Statistical properties and linguistic coherence in noncoding DNA sequences

B. Cantú-Bolán
Departamento de Biología, Facultad de Ciencias, Universidad Nacional Autónoma de México, Coyoacán 04510, México D.F., México

E. Hernández-Lemus*
Departamento de Física y Química Teórica, Facultad de Química, Universidad Nacional Autónoma de México, Coyoacán 04510, México D.F., México

Recibido el 29 de marzo de 2005; aceptado el 16 de junio de 2005

It has generally been thought that the vast majority of the DNA of living organisms (about 95%) was constituted of what is now called non-coding DNA (*NC-DNA*). No mechanisms of the genetic expression were known for this *NC-DNA*, as opposed to the protein expression for coding DNA (*C-DNA*). So *NC-DNA* was traditionally assigned a role as a cover-up (with no biological function of its own) against the random attack of mutagenic elements on the *C-DNA*. Nevertheless (and in some sense motivated by the discovery of the tertiary structure of the genetic code), studies into the nature and biological function of *NC-DNA* began. Some of the tools of multifractal theory and statistical linguistics were recently applied to the analysis of coherence and correlation in non-coding DNA fragments. As a result, the presence of long-range correlations, coherent patterns, and even some well defined structural features, showed up. This structure and correlation would be impossible to find in a random nucleotide sequence (as *NC-DNA* was originally thought to be constituted).
Keywords: DNA; genomics; statistical linguistics; fractal dimension.

Hasta hace poco tiempo existía la creencia general de que la gran mayoría del ADN de organismos vivos (alrededor de un 95%) estaba constituido por lo que se ha dado en llamar ADN no codificante (NC-DNA, por sus siglas en inglés). No se conocen mecanismos de expresión para este *NC-DNA*, en contraste con lo que sucede con la expresión en forma de proteínas que posee el ADN codificante *C-DNA*. Así las cosas, el papel tradicional que se asignaba al *NC-DNA* era el de una protección (sin función biológica propia) contra el ataque aleatorio de elementos mutagénicos sobre el *C-DNA*. Sin embargo, algunos estudios comenzaron a realizarse sobre la naturaleza y función biológica del *NC-DNA*, estudios en cierto modo motivados por el descubrimiento de la estructura terciaria del código genético. Algunas herramientas de la teoría de multifractales y de la lingüística estadística han sido aplicados recientemente al análisis de coherencia y correlación en fragmentos de ADN no codificante. Como resultado, se ha mostrado la presencia de correlaciones de largo alcance, patrones coherentes y aún ciertas características estructurales. Tal estructura y correlación serían imposibles de hallarse en la secuencia aleatoria de nucleótidos que se pensaba, constituía el *NC-DNA*.
Descriptores: ADN; genomica; lingüística estadística; dimensión fractal.

PACS: 87.10+e

1. Scope

In this work several statistical analysis were carried out on DNA fragments (coding and non-coding) of some representative species of animals (*Felis catus*, *Drosophila melanogaster*), plants (*Pinus thunbergii*) and bacteria (*Mycoplasma pneumoniae*); we also analyzed a random generated *genetic* sequence in order to highlight the common features and main differences. We studied the probability distribution for the frequency of *genetic words* (*i.e.* permutations of nucleotides A, C, T and G) of a size up to five, in order to construct the frequency vs rank plots also known as Zipf plots. We also studied the *time* series associated with the position of each one of these bases. A renormalization procedure was carried out and the resulting renormalized time series were studied as multifractals. The Hausdorff dimension and Shannon-Weaver entropy were calculated for these sets. A quantitative measurement of the concentration of nucleotide bases in NC-DNA was taken and compared with a theoretical estimate for C-DNA (the so called *Chargaff's rules*).

This paper is organized as follows, Sec. 2 is an introduction to the statistical linguistic description of genomes and the major achievements of this approach up to this day. In Sec. 3 we review and describe some of the more common methods used in quantitative linguistics and statistical physics that have proven to be useful in the description of DNA sequences. Section 4 is devoted to giving some general results for the cases studied; it must be stressed that we took genomic fragments of very different living beings and, in the case of *Mycoplasma pneumoniae*, the whole genome was analyzed. In all cases the genomic sequences consisted of approximately one hundred thousand base pairs (100,000 bp). The choice of genomes to be studied was made on the basis of three fundamental criteria:

i) We wished to study the broadest possible variety of living beings available (animals both vertebrate and invertebrate, plants, microorganisms, etc.) since in this case we looked for *universal linguistic features* in the genomic texts.

ii) Our selection had to be included in the GenBank database [1] since this one of the most reliable sources for genomic data.

iii) As we said, the size of all genomic sequences is about 10^5 bp, that is, the size of the entire genome of the *My-*

coplasma pneumoniae bacterum, an organism which at the same time is an already *complex* form of life (from the metabolic and physiological point of view) and possesses a genome which is of a manageable size.

The computational calculations and word countings were made on a personal computer (single Intel Pentium V processor), and only some very specific tasks required a unix workstation (Alpha Twin Peaks running under RH Linux for generating the nucleotide combinatorics for the word list and the random generated sequence), and in no case did a single calculation take more than a few hours, and usually just a few minutes. These computer requirements are far, below the usual computational genomic standards, which in this case points to an optimized procedure. Section 5 includes an abridged but systematic discussion of results, as well as some conclusions and perspectives.

2. Introduction

We usually consider DNA molecules as the main mechanism for the storage of information about any organism. DNA molecules are long sequences (linear or even closed into a loop) contained in every cell of an organism. DNA sequences are represented by a string of four letters (A, C, G, T), each of which corresponds to a definite type of nucleotides: adenine, cytosine, guanine and thymine, respectively. These letters can be cast into different combinations and hence form a *vocabulary*. Some combinations in DNA texts are deliberately non-random. They should thus reflect the structure and function of DNA and proteins. From this the question then arises: what kind of characteristic features of nucleotide sequences correspond to known DNA properties? As we have stated above, DNA molecules are usually separated into fragments that are either catalysts for known protein reaction kinetics (*C-DNA*) or not (*NC-DNA*).

For example, recent studies into the peculiarities of bacterial DNA revealed that the word ranked distributions are quite well approximated by logarithmic law [2]. The results obtained then (in the so-called *absent word investigation*) showed the considerably nonrandom character of DNA texts. Characteristic of the autocorrelation function behavior of several genomes was the presence of period-3 oscillations. Short-range autocorrelations were shown to be present in short ($n = 3$) words and practically absent in longer words.

One of the most interesting problems in DNA analysis is to find principles for computational (automated) differentiation between coding (exons) and non-coding (intergenomic and intron) regions in DNA [3]. Now most of the computational approaches for identification of coding regions in DNA have strong implementation limitations since they need a training set of already known examples of coding and noncoding regions. Limited by the lack of availability of data, researchers work with a much shorter subsequence of DNA rather than the whole sequence and they are hardly able to recognize mainly protein coding regions. Some approaches seemed to promise to be free of these limitations. Some of these approaches may take advantage of the statistical linguistic differences between *NC-DNA* and *C-DNA* in order to improve the process of *learning*.

A variety of methods has been used to statistically study DNA sequences by means of *word counting*. Here we will use the combined approach of extended Zipf plots [4, 5] and renormalized time series multifractal analysis [6, 35]. Following the pioneer work of Stanley, et al [9] we will analyze statistical similarities between different resolution scanning genomes and hence scaling properties. Scaling features represent some of the most outstanding *emerging properties* of complex systems[i] [13, 14]. Scale invariance in systems with a vast amount of degrees of freedom is a signature of long-range correlation between individual units [15, 16].

By means of a scaling analysis, Stanley *et al.* [9] found evidence supporting the idea that the DNA sequence in genes containing noncoding regions is correlated, and that the correlation is remarkably long-range. Indeed, base pairs that are thousands of base pairs distant are correlated. They did not find this type of long-range correlation in the coding regions of the gene, and quantified the *redundancy* of a linguistic text in terms of a measurable Shannon-like entropy function [17], reporting that noncoding regions in eukaryotes display a larger redundancy than coding regions. It was shown that the cytosine-guanine (CG) concentration does have a strong *background* effect on the redundancy structural feature also mentioned in connection with a melting temperature signature by Resendis and García-Colín [18, 19] (see specially equation (20) of [18]). The relation between thermodynamic properties of DNA such as its melting temperture and composition domains evidencing internal structure has been also studied by means of the theory of stochastic processes by Dagdug and coworkers [21, 22]. In these cases, the fundamental linguistic (or structural) units under study were divided into purines and pyrimidines, looking for the behavior of clustering homopolimers and oligopolimers among these classes. In a certain sense, this work is related to our search for microsatellites or dimeric tandem repeats (DTR). This search for recursion and clustering of units has recently been reviewed by the group headed by Sun [23].

A related approach has been taken to look for the kinetic behavior of these phase changes [24]. From the standpoint of computational molecular biology, Collado-Vides, et al [25] have developed some systematic ways of dealing with the relation between linguistic optimality and biological functionality related to the denaturation process. With regard to (CG)-content and physical properties of DNA, an interesting spectral analysis of the PDF has been done recently by Li and Holste [28].

This connecting effect between redundancy and thermal stability may imply a strong connection between linguistic optimality (via-long range correlations) and the tertiary (spatial) structure of DNA incident on the melting temperature. Ultimately, this optimality stage could be connected with empirical rules regarding the average DNA concentration, such

as *Chargaff's Rules.* While studying the denaturation process, Chargaff found out that the concentration of adenine $[A]$ was approximately equal to thymine concentration $[T]$ for many organisms and the same occurred with the $[C]$ and $[G]$ bases. A direct consequence of this observation is that the ratios $[G + A]/[C + T]$ and $[G + T]/[A + C]$ are close to unity [20]. Also related to CG concentration, Shannon's redundancy for the set of analyzed sequences in [9] was greatr for noncoding regions than for coding regions.

Analytical DNA ultracentrifugation revealed that eukaryotic genomes are mosaics of isochores: long DNA segments (> 300 kb on average) relatively homogeneous in CG concentration [26]. Important genome features are dependent on this isochore structure, since genes are found predominantly in the CG-richest isochore classes. The entropic segmentation method is proven to be able to divide a DNA sequence into relatively homogeneous, statistically significant domains.

Genomic sequences consist in several levels of information. These include specifications for sequences responsible for protein structure, identification of coding and non-coding parts of the sequence, information necessary for the specification of regulatory (promoter, enhancer) sequences, information directing protein-nucleic acid interactions, and directions for DNA folding and unfolding. A remarkable mechanism is provided for the transformation of different information levels (replication, decoding, etc.) that occurs over a short time interval. The means of encoding some of this information are understood, but for the vast majority of the layers of information encrypted in a DNA molecule, relatively little is known [27]. As mentioned above, genomes of high eukaryotic organisms have just a small portion of the total genome used for protein coding.

The role of introns (continuous non-coding regions in DNA) and intergenomic sequences (*NC-DNA* fragments intertwined between *C-DNA* regions) constituting a large portion of the genome thus remains largely unknown. Nevertheless, the presence of long-range correlations (as evidenced by the scaling character of *NC-DNA* sequences) points to the presence of an underlying structural order in the intron and intergenomic segments. As we already noted, the spatial structure of DNA responsible for its thermal stability has been related to *global* features, such as CG concentration [18]. Since this concentration depends not only on local correlations (such as those present in *C-DNA*) but also on long-range correlations due largely to introns and intergenomic fragments (with a larger concentration on the DNA of living organisms), *NC-DNA* plays an important biological role in the sense of stabilizing DNA. As is well known, the process of melting *denaturates* a protein (such as DNA), depriving it of its biological functions such as enzymatic catalysis. A protection against mechanisms of thermal attack, such as that associated with long range correlations in *NC-DNA* will preserve biological functions of the protein.

On the other hand, many fruitful examples of the application of Zipf's law to DNA statistics are available [2, 5, 9]. Its practical importance for the biomedical sciences has been recently stressed by Li [31], since a difference in the expression level of a gene for two different conditions/phenotypes, such as cancerous versus non-cancerous, one subtype of cancer versus another, before versus after a drug treatment, is indicative of the relevance of that gene to the difference of the high-level phenotype. Each gene can be ranked by its ability to distinguish between the two conditions. Li studied how the single-gene classification ability decreases with its rank (Zipf's plot). A power-law distribution function in Zipf's plot was observed for several microarray data sets obtained from actual cancer studies. The presence of this power-law behavior turns out to be very important for deciding the number of genes to be used for a discriminant microarray data analysis[ii], and hence facilitates the microarray study.

3. Statistical analysis: word counting in DNA

DNA sequences have been studied by means of a variety of word counting models that can basically be grouped into two large categories. The first types are local analysis that take into account the fact that DNA sequences are produced in sequential order; hence neighboring base pairs will affect a closely-attached base pair. This type of analysis, such as typical Markov models, can indeed take into account some short-range correlations observed in genomic sequences. The second category of analysis is global in nature, and it concentrates on the presence of repeated patterns founded commonly in eukaryotic DNA.

3.1. Language generation and factorization

3.1.1. Definitions

Let us consider a finite alphabet $\Sigma = \{A, C, G, T\}$. We can collect all possible text strings within this alphabet in an infinite set Σ^*; for completeness this set Σ^* should include the empty string (*i.e.* the chain with no letter). Any subset L of Σ^* is called a *language* on Σ. In order to define the class of language, we must give the generating rule:

1. If L is a finite subset, this can be done by enumerating its elements.

2. It is possible to develop some *production rules* and apply them systematically to some initial letters (called *breeders* or *generators*) in order to develop the language completely. This is the most important and well-defiend procedure for language generation. If these rules are applied on a sequential basis, they lead to Chomsky's generative grammar. When applied in a parallel fashion, they lead to Lindenmayer systems [7, 8].

3. For a special class of languages called *factorizable languages*, it is possible to define a language by indicating a subset of forbidden words. This approach is usually followed in DNA analyses.

3.1.2. *Segmental languages*

A special class of factorizable languages (called segmental languages) could be defined over a complete genome. Given the complete genome $\mathcal{G}$ of an organism, we can *cut-out* all possible subsequences and form a language $L = sub(\mathcal{G})$. This language should include the empty string or null-word. This language is factorizable by definition (it was constructed through its elementary segments) and it is, in principle, possible to construct a deterministic language from it [7].

The last paragraph implies that it is possible, at least in principle, to construct the set of *writing rules* (in the simplest case, the set of forbidden words, but also other more specific linguistic features) for a given genome.

Now that we have seen that DNA forms a factorizable but very large segmental language, the only reasonable approach to constructing these writing rules is statistics. Nevertheless, since we have recognized a linguistic structure in DNA, it has become possible to apply linguistic characterization procedures to simplify the task of generating the writing rules.

3.2. Zipf plots

3.2.1. *Zipf's law: Power laws and linguistics*

A number of statistical relationships between frequency of occurrence and rank in linguistics were noticed in the early 1930s by George Kingsley Zipf, who taught German at Harvard, and they are all aspects of what is now called Zipf's law [29],

$$f_n = f_1 n^{-\alpha}. \tag{1}$$

In Eq. (1) f_n is the frequency of the nth most frequent word, f_1 is the frequency of the most common word, n is the rank, and α is a characteristic exponent. In the probability representation,

$$p_\nu = X^{-\alpha} \tag{2}$$

p_ν is the probability of the ν-th word and X is proportional to the rank n (in fact $X = (p_1)^\alpha\ n$). This representation will be useful later on and highlights the role of X as a generator for the *extended moments* of the associated PDF.

For well structured languages, Zipf's studies reveal $|\alpha|>1$ and for low level languages (*e.g.* in children's vocabulary) $|\alpha| < 1$. Zipf concluded that a larger value of $|\alpha|$ implie a better structured, more coherent language [4].

A careful analysis of Eq. (1) reveals interesting statistical properties. Since frequency is proportional to probability, Zipf's law represents, indeed, a PDF in rank representation. The functional form of a power law implies, then, the presence of statistical persistence. If we think of the DNA sequence as a Markov chain, a power law on its PDF means a strong non-markovian character, whereas an exponential decay would define a Markov process [30]. A larger exponent value would expand the statistical correlation length within the chain. In this sense, for a segmental language as DNA is supposed to be, a non-markovian character in the PDF implies linguistic coherence and structured information behavior. It is in this sense that this work deals with linguistic properties in DNA.

3.3. Fractal methods: Hausdorff dimension and renormalized time series

3.3.1. *Hausdorff dimension and scale invariance*

Since we proposed the probability distribution functions (PDF) for the frequency of *words* in DNA to be scale invariant, another method of characterizing genomic sequences could be found in a fractal-like analysis of the self-similar set associated with DNA's scale invariance.

3.3.2. *Definitions*

Let A be a subset of metric space X. Then the *Hausdorff dimension* $dim_H(A)$ of A is the infimum of $\alpha \geq 0$ such that the α-dimensional *Hausdorff measure* of A, $m^\alpha(A)$ is 0, and

$$m^\alpha(A) = \lim_{r\to 0^+} m^\alpha_r(A) \tag{3}$$

where

$$m^\alpha_r(A) = \inf_\sigma \left\{ \sum_{k=1}^{\infty} (r_k)^\alpha \right\}.$$

The α-dimensional Hausdorff measure of A, $m^\alpha(A)$, is the infimum of positive numbers σ such that for every $r > 0$, A can be covered by a countable family of closed sets, each of diameter less than r, such that the sum of the α^{th} powers of their diameters is less than σ. Note that $m^\alpha(A)$ may be infinite, and α need not be an integer [34].

A quantity related with the Hausdorff dimension of a set is called the Kolmogorov capacity $dim_K(A)$ of this set:

$$dim_K(A) = \lim_{r\to 0} \sup \frac{\log N(r, A)}{\log(1/r)}, \tag{4}$$

where $N(r, A)$ is the number of open balls of up to radius r needed to cover the set A in the topological sense. There are several conditions relating the Hausdorff dimension of a set and its Kolmogorov capacity, namely:

$$dim_H(A) \leq dim_H(B) \quad \text{if} \quad A \subset B \tag{5}$$

$$dim_H(A) \leq dim_K(A) \tag{6}$$

It is also possible to write an expression for $dim_H(A)$ simmilar to that of the Kolmogorov capacity $dim_K(A)$:

$$dim_H(A) = \lim_{\alpha\to\alpha_0} \frac{\log N(A, \alpha)}{\log(1/r)} \tag{7}$$

Eq. (7) could be inverted to give

$$N(A, \alpha) = r^{dim_H} \tag{8}$$

Eq. (8) relates a measurable quantity $N(A,\alpha)$ with the Hausdorff dimension of a set. Based on similar considerations, Procaccia and Grassberger [35] proposed a related algorithm to calculate the fractal dimension of a chaotic signal given as a time series. We shall use this Integral Correlation Function (ICF) to calculate the fractal dimension for the time series for each nucleotide base.

The ICF is given by:

$$C(r) = \frac{1}{N^2}\sum_{i=1}^{N}\Theta\left(r - \|X_i - X_j\|\right), \tag{9}$$

where N is the size of the set (number of points), X_i are the binary-valued vectors for the time series for each nucleotide[iii] and Θ is Heavyside's unitary step function.

The afore-mentioned method makes it possible to construct the plot of $C(r)$ vs r, which possesses a power law representation. The power involved is a lower bound for the Hausdorff dimension of the set given by a time series [35].

Less formally speaking, if a self-similar object with parameters N and s is described by a power law such as $N = s^{\alpha}$, then α is the *dimension* of the scaling law and it is known as the *Hausdorff dimension*. Sometimes the Hausdorff dimension is also called the *fractal* dimension. These two concepts are not really the same. Fractals are objects that possess self-similarity *on all scales*. In the natural sciences, however, the terms "fractal" and "self-similar object" are often used as synonyms.

DNA sequences have been handled using by fractal methods in the past, either by direct calculation or by means of a *random walker* diffeomorphism [37, 38]. For example Berthelsen *et al.* [38] used a pseudo-random walk representation in a four-dimensional embedding to estimate the global fractal dimension of 164 GenBank sequences. In recent times, a debate has risen whether long-range correlations found in DNA are present in *NC-DNA*, *C-DNA* or both. For an interesting review of the debate and a global fractal-like analysis of some sequence see the work of Yu *et al.* [39]. In any case, in this work we are more interested in characterizing *NC-DNA* and, as is shown in Buldyrev *et al.* [40] by using all the DNA sequences available, long-range correlations appeared mainly in non-coding DNA. This being the case we made a Hausdorff dimension analysis of genomes looking for a *NC-DNA signature* (*i.e.* we would try to relate a Hausdorff dimension characteristic to a given sequence).

3.4. Shannon-Weaver entropy: information content at the genetic level

In order to make quantitative statements regarding the information content of a DNA segment, it is useful to apply the algorithm developed by Shannon and Weaver to analyse the information that could be transmitted in a message. The notion of Shannon's information I is defined by:

$$I = -K_s\sum_{\nu} p_\nu \log p_\nu, \tag{10}$$

here I is Shannon-Weaver's information (SWI), K_s is a constant and p_ν is the probability (frequency) for the ν-th word. In the case of NC-DNA analysis, we consider SWI to be the better information measure since it has no adjustable parameters (a valuable characteristic if one has no previous knowledge of the codification procedure) and the value of SWI is a lower bound for information content, so any other measure will be greater than SWI (*i.e.* we are studying NC-DNA in a worst-case scenario).

In addition in the case of SWI, it is possible to prove that a local maximum on the information is achieved if the PDF follows a power law, *e.g.* Zipf's law behavior. The proof reads as follows: If $p_\nu = X^{-\alpha}$ (Zipf's law according to Eq. (2)), then

$$\frac{dp_\nu}{dX} = -\alpha X^{-\alpha+1}; \quad \frac{d^2p_\nu}{dX^2} = \alpha(\alpha+1)X^{-\alpha+2};$$
$$\ln(p_\nu) = -\alpha\ln X \tag{11}$$

Maximization of SWI implies:

$$\frac{dI}{dX} = 0; \quad \frac{d^2I}{dX^2} < 0. \tag{12}$$

In this particular case,

$$\frac{dI}{dX} = \sum_{\nu} K_s\,(1-\alpha\ln X)(\alpha X^{-\alpha+1}) \tag{13}$$

which, combined with the first equation in (12) gives rice to a condition relating the rank X and the exponent α for the maximized case, namely $(1-\alpha\ln X) = 0$. Also,

$$\frac{d^2I}{dX^2} = -\sum_{\nu} K_s\Big[(1-\alpha\ln X)\alpha(\alpha+1)X^{-\alpha+2} + \frac{1}{X^{-\alpha}}(-\alpha X^{-\alpha+1})^2\Big]. \tag{14}$$

After considering the condition between X and α given by the first derivative and rearranging, we obtain

$$\frac{d^2I}{dX^2} = -\sum_{\nu} K_s\left(\alpha^2\,X^{-(\alpha+2)}\right). \tag{15}$$

Since $K_s > 0$, $X > 0$, $\alpha^2 > 0$ and $X^{-(\alpha+2)} > 0$, $\forall\alpha$, the second derivative d^2I/dX^2 is negative, indicating that Zipf's law implies a local (at least) maximum of information content. If, on the other hand, we were to consider the exponential decay of the PDF, namely $p_\nu = e^{-\beta X}$ (typical of Markov statistics) it can be proved that a maximum of information content could not be reached, since this functional form induces a saddle point at $X = 1/\beta$. Equation (15) also shows that the peak in the maximum information content is proportional to α^2, so that a language with a larger value of α will be able to carry more information. These facts will be taken into account later on.

4. Results

4.1. Shannon's Information Analysis

When looking at a plot of accumulated Shannon entropy vs. position in the renormalized time series[iv] (Fig. 1), it is possible to notice the significant contrast between the random generator and other experimental subjects. In the case of the DNA of living beings, the series plots all start with a high (absolute) value of Shannon's entropy (*i.e.* a low information content) that gradually grows, thus enlarging the information content of the chain. The presence of an almost continuous curve is noticeable in the case of all living beings.

This fact contrasts with the presence of discontinuities in the curve corresponding to the random-generated sequence which exhibits four well-defined clusters in the form of plateau. This behavior is due to the fact that our random string of nucleotides is not absolutely lacking in information but has some redundant information in its structural formation, *i.e.* there are words of four different sizes plus the overall count. It can be seen that the information content of NC-DNA increases along the chain instead of decreasing, as would be expected for random sequences. This fact is in no way a violation of some form of the second law of thermodynamics since DNA codification is based on a series of biochemical reactions with complex molecular behavior and dynamics which obviously dissipate energy [12, 26].

4.2. Probability distribution function (PDF) and Zipf's analysis

The results shown in Fig. 2 indicate the very low coherence of the random generated sequence, as opposed to the coherent character of the DNA of living beings. Since we have already mentioned the relation between linguistic coherence and the presence of long-range order as given for the tails of the distribution in a frequency vs. rank plot, we may notice a contrast between the long asymptotic tails of every curve but the one corresponding to the generated sequence since in this specific case the decay of the correlation occurs at a very short distance, showing a functional behavior closer to exponential decay than to the presence of a power law. As we have seen, a power law in a frequency vs. rank plot indicates coherence through its long *time* tails. In fact, it is possible to relate the presence of long tails in the PDF with a non-markovian character of the transition probabilities between states (*i.e.* nucleotides in a genomic string) [10,12,14]. This feature could be derived from the fact that, as we have already mentioned, a power law behavior such as the one in the frequency vs. rank plots of NC-DNA implies, on one hand, a coherent structure through Zipf's linguistic analysis as well as long-range order derived from its non-markovian character, and on the other hand a local maximum on its information content, as seen from Shannon's entropic analysis given by Eq. (15) and the paragraph below. In the preceding subsection, we also pointed out to the connection between coherence, order, and information, as given by the power law representation of the PDF of NC-DNA, with all features also present in a linguistic (human produced) text. The next subsection will touch on another characteristic property of language complexity. We address on the complexity issue by considering a multifractal set representing the message [36] in the DNA sequence and characterizing it by means of its associated Hausdorff (or fractal) dimension as a measure of complexity.

4.3. Hausdorff dimension

By observing the behavior of the Integral Correlation Function [Eq. (9)] as a function of the radius of the window used to probe the time series (Fig. 3) one can see the presence of an almost power law character in most cases; and since we made this plot on a log-log basis, the slope of the graph corresponds to the Hausdorff dimension of the associated sequence. In the case of the random generated genome for Adenine, this slope

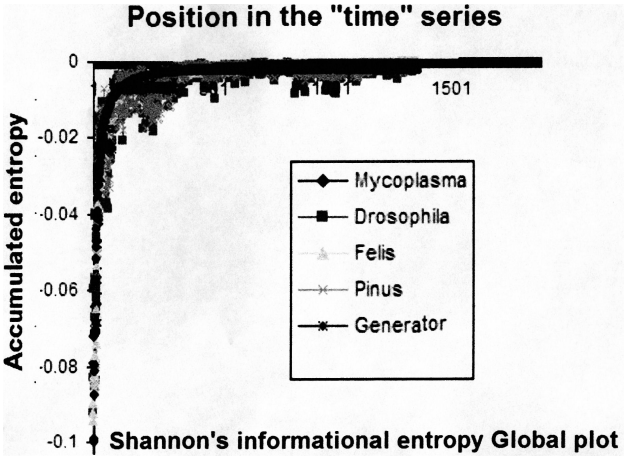

FIGURE 1. **Shannon-Weaver Informational entropy** versus position in the time series for the complete set of genomic words up to size 5 of all species plus the random generator. High absolute values of entropy mean low information content.

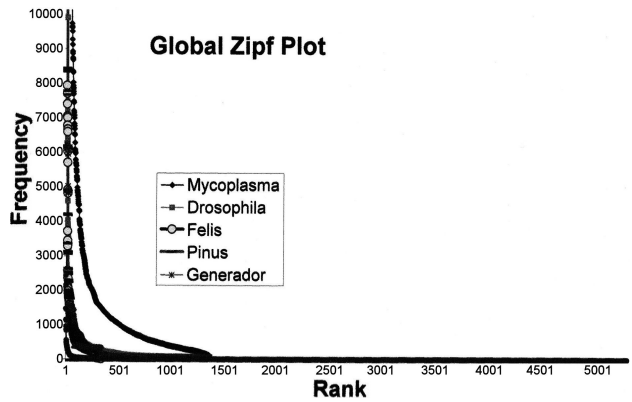

FIGURE 2. Global plot of **frequency versus rank** (**Zipf plot**) for the complete set of genomic words up to size 5 of all species plus the random generator. A larger exponent implies greater coherence, a closer correlation, and more structure.

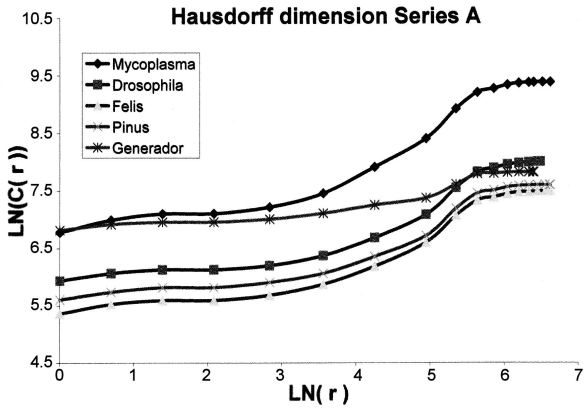

FIGURE 3. Logarithm of the **Integral Correlation Function** of Procaccia & Grassberger versus logarithm of the size of the observation window for the multifractal representing the renormalized time series for Adenine of the genomes for all species studied plus the random generator. Greater slopes imply a higher density of phase space points, hence more complexity. Horizontal line behaviour implies trivial complexity.

is almost zero (0.17 to be precise), indicating low complexity on its associated quasifractal in sharp contrast with the case of real genomic fractions. It is also worth noticing that the slope of the ICF plot was about the same for all the time series associated, whereas in the case of living beings there were higher slopes for Adenines and Thymines than for Citosines and Guanines, a fact related with the higher [A+T] content as related to the [G+C] content in DNA (some 60/40 percentage of [A+T] with respect to [C+G]), a property that in the case of S-DNA has been called *Chargaff's Rule*. We must also stress the fact that the highest complexity was found in all cases for *Mycoplasma pneumoniae*, for in this case we used the complete genome (about 100, 000 bp).

5. Conclusions

We could summarize our results for the statistical linguistic analysis of DNA sequences in the following conclusions.

1. An information theoretical analysis showed a higher information content in NC-DNA than in allother cases, including random sequences and coding sequences represented by genetic words of size three (codons).

2. The PDF of NC-DNA word countings on a frequency vs. rank representation indicated a clear non-random behavior; what is more, a highly coherent and structured linguistic character was devised.

3. Due to the long tails in the PDF's of NC-DNA, it is possible to talk about the presence of long-range correlations as opposed to the medium to short-range interaction of C-DNA [11], or to the exponential decay of a random generated genomic sequence.

4. A multifractal analysis showed a highly complex landscape in the quasifractal sets related to nucleotide distribution, thus showing the non-trivial nature of the coding [36].

5. The concentration of [A+T] content as related to the [G+C] content in NC-DNA presented values of about 60/40 percentage of [A+T] with respect to [C+G], in all genomes of living beings, whereas the random generated one presented equal parts (50 % approx.) as should be obvious, given its homogeneous random distribution.

All these reasons lead us to conclude that "non-coding" DNA contains a large quantity of information arranged in a coherent, structured and complex manner whose linguistic character (coding ?) lead us to think of some kind of biological function. Interesting advances, however, have been made in understanding the possible function of NC-DNA relating its presence with the synthesis of primordial ribosomal RNA [41, 42]. Nevertheless, a great deal of research must be done in order to clarify the complex biochemical and genetic mechanism present behind the statistical information and linguistic character of NC-DNA as revealed through this and similar studies.

$*$. Corresponding author: enrique@eros.pquim.unam.mx.

i. A scale invariant function $f(x)$ has the remarkable property that each time x is doubled, tripled, etc., the function $f(x)$ changes by the same factor. There is thus no way to set a characteristic scale for such a function. Stated mathematically, if the variable x is increased by an arbitrary factor λ, then the function is changed by a factor λ^p which is independent of the value of x, and $f(\lambda x) = \lambda^p f(x)$ for all λ. A functional equation like this one, constrains the set of possible functional forms of $f(x)$: any function $f(x)$ satisfying this equation must possess a power-law representation.

ii. The number of *relevant* genes is related to the statistic correlation lenght.

iii. In the associated time series a number one is assigned to every position containing the given nucleotide base, a zero is assigned otherwise.

iv. In order to work with time series of a more manageable size and using the fact that DNA segments could be treated as multifrac-

tals [10] we applied a fixed point renormalization approach to the original genomic time series, the renormalized series have, of course, the same properties (statistically talking) as the originals.

1. http://www.ncbi.nlm.nih.gov/Genbank/
2. Kirillova, O. V.; *Physica A* **290**, 453-463, (2001)
3. A. Gorban, A. Zinovyev, T. Popova,, *Statistical approaches to automated gene identification without teacher*, *Institut des Hautes Etudes Scientifiques Preprint*, IHES/M/01/34 (2001).
4. G.K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley Press Inc., Cambridge, (1949).
5. R.N. Mantegna *et al.*, *Phys. Rev. Lett.* **73** (1994) 3169.
6. B.B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman , New York, (1982); B.B. Mandelbrot, *An informational theory of the statistical structure of languages*, in *Communication Theory*, ed. W. Jackson, Betterworth, (1953) 486.
7. D. Welsh, *Codes and Criptography*, Oxford University Press, Oxford, (1988) 257.
8. P. Jauralde, *Introducción al conocimiento de la lengua española*, Everest, España, (1982) 444 (In spanish).
9. H.E. Stanley *et al.*, *Physica A* **273** (1999) 1.
10. H.E. Stanley, *et al. Physica A* **281** (2000) 60.
11. H.E Stanley, *Physica A* **285** (2000) 1.
12. H.R.R. Stanley *et al.*, *Journal of Biomolecular Structure & Dynamics* **17** (1999) 79.
13. E.W. Montroll, M.F. Shlesinger, in: J.L. Lebowitz, E.W. Montroll (Eds.), *Nonequilibrium Phenomena II. From Stochastics to Hydrodynamics*, North-Holland, Amsterdam, (1984) 1.
14. H.E. Stanley, N. Ostrowsky, (Eds.), *On Growth and Form: Fractal and Non-Fractal Patterns in Physics*, Martinus Nijhoff Publishers, Dordrecht, (1986).
15. A. Bunde, S. Havlin, (Eds.), *Fractals and Disordered Systems*, Springer, Berlin, (1991).
16. P.G. de Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca NY, (1979).
17. C.E. Shannon, *Bell Systems Technol. J.* **80** (1951) 50.
18. O. Resendis-Antonio and L.S. García-Colín, *Physica A* **290** (2001) 203.
19. O. Resendis-Antonio and L.S. García-Colín, *Physica A* **310** (2002) 212.
20. M.V. Volkenstein, *Molecular Biophysics*, Academic Press, New York, (1977).
21. L. Dagdug and E. Vázquez-Contreras, *Rev. Mex. Fís.* 48(S) (2002) 168.
22. L. Dagdug and L. Young, *Rev. Mex. Fís.* **50** (2004) 594.
23. T.T. Sun *et al.*, *Chaos Solitons & Fractals* **20** (2004) 1075.
24. R. Murugan, *Biophysical Chemistry* **104** (2003) 535.
25. *Gene regulation and metabolism. Post-Genomic computational approaches* eds. Collado-Vides J., y Ralf Hofestadt, MIT Press. (2002).
26. J.L. Olivera, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, *Gene* **276** (2001) 47.
27. D.B. Searls, *Nature* **420** (2002) 211.
28. W.T. Li, D. Holste, *Fluctuation And Noise Letters* **4** (2004) L453; W.T Li, D. Holste, *Phy. Rev. E* **71** (2005) 041910.
29. M. Gell-Mann, *The Quark and the Jaguar*, Freeman & Co, (1994)
30. N. van Kampen, *Stochastic processes in physics and chemistry*, North Holland, (1992)
31. W. Li, *Zipf's law in importance of genes for cancer classification using microarray data*, arxiv.org e-print : physics/0104028, (April 2001).
32. O. Dreyer, R. Puzio, *J. Math. Biol.* **43** (2001) 144.
33. T.A. McMahon, *Science* **179** (1973) 1201.
34. G.A. Edgar, *Measure, Topology, and Fractal Geometry*, New York: Springer-Verlag, (1990)
35. P. Grassberger and I. Procaccia, *Phys. Rev. Lett.* **50** (1983) 346.
36. B.H. Lavenda, *J. Phys. A. Math. Gen.* **31** (1998) 5651.
37. S. Tavare, B.W. Giddings, in *Mathematical Methods for DNA Sequences*, edited by Michael S. Waterman, CRC, Boca Raton, (1989)
38. C.L. Berthelsen, J.A. Glazier, M.H. Skolnick, *Phys. Rev. A* **45** (1992) 8902.
39. Z. Yu, V.V. Anh, B. Wang, *Phys. Rev. E* **63** (2001) 11903.
40. S.V. Buldyrev *et al.*, *Phys. Rev. E* **51** (1995) 5084.
41. S. Duga *et al.*, *Biochim. Biophys. Acta* **29** (2000) 1490, 225.
42. D. Filippini *et al.*, *Biochem. Biophys. Research Comm.* **288** (2001) 16.