Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics

R. N. Mantegna,[1,2] S. V. Buldyrev,[1] A. L. Goldberger,[3,4] S. Havlin,[1,5]
C.-K. Peng,[1,3] M. Simons,[3] and H. E. Stanley[1]

[1] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2] *Dipartimento di Energetica ed Applicazioni di Fisica, Università di Palermo, Palermo, I-90128, Italy*
[3] *Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215*
[4] *Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215*
[5] *Department of Physics, Bar-Ilan University, Ramat Gan, Israel*

(Received 17 April 1995)

We compare the statistical properties of coding and noncoding regions in eukaryotic and viral DNA sequences by adapting two tests developed for the analysis of natural languages and symbolic sequences. The data set comprises all 30 sequences of length above 50 000 base pairs in GenBank Release No. 81.0, as well as the recently published sequences of *C. elegans* chromosome III (2.2 Mbp) and yeast chromosome XI (661 Kbp). We find that for the three chromosomes we studied the statistical properties of noncoding regions appear to be closer to those observed in natural languages than those of the coding regions. In particular, (i) an n-tuple Zipf analysis of noncoding regions reveals a regime close to power-law behavior while the coding regions show logarithmic behavior over a wide interval, while (ii) an n-gram entropy measurement shows that the noncoding regions have a lower n-gram entropy (and hence a larger "n-gram redundancy") than the coding regions. In contrast to the three chromosomes, we find that for vertebrates such as primates and rodents and for viral DNA, the difference between the statistical properties of coding and noncoding regions is not pronounced and therefore the results of the analyses of the investigated sequences are less conclusive. After noting the intrinsic limitations of the n-gram redundancy analysis, we also briefly discuss the failure of zeroth- and first-order Markovian models or simple nucleotide repeats to account fully for these "linguistic" features of DNA. Finally, we emphasize that our results by no means prove the existence of a "language" in noncoding DNA.

PACS number(s): 87.10.+e

I. INTRODUCTION

Hereditary genetic information is stored in DNA. The sequences of amino acids for a given protein are encrypted by the coded correspondence between codons (triplets of nucleotides) and amino acids. This is called the *genetic code.* In higher organisms, the protein coding sequences comprise a small fraction of the total DNA (the genome) [1]. Experimental evidence of important functions of noncoding sequences has been reported in recent years [2–4]. Moreover, statistical analysis has shown that long-range power-law correlations between nucleotides are present in noncoding regions [5–7]. A fundamental question is whether information not related to the structure of proteins can be stored in these noncoding DNA sequences [8,9].

In this paper we study long DNA sequences using tools mainly developed for quantitative analysis of natural languages and symbolic sequences. Our analysis is performed on eukaryotic and viral DNA sequences. We investigate the statistical properties of all sufficiently long DNA sequences from the current version of the GenBank (Release No. 81.0) by analyzing the complete sequences as well as their separate coding and noncoding parts.

In particular, we study the frequency of n-tuples observed in each sequence in the same way that Zipf [10] analyzed the frequency not of n-tuples but rather of words appearing in texts of natural languages. We also study the Shannon n entropy [11] of DNA sequences. In natural languages the frequency of occurrence f of a given word is related to its rank R (i.e., to the position of the word in a list ordered in terms of the word frequency) by an approximate power-law relation characterized by an exponent $\zeta \simeq 1$ [10,12]. The Shannon n entropy (i.e., the entropy of the n-tuples observed in a given text string) is a nonlinear function of n for natural and artificial languages, indicating that languages are non-Markovian processes [13,14]. The non-Markovian nature of natural languages is also supported by the discovery of long-range correlations in the binary mapping of literary texts [15,16].

We find that the linguistic properties of DNA sequences of different organisms are significantly different.

(a) For coding DNA sequences a plot of $f(R)$ can be fit by a logarithmic function, while for noncoding sequences a plot of $f(R)$ may deviate significantly from a logarithmic behavior. In the investigated chromosomes and in several other DNA sequences the Zipf plot of noncoding DNA can be fitted by a power law in a relatively large interval of ranks.

(b) In the analyzed chromosomes, the noncoding regions show a lower n-gram entropy (a higher n-gram redundancy) than the coding regions. For the sequences of vertebrate and viral DNA the difference in redundancy of coding and noncoding sequences is less pronounced and less systematic.

It is known that different organisms, and even different regions of DNA from the same organism, may show a nonuniform concentration of the four DNA bases. We compare our experimental results with the predictions of low-order Markovian processes. The checks performed allow us to conclude that zeroth-order and first-order Markovian processes cannot explain our experimental findings.

The paper is organized as follows. Section II provides basic biological background. In Sec. III we discuss the "n-tuple" Zipf analysis and report the results of our comparison between coding and noncoding DNA. Section IV deals with n-gram entropy and n-gram redundancy calculations on coding vs noncoding DNA. We conclude in Sec. V with a discussion of our findings. In particular, we emphasize that our results by no means prove the existence of a "language" in noncoding DNA.

II. BIOLOGICAL MOTIVATION

In this paper we consider DNA as a symbolic sequence of a four-letter alphabet. The four letters are A, C, G, and T indicating the four bases (nucleic acids) that are the building blocks of DNA, i.e., adenine, cytosine, guanine, and thymine, respectively. Even for very simple organisms, the complexity of DNA sequences is remarkable [1].

A. Genetic code

The 20 amino acids that are the building blocks of proteins are coded by 3-tuples (strings of three successive nucleotides) of DNA called codons. There are 64 possible combinations for the four bases (AAA, AAC, AAG, AAT,...). Sixty-one codons are used to code 20 amino acids, the remaining three codons (TAA, TAG, and TGA) are the *stop* signals indicating the termination of a protein sequence. Since there are 61 codons code for 20 different amino acids, more than one codon is used to specify the same amino acid in many cases. Thus the genetic code is *degenerate.*

B. Genome complexity

In our study we mainly focus attention on eukaryotic cells (organisms in which the cells have a nucleus and the DNA is inside the nucleus). A typical gene has its coding information stored in "pieces" (exons) interspersed with a number of noncoding regions (introns). The length of an intron can vary over many orders of magnitude, e.g., from 31 nucleotides in the viral SV40 gene to over 210 000 nucleotides in the human dystrophin gene [1]. In addition to exons and introns, genomic DNA contains *intergenic* sequences that separate different genes and form more than half the human genome.

C. Two paradoxes

1. C paradox

A comparison of genome sizes from different species reveals a surprising and important finding known as the *C* paradox ("complete genome size" paradox) [17]. The

TABLE I. Sequences analyzed here.

Organism	Type of sequence	No. of nucleotides	Coding (%)
	Phage		
T 7	complete genome	39 936	91.7
phage λ	complete genome	48 502	83.0
	Bacterial		
E. coli	6 sequences	687 329	82.2
	Viral		
Herpes simplex	complete genome	152 260	78.4
Epstein Barr	complete genome	172 281	71.0
	Invertebrate		
Sacc. cere. (yeast) chromosome XI	complete chromosome	666 448	71.9
Sacc. cere. (yeast) chromosome III	complete chromosome	315 338	67.0
C. elegans chromosome III	complete chromosome	2 176 983	29,0
	Rodent		
Mus musculus (mouse)	three sequences	201 894	5.32
	Human		
Homo sapiens	nine sequences	748 843	5.33

paradox is that the size of complete genomes does not seem to be correlated with the phenotypic complexity of the species. For example, the size of the complete genome for the species *Homo sapiens* is much smaller than the one of *Amoeba dubia* (although it is much larger than the one of several simple organisms such as, for example, the *Paramecium aurelia*). Even after taking into account anthropomorphic bias, it is quite improbable that the phenotypical complexity of the lungfish ($C \approx 1.4 \times 10^{11}$ bp) is higher than the phenotypical complexity of *Homo sapiens* ($C \approx 3.4 \times 10^9$ bp).

2. Coding to noncoding ratio

A second paradox concerns the coding to noncoding DNA ratio. By analyzing the coding-noncoding ratio in several published complete genomes, complete chromosomes, and very long DNA sequences, a clear pattern emerges: The coding-noncoding DNA ratio generally *decreases* from simpler to higher organisms. An illustrative example can be obtained from the ensemble of sequences studied in this paper (see Table I).

III. ZIPF ANALYSIS OF CODING VS NONCODING DNA

Texts written in natural languages store information in a *hierarchical* fashion. In a literary text, the message is usually partitioned in sections, paragraphs, sentences, and words. Texts in natural languages can be interpreted as symbolic sequences with non-Markovian statistical properties (i.e., long-range power-law correlations). Long-range correlations are also present in artificial (e.g., computer) languages and in formal languages [13,15,16]. The discovery of long-range correlations in noncoding DNA sequences [5,6] motivated the present analysis to answer the following question: Are linguistic features present in the noncoding regions of eukaryotic and viral DNA?

A. Genetic code and genetic language

Since DNA is the most important source of genetic information, a plausible hypothesis is that additional biological information (not directly related to the structure of proteins) is stored in noncoding DNA. The genetic code is very efficient in storing the sequence of amino acids that constitutes a given protein. The genetic code is powerful but extremely specialized: It is usually used to store information about protein sequences. A more flexible tool for storing biological information could be a *genetic language*. By "language" we mean something that could, for example, direct biological procedures performed within the cell. A sequence of procedures and/or instructions with a well-defined grammar is generically referred to as a language. Here we use the term *language* in this sense.

In a specialized code, the statistical properties of the symbolic sequences are related to the properties of the coded objects (in our case the linear sequence of nucleotides in a coding region). In a language, however, the statistical properties reflect the underlying structure (grammatical and semantic) of the communication system. Quantitative linguistics and information theory have provided two powerful tools for studying the statistical properties of natural and artificial languages: (i) Zipf analysis and (ii) the entropy (or redundancy) of a source of information.

B. Conventional Zipf analysis vs n-tuple Zipf analysis

In conventional Zipf analysis, the frequency of occurrence of words present in a given text is measured by counting the number of occurrences of each word throughout the text and dividing this value by the number of words. The frequency of occurrence f of each word is then ordered from the most frequent to the least frequent value. The position of each word in this ordered list is called its rank R.

By studying log-log plots of word frequency versus word rank, Zipf discovered a heuristic power-law relation between them

$$f = \frac{a}{R^{\zeta}}. \tag{1}$$

The exponent ζ was found to be close to 1 in several texts written in different natural languages [10,12]. Equation (1) is called the Zipf law. From the publication of Zipf's seminal work there have been several attempts to prove, as well as disprove, the Zipf law [12,18,19].

Zipf behavior has been universally observed in analyses of natural and technical languages. It is important to note that since Zipf analysis is a statistical technique, it can be performed on texts of unknown languages (with the only limitation of being able to recognize the basic semantic unit: the word). Knowledge of the investigated language is not required.

On the other hand, conventional Zipf analysis has been criticized [12] since Zipf scaling can emerge in a purely random symbolic sequence if one character is defined as a "word" delimiter [12,19]. Hence, while the observation of power-law behavior in a conventional Zipf analysis is *necessary* in natural and formal languages, it is *not sufficient* to prove the existence of non-Markovian correlations in the analyzed symbolic sequence.

Here we use n-tuple Zipf analysis to investigate the statistical properties of coding and noncoding regions of eukaryotic and viral DNA. The n-tuple Zipf analysis of a symbolic text differs from the conventional Zipf analysis performed in natural languages. In a symbolic text, the elementary semantic unit (the word) is not immediately recognizable (if present). In the study of the complexity of symbolic sequences the usual approach is to investigate the statistical properties of the substrings of length n obtained from the symbolic text by progressively shifting over the text a window of n characters (see, for example, [20] and references therein). Zipf scaling does not emerge in n-tuple Zipf analysis of a pure random symbolic sequence if the occurrence frequency is the same for all symbols, while if the occurrence frequency is unequal, a log-normal Zipf plot can emerge [21]. The practical usefulness of the n-tuple analysis of natural languages (in spite of the theoretical possible shallowness, which is still also present in the n-tuple analysis) is exemplified in a recent paper in which information derived from n-tuple frequency combined with a simple vector-space technique allows a language-independent categorization of topical similarity in unrestricted text [22].

C. n-tuple Zipf analysis of artificial and natural languages

To study the differences between the conventional Zipf analysis and the n-tuple Zipf analysis, we perform an n-tuple Zipf analysis of two known "texts." The first is a

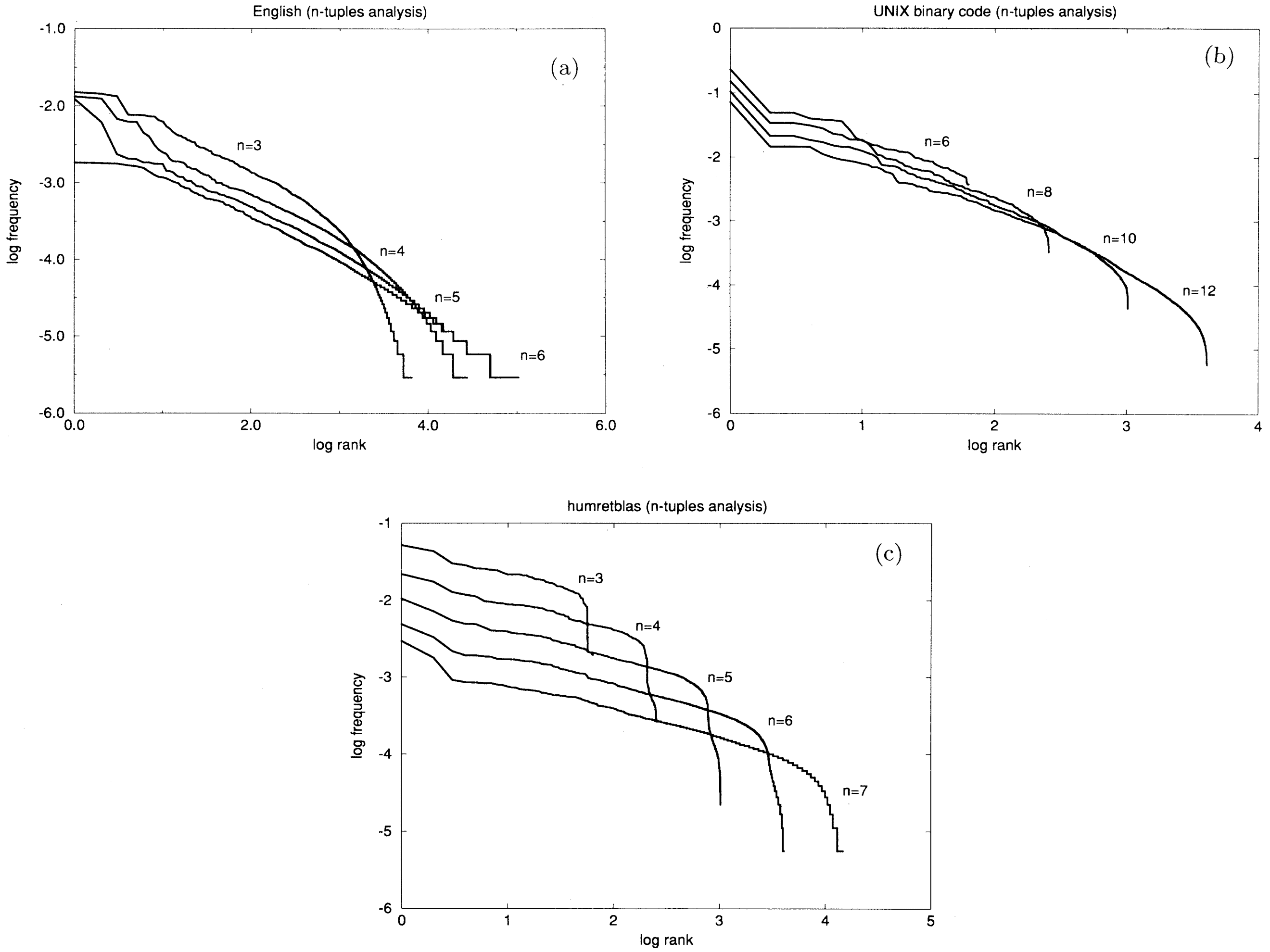

FIG. 1. (a) n-tuple Zipf analysis of a collection of English texts comprising $\approx 10^6$ words. The texts are selected from an encyclopedia. In our analysis we use a 32-character alphabet, consisting of the 26 letters of the English alphabet and 6 punctuation symbols (including the "blank" character). Characters different from the chosen 32-character alphabet are not taken into account. For $n > 3$ a power-law region extending over roughly two decades is observed. When $n = 6$ the best linear fit of the log-log plot gives $\zeta = 0.57$. Roughly the same value of ζ is obtained when $n = 4$ and 5. (b) n-tuple Zipf analysis of the compiled version of the UNIX computer operating system comprising 9×10^6 characters. The alphabet is binary. A power-law behavior is observed for a rank interval of more than two decades. In the power-law region when $n = 12$, the best linear fit of the log-log plot gives the value of $\zeta = 0.77$. Roughly the same value of ζ is obtained when $n = 8$ or $n = 10$. (c) n-tuple Zipf analysis of a primarily noncoding (1.5% coding regions) DNA sequence, the sequence HUMRETBLAS (GenBank accession code) belonging to the genome of *Homo sapiens*. From top to bottom we present the Zipf plot measured for values of n ranging from 3 to 7. A power-law behavior is observed when $R < 4^{n-1}$ and the exponent of the power law is roughly constant as a function of n. The best linear fit of the log-log plot gives the value of $\zeta = 0.34$ ($n = 6$). A deviation from the power-law regime is observed for $R < 10$ when $n = 7$. The notation "log" denotes $\log_{10}$.

collection of articles taken from an encyclopedia (written in English) and the second is a compiled file of the UNIX computer operating system. In the first case for $n > 3$ we observe a Zipf-like plot showing a power-law behavior on more than two decades. In Fig. 1(a) we report the histogram observed when $n = 3,4,5,6$. A power-law behavior characterized by an exponent -0.57 is observed when $n = 6$ in the interval $10 \leq R \leq 1000$. Roughly the same exponent is observed when $n = 3, 4, 5$.

We found that the value of the power-law exponent ζ observed in the n-tuple analysis is different from the one obtained by performing the conventional Zipf analysis of the real words of the same text. For the investigated text, by performing the conventional Zipf analysis of real words, we obtain a slope equal to -0.85, a value substantially different from the -0.57 observed in the n-tuple Zipf analysis.

This test suggests that the Zipf-like analysis of the n-tuples of a given text show a region of the Zipf plot where a power-law behavior is observed. A more detailed study of the n-tuple Zipf analysis of Markovian and non-Markovian symbolic sequences has been reported else-

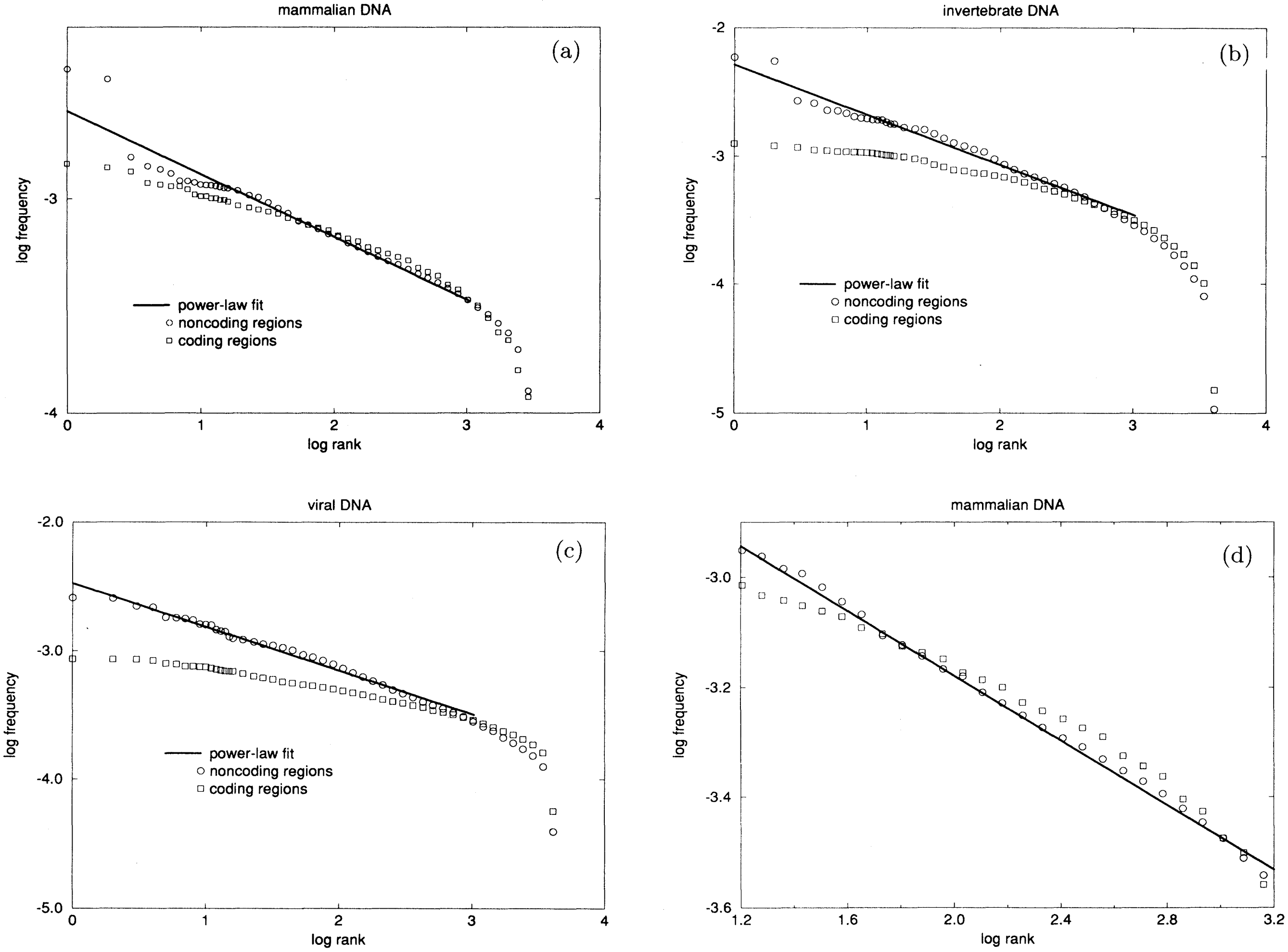

FIG. 2. The log-log plot of the Zipf plot of mammalian, invertebrate, and viral DNA sequences. Coding and noncoding DNA regions are obtained from the 14 mammalian DNA sequences (HSG6PHDH, HSMHCAPG, HUMGHCSA, HUMHBB, HUMHDABCD, HUMHPRTB, HUMMMDBC, HUMNEUROF, HUMRETBLAS, HUMTCRADCV, HUMVITDBP, MMBGCXD, MUSTCRA, and RATCRYG for a total number of 1 078 100 bp and a number of 50 687 coding DNA bp) (a) from 4 invertebrate DNA sequences (CEC07A9, CELTWIMUSC, DROABDB, SCCHRIII, and SCCHRXI for a total number of 1 102 752 bp and a number of 728 998 coding DNA bp) (b) and from 11 viral DNA sequences (ASFV55KB, EBV, HE1CG, HEHCMVCG, HEVZVXX, HS1ULR, HSECOMGEN, HSGEND, IH1CG, VACCG, and VVCGAA for a total number of 1 616 928 bp and a number of 1 361 411 coding DNA bp). (c) The straight line (power-law behavior) is the best fit of the Zipf plot for noncoding DNA sequences. The fitting procedure is performed in the interval $R \leq 1000$. The curvature of the Zipf plot of the coding DNA is evident for the three groups. (d) An enlargement of an important region of (a). The notation "log" denotes $\log_{10}$.

where [23]. The difference observed in the exponent ζ measured in the conventional Zipf and in the n-tuple Zipf analysis may arise with the enlargement of the vocabulary with the n-tuple Zipf analysis: Only real words constitute the vocabulary of conventional Zipf analysis, while all possible n-tuples constitute the vocabulary of the n-tuple Zipf analysis.

In Fig. 1(b) we show the result of an n-tuple Zipf analysis performed on a string of an artificial language. Our text consists of 9 000 000 of bits of a binary executable file of the UNIX computer operating system. In this case the alphabet is composed of only two characters (0,1). To have a set of n-tuples of a size comparable with the one used in the analysis of a natural language and of DNA sequences we report the histogram of the frequency of occurrence of n-tuples with n = 6, 8, 10, and 12. Once more a power-law behavior is observed by analyzing the histogram. In this case, we find $\zeta = -0.77$ for the case $n = 12$.

D. Methods: The investigated DNA ensemble

We studied all 46 DNA sequences longer than 50 000 base pairs present in the GenBank release 81 of 15 February 1994. This includes 14 mammalian DNA sequences, 4 invertebrate DNA sequences, 5 chloroplast DNA sequences, 3 mitochondrial DNA sequences, 7 bacterial DNA sequences, 11 viral DNA sequences, and 1 phage DNA sequence. In this study we consider only eukaryotic and viral DNA. We do not consider prokaryotic (bacteria and phage) DNA (8 sequences) and DNA from organelles (8 sequences of mitochondrion and chloroplast DNA). This choice is motivated by the fact that we are mainly interested in the comparison of the statistical properties

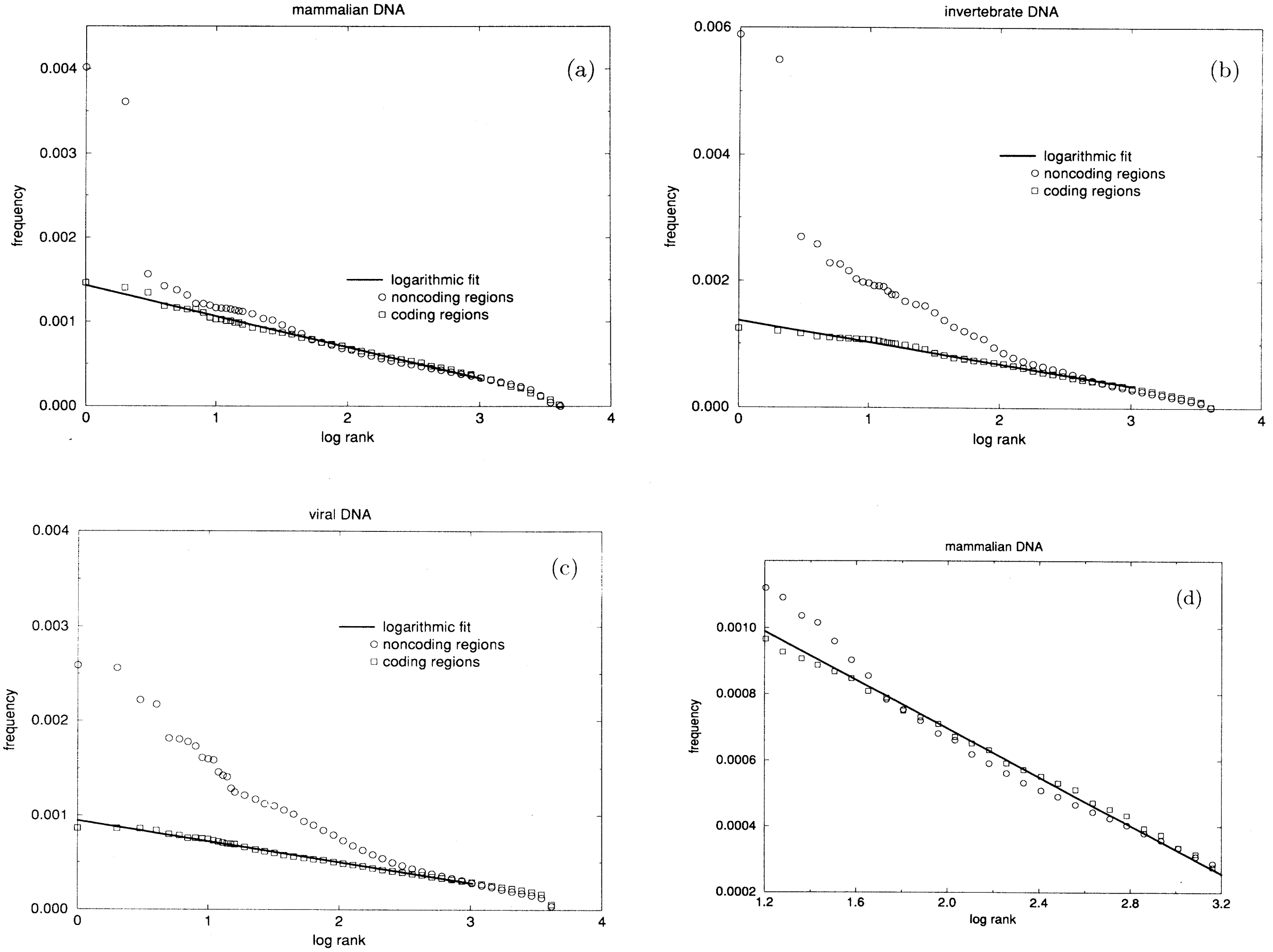

FIG. 3. Semilogarithmic Zipf plot of (a) mammalian, (b) invertebrate, and (c) viral DNA sequences of the results plotted in Figs. 2(a)–2(c). The straight line (logarithmic behavior) is the best fit of the Zipf plot for coding DNA sequences. The fitting procedure is performed in the interval $R \leq 1000$. The curvature of the Zipf plot of the noncoding DNA is evident for the three groups. (d) A blowup of an important region of (a). The notation "log" denotes $\log_{10}$.

observed in the coding and noncoding regions of a complete genome, a complete chromosome, and the longer available DNA sequences. In prokaryotic DNA and in the DNA of organelles, noncoding regions comprise usually only a very small subset of the complete sequence and they may often actually contain still unidentified coding regions. We also investigate the DNA of two chromosomes published after 15 February 1994: the sequence of complete chromosome XI of the yeast (666 448 bp) [24] and the chromosome III of the *C. elegans* (2.2×10^6 bp) [25].

Here we systematically investigate the longest sequences of all eukaryotic and viral DNA available today. In particular, we emphasize the analysis of chromosomes because they are well defined biological units. Unfortunately, the number of complete chromosomes is still limited, but a rapid increase is expected in the near future. A complementary approach, the investigation of the complete GenBank database, has also been performed [21]. The advantage of analyzing the complete GenBank is that one considers all the biological information available. However, the rather severe disadvantage is that the entries in the entire GenBank are an extremely "biased" set since the choice of what to sequence is not done at random.

A "blind" analysis of the entire GenBank database is inevitably affected by the nature of the entries in that database. GenBank is not a random data set, as the entries are heavily biased towards coding regions; indeed, the proportion of coding sequences in the GenBank is much higher than in the genome. Furthermore, there are numerous examples of closely related sequences in GenBank, thus further contributing to bias. Finally, the range of species represented by their entries in GenBank is grossly imbalanced. For example, the peculiarity of the mouse genome (in terms of repetitive elements, codon bias, etc.) weighs heavily on the overall data set. It follows, then, that a blind analysis of GenBank sequences, while giving a superficial impression of "thoroughness" and "completeness," is in fact biologically unsatisfactory.

E. *n*-tuple Zipf analysis of DNA sequences

We perform an n-tuple Zipf analysis of the DNA sequences of our ensemble by counting the occurrence of the set of n-tuples extracted from a source string. The different n-tuples are obtained from the source string by shifting progressively by one base a window of length n. With this method from a source string of length L we obtain $L-n+1$ different substrings. For an alphabet of four characters (A,C,G,T) the number of possible n-tuples is 4^n.

We perform n-tuple Zipf analysis by varying n from $n = 3$ to $n = 7$. At the moment, higher values of n are not analyzable due to the limited length of the published sequences. In fact, the results of n-tuple Zipf analysis are reliable only if $L \gg 4^n$ [26]. In this study we always fulfill the condition $L > 10 \times 4^n$. We note that the same functional form of the n-tuple Zipf plot is observed when a given sequence is analyzed for different values of n (ranging from $n = 3$ to the maximal value $n_{\rm max}$ satisfying $L > 10 \times 4^{n_{\rm max}}$).

An example of the measured n-tuple Zipf plots is given in Fig. 1(c). In this case, the n-tuple Zipf plot shows a power-law behavior for $R \lesssim 4^{n-1}$ and a rapid decrease of the frequency for $R \gtrsim 4^{n-1}$ (a few n-tuples are not present in the 180 000-character sequence). The exponent ζ of the approximate power-law region is roughly the same for different values of n ($\zeta = 0.33, 0.34$, and 0.34 when $n = 4, 5$, and 6, respectively).

Within the present accuracy of the statistical analysis

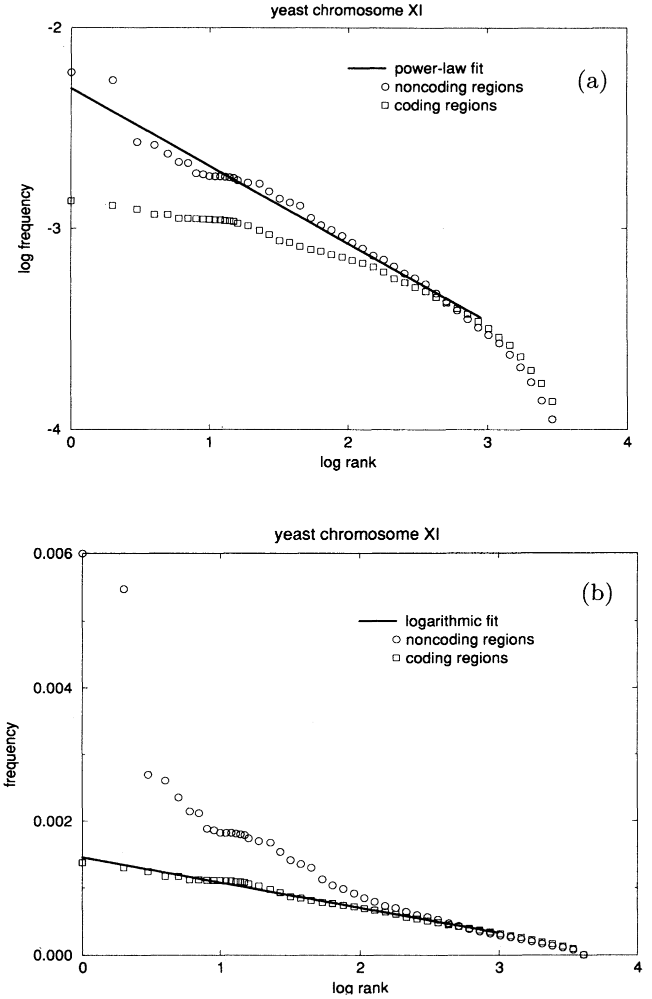

FIG. 4. Log-log plot of the frequency of occurrence of 6-tuples observed in the coding (○) and noncoding (□) regions of the yeast chromosome XI. The straight line is the power-law best fit of the noncoding regions [Eq. (1)] performed in the interval $0 < R \leq 1000$. The best fitting value of the exponent ζ is 0.39. (b) Semilogarithmic plot of the frequency of occurrence for the same set of data shown in (a). The straight line is the logarithmic best fit of the coding regions [Eq. (2)] performed in the interval $0 < R \leq 1000$. The notation "log" denotes $\log_{10}$.

we do not have a reliable way to infer about the possible functional form of the n-tuple Zipf plot for higher values of n. It is also an open question whether the slope ζ of the power-law region is constant or monotonically increasing with n. A detailed study of the scaling properties of ζ is, at the moment, not possible due to the limited length of the published sequences.

In the present study, we focus our attention on the differences of the statistical properties of DNA sequences in coding and noncoding regions. To this end, we separate each long sequence (e.g., yeast chromosome III [27]) into two sequences: the first is obtained by "stitching together" all the known and putative coding regions, whereas the second concatenates the remaining regions of the DNA sequence. The information used to perform the separation is taken from the heading present in almost all the files stored in the GenBank. By using this procedure we obtain separate *coding* and *noncoding* sets of DNA sequences. The results of the analysis of these DNA strings are summarized in Figs. 2 and 3.

We show the results of the n-tuple Zipf analysis performed on the coding and noncoding regions of mammalian DNA [Figs. 2(a) and 3(a), 14 sequences of our ensemble], invertebrate DNA [Figs. 2(b) and 3(b), 4 sequences of our ensemble], and viral DNA [Figs. 2(c) and 3(c), 11 sequences of our ensemble]. For each group of DNA sequences, we show the Zipf plot as a log-log plot (Fig. 2) and as a semilogarithmic plot (Fig. 3).

Our analysis of the selected sequences as well as of the groups of the sequences investigated suggests that the *functional form* of the n-tuple Zipf plot is different in coding and noncoding regions. In the chromosomes investigated the difference is clear and statistically reliable. In particular, the n-tuple Zipf plot of noncoding regions is fitted better by a power law (straight line in a log-log plot) when $R \lesssim 4^{n-1}$ and the n-tuple Zipf plot of coding regions is fitted better by a logarithmic behavior (straight line in a semilogarithmic plot) when $R \lesssim 4^{n-1}$. As an illustrative example, in Figs. 4(a) and 4(b) we show the n-tuple Zipf plots of the coding and noncoding regions of the complete chromosome XI of the yeast *S. cerevisiae* [24]. On the other hand, for vertebrates such as human the difference between Zipf plots of coding and noncoding DNA is less conclusive because the frequencies of the n-tuples are close to each other with the exception of the most frequent and most rare n-tuples.

It is also worth noting that the noncoding DNA sequences of the chromosomes investigated (yeast and *C. elegans*) show a statistical property that is characteristic of the natural and artificial languages: a power-law regime in the Zipf plot of the frequency of occurrence of n-tuples. This finding is not sufficient by itself to prove that the n-tuples in the noncoding regions are words of a given structured genetic language. To find genuine evidence concerning whether or not a hierarchical language (or more than one language) indeed is present in the noncoding DNA a measure of complexity would be needed. Unfortunately, it is well known that a robust measure of complexity is still lacking. Our analysis uses the statistical tools presently available to us.

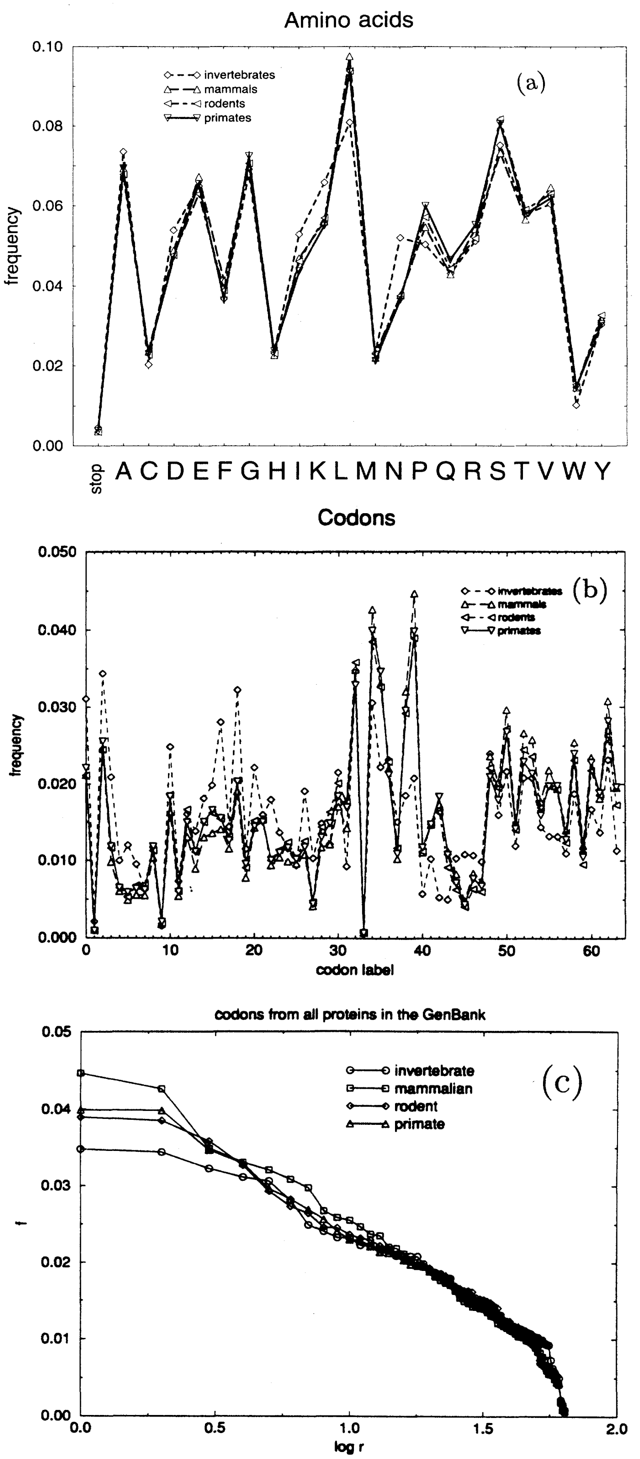

FIG. 5. (a) Relative frequency of occurrence of amino acids using all available GenBank data for each of four categories (invertebrates, rodents, primates, and mammals other than rodents and primates). Note that there are only small differences among categories. (b) Same as (a), except for the 64 3-tuple codons. Note that there are larger differences among categories. (c) Same as (b), except that the 64 codons are arranged in rank order and a semilogarighmic plot is made. Note the wide regime of linear behavior [see Eq. (2)]. The notation "log" denotes $\log_{10}$.

Our analysis suggests (Fig. 3) that in the coding regions, for all sequences studied the Zipf plot is well approximated by

$$f = b - c \log_{10} R \tag{2}$$

for $R \lesssim 4^{n-1}$. This functional form, previously observed for $n = 3$ in a limited number of (relatively) short sequences [28], seems to be also valid for n ranging from 3 to 7 in all the investigated (sufficient long) coding DNA sequences of our ensemble.

This result (2) is consistent with the fact that the frequency of amino acids is conserved in a large ensemble of eukaryotic proteins. In Fig. 5(a) we show the frequency of the 20 amino acids measured in all the invertebrate, mammal, rodent, and primate sequences present in the GenBank (June 1994). We see that the fluctuations between the different groups are small. In contrast, in Fig. 5(b) the measured frequency of codons for the same groups shows larger fluctuations. However, the Zipf analysis [Fig. 5(c)] of the codons measured in the same ensemble shows a similar functional form [i.e., a wide range of logarithmic behavior, see Eq. (2)]. It is interesting to note that the distribution of letters of the alphabets of human languages also follow the functional form of Eq. (2).

F. Low-order Markovian models

A first-order Markovian process describes a symbolic sequence in which the probability of occurrence of a character j is determined only by the previous character i with a probability $p(i,j)$. For a four-letter alphabet (such as DNA), the $p(i,j)$ matrix contains 16 elements.

It is known that in different organisms (and within the same organism in different regions of the same genome) the DNA has different C+G content and different first-order Markovian matrices [i.e., different probabilities $P(i,j)$], see, e.g., [9]. A possible explanation of the difference in functional form observed in the Zipf plot could be due to the differences in the CG content and/or in the Markovian matrices characterizing the investigated sequences and their coding and noncoding regions. To check this hypothesis, we measure for each investigated sequence the experimental first-order Markovian matrix. By using this matrix, we calculate the probability of a given n-tuple under the hypothesis of Markovian process. For example, we calculate the probability to observe the 6-tuple AACTGA by using the relation

$$\begin{aligned} P(AACTGA) = {} & P(A)P(A,A)P(A,C) \\ & \times P(C,T)P(T,G)P(G,A). \end{aligned} \tag{3}$$

For all the analyzed sequences, we compare the measured Zipf plots with the corresponding first-order Markovian process. The results are illustrated by the following example. In Figs. 6(a) and 6(b) we show the Zipf plot measured in the (a) coding, and (b) noncoding regions of the longest DNA sequence available today (chromosome III of *C. elegans*). From the figures the discrepancy between the measured and the first-order Markovian Zipf plot of the (a) coding and (b) noncoding regions is evident. On the basis of the observed results we conclude that a first-order Markovian process cannot fully *explain* the experimental findings we observe in the histograms of the n-tuple Zipf plots for noncoding and intron-rich DNA sequences. For a detailed discussion about the effect of higher-order Markovian processes, see Ref. [21].

G. Role of simple and tandem repeats

A Zipf analysis is affected by the presence of repeats in a symbolic sequence. Repeats are inevitably present in all natural and artificial languages; indeed, they are one of the sources of the redundancy of languages. Re-

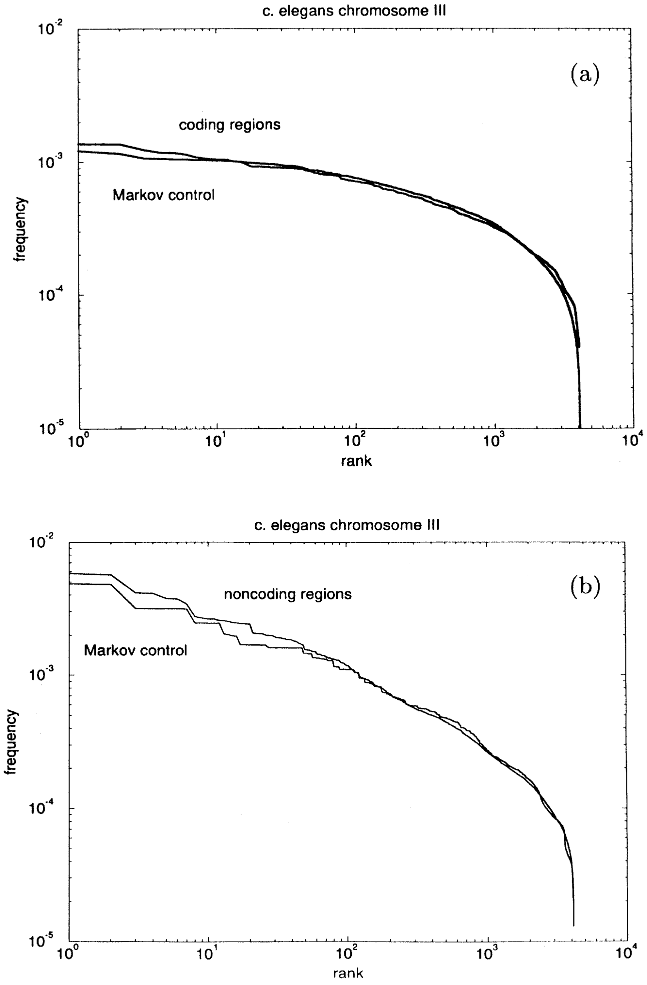

FIG. 6. Comparison between the measured Zipf plot and the corresponding first-order Markovian approximation of the full *C. elegans* chromosome III, (a) coding and (b) noncoding regions. The first-order Markovian approximation is calculated by measuring the first-order Markovian matrix in the studied sequences. The deviation from the first-order Markovian approximation is maximal in noncoding regions.

peated nucleotide sequences (called tandem repeats) are a large part of eukaryotic genomes: in introns and intergenic DNA, one finds many n-tuples that are simple repeats (e.g., poly A- and poly T-tuples) [29]. Both simple and tandem repeats are present in noncoding DNA; of course, simple repeats do not occur in natural languages for $n > 2$.

We perform a test in order to check if simple repeats (AAAA... and TTTT...) can be solely responsible for the features of the Zipf plot observed in noncoding DNA. Specifically, we delete all the simple repeats longer than six characters and then we perform the n-tuple Zipf analysis of the new sequence. The results of our test show that simple repeats strongly affect the most frequent n-tuples (approximately the first decade), but do not affect the wide region of the Zipf plot ranging from $R \approx 10$ to $R \approx 4^{n-1}$ (see Ref. [21] for details). In summary, the Zipf plot features cannot be ascribed solely to the known presence of the long simple repeats present in noncoding DNA.

IV. n-GRAM ENTROPY AND n-GRAM REDUNDANCY OF CODING AND NONCODING DNA

Another statistical measure giving partial information about of the degree of complexity of a symbolic sequence is obtainable by calculating the n-gram entropy of the analyzed text. The Shannon n-gram entropy [11] (i.e., the entropy of the n-tuples observed in a given text string) is defined by

$$H(n) = -\sum_{i=1}^{\lambda^n} p_i \log_2 p_i, \tag{4}$$

where p_i is the probability of the n-tuple labeled by index i and λ is the number of letters of the alphabet.

From the n-gram entropy one can obtain a quantitative measure of the redundancy R present in any text. The redundancy is defined by

$$R \equiv 1 - \lim_{n\to\infty} \frac{H(n)}{kn}, \tag{5}$$

where $k \equiv \log_2 \lambda$. The redundancy is a manifestation of the *flexibility* of the underlying language.

We calculated the Shannon n-gram entropy $H(n)$ for different values of n. The maximum value of n for which it is possible to determine $H(n)$ is $n = 7$—even for very long sequences (e.g., *C. elegans*)—due to the extremely slow convergence to the final value. For shorter sequences, reliable values of $H(n)$ are obtainable only up to a value of n smaller than 7. In this paper, for each DNA sequence of length L, we report entropy values for values of n fulfilling the condition $L > 100 \times 4^n$.

In general, coding and noncoding regions of a given sequence (or group of sequences) are not equal in size. For this reason the maximal value of n for which we calculate $H(n)$ [and then $R(n)$] is in general different for our three parallel analyses of complete, coding, and noncoding se-

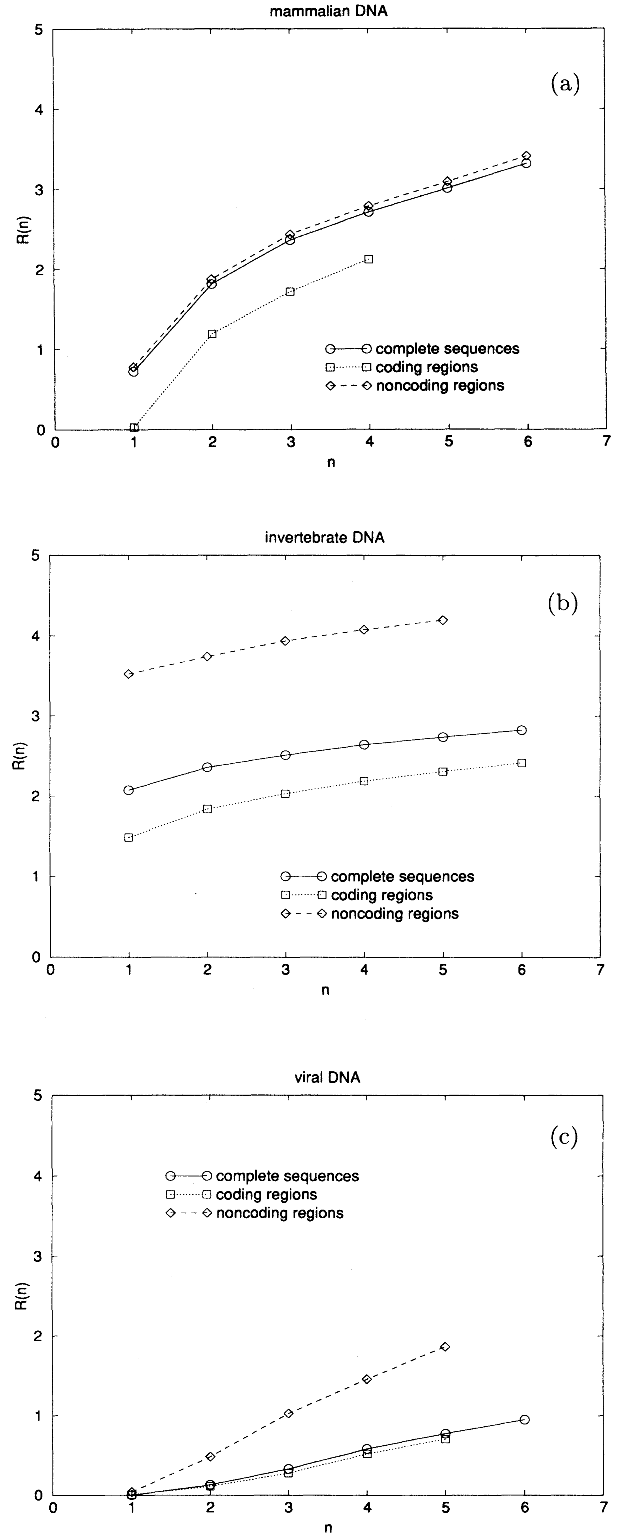

FIG. 7. $R(n)$ as a function of n measured in the three groups of sequences investigated (see the list of GenBank accession code in the caption of Fig. 2): (a) mammalian, (b) invertebrate, and (c) viral DNA. In all the three groups, the n redundancy is higher in noncoding regions and lower in coding regions.

quences.

In Figs. 7(a)–7(c), we show the measured values of the n-gram redundancy, expressed as a percent,

$$R(n) \equiv \left(1 - \frac{H(n)}{2n}\right) \times 100. \qquad (6)$$

The values of $R(n)$ are measured in the complete, coding, and noncoding sequences of (a) mammalian, (b) invertebrate, and (c) viral DNA.

We calculate $H(n)$ and $R(n)$ for all the sequences of our ensemble. We find that in the chromosomes investigated $R(n)$ is significantly higher in noncoding DNA than in coding DNA. A typical example is shown in Fig. 8(a), where we plot the values of $R(n)$ measured in the complete, coding, and noncoding regions of the chromosome III of the *C. elegans*. However, among vertebrate and viral sequences we find many exceptions of this rule. As an example, in Fig. 8(b) we plot the values of $R(n)$ measured in one of such exceptions, the viral DNA sequence HEHCMVCG.

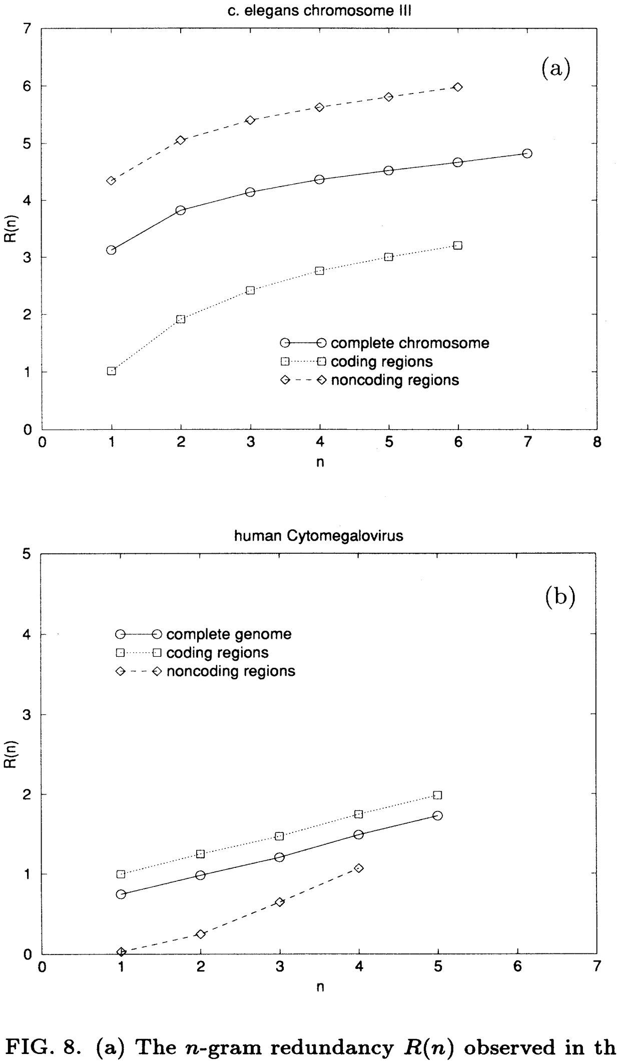

FIG. 8. (a) The n-gram redundancy $R(n)$ observed in the longest DNA sequence today available: chromosome III of the *C. elegans* (2.2 Mbp). Noncoding regions show higher redundancy than coding regions. (b) $R(n)$ for one of the exceptions encountered in our ensemble of DNA sequences. The virus *human Cytomegalovirus Strain AD169* shows a higher redundancy in the coding regions than in the noncoding regions.

It is probably worth pointing out that the severe limitations on the maximal value of n presently reachable by our analysis unavoidably limit this kind of analysis to mainly detect the concentration and the Markovian properties of the DNA sequences. With a maximal value of $n = 7$ it is impossible, for example, to take into account the role of the long (such as short) interspersed repeat sequences, which affects the entropy value (and then the redundancy) of DNA sequences [26]. Our preliminary results suggest that much of the difference in n-gram redundancy between coding and noncoding sequences can be ascribed to the difference in their CG content. A detailed study of the role of the CG content and higher-order Markovian properties on the n-gram redundancy is given in [21].

V. DISCUSSION

The central finding of the present study is our discovery of a difference in the statistical properties of coding vs noncoding sequences of the largest DNA sequences currently available. In particular, by adapting statistical methods developed for the analysis of natural languages and symbolic sequences, we observe that, in coding sequences, the n-tuple Zipf plots can be well approximated by a logarithmic function. In contrast, noncoding regions have qualitatively different Zipf behavior, displaying in the case of the chromosomes investigated and in several other long DNA sequences *power-law scaling* over a wide interval.

These findings raise some intriguing biological speculations about the role of noncoding DNA, since in some cases noncoding DNA exhibits some features found in languages. While the function of the coding regions is well known (coding regions store biological information about the sequence of amino acids present in a protein), the function of the noncoding regions is not well understood. While we must emphasize that our results by no means prove the presence of a "language" in noncoding DNA, we nonetheless note that the hypothesis of the existence of a "genetic language" in the noncoding regions might provide a different interpretative framework for the C paradox and for the observation of a variable coding to noncoding ratio. Under the hypothesis that structured biological information is stored in the noncoding regions, we might conjecture that the overall size of these regions must be related to the phenotypical complexity of the organism. For example, the noncoding regions of vertebrates are longer than in invertebrates and have lower n-gram redundancy. This could mean that the amount of information stored in noncoding regions of vertebrates is larger than in those of invertebrates.

Finally, our present analysis shows that a first-order Markovian process cannot explain the experimental findings observed in the noncoding regions. Higher-order Markovian processes can mimic with increasing accuracy the observed statistical properties [21]. However, by considering the present results together with the results of previous studies performed using different statistical tools [5], we can conclude that the noncoding sequences cannot be described by a Markovian stochastic process.

ACKNOWLEDGMENTS

We wish to thank I. Grosse, H. Herzel, F. Sciortino, and E. N. Trifonov for helpful discussions and the NSF, the NIH, the Mathers Charitable Foundation, NASA, the American Heart Association, and the Israel–U.S. Binational Science Foundation for support.

[1] J. D. Watson, M. Gilman, J. Witkowski, and M. Zoller, *Recombinant DNA* (Scientific American, New York, 1992).

[2] L. L. Sandell and V. A. Zakian, Cell **75**, 729 (1993).

[3] T. G. Kontiris, B. Devlin, D. D. Karp, N. J. Robert, and N. Risch, N. Engl. J. Med. **329**, 517 (1993).

[4] A. M. Lambowitz and M. Belfort, Annu. Rev. Biochem. **62**, 587 (1993).

[5] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992). The difference in long-range correlation properties between coding and noncoding DNA sequences has recently received confirmation from the wavelet analysis of A. Arneodo, E. Bacry, P.V. Graves, and J.F. Mugy, Phys. Rev. Lett. **74**, 3293 (1995), and by an analysis of the entire GenBank database [see, e.g., S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **51**, 5084 (1995)].

[6] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[7] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[8] 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).

[9] E. N. Trifonov, Bull. Math. Biol. **51**, 417 (1989).

[10] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Redwood City, CA, 1949).

[11] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **30**, 50 (1951); see also L. Brillouin, *Science and Information Theory* (Academic, New York, 1956).

[12] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983).

[13] S. Wolfram, Commun. Math. Phys. **96**, 15 (1984).

[14] P. Grassberger, IEEE Trans. Inf. Theory **35**, 669 (1989).

[15] A. Schenkel, J. Zhang, and Y.C. Zhang, Fractals **1**, 47 (1993).

[16] M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, and N. Shnerb, Fractals **2**, 7 (1994); W. Ebeling and A. Neiman, Physica A **215**, 233 (1995); S. Havlin, *ibid.* **216**, 148 (1995).

[17] W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA, 1991).

[18] H. A. Simon, Biometrika **42**, 435 (1955).

[19] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992).

[20] G. D'Alessandro and A. Politi, Phys. Rev. Lett. **64**, 1609 (1990).

[21] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and H. E. Stanley (unpublished).

[22] M. Damashek, Science **267**, 843 (1995).

[23] A. Czirók, R. N. Mantegna, S. Havlin, and H. E. Stanley, Phys. Rev. E **52**, 446 (1995).

[24] B. Dujon *et al.*, Nature **369**, 371 (1994).

[25] R. Wilson *et al.*, Nature **368**, 32 (1994).

[26] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos, Solitons Fractals **4**, 97 (1994); H. Herzel, W. Ebeling, and A. O. Schmitt, Phys. Rev. E **50**, 5061 (1994); H. Herzel and I. Groſse, Physica A **216**, 518 (1995).

[27] S. G. Oliver *et al.*, Nature **357**, 38 (1992).

[28] M. Yu. Borodovsky and S. M. Gusein-Zade, J. Biomol. Struct. Dyn. **6**, 1001 (1989).

[29] A well-known example of a repetitive element that is not a simple repeat is the ALU family, which appears in the human genome.