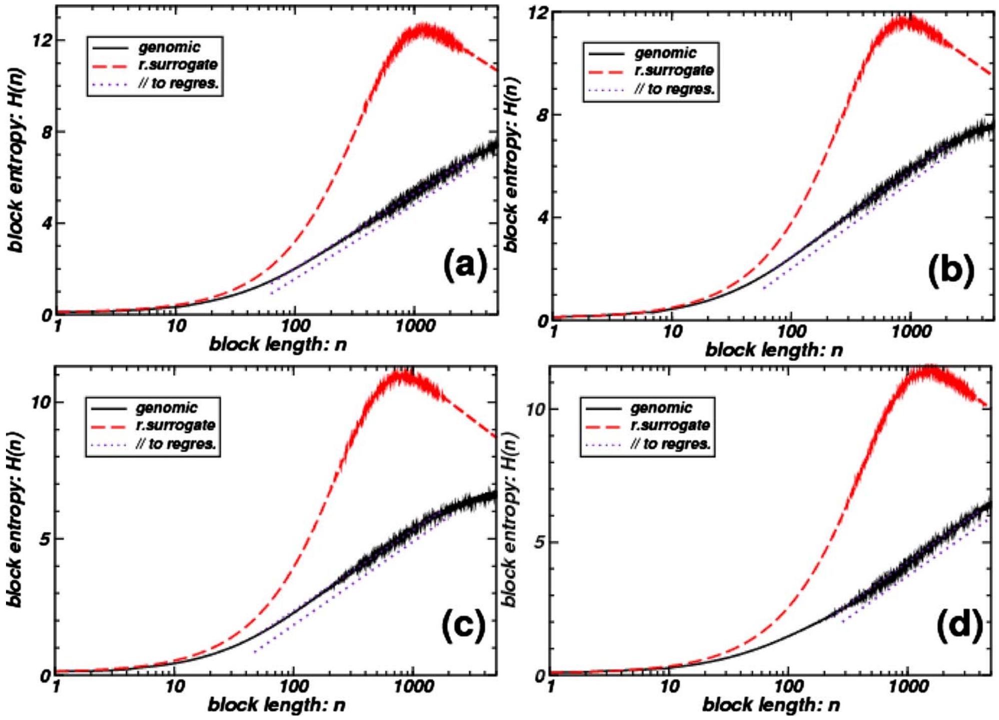

FIG. 4. (Color online) Examples of human chromosomes' block entropy $H(n)$ plots. In all cases s.f. is taken equal to 30. Random surrogates are also included. In (a)–(d) chromosomes 2, 12, 20, and X are shown, respectively. The complete set of human chromosome plots is given in the auxiliary material [27], see also Table I.

in the case of the Feigenbaum attractor for the logistic map and for $n=2^k$ ($k=2,3,4,\ldots$), Grassberger [12] (see also [15,19]) showed that for reading the sequence by gliding, the following scaling holds:

$$H(n) = \log_2(3n/2). \tag{3}$$

In this system, it is admitted that linearity in semilogarithmic plot holds (see [20] and references given therein), which in terms of Eq. (2) corresponds to $g \neq 0$, $h=0$, $\mu_0=0$, and $\mu_1 > 0$ (see [21]). This type of scaling is conjectured to hold for a large class of symbol sequences with fractal properties. Thus, the $H(n)-\log n$ linearity is related to the scale-free structure of such sequences entailing the existence of long-range correlations.

In order to test the suitability of this approach in the study of genomic sequences (estimated to present fractality in the alternation of their coding and/or noncoding segments [6]), in Sec. II B, we proceed with a preliminary study of sequences which are "by construction" fractal. Here, we consider a deterministic and a probabilistic Cantor-like symbol sequence.

B. Case of Cantor-like symbol sequences

For the construction of a ternary deterministic Cantor sequence of length 3^R our starting symbol is 1. We then apply for R consecutive times the substitutions of every 1 by 101 and of every 0 by 000. In the probabilistic case we substitute again every 0 by 000, while 1 is substituted by 110, 101, and 011 with probabilities p_1, p_2, and p_3, respectively. Then for both cases (considered as "perfect" Cantor sequences) we generate additional "noisy" sequences after making a number of random insertions and deletions of symbols. The reason for the construction of such sequences is that random insertions and deletions (termed collectively: indel) are widespread phenomena in molecular dynamics with important consequences for the genomic architecture (see, e.g., [22]). The entropy scaling for those sequences is shown in Figs. 1 and 2. Notice that in a different context, entropic analysis in symbol sequences after the addition of several forms of noise has been done initially by Freund *et al.* [23,24]. In the figures presented in the sequel, surrogate random sequences are also included, which are generated in the following way: for each sequence its surrogate is of equal length and equal number of symbols (0 and 1) that are positioned randomly.

In Figs. 1(a) and 1(b) the block entropy $H(n)$ is plotted against the block (word) length n for a deterministic and a probabilistic Cantor-like symbol sequence, respectively, alongside with their noisy counterparts. In both cases, curves for surrogate random sequences are included. First, we verify the nonmonotonicity for the graphs of the unaltered sequences as expected on the basis of analogous findings [17]. Most importantly for the analysis of genomic sequences which follows, we observe that a weak noise suffices for the disappearance of this nonmonotonic pattern and the transformation of the curve into its upper envelop. This could be expected intuitively because the nonmonotonicity is due to a "stroboscopic" combination of the scale-free structure of the sequence with increasing values of n (see explanation in [15–17]). This may tolerate probabilistic features as shown in our Fig. 1(b), but it is killed out due to multiple frame shifts caused by random insertions and deletions. We also

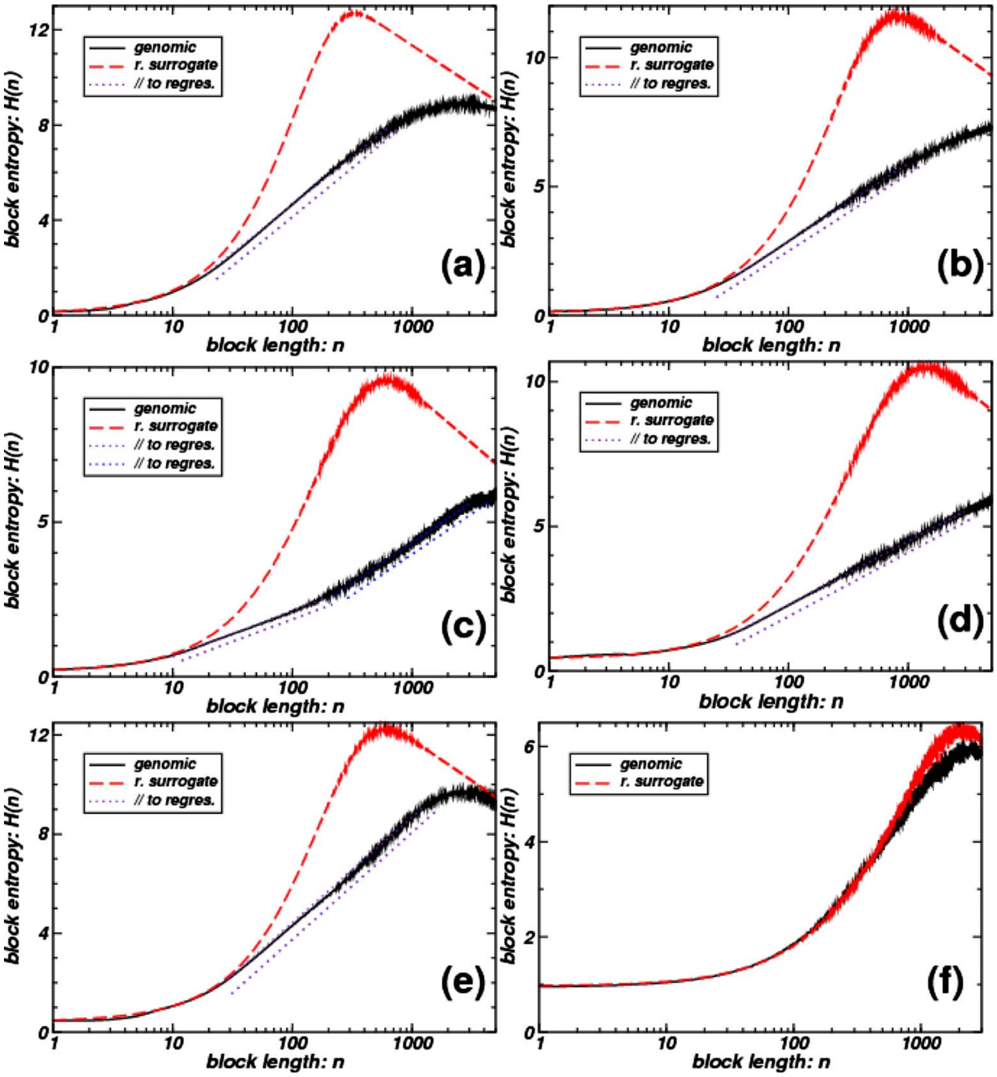

FIG. 5. (Color online) Examples of chromosomes' block entropy $H(n)$ plots from several organisms. In all cases random surrogates are included. In (a)–(f) *R. norvegicus*, chromosome 2, s.f.=100; *G. gallus*, chromosome 4, s.f.=30; *A. mellifera*, chromosome LG6, s.f.=30; *T. castaneum*, chromosome LG2, s.f.=5; *O. sativa*, chromosome 2, s.f.=10; and *S. cerevisiae*, chromosome I, s.f.=1 are shown, respectively. The plots of all examined nonhuman chromosome are given in the auxiliary material [27], see also Table II.

remark the linearity of the $H(n)$-n curve for the random surrogate sequence. This form of scaling has been shown for the Bernoulli systems (see, e.g., [13,25]).

In Fig. 2 the noisy probabilistic Cantor-like symbol sequence [shown in Fig. 1(b)] is depicted in semilogarithmic scale, alongside with its surrogate random sequence. The corresponding figure for the deterministic Cantor-like construction is omitted, as it is qualitatively similar and almost identical with the presented plot. We verify here that Cantor-like sequences exhibit the expected linearity in semilogarithmic scale in $H(n)$-n graphs, which is shown to occur in other cases of fractal-like symbol sequences, as discussed in Sec. II A.

Here, we have to emphasize the inclusion of high values of word length n in the graphs presented in Fig. 2 and in the figures including genomic data presented in Sec. III. Thus, the range of n is extended much further than the limit usually set in order to guarantee a good statistics in a random symbol sequence and avoid finite-size effects. This limit is typically taken equal to $n_0=\log_2 N$ if N is the length of a binary sequence in order to have assured the presence of all possible words of length n in the sequence. This is done on the basis of two assumptions: (i) the equiprobability of all possible words in random sequences and (ii) that the sequence has to include each possible word at least once [26]. However, the approximative estimation that a sequence of length 2^n includes each n word once is compromised, especially when nonequiprobability of symbols and fractality hold. There, the number of "frequent" words (words contributing the most in the value of entropy; see [14]) depends strongly on the underlying scale-free pattern. Consequently, in the present study, we have chosen to include high values of n. We observe a long extent of linearity in the semilogarithmic plot for the fractal-like symbol sequences. This applies both in the Cantor-like constructions, like that of Fig. 2 and in the genomic sequences presented in Sec. III. Notice the distortion (leading to the appearance of a maximum) of the random surrogate symbol sequence, which, however, occurs at

TABLE I. Quantitative information for all human chromosomes. The extent of the linear region in semilogarithmic scale and the slope are included. The square of the correlation coefficient resulting from a logarithmic regression analysis is in all cases (in this table and in Table II) higher than 0.98. s.f. equals 30 in all chromosomes. Chromosome lengths are given in Mbp (millions of nucleotides).

Chromosome (*H.sapiens*)	Chromosome length	Number of coding segments	Coding percent	Extend of linearity	Slope
1	247.2	26121	1.77	1.42	3.73
2	242.7	19861	1.39	1.76	3.30
3	199.4	15063	1.23	1.79	3.15
4	191.2	9491	0.87	1.38	2.92
5	180.1	11493	1.16	1.73	2.86
6	170.7	12585	1.28	1.33	3.48
7	158.6	12578	1.31	1.39	3.52
8	146.3	8749	1.07	1.64	2.92
9	140.2	10134	1.24	1.70	2.67
10	135.3	11876	1.35	1.36	3.38
11	134.3	14328	1.85	1.43	3.33
12	132.3	13752	1.64	1.62	3.34
13	114.1	4197	0.66	1.78	2.26
14	106.4	8330	1.35	1.79	2.62
15	100.3	9624	1.60	1.64	2.86
16	88.8	10764	2.01	1.82	2.67
17	78.7	15898	3.29	1.45	3.96
18	76.1	3775	0.86	1.60	2.79
19	63.8	13646	3.95	1.45	4.10
20	62.4	6952	1.80	1.84	2.92
21	46.9	3429	1.14	2.03	1.75
22	49.6	6030	2.04	1.64	2.72
X	154.9	9802	1.16	1.45	3.20
Y	57.8	810	0.24	1.13	0.37

values much higher than the aforementioned limit $n=n_0$. This limit does not apply here due to the strong inequality of zero and one populations, although surrogate data lack by construction any internal order. The inclusion of lengthy words in our analysis ($n \gg n_0$) is important for the study of genomic sequences where linearity appears in an intermediate range of word lengths, as we discuss further in Sec. III.

III. SCALING PROPERTIES OF BLOCK ENTROPY IN GENOMIC SEQUENCES

For the study of the coding or noncoding segments' alternation in entire eukaryotic chromosomes we proceed in the following way. We downloaded data of genomic annotation for several genomes from the National Center for Biotechnology Information ftp site (for further details see in the auxiliary material [27]). Using these data, we pass from the sequence written in the four letter alphabet A, G, C, and T to the two numbers (0, 1) where "1" stands for every coding nucleotide and "0" for every noncoding one. The downloaded data include coordinates of every coding segment in a chromosome, thus allowing the described construction.

Now, notice that the smallest coding segments (coding exons) are of a length of a few tens of nucleotides. In order to avoid the study of short words formed in the region of a few nucleotides not contributing meaningfully to the entropy scaling (because alternation of coding and noncoding is not expected there), we introduce a "shrinkage factor" (s.f.) allowing compression of the genomic sequence. Thus, we proceed in the following way: for s.f. equal to, e.g., 30 symbols, we start from the beginning of the chromosome and we substitute every 30 zeros (0) by one zero and every 30 units (1) by one unit. When we meet a 30-letter string of mixed composition we substitute it by a single 1. Notice that the coding segments consist of a population of small dispersed parts, in comparison to noncoding intervening sequences in the eukaryotic genome (coding space spans only ~1.5% of the human genome). Thus, in our approach ones correspond to the almost "zero-measure" component of a Cantor-like construction. In this way, we perform a "coarse graining," mainly retaining the alternation of the coding and noncoding segments. We have tested a variety of values of s.f., ranging from 10 to 10 000 (see example curves in the auxiliary material [27]). We chose to present our results of *H. sapiens* for

TABLE II. Quantitative information for all nonhuman chromosomes treated herein. Column content is similar to that of Table I, while in the extra column are given the used shrinkage factors.

Organism, chromosome	Chromosome length	Number of coding segments	Coding percent	Extend of linearity	Slope	s.f.
B. taurus, chromosome 9	95.0	3996	10.75	1.48	2.65	30
B. taurus, chromosome 20	68.3	2911	0.72	1.43/0.93	2.69/1.96	30
M. musculus, chromosome 3	166.7	32195	3.28	1.06	4.44	50
M. musculus, chromosome 5	152.0	43574	4.93	0.82	6.06	30
M. musculus, chromosome 12	120.5	23455	3.43	1.00	4.90	30
M. musculus, chromosome 15	106.3	28923	4.73	0.90	4.86	50
M. musculus, chromosome 19	63.5	23033	6.09	0.95	5.32	20
R. norvegicus, chromosome 2	258.1	26000	1.72	1.56	4.05	100
R. norvegicus, chromosome 3	171	30705	3.23	1.26	4.77	50
R. norvegicus, chromosome 16	90.1	11236	2.12	1.57	3.22	30
G. gallus, chromosome 4	94.2	10869	1.94	1.94	3.02	30
G. gallus, chromosome 11	21.9	3588	2.71	1.74	2.30	10
D. melanogaster, chromosome 2L	23.0	15883	26.8	0.70	4.90	10
D. melanogaster, chromosome 2R	21.1	19711	32.8	0.78	5.34	10
A. mellifera, chromosome LG2	16.0	3201	4.85	1.34/0.90	2.67/1.89	30
A. mellifera, chromosome LG6	17.7	2524	3.39	1.40/0.76	2.62/1.47	30
O. sativa, chromosome 2	35.9	15276	10.2	1.67	4.26	10
O. sativa, chromosome 6	30.7	10161	8.44	1.33/0.67	4.19/3.43	10
O. sativa, chromosome 12	27.6	7080	6.52	0.97/0.84	2.67/3.62	10
A. thaliana, chromosome 4	18.6	25215	31.25	Only traces of linearity		5
A. thaliana, chromosome 5	27.0	38632	32.73	Only traces of linearity		10
T. castaneum, chromosome LG1	8.1	2275	8.05	0.82	1.82	5
T. castaneum, chromosome LG2	12.9	4510	9.47	2	2.19	5
T. castaneum, chromosome LG8	15.8	4629	9.19	0.88/0.77	2.10/2.56	5
S. cerevisiae, chromosome I	0.23	97	62.3			1
S. cerevisiae, chromosome IV	1.53	793	73.82			1

s.f.=30, length which roughly corresponds to the shortest coding exons in the human genome.

In Fig. 3 we plotted $H(n)$ versus n for the human chromosome 21 alongside with a "noisy" Cantor-like construction, the length, and percentages of ones of which almost coincide with those of the chromosomal sequence. We chose to compare entropic scaling of genome sequences with the one of noisy Cantor-like sequences. This is done because insertions and deletions occur regularly in the noncoding regions, while coding space is highly conserved due to its role, which is crucial for the organism's survival. Insertions and deletions considered here comprise several molecular (genomic) events, such as formation of clusters of similar nucleotides (mostly in the noncoding) [28,29], usually due to slippage errors during replication [30,31], as well as insertions (and less often deletions) of the so-called repeated elements [32,33] which, in many genomes, consists of a large part of the genome (~45% in human genome). These genomic modifications happen in the slow evolutionary time (slow if compared to the "fast" time of individual organism's life span). Thus, a continuous short-distance (irregular) shift of the coordinates of coding segments underlies genome structure generating a "noise" analogous to the one added in the Cantor structures depicted in previous figures. Due to this property, the nonmonotonicity discussed in Sec. II B is not found in genomic sequences even if part of the underlying genome dynamics is perhaps characterized by some sort of multiplicative processes.

Inspection of Fig. 3 shows that the genomic sequence presents linearity in a semilogarithmic plot, although shorter than that observed for the Cantor-like sequence, as it deviates from linearity for both limits of short and very long words. Linearity starts at $n \cong 15$ and ends at $n \cong 900$, which corresponds (given the s.f. value, which here equals to 100) to noncoding spacers of lengths ~1500 and ~90 000 nucleotides, respectively. This result has its counterpart in the size distribution of the distances between coding segments in eukaryotic genomes, which are found to be power-law-like [8]. Notice that these distances correspond to runs of zeros in our symbol sequences, which in the case of the Cantor-like sequences follow strictly a power-law statistics. The curve of the surrogate random sequence lacks a considerable linear part, while the finite-size effects start at relatively low n values, as already mentioned.

In Figs. 4 and 5 examples of the block entropy scaling for chromosomes of *H. sapiens* and of a collection of several

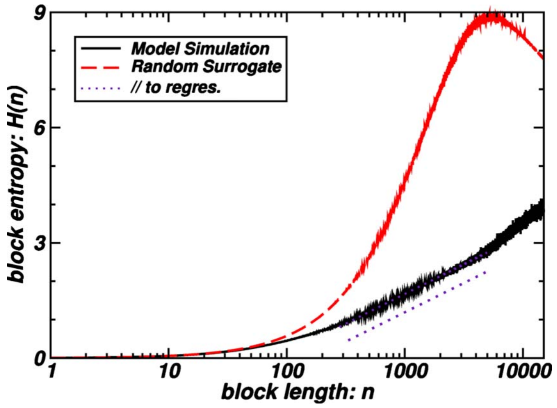

FIG. 6. (Color online) Block entropy $H(n)$ is plotted in semilogarithmic scale as a function of the word length n for a sequence generated by the "segmental duplication-gene elimination" model. For details, see in the text.

other organisms are shown, respectively. In Tables I and II quantitative features of the full set of such plots are included for all examined chromosomes. In the auxiliary material [27] the full set of these plots is provided. It is obvious as a general trend that linearity of the entropy scaling in semilogarithmic plots is typical for eukaryotic chromosomes seen at the level of coding or noncoding alternation. Deviations from this trend are always observed at the limits of short and very lengthy words.

For high coding percentages, this linearity is considerably restrained (cases of *A. thaliana* and *D. melanogaster*), while it completely disappears in "coding dense" genomes, like in *S. cerevisiae* (baker's yeast).

IV. SCALING PROPERTIES OF THE BLOCK ENTROPY FOR THE "SEGMENTAL DUPLICATION–GENE ELIMINATION" MODEL

An "expansion-modification" model, expanding a sequence and simultaneously generating long-range correlations in its nucleotide composition, has been proposed by Li [34] for noncoding DNA where the expansion process rewrites one symbol to two identical symbols and the modification process switches one symbol to another symbol. More recent findings on strand slippage during replication combined with point mutations shed light into homonucleotide tracts and microsatellites' evolution and may be a realistic implementation of the above model to genome dynamics (see [30,31] and references given therein). It has to be shown if simulations of the combination of nucleotide-clusters' expansion with point mutations do generate long-range correlations (as the expansion-modification model does) and whether this dynamics may lead to intercoding spacers' power lawlike size distributions and entropic scaling similar to that of genomic sequences. Thus, the mechanism described in [34] probably contributes significantly to fractal properties of the size distribution of noncoding genomic spacers.

Buldyrev *et al.* [35] also proposed an "insertion-deletion" model for the explanation of long-range correlations in the nucleotide composition of noncoding DNA. However, the production of long-range correlations in this model depends on the assumption of the occurrence of transpositions and insertions of DNA stretches, whose lengths are chosen from a power-law distribution. This is justified on the grounds of the theoretical prediction of such size distributions in loop formation in "very dilute solutions, i.e., isolated polymers" (see [36]). The extrapolation of such a prediction to the dynamics of chromatin inside the nucleus would require additional experimental or theoretical evidence in order to be supported.

Recently, we have proposed an evolutionary mechanism for the explanation of the linearity in log-log plots observed in the size distribution of the noncoding regions in many eukaryotic genomes [8]. This mechanism has been based on a model proposed by Takayasu and coworkers [37] explaining fractal structures obtained by aggregative growth in physicochemical systems. The proposed evolutionary mechanism includes well-studied events of the genomic dynamics. These are (i) *segmental duplications* (ubiquitous in all studied eukaryotic genomes). These are regions of the genome (each may include several genes), which are randomly copied, and the copy is reinserted in a new position (again randomly); (ii) the subsequent (in the slow course of evolutionary time) *elimination of most of the duplicated genes*, while some of the duplicated genes may survive if they gain a new functional role [38–40]. In Ref. [8] a detailed analysis based on the related biological literature is provided for the justification of the specific types of events included in the proposed mechanism. Notice that the model described in [37] is analytically solvable, while the "segmental duplication-gene elimination" model proposed for the genomic evolution may be studied only by means of computer simulations.

In this section we test if this model may also generate the pattern of entropic scaling studied thus far. In the simulation presented in Fig. 6, the model acted upon a sequence of initial length 2×10^6, mainly formed by zeros (0), which includes 1000 randomly distributed short islands of ones (1) representing coding segments. Then, 84 events of segmental duplication of mean length equal to 10% of the sequence length occurred until a final length of 10^8 symbols was reached. Hence, 90% of the duplicated coding segments (included in the duplication regions) have been deleted randomly (this percentage is inferred by the biological literature [38–40]). In Fig. 6 the $H(n)$-n plot in semilogarithmic scale for a model-generated sequence is depicted, alongside with its random surrogate. Linearity in semilogarithmic scale is observed only for the curve corresponding to the simulated sequence. The size distribution of the distances between localizations (denoting coding segments), corresponding to the sequence presented in Fig. 6, may be found in Fig. 4(b), in Ref. [8] where a clear-cut linearity in double-logarithmic scale is formed.

Notice that the same qualitative picture is obtained for a variety of choices for the model parameters. In some of the simulations we have also included other events which are common in genomic dynamics [repeat insertions, intrachromosomal translocations, etc. (see [8])] again without qualitative changes in the emerging picture (figures not shown).

This result corroborates the hypothesis that the interplay of segmental duplications and gene eliminations may be an important component for the generation of long-range correlations and fractal features in the eukaryotic genome at the coding or noncoding level.

V. DISCUSSION

In the present paper we study how the block (Shannon) entropy scales with the word length in genomic sequences when focusing on the coding or noncoding structure of eukaryotic chromosomes. Initially, we studied binary symbol sequences constructed following a Cantor pattern. It is concluded that the Cantor structure entails nonmonotonic scaling of the block entropy when the sequence is read by lumping. However, this nonmonotonicity does not persist when a weak noise, having the form of symbols' insertions and deletions, is added. Furthermore, linearity in semilogarithmic scale is observed in $H(n)$-n plots, extended in the region of high values of n. For reasons of comparison, random surrogate symbol sequences have been included in each case. No linearity in semilogarithmic scale is observed in the random surrogates and their block entropy values are always considerably higher than the values of their genomic counterparts. Instead, block entropy scales linearly in the original plot in accordance with the cited literature until finite-size effects appear. Then, we examined genomic sequences of eukaryotic origin, transformed to binary symbol sequences, with one and zero denoting the protein-coding or noncoding characters of each nucleotide, respectively. Resemblance has been found between the scaling properties of genomic sequences and of sequences which are fractal by construction. Our findings indicate that the block entropy of symbol sequences is a suitable tool for the analysis of fractality or self-similarity features, even in cases of sequences highly "imperfect" and noisy, as is the genomic DNA. Using a model based on well-known events of genomic dynamics [8], we have reproduced the qualitative features of the entropic scaling of genomic sequences. The proposed evolutionary scenario implies that fractality and long rangeness, at least at the level of coding or noncoding structure, can emerge as results of genomic dynamics.

The deviation of the genomic $H(n)$-n curve from linearity in semilogarithmic plot at the limit of low word length (see Figs. 3–5) may be understood given that short introns or intergenic regions are denser in regulatory sequences than other regions and therefore have to be under evolutionary constraints. This means that the organism's viability would be affected if these short spacers are subject to the molecular dynamics which eventually drives the rest of the genome to the quasifractal structure generating the linear entropic scaling adopted by the lengthier spacers. At the limit of the very lengthy words we observe again deviation from linearity, which can be related ultimately to finite-size effects.

The question of existence of eventual benefits for the organism from a fractal-like genomic organization as the one described herein remains open. Recent findings obtained using powerful experimental techniques combined with computational treatment (the Hi-C method; see [41]) shed light to the spatial arrangement of the genome in the very confined condition of the eukaryotic nucleus and show that the genome very probably adopts the so-called form of the "fractal globule." Such a structure for the genome inside the nucleus has been predicted theoretically [42,43] and offers important benefits to cellular functioning. More specifically, it facilitates the quick and repetitive transcriptional switching on and off of specific genes, possibly in a coordinated way when genes cooperate in the same cellular task even if they are separated by large distances intrachromosomally or they belong to different chromosomes. In addition, the structure of the fractal globule allows the repetitive winding and unwinding of the genomic thread during the consecutive cell cycles, each cycle mediated by complete replication of the genomic material. A knot-free structure [44] would enormously facilitate this function. Notice that quantitative results from the application of the Hi-C and three-dimensional fluorescence *in situ* hybridization methods verify the predicted scaling features of the fractal globule [41,45].

A large amount of results about fractality and long-range order in the genome are actually available. Only indicatively we could refer to (i) the clustering of similar nucleotides generating long-range correlations in the nucleotide constitution of large noncoding eukaryotic sequences [1–3,46,47]; (ii) the findings presented herein and in related works about the structure generating by coding or noncoding alternation [5–8]; (iii) the pattern observed in the distribution of repeats in eukaryotic genomes which is shown to be the product of genomic dynamics in evolutionary time [48]; (iv) the power-law size distribution of the isochores (regions of relatively homogeneous G+C content) (see [49] and supplementary Fig. 2 of Ref. [50]); and (v) long-range correlations and fractality due to the nucleosomal structure, strand-asymmetry, replication origins' distribution, and localization of other genomic functional or structural sites (see [51–54] and other works of the same group).

These geometrical genomic features, often found at the level of the sequence primary structure, have been probably used by natural selection for the formation and maintenance of the overall fractal structure of the nuclear content (the fractal globule) which itself bears concrete functional roles. The recruitment (exaptation) of structures and features, which initially emerged accidentally, into functions and roles crucial for the survival and development of organisms, occurs frequently with important repercussions for biological evolution.

ACKNOWLEDGMENTS

The authors would like to thank Dr. A. Provata and D. Sellis for helpful discussions, Professor K. Eftaxias and Dr. C. Nikolaou for critical reading of the paper, and the anonymous referees for suggestions which have considerably improved the presented work.

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[2] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[3] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[4] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[5] Y. Almirantis and A. Provata, J. Stat. Phys. **97**, 233 (1999).

[6] A. Provata and Y. Almirantis, Fractals **8**, 15 (2000).

[7] A. Provata and T. Oikonomou, Phys. Rev. E **75**, 056102 (2007).

[8] D. Sellis and Y. Almirantis, Gene **447**, 18 (2009).

[9] A. O. Schmitt and H. Herzel, J. Theor. Biol. **188**, 369 (1997).

[10] M. A. Jiménez-Montaño, W. Ebeling, T. Pohl, and P. E. Rapp, BioSystems **64**, 23 (2002).

[11] J. Kim, S. Kim, K. Lee, and Y. Kwon, Chaos, Solitons Fractals **39**, 1565 (2009).

[12] P. Grassberger, Int. J. Theor. Phys. **25**, 907 (1986).

[13] W. Ebeling and G. Nicolis, Europhys. Lett. **14**, 191 (1991).

[14] W. Ebeling and G. Nicolis, Chaos, Solitons Fractals **2**, 635 (1992).

[15] K. Karamanos and G. Nicolis, Chaos, Solitons Fractals **10**, 1135 (1999).

[16] K. Karamanos, J. Phys. A **34**, 9231 (2001).

[17] K. Karamanos, Kybernetes **38**, 1025 (2009).

[18] W. Ebeling, J. Freund, and K. Rateitschak, Int. J. Bifurcation Chaos Appl. Sci. Eng. **6**, 611 (1996).

[19] W. Ebeling, in *ICCS 2002, LNCS 2331*, edited by P. M. A. Sloot *et al.* (Springer-Verlag, Berlin, 2002), p. 1209.

[20] W. Ebelings and K. Rateitschak, Discrete Dyn. Nat. Soc. **2**, 187 (1998).

[21] G. Nicolis and P. Gaspard, Chaos, Solitons Fractals **4**, 41 (1994).

[22] D. A. Petrov, Theor Popul. Biol. **61**, 531 (2002).

[23] J. A. Freund and K. Rateitschak, Int. J. Bifurcation Chaos Appl. Sci. Eng. **8**, 933 (1998).

[24] J. A. Freund, W. Ebeling, and K. Rateitschak, Phys. Rev. E **54**, 5561 (1996).

[25] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos, Solitons Fractals **4**, 97 (1994).

[26] A. O. Schmitt, H. Herzel, and W. Ebeling, Europhys. Lett. **23**, 303 (1993).

[27] See supplementary material at http://link.aps.org/supplemental/10.1103/PhysRevE.82.051917 for auxiliary material.

[28] Y. Almirantis and A. Provata, Bull. Math. Biol. **59**, 975 (1997).

[29] Y. Almirantis, J. Theor. Biol. **196**, 297 (1999).

[30] K. J. Dechering, K. Cuelenaere, R. N. Konings, and J. A. Leunissen, Nucleic Acids Res. **26**, 4056 (1998).

[31] H. Gragg, B. D. Harfe, and S. Jinks-Robertson, Mol. Cell. Biol. **22**, 8756 (2002).

[32] J. Jurka, O. Kohany, A. Pavlicek, V. V. Kapitonov, and M. V. Jurka, Proc. Natl. Acad. Sci. U.S.A. **101**, 1268 (2004).

[33] J. Jurka, V. V. Kapitonov, O. Kohany, and M. V. Jurka, Annu. Rev. Genomics Hum. Genet. **8**, 241 (2007).

[34] W. Li, Phys. Rev. A **43**, 5240 (1991).

[35] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993).

[36] J. des Cloizeaux, J. Phys. (France) **41**, 223 (1980).

[37] H. Takayasu, M. Takayasu, A. Provata, and G. Huber, J. Stat. Phys. **65**, 725 (1991).

[38] J. A. Bailey *et al.*, Science **297**, 1003 (2002).

[39] M. Lynch and J. S. Conery, Science **290**, 1151 (2000).

[40] A. McLysaght, K. Hokamp, and K. H. Wolfe, Nat. Genet. **31**, 200 (2002).

[41] E. Lieberman-Aiden, N. L. van Berkum, L. Williams *et al.*, Science **326**, 289 (2009).

[42] A. Yu. Grosberg, S. K. Nechaev, and E. I. Shakhnovich, J. Phys. France **49**, 2095 (1988).

[43] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Europhys. Lett. **23**, 373 (1993).

[44] O. A. Vasilyev and S. K. Nechaev, Theor. Math. Phys. **134**, 142 (2003).

[45] J. Mateos-Langerak, M. Bohn, W. deLeeuw *et al.*, Proc. Natl. Acad. Sci. U.S.A. **106**, 3812 (2009).

[46] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[47] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **52**, 2939 (1995).

[48] D. Sellis, A. Provata, and Y. Almirantis, Mol. Biol. Evol. **24**, 2385 (2007).

[49] N. Cohen, T. Dagan, L. Stone, and D. Graur, Mol. Biol. Evol. **22**, 1260 (2005).

[50] M. Costantini, O. Clay, F. Auletta, and G. Bernardi, Genome Res. **16**, 536 (2006).

[51] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).

[52] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, J. Mol. Biol. **316**, 903 (2002).

[53] S. Nicolay, E. B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Phys. Rev. Lett. **93**, 108101 (2004).

[54] S. Nicolay, E. B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Phys. Rev. E **75**, 032902 (2007).