An evolutionary model for the origin of non-randomness, long-range order and fractality in the genome

Yannis Almirantis[1]* and Astero Provata[2]

Summary

We present a model for genome evolution, comprising biologically plausible events such as transpositions inside the genome and insertions of exogenous sequences. This model attempts to formulate a minimal proposition accounting for key statistical properties of genomes, avoiding, as far as possible, unsupportable hypotheses for the remote evolutionary past. The statistical properties that are observed in genomic sequences and are reproduced by the proposed model are: (i) deviations from randomness at different length scales, measured by suitable algorithms, (ii) a special form of size distribution (power law distribution) characterising different levels of genome organisation in the non-coding, and (iii) extensive resemblance in the alternation of coding and non-coding regions at several length scales (self-similarity) in long genomic sequences of higher eukaryotes. ***BioEssays*** **23:647–656, 2001.**

Introduction

The recent developments in molecular biology, leading to the determination of entire genome sequences, have been paralleled by systematic investigation of the statistical and probablistic aspects of genome organisation. Considerable effort was concentrated at first on the search for systematic differences between coding and non-coding sequences. At this level of organisation, the principal factor affecting the statistics of the ''biological text'' — written in the four letter alphabet of the nucleotides — is the use of the ''grammar and syntax'' of the triplet code. Several algorithms based on oligonucleotide statistics, codon usage, etc were devel-

[1]Institute of Biology, National Research Centre for Physical Sciences ''Demokritos'', Athens, Greece.
[2]Institute of Physical Chemistry, National Research Centre for Physical Sciences ''Demokritos'', Athens, Greece.
*Correspondence to: Yannis Almirantis, Institute of Biology, National Research Centre for Physical Sciences ''Demokritos'', 15310 Athens, Greece.
E-mail: yalmir@mail.demokritos.gr

Box 1: Definitions

Random processes: the output of die- or coin-tossing experiments and any process that could be put in one-to-one correspondence to them. For the purposes of our analysis it is useful to list here the following immediate implications of plain randomness on a symbol sequence (even if the involved symbols are not equiprobable):

- In a random sequence, the possibility of finding a symbol at a given position does not depend on the previous symbols.
- The size distributions of similar-symbol clusters are exponentially decaying in random sequences.

Non-random processes: any process that deviates non-trivially from the above.

Detrending: the procedure of filtering only specific length scales of interest and ignoring larger and/or shorter length scale features.

Long-range correlations (LRC). Correlations are the result of interactions between different constituents of a system. When these interactions extend within the entire system, then, the correlations are called *long-range.*

Power law distributions. LRC often characterise symbol sequences with over-represented long tracks (clusters) of similar symbols. More rigorously, indication of LRC is linearity in double logarithmic scale of the cluster size distribution of similar symbols for some length scales. This is the so-called *power law distribution*.

Fractal: an object whose characteristic features grow (scale) with the object's size with a power D_f smaller than the spatial dimensionality of the object.

Self-similar: any object whose statistical properties are independent of the observation scale. Intuitively, self-similar objects resemble themselves seen in different magnifications.

oped.[1–3] These were called "content methods" while the so-called "signal methods" use information derived from the search for signals known to surround coding and transcribed sequences.[4] Combinations[5] of these two methods have generated powerful analytical tools, which are still in use for the annotation of the new sequences produced by the ongoing genome projects.

In contrast, the statistical treatment of the middle- and large-scale features of genomic sequences has been undertaken systematically only in the last decade.[6] This only became possible when long sequences, i.e. whole chromosomal regions or chromosomes, and whole genomes of lower organisms, had been decoded. Two types of quantitative analysis are particularly useful in the study of long DNA sequences and will be reviewed here. These are: (i) the introduction of a suitable quantification of non-randomness, and (ii) the search for long-range order and fractality in a given sequence. In addition, the development of these methods and algorithms in the study of DNA sequences has benefited considerably from related experience in fields such the Statistical Physics of critical phenomena, the dynamics of chaotic and complex systems, the theory of signal transmission, etc. More specifically, the quantitative estimation of randomness in nucleotide sequences is considered in the framework of such quantifications in symbol sequences. The quantities developed in Statistical Physics, for example the mean value, and the standard deviation, which are used to describe random processes, may be suitably formulated to include the frequencies of occurrence of symbols, and furthermore for the characterisation of symbolic sequences.[10,11,24,25]

In the next section, we review some middle- and large-scale features of DNA sequences. We focus on several structural and statistical properties that may identify DNA sequences that play different roles in the genome, such as protein coding, t-/r-RNA coding, intronic and other non-coding DNA. These properties include the "DNA walk" approach, the "modified standard deviation" method, long-range order and fractality estimations. In the third section an evolutionary model is formulated and the biologically plausible evolutionary rules that are incorporated in the model parameters are presented. This scenario mainly consists of external sequence incorporations and internal transpositions, both of which are events with widely recognised participation in genome evolution. In the fourth section, a comparison of the properties of real sequences and of properties of model-derived sequences is presented. This comparison indicates that the types of events included in our model may reproduce the principal middle- and large-scale statistical features of DNA sequences as reported in the literature.[6,7,10–12,22,24] The final section is devoted to a discussion on the main results of the presented work and on the fitness of the statistical approach for the systematic study of genome structure and function.

Large-scale statistical features of DNA sequences

Non-randomness at the nucleotide level

In a pioneering work, B.E.Blaisdell[7] has pointed out for the first time that a "persistent global non-randomness distinguishes coding and non-coding eukaryotic nuclear DNA sequences". Several other workers have applied tools from standard Information Theory (Shannon Entropy, Kolmogorov's theory of randomness etc) to the study of genomic DNA sequences. The early works of Gatlin[8] and the demonstration[9] that, from the information theory perspective, mitochondrial DNA is closer to bacterial than to nuclear DNA, are examples.

In previous works, the current authors and other researchers concluded that the non-randomness of non-coding sequences is high, while coding sequences are mostly random. More specifically, a measure of the non-randomness of a DNA sequence has been introduced,[10] in terms of local fluctuations using the concept of the "DNA walk";[11] and an alternative measure of non-randomness of DNA sequences has also been introduced, based on a "modified standard deviation".[12] In both of these attempts, the scale-dependent non-randomness is measured by means of appropriate quantities, which are functions of a given length scale or block length m.

In the case of the modified standard deviation, the sequence is divided into segments of length m and, for every such segment, the standard deviations of the four nucleotide frequencies of occurrence are computed. In these standard deviation computations, the role of mean values is played by "local" means, calculated for a sequence region including the segment under consideration and its close neighbours. The "modified" standard deviations calculated in this way represent quantities related to the degree of local non-randomness inside the sequence. Averaging for all blocks of the sequence, and for the four nucleotides, we obtain the quantity $MSD(m)$ which relates to the mean non-randomness of the sequence (See Box 2).

Box 2 shows the use of blocks of a chosen length m and local mean values to define quantities characterising a sequence probabilistically. This characterisation stands for a given length scale, determined by a given value of m. We have discussed in depth elsewhere[10,12] that measures of the "scale-dependent non-randomness" are particularly useful in correlating probabilistic features to the functional role of the sequence. The measurement of scale-dependent non-randomness becomes possible through the introduction of "filters". These filters allow for "detrending" of a given sequence and the determination of its degree of randomness around a specific length scale (this is the role of the construction in Box 2). If such precautions are not taken, the large-scale patchiness (particularly present in long coding sequ-

Box 2: MSD Computation

The algorithm for the computation of *MSD(m)* may be described as follows.

A: We divide our sequence (of length *L*), into equal segments of a chosen "Block Length" (*m*), and we compute the four nucleotide frequencies of occurrence for every segment: $P(m)_{s,j}$ ($s = $ A, G, C, T and $j = 1, 2, 3, \ldots, N : N = L/m$).

B: We associate with each of these segments *j* of length *m*, a longer segment *j'*: $j' = 1, 2, 3, \ldots, N$ which is always centred at the centre of segment *j*. The length of these segments *j'* is denoted by *MV*, the "mean value computation segment length". We define the quantity:

$$\mathcal{Q}^2(m, MV)_{s,j} = (P(m)_{s,j} - P(MV)_{s,j})^2 \qquad (1)$$

for each j ($j = 1,2,3, \ldots N$) and each nucleotide ($s = $ A, G, C or T), where $P(m)_{s,j}$ is the same-nucleotide-frequency of the associated *j'* segment of length *MV*.

C: Then, we average over all *N* segments, defining in this way the "modified standard deviation" $MSD(m,MV)_s$ for a given block length, a mean value computation segment length and a given nucleotide:

$$MSD^2(m, MV)_s = \sum_{j=1}^{N} \mathcal{Q}^2(m, MV)_s / N \qquad (2)$$

D: Finally, we average over the four nucleotides in order to obtain *MSD(m,MV)* used in the following to characterise sequences:

$$MSD(m, MV) = \sum_{s=A,G,C,T} MSD(m, MV)_s / 4 \qquad (3)$$

m is the main parameter allowing us to fine tune the method for the search of the non-randomness of nucleotide sequences on a specific length scale. When a collection of DNA sequences is studied for various *m* values (see next section), the ratio $n = MV/m$ is kept constant and we use the notation *MSD(m,n)*. Very good results are obtained for *n* as small as 3 or 5. In the following, *n* is taken equal to 3 and dropped out in the notation.

ences) would introduce ambiguity in any correlation of the probabilistic characteristics of a sequence with its functional role.[13–16]

Two collections of 50 sequences each are used in Fig. 1A. These sequences originate from a variety of organisms and their lengths range from thousands to tenths of thousands of base pairs (bps). In one of these collections, sequences are mainly coding (i.e. coding $> 80\%$), while in the other collection, sequences are mainly non-coding (i.e. coding $< 20\%$). The continuous line histogram represents the distribution of the *MSD(100)* values for the non-coding sequences, and the dashed line histogram stands for the coding ones. Figure 1B shows data produced using the model introduced in the following sections.

The arrow in Figure 1A specifies the *MSD(100)* value of a random sequence, which is the output of a "random 4-symbol sequence generator". This algorithm is based on the standard Knuth[32] method for random number generation. This procedure guarantees independent, non-correlated, random consecutive symbols. The "frequencies of occurrence" of the four symbols in the sequence may be all equal or not.

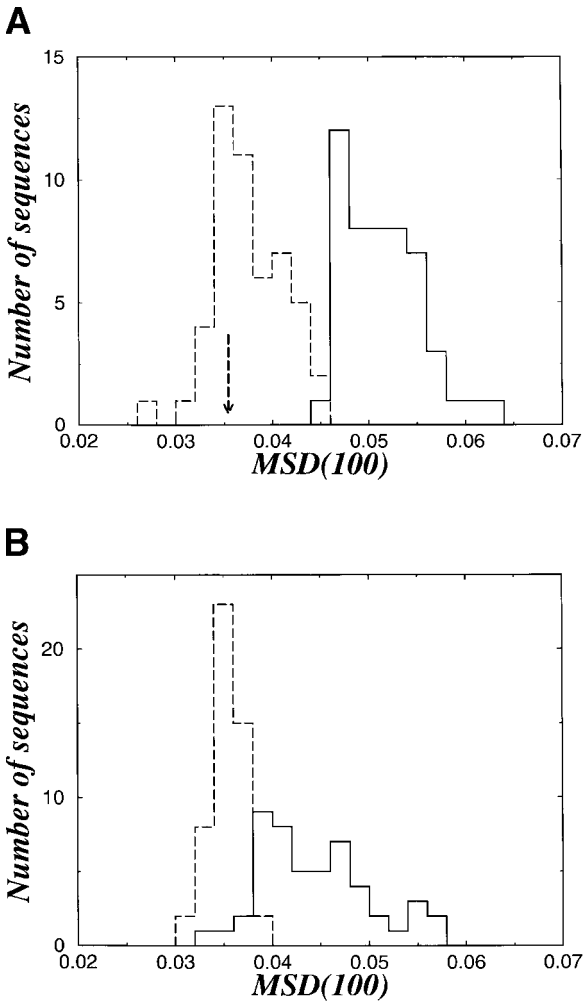

Figure 1. A: Distribution of the *MSD(100)* values of a collection of 50 coding sequences (dashed line) and a collection of 50 non-coding sequences (continuous line). **B:** Distribution of the *MSD(100)* values of a collection of 50 artificial random sequences (dashed line) and a collection of 50 sequences produced by the proposed model (continuous line).

Around the value $m = 100$, we obtain the maximum of prediction power of this method allowing an almost complete separation of coding-rich from coding-poor sequences. In addition, it has been observed that, for a large width of length scales ranging from tenths to hundreds of bps, non-coding sequences present clear non-randomness, related to their "mosaic structure". This means that non-coding DNA is composed of a juxtaposition of patches of various lengths with different local nucleotide constitutions.

Long-range order at the nucleotide level

Tools developed by Statistical Physics to study collective phenomena in physical systems can be used to determine the extent and uniformity of important statistical features in lengthy DNA sequences. "Long-range order" has been detected initially in non-coding DNA sequences by Peng et al.[11] using the so-called "DNA walk" construction, by R.F.Voss,[17] W.Li and K.Kaneko[18,19] while studying sequence periodicities, and later by other authors using different approaches.

We have examined coding and non-coding sequences for the existence of long-range correlations by studying the size distribution of clusters (or trains or islands) of similar nucleotides.[20] Let us present briefly the main idea in this approach. If a coin is tossed many times and the outcomes (heads = H, tails = T) are registered, HTTHHTTTHTHHHT-HHTHHH... might be a typical random result without any memory or correlation in the sequence, neither in the case of an ideal coin ($P_H = P_T$) nor in cases of falsified coins where $P_H \neq P_T$. Such cases of size distribution of trains of similar tossings are always of exponential form, with individual H or T outcomes more common that HH or TT, and so on for longer trains. The probability *P(s)* (frequency of occurrence) of a train of similar symbols of length *s* takes the form $P(s) \sim e^{-cs}$ (c is a constant). Such decaying distributions produce a linear graph in a linear-logarithmic scale. This linearity is a good criterion of randomness, or more precisely, of lack of long-range order, or "memory", in a long series of symbols. However, nature is also rich in another form of size distributions, the so-called "power law distributions", which may be represented by $P(s) \sim s^{-1-\mu}$ for $0 < \mu < 2$. For such distributions, the curve becomes linear if drawn in a double logarithmic scale. This behaviour can be seen in many systems, e.g., the total area of lungs measured at different length scales, the diameter distribution of blood vessels in a bat's wing or the diameter distribution of the craters on the moon surface. There are many other examples of these size distributions taken from various scientific fields.[21]

We have found that the size distribution of continuous clusters (trains) of purines (Pu) or pyrimidines (Py) in coding DNA sequences are of exponential (i.e. random-like or uncorrelated) form, and that the corresponding distributions in non-coding sequences are of the power law type.[20] This means that, in a large non-coding sequence, there is an overrepresentation of long continuous clusters (trains) of Pu or Py, if compared to the random two-symbol sequence that results from a coin-tossing experiment where: $P_H = P_{Pu}$ and $P_T = P_{Py}$. As an example, the size distribution of the Pu clusters of a coding-poor human sequence is depicted in Figure 2A. The linearity in the non-coding DNA Pu- or Py-cluster size distributions, observed in graphs drawn in double logarithmic scale, expresses the spatial self-similarity at the level of nucleotide juxtaposition of such sequences. The extent of this linearity expresses the width of the range of the self-similarity in the nucleotide juxtaposition of the examined sequence. Self-similarity is the property of an object to "resemble" a part of it (ideally, to "resemble" any, however

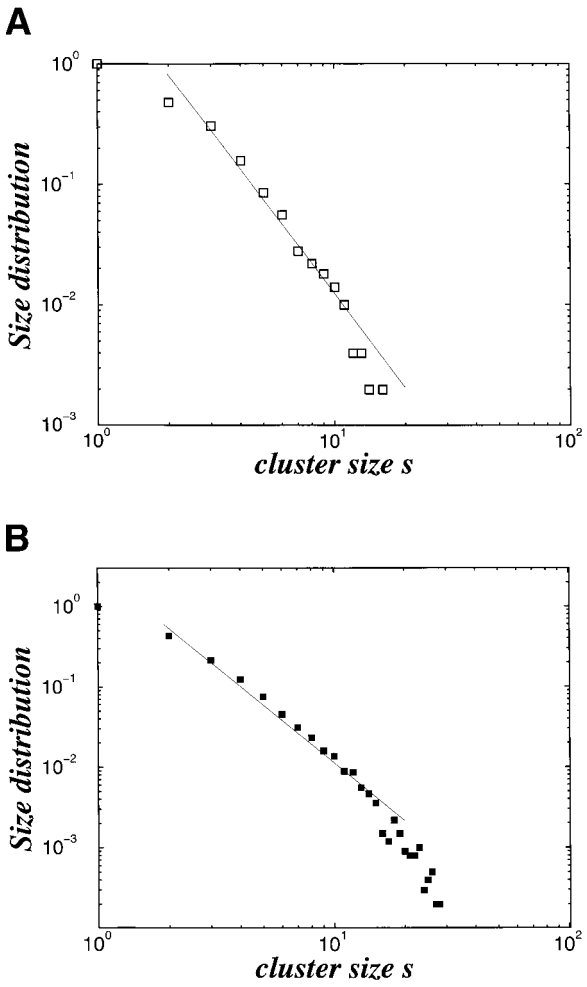

Figure 2. A: The size distribution of "Purine" clusters in double logarithmic scale, of Human 50 kDa type I epidermal keratin gene (EMBL name: HSEPKER). $\mu = 1.6$. **B**: The size distribution of "Pyrimidine" clusters in double logarithmic scale, of the model-produced sequence described in section 3. $\mu = 1.5$.

small, part of it). The power exponent -μ of the nucleotide size distribution offers a quantified expression of the degree of "long-range order/correlation". The lower the value of μ the stronger the degree of correlation, and the more marked the "long tail" of the distribution is.

Long-range order at the level of alternation of coding and non-coding regions

Investigating a higher level of organisation of the genome, we examined systematically the size distributions of the coding segments and the intervening non-coding regions[22] (introns or intergenic spacers). We conclude that the coding segment size distributions are of the exponential type. In contrast, non-coding segments, at least in higher eukaryotes, follow power law size distribution (see Fig. 3A). Figure 3A uses the "cumulative size distribution", where the number of all segments of length higher or equal to *s* is plotted for each value of *s*. Depending on whether the original distribution is short ranged or power law, the corresponding cumulative distribution is of the same type. Cumulative distributions are "integral forms" of the original ones and are more suitable for illustrating the features of rather small collections. If the exponent of the original power law size distribution is $-\mu$-1 then, the exponent of its cumulative form will be $-\mu$. Note that, despite the very restricted sizes of the non-coding spacers in prokaryotic genomes, in at least half of the examined cases remarkable linearity is obtained in double logarithmic scale. This residual "long-range order" could be attributed to the existence of an ancestor of prokaryotes with extended non-coding parts in its genome.[23]

Fractality in the coding/non-coding structure

Fractality has been detected in the juxtaposition of coding and non-coding regions in higher eukaryotic genomes.[24] As (linear) fractals, we consider here geometrical objects comprising a large (ideally infinite) number of short "black" segments, interrupted by large "white" spacers and characterised by self-similarity at some (ideally at any) length scales. In Figure 4A, the so-called box-counting method is used to determine the existence and the degree of fractality in three long genomic sequences. Sequences are viewed as juxtapositions of coding (black) and non-coding (white) regions. For every size *s* of (one-dimensional) boxes, the plot gives the number of boxes necessary to cover the total black space. Linearity in double logarithmic scale with slope value -$D_f \neq -1$ reveals fractality. The value of D_f is called "fractal dimension" of the examined sequence. In Fig. 4A, the straight line fits the data of the human sequence and has a fractal dimension $D_f = 0.83$. Large genomic regions of higher eukaryotes are in general characterised by clear fractality and D_f around the value 0.85. For more details see Ref. 24.

The probabilistic and statistical features of coding DNA presenting quasi-randomness at intermediate and large scales are not in contradiction with its functional role. Deviations from randomness due to the "protein-coding linguistics" occur mainly at the length scale determined by the "word length" (three for the principal protein coding procedure, see also Ref. 25). In contrast, the highly non-random features of non-coding DNA and, more specifically, its tendency to give rise to power law size distributions needs to be explained on the grounds of its history and function. In the next section, a minimal model for early genome evolution is introduced aiming at the clarification of the origin of these middle- and larage-scale statistical properties of the non-coding DNA.

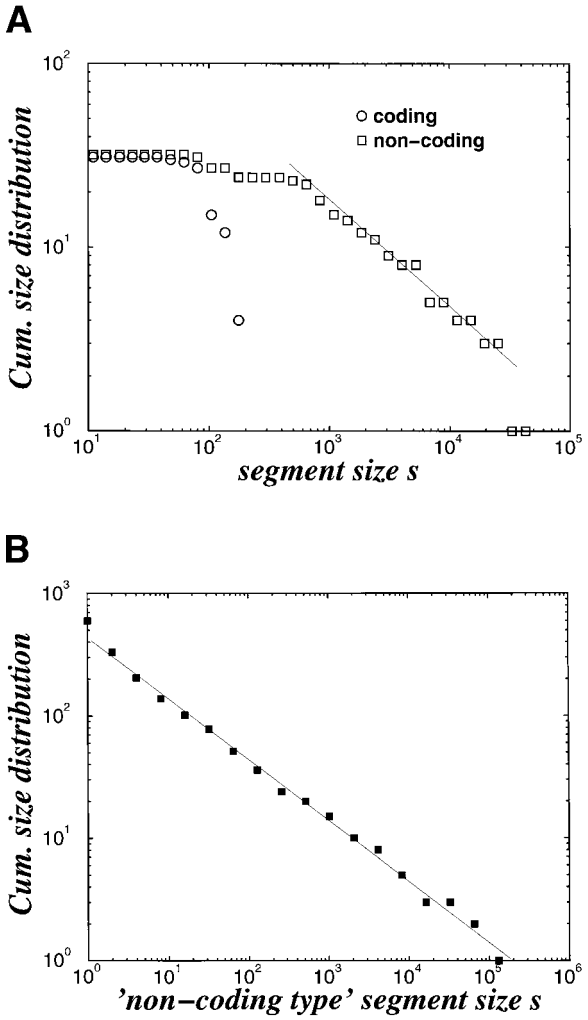

Figure 3. A: The cumulative size distribution of the coding (circles) and non-coding (squares) segments for the *Homo sapiens* chromosome 16 BAC clone CIT987SK-334D11, complete sequence (GenBank name: HSAF001550). $\mu = 0.6$. **B:** The cumulative size distribution of the "non-coding type" segments of the artificial sequence described in section 2c. $\mu = 0.5$.

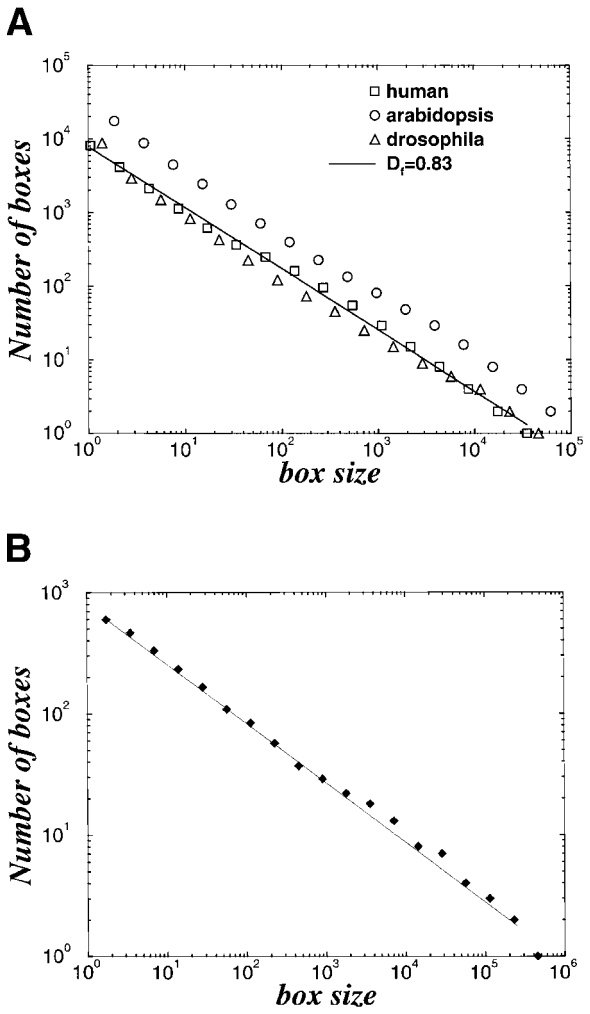

Figure 4. A: The box-counting curve for the determination of the fractal dimension of three long higher eukaryotic sequences: *Homo sapiens* collagen type VII intergenic region and (COL7A1) gene, *Arabidopsis thaliana* chromosome II BAC F4L23 genomic sequence, and *Drosophila melanogaster* cosmid clone 80H7 (from the distal X chromosome), with GeneBank names: HUMCOL7A1X, AC002387 and DMC80H7 respectively. **B:** The box-counting curve for the fractal dimension determination for the model-derived sequence of Figure 3B.

The evolutionary model

An evolutionary model is developed in order to interpret some key statistical properties of the genome. The model takes into consideration biological mechanisms/events that are commonly accepted. These events are capable of progressive modification of initially random nucleotide chains. As a result, several non-trivial statistical properties appear, which are indeed observed in real genomic sequences. The complexity of the proposed model has been kept to a minimum, in order to avoid ad hoc hypotheses for the remote evolutionary past. It aims principally to explain the discussed genomic features.

The types of events included in the proposed model are:

1) Insertion of exogenic segments of variable length at random positions in a nucleotide chain. The initial form of this chain is an heteropolymer comprising of four types of monomers placed in a random manner. Intrusions of relatively short segments (e.g. due to viral infections, viroids etc) are considered to occur more often than the incorporation of very lengthy foreign sequences. The probability of insertions interrupting coding regions is considered to be relatively small due to their effect on the viability of the organism.

A realistic hypothesis corroborating the use of insertion events in this model is that during the prebiotic past of the organisms and during the first stages of their evolution, the "fluidity" of genomes was considerably higher than now. Extensive "lateral gene transfer" is now accepted as having been important at the origin of several evolutionary events, like the formation of the present day eukaryotic organisms.[26]

2) Events of transposition or duplication-and-transposition inside the genome. Internal transposition events take place at random positions and are considered to occur much more often than events of external sequence segment insertions. This continuous (at the evolutionary time scale) mixing or "shuffling" of the genome occurs with a considerable probability only if coding regions are kept intact. The selective mixing causes a progressive homogenisation of the largest part of the genome, which is non-coding. Inhomogeneities reappear in the genome due to new intrusions of segments of different nucleotide constitution. Meanwhile, a mosaic constitution pattern is adopted by the genome with features of non-randomness and self-similarity (see following section).

It is well known[10,11,27] that "large-scale patchiness" is a common characteristic of the lengthy coding regions. The scale of these patches is of the order of thousands of bps. They consist of regions with different nucleotide constitutions and relatively sharp boundaries.[11] This seems to occur because the shuffling due to the transpositions leaves coding regions relatively intact. Thus, a more "ancient" (preserved in evolutionary time) pattern of genome constitution is still present in the coding regions. The large-scale patchiness must be distinguished from clustering of the nucleotide constitution at several length scales which is present in non-coding regions and may be seen as "mosaic-like structure". Here, instead of a characteristic length scale (a prevailing range of magnitude for the same-constitution regions), we have clustering present at several length scales (self-similar) due to the aggregative mixing of sequence parts with different constitutions, thus producing the typical for the non-coding power law distributions (see Fig. 2A).

We have applied the above scenario using various choices of the parameters: number of genome fusion events, number of transposition events, mean length of the transposed segments etc. It has been verified that the "good" behaviour (reproduction of the statistical properties appearing in genomic sequences) is a structurally stable feature of the model. That is, it applies to large regions in the parameter space and is resistant to fluctuations of parameter values. Moreover, its "robustness" for several rates of ubiquitous point mutation processes has been checked. The following choice of settings was used in a concrete simulation scenario of mixing and evolution of initially random sequences. A random nucleotide sequence of length 25 kbp and nucleotide densities (for the one strand), $P_A = P_G = 0.40$; $P_C = P_T = 0.10$, is fused end-to-end with another, of equal length, and with $P_A = P_G = P_C = P_T = 0.25$. In the resulting sequence, 200 transposition events are imposed: pieces of the nucleotide chain of length 50–200 base pairs are subtracted at random positions and are again randomly incorporated in new positions. Let us call the resulting sequence S_1. Another sequence S_2 is formed in the following way. A random nucleotide sequence of length 25 kbp and nucleotide densities (for the one strand), $P_A = P_G = 0.10$; $P_C = P_T = 0.40$, is fused end-to-end with another sequence of equal length, and with $P_A = P_G = P_C = P_T = 0.25$. In the resulting sequence, 200 transposition events occur, just like in sequence S_1. Then, sequences S_1 and S_2 are fused end-to-end and in the resulting new sequence 800 transposition events are allowed to take place. The resulting artificial sequence will be used in the following to compare its statistical features to those of real genomes. It has been confirmed that the statistical features of the above artificial sequence remain qualitatively unchanged if, instead of a few intrusions of huge external sequence (two in the above version), more insertions of smaller external macromolecules of different nucleotide constitutions are assumed.

Results

Scale-dependent non-randomness of the model- derived sequences

We have produced 50 randomly generated sequences as a first approximation to coding sequences and 50 sequences, using the model and the parameter settings introduced in the previous section, assuming non-coding conditions (i.e. interruptions allowed). All these sequences are of the same length (6 kbp). In Figure 1B, the histograms for the random set (dashed line) and for the model-derived set (continuous line) illustrate the distributions of the corresponding *MSD(100)* values. The qualitative resemblance to Figure 1A is obvious. More precisely, we observe that the two distributions are in both cases almost completely separated and that the range of *MSD(100)* values for the model-derived sequences is very similar to the genomic non-coding sequences range. It should be noted, however, that the dispersion of the set of coding sequences presented in Figure 1A is higher than the dispersion of the set of random sequences used here. This particular feature of coding sequences is discussed elsewhere[10,12] and implies that coding sequences behave as rather near-random than purely random sequences. Their deviations from randomness are due to different causes and vary between coding sequences of various origins. Figure 1A is principally affected by the already mentioned tendency of the coding DNA for large-scale patchiness. Such a patchy form is unevenly shared between sequences of different origin and is particularly visible in long prokaryotic and viral genomes. This is mainly a large-scale characteristic, but it also affects moderately the sequence randomness at the length scale of $m = 100$ bps. For reasons of simplicity and clarity, only results related to the length scale $m = 100$ are presented in Figure 1A,B. It has to be noted, however, that sequences produced by the proposed model present high *MSD(m)* values (i.e. values out of the range found in coding or random-like sequences) for all examined length values (from $m = 20$ to $m = 1600$), just as do the real non-coding sequences (see figures 6 and 7 in Ref. 12).

Long-range correlations in the nucleotide clustering of the model derived sequences

The size distribution of "pyrimidine clusters" of a sequence produced using our model is presented in Figure 2B, assuming again non-coding conditions (i.e. interruptions due to sequence shuffling allowed). Linearity in double logarithmic scale is again present, with $\mu = 1.5$. This means that the combination of fusion between large sequences of different initial nucleotide constitutions with mixing (due to transposition events for long evolutionary time intervals), produces the self-similar characteristics of the non-coding. The deviation from linearity (i.e. from the typical "power law" distribution) for large values of cluster size s is due to several reasons. Most essential is the finiteness of the examined sequence. Another reason is that, in the concrete scenario that we have simulated in Figure 2B, only two "end-to-end genomic fusion" events (i.e. end-to-end concatenation of two sequences) are considered. Our numerical simulations suggest that a longer maturation with more events of long external sequence intrusions, combined with transpositions, may further extend the linear region in the log-log plot. We have also to take into account the fact that the extent of the linear region in Figure 2B is as good as the best of our plots for non-coding genomic sequences. In the simulations, clearly more extended linearity is reached if the concentrations of the initially mixed artificial sequences are close to pure pyrimidine or purine polymers (data not shown). However, the presented model simulation results fit better to the genomic sequence properties. This is an indication that the departure concentrations of (primordial) genomic sequences were not (and of course, had not to be)

almost homopolymer-like. In addition, in long genomic sequences, probably ''impurities'' of several types further prevent the power law behaviour appearing in higher length scales. Such ''impurities'' may be the existence of some unidentified coding space, pseudogenes, microsatellites, repeats and so on.

We have included repeats in the list of ''impurities'' which may decrease the long-range order in large sequences. This is true for the repeats usually met in huge numbers in higher eukaryotic genomes like Alu and Line1 (see for example the complete sequence of human chromosome 22, Ref. 28) and microsatellites. These repeats alter the size distributions of nucleotide clusters of a sequence in a short-ranged way, depending on their abundance. On the contrary, it has been pointed out that replication events of a different type combined with mutations may generate long-range correlations in long DNA sequences.[18,19] In the description of our model, we have included duplications as a natural companion of transposition events in the molecular dynamics of the genome.[23] Moreover, in simulation experiments, we have verified that duplications may facilitate the long-range clustering in genome evolution. In the numerical examples described here, however, we always ignore the duplication possibility. This is done in order to illustrate clearly the long-range potentialities of the combination in the genome history of simple transpositions with aggregation, at various length scales.

Long-range correlations in the non-coding spacers distribution in model-derived sequences

Let us consider the proposed model acting on a long sequence where an alternation of coding segments and non-coding spacers is assumed. The initial size distributions of both populations are random by construction, thus of exponential form. Here, it is not necessary to include the nucleotide constitution in the simulation procedure, but even at this higher (functional) level description, the model action involves both external intrusions and internal events of transposition.

With reference to the proposed model, the difference between sequence parts marked as ''coding'' and others marked as ''non-coding'', is that, in the former, the probability to be interrupted from the intrusion of an external (or transposed) segment is taken to be very small or zero. Instead, in the non-coding parts, incorporation of the intruder occurs with a relatively high probability. This reflects the difference in the time scales of modifications allowing for a viable progeny, between the non-coding and the coding (or other highly conserved regions) in real genomes.

It has been proven theoretically that such an evolutionary sequence of events, seen in the framework of the aggregation dynamics, spontaneously creates long-range correlations.[29] The result of those correlations is a power law distribution in the sizes within the non-conserved subpopulation of segments. As mentioned, this subpopulation represents the evolution of the non-coding sequences.

The cumulative size distribution of ''non-coding'' segments of such a model-derived long sequence is depicted in Figure 3B in double logarithmic scale. During the simulation of the model, 9×10^6 insertion and mixing events (with insertion/mixing ratio 0.05) have taken place, and 3000 ''coding'' and 3000 ''non-coding'' segments have been formed. The distribution for the ''non-coding'' has clearly adopted the power law form with exponent $-\mu = -0.5$, although originally the size distributions of both ''coding'' and ''non-coding'' segments were constructed to be indistinguishable and random. This exponent is comparable with the one shown in Figure 3A obtained from a human sequence and other data presented in Ref. 22.

Fractality in the coding/non-coding structure in model derived sequences

The alternating ''coding/non-coding'' structure produced by the model presents both short range features in the coding size distribution and long-range features in the non-coding size distribution. This alternation of relatively narrow-distributed coding parts separated by non-coding parts whose size distribution covers several length scales (in both genomic and model-generated sequences) suggests a fractal structure. To calculate the fractal dimension of the artificial sequence used in Figure 3B, we again apply the box-counting method used earlier for the calculation of D_f of real sequences. The application of this method produces a linear curve in double logarithmic scale (see Fig. 4B) whose slope is $-D_f = -0.5$. The same value is obtained theoretically using open aggregative/mixing models[29]. Since for the production of Figures 3B and 4B, the same simulation model and the same parameters are used, the computed values for the power exponent $-\mu$ and the fractal dimension D_f are compatible, as theory predicts.[34] The value of D_f produced by this minimal model is in fairly good agreement with values obtained from eukaryotic sequences (see Fig. 4A).

Notice that the robustness of the long-ranged and fractal features in the coding/non-coding structure, as produced by the proposed model, is tested against a variety of realistic situations that have probably contributed to the structure and complexity of present day genomes. Such situations may be: (i) Intrusion of lengthy parts including coding segments (lateral gene transfer), (ii) dependence of the incorporation rate into an individual non-coding region, on its own size and (iii) various ratios between insertion / transposition events.

Discussion

Some intermediate and large-scale statistical properties of genome organisation are reviewed and a minimal model is proposed based on biologically motivated mechanisms, which may account for the aforementioned statistical properties.

There is obviously a long way to go until a satisfactory understanding of the genome structure and its role as the principal information carrier in living organisms. The genome, as a whole, seems to possess an outstanding ability to retain functionality during profound developmental modifications (which accompany differentiation and ageing) and during major evolutionary events. Such events are the assumed genome compactification of the ancestor of present day prokaryotes (which are supposed to have had genomes with extended non-coding parts, Ref. 23) and the genome size reduction of the pufferfish.[30,31] These global events, just like the step-by-step evolutionary modifications of the genome, prove its high "stability"; i.e. the genome as a whole retains, after such changes, the functionality of its coding parts and the efficiency of its complex addressing index which is responsible for cellular differentiation and thus for coherent development. In contrast, the extended non-coding part of the genome seems to serve as a source of "biological text", and also seems to facilitate exon shuffling and generation of new proteins.

Self-similarity, which was demonstrated for the genome structure, is also found in several systems, such as the form of the blood circulation pattern and the size distribution of the vein diameters. Such a circulation pattern (fractal and self-similar for several length scales, Ref. 21) guarantees that the blood stream passes in the proximity of any single cell. The existence of only thick veins would make this impossible, while only capillary circulation would cause a rise in internal friction to unacceptable levels. The occurrence of a self-similar, power law distributed, non-coding segment pattern in the genome also seems to offer evolutionary advantages. As we have described in this article (for the theoretical deduction of this property see Ref. 29), the aggregative interplay of insertions/translocations in the genome of higher eukaryotes may produce this genomic pattern. Such a non-coding size distribution offers: (i) a large number of non-coding spacers (introns) for the separation of exons of the same protein, (ii) more lengthy spacers for the separation of proteins belonging to the same group of genes, often under a common transcriptional control, and (iii) large intergenic non-coding regions. Such a hierarchy of lengths typically obeys a power law distribution.[21,34] Probably, the localisation of coding segments and regulatory elements close to non-coding segments of various lengths, offers to the evolution the opportunity to use this proximity, occasionally transforming small parts of the non-coding into coding regions, if needed.

We have presented several middle- and large-scale statistical features of genomic DNA that distinguish functionally different sequences. The predictive power of our approach may be assessed if we form test collections of coding-rich and coding-poor sequences and apply the algorithms of randomness quantification discussed so far (see Fig. 1A,B and for further discussion in Refs.10,12). The undertaken statistical approach principally aims at understanding of aspects of genome evolution. Thus, we propose the model presented in the previous sections. On the one hand, this model describes a minimal plausible evolutionary scenario consistent with the aforementioned statistical properties. On the other hand, this statistical treatment does not seem to fit directly with the need for precise localisation of coding segments during the annotation of new sequences. The length scale in our approach clearly exceeds the word length of the protein coding and thus the length scale in the grammar and syntax of the coding procedure. Signal methods,[4,5] assessment of the extent of open reading frames, and related techniques are largely used for the detection of protein coding in sequence annotation during the genome projects.

For the same reason, a variety of events of paramount importance for biological evolution, like codon bias, mutation pressure, replication slippage and microsatellite dynamics, etc are not directly influencing the genome statistics at the length scale principally addressed here. Their influence on genome statistics is mostly expressed at the length scale of the triplet formation thus determining the "meaning" of the genetic message at the level of the protein aminoacid sequence.

As we have already remarked (for more details see Ref. 10 and 12), what we describe in this article as "coding" DNA is not only the protein-coding segments but equally the r- and t-RNA coding parts of the genome. Recent results[35] indicate, however, that promoter sequences, despite a first intuition, have non-randomness measures close to that of the plain non-coding, despite their important role in the gene expression regulation. In contrast to coding sequences, the promoter functionality relies on weaker prerequisites: only conservation of very short "consensus sequences" and their relative distances generally need to be guaranteed. Thus, a considerable fraction of all the occurring transposition/insertion "accidents" may leave intact these relative distances and the functionality of the promoter. The effect of such tolerance in long evolutionary time would be the increase of the non-randomness in promoter regions. Further work on collections of similar-functionality sequences collected from the complete content of databanks is needed in order to characterise completely the degree of randomness of different components of the genome, especially when homogeneity in function does not extend throughout lengthy sequences, as is the case in regulatory elements.

A better understanding of the genome structure and functionality may be gained by the introduction of quantitative markers formulated on the basis of statistical considerations. Such markers may be suitable measures of non-randomness, values of exponents of power law distributions, the extent of the self-similarity of such distributions (extent of linearity in log-log scale), several fractal features, measures of the repetitivity of sequences[33] etc. The application of such large-scale markers in the annotation and the pictorial description of whole chromosomes and genomes of higher eukaryotes will con-

tribute to the classification and better understanding of the growing amount of genetic information. In addition, the use of such markers can serve as an alternative way to make structural and functional comparison of the genetic material of different organisms and for the understanding of the chromosomal architecture.

References

1. Mani GS. Correlations between the coding and non-coding regions in DNA. J Theor Biol 1992;158:429–445.
2. Mani GS. Long-range doublet correlations in DNA and the coding regions. J Theor Biol 1992;158:447–464.
3. Staden R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucleic Acids Res 1984;12:551–567.
4. Mulligan ME, Hawley DK, Entriken R, McClure WR. E. coli promoter sequences predict in vitro RNA polymerase activity. Nucleic Acids Res 1984;12:789–800.
5. Staden R. Graphic methods to determine the function of nucleic acid sequences. Nucleic Acids Res 1984;12:521–538.
6. Li W. The study of correlation structures of DNA sequences:a critical review. Computers and Chemistry 1997;21:257–271.
7. Blaisdell BE. A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences. J Mol Evol 1983;19:122–133.
8. Gatlin LL. The information content of DNA I, II. J.Theor.Biol., 1966;10: 281–300.
9. Granero-Porati MI, Porati A. Information parameters and randomness in mitochondrial DNA. J Mol Evol 1988;27:109–113.
10. Almirantis Y, Provata A. The ''clustered structure'' of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences. Bull Math Biol 1997;59:975–992.
11. Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE. Long-range correlations in nucleotide sequences. Nature (London) 1992;356:168–170.
12. Almirantis Y. A standard deviation based quantification differentiates coding from non-coding DNA sequences and gives insight to their evolutionary history. J Theor Biol 1999;196:297–308.
13. Nee S. Uncorrelated DNA walks. Nature (London) 1992;357:450.
14. Prabhu VV, Claverie JM. Correlations in intronless DNA. Nature (London) 1992;359:782.
15. Munson PJ, Taylor RC, Michaels GS. DNA correlations. Nature (London) 1992;360:636.
16. Almirantis Y, Papageorgiou S. Long or short range correlations in DNA sequences? In the ''Proc Eur Conf Art Life'', Sept. 1993, Brussels, 9–14.
17. Voss RF. Evolution of long-range fractal correlations and 1/f noice in DNA base sequences. Phys Rev Lett 1992;68:3805–3808.
18. Li W, Kaneko K. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. Europhysics Letters 1992;17:655–660.
19. Li W. Generating nontrivial long-range correlations and 1/f spectra by replication and mutation. Int J of Bifurcation and Chaos 1992;2:137–154.
20. Provata A, Almirantis Y. Scaling properties of coding and non-coding DNA sequences. Physica A 1997;247:482–496.
21. Takayasu H. Fractals in the physical sciences. Manchester University Press, Manchester and New York, 1990.
22. Almirantis Y, Provata A. Long and short-range correlations in genome organisation. Journal of Statistical Physics 1999;97:233–262.
23. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular Biology of THE CELL, Garland Publishing, Inc., New York & London, 3rd edition, 1994.—For the existence of a prokaryotes' ancestor with extended non-coding DNA see p.389–391.
24. Provata A, Almirantis Y. Fractal Cantor patterns in the *sequence structure of DNA. Fractals 2000;8:15–27.
25. Trifonov EN. The multiple codes of nucleotide sequences. Bull Math Biol 1989;51:417–432.
26. Doolittle WF, Logston JM Jr. Archaeal genomics:Do archaea have a mixed heritage? Curr Biol 1998;8:R209–R211.
27. Karlin S, Brendel V. Patchiness and correlations in DNA sequences. Science 1993;259:677–680.
28. Dunham I et al. The DNA sequence of human chromosome 22. Nature (London) 1999;359:489–495.
29. Provata A. Random aggregation models for the formation and evolution of coding and non-coding DNA. Physica A 1999;264:570–580.
30. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S. Characterisation of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature (London) 1993;366:265–268.
31. Lewin B. Genes VI, Oxford University Press, Oxford, New York, 1997. For a general discussion about the genome size variability in several taxonomic groups see chapter 23.
32. Knuth DE. *The art of computer programming*. Addison-Westley Pub. Compagny. 1981.
33. Hancock JM, Armstrong JS. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput Appl Biosci 1994;10:67–70.
34. Vicsek T. Fractal growth phenomena. World Scientific, Singapore, London, 1989.
35. Nikolaou C, Almirantis Y. A classification of functionally different DNA sequences according to their degree of randomness. In Preparation. 2001.