Journal of Statistical Physics, Vol. 97, Nos. 1/2, 1999

Long- and Short-Range Correlations in Genome Organization

Y. Almirantis[1] and A. Provata[2]

Received January 27, 1999; final June 2, 1999

We study the size distribution of coding and non-coding regions in DNA sequences. For most organisms we observe that the size distribution $P_c(S)$ of the coding regions of size S shows short range distribution, whereas the size distribution of the non-coding regions follows a power-law decay $P_{nc}(S) \sim S^{-1-\mu}$, with power exponents indicating clear long-range behavior. We argue, using the Generalized Central Limit Theorem, that the long-range distributions observed in the non-coding are related to the lower level clustering of purines and pyrimidines ($1d$ islands) which follow similar long-range laws. We also address the question of clustering of coding segments in the two complementary strands of DNA. We observe a short-range clustering of coding regions in both strands, expressed by an exponential decay in the clustering size distribution. The decay exponent expresses the degree of short-range correlations and the deviation from random clustering.

KEY WORDS: Power law distributions; long-range correlations; coding/non-coding DNA sequences; DNA strand partition.

1. INTRODUCTION

During the past few years, there has been an increasing number of works concerning the study of the intriguing statistical behaviour of DNA sequences, and especially, the occurrence of important statistical correlations,[1–6] which in several cases, have been related to the biological role of the examined DNA. Some examples of statistical properties of DNA sequences recently investigated are imperfect periodicities[7] and their contribution to the long range order,[8] and multiple scaling.[9]

[1] Institute of Biology, NRCPS Demokritos, 15310 Athens, Greece.

[2] Institute of Physical Chemistry, NRCPS Demokritos, 15310 Athens, Greece.

The first attempt to examine DNA sequences from the Statistical Physics point of view was due to Peng *et al.* In a 1992 publication[1] they map a DNA sequence consisting of Purines (Pu) (Adenine or Guanine) and Pyrimidines (Py) (Cytosine or Thymine) onto a 1-dimensional walk. The model is called "DNA walk." In the same publication they show that coding DNA sequences (the parts of the DNA sequences which code for protein synthesis) correspond to regular random walks, while non-coding sequences correspond to correlated walks showing superdiffusive behaviour. This finding, which has now been verified by several sources, comes in apparent contradiction with the constraints acting on the structure and functioning of coding DNA and with the evidence that there are less constraints acting on the non-coding, the evolutionary history and the role of which remain still not completely understood.[1, 3–13]

In an earlier publication we have examined the nature of these correlation by looking at the clustering of consecutive Pus and consecutive Pys.[3] We have observed that the size distribution of Pu and Py clusters in coding DNA follows an exponential decay, while the size distribution of Pu and Py clusters in non-coding DNA follows a power law decay. Recently, similar power law behaviour is observed in the repeats of identical dimers of nucleotides belonging to the non-coding regions of eucaryotic DNA.[5] This additional evidence of correlations motivates the search for the degree of universality of the statistical characteristics of genomes and for their (ultimate) origin.

In the current work, we examine the size distributions of coding and non-coding regions in several taxonomic groups. Using the clustering of Pus and Pys observed in ref. 3 we predict, using the Generalized Central Limit Theorem (GCLT), that the non-coding regions follow long-ranged power law length distributions, whereas the coding regions follow short ranged distributions. These theoretical predictions are tested successfully against experimental evidence from DNA sequences of variable origin.

Another problem tackled in this work is the search for the statistical properties of the partition of coding segments between the two complementary strands of the DNA molecule. These two poly-nucleotide chains, called usually "Watson" and "Crick," or "direct" and "complementary," are in general statistically equivalent. We have studied the "DNA stand partition" examining the clustering in the two strands of the length distribution of 1-dimensional islands, i.e. uninterrupted coding or non-coding segment juxtapositions. Only whole genomes of procaryotic organisms have been used, where almost all the DNA is coding. From this study follows that the DNA strand partition in most cases behaves in a quite uniform way, expressed by an exponential length distribution.

The outline of the work is as follows: in the next section we state the GCLT and apply it to coding and non-coding sequences. In Section 3, we examine large DNA sequences obtained from several taxonomic groups, we elaborate on the similarities and differences of their statistical properties and we test our theoretical predictions. In Section 4 we study the strand partitioning for several procaryotic complete genomes. In the final section we recapitulate the main conclusions of this work and we discuss new perspectives.

2. PROBABILISTIC DESCRIPTION OF DNA SEQUENCES

A DNA sequence, seen as a biological text written by means of a four letter alphabet, has interesting statistical properties as manifested by many recent reports and as discussed in the Introduction. In a coarse grained way, we can see a DNA strand as a sequence of coding and non-coding regions each of which consists of a juxtaposition of Pus and Pys.

In the lower level of organization, the Pu/Py level, several authors have confirmed the existence of long-range correlations between homologous nucleotides, especially in the non-coding parts of higher organisms. These correlations were first observed using the notion of "DNA walk"(1) and later were confirmed using the size distribution function of clusters of homologous nucleotides.(3, 5) It has also been reported that several types of scale dependent non-randomness, related to the observed long-range order, is present in non-coding sequences.(12, 13)

In a higher level of organization, the level of coding/non-coding, long range distributions of the non-coding regions for higher organisms were recently reported.(4) At the same level, the coding size distribution of all organisms shows only short range order.

In the current section, we show that there is an intrinsic connection between the two levels of organization and the key factor is the Central Limit Theorem (CLT), which explains how statistical characteristics are transferred between the different levels of coarse graining.

The original form of the CLT states that if we have independent, random variables $X_1, X_2,..., X_n$ following the same distribution $\mathfrak{R}$, which has finite mean and variance (short range distributions), then the sum of a large number N of such variables will follow, the Gaussian distribution. Note that the Gaussian distribution is a short ranged distribution, as were the original distributions $\mathfrak{R}$. Consider now the composition of coding DNA: In ref. 3 the coding DNA is shown to be a collection of 1-dimensional clusters of Pus and Pys. All these Pu and Py clusters follow the same exponential distribution

$$P_c(s) \sim e^{-s\,|\ln p|} \tag{1}$$

where s is the cluster size and p is the probability to find a single Pu or Py on the DNA chain. Statistically, the frequencies of Pu and Py are approximately equal on a DNA chain, i.e., $p = 1/2$. Since the exponential distribution has finite mean and variance we expect that the probability distribution of the coding regions of DNA will be a Gaussian in the large size limit. Notice that for large values of the distributed variable the Gaussian may be well approximated by an exponential distribution. Both Gaussian and exponential distribution are short ranged[14] and thus for large values of the distributed variable they fall equally fast.

On the other hand, the Generalized Central Limit Theorem concerns distributions $\mathfrak{R}$ with infinite moments. For independent, random variables $S_1, S_2, \ldots, S_n$ following the same distribution $\mathfrak{R}$, the distribution $\mathfrak{R}$ is called *stable* if the sum of the variables $S_n = S_1 + S_2 + \cdots + S_n$ follows the same distribution $\mathfrak{R}$ (with some appropriate linear transformations).[14–16] Stable distributions with infinite moments have long power law tails of the form

$$P(s) \sim s^{-1-\mu}, \qquad \text{for} \quad 0 \leqslant \mu \leqslant 2, \quad s \gg 1 \tag{2}$$

The size distributions of Pu and Py clusters observed in higher eucaryotes follow power laws of the form (2). Going to the next level of organization, the non-coding regions are composed of many Pu and Py clusters interchanged. According to the GCLT, we expect that the size distribution of the non-coding parts P_{nc} will also follow the same power law distribution

$$P_{nc}(S) \sim S^{-1-\mu}, \qquad \text{with} \quad 0 \leqslant \mu \leqslant 2, \quad S \gg 1 \tag{3}$$

where S is the size of the non-coding regions and it is regarded as a large collection of Pu and Py clusters, each of which has size s_i, thus $S = s_1 + s_2 + \cdots + s_n$, with $n \to \infty$.

In the previous paragraph we have silently assumed that all the variables S_i follow the same cluster size distributions with the same exponent μ. This is certainly not the case as may be seen in previous references.[2, 3] In most cases the exponents μ observed for the Pu and Py distribution of the same organism are slightly different. When macromolecules with different power law distributions randomly join together to form larger and larger macromolecules, the exponent which will dominate and will ultimately characterize the tails of the resulting distribution is the one with the lowest value of μ (see Appendix). As we will see in the next section, the exponents obtained for the size distribution of the different non-coding sequences are always smaller or equal to the exponents obtained from the size distribution of the Pu and Py clusters of the same organism. This outcome is a direct conclusion of the calculations of the Appendix. The distribution resulting from the aggregation has chosen the smallest exponent of all the distributions which contributed in the process.

For the analysis of the sequences we use the cumulative size distribution function $\tilde{P}(s)$ defined as[14]:

$$\tilde{P}(s) = \int_s^{\infty} P(r)\, dr \tag{4}$$

where $P(r)$ is the original distribution of coding or non-coding regions of size r. In general the cumulative distributions have better statistical properties than the original distribution functions. Notice that if the distribution $P(r)$ has the exponential form its cumulative $\tilde{P}(s)$ will also have the exponential form. If the distribution function has a power law form of the type Eq. (2) then the cumulative distribution will have a power law form with exponent $-\mu$

$$\tilde{P}(s) = \int_s^{\infty} l^{-1-\mu}\, dl = s^{-\mu}, \qquad 0 \leqslant \mu \leqslant 2 \tag{5}$$

In the following section we apply these ideas to real DNA sequences. We pay particular attention to the study of the non-coding region size distribution and the occurrence of long range correlations in long DNA sequences of higher organisms and whole genomes of simple organisms.

3. ANALYSIS OF DATA

The sequences analyzed in this section were obtained from the Molecular Data Banks and from individual institutions working on genome projects. For information on the origin of each analyzed sequence see Table I. The analyzed sequence presented here were selected under the following basic criteria:

1. Complete genomes or entire chromosomes of organisms when possible.

2. If complete genomes are not available, lengthy sequences are selected containing many coding and non-coding regions. Longer sequences give in general better statistical results.

3. From all the sequences meeting the above two criteria only the fully annotated ones are useful for this analysis since it requires exact knowledge of all coding regions in a sequence.

4. Special cases with particular statistical interest are also shown.

Table I. Quantitative Features for Sequences Analyzed in Section 3

Fig. no.	Organism	Sequence origin	Length (base pairs)	Coding percent	Coding segments	Exp $-\mu$
1a	Human	HUAC004384/GenBank	213541	1.4%	29	-0.9
1b	≫	HSAF001550/GenBank	173882	2.1%	31	-0.6
1c	≫	HUMCOL7A1X/GenBank[a]	36631	24%	117	-1.4
2a	*Drosophila*	DMU31961/GenBank	338234	4%	42	-0.5
2b	≫	DMC22E5/GenBank	45672	36%	33	-0.4
2c	≫	DMC80H7/GenBank	45861	19%	30	-0.7
3a	*C. elegans*	CEY17G7B/EMBL	143092	19%	121	-1.3
3b	≫	CEY41E3/EMBL	150641	14%	78	-1.0
3c	≫	CEY57G11C/EMBL	313573	14%	133	-1.3
4a	*A. thaliana*	ATAC002387/GenBank	122928	28%	125	-0.8
4b	≫	ATAC002335/GenBank	102057	36%	145	-1
4c	≫	ATAC001645/GenBank	91714	30%	123	-0.8 & -1
5a	*S. cerevisiae*	Chromosome I, left arm/EMBL	103678	66%	47	-0.8
5b	≫	Chromosome IV (partial)/EMBL	552013	75%	288	-1.8
5c	≫	Chromosome XIII (complete)/EMBL	924430	75%	501	-1.3
6a	*M. janaschii*	Complete genome/TIGR	1664977	88%	1723	-1.3
6b	*H. influenza*	Complete genome/TIGR	1830135	87%	1792	-1.4
6c	*E. coli*	Complete genome/TIGR	4639221	89%	4400	—
7a	Nuclear Polyhedrosis Virus	Complete genome/GenBank	133894	93%	154	-1.1
7b	Bacteriophage sk1	Complete genome/GenBank	28451	93%	54	-1.0
7c	Adenovirus type2	Complete genome/GenBank	35937	83%	25	—

[a] Sequence obtained around a given gene (otherwise genome project products).

In this work all large, annotated, publicly available sequences are practically considered. For viruses and procaryotic organisms the entire sequences were available in many cases and they have been used here. For higher organisms entire genomes are not available at this time and we only consider annotated sequences containing a large number of coding and non-coding regions to obtain better statistics. In the latter case, sequences meeting these requirements are: (a) extended gene clusters, including intergenic non-coding regions and introns, or (b) the primary annotated products (cosmids or other types of genomic clones) of genome projects. Such projects are in progress for several test organisms such as: human, *Drosophila*, *C. elegans*, *S. cerevisiae*, etc.

We note here that in bacteria and viruses DNA overlapping genes (overlapping coding regions) are often met. In the current treatment we count a continuous coding region as one coding entity. However, if we choose an alternative treatment, counting each individual coding region as a coding entity (counting the overlaps more than once) no qualitative difference in our statistical characterization of the size distributions is observed.

We summarize the quantitative aspect of our results on the segment length distribution in Table I and in the corresponding Figs. 1–7. In Table I, column 1 contains the corresponding figure number, column 2 contains the organism name, column 3 gives the sequence coordinates and the DataBank where it was obtained, column 4 gives the sequence length in base pairs, column 5 contains the percentage of coding length in the sequence, column 6 gives the number of coding segments included and column 7 gives the exponent $-\mu$ corresponding to the non-coding segment size distribution. Most sequences are *randomly* cut parts of chromosomes, products of the genome sequencing project of the corresponding organism. One exception is shown in Fig. 1c, where the selected sequence is centered around a given gene. When complete genomes/chromosomes are used it is indicated in the second column of Table I. In Figs. 1–7 we present the size distribution of coding and non-coding segments obtained from the DNA sequences as described in Table I. All diagrams are in a double logarithmic scale and the straight lines (when they exist) indicate the power law behaviour corresponding to the non-coding segment length distribution. In these diagrams Circles (○) correspond to the cumulative distribution of coding regions while Squares (□) correspond to the cumulative distribution of non-coding regions.

3.1. Higher Eucaryotes

Coding and Non-coding sequences belonging to higher eucaryotes, such as vertebrates, insects and plants are first presented. These organisms

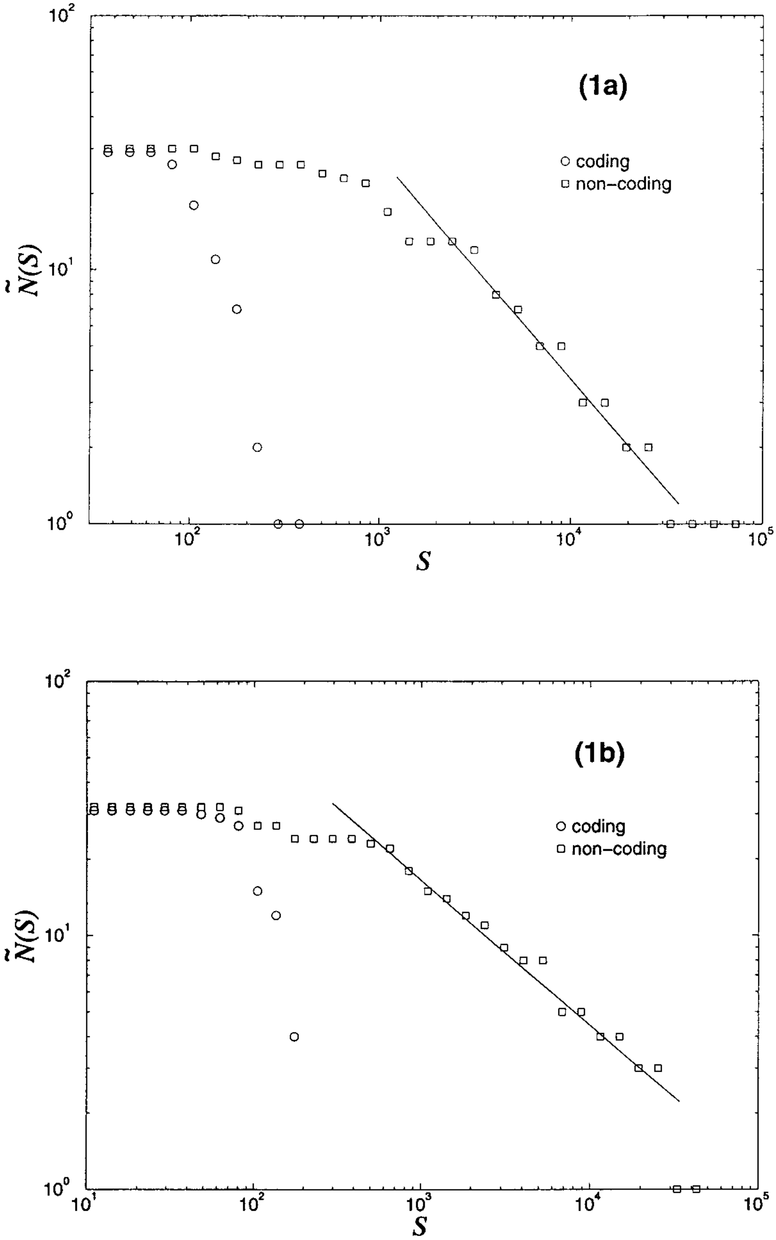


Fig. 1. The number of coding (circles) and non-coding (squares) regions of size $\geqslant S$, $\tilde{N}(S)$, for three human DNA sequences (for details see Table I). The straight lines have the following slopes: (1a) $-\mu = -0.9$, (1b) $-\mu = -0.6$ and (1c) $-\mu = -1.36$.

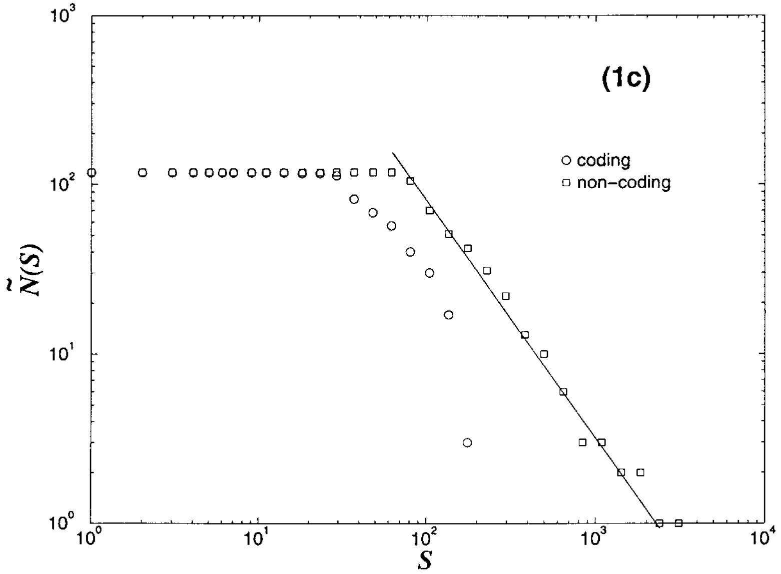


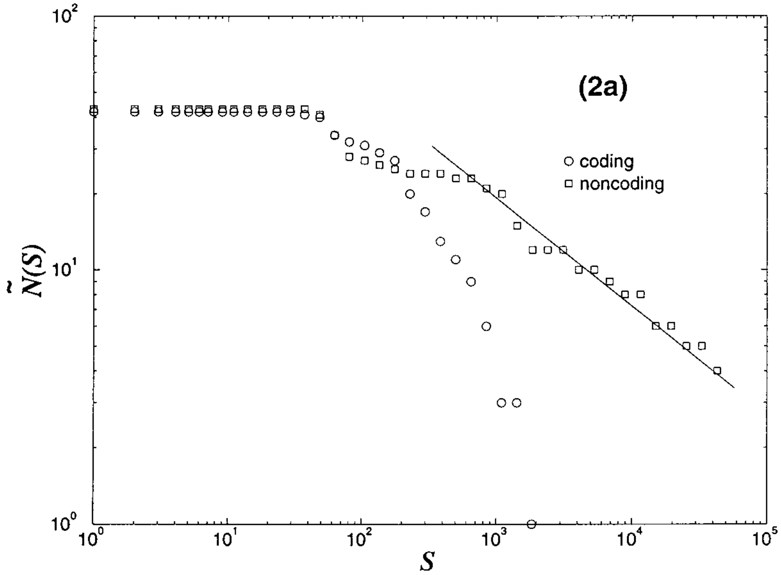
Fig. 1. (*Continued*)

have very large genome sequences, of the order of 3×10^9 base pairs for human, and the complete genome sequences are not completely decoded as yet. However, we were able to find some fully annotated sequences which were large enough for statistical credibility.

In Figs. 1a–c we present data from human sequences, see Table I. In all three sequences we observe a well defined power law decay in the non-coding with the corresponding exponents for the cumulative size distributions taking the values $-\mu = -0.9$, -0.6, -1.36. The corresponding values of the exponent $-\mu - 1 = -1.9$, -1.6, -2.36, which characterize the original non-coding segment size distributions, show clearly the long range nature of the correlations. In particular, sequence 1*c* is obtained in a targeted way around a concrete gene. In contrast, the majority of other sequences are randomly cut pieces of chromosomes. This is the reason why in sequence 1*c* the coding percentage is high (23.5%), which contributes to the high value of $|\mu|$. However, even with this particularity the sequence conserves its long range nature in the non-coding.

In all three cases the coding part does not indicate long range correlations, but the data falls abruptly in a short range manner.

In Figs. 2a–c the cumulative size distribution of coding and non-coding regions for *Drosophila melanogaster*, serving typically as a model organism for insects, is shown. The behaviour looks very similar to the cases of



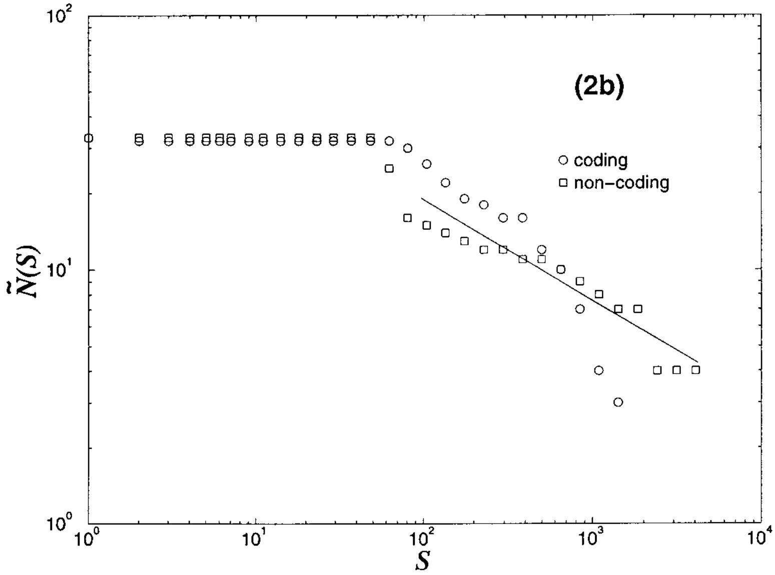


Fig. 2. The number of coding and non-coding regions of size $\geqslant S$, $\tilde{N}(S)$, for three insect DNA sequences (see Table I). The straight lines have the following slopes: (2a) $-\mu = -0.5$, (2b) $-\mu = -0.43$ and (2c) $-\mu = -0.7$.

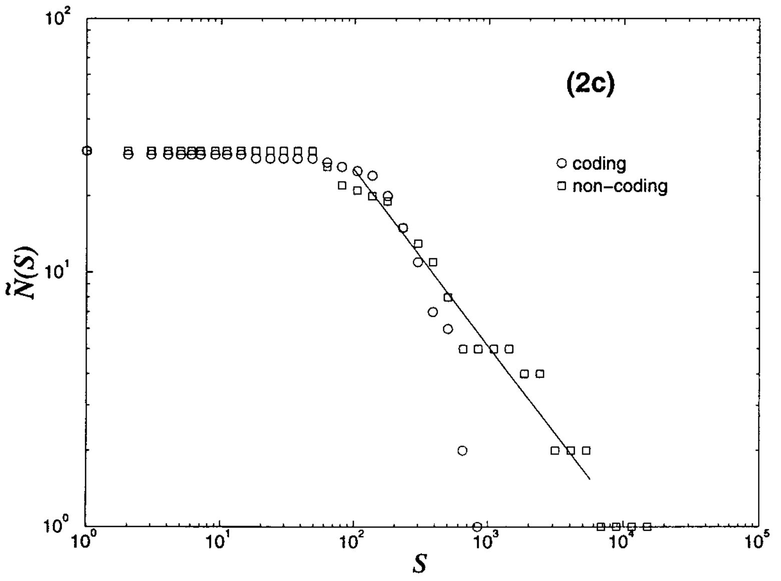
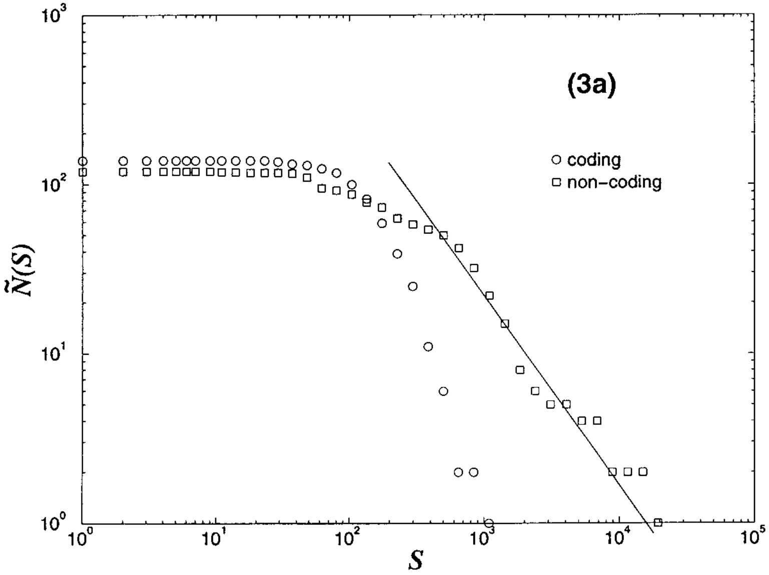


Fig. 2. (*Continued*)

Figs. 1, for the human DNA, see also Table I. This is not surprising since vertebrates, insects and all higher eucaryotes share the same general architectural characteristics in their genome organization. The same is true for an even simpler organism, the nematode *Caenorhabditis elegans*, which is extensively studied as a test organism. This simple organism has a genome rich in non-coding space and thus is ranked here with higher eucaryotes, see Figs. 3a–c.

In Figs. 4a–c the cumulative size distributions of coding and non-coding regions for *Arabidopsis thaliana* is shown, which serves as a test organism for plants. *A. thaliana* has the additional advantage of not having repetitive elements in the non-coding and thus the distributions are smoother than in the case of most vertebrates. In Fig. 4c, we have chosen to present a particular case which corresponds to a 91.7 kb segment of chromosome III, containing 123 exons. In this case we observe two regions of scaling in the non-coding. A clear power law decay with exponent $-\mu = -0.8$ is observed for non-coding sizes between 80 and 1000 base pairs, a transition takes place around size 1000 and a different power law prevails for sizes larger than 1000 with a slightly bigger exponent $-\mu = -1.0$. The existence of two different scales can be observed in other sequences of higher eucaryotes as well, but it rarely is as marked as it is here. The different scaling behaviour may be attributed to the different length scales of the non-coding spacers met:



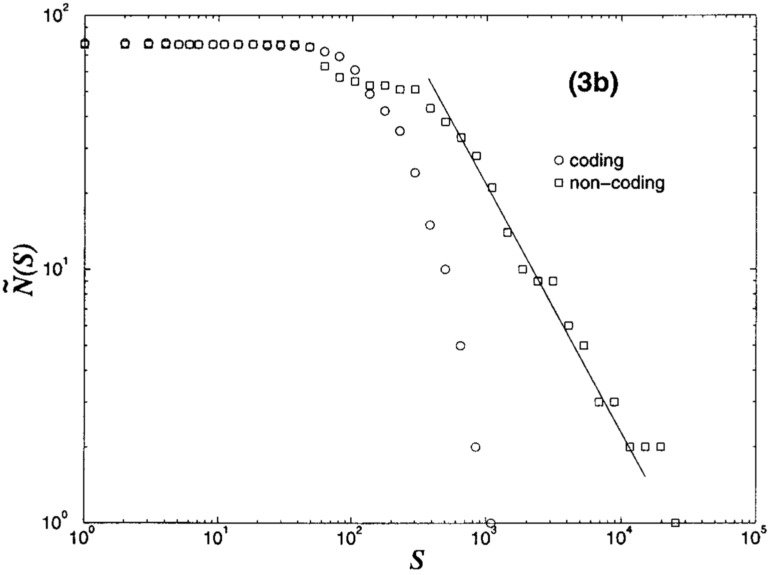


Fig. 3. The number of coding and non-coding regions of size $\geqslant S$, $\tilde{N}(S)$, for three nematode DNA sequences (see Table I). The straight lines have the following slopes: (3a) $-\mu = -1.3$, (3b) $-\mu = -1.0$ and (3c) $-\mu = -1.3$.

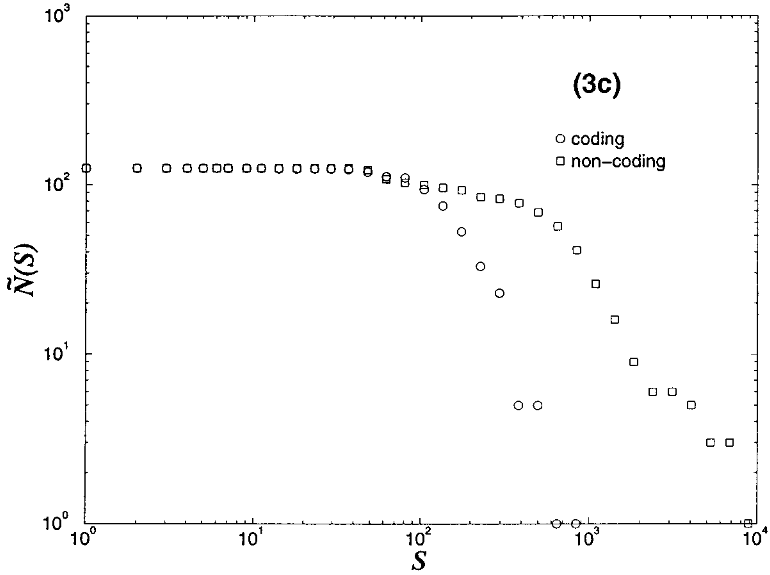


Fig. 3. (*Continued*)

(a) between coding segments (exons) of the same gene, and (b) between different aggregates of exons usually belonging to the same gene cluster.

We note here, that there is a hierarchy in the length scales of distances between coding regions in whole chromosomes of higher eucaryotes:

1. Small non-coding spacers separating exons belonging to the same protein.

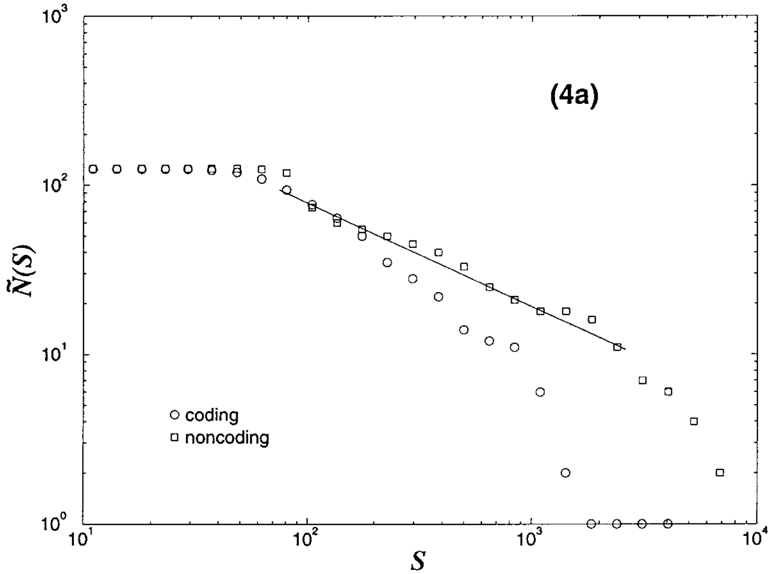
2. Intermediate size non-coding spacers separating genes belonging to the same cluster.

3. Large size non-coding spacers, intergenic regions, separating gene clusters.

This multitude of scales may result in multimodal distributions, as we have seen in Fig. 3c. It reflects the complexity of gene interaction/organization in these organisms and it may account for potential self-similarity features and multiscaling. Further examples may appear from the study of entire chromosomes of higher eucaryotes, when available.

3.2. Lower Eucaryotes

Lower eucaryotes are organisms with restricted genome such as fungi. One important difference in the DNA between lower and higher eucaryotes



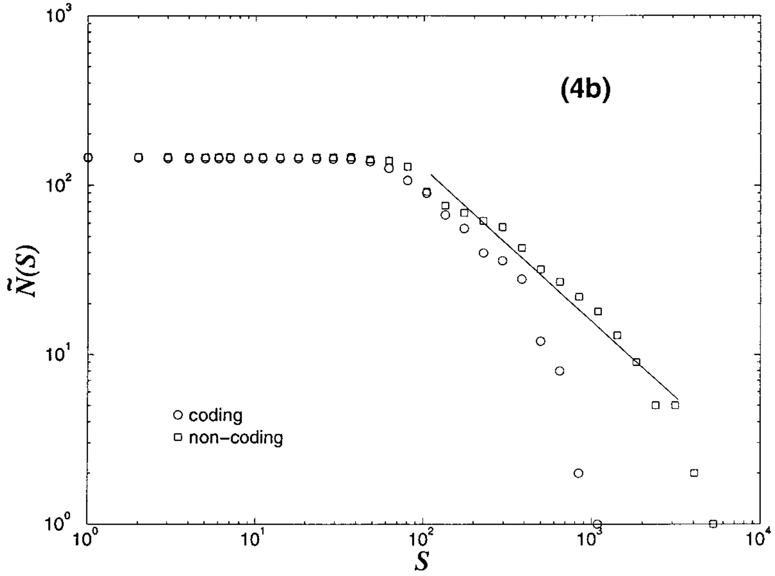


Fig. 4. The number of coding and non-coding regions of size $\geqslant S$, $\tilde{N}(S)$, for three plant DNA sequences. The straight lines have the following slopes: (4a) $-\mu = -0.8$ and (4b) $-\mu = -1.0$. In (4c) we notice two distinct regions with different slopes (see details in the text and in Table I).

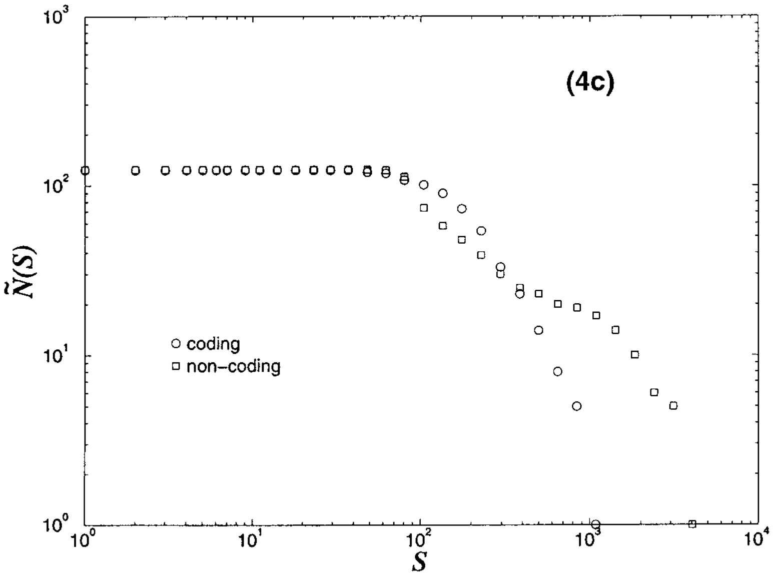


Fig. 4. (*Continued*)

is that the coding part of the former covers between 60–80% of the total length of the DNA, while in the latter the genome contains large non-coding spacers, as we have seen in he previous section and in Table I. The evolutionary reason for this difference is not known as yet. As an exception, Brenner *et al.*(17) have recently found that a typical higher eucaryote, the pufferfish *Fugu rubripes rubripes*, presents a particularly condensed genome. The origin of this particularity remains still unclear.

This far, only the genome of the *Saccharomyces cerevisiae*, serving as a test organism, has been decoded and partially annotated; we will study it hereafter. Eventhough its non-coding parts are reduced, one can easily see traces of long range correlations. In Fig. 5a we present the cumulative size distribution of coding (circles) and non-coding (squares) segments for the *S. cerevisiae*, chromosome I, left arm, complete sequence. Contrary to higher eucaryotes, we observe here that the curve corresponding to the non-coding is below the curve corresponding to the coding distribution. This is not surprising, since the coding is covering 66% of the total length of this sequence. In a double logarithmic scale we can see that the non-coding presents a clear, power law, cumulative size distribution with exponent $-\mu = -0.8$, while the coding shows only short range behaviour. We also present data from *S. cerevisiae*, chromosome XIII, complete sequence, and chromosome IV, partial sequence, in Figs. 5b, c.

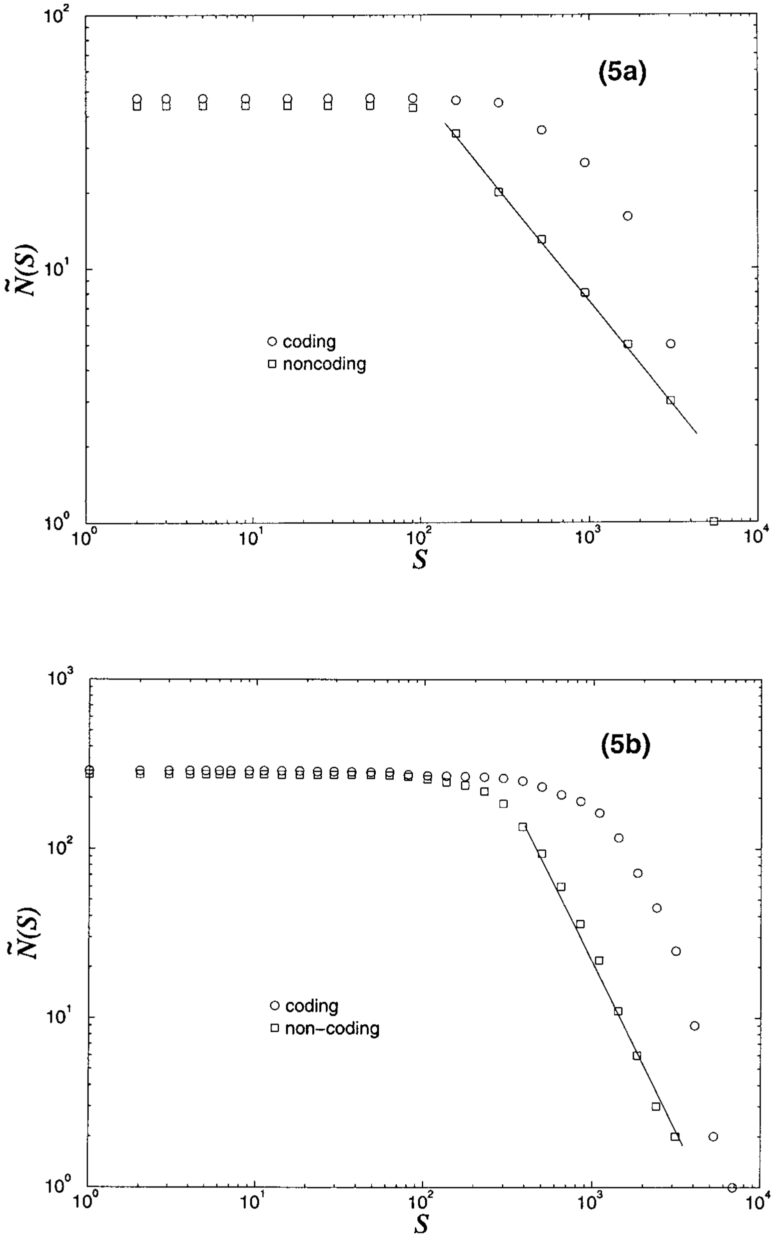


Fig. 5. The number of coding and non-coding regions of size $\geqslant S$, $\tilde{N}(S)$, for three fungal DNA sequences. The straight lines have the following slopes: (5a) $-\mu=-0.8$, (5b) $-\mu=-1.8$ and (5c) $-\mu=-1.3$.

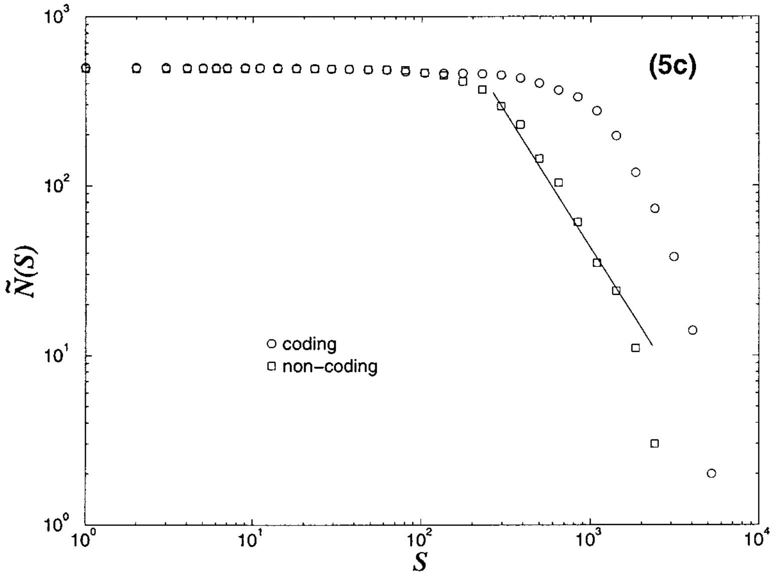
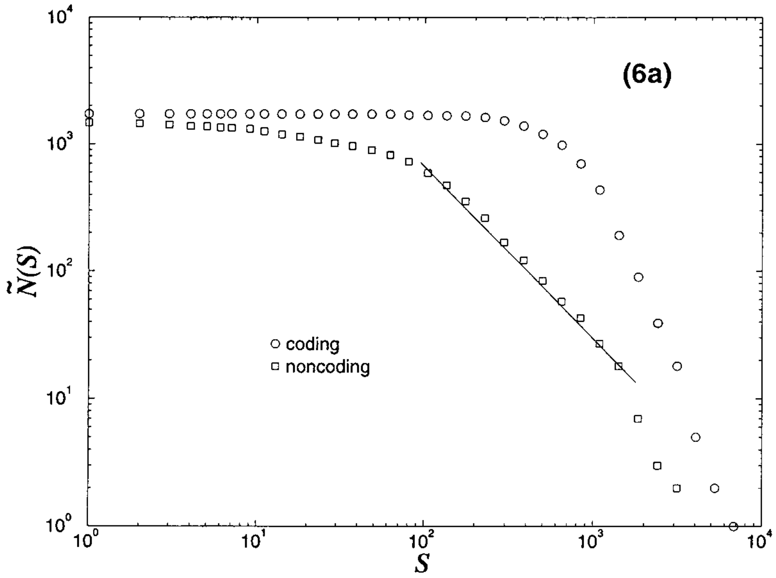


Fig. 5. (*Continued*)

The non-coding of theses chromosomes (squares) show also power law behaviour but the overall behaviour of the coding and non-coding are very similar. Such behaviour is met often in the cases where the non-coding is restricted.

3.3. Procaryotes

Procaryotes are simple unicellular organisms with reduced genome size ranging from 10^4 to 10^6 base pairs. The complete annotated genome of a few of these organisms is already available while several others are in the process of sequencing. We have found and examined eight fully annotated complete genomes of procaryotes available at this point. Given the reduced non-coding percentage of these organisms, the identification of long range correlations was doubtful. However, we have found in four cases well determined power law decay and in the other four borderline behaviour of the non-coding. In all eight cases the coding has shown clear short range coding size distributions. In Fig. 6 and Table I we present two cases of procaryotes with clearly identified long range distributed non-coding segments and one case of border line behaviour. More specifically, in Figs. 6a, b we present the organisms *Methannococcus janaschii* and *Haemophilus influenza* which have a power law decay with exponent values $-\mu = -1.3$ and -1.4 respectively.



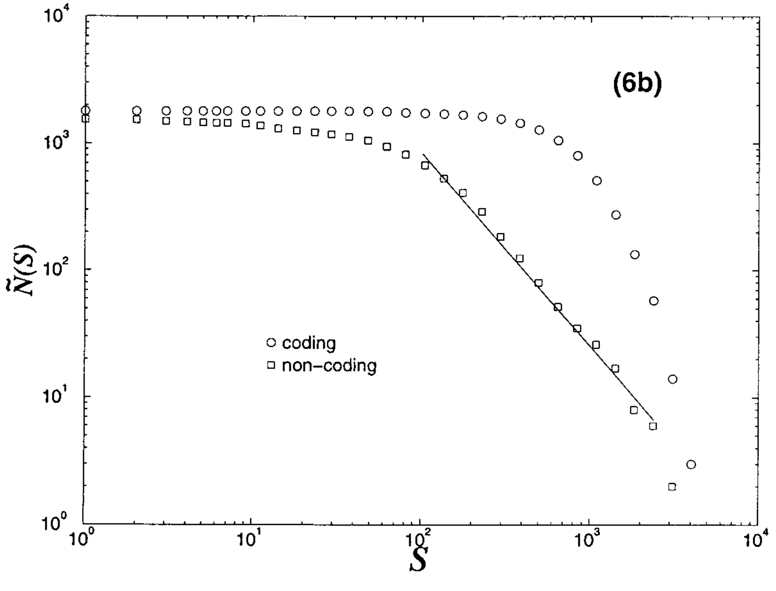


Fig. 6. The number of coding and non-coding regions of size $\geqslant S$, $\tilde{N}(S)$, for three procaryotic sequences. The straight lines have the following slopes: (6a) $-\mu = -1.3$ and (6b) $-\mu = -1.4$. Sequence (6c) presents a borderline behaviour.

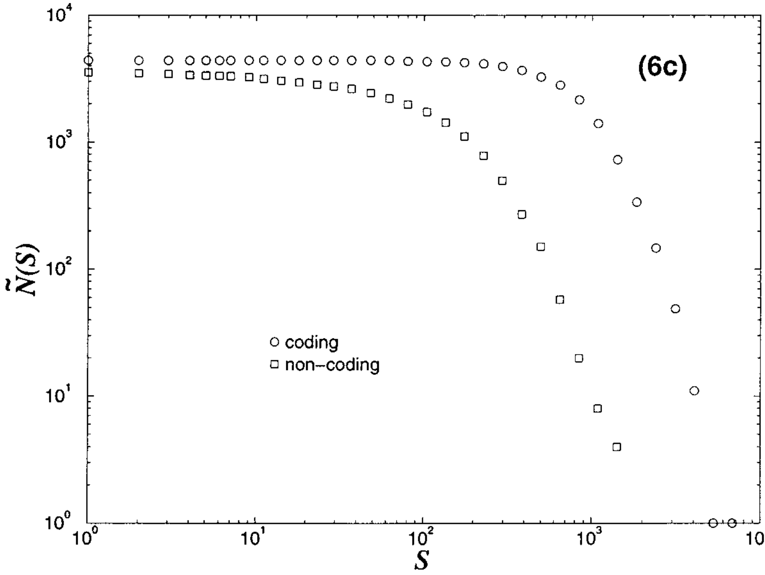


Fig. 6. (*Continued*)

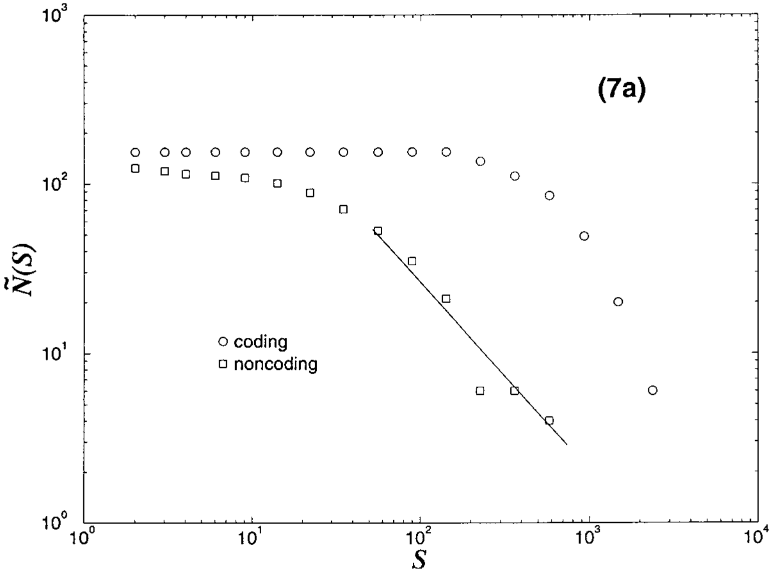
In both cases the coding size distribution shows clear short range behaviour. Equivalently to the above presented cases the organisms *Mycoplasma genitalium* and *Pyrococcus horicoshii* (all complete genomes) show also clearcut long range behaviour in the non-coding with exponents $-\mu = -1.3$ and -1.6 respectively.

In Fig. 6c we also present the boarderline case of *Escherichia coli.* We do not observe power law behaviour in the non-coding but clearly the size distribution falls more abruptly in the coding than in the non-coding. Similar borderline behaviour is seen for *Mycoplasma pneumonia*, *Synechocystis sp.* and *Bacillus subtilis.*

The existence of long range correlations in the non-coding part of procaryotic genomes confirms the predominant theory that the ancestral forms of these organisms were originally endowed with a large amount of non-coding but they have lost the largest part of it at some stage of evolution. As described elsewhere[4, 13] transpositions, early genome merging and aggregative dynamics are the standard prerequisite necessary for the observed long range behavior.

3.4. Viruses

Viruses, being cellular parasites, have a very restricted genome most of which codes for proteins. Thus the non-coding spacers in viral genome are



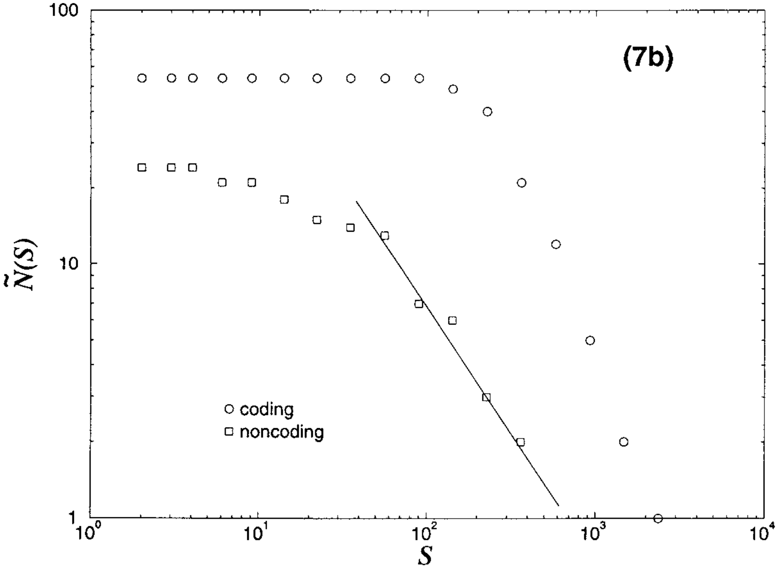


Fig. 7. The number of coding and non-coding regions of size $\geqslant S$, $\tilde{N}(S)$, for three viral DNA sequences. The straight lines have the following slopes: (7a) $-\mu = -1.1$, (7b) $-\mu = -1.0$. In (7c) the exponent is hardly observable.

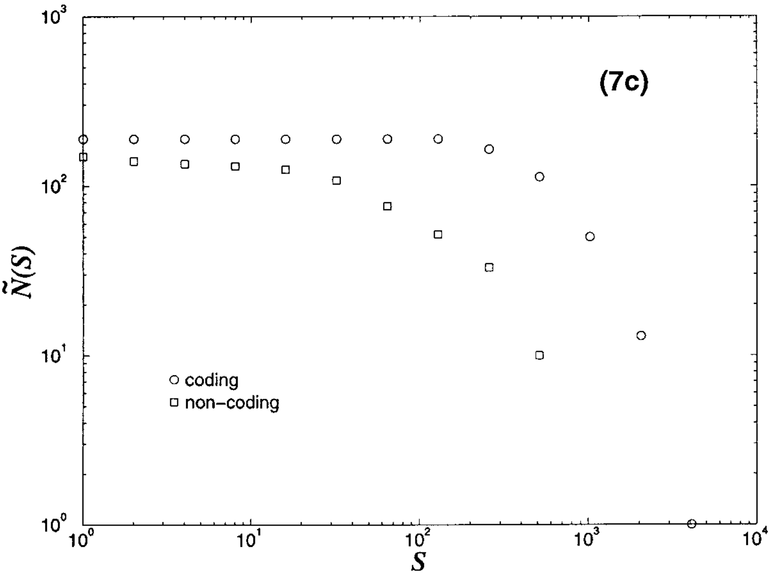


Fig. 7. (*Continued*)

much smaller than in all other organisms. Due to the restricted size of their genome many of them have already been sequenced. We present in Fig. 7 three characteristic cases. In all of them traces of long range distributions in the non-coding can be seen and they are more prominent in Figs. 7a, b while in Fig. 7c a clear exponent cannot be identified. The reduced size of viral genome does not allow the correlations to be developed in an extended way.

4. STRAND PARTITION OF CODING SEGMENTS

As it has already been discussed the genome of procaryotes is very rich (almost 90%) in coding regions. These coding regions are sequences of nucleotides beginning always with a triplet AUG (serving as **start** codon) and finishing with one the three **stop** codons UAG, UAA or UGA. The coding sequences of the chromosome are distributed, almost equiprobably, along the "Watson" and "Crick" DNA strands, which are in principle of equal potentiality. An exception is the mitochondrion where the "heavy" strand hosts practically all the protein coding segments. Note that mitochondrial DNA can be separated into heavy and light single strands on the basis of their density in centrifugation experiments.(11) In the present section we examine the properties of the partition of the coding segments between the two strands for eight procaryotic organism genomes for which the entire DNA sequence is known and annotated.

The question we would like to address is whether the coding regions were placed on the Crick or the Watson strand in a random manner or if there is some non-random dynamics on their placement. We first note that the coding regions have different lengths, but the distribution of these lengths is short ranged[3] and thus we may consider them as being roughly equivalent. We will refer to the coding regions as "segments" for simplicity. Segments are separated by non-coding regions, which are only regarded here as spacers, and they are not involved in the dynamics at this stage.

Let us suppose that the segments are "deposited" either on the Crick or the Watson strand in a random manner. If we have a long DNA sequence containing N segments, on the average half of them will be deposited on the Crick and half of them on the Watson strand. The probability $P_W(s)$ to find s consecutive coding segments on the same strand (for example on the Watson strand) is[18]:

$$P_w(s) = cp_w^s(1 - p_w)^2 \tag{6}$$

where p_w is the probability that one segment is located on the Watson strand. Since we have equiprobable deposition on the two strands $p_w = 1/2$. The constant c is set for normalization reasons. If the number of deposited segments $N \to \infty$, then $c = |\ln p_w|/(1 - p_w)^2$. The arguments are similar for the Crick strand.

The probability distribution P_w represents clearly an exponential decay since it can be written as:

$$P_w(s) = c(1 - p_w)^2 \, \mathrm{e}^{\,-s\,|\ln p_w|} \tag{7}$$

and the decay exponent for random and equiprobable deposition on the Crick or the Watson strand is $-|\ln p_w| = |-\ln(1/2)| = -0.693$.

It is well known that genes able to produce proteins only after their activation by an environmental factor (like externally applied temperature, salinity, starvation and other stresses) are clustered together in the procaryotic genome. These are clusters of consecutively (on the same strand) encoded genes, which share the same "promoter". The environmental influence on the activation of a gene is mediated by the promoter, which is a non-coding DNA region located upstream to the "inducible" gene cluster and allowing its transcription when a stimuli depended protein is bound on it. On the other hand the so-called "house-keeping" genes are not under the control of switchable promoters.

Protein-coding-segments often form same-strand-concatenations when their functions are just related to the same cell activity. In these cases there is a profit for the organism to have proteins transcribed on the same "polycistronic" mRNA molecule because their translation in protein molecules occurs simultaneously and in the right proportions.

All the above constraints compromise any expectation of a plain random picture for the strand partition. Only the examination of concrete genomes may determine the validity of a description in terms of exponential decay and may give the numerical values of the corresponding exponents.

We have examined the form of the decay for eight procaryotic genomes and have found mostly exponential decay. The corresponding exponents are given in the 3rd column of Table II. As we observe, in most cases the exponents deviate from the value calculated for the completely random case.

Let us now assume that there are indeed correlations between a number of segments. The most natural assumption is that the correlations are important between g consecutive units, where g is a positive, finite and fixed integer. If we now group the total number of segments N (on the Watson and Crick strands), in groups of g (or in groups of a number greater than g), then these groups are uncorrelated with each other.(19) The number of uncorrelated groups is now N/g. Assuming self-similarity in the scale of units and the scale of groups, the N/g groups are uncorrelated and random. They might be placed with the same probability on the Watson or the Crick $p_w = p_c = 1/2$ as the original units (self-similarity). Then the probability to find r consecutive groups on the Watson strand is

$$\bar{P}_w^{\text{correl}}(r) = cp_w^r(1-p_w)^2 = c(1-p_w)^2 \, \mathrm{e}^{-r\,|\ln p_w|} \tag{8}$$

and similarly for the Crick strand. However, the r consecutive groups contain in reality $s = r \cdot g$ segments. Then, the probability $P_w^{\text{correl}}(s)$ (unit

Table II. Characteristic Exponents in the Strand Dynamics of Procaryotes

No.	Organism	Length (Kbp)	Exponent	g
a	*Escherichia coli*	4639	0.28	2.5
b	*Pyrococcus horikoshii*	1739	0.60	1.2
c	*Haemophilus influenzae*	1830	0.30	2.3
d	*Methanococcus jannaschii*	1665	0.31	2.2
e	*Synechocystis sp*	3572	0.41	1.7
f	*Bacillus subtilis*	4215	[a]	[a]
g	*Mycoplasma pneumoniae*	816	0.20	3.5[b]
h	*Mycoplasma genitalium*	580	0.14	5.0[b]

[a] Sequence with power law dependence.

[b] Sequence with power law-like dependence.

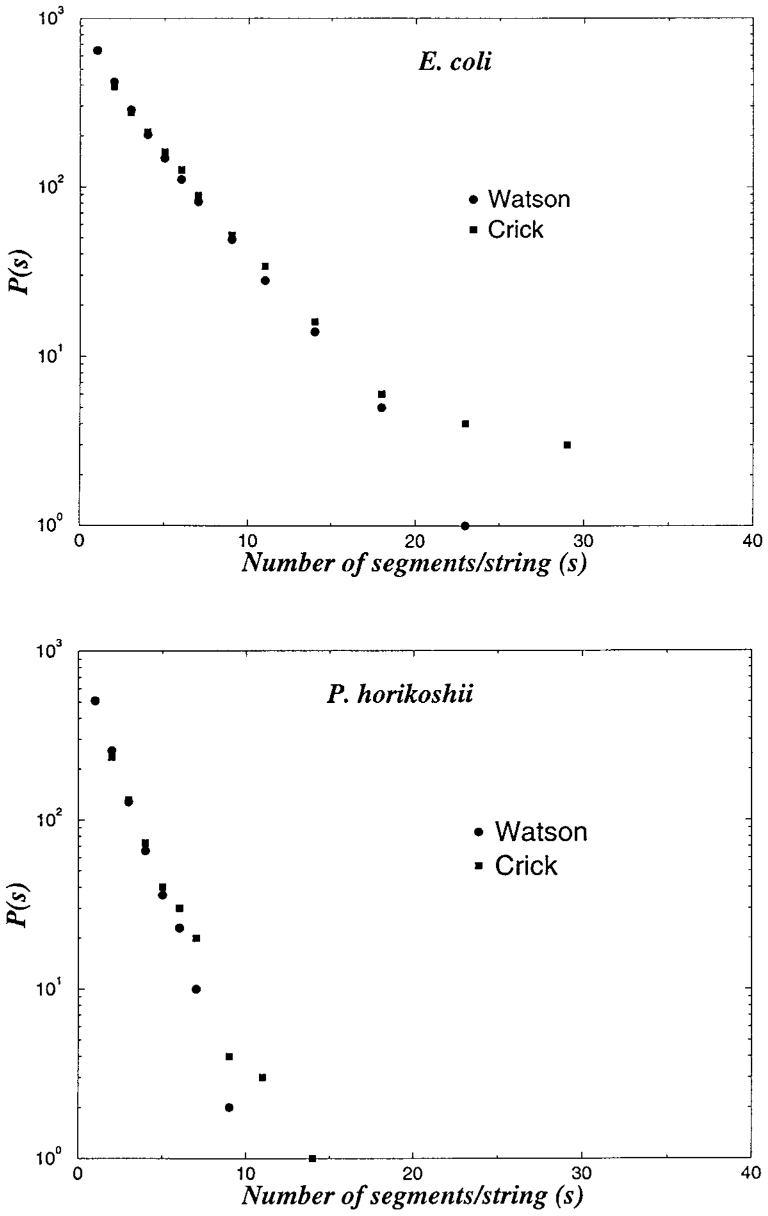


Fig. 8. The partition of coding segments in the two strands of three complete procaryotic genomes: (a) *E. coli*, (b) *P. horikoshii* and (c) *H. influenzae*. Squares and circles stand for the partitioning of each strand. The frequency of occurrence $P(s)$ is plotted as a function of the number of segments s.

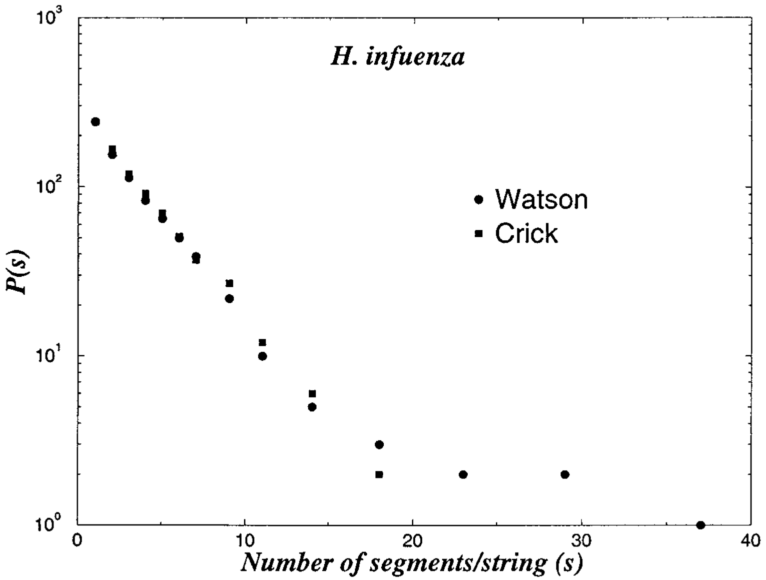


Fig. 8. (*Continued*)

distribution function) to find s consecutive segments on the same strand is related to $\bar{P}_w^{\text{correl}}(r)$ (the group distribution functions) as

$$\bar{P}_w^{\text{correl}}(r)\,dr = P_w^{\text{correl}}(s)\,ds \tag{9}$$

and thus

$$P_w^{\text{correl}}(s) = \frac{c}{4g}\,\mathrm{e}^{-s\,|\ln(1/2)|/g} \tag{10}$$

Eq. (10) indicates that the decay exponent is modified by the correlation length g. From the exponents computed in Figs. 8 we find, e.g., that for *E. coli*, Fig. 8a, $g = 2.6$, which indicates a clear shift to higher clustering.

The divergence from the value -0.693 of the determined exponents expresses the existence in the procaryotic genome of a rich organization corresponding to the biological constraints mentioned earlier and probably to several additional characteristics not yet identified. From the examined sequences we have found one case of almost random partition of coding segments between the two strands with corresponding value $g = 1.2$ for

P. horikoshii, Fig. 8b. Typical clustering (short range) of coding is observed in four cases with g ranging from 1.7 to 2.5. In two cases of mycoplasma, due to their particularly restricted genome, small deviations from exponential behaviour were observed giving rise to relatively high values of $g = 3.5$ and 5.0.

In particular, one of the two strands of *B. subtilis* shows long range (power law type) of gene clustering. Simple inspection of the *B. subtilis* genome reveals the frequence of particularly long concatenations of coding segments on the same strand. A power law decay in this strand seems to suggest a higher degree of internal organization of this genome. When more entire procaryotic genomes will be available we could conclude whether this type of strand partition is an exception or if *B. subtilis* is representative for a whole class of organisms.

Further investigation, both experimental and statistical is needed in order to reveal eventual advantages for the organism due to the adoption of this particular distribution. The same is true for the understanding of the evolutionary origin of strand architecture. The current opinion about the remote ancestors of the present-day procaryotes is that they were endowed with large non-coding regions, much similarly to eucaryotic cells.(10) Relatively frequent occurrence of transposition events, lateral gene transfer and extended genome parts fusion seem to have accompanied the primitive forms of life.(4, 12, 13) These events should have a direct effect in the Watson/Crick clustering of the coding segments as they imply double strand ruptures followed by ligation events. Further contributions to this dynamics may be due to the conjectured procedure of genome compactification which has eliminated the largest amount of non-coding DNA and may have lead to the clustering of the coding observed in the known procaryotes.

The study of the strands partition of coding segments for eucaryotic organisms is not included here. This is due to the following three reasons: (a) Eucaryotic genome is mainly non-coding (order of 10% coding) and any approach treating the juxtaposition of the coding segments and ignoring the intervening spacers seems questionable. (b) The DNA coding for most proteins in eucaryotic genomes is segmented in several exons which are juxtaposed only at the stage of the mRNA processing called "splicing."(10) Exons, coding for parts of the same protein, are usually located on the same strand. Thus the clustering of coding segments in eucaryotes is quite extended and mainly expresses the continuity of exons belonging to the same protein. (c) While the longest fully annotated sequences of higher organisms (available today) have number of coding (and non-coding) segments sufficient for the study of their length distribution, they do not have enough "Watson" and "Crick" clusters in order to assess the degree of their clustering.

5. CONCLUSIONS–PERSPECTIVES

The search for the statistical features of the genomic DNA sequences is a domain with a rapid growth both in the amount of the obtained results as well as in the range of their functional and evolutionary implications for organisms. The development of this domain is directly influenced by the exponential growth of the content of Molecular DataBases and the massive, computer-assisted annotation of the newly obtained large DNA sequences. Main sources for these sequences are the Genome Projects for several higher organisms.

In the present work we have investigated two novel statistical aspects of the organization of large genomic sequences:

First we study the size distribution of coding and non-coding DNA sequences belonging to different organisms ranging from higher eucaryotes to lower eucaryotes and procaryotes. Theoretically, starting from the long range distributions observed for the clusters of Pu and Py in the non-coding DNA and using the Generalized Central Limit Theorem, we concluded that the size distributions of the non-coding regions must also be long ranged. Using the same reasoning we conclude that only short range distributions must be observed in the length distribution of the coding regions. These theoretical conclusions are tested successfully by analyzing large DNA sequences from different taxonomic groups. These results present a qualitative aspect, i.e., the form of the size distribution decay (exponential or power law), and a quantitative one related to the order of magnitude of the obtained power law exponents.

The second part of our investigation deals with the partition of coding segments in the two complementary DNA strands, and as we have pointed out in Section 4 short range clustering characterizes a variety of compact-genome organisms (procaryotes).

As already discussed in Section 2, the long range correlations observed in the size distribution of non-coding DNA segments may be attributed to the similar statistical characteristics of the nucleotide islands which form the non-coding DNA. The problem of the biological origin of this feature is partly undertaken in refs. 13 and 4. In ref. 13 a minimal evolutionary model is put forward describing the past of present day genomes as a succession of two types of biologically plausible events belonging to two different time scales: (a) Fusions of primitive genomes or of large genomic parts with different nucleotide constitution and (b) Transposition events shuffling continuously (in evolutionary time) the non-coding parts while leaving mostly intact the coding segments. In ref. 4 a random open aggregation mechanism is proposed for the growth and evolution of genomic sequences. The random incorporation of external intruding macromolecules or of partial selfcopies creates long range correlations in the non-coding, regardless of the nucleotide constitution of the aggregating macromolecules. Computer simulations corroborate that the above biologically motivated, minimal models may generate probabilistic features similar to the ones observed on current coding and non-coding DNA.

Long range correlations have recently been observed in many characteristic distributions related to the non-coding DNA. We just cite here the Pu and Py cluster distributions,[1–3] the repeats of identical dimers[5] and now the size distribution of non-coding segments. Long range distributions and power laws are normally observed in closed thermal systems at equilibrium near the critical point. These systems approach criticality via a control parameter. Far from equilibrium situations accompanied in some cases by power law behaviour are met in the so called far-from-equilibrium phase transitions, where criticality is again approached by means of a control parameter, such as in turbulence.[20] Critical behaviour and power laws may appear automatically in irreversible dissipative systems without the need of a control parameter (such as aggregation).[15–16] Such critical behaviour is expected to appear in biological system, which are by nature irreversible and dissipative. We believe that the study of the statistical properties of biological systems will lead to a better understanding of the mechanisms underlying important biological processes such as molecular and species evolution, development and cell differentiation control.

APPENDIX. CALCULATION OF THE DOMINANT EXPONENT

Let us consider the coagulation of N macromolecules of sizes s_i, $i=1,..., N$. The macromolecule s_i has a size distribution function of the form

$$p(s_i) \sim s^{-\mu_i - 1}, \qquad i=1,..., N \tag{11}$$

For simplicity choose

$$\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_N \tag{12}$$

In ordinary cases we only have two or three different values of μ. The probability to find a cluster of size S after the aggregation of the N macromolecules is

$$p(S) = \prod_{i=1}^{N} p(s_i)|_{\sum_i s_i = S} \tag{13}$$

by Fourier transforming Eq. (13) we obtain

$$\begin{aligned} Z(\rho) &= \int dS\, \mathrm{e}^{-i\rho S} \prod_{i=1}^{N} p(s_i)\, ds_i\, \delta\left(S - \sum_i s_i\right) \\ &= \prod_{i=1}^{N} \int ds_i\, \mathrm{e}^{-i\rho s_i} p(s_i) \end{aligned} \tag{14}$$

Using Eqs. (14) and (11) we find the Fourier Transform of the probability distribution function[16] as

$$Z(\rho) \sim \prod_{i=1}^{N} (1 - i\rho^{\mu_i}), \qquad \text{for} \quad \rho \ll 1 \tag{15}$$

Reducing further Eq. (15), for small values of ρ,

$$Z(\rho) \sim (1 - i\rho^{\mu_1}), \qquad \text{for} \quad \rho \ll 1 \tag{16}$$

since μ_1 is the smallest of the μ's. By inverse Fourier transform of Eq. (16) we find[16] that the size distribution of the aggregate macromolecules follows a power law with the smallest exponent μ_1, for large values of S, as

$$p(S) \sim S^{-\mu_1 - 1}, \qquad \text{for} \quad S \gg 1 \tag{17}$$

REFERENCES

1. C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356**:168 (1992).
2. 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); A. Czirók, R. N. Mantegna, S. Havlin, and H. E. Stanley, *Phys. Rev. E* **52**:446 (1995); C. A. Chatzidimitriou-Dreismann, R. M. Streffer, and D. Larhammar, *Nucleic Acid Research* **24**:1676 (1996); Y. Almirantis and S. Papageorgiou, *Proceedings of the European Conference on Artificial Life*, J.-L. Deneubourg, page 9, ed. (U.L.B, Brussels, 1993); W. Li and K. Kaneko, *Europhys. Lett.* **17**:655 (1992).
3. A. Provata and Y. Almirantis, *Physica A* **247**:482 (1997).
4. A. Provata, *Physica A* **264**:570 (1999).
5. S. V. Buldyrev, A. L. Goldberger, C.-K. Peng, M. Simons, and H. E. Stanley, *Phys. Rev. E* **47**:4514 (1993).
6. N. V. Dokholyan, V. Sergey, S. V. Buldyrev, S. Havlin, and H. E. Stanley, *Phys. Rev. Lett.* **79**:5182 (1997).
7. A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, *J. Theor. Biol.* **151**:323 (1991).
8. H. Herzel and I. Grosse, *Physica A* **216**:518 (1995).
9. O. Popov, D. M. Segal, and E. N. Trifonov, *BioSystems* **38**:65 (1996); E. N. Trifonov, *Bull. Math. Biol.* **51**:417 (1989).

10. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell* (Garland Publishing, Inc., New York, 1994).
11. B. Lewin, *Genes VI* (Oxford University Press, Oxford, 1997).
12. Y. Almirantis and A. Provata, *Bull. Math. Biol.* **59**:975 (1997).
13. Y. Almirantis, *J. Theor. Biol.* **196**:297 (1999).
14. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1966).
15. H. Takayasu, *Fractals in the Physical Sciences* (Manchester University Press, 1990).
16. H. Takayasu, M. Takayasu, A. Provata, and G. Huber, *J. Stat. Phys.* **65**:725 (1991); H. Takayasu, *Phys. Rev. Lett.* **63**:2563 (1989).
17. S. Brenner, G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, and S. Aparicio, *Nature* **366**:265 (1993).
18. D. Stauffer, *Introduction to Percolation Theory* (Taylor and Francis, London, 1985).
19. P.-G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, London, 1979).
20. G. Nicolis, *Self-organization in non-equilibrium systems* (Wiley, New York 1977).