Journal of Statistical Physics, Vol. 106, Nos. 1/2, January 2002 (© 2002)

Statistical Dynamics of Clustering in the Genome Structure

A. Provata[1, 3] **and Y. Almirantis**[2]

Received August 8, 2000; accepted August 6, 2001

Clustering and long-range correlations in the nucleotide sequences of different categories of organisms are discussed. Clustering, mostly observed in higher eucaryotes, can be found at different length scales in DNA and Central Limit Theorems are used as links between these length scales. Several dynamical, statistical, mean-field models are proposed based on biologically motivated dynamical mechanisms and they successfully reproduce both the short range behavior observed in coding DNA and the long range, out-of-equilibrium features of non-coding DNA. Such dynamical mechanisms include aggregation of oligonucleotides, influx and DNA length reduction schemes, transpositions, and fusions of large DNA macromolecules. Fractality can be inferred from the short and long range correlations observed in the sequence structure of higher eucaryotes, where the non-coding part is relatively extended. In these organisms the DNA coding/non-coding alternation has the characteristics of finite size, fractal, random sets.

KEY WORDS: Power law distributions; long-range correlations; aggregation; coding/non-coding DNA sequences; fractal dimensions; Cantor sets.

1. INTRODUCTION

Statistical Dynamical approaches used first in the description of "large" complex systems[1–4] become now particularly fruitful in the description and understanding of the genome organization in living organisms. Living systems present several unique features:

1. Storage of large amount of information in a "digital" form (DNA primary structure) which may be transformed in a very versatile way to

[1] Institute of Physical Chemistry, National Research Center "Demokritos," 15310 Athens, Greece; e-mail: aprovata@mail.demokritos.gr

[2] Institute of Biology, National Research Center "Demokritos," 15310 Athens, Greece.

[3] To whom correspondence should be addressed.

functionally active forms: proteins and the subsequently formed cellular machinery.

2. Formation of an extended system of "indexing" of this large corpus of genomic information. This indexing system is the whole structure of gene interactions. The network of mutually controlled and regulated cascades of genes, becomes progressively known by the recent findings of Molecular Biology, and presents highly non-trivial features of a statistical dynamical nature. The significance of characteristics of gene networks such as stability, memory and learning ability becomes more and more obvious, thus progressively transforming Molecular Biology to a discipline equally based on experimental research and on a theoretical/statistical framework.

3. The coupling between the genomic information pool and the environmental pressure, through selection at the phenotype level, acts as an open system, constantly generating new information in evolutionary time. Statistical approaches are particularly fruitful both, for the description of relevant features of the evolving biological "text" written in the DNA sequences, and for the formulation of testable hypotheses concerning evolutionary critical mechanisms for the creation of new biological complexity/information.

Modern approaches to statistical mechanics and chaotic dynamics offer new tools for the description and comprehension of a variety of biological events. Such events include cell differentiation and development, responses of biological systems to complex external influences, electrical brain activity, circadian rythmes and even the spatial and temporal behavior of extended gene-regulation networks mentioned earlier.(1–4) Symbolic Dynamics, is also a powerful tool in the description of complex systems and combined with Statistical/Informational approaches offers new insights to the richness of information content encrypted by nature in DNA and protein sequences.(5)

Along these lines the current work first reviews and discusses some results related to the search for organization in the structure of DNA, and proposes several biologically plausible statistical dynamical mechanisms which can account for the formation and evolution of such organized structures.

The genomic structure of living organisms has been the subject of numerous studies in the recent years while its functions are still an intriguing question, vital in the understanding of the development and evolution of life.(6–8) The detailed nucleotide sequence of several test organisms is either already known or will be shortly available through the on-going genome projects. Instead, we are still far from understanding the mechanisms which govern the transition from this digital record to the complexity of living organisms.

The folded 3-D configuration of DNA, currently under intense investigation, is not in the direct scope of this article. Instead, we concentrate on the 1-D structure of DNA, which for humans reaches the 1.2 meters of length, or 3×10^9 nucleotides. More specifically we examine the way the individual nucleotides are placed in sequence to form clusters (1-d islands of similar nucleotides) which in turn form functional regions (coding and non-coding). Both the way the nucleotides and/or the functional regions are placed next to each other in a sequence, as well as the folding of the DNA helix and its combination with special proteins in order to form active chromosomes, are of major importance for the accomplishment of its biological role.

There have been many attempts to quantify the degree of randomness/organization in the structure of DNA.(5, 9–14) An elegant way to represent random DNA chains was introduced by Peng *et al.*(9) and was called the "DNA walk" model. This model is in reality a map which associates the DNA sequence to a "random" walk and thus, the degree of randomness in the nucleotide sequence is mirrored on the properties of the corresponding walk. On the same footing Li *et al.*(10) and Voss(11) calculated the 1/f spectrum of DNA sequences. As a result, these studies showed that while coding sequences have a non-correlated random nature, the non-coding spacers show long range correlations. It was also shown that these correlations are due to non-trivial clustering of similar nucleotides which is observed only in the non-coding regions of DNA.(14, 15)

In a series of recent publications(15–18) the current authors have shown, amongst other statistical properties, that the clustering of homologous nucleotides is linked to a higher level of organization, the level of coding/non-coding. For completeness the two levels of clustering are briefly reviewed in the sequel and the link between them is presented. The special features of these two organization levels and the links between them are important for the comprehension of genome organization and for the formulation of the dynamical evolutionary mechanisms presented in the next section.

Clustering at the Level of Nucleotides

The clustering of similar nucleotides is manifested by studying the properties of the Cluster Size Distribution, on various real DNA sequences, ranging from viral to higher eucaryotic sequences. The term "cluster of similar (homologous) nucleotides" means a string in a sequence containing only one type of nucleotides, e.g., only A's or only C's etc. In a more coarse grained manner the four nucleotides can be divided into two categories: the A and G nucleotides are the Purines (Pu) while the C and T are the Pyriminides (Py). This categorization reflects not only the similarity in their chemical structure but each category retains only the most primitive genomic characteristics: the most common "point mutations" (transitions) are the Pu → Pu or Py → Py transformations and not the so called "transversions." In the sequel, by the term "clusters of homologous nucleotides" we will refer to Pu clusters or Py clusters.

In recent studies[14, 15] it was shown that (a) in the non-coding regions of DNA the Cluster Size Distribution of homologous nucleotides follows a power law; (b) in the coding regions of DNA the Cluster Size Distribution follows a short range distribution. This non-trivial clustering of similar nucleotides is in the origin of the superdiffusive behavior observed in ref. 9 for DNA walks corresponding to non-coding sequences of higher eucaryotes.

The long power law tails are more clearly observed in the non-coding of higher eucaryotes. In lower eucaryotes (e.g., fungi), procaryotes, and viruses, it is difficult to observe the long tails in the size distributions, because the non-coding regions are relatively small and the tails are not always clearly manifested.

Organization at the Level of Coding/Non-Coding

The functional units in a DNA sequence (genes) often are formed by several coding segments interrupted by non-coding spacers. Furthermore, genes are separated by extended non-coding regions. To have a better understanding of this higher level of organization we have studied the properties of the Size Distribution of Coding and Non-coding DNA regions.[16] Our results, based on studies of organisms of various level of complexity, ranging from viruses to higher eucaryotes, may be summarized as follows: (a) the non-coding DNA regions (or spacers) follow a power law size distribution; and (b) the coding regions follow a short ranged, Gaussian or exponential type distribution.

The last comment of organization level I applies also here. For lower organisms, with small non-coding percentage, it is difficult to obtain a wide range of length scales where the power law behavior is expressed and short range behavior is sometimes observed.

Connection Between Levels I and II

In ref. 15 the coding DNA is regarded as 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 a DNA single strand and is approximately equal to $1/2$ (statistically same number of Pu and Py in DNA). It is well known and can be easily shown that the exponential distribution has finite mean and variance. Using the Central Limit Theorem, one thus expects that the probability distribution of the coding regions of DNA will be a Gaussian in the large size limit (see ref. 16). Notice that for large values of the distributed variable the Gaussian may be well approximated by an exponential distribution. Both Gaussian and exponential distributions are short ranged[40] and thus for large values of the distributed variable they fall equally fast.

On the other hand, the formulation of the Generalized Central Limit Theorem for the case of distributions with infinite variances (e.g., ref. 40) reminds us of the composition of the non-coding regions of DNA, especially in higher eucaryotes. In refs. 14 and 15 the non-coding DNA is regarded as a collection of 1-dimensional clusters of Pus and Pys whose size s_i follows a power law distributions of the form

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

for large values of the cluster sizes s_i. These are limiting cases of stable distributions as was stated in the Generalized Central Limit Theorem (see ref. 16). It is thus expected that the size distribution of the non-coding regions of DNA will fall in the basin of attraction of one of the stable distribution for large values of the distribution variable $S = s_1 + s_2 + \cdots + s_n$, with $n \to \infty$,

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

where the value of μ is equal in Eqs. (2) and (3). From studies of the genome of different organisms,[15, 16] it was shown that the value of μ varies between organisms of different complexity. Slight differences have been found even for different chromosomal regions within the same organism. In addition, the values of μ obtained from genomic data at the level I are usually larger than the corresponding values at the level II. This is an indication that the building of the non-coding is not just a simple juxtaposition of nucleotide clusters, but additional dynamical mechanisms must have acted during evolution, as will be analyzed in sections 2 and 3.

The difficulty in obtaining the correct power law behavior in the non-coding of lower organisms can be attributed to the limitations of the Generalized Central Limit Theorem. The correct power law behavior is obtained when the limit $n \to \infty$ is valid, which applies only to organisms with extensive non-coding regions, i.e., higher eucaryotes mainly.

In the current work we study biologically plausible dynamical mechanisms, originating from aggregation processes, which can lead to the observed statistical long range and short range characteristics, and to fractality. In the next section some simple dynamical mechanisms are introduced which may account for the observed statistical properties of real genomes. For all the considered mechanisms there is solid biological evidence that such events occur or have occurred during the evolutionary history of organisms. For clarity, Section 2 is divided into four subsections. We discuss: (1) the closed random aggregation mechanism; (2) the aggregation with influx; (3) outflux of different types; and (4) fusion/transposition mechanisms. For these mechanisms analytical and numerical results are presented and the different models are compared with statistical studies of data obtained from real biological sequences.

Section 3 is devoted to another aspect of DNA organization, fractality. We show how the fractality emerges from the different dynamical evolutionary mechanisms introduced in Section 2 and we compare real and model originated data. Recapitulation of the main conclusions and evolutionary hypotheses resulting from this study are proposed in the last section.

2. DYNAMICAL MECHANISMS

In this section we investigate the dynamical mechanisms which give rise to the complex genomic structures as described in Section 1. Since often genomic sequence are composed as juxtapositions of long range distributed non-coding segments with short ranged distributed coding segments, it is natural to search in the direction of dynamical mechanisms producing such patterns. From all the possible such mechanisms the ones which are relevant to this work must have a biological plausibility.

From the biological point of view, DNA evolution has been the subject of long and extensive studies and most evolutionary mechanisms which will be discussed in the sequel are now textbook knowledge.[19] Molecular Biology has revealed various genomic entities, such as repeats, transposable elements, and events like lateral gene transfer, replication of "selfish" DNA etc., which point to a variety of mechanisms active during evolution.

When describing the various mechanisms, a distinction must be made on the evolution of DNA regions with different functionality (coding versus non-coding). The coding part is much less tolerant to external perturbations of the sequence structure which is highly conserved during evolution ("quasi-closed system"). The coding regions may be described as "frozen states," which were formed in the remote past and they remain almost unchanged since. On the other hand, the non-coding DNA can suffer major modifications and can change considerably in evolutionary time, without major damage to the organism. The non-coding part may be considered as an "open system" in constant exchange with a genomic environment and, in the course of the time, it may reach a non-equilibrium steady state. Next to the coding regions and within the non-coding parts, one can find promoters, enhancers and several other cis-acting elements. These regulatory regions have known functional properties, but their structure is less conserved than that of pure coding regions. Their functionality lies on weaker prerequisites and only conservation of short "consensus sequences" and their relative distances is generally required. Thus these "next-to-coding," regulatory regions, are expected to behave as "moderately open system" with intermediate statistical characteristics.

In the sequel, several evolutionary mechanisms are described and their contribution to the formation of coding or non-coding sequences is discussed:

1. *The "aggregation" or "synthesis" mechanism:* to form primitive genomic sequences different oligomers or macromolecules mix and aggregate, under evolutionary constraints, to form larger and larger sequences. This process is appropriate both for the coding and the non-coding regions.[20]

2. *The "transposition" or "cut-and-paste" mechanism:* segments are constantly cut from one part of a sequence and are transposed onto another. Transpositions are frequent in genome dynamics and proceed through several mechanisms. Here we consider simple cut-and-paste, usually occurring from a donor site by double strand breaks at both ends of the "transposon" and a random jump to a target site.[21] Another usual alternative is via replication (see later). Transpositions are involved in several evolutionary events like the development of the vertebrate immune system[22] and the incidence of genetic diseases.[23] For further reading on transpositions see ref. 24.

3. *The "replication" or "copy-and-paste" mechanism:* a DNA segment is replicated and the replica randomly incorporates itself within the sequence. Particularly widespread are the retroposons (selfish DNA propagating via the inverse transcription of an RNA replica).[25,26] Among mobile segments of genetic material self-splicing introns (replicating at high rates) are of particular interest.[28] Cases of mechanisms switching from simple cut-and-paste to replication have also been reported.[29] Several forms of repetitive DNA originate from consequitive replication events.[23,26]

4. *The "influx" mechanism:* DNA segments of external origin incorporate themselves in the sequence.[30,31] Due to influx the sequence size

increases. Usual cases are insertions of parasite DNA, such as viral DNA or viroids.[32] Note that the replication mechanism, which mostly takes place in non-coding DNA, is kind of influx, since it causes increase of the sequence size. Replication and insertion events, when occurring between different chromosomes maybe considered as influx because the size of the target chromosome increases. “Infection” of an organism with genomic material of another organism may also occur. When coding regions are transported within the incoming segment one may speak of “lateral gene transfer.” The sequencing of complete genomes has contributed substantial evidence that lateral gene transfer and thus extended exchange of genetic material between organisms has been frequent in the evolutionary past,[33] especially in procaryotes[34] but also from procaryotes to eucaryotes.[35] It is also conjectured that lateral gene transfer has occurred between procaryotes and the ancestor of the present day eucaryotes.[36]

5. *The “outflux” mechanism:* there are known cases of organisms that have lost large parts of their DNA. The ancestor of the present day procaryotes is considered to have had extended non-coding regions, eliminated by yet unspecified genomic outflux events.[20] Similarly, the pupperfish[37] has undergone genome compactification by loosing a great part of its non-coding DNA, still present in its close relatives. Precise loss of introns may occur as a low probability event of random deletion of the intronic sequence or by several recombination assisted events, like interaction between an intact gene and a processed pseudogene or reverse-transcribed cDNA.[38]

Obviously the list of biological evolutionary mechanisms proposed above is neither exhaustive nor complete. Moreover, all these mechanisms are not simultaneously active, neither in time nor in space (locally onto the sequence). For example the “influx” and “cut-and-paste” mechanisms are rather usual in the non-coding parts of the DNA but they are proved mostly fatal for the coding parts. A parasite macromolecule intruding in a non-coding region can remain “silent” there and even propagate to the descendants through several forms of reproduction. An intrusion in a coding region, most probably interrupts the production of a protein and such a loss can be lethal for the cell. There are rare cases where such an intrusion does not damage or even improves the protein producing region, and the modified cell might survive and propagate.

In particular, the question of repeats, their origin and reason of appearance, has been extensively studied in the literature.[23, 26] Repetitions of nucleotide strings of sizes ranging from a few to several hundred base pairs (bps), are frequently found in non-coding DNA. In the current study repeats can be considered as results of an influx mechanism, resulting usually from multiple replication. Thus the presence of repeats not only does not destroy the long range correlations found in non-coding DNA, but, on the contrary, it may enhance them. As will be explained in Section 2.2, the influx of external macromolecules and the proliferation of repeats, are essential for the development of long range correlations in the non-coding.

It seems rather premature to currently construct mechanisms with so fine details that can really account in detail for the statistical structure of the present day genomes. Minimal models containing simplified versions of the above mechanisms can capture the most important statistical characteristics of today's DNA. Such mechanisms were first introduced in ref. 17. These are mean field minimal models which involve directly the aggregation and the influx mechanism. The "cut-and-paste" mechanism may be incorporated as a special case of aggregation while the replication mechanism can be considered as influx (replica additions which increase the chain lengths). In ref. 18, genome fusion and transposition mechanisms are considered. They may account for scale dependent non-randomness in DNA sequences. Our recent findings indicate that these mechanisms, having a mixing nature, also produce long range features. We present here four minimal models: each of them includes some of the dynamical steps described above and aims to the understanding of the statistical attributes of concrete genomic structures. These models can reproduce some of the statistical long range and short range features found in the coding and/or non-coding size distributions as described in the introductory section.

2.1. The "Closed Aggregation" Model

As described in the previous sections, the statistical nature of the coding and the non-coding regions are significantly different and thus the mechanisms which have acted during their formation and evolution must also be different. For example it is known that coding parts are highly conserved through evolution; if a slight change occurs in the coding the cell most probably dies. Thus influx events interrupting coding segments are highly improbable.

A coding segment is thus created at a certain evolutionary stage and is conserved (with minor changes) ever after. A reasonable scenario to describe the coding segment evolution, compatible with a simple statistical treatment, is a model based on random aggregation of nucleotide oligomers or even larger macromolecules. Oligomers or larger molecules of various sizes have aggregated at random in the evolutionary past and when a functional segment is created its structure does not change further and can be considered as steady state. [Rare modifications/mutations may occur but their statistical influence on the steady state are considered negligible]. A cascade of aggregating events finally creates a DNA sequence which codes for the functionally active aminoacid sequences. For this scenario the most obvious assumption about the size distribution of the aggregating segments (oligomers or larger molecules) would be any short ranged distribution. Typical such distribution is the Gaussian which takes the form

$$P(s) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(s-\langle s\rangle)^2/2\sigma^2}, \tag{4}$$

where $\langle s\rangle$ is the average size of the aggregating oligomers and σ is the mean square deviation. A coding segment is then considered as a collection of a large number of these aggregating segments. According to the Central Limit Theorem the sum of a large number of variables each of which follows any short ranged distribution (including the Gaussian distribution) will follow a Gaussian distribution. In addition, the Gaussian is the only short ranged stable distribution endowed with the property that the sum of a large number $N \to \infty$ of variables following the Gaussian distribution (4), will also follow a Gaussian distribution (see Appendix I). As a consequence of this property one can immediately calculate the probability $P(S)$ to find a coding segment of size S as:

$$P(S) = \frac{1}{\Sigma\sqrt{2\pi}} \mathrm{e}^{-(S-\langle S\rangle)^2/2\Sigma^2}, \tag{5}$$

where this new distribution has as average value $\langle S\rangle = N\langle s\rangle$ and as mean square deviation $\Sigma^2 = N\sigma^2$.

In the case of aggregation of oligomers to create the coding segments, not only the size s of the oligomers varies, but also the number N of oligomers involved in the formation of the coding may vary from one coding region to another. It is also plausible to assume that the distribution $\mathscr{P}(N)$ of the number of oligomers N follows also a short range, Gaussian like distribution, centered around an average value $\langle N\rangle$. In this case the probability to find a coding part of size S is given by the probability to find N oligomers $\mathscr{P}(N)$ times the product of the probabilities that the ith oligomer has size s_i, $P(s_i)$, under the condition that the sum of the s_i's will be equal to S.

$$P(S) = \sum_{N=1}^{\infty} \mathscr{P}(N) \prod_{j=1}^{N} P(s_j)|_{\Sigma_{j=1}^{N} s_j = S} \tag{6}$$

The form of $\mathscr{P}(N)$ may be assumed Gaussian

$$\mathscr{P}(N) = \frac{1}{\Sigma_N \sqrt{2\pi}} \, \mathrm{e}^{-(N-\langle N\rangle)^2/2\Sigma_N^2}, \tag{7}$$

where Σ_N^2 is the mean square deviation of the number distribution of oligomers. Eq. (6), with the number distribution Eq. (7) is solved in Appendix I for large values of S. The probability distribution approaches the Gaussian in the large S limit

$$P(S) = \frac{1}{\Sigma_{s/N} \sqrt{2\pi}} \, \mathrm{e}^{-(S-\langle N\rangle\langle s\rangle)^2/2\Sigma_{s/N}^2}, \tag{8}$$

where

$$\Sigma_{s/N}^2 = \langle N\rangle \, \sigma^2 + \langle s\rangle^2 \, \Sigma_N^2 \tag{9}$$

In Eq. (9) $\Sigma_{s/N}$ plays the role of combined fluctuations when the size and number distribution of the merging oligomers are varied simultaneously. Since the probability distribution $P(S)$ approaches the Gaussian for large values of S, $P(S)$ is obviously short ranged.

This is indeed the case of the coding segments of all organisms. Their size distributions in the genome are short ranged, Gaussian like. This finding holds both for procaryotic genomes where every coding segment codes for a protein chain and for higher organisms where most coding segments (called "exons") code only for small parts of proteins.[20] To test this theoretical conclusion we have examined many genomic sequences originating from different organisms. All the sequences presented in this section and throughout this work were chosen under the following four criteria: (a) the sequences must be fully and reliably annotated; (b) for statistical reasons the sequences must be as long as possible; (c) complete genomes or chromosomes whenever possible; and (d) must contain a relatively large number of coding and non-coding regions. These four criteria are highly important especially in the case of open aggregation models (aggregation with influx and outflux) which model the statistical dynamics of the non-coding. For the case of the closed aggregation models shorter sequences give also good results.

In Fig. 1 is plotted the cumulative size distribution $\tilde{P}(S)$ of size greater than S for a coding DNA sequence for *A. thaliana* BAC TM021B04 clone, (ATAF7271, EMBL, 90.0 kbps, 35,1% coding) and *C. elegans* sequence (Genbank, chromosome I, 16183 kbps, 19.7% coding) and *E. coli* complete sequence (4639 kbps 89% coding). All sequences presented in the current work are obtained either from EMBL or from GenBank. The cumulative size distributions $\tilde{P}(S)$ have been all normalized to $\tilde{P}(S=1)=1$ for comparison between the different sequences. The results in Fig. 1 are plotted in double logarithmic scale mainly for comparison with Figs. 2, 3, and 4. The straight line represents a power law with exponent $-\mu=-2$ which denotes the border lines between short and long range behavior (for the cumulative distribution). Observation of the results in Fig. 1 leads to the following conclusions: (a) the presented data comes from different categories of organisms but only the coding cumulative size distributions are shown; (b) the cumulative distribution of the coding parts drops very abruptly for all sequences; (c) the cumulative forms are sigmoid, almost step-like functions. The sigmoid form of all curves indicate that the original (non-cumulative) distributions $P(S)$ are short ranged, Gaussian-like distributions. Note that the cumulative of the Gaussian is the complementary Error function *erfc* which has a sigmoid form (see Appendix II). The sigmoid form in the cumulative size distribution of the coding is found in all categories of organisms and it agrees with the theoretical predictions of the "closed aggregation" model.

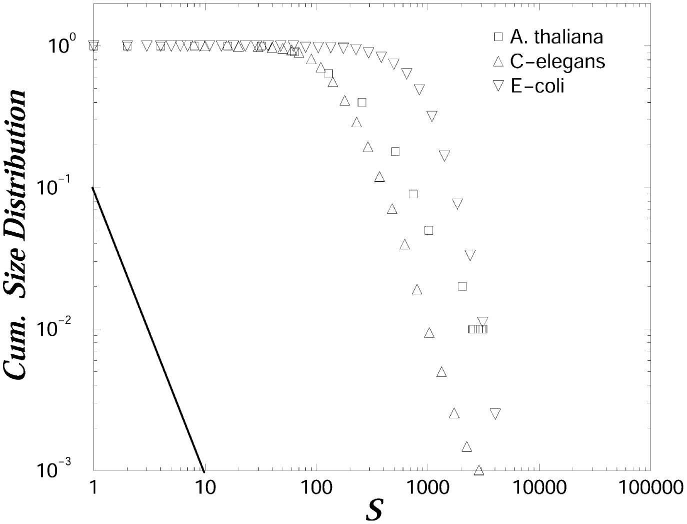

Fig. 1. The cumulative size distribution $\tilde{P}(S)$ of coding regions of size $\geqslant S$, with data from three different organisms. All distributions decay abruptly following short range laws. The straight line represents a power law with exponent $-\mu=-2$.

2.2. The "Aggregation with Influx" Model

The requirement of conservation of the genomic structure does not hold for the non-coding DNA where it is known that "parasite macromolecules" are frequently incorporated together with replication events which cause the gradual increase of the non-coding size. In general, all the mechanisms which cause the increase of the size of the non-coding may be considered as influx mechanisms.(26, 28, 31) For example, the influx of parasite macromolecules and the replication/transposition mechanism may have different biological origin and implications, but their final effect on the size distribution of the non-coding segments is the same: both increase the size of the non-coding.

"Repeats," which are mostly met in the noncoding, may also be considered as influx of many replicas of the same segment in a given sequence.(23, 26, 27) What is important for the current study, is the size distribution of the repeats (or more generally of the incoming segments) and not their detailed structure. Thus, the presence of repeats enhances further the idea of considering the non-coding sequences as open systems in dynamical equilibrium with their environment.

The notion of the characteristics of the influx and outflux size distributions is enough for the study of the characteristics of the non-coding undertaken in Sections 2.2 and 2.3. For a further detailed study of the Pu/Py underlying structure of the non-coding, the specific Pu/Py concentration must be explicitly taken into account.(15, 18) This is most important when repeats are involved because they import specific motifs. Influx of segments with detailed structures is undertaken in Section 2.4, where the specific on the Pu/Py concentrations/structures of the incoming macromolecules (fusing segments) is a prerequisite of the model.

To study the evolution of a given number of non-coding macromolecules in constant exchange amongst them and with incorporation of incoming segments, an explicit assumption needs to be made about the size distribution of the influx $P_{\text{in}}(I)$, where I is the size of an incoming "parasite" macromolecule or the "replica element" (repeat) added to the sequence. Again a normal assumption for $P_{\text{in}}(I)$ would be a Gaussian, or any short ranged distribution.

We start from the general formula describing aggregation of n macromolecules of sizes s_j, $j = 1,\ldots n$ and influx of macromolecules of size I. The evolution of the probability distribution $P(S, t)$, i.e., the probability to find a non-coding region of size S at time $t + \Delta t$ is given by

$$P(S, t+\Delta t) \\ = P_{\text{in}}(I) \sum_{n=1}^{N} \binom{N}{n} \left(\frac{1}{N}\right)^{n} \left(1 - \frac{1}{N}\right)^{N-n} \prod_{j=1}^{n} P(s_j, t)\Big|_{\sum_{j=1}^{n} s_j + I = S} \tag{10}$$

Formula (10), used by Takayasu *et al.*[41] is a generic formula for open aggregation of N objects with input whose size distribution is $P_{\text{in}}(I)$ (I is the size of the incoming object). The meaning of Eq. (10) is the following: from a pool of N macromolecules n of them meet and aggregate at time t and together with a potential influx of size I they form a larger macromolecule of size $\sum_{j=1}^{n} s_j + I = S$. It is reasonable to seek the steady state behavior $P(S, t+\Delta t) = P(S, t) = P(S)$. Taking the Fourier transform $Z(\rho) = \int e^{-i\rho S} P(S)\, dS$ of Eq. (10) and for $N \to \infty$ we obtain a closed equation for the Fourier characteristic function $Z(\rho)$:

$$Z(\rho) = e^{Z(\rho)-1} \Phi(\rho), \tag{11}$$

where $\Phi(\rho) = \int P_{\text{in}}(I)\, e^{-i\rho I}\, dI$ is the Fourier transform of the influx distribution. For the influx distribution we may consider the following two cases:

1. With the assumptions of $P_{\text{in}}(I)$ being short-ranged (Gaussian, Exponential, etc.) one can show that $\Phi(\rho) = \int \mathrm{e}^{-i\rho I} P_{\text{in}}(I)\, dI = 1 - c\, |\rho|^{1/2}$, where $\rho \ll 1$ and c is a constant. By substituting the explicit form $\Phi(\rho)$ in Eq. (11) and taking the inverse Fourier transform the resulting distribution of the non-coding segments in a DNA chain takes the form

$$P(S) \sim S^{-3/2}, \qquad \text{for} \quad S \gg \langle I \rangle. \tag{12}$$

2. If the distribution of influx DNA has long tails, for example $P_{\text{in}}(I) \sim I^{-1-\beta}$, for $0 \leqslant \beta \leqslant 2$ then $\Phi(\rho) = 1 - c\, |\rho|^{\beta}$ and the form of the non-coding size distribution becomes

$$P(S) \sim S^{-1-\beta/2}, \qquad \text{for} \quad S \gg 1. \tag{13}$$

A power law tail of the size distribution of the influx may modify the exponent of the non-coding size distribution from $-3/2$ in the case of random non-correlated influx to the value $-1-\beta/2$.

In Fig. 2 we present the cumulative size distribution $\tilde{P}(S)$ of non-coding sequences of: human, chromosome 16, BAC clone CIT987sk-334D11, (Genbank, HSAF001550, complete clone sequence, 173.9 kbps, 2.1% coding); *A. thaliana* BAC TM021BO4 clone (ATAF7271, EMBL, 90.0 kbps, 35.1% coding); *C. elegans* sequence (CELD1007, Genbank, cosmid D1007, complete sequence, 47.6 kbps, 32.5% coding); and the Adh region of *Drosophila melanogaster* (2918.5 kbps, 17.5% coding[42]). Here, it is particularly important to observe strictly the four criteria announced earlier (see Section 2.1) for the selection of the examined sequences. Especially, the length of the sequences must be as large as possible for the observation of

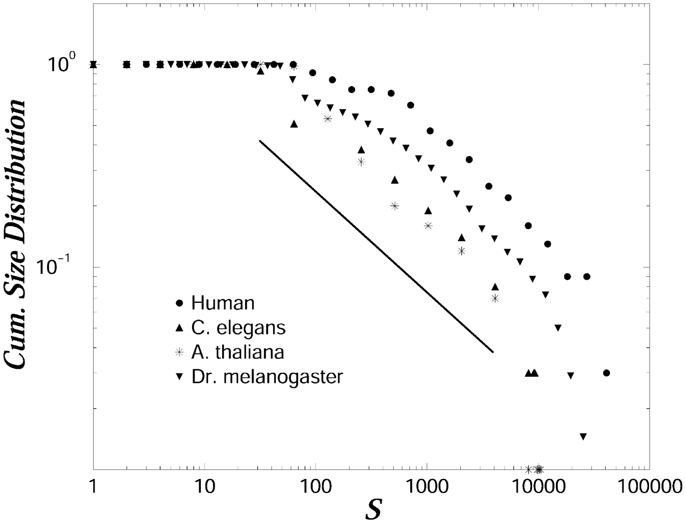

Fig. 2. The cumulative size distribution $\tilde{P}(S)$ of non-coding regions of size $\geqslant S$, with data from four different higher organisms. The solid line, in a double logarithmic scale, represents a power law region with exponent equal $-\mu = -1/2$.

the long power law tails. The straight line corresponds to a power law with exponent $-\mu = -1/2$. Note that if the original size distribution follows a power law with exponent $(-1-\mu)$ then the cumulative distribution follows a power law with exponent $-\mu$ (see Appendix II). Eventual deviation from this value may be attributed to "anomalies" of power law type in the influx distribution. Another source of the deviation of the exponent $(-1-\mu)$ from the value $-3/2$ could be dynamical irregularities in the rate of influx of the parasite macromolecules in biological time scales.

As mentioned at the beginning of the current section, repeats represent a special case of influx met mostly in the non-coding of higher eucaryotes and are particularly frequent in human genome. To check the influence of the repeats in the formation of the long range correlations of the non-coding we have examined several human sequences, with sizes ranging from 37 kbps up to 3500 kbps and with repeat percentage ranging from 3% up to 44% of the total sequence length. We have used the RepeatMasker tool,[43] developed at the University of Washington to mask all known repeats. The size distributions of non-coding parts of the masked sequences were examined and the results have shown that the power law behavior is indeed robust under the removal of repeats. More specifically, for the most drastic repeat removal (covering the 44% of the sequence) a change of the value of the exponent μ of the order of only 5% was observed and the exponent μ remained well inside the range of long range behavior ($0 \leqslant \mu \leqslant 2$). This robustness of the exponent is indeed striking since the data was reduced to almost half of its original size. For sequences containing smaller percentage of repeats the exponent μ remained almost identical under repeat removal. The robustness of the distribution is expected since the power law behavior is observed not only in higher eucaryotes but also in nematode, fungal and some bacterial genomes,(16) where repeats are clearly less abundant. At the beginning of the current section we have described several mechanisms which may account for the increase of the non-repeated, non-coding part of DNA and which can be responsible for the persistence of the power laws.

2.3. The "Aggregation with Influx and Outflux" Model

Lower organisms have relatively short non-coding DNA and thus only occasionally long range correlations are observed in these organisms. As discussed earlier, evidence from Molecular Biology of evolution suggests that the ancestors of these organisms had also extended genome which included large non-coding spacers but was reduced later.(20) An outflux mechanism seem to take place rarely in higher eucaryotes, see discussion of refs. 17 and 37. Indeed, such a mechanism may sometimes reduce the power law distribution of the non-coding into a short-ranged distribution. Starting again from Eq. (10) let us consider the following outflux mechanisms:

2.3.1. The Relative Outflux

This is an outflux mechanism such that: each time the n macromolecules (plus the influx I) join together to form a larger chain of size $S = \sum_{j=1}^{n} s_j + I$, some percentage λS of the chain dissociates. Then the necessary summation condition in Eq. (10) is:

$$\sum_{j=1}^{n} s_j + I = S + \lambda S = (1+\lambda)\, S \tag{14}$$

Using a short ranged influx size distribution, one can show that Eq. (11) under condition (14) is now reduced to

$$Z(\rho) = e^{Z(\frac{\rho}{1+\lambda})-1} \Phi\left(\frac{\rho}{1+\lambda}\right). \tag{15}$$

In lowest order of ρ, (thus large values of S) Eq. (15) reduces to

$$Z(\rho) = 1 - i\,\frac{\langle I\rangle}{\lambda}\,|\rho| - \mathcal{O}(\rho^2). \tag{16}$$

The original size distribution $P(S)$ maybe found by taking the inverse Fourier transform of Eq. (16) which, at this lowest order, indicates a *δ-function* centered around the mean value

$$\langle S\rangle = \langle I\rangle/\lambda \tag{17}$$

If we also keep the second order term in Eq. (16) then the size distribution will become a "fat"-delta, Gaussian distribution. This simplistic "outflux" mechanism reduces the power law distribution obtained earlier into a short range law which posses a well defined mean value given by Eq. (17). This reduction mechanism may be one of the reasons for not observing frequently power laws in the non-coding of lower organisms. This situation is shown in Fig. 3, where the cumulative non-coding size distribution is plotted for procaryote *Bacillus Subtillis* (BSUB0010, EMBL, 233.8 kbps = 5% of the complete genome, 89.3% coding), procaryote Synechocystis sp. (SYCSLLLH, Genbank, 132.1 kbps = 4% of complete genome, 88.6% coding) and Nuclear Polyedrosis virus (OPU75930, EMBL, 132.0 kbps,

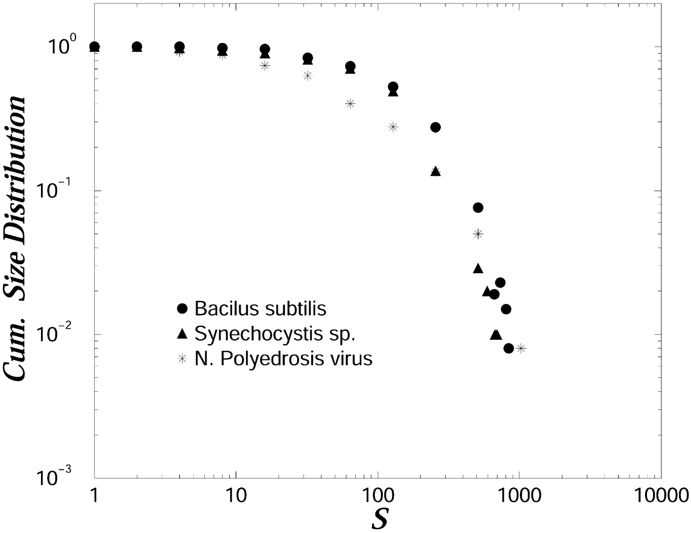

Fig. 3. The cumulative size distribution $\tilde{P}(S)$ of non-coding regions of size $\geqslant S$, with data from three different lower organisms with relatively short non-coding DNA.

complete genome, 93% coding). These organisms have relatively small non-coding and it is possible that reduction mechanisms have acted during evolution and have driven the non-coding to a state described by a short range distribution.

2.3.2. The Constant Outflux

This mechanism although at first site is a very simple variation of the relative outflux, the resulting non-coding size distribution is intrinsically different. Consider that the outflux takes constant values, independent of the total size of the non-coding. Consider that the size of the outcoming macromolecules $P_{\text{out}}(C)$ follows short ranged distributions with finite mean value $\langle C \rangle$ and finite variance. Then Eq. (10) reduces to

$$P(S) = P_{\text{in}}(I)\, P_{\text{out}}(C) \sum_{n=1}^{N} \binom{N}{n} \left(\frac{1}{N}\right)^{n} \left(1 - \frac{1}{N}\right)^{N-n} \prod_{j=1}^{n} P(s_j, t) \tag{18}$$

while the necessary summation condition is:

$$\sum_{j=1}^{n} s_j + I = S + C \tag{19}$$

The Fourier transform of Eq. (18) is

$$Z(\rho) = \Phi(\rho)\, \mathscr{F}(\rho)\, e^{-1+Z(\rho)} \tag{20}$$

where $\mathscr{F}(\rho) = 1 - i\langle C \rangle \rho + \mathscr{O}(\rho^2)$ is the Fourier transform for the outflux distribution. The solution for small values of ρ is

$$Z(\rho) = 1 - i^{1/2}(\langle I \rangle - \langle C \rangle)^{1/2}\, |\rho|^{1/2} - i\theta\rho, \tag{21}$$

where θ is a constant, dependent on the second moments of the influx and outflux distributions. Since both $P_{\text{in}}(I)$ and $P_{\text{out}}(C)$ have by assumption finite averages and variances, θ takes a finite value. Taking the inverse Fourier transform of Eq. (21) and keeping the lowest order of $|\rho|^{1/2}$, one obtains a power law distribution similar to the one found in the only-influx case.

$$P(S) \sim S^{-3/2}, \qquad \text{for} \quad S \gg (\langle I \rangle - \langle C \rangle). \tag{22}$$

Indeed, in many cases of lower organisms, where outflux mechanisms could have played an important role, power laws are observed in the non-coding distribution, Fig. 4. The cumulative non-coding size distribution is shown here for sk1 bacteriophage (AF011378, 28.5 kbps, complete genome, 93% coding), *S. cerevisiae*, (Chromosome I, complete left arm genome, 103.7 kbps, 66% coding) and the procaryote *H. influenza*, (complete genome, 1830.1 kpps, 87% coding) are presented. The dotted line, plotted for comparison, has slope $-1/2$. The *S. cerevisiae* and the sk1 bacteriophage distributions present exponents $-\mu=-0.8$ and -0.65 respectively, and approach relatively well the mean field prediction (dotted line), while the distribution of *H. influenza* still follows a power law but with smaller exponent (compares well with -1). As a general conclusion, the size distributions of non-coding spacers in the genomes of lower organisms frequently follow power laws even though their total non-coding length is relatively small. The proposed mean field model shows a simple mechanism of an open system with influx and ouflux which still produces power laws. Note that the case $\langle I\rangle<\langle C\rangle$ may not be observed; when the outflux is larger than the influx the total size reduces to 0 at the steady state.

2.3.3. The Constant Outflux with $\langle I\rangle=\langle C\rangle$

Of marginal importance is the case where the average influx and outflux are the same, so that on the average the size of the macromolecule remains unchanged. This would be of importance in cases where the size of the DNA needs to be conserved: when the genome size exceeds some threshold of tolerance, due to massive external intrusions, evolution can favor outflux resulting in approximately constant genome size. This type of "influx–outflux" can be understood also as a "birth-and-death" process. When the average influx and outflux rates have equal mean value $\langle I\rangle=\langle C\rangle$, the lowest order term in $|\rho|^{1/2}$ vanishes and the next term is of order ρ. Eq. (21) now reduces into

$$Z(\rho)=1-i\theta\,|\rho|, \tag{23}$$

and from the inverse Fourier transform one obtains

$$P(S)\sim S^{-2}. \tag{24}$$

This is also a power law but with a more abrupt decay for large values of S. This behavior is intermediate between the slow Gaussian decay of the "relative outflux" mechanism and the steep power law in the case of the constant outflux "($\langle I\rangle>\langle C\rangle$)."

In Fig. 4 the cumulative size distribution of non-coding sequences of size S, $\tilde{P}(S)$, for *H. influenza*, follows a power law with exponent approximately equal to $-\mu=-1$, as predicted by the current mechanism. Certainly, there is no biological evidence that this precise influx-outflux mechanism is responsible for the genomic structure of this organism.

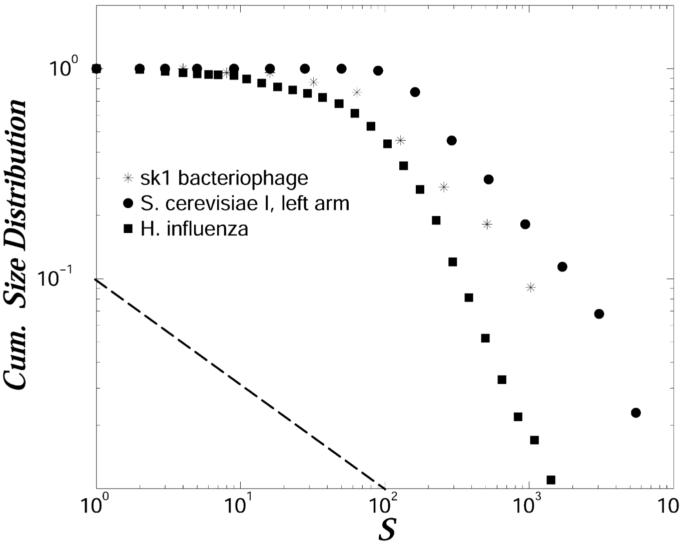

Fig. 4. The cumulative size distribution $\tilde{P}(S)$ of non-coding regions of size $\geqslant S$, with data from three different lower organisms. For comparison, the dashed line represents a power law with exponent equal $-\mu = -1/2$.

Instead, this is an example of the rich structural genomic behavior and of the constructive dynamical mechanisms which can give rise to such structures. In many cases more than two exponents may appear during the study of DNA sequences. It is possible that the different influx and outflux mechanisms may act simultaneously in different scales of sizes and thus two different exponents maybe observed. The different evolutionary time scales during which the influx and outflux mechanisms have acted can also affect the non-equilibrium steady states presented in Eqs. (17), (22) and (24). Additional power law influx and outflux processes may modify the exponents themselves, as we have seen in Eq. (13). The resulting distribution will strongly depend on the choice of the participating mechanisms.

2.4. The "Fusion/Transposition" Model

The minimal dynamical models presented thus far explain the long range distribution at the level of coding/non-coding regions. In the current version they do not address explicitly the clustering at the level of nucleotides. However, mechanisms involving combination of long sequences with different constitutions, followed by transposition events causing mixing at the nucleotide scale may account for the long range behavior in the clustering of homologous nucleotides.

In particular, end-to-end genomic fusion events are known to have taken place in early evolutionary times, between long molecular DNA and/or RNA chains.[36, 39] The term "fusion" is used here to denote the mechanism by which two (or more) macromolecules meet and join end-to-end to form a larger macromolecule. The terms "fusion" and aggregation are used in a similar general context; however, when we use the term "fusion" we consider combining, relatively large macromolecules which have underlying structures of Pus and Pys, which mix via the transposition events. In the previous sections, we mainly considered "aggregating" macromolecules without underlying structure, the aggregating macromolecules were either just coding or non-coding.

When two or more of macromolecules fuse end-to-end, the resulting long chain has in general different local Pu/Py concentrations. This may apply for single and double stranded macromolecules. There is considerable evidence that continuous shuffling of the genome occurs due to transposition events, see ref. 24. This mixing leads towards a restoration of homogenization in the local base constitution. However, before the final, totally homogenized steady state, intermediate stages presenting high patchiness and clustering of homologous nucleotides are formed.

It is also possible to assume that, during evolutionary time, additional influx events of large macromolecules occasionally take place, and the new incoming/fusing macromolecules have differences in their Pu/Py concentrations. Continuous transposition events would then start dissolving them gradually, creating new patchy structures, while new incomers keep the system in a non-equilibrium steady state. The feeding and dissolution of external macromolecules helps the system conserve in time the clustered, at several length scales, character of the non-equilibrium steady state.

One must remind here, that incoming events principally take place in the non-coding parts of the DNA. The coding parts do not present the clustered structure of the non-coding. Instead, they are locally at statistical equilibrium, cf. paragraph 2.1. However, a large scale patchiness is visible in lengthy coding sequences.[9] This seems to relate to the rare cut-and-paste and fusion events which have occurred during the slow-time evolution of the coding.

The "fusion/transposition" model we propose can be described in the following two versions:

2.4.1. The "Equilibrium Fusion/Transposition" Model

1. Start with two macromolecules S_1 and S_2 of similar length L, containing Pus with probability frequencies p_1 and p_2 and Pys with frequencies $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$, respectively. Let us assume that $p_1 < p_2$.

2. To mimic the event of fusion, the two macromolecules are placed end-to-end to form a larger macromolecule S of size $2L$.

3. A random position on the macromolecule S is chosen, one segment of length l is cut and is inserted, in the forward or inverse orientation, into the randomly chosen position on S.

4. Step 3 is repeated many times.

5. Calculate occasionally the cluster size distribution of Pus and Pys.

This model is “closed” (only one fusion event takes place at the beginning). For intermediate times patchy structures can be created, while for longer times the model predicts an homogeneous structure with randomly distributed Pus and Pys.

Following the above algorithm a numerical experiment is performed starting with pure initial states. The initial nucleotide probabilities were $p_1 = 1$, $q_1 = 0$ for the macromolecule S_1 (pure Purines) and $p_2 = 0$, $q_2 = 1$ for the macromolecule S_2 (pure Pyrimidines). The initial macromolecules had equal size $S_1 = S_2 = L = 25000$ units (or base pairs). After the end-to-end fusion, a large number of transposition events take place, while the sizes of the transposed regions are randomly selected in the range 5–50.

In Fig. 5 the numerical results of the closed “fusion/transposition” model are presented. The cumulative cluster size distribution of the Pus are plotted as a function of the cluster size s, after 400 transposition events. For such intermediate times, the cumulative size distribution follows a power law with exponent $-\mu = -1.12$ (solid line). Thus the ordinary cluster size distribution follows a power law with exponent $-\mu - 1 = -2.12$. This power law behavior persists in the intermediate time scales. Similar exponents were found in ref. 15 for the cluster size distribution of Pu and Py in the non-coding of higher eucaryotes. As $t \to \infty$ the system homogenizes and the cluster size distribution takes a Gaussian-like form. Equivalently, the cumulative distribution takes an abrupt sigmoid form (see Appendix II), almost a step like decay (dashed line).

2.4.2. The “Out-of-Equilibrium Fusion/Transposition” Model

A more complete algorithm would also include a sixth step to keep the process out-of-equilibrium and to account for occasional fusions of “incoming macromolecules” with variability in the concentrations of Pu and Py:

6. Occasionally, long macromolecules with high Pu density, (or high Py density) are incorporated in a random position of the system and the algorithm returns to step 3 to continue with transposition events.

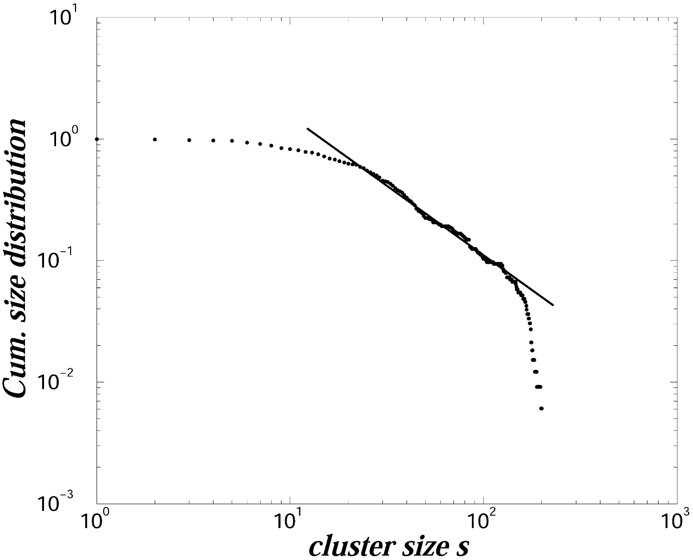

Fig. 5. The cumulative size distribution of Pus is plotted as a function of the cluster size s for the "equilibrium fusion/transposition" model. The solid line represents a power law with exponent $-\mu = -1.12$.

Two discrete time scales are thus introduced: the slow scale for fusion events and the fast time scale for transposition events. This open fusion/transposition model is expected to produce the non-equilibrium steady state observed in the Pu/Py size distributions in the non-coding.

For this numerical experiment one macromolecule was first created by fusion of two macromolecules of equal size $L = 400000$. The initial nucleotide densities were ($p_1 = 0.96$, $q_1 = 0.04$) for the macromolecule S_1 (almost pure Purines) and ($p_2 = 0.5$, $q_2 = 0.5$) for the macromolecule S_2 (homogeneous). The resulting macromolecule S underwent 2500 random transposition events with transposition segment length randomly chosen in the range 50–200. An additional macromolecule S' was created by fusion of two macromolecules S'_1 and S'_2. These two macromolecules had the same size as before $L = 400000$ and initial nucleotide densities ($p'_1 = 0.04$, $q'_1 = 0.96$ for the macromolecule S'_1 (almost pure Pyrimidines) and $p'_2 = 0.5$, $q'_2 = 0.5$ for the macromolecule S'_2 (homogeneous). S'_1 and S'_2 were combined end-to-end and were mixed via 2500 random transposition events. Note that the fusing macromolecules S_2 and S'_2 may also be large repeats of the same segment. Finally the macromolecules S and S' were fused together and an additional 25000 random transposition events took place.

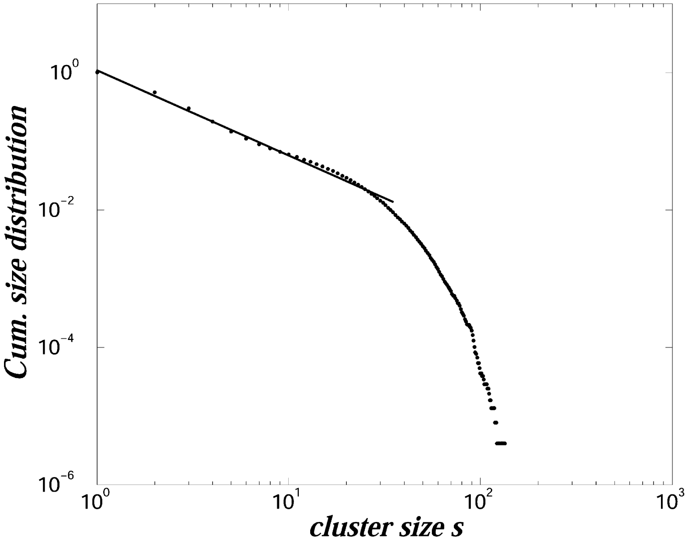

Fig. 6. The cumulative size distribution of Pus is plotted as a function of the cluster size s for the "out-of-equilibrium fusion/transposition" model. The solid line represents a power law with exponent $-\mu = -1.22$.

In Fig. 6 the results of this numerical experiment are presented. The cumulative cluster size distribution of the Pus are plotted as a function of the cluster size s. In a double logarithmic scale a power exponent is observed $-\mu = -1.22$. Accordingly, the original cluster size distribution follows a power law with exponent $-1-\mu = -2.22$. In Fig. 6, due to the large number of transposition events, the power law behavior was pushed down to the lower length scales. These results are thus in accord with the statistical properties of real genomic sequences.[15] In addition, the competition during evolution, between a continuous influx of large macromolecules (diversification factor) and the mixing caused by transposition (homogenization factor) will keep "alive"the non-equilibrium power law steady state.

A partial description of this model is also possible via Eq. (10). If one regards an aggregation procedure backwards in time, the large segments dissociate in smaller ones, much like the effect of a Pu segment of size l_1 cut from a Pu-rich region of size $l_2 > l_1$ and incorporated inside a Py region. This effect is a break of large Pu island of size l_2 into two smaller parts, one of size l_1 and one of size $l_2 - l_1$. Influx also takes place at times $t-1$. It is thus enough, in Eq. (10) to replace the time t on the left hand side with $t-1$ and the time $t-1$ on the right hand side with t. The summation condition will now become $\sum_j^n s_j = S - I$. It is also reasonable to assume $P(I)$ being Gaussian like, or any short range distribution. The steady state equation for this inverse model is exactly the same as in the direct aggregation model, since $P(s,t) = P(s,t-1) = P(s)$. Thus the solution is also given by Eq. (12), indicating a power law with exponent $-3/2$, exactly the same as the one obtained in direct aggregation. This intuitive argument predicts that (a) the distribution of the Pus and Pys in the non-coding follow a power law as also seen in real DNA sequences and in the numerical simulations obtained in Figs. 5 and 6 and (b) the exponent of the power law is $-3/2$ as seen also for the distribution of coding/non-coding. In earlier works,(14, 15) it was observed that the Pu and Py cluster size distribution of non-coding sequences in higher eucaryotes follows a power law with exponent smaller than $-3/2$. As an example values as small as -2.4 and -2.6 have been observed.(15) Thus the simulation results agree with the statistical results obtained from real sequences. The mean-field argument predicts qualitatively the same behavior but the power law exponent is slightly larger. It is interesting to note here, that the mean field exponent $-3/2$ can also be recovered using the Generalized Central Limit Theorem, which connects the small scale clustering (Pu and Py clusters) with large scale characteristics (coding/non-coding character).(16)

3. FRACTAL FEATURES OF DNA

Aggregation processes taking place on substrates of defined dimensionality are known to produce fractal spatial structures. A well known example of aggregation mechanism producing fractal patterns is the Diffusion Limited Aggregation Model (DLA),(44, 45) where aggregation of small particles on a 2-d surface produce fractal patterns with fractal dimension $d_f = 1.7$.(46) The diffusion limited Cluster-Cluster Aggregation Model (CCA)(47) gives also rise to spatial fractal structures with fractal dimensions $d_f = 1.43$ in 2-d (calculated from the radius of gyration). In both DLA and CCA models, simple modifications on the dynamical generating mechanisms affect the spatial structure of the aggregates and thus the fractal dimensions. Inverse aggregative mechanisms are responsible for processes related to crack propagation in material breakdown also create spatial fractal patterns.(48–51) These crack mechanisms are an examples of breakup mechanisms which are also acting in fusion-transposition models.

The above models are not directly related to the dynamical mechanisms described in the previous sections mainly because the DLA and CCA models take explicitly into account the spatial restrictions of the substrates where the aggregation takes place, while the models developed in the previous sections are mean-field models. However, in reality the DNA oligomers or larger segments move in 3-dimensional space when they are transposed, replicated, aggregate or fuse together. Since all these dynamical processes take place on the 3-dimensional space it not surprising to find fractal characteristics in the linear structure of DNA.

Fractality in genomic sequences has been the subject of several earlier studies such as the wavelet fractal analysis by Arneodo *et al.*(12) and the graphical representation of oligonucleotide strings by Hao *et al.*(52) Our earlier studies on the size distribution of coding and non-coding regions of higher eucaryotes in particular, showed that the non-coding regions follow long range distributions while the intervening coding parts follow short range distribution.(16) These are certainly further indications about the fractality on the linear structure of DNA, especially in higher eucaryotes.

Motivated by the presence of fractality in most aggregative processes(44–47) we have examined long sequences belonging to different classes of organisms(53) searching for evidence of fractality. In particular, we regard the coding and non-coding segments as two different phases on a 1-d line and we use the normal box counting method(40) to find the fractal dimensionality of the sequence.

In Fig. 7 we present the number of boxes $M(r)$ of size r needed to fully cover DNA sequences of sizes covering several length scales: Human HUMCOL7A1X (squares) [collagen type VII, intergenic region and (COL7A1) gene], the plant *A. thaliana* ATAC002387 (triangles up) [chromosome II, clones T14P1, F4L23] and the insect *Dr. melanogaster* (triangles down) [scaffold, accession number AE002708, complete genome]. The human sequence contains 36.6 kbps distributed in 117 coding and 118 non-coding regions; the fractal dimension is computed using the box-counting technique as $D_f = 0.83$ (dashed line). The *A. thaliana* sequence contains 122.9 kbps distributed into 125 coding and 126 non-coding regions; the fractal dimension here is slightly higher that the human $D_f = 0.85$. The *Dr. melanogaster* sequence contains 24248 kbps distributed in 1180 coding and 1181 non-coding regions; the fractal dimension is also $D_f = 0.85$. In the same plot we show the result for simulation of an artificial random Cantor fractal of the same fractal dimension $D_f = 0.83$ and similar linear size as the human sequence. Simulation of an artificial sequence with the fractal dimension and the size of the *A. thaliana* sequence give equally good results (not shown, ref. 53). We observe that the structure of both DNA sequences are well described by fractals with nearly equal exponents. We have examined a number of large sequences originating from higher eucaryotes and all of them showed similar behavior with fractal dimensions in the region $D_f = (0.80-0.85)$. Some of the examined sequences have shown several regions of different slope, indicating that different fractal exponents may be dominant at different scales.

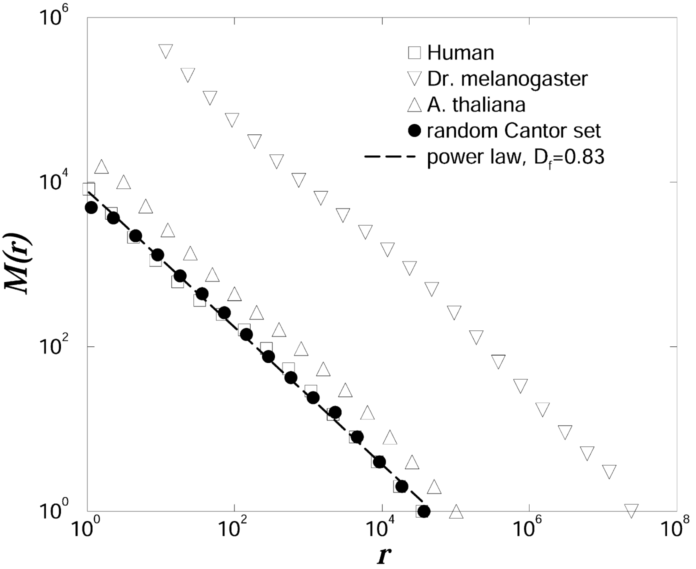

Fig. 7. The number of boxes $M(r)$ of size r needed to cover Human HUMCOL7A1X sequence (squares), plant *A. thaliana* ATAC002387 sequence (triangles up) and *Dr. melanogaster* AE002708 scaffold (triangles down). The black spheres correspond to an artificial random sequence simulating a fractal Cantor set with fractal dimensions equal to $D_f = 0.83$ as in the human sequence.

Our results on fractality of the linear structure of higher organisms are obtained for real sequences of finite sizes. In the mathematical sense however, fractals are considered as dust, whose measure is zero. In our calculations, even if we consider entire chromosomes or genomes of organisms, still the obtained sequences will be of finite size and with finite coding percentage. For example, we know today that in human DNA the coding regions cover approximately 5%, thus one cannot have the complete mathematical fractal behavior which can only be obtained at the infinite size limit.

Following the previous sections lower organisms have reduced non-coding, and thus fractality, if it exists, is difficult to be observed. Most box-counting plots of lower organisms give a fractal dimension equal to 1, indicating homogeneous fusion of coding and non-coding regions.[53] This finding of homogeneity of course does not exclude the possibility of the genome of lower organisms to have the structure of a "fat fractal."[54] This idea may be supported by the long range statistics found even in the restricted non-coding of these organisms.

At the level of nucleotide clustering (Pu/Py alternations) different authors have searched for fractal features by mostly regarding the structure of a DNA walk and by attributing to it a fractal dimension. The DNA walk is then a curve embedded on a 2-D substrate, as can be seen in refs. 9 and 12. The fractality introduced in the current study is more macroscopic and concerns structures with alternating coding and non-coding segments, while the substrate, genome, is 1-dimensional. The fractality presented here represents therefor a different genomic scaling property than cases reported in refs. 9 and 12.

The non-trivial clustering and fractality at the level of coding/non-coding presented in this work and attributed to the action of the dynamical mechanisms may also be inferred from earlier studies of the coding/non-coding cluster size distributions.[16,17] Namely, in a single DNA chain, segments which follow a short range distribution (coding) are alternated with segments which follow a long range distribution (non-coding). This complex picture may be found in random Cantor fractals embedded in 1-D substrates and one can establish a mapping between DNA sequences and Cantor fractals. A random Cantor fractal visually presents empty (white) and filled (or black) regions. The size distribution of filled regions is short ranged, while the distribution of empty spacers is long ranged power law.[40] The correspondence between (DNA $\Leftrightarrow$ Cantor) maps the (coding $\Leftrightarrow$ filled, black) regions and (non-coding $\Leftrightarrow$ empty, white) regions. This mapping has been first introduced in ref. 53. Thus a fractal dimension may be attributed to DNA sequences. A dimensionality D_f close to unity would correspond to a uniform, non-fractal structure, while $D_f \neq 1$ indicates self-similar structures. As a consequence, the fractality produced during evolution via the various dynamical mechanisms is also mirrored in the short and long range distributions found earlier in the coding and non-coding parts of higher eucaryotes, respectively.

4. CONCLUSIONS

Evolutionary dynamical mechanisms inspired from biological processes, which can reproduce the statistical characteristics of real DNA sequences have been discussed. Minimal models are constructed which include aggregation of oligonucleotides, influx of “parasite” macromolecules and replicas of DNA, and DNA reduction through outflux of DNA segments. These simple models may account for many of the complex statistical features of genomic sequences at the level of coding/non-coding juxtaposition. In addition, fusion/transposition mechanisms have been introduced, which may produce non-trivial, out-of equilibrium, correlated structures at the level of Pu/Py clustering. Fusion/transposition dynamics bridges the gap between the clustering at the higher (functional) level of coding/non-coding with the lower (composition) Pu/Py level.

Fractal features at the level of coding/non-coding alternation are studied using the box counting technique. Sequences originating from higher eucaryotes and which contain mostly non-coding present a definite fractal structure with fractal dimension D_f ranging in the region (0.80–0.85). Sequences originating from lower organisms, where the non-coding is suppressed, present $D_f = 1$, as expected since the coding covers almost completely the entire genome.

The fractal features have been connected to the short and long ranged size distributions observed in the coding and non-coding regions of higher eucaryotes. The alternation of regions of different functionality, i.e., of regions which follow short range behavior (coding segments) with ones which follow long range behavior (non-coding spacers), leads to a direct analogy with finite Cantor-like fractals. A long DNA sequence is then described as a finite, random Cantor fractal, while the infinite size limit cannot be obtained due to the finiteness of genomic sequences.

Repeats have been discussed in the previous sections and their contribution was considered as normal influx which increases the size of non-coding regions. Thus from the point of organization at the level of coding non-coding the influx of repeats helps in the establishment of the power law behavior of the non-coding size distribution. However, extended occurrence of repeats might diminish the power law character of the distribution of clusters of homologous nucleotides in the non-coding (Pu clusters and Py clusters). From the study of different organisms[14, 15] it was shown that the Pu or Py cluster size distribution of higher eucaryotes follows a power law with exponents μ considerably larger than the exponents found in the corresponding non-coding size distribution.[16] Because the non-coding regions of DNA are composed by a large number of Pu and Py clusters, and the Pu/Py clusters follow a power law behavior, according to the Generalized Central Limit Theorem the non-coding regions of DNA must follow also power law with the same exponent. The difference in the value of the exponents μ of the Pu/Py cluster size distributions with the corresponding exponents of the non-coding size distribution may be attributed not only to point mutations but also to repeats of identical segments which may modify the long range statistics introducing many copies of Pu/Py clusters with identical lengths.

In the near future the products of the different genome projects will offer complete genomes of higher organisms which will be both sequenced and reliably annotated (i.e., the coding/non-coding and functional character of the sequence segments will be determined). These long sequences will help to explore larger length scales and to investigate differences and similarities between the different classes of organisms. Large scale statistical markers, based on the values of the distribution exponents, fractal dimensions, non-randomness measures etc. can be introduced as global characteristics of whole chromosomal sequences. From the evolutionary point of view, more detailed models beyond mean field need to be constructed taking into account the spatial organization and constraints, in order to further reproduce the local features and irregularities that are observed in DNA.

APPENDIX I: CODING SIZE DISTRIBUTIONS

The size distributions of oligomers merging together to form a coding sequence is assumed to follow a Gaussian distribution (short range), Eq. (4). The probability to find a coding segment of size S resulting from the juxtaposition of $N \rightarrow \infty$ oligomers, takes the form:

$$P(S) = \prod_{j=1}^{N} P(s_j)|_{\sum_{j=1}^{N} s_j = S} \tag{25}$$

where s_i is the size of the ith oligomer. Let us call $Q_s(\rho) = \int P(S) \times \exp[-i\rho S]\, dS$ the Fourier transform of the size probability distribution, where the subscript s denotes that there is a variation on the size s of the merging oligomers. By taking the Fourier transform in both hand sides of Eq. (25) and using the Fourier transforms $Q_i(\rho)$ of the Gaussians $P(s_i)$, Eq. (25) reduces to

$$Q_s(\rho) = Q_1(\rho)\, Q_2(\rho) \cdots Q_N(\rho) \tag{26}$$

where the Fourier transform of the Gaussian Eq. (4) is

$$Q(\rho) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{\rho^2 \sigma^2 - 2i\rho\langle s\rangle}{2}} \tag{27}$$

Inserting Eq. (27) for the various values of s_i in Eq. (26) and taking the inverse Fourier transform we reach Eq. (5).

Let us now consider the case where the oligomers merging together to form a coding sequence have a distribution $P(s)$ in their length size and also a distribution $\mathscr{P}(N)$ in their number size. Moreover consider a Gaussian, short ranged, number distribution as in Eq. (7). Substituting expressions (4) and (7) in Eq. (6), and taking the Fourier transform with respect to the variable S we obtain

$$Q_{s/N}(\rho) = \frac{1}{\sqrt{2\pi}} \int \mathscr{P}(N)\, dN (2\pi)^{N/2}\, Q^{N}(\rho), \tag{28}$$

Where $Q_{s/N}(\rho)$ is the Fourier transform of $P(S)$ when both the oligomer size distribution and number distribution vary. Using the Gaussian form Eq. (7), Eq. (28) reduces further to

$$Q_{s/N}(\rho) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-i\rho\langle N\rangle\langle\sigma\rangle - \rho^2\left[\frac{\langle N\rangle\sigma^2 + \langle s\rangle^2 \Sigma_N^2}{2}\right]} \tag{29}$$

In Eq. (29) only terms of order up to ρ^2 have been retained, and thus the expression for the coding size distribution will be valid in the limit of large values of S. By Fourier inverting Eq. (29) the size probability distribution is obtained as

$$P(\rho) = \frac{1}{\Sigma_{s/N}\sqrt{2\pi}} \mathrm{e}^{-(S - \langle N\rangle\langle s\rangle)^2 / 2\Sigma_{s/N}^2}, \tag{30}$$

where $\Sigma_{s/N}^2 = \langle N\rangle\, \sigma^2 + \langle s\rangle^2\, \Sigma_N^2$ accounts for the combined effect of varying simultaneously the size and number distribution of the merging oligomers.

APPENDIX II: CUMULATIVE SIZE DISTRIBUTIONS

The cumulative distribution of a given distribution function $P(s)$ is defined as

$$\tilde{P}(S) = \int_S^\infty P(s)\, ds \tag{31}$$

The forms of the cumulative distributions for different forms of $P(s)$ are presented in the next paragraphs:

1. For a δ distribution centered around the value s_0, the cumulative size distribution is a Heaviside, or step function:

$$\tilde{P}(S) = \int_S^\infty \delta(s - s_0)\, ds = H(S - s_0) = \begin{cases} 1 & \text{if} \quad S \leqslant s_0 \\ 0 & \text{if} \quad S > s_0 \end{cases} \tag{32}$$

2. If $P(s)$ is a Gaussian distribution

$$\tilde{P}(S) = \int_S^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-s^2} \, ds = \frac{1}{\sqrt{2\pi}} erfc(S) \tag{33}$$

where $erfc(S)$ is the complementary Error function. The actual form of the $erfc(S)$ is a sigmoid, centered around 0. If the Gaussian distribution is centered around s_0, then the cumulative, the Error function will also be centered around s_0. Visually one may consider the Gaussian distribution as a "fat" delta function. Similarly, the cumulative, of the Gaussian will be an inclined Heaviside-function, a sigmoid.

3. If $P(s)$ has a power law form $P(s) \sim s^{-1-\mu}$ then

$$\tilde{P}(S) \sim \int_S^{\infty} s^{-1-\mu} \, ds \sim S^{-\mu}, \qquad 0 < \mu \leqslant 2 \tag{34}$$

The cumulative distribution of a power law is also a power law with a different exponent. Note that the stable distributions have power law tails, thus the cumulative functions of the stable distributions behave as power laws for large values of their variables. The limitations to the values of μ between 0 and 2 are due to normalisability of the distribution for the lowest limit and the long tail features for the upper limit.

ACKNOWLEDGMENTS

The authors would like to thank the referees for helpful suggestions and for bringing to our attention the large annotated Adh region of *Dr. melanogaster*. The authors would like to thank A. Smit and P. Green for allowing the use of the RepeatMasker, at the University of Washington.

REFERENCES

1. G. Nicolis and I. Prigogine, *Exploring Complexity* (Freeman, New York, 1989).
2. G. Nicolis, *Introduction to Non-linear Science* (Cambridge University Press, Cambridge, 1995).
3. G. Nicolis, *Progr. Theoret. Phys.* **49**, 825 (1986).
4. S. A. Kauffman, *The Origins of Order: Self-Organisation and Selection in Evolution* (Oxford University Press, New York, 1995).
5. W. Ebeling and G. Nicolis, *Chaos, Solitons and Fractals* **2**:635 (1992); H. Herzel and I. Grosse, *Physica A* **216**:518 (1995).
6. M. Kostianovski, *Ultrastruct. Pathol.* **24**:59 (2000).
7. I. Dunham *et al.*, *Nature* **402**:489 (1999).
8. M. Hattori *et al.*, *Nature* **405**:311 (2000).

9. C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356**:168 (1992).
10. W. Li and K. Kaneko, *Europhys. Lett.* **17**:655 (1992).
11. R. F. Voss, *Phys. Rev. Lett.* **68**:3805 (1992).
12. A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, *Phys. Rev. Lett.* **74**:3293 (1995); A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Thermes, *Physica D* **96**:291 (1996).
13. 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); A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, *J. Theor. Biol.* **151**:323 (1991); S. Karlin, B. E. Blaisdell, R. J. Sapolsky, L. Cardon, and C. Burge, *Nucleic Acids Res.* **21**:703 (1993); S. Karlin and V. Brendel, *Science* **259**:677 (1993); H. Herzel, E. N. Trifonov, O. Weiss, and I. Grosse, *Physica A* **249**:449 (1998).
14. S. V. Buldyrev, A. L. Goldberger, C.-K. Peng, M. Simons, and H. E. Stanley, *Phys. Rev. E* **47**:4514 (1993).
15. A. Provata and Y. Almirantis, *Physica A* **247**:482 (1997).
16. Y. Almirantis and A. Provata, *J. Statist. Phys.* **97**:233 (1999).
17. A. Provata, *Physica A* **264**:570 (1999).
18. Y. Almirantis, *J. Theor. Biol.* **196**:217 (1999); Y. Almirantis and A. Provata, *Bull. Math. Biol.* **59**:975 (1997).
19. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell* (Garland Publishing, New York, 1994); B. Lewin, *Genes VI* (Oxford University Press, Oxford, 1997).
20. M. Hoglund, T. Sall, and D. Rohme, *J. Mol. Evol.* **30**:104 (1990).
21. K. Mizuuchi, *Ann. Rev. Biochem.* **61**:1011 (1992); N. L. Craig, *Curr. Topics Microbiol. Immunol.* **204**:27 (1995).
22. A. Agrawal, *Science* **290**:1715 (2000).
23. M. R. Wallace, L. B. Andersen, A. M. Saulino, P. E. Gregory, T. W. Glover, and F. S. Collins, *Nature* **353**:864 (1991).
24. D. E. Berg and M. M. Howe (eds), *Mobile DNA* (American Society for Microbiology, Washington DC, 1989).
25. G. P. Holmquist, *J. Mol. Evol.* **28**:469 (1989).
26. K. Kobayashi *et al.*, *Nature* **394**:388 (1998).
27. B. Charlesworth, P. Sniegowski, and W. Stephan, *Nature* **371**:215 (1994).
28. M. W. Mueller, M. Allmaier, R. Eskes, and R. J. Schweyen, *Nature* **366**:174 (1993); C. H. Sellem, G. Lecellier, and L. Belcour, *Nature* **366**, 176 (1993).
29. E. W. May and N. L. Craig, *Science* **272**:401 (1996).
30. H. Zischler, H. Geisert, A. von Haeseler, and S. Paabo, *Nature* **378**:489 (1995).
31. R. V. Collura and C.-B. Stewart, *Nature* **378**:485 (1995).
32. M. A. Matzke, M. F. Mette, and A. J. Matzke, *Plant. Mol. Biol.* **43**:401 (2000).
33. S. Garcia-Vallve, A. Romeu, and J. Palau, *Genome Res.* **10**:1719 (2000).
34. P. Worning, L. J. Jensen, K. E. Nelson, S. Brunak, and D. W. Ussery, *Nucleic Acids Research* **28**:706 (2000); H. Ochman, J. G. Lawrence, and E. A. Groisman, *Nature* **405**:299 (2000).
35. J. Field, B. Rosenthal, and J. Samuelson, *Mol. Microbiol.* **38**:446 (2000).
36. W. Ford Doolittle and J. M. Logsdon, Jr., *Current Biology* **8**:R209 (1998).
37. S. Brenner, G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, and S. Aparicio, *Nature* **366**:265 (1993).

38. R. L. Small and J. F. Wendel, *Genetics* **155**:1913 (2000); G. Drouin and M. Moniz de Sa, *J. Mol. Evol.* **45**:509 (1997); J. A. Frugoli, M. A. McPeek, T. L. Thomas, and C. R. McClung, *Genetics* **149**:355 (1998); L. L. Longuercio and T. A. Wilkins, *Current Genetics* **34**:241 (1998).
39. Y. I. Wolf, A. S. Kondrashov, and E. V. Koonin, *Genome Biol.* **1**:6 (2000).
40. H. Takayasu, *Fractals in the Physical Sciences* (Manchester University Press, 1990); W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1966).
41. H. Takayasu, M. Takayasu, A. Provata, and G. Huber, *J. Statist. Phys.* **65**:725 (1991); H. Takayasu, *Phys. Rev. Lett.* **63**:2563 (1989).
42. M. Ashburner *et al.*, *Genetics* **153**:179 (1999).
43. A. F. A. Smit and P. Green, RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html
44. T. A. Witten and L. M. Sander, *Phys. Rev. Lett.* **47**:1400, (1981); T. A. Witten and L. M. Sander, *Phys. Rev. B* **27**:5686 (1983).
45. T. Vicsek, *Fractal Growth Phenomena* (World Scientific Publishing, Singapore, 1989).
46. P. Meakin *Phys. Rev. A* **26**:1495 (1983).
47. P. Meakin *Phys. Rev. Lett.* **51**:1119 (1983); R. Botet and R. Jullien, *J. Phys. A* **17**:2517 (1985).
48. H. J. Herrmann, J. Kertesz, and L. de Arcangelis, *Europhys. Lett.* **10**:147 (1989).
49. L. Fernandez, F. Guinea, and E. Louis, *J. Phys. A* **21**:L301 (1988).
50. E. Louis and F. Guinea, *Europhys. Lett.* **3**:871 (1987).
51. P. Meakin and A. T. Skjeltorp, *Advances in Physics* **42**:1 (1993).
52. B. L. Hao, *Physica A* **282**:225 (2000); B. L. Hao, H. C. Lee, and S. Y. Zhang, *Chaos, Solitons and Fractals* **11**:825 (2000).
53. A. Provata and Y. Almirantis, *Fractals* **8**:15 (2000).
54. R. Eukholt and D. K. Umberger, *Physica D* **30**:43 (1988).