Power law exponents characterizing human DNA

A. Provata[1,*] and Th. Oikonomou[1,2]
[1]*Institute of Physical Chemistry, National Center for Scientific Research "Demokritos," 15310 Athens, Greece*
[2]*Department of Biological Chemistry, School of Medicine, University of Athens, 11527 Athens, Greece*
(Received 7 February 2006; revised manuscript received 9 February 2007; published 3 May 2007)

The size distributions of all known coding and noncoding DNA sequences are studied in all human chromosomes. In a unified approach, both introns and intergenic regions are treated as noncoding regions. The distributions of noncoding segments $P_{nc}(S)$ of size S present long tails $P_{nc}(S) \sim S^{-1-\mu_{nc}}$, with exponents μ_{nc} ranging between 0.71 (for chromosome 13) and 1.2 (for chromosome 19). On the contrary, the exponential, short-range decay terms dominate in the distributions of coding (exon) segments $P_c(S)$ in all chromosomes. Aiming to address the emergence of these statistical features, minimal, stochastic, mean-field models are proposed, based on randomly aggregating DNA strings with duplication, influx and outflux of genomic segments. These minimal models produce both the short-range statistics in the coding and the observed power law and fractal statistics in the noncoding DNA. The minimal models also demonstrate that although the two systems (coding and noncoding) coexist, alternating on the same linear chain, they act independently: the coding as a closed, equilibrium system and the noncoding as an open, out-of-equilibrium one.

DOI: 10.1103/PhysRevE.75.056102 PACS number(s): 89.75.Da, 87.14.Gg, 05.40.−a, 87.15.Aa

I. INTRODUCTION

With the rapid growth of genomic data in the past decade, it now becomes possible to analyze the complexity and diversity in the primary structure of DNA and to compare between chromosomes of the same species and across different classes of organisms. An aspect of genomic complexity which has attracted considerable attention since the early 1990s is the discovery of long-range correlations in genomic sequences. The long-range correlations were initially demonstrated using the $1/f$-spectrum analysis and the DNA walk model of nucleotide sequences [1–3]. This demonstration was based on partial sequences of higher eucaryotes available at the time or on complete sequences of organisms with limited genomes (procaryotes and viruses). Today, it is possible to analyze entire chromosomes in most classes of organisms and thus to look for global as well as local statistical characteristics.

For the analysis of genomic correlations, the understanding of their origins and their development during evolution, and the search for common and different features in the DNA of different classes of organisms, a variety of theoretical, statistical, and numerical approaches are used, ranging from linguistic and entropic approaches to wavelets [4–12]. The outcome of these studies on many classes of organisms has resulted in detecting nontrivial statistical characteristics, such as patchiness, nonhomogeneity, extensive repetitions, and long-range, short-range, and fractal features. In particular, in recent studies of the size distribution (SD) of coding and noncoding sequences, one of the authors (A.P.) and collaborators have explored the genomes of diverse organisms (*Homo sapiens, Drosophila Melanogaster, C. elegans, Arabidopsis thaliana, S. cerevisiae*, various procaryotes and viruses). In all cases, short-range correlations were detected in the coding (exon) size distributions, while long-range correlations were obtained for the noncoding size distributions of evolutionary newer organisms [10]. Traces of long-range behavior were also found in the noncoding size distributions of some procaryotes and viruses. Similar types of correlations were shown in the distance distribution between promoters [13].

The aim of the current study is to explore the existence of long-range correlations globally in the human genome, which is now nearly complete and available in the international databases. As will be explained in the next section, in this approach we choose to divide the human genome in two categories of segments: (a) segments which may support extensive or moderate modifications during evolutions and these are the intergenic regions and the introns, respectively, and (b) segments which cannot be modified without damaging or changing drastically the cell functions and these are the coding regions which are identical with the exons situated within the genes. In the rest of this study the terms "exon" and "coding segments" ("regions") will be used alternatively to indicate the segments which are translated for the production of proteins. As shown in Sec. II, the coding (exon) and noncoding size distributions have striking statistical differences, mirrored in the corresponding correlations, and these features are common in all chromosomes. To understand the emergence of the different types of distributions and correlations, minimal, stochastic, mean-field models are proposed in the last part of this study, which take into account elementary evolutionary processes. The proposed models, on the one hand, generate the correlations observed in the human genome and, on the other hand, connect the coding and noncoding formation and evolution with closed (equilibrium) and open (out-of-equilibrium) systems, respectively.

In the next section statistical characteristics of the human data are first presented as obtained from the international databases and the size distributions of coding (exon) and noncoding sequences are studied in a unified way. A pure power law is observed to characterize the noncoding size

*Author to whom correspondence should be addressed. Electronic address: aprovata@limnos.chem.demokritos.gr

TABLE I. Characteristic features and exponents describing the coding and noncoding SDs of the human genome.

				Coding		Noncoding	
Chromosome No.	Size (10^8 bps)	Seq. (%)	Coding (%)	α_c ($\times 10^{-3}$)	μ_c	α_{nc} ($\times 10^{-7}$)	μ_{nc}
1	2.45	90.8	1.41	0.32	1.85	8.0	0.92
2	2.43	97.8	1.03	0.10	1.80	7.0	0.89
3	2.00	97.6	0.98	0.10	1.95	7.0	0.83
4	1.91	97.8	0.72	0.15	1.80	6.0	0.82
5	1.81	98.3	0.87	0.15	1.75	7.0	0.82
6	1.71	97.9	1.04	0.30	1.80	7.0	0.82
7	1.59	97.6	1.05	0.05	1.85	12.0	0.82
8	1.46	97.5	0.80	0.10	1.85	9.0	0.80
9	1.38	85.1	1.01	0.15	1.65	6.0	0.85
10	1.35	97.2	1.08	0.13	1.85	9.0	0.82
11	1.34	97.5	1.49	0.10	1.95	9.0	0.86
12	1.32	98.4	1.32	0.10	2.00	9.3	0.83
13	1.14	83.7	0.58	0.08	1.80	7.0	0.71
14	1.06	83.0	1.04	0.10	1.95	0.5	1.05
15	1.00	81.1	1.24	0.13	1.85	9.0	0.85
16	0.89	88.8	1.67	0.18	1.85	7.0	0.85
17	0.79	98.9	2.52	0.20	1.95	30.0	0.90
18	0.76	98.1	0.72	0.02	1.75	7.0	0.81
19	0.64	87.4	3.42	0.02	1.95	2.0	1.20
20	0.62	95.3	1.37	0.25	1.90	10.0	0.89
21	0.47	72.8	0.75	0.05	2.00	6.0	0.74
22	0.50	70.2	1.52	0.05	1.90	0.5	0.98
X	1.55	97.1	0.87	0.15	1.80	9.0	0.73
Y	0.58	43.1	0.22	0.35	1.50	20.0	0.40

distributions while in the coding the exponentially decaying terms dominate. In the same section a short discussion is devoted to the size distribution of introns and intergenic regions. In Sec. III a unified model is presented, based on aggregative formation of DNA from macromolecular segments, which accounts for the observed short- and long-range characteristics of coding and noncoding DNA, respectively. Finally, the main conclusions of this work are drawn and open issues are addressed.

II. CODING AND NONCODING SIZE DISTRIBUTION STATISTICS

The human genomic data were obtained from the GenBank. In this database the primary DNA structure is given together with information about functionality, such as precise positions of coding segments (CDS), mRNA positions, genes, repetitive elements, etc. Some of these characteristics are collectively shown in Table I for comparison.

In column 2 of Table I, the size of the human chromosomes is given in numbers of base pairs (bps). The largest chromosomes are the first in order, and their size decreases from top to bottom (with the exception of the last three chromosomes 22, X, and Y). In the GenBank database each base pair (bp) is denoted as A (adenine), G (guanine), C (cytosine), T (thymine), or N (unidentified). The unidentified bps may be pairs completely unknown (resisting to sequencing) or partly known (e.g., known to be purines or pyrimidines) or the various sequencing laboratories disagree on their identification. In column 3 of Table I, the percentage of identified bps is presented, called "Seq. %." This varies from 98.7% in chromosome 17 to 43% for chromosome Y. In the fourth column the coding percent, known up to date, of each chromosome is recorded. This is as low as 0.72% for chromosomes 4 and 18, with a maximum value of 3.42% for chromosome 19.

The size distributions of exons and introns have been the subject of many previous investigations [14–18]. In this study we choose, mainly, to treat in a uniform way all sequences which do not code for protein production. Thus the noncoding sequences include both introns and regions between consecutive genes (intergenic regions). This unified approach of the noncoding is followed because all noncoding DNA segments have common characteristics in their evolution. For example, they support major or moderate influences

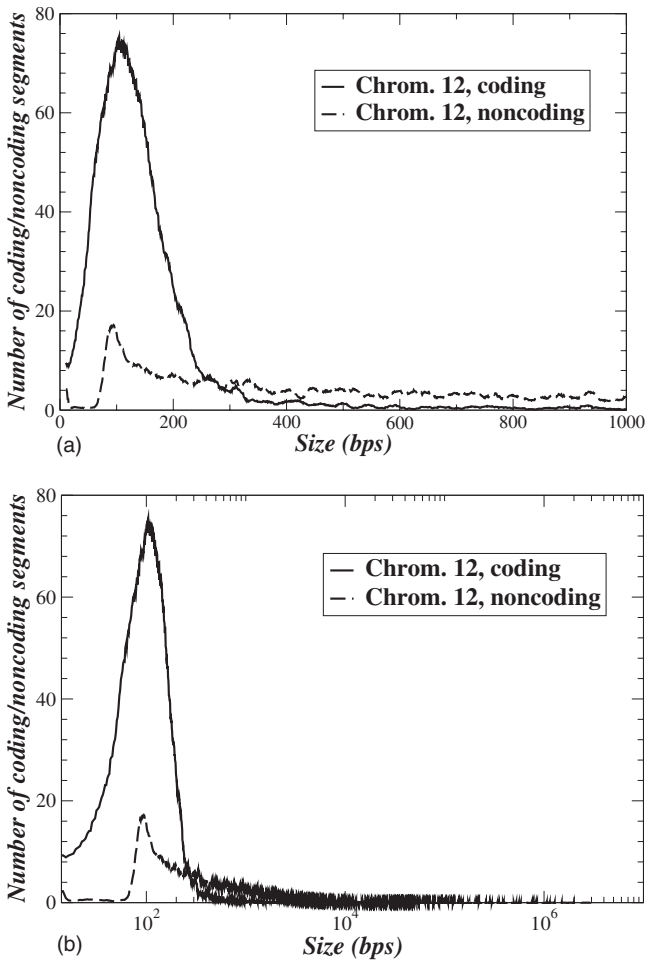

FIG. 1. (a) Number of coding and noncoding segments for chromosome 12. (b) Same data as in (a) in logarithmic scale to demonstrate the size difference in the coding and noncoding sequences. In all data running averages are considered in 20 bins.

from the environment (e.g., influx or outflux of genomic material, mutations, duplications, etc.); thus they can be both classified as out-of-equilibrium systems. Alternatively, the coding segments (exons) rarely support external influences. Any modification to the structure of a coding segment will most probably be fatal for the cell (interruption of the production of some protein), and thus the coding DNA behaves as a closed system at equilibrium and does not support environmental influences in the primary structure.

After separating the coding and noncoding DNA sequences in each chromosome, their respective SD of coding and noncoding segments were studied as a function of the sequence size S, which is taken equal to the number of bps in the sequence. For clarity and to better distinguish the coding from the noncoding data, the number of coding and noncoding segments instead of the segment percentage is depicted in the following five figures. Representative SD plots are shown in Fig. 1. It is noted, confirming also earlier reports, that both plots present a maximum around 100 bps and that the coding sequences rarely reach sizes of the order of 10^4 bps, while the noncoding ones easily reach easily sizes of 10^6 bps [15,16]. From the same figure it is also evident that the tails of the coding fall much faster than the noncoding ones. Similar behavior is demonstrated in all other human chromosomes.

To avoid statistical fluctuations the cumulative size distributions are studied which are defined as

$$Q_c(S) = \int_S^\infty P_c(r)dr, \quad Q_{nc}(S) = \int_S^\infty P_{nc}(r)dr \tag{1}$$

for coding and noncoding sequences, respectively. Equations (1) represent a cumulative distribution from size S to ∞, whereas the terms in statistical physics often refer to that of sizes from 0 to S. The current definition is preferable for long-range distributions, because the power law in the tail characteristics of the original distribution is conserved in the cumulative. Namely, if a distribution $P(r)$ has a power law tail with exponent $-1-\mu$, the cumulative distribution $Q(S)$ acquires a power law tail with exponent $-\mu$, while if $P(r)$ follows an exponential decay (more generally a short-range tail), the cumulative distribution will also follow an exponential decay (more generally short-range decay) [10].

In Fig. 2 the cumulative number of coding and noncoding segments of the first 22 human chromosomes are depicted in a double-logarithmic scale (coding and noncoding SDs are depicted in the same plots). Since the

$$\text{(number of coding segments)} = \text{(number of noncoding segments)} \pm 1, \tag{2}$$

both SDs start from the same value (± 1 unit) at $S=1$. The coding SDs drop abruptly (shorter tails) while the noncoding ones (longer tails) fall more smoothly. In all four plots, for comparison, the dotted and dashed lines represent pure power laws with exponent $\mu=2$ and $\mu=1$, respectively. It is striking that the noncoding SDs of the 22 chromosomes present a clear power law region for over two orders of magnitude in the 10^4-10^6 bp scales, with exponent μ_{nc} mostly less than 1. Comparing the coding SDs with the dotted line, it is obvious that they decay following a power law with exponent μ_c greater or equal to 2. Power law functions with large exponents ($\mu>2$) cannot be distinguished from exponential (short-range) functions when the sample size and the x-axis extension are limited. Accordingly, short-range, exponential decay behavior is manifested by the coding (exon) size distributions in human DNA.

For completeness, in Fig. 3 we depict the cumulative number of coding and noncoding segments of chromosomes X and Y separately. We note a significant deviation in the power law exponent $\mu_{nc,Y}$ compared to other chromosomes, which is attributed to its low sequencing percentage, only 43% (see Table I).

To capture the long- and short-range characteristics in the SDs of coding and noncoding sequences we treat them in a unified way by fitting them to the same general formula containing both exponential (short-range) terms and power law (long-range) terms:

$$Q(S) \sim BS^{-\mu} \exp[-\alpha S] \quad \text{for } S \gg 1, \tag{3}$$

where B is a normalization factor and α is the exponential decay exponent. The results are shown in Table I.

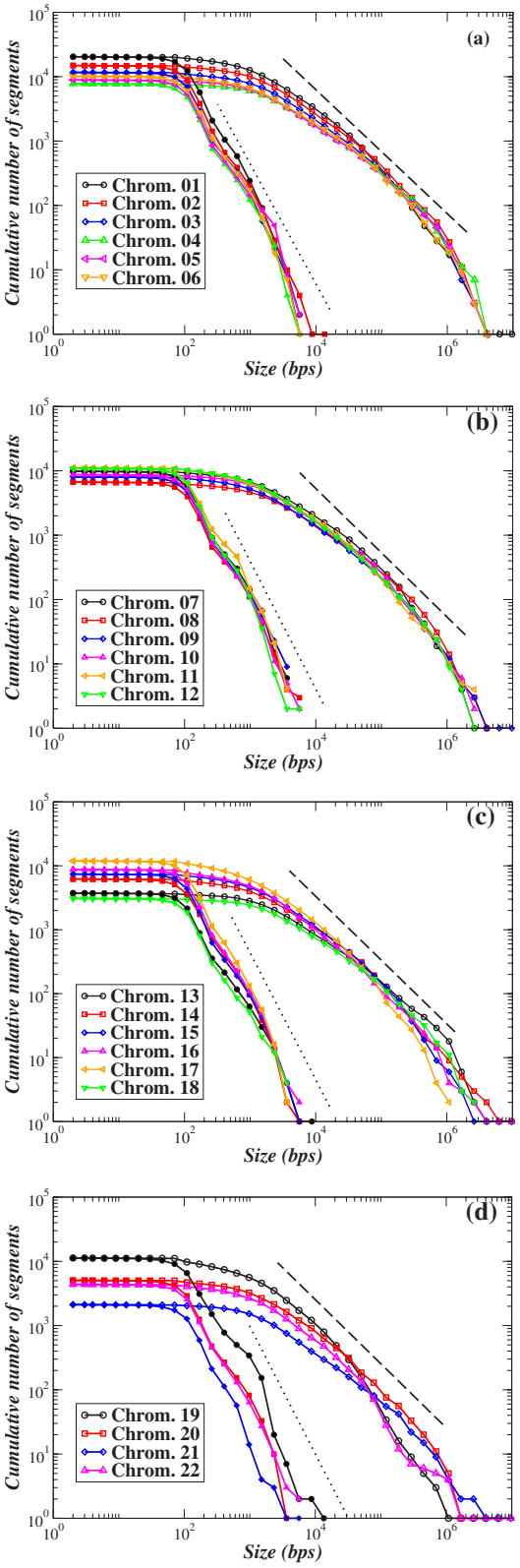

FIG. 2. (Color online) Cumulative coding and noncoding SDs in 22 human chromosomes. The solid symbols correspond to coding segments while the open symbols correspond to noncoding ones. In all graphs the dashed (dotted) lines represent pure power laws with slope −1 (−2). Coding sequences show power law behavior with exponent $\mu_c \geqslant 2$, while noncoding ones show power law behavior with $\mu_{nc} \leqslant 1$.

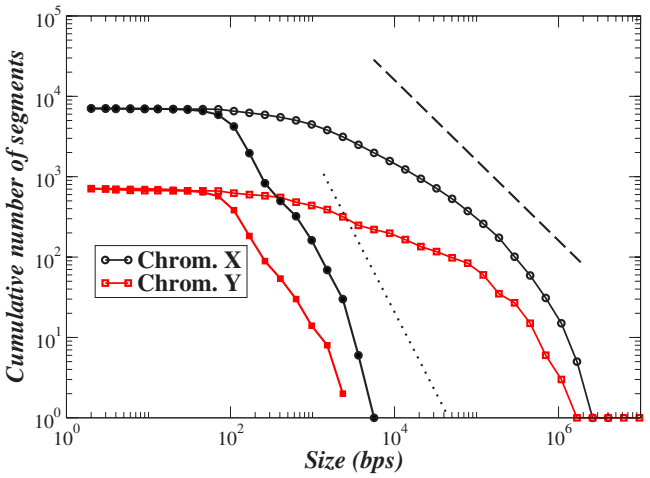

FIG. 3. (Color online) Cumulative coding and noncoding SDs of chromosomes X (circles) and Y (squares). The solid symbols correspond to coding segments while the open symbols correspond to noncoding ones. The dashed (dotted) lines represent pure power laws with slope −1 (−2).

As we can see in Table I, in the noncoding SDs the exponential decay exponent α_{nc} is of the order of 10^{-7}, while the main contribution comes from the power law term $S^{-\mu_{nc}}$, with μ_{nc} mostly around the value 0.85. The values of μ_{nc} have error bars of the order of $\Delta\mu_{nc} \sim \pm(0.05-0.10)$ depending on the sequence while $\Delta\alpha_{nc} \sim \pm 0.5 \times 10^{-7}$. The highest power law exponent is currently presented by chromosome 19, $\mu_{nc,19}=1.2\pm0.08$, with sequenced percent 87.43% and coding percent 3.42% (highest). The lowest power law exponents are presented by chromosomes X and Y, $\mu_{nc,X}=0.73\pm0.05$, $\mu_{nc,Y}=0.40\pm0.10$, and $\mu_{nc,13}=0.71\pm0.05$, with sequenced percents $\text{seq}_X=97.14\%$, $\text{seq}_Y=43.10\%$, and $\text{seq}_{13}=83.73\%$ and coding percent $\text{cod}_X=0.87\%$, $\text{cod}_Y=0.22\%$, and $\text{cod}_{13}=0.59\%$, respectively. Minor changes are expected in the exponent values (especially for chromosome Y) after the full human genome annotation.

From the same table, we note that the functional form of coding SDs includes an exponential term with exponent α_c of the order of 10^{-3}—i.e., three orders of magnitude larger than in the noncoding sequences. Thus the exponential part in the coding sequences is non-negligible. The power law exponent is of the order of $\mu_c \sim 2$ which is borderline between short- and long-range behavior. Error bars on μ_c are of the order of $\Delta\mu_c \sim \pm(0.05-0.10)$ depending on the sequence, while $\Delta\alpha_c \sim \pm 0.05 \times 10^{-3}$. Taking into account the exponential term which dominates in the large scales, we conclude that the coding SDs follow short-range laws.

In the analysis performed so far, both introns and intergenic regions are treated collectively as noncoding regions. As was discussed earlier, the introns are located within the gene regions and they are known to include some functional units, such as promoters and regulatory elements. These segments can suffer moderate modifications, but not extensive ones. On the other hand, such restrictions do not hold on the evolution of intergenic regions which can absorb extensive changes without damaging the cell functions. Thus the intron size distribution is expected to present some characteristics

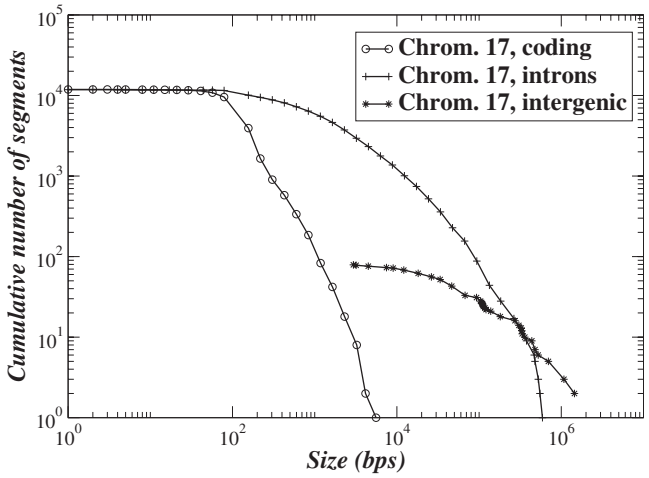

FIG. 4. Cumulative coding (circles) and intron (crosses) and intergenic region (stars) SDs of chromosome 17.

similar to the coding SDs as has been confirmed by other authors [15–18]. One of these characteristics is the maximum on the intron size distribution, which is situated around the value of 100 bps as in the case of coding. However, since modifications do take place in introns, long-range tails are also expected in their size distribution.

To clarify the difference between intron size distributions and intergenic regions, the GenBank data were analyzed as follows: from the gene regions (CDS) the positions of introns and exons were obtained separately and their size distributions were found. The segments which separate the different gene regions were considered as intergenic regions, and their size distributions were calculated. Representative results are shown in Fig. 4 for chromosome 17. Here, the three cumulative distributions of exons (coding), introns, and intergenic regions are shown for comparison. It is clearly observed that the size distribution of the coding (open circles) drops abruptly and the tails demonstrate exponential decay, while the intron size distribution demonstrates longer tails with exponent $\mu \sim 1.2$ which clearly indicates a power law tail. Thus the pure intron size distributions present dual characteristics: short range (such as the maximum in the 100 bps area) complimented by longer tails. As for the intergenic size distributions the tail is clearly longer and the approximate exponent is calculated as $\mu \sim 0.7$. When the two distributions of introns and intergenic regions are considered together, the intergenic distribution drags the intron one toward the right, producing one single composite exponent $\mu \sim 0.9$, as was reported in Table I. This separation in introns and intergenic regions is only approximate, since the annotation of the human genome is not complete and all the genes have not been identified, yet the general trends are clearly manifested. The same general behavior is observed in all other human chromosomes.

III. MODELING THE SHORT- AND LONG-RANGE FEATURES OF DNA

The emergence of long-range correlations in DNA has been previously studied using minimal models [1–3,8,11]. Most minimal models refer to long-range correlations related to DNA nucleotide chains converted to binary sequences. In the sequel we explore the emergence of long- and short-range correlations in the alternating noncoding and coding segments and we introduce an appropriate minimal, stochastic model of genomic evolution which takes into account elementary evolutionary processes such as mixing of genomic material (through a cut-and-paste mechanism), insertion of random segments, outflux of segments and duplication.

The model is inspired by the Takayasu aggregation model [11,19] and can be described as follows.

(i) Consider a linear chain which initially consists of alternating coding and noncoding strings. The initial size of the strings is not important (especially in the noncoding) as it will be soon forgotten due to aggregation and mixing events. In any case all segments may be chosen to have random initial size.

(ii) At every elementary time step a segment of random size (random between an upper and a lower size) is added on a randomly chosen noncoding string on the line. Thus the length of chosen noncoding string increases. This is an insertion event.

(iii) A mixing, or cut-and-paste, event consists of reducing the size of a given random noncoding string to a minimum size and adding its content to another randomly chosen noncoding string.

(iv) A duplication event consists of doubling the size of a randomly chosen noncoding string.

(v) An outflux event consists of eliminating the contents of a randomly chosen noncoding string.

(vi) In coding strings only one insertion is allowed at the beginning and very seldom additional insertions are supported. (Biological evidence indicates that modifications are rarely supported by "working" coding segments.)

(vii) After having taken into account the above events with appropriate probability factors (relative frequencies) the algorithm returns to step (ii) for a new elementary time step to start.

Note that insertion of random segments, cut-and-paste events, and outflux events are allowed only for noncoding strings.

Previous analytical mean-field calculations have shown that in the absence of the outflux mechanism the noncoding size distributions follow a pure power law with exponent $\mu = 0.5$ [19]. In the case where a random outflux of segments is considered, the power law takes the exact exponent $\mu = 1$ only when the random outcoming and incoming segments have equal sizes on the average. The pure power laws are modified if we further assume that the outflux mechanism is weaker in comparison to the influx. This is biologically acceptable, because the current view of biology admits that the genome of evolutionary newer organisms increases in size, in spite of the occasional, spontaneous removal of genomic material.

Using the above algorithm linear chains of sizes up to 3×10^8 bps have been created. In Fig. 5 representative graphs of cumulative noncoding size distributions are presented. In the simulation the inserted segments have random sizes between 1 and 50 bps and the relative frequency of (cut-and-

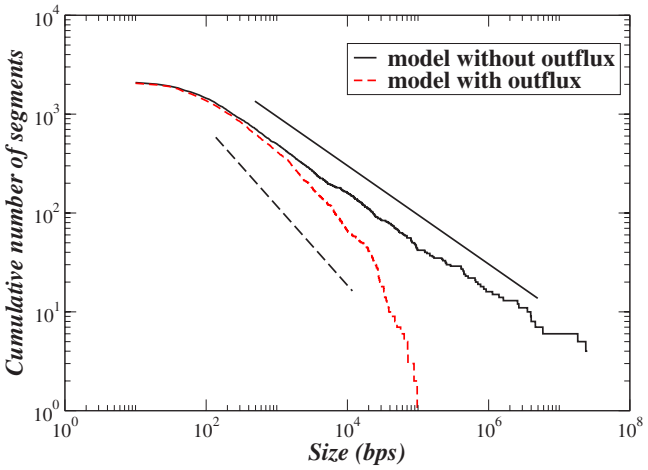

FIG. 5. (Color online) The mean-field model results for the noncoding size distributions. The dashed line corresponds to results with outflux of genomic segments while the solid line indicates results with both influx and outflux mechanisms. The straight lines are guides to the eye.

past)/(insertion) events is 0.1. The results without outflux are presented with the black solid line and compare well with the straight solid line which represents a power law with exponent $\mu=0.5$. If a weak outflux process is added (outflux/insertion events is 0.01), then the power law exponent is modified and a region with exponent $\mu \sim 0.8$ emerges, which is consistent with the exponents shown in Table I, for the noncoding. The same algorithm gives a short-range distribution of the number of coding segments, in accordance also with the human genome data.

IV. CONCLUSIONS

In summary, we have analyzed all human chromosomes and have demonstrated that the noncoding size distributions present power law behavior and are consistent with open, out-of-equilibrium systems, while the coding size distributions present short-range behavior as in closed systems, at equilibrium. We also explore the intron and intergenic size distributions. The latter present strong long-range tails while the former demonstrate intermediate behavior having some short-range features as the coding (exon) distributions with additional long-range tails as found in intergenic regions. A minimal, stochastic, mean-field model is proposed, which takes into account elementary, evolutionary processes (insertion of random segments, duplication, outflux, and mixing events) and produces the observed nontrivial statistical characteristics.

From preliminary studies, we have seen that these long-range features are also observed in other higher eucaryotic genomes. With the rapid growth of genomic data and with the completion of chromosomes from lower animalia and plants the persistence of these characteristics can soon be tested across organisms.

ACKNOWLEDGMENTS

The authors would like to thank Dr. G. A. Tsekouras, Dr. P. Katsaloulis, Professor E. Kanavakis, Professor S. Kouloheri, and Professor K. Trougkos for helpful discussions.

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); W. Li, *ibid.* **10**, 395 (1989).
[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] 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).
[5] H. Herzel, E. N. Trifonov, O. Weiss, and I. Grosse, Physica A **249**, 449 (1998).
[6] R. Roman-Roldan, P. Bernaola-Galvan, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
[7] Z. Ouyang, C. Wang, and Z.-S. She, Phys. Rev. Lett. **93**, 078103 (2004).
[8] P. W. Messer, P. F. Arndt, and M. Lassig, Phys. Rev. Lett. **94**, 138103 (2005); P. W. Messer, M. Lassig, and P. F. Arndt, J. Stat. Mech.: Theory Exp. (2005) 10004.
[9] B. L. Hao, Physica A **282**, 225 (2000); B. L. Hao, H. C. Lee, and S. Y. Zhang, Chaos, Solitons Fractals **11**, 825 (2000).
[10] Y. Almirantis and A. Provata, J. Stat. Phys. **97**, 233 (1999); A. Provata and Y. Almirantis, Fractals **8**, 15 (2000).
[11] A. Provata and Y. Almirantis, J. Stat. Phys. **106**, 23 (2002).
[12] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo, Y. d' Aubenton-Carafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Thermes, Physica D **96**, 291 (1996).
[13] P. Katsaloulis, T. Theoharis, and A. Provata, Physica A **316**, 380 (2002); P. Katsaloulis, T. Theoharis, W. M. Zheng, B. L. Hao, T. Bountis, Y. Almirantis, and A. Provata, *ibid.* **366**, 308 (2006).
[14] J. D. Hawkins, Nucleic Acids Res. **16**, 9893 (1988).
[15] M. Deutsch and M. Long, Nucleic Acids Res. **27**, 3219 (1999).
[16] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[17] L. P. Lim and C. B. Burge, Proc. Natl. Acad. Sci. U.S.A. **98**, 11193 (2001).
[18] M. Sakharkar, F. Passetti, J. E. de Souza, M. Long, and J. S. de Souza, Nucleic Acids Res. **30**, 191 (2002); M. K. Sakharkar, V. T. K. Chow, and P. Kangueane, In Silico Biology **4**, 387 (2004); M. K. Sakharkar, B. S. Perumal, K. R. Sakharkar, and P. Kangueane, *ibid.* **5**, 347 (2005).
[19] H. Takayasu, M. Takayasu, A. Provata, and G. Huber, J. Stat. Phys. **65**, 725 (1991); H. Takayasu, Phys. Rev. Lett. **63**, 2563 (1989).