Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis

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

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

(Received 5 January 1995)

An open question in computational molecular biology is whether long-range correlations are present in both coding and noncoding DNA or only in the latter. To answer this question, we consider all 33 301 coding and all 29 453 noncoding eukaryotic sequences—each of length larger than 512 base pairs (bp)—in the present release of the GenBank to determine whether there is any statistically significant distinction in their long-range correlation properties. Standard fast Fourier transform (FFT) analysis indicates that *coding* sequences have practically no correlations in the range from 10 bp to 100 bp (spectral exponent $\beta = 0.00 \pm 0.04$, where the uncertainty is two standard deviations). In contrast, for *noncoding* sequences, the average value of the spectral exponent β is positive (0.16 ± 0.05) which unambiguously shows the presence of long-range correlations. We also separately analyze the 874 coding and the 1157 noncoding sequences that have more than 4096 bp and find a larger region of power-law behavior. We calculate the probability that these two data sets (coding and noncoding) were drawn from the same distribution and we find that it is less than 10^{-10}. We obtain independent confirmation of these findings using the method of detrended fluctuation analysis (DFA), which is designed to treat sequences with statistical heterogeneity, such as DNA's known mosaic structure ("patchiness") arising from the nonstationarity of nucleotide concentration. The near-perfect agreement between the two independent analysis methods, FFT and DFA, increases the confidence in the reliability of our conclusion.

PACS number(s): 87.10.+e

I. INTRODUCTION

Recently, Peng *et al.* [1] observed long-range power-law correlations of nucleotides in DNA sequences. Applying to 24 different sequences the technique of mapping DNA onto a random walk, they found that the noncoding sequences (introns and intergenic sequences) display long-range correlations while coding sequences do not. Similar observations were reported independently by Li and Kaneko [2], who applied standard Fourier analysis to a sample consisting of seven genes. Subsequently, the observation of long-range power-law correlations was confirmed by Voss [3], who studied all coding and noncoding sequences longer than 512 base pairs (bp) from the entire GenBank using power spectral analysis with subtraction of the white noise level. However, Voss failed to detect any statistically significant difference between long-range correlation properties of coding and noncoding sequences. The goal of the present article is to resolve this discrepancy between the results of Refs. [1,2] and Ref. [3] by answering the the question: Are the long-range correlation properties of coding and noncoding sequences different? This question is important because of its implications for understanding the structure and evolution of DNA [4–9,15], as well as for practical considerations in distinguishing coding and noncoding sequences [10,11].

II. METHODS

To resolve this fundamental issue, we systematically applied two different scaling methods to all coding and noncoding sequences larger than 512 bp in the GenBank release of 15 August 1994. The first method is a standard power spectrum analysis used by Voss [3], but without the ambiguous procedure of subtracting the "white noise" level. The second method is the detrended fluctuation analysis (DFA) method developed by Peng *et al.* [12], which is a modification of a standard rms analysis of a random walk. The advantage of the DFA method over the original analysis method of Peng *et al.* [1] is that without eliminating true power-law correlations, it systematically corrects the results for nonstationarity of nucleotide concentration (so-called DNA patchiness), which may cause spurious correlations [12].

A. Mapping rules

A nucleotide sequence $\{n_i\}$ $(i = 1, 2, ..., L)$ of length L is comprised of the base pairs A (adenine), C (cytosine), T (thymine), and G (guanine). In order to apply numerical methods to a nucleotide sequence, we first prepare seven numerical sequences $\{u_i\}$, corresponding to seven

ways of mapping the original nucleotide sequence onto a one-dimensional numerical sequence.

(i) Purine-pyrimidine (RY) rule. If n_i is a purine (A or G) then $u_i = 1$; if n_i is a pyrimidine (C or T) then $u_i = -1$.

(ii) $A\bar{A}$ rule. If $n_i = A$ then $u_i = 1$; in all other cases $u_i = -1$.

(iii) $T\bar{T}$ rule. If $n_i = T$ then $u_i = 1$; in all other cases $u_i = -1$.

(iv) $G\bar{G}$ rule. If $n_i = G$ then $u_i = 1$; in all other cases $u_i = -1$.

(v) $C\bar{C}$ rule. If $n_i = C$ then $u_i = 1$; in all other cases $u_i = -1$.

(vi) Hydrogen bond energy rule (called the SW rule, [13]). $u_i = 1$ for "strongly bonded" pairs (G or C); $u_i = -1$ for "weakly bonded" pairs (A or T).

(vii) Hybrid rule (called the KM rule, [13]). $u_i = 1$ for A or C; $u_i = -1$ for G or T.

The RY rule has been perhaps the mostly widely used rule, but the other rules have also been applied [1,3,11]. We have also considered other rules: e.g., each base pair can be weighted by any characteristic of that base pair, so u_i can be any number, e.g., molecular mass, hydrophobicity, etc. (see also [2,14,15]).

B. Fast Fourier transformation method

For a fast Fourier transformation (FFT) analysis, we divide each sequence of L nucleotides into $K = [L/N]$ nonoverlapping subsequences of size $N = 512$ starting from the beginning and K nonoverlapping subsequences starting from the end of the sequence. For each subsequence we compute the Fourier transform

$$q_f \equiv \sum_{k=0}^{N-1} u_k \exp(ikf2\pi/N) \tag{1a}$$

and the power spectrum

$$S(f) = |q_f|^2 + |q_{N-f}|^2. \tag{1b}$$

Then we average $S(f)$ over the K subsequences of a given sequence, obtained from starting at one end, and K subsequences starting from the other end.

If a sequence has long-range power-law correlations, then

$$S(f) \sim f^{-\beta} \tag{2}$$

and consequently a log-log plot of $S(f)$ versus f is a straight line with slope $-\beta$. This analysis was performed by Voss [3]. Voss found that for almost all sequences this line is not straight but has a changing slope. In order to make it straighter he applied the procedure called white noise level subtraction. Then he computed β as a slope of a linear fit to an arbitrarily selected part of the resulting data. Voss selected a variable fitting range including in some cases the lowest possible frequencies (as low as 10^{-4} bp). We exclude such low frequencies in our analysis (see the discussion of our procedure below) since the average protein coding sequence is only several hundred bp in length. Also, the procedure used by Voss involves several unknown parameters (including the white noise level as well as the fitting range), which strongly affect the resulting value of β and thereby render problematic any systematic comparison of correlation properties of coding and noncoding sequences.

An alternative approach that we have developed is to identify the physical and biological factors that cause the changes of the slope of the power spectra and then to select a fitting region where the slope of the data is less affected by these factors. It is important that this region should be the same for coding and noncoding sequences and for all groups of organisms. For reasons described below, we select a fitting region from $f = 0.012$ bp^{-1} to $f = 0.097$ bp^{-1} and compute β as the slope of the least-squares linear fit in this region.

We analyze all coding and noncoding sequences of the current GenBank release, subdividing them into large groups of organisms including plants, invertebrates, mammals, rodents, and primates. We restrict our analysis to studies of eukaryotic DNA sequences. In prokaryotes such as bacteria or phages, almost all of the genome is coding and the noncoding regions are often ambiguously identified. For the same reason we exclude from our analysis the sequences of chloroplasts and mitochondria [16].

For each group we compute the weighted average

$$\bar{\beta} \equiv \sum_{i=1}^{M} \beta_i L_i \Big/ \sum_{i=1}^{M} L_i \tag{3a}$$

and the variance

$$\sigma_\beta^2 \equiv \sum_{i=1}^{M} (\beta_i - \bar{\beta})^2 L_i \Big/ \sum_{i=1}^{M} L_i, \tag{3b}$$

where β_i is the value of β for each sequence, M the number of sequences in each group, and L_i the length of each sequence.

C. Detrended fluctuation analysis

In Ref. [1], a "min-max" method was proposed to take into account "nucleotide heterogeneity." A potential drawback of this method is that it requires the investigator to judge how many local maxima and minima of a landscape to utilize in the analysis. In [12] we presented another method—*detrended fluctuation analysis*—that is independent of investigator input and permits the detection of long-range correlations embedded in a patchy landscape and also avoids the spurious detection of apparent long-range correlations that are an artifact of patchiness.

The original DFA method comprises the following steps.

(i) For each numerical sequence $\{u_i\}$ compute a running sum

$$y(n) \equiv \sum_{k=1}^{n} u_k, \qquad [y(0) \equiv 0] \ , \tag{4}$$

which can be presented graphically as a one-dimensional landscape or DNA walk [1].

(ii) Divide the entire sequence of length L into L/ℓ *nonoverlapping* boxes, each containing ℓ nucleotides, and define the "local trend" in each box (proportional to the compositional bias in the box) to be the ordinate of a linear least-squares fit for the DNA walk displacement in that box.

(iii) Define the "detrended walk," denoted by $y_\ell(n)$, as the difference between the original walk $y(n)$ and the local trend. Calculate the variance about the local trend for each box and calculate the average of these variances over all the boxes of size ℓ.

In this work we use the original DFA method, but with a sliding box, in order to obtain better statistics. Specifically, we define a sliding observation box of size ℓ that starts at base pair i and ends at base pair $i+\ell$. Then we compute the least-squares linear fit $y_{i,\ell}(n) = na + b$ such that the sum of $\ell + 1$ squares for this box

$$E_{i,\min}(\ell) \equiv \sum_{n=i}^{i+\ell} [y(n) - y_{i,\ell}(n)]^2 \tag{5}$$

is a minimum.

Finally, we average $E_{i,\min}(\ell)$ over all positions of the observation box from $i = 0$ to $i = L - \ell$ and define the "detrended fluctuation function" as

$$F_{\mathcal{D}}^2 \equiv \frac{1}{(L-\ell+1)(\ell-1)} \sum_{i=0}^{L-\ell} E_{i,\min}(\ell). \tag{6}$$

For sequences with power-law long-range correlations for $\ell > 10$ the detrended fluctuation can be well approximated by a power law [17]

$$F_{\mathcal{D}} \sim \ell^{\alpha} \ , \tag{7}$$

where, for an infinite sequence α is related to β through

$$\alpha = (\beta + 1)/2. \tag{8}$$

For each nucleotide sequence and for each of the seven mapping rules, we compute α by fitting the double logarithmic plot of $F_{\mathcal{D}}(\ell)$ in the range $\ell = 10 - 100$, which approximately corresponds to the range of frequencies for identifying β. According to the theoretical relationship (8) between α and β, we calculate for each sequence an exponent $\beta' \equiv 2\alpha - 1$, which we compare with the spectral exponent β, obtained by the FFT method. We compute the average $\bar{\beta}'$ and the standard deviation $\sigma^2_{\beta'}$ using Eq. (3), for the same groups of organisms as above.

III. RESULTS

A. FFT analysis

We analyze sequences for all seven mapping rules. We show in Fig. 1(a) a log-log plot of the averaged power spectra for all 33 301 coding and for all 29 453 noncoding sequences of the GenBank larger than 512 bp. We find that various large subgroups such as plants, invertebrates, and primates have similar qualitative pictures, but have slightly different slopes. We note in Fig. 1 the presence of three spectral regimes denoted H, L, and M, corresponding, respectively, to relatively high-frequency, low-frequency, and mid-frequency scaling regions.

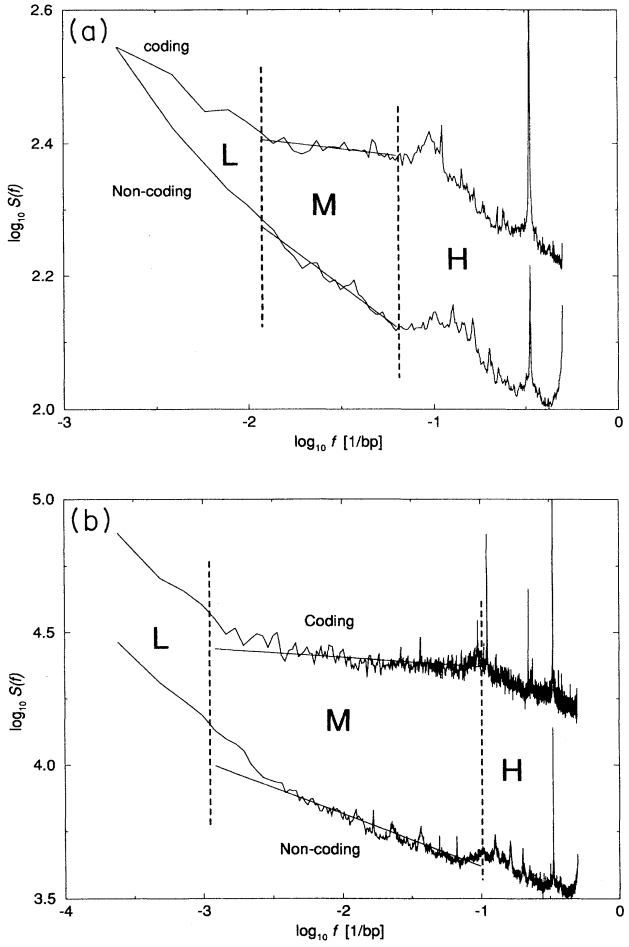

FIG. 1. (a) Power spectra averaged over all eukaryotic sequences longer than 512 bp, obtained by a FFT with a window size of 512. The upper curve is the average over 29 453 coding sequences; the lower curve is the average over 33 301 noncoding sequences. The straight lines are least-squares fits for the second decade (region M). The values of β measured as the slopes of the fits are 0.03 and 0.21, respectively. (b) Same data for all sequences larger than 4096 bp, obtained by a FFT with a window size of 4096. The average is computed over 874 coding and 1157 noncoding sequences. Note that for high frequencies, the power spectra for both window sizes practically coincide. In the region of frequencies $f < 1/100$ bp^{-1} [region H of (a)], the power spectra in (a) bend upward from the apparent straight line. For (b) (larger windows) the $S(f)$ spectra have a constant slope over more than one decade (region M). The fits are the same for both (a) and (b): for coding, $\beta = 0.04$, while for noncoding, $\beta = 0.21$.

1. High-frequency range: Regime H

In the region of high frequencies, there are two major peaks, corresponding to the frequencies $1/3$ bp^{-1} and $1/9$ bp^{-1} [3]. These peaks are much more pronounced in coding than in noncoding sequences and are probably related to the codon structure, which consists of three nucleotides [18]. It is interesting to note that in the noncoding sequences the peaks are still present, but are much weaker than in the coding sequences. This may support the hypothesis (see [19,20] and references therein) that many noncoding sequences are the result of the insertion of formerly coding sequences that later mutate.

Since the high-frequency region ($f > 1/10$ bp^{-1}) is strongly affected by short-range correlations related to the codon structure, it cannot be used for investigating long-range correlations. In an attempt to obviate this problem, Voss subtracted a "white noise level" from power spectra, which he had to estimate subjectively [21].

2. Low-frequency range: Regime L

At the lowest frequencies, the signal is distorted by artifacts of the Fourier transformation method. Specifically, at frequencies smaller than roughly $5/N$, there is a spurious contribution arising from the fact that in the FFT method data that are not periodic are treated as periodic with period N. As a result, there is a contribution to $S(f)$ that varies as $1/f^2$. For this reason, we show in Fig. 1(b) our FFT results for the same data set but using a larger window ($N = 4096$). We note that the spurious behavior is shifted to even lower frequencies [Fig. 1(b)].

3. Mid-frequency range: Regime M

The only region for which a log-log plot of $S(f)$ vs f is consistently straight for both coding and noncoding sequences is the intermediate region; for $N = 512$,

TABLE I. The top line gives the average value of β, measured by a FFT for all eukaryotic DNA noncoding segments larger than 512 bp for all seven mapping rules defined in Sec. II of the text. The fitting range for β is from $f = 0.012$ bp^{-1} to $f = 0.097$ bp^{-1}. The bottom line gives the value of $\beta' \equiv 2\alpha - 1$, where α is computed by the DFA method (the fitting range is from $\ell = 10$ to $\ell = 100$). The standard errors of β and β' are all less than 0.01. The asterisk denotes all other mammals except primates and rodents.

			Seven rules of mapping						
Group	Segments	length (kbp)	RY	$A\bar{A}$	$T\bar{T}$	$G\bar{G}$	$C\bar{C}$	SW	KM
I. Plants	6843	8155	0.14	0.13	0.12	0.05	0.06	0.03	0.11
			0.13	0.12	0.11	0.05	0.06	0.03	0.09
1. Fungi	3222	3459	0.14	0.13	0.12	0.02	0.03	0.01	0.08
			0.15	0.13	0.12	0.03	0.04	0.02	0.08
2. Embryophytes	2632	3418	0.15	0.13	0.12	0.08	0.08	0.04	0.12
			0.13	0.11	0.10	0.06	0.07	0.03	0.10
II. Invertebrates	5691	8893	0.09	0.10	0.09	0.08	0.06	0.08	0.09
			0.11	0.12	0.11	0.08	0.07	0.08	0.10
1. Insects	2573	3994	0.08	0.13	0.10	0.09	0.06	0.11	0.11
			0.09	0.13	0.11	0.08	0.06	0.11	0.11
(a) Drosophila	1627	2687	0.10	0.13	0.11	0.09	0.08	0.10	0.12
			0.11	0.13	0.13	0.08	0.08	0.10	0.12
2. Nematodes	1363	2602	0.04	0.04	0.03	0.05	0.04	0.05	0.03
			0.13	0.11	0.11	0.07	0.07	0.07	0.08
3. Protozoa	1144	1553	0.15	0.14	0.14	0.14	0.09	0.05	0.14
			0.14	0.13	0.13	0.08	0.09	0.06	0.13
III. Vertebrates	20638	29198	0.20	0.15	0.14	0.13	0.14	0.12	0.10
			0.18	0.13	0.12	0.12	0.12	0.07	0.11
1. Fishes and amphibians	1046	1299	0.17	0.13	0.12	0.08	0.08	0.05	0.10
			0.15	0.11	0.12	0.08	0.09	0.05	0.09
2. Birds	1060	1418	0.21	0.16	0.14	0.11	0.12	0.05	0.12
			0.20	0.14	0.13	0.09	0.10	0.03	0.10
3. Mammals	18661	26685	0.20	0.15	0.14	0.13	0.14	0.13	0.10
			0.18	0.13	0.12	0.12	0.12	0.07	0.11
(a) Rodents	7073	9837	0.19	0.16	0.16	0.12	0.13	0.12	0.12
			0.18	0.14	0.12	0.11	0.12	0.06	0.12
(b) Primates	9801	14646	0.21	0.14	0.13	0.15	0.15	0.14	0.09
			0.18	0.13	0.12	0.13	0.13	0.08	0.11
(b) Others*	1787	2201	0.18	0.14	0.14	0.09	0.10	0.10	0.09
			0.17	0.11	0.12	0.09	0.10	0.05	0.08
Total	33301	46449	0.16	0.14	0.13	0.11	0.11	0.10	0.10
			0.16	0.12	0.12	0.10	0.10	0.06	0.10

this regime is roughly the decade $1/100\ \text{bp}^{-1} \leq f \leq 1/10\ \text{bp}^{-1}$. Therefore, for this region, the slopes can be used reliably to test for power-law correlations. We also note that the size of a protein is usually limited to a few hundred amino acids, so the coding region rarely exceeds $\approx 10^3$ bp. Consequently, comparison of the correlation properties of coding and noncoding sequences is valid only for length scales below $\approx 10^3$ bp. In fact, a majority of the coding sequences are smaller than 10^3 bp. In order to obtain good statistics and a consistent procedure for all sequences analyzed, we choose $N = 512$ bp and compare long-range correlation properties of coding and noncoding sequences only on the second decade of the power spectra, fitting the data from $f = 0.012\ \text{bp}^{-1}$ to $f = 0.097\ \text{bp}^{-1}$.

We find that for the eukaryotic sequences for each of the mapping rules the average value of β is significantly smaller for coding sequences than for noncoding. The value of β is very close to zero for coding sequences, indicating almost no correlations in the region of 1/10–1/100 bp. The results are summarized in Tables I and II and in Fig. 2. The histograms of distributions of β for coding and noncoding sequences for several groups of organisms are presented in Fig. 3. The width of these histograms is equal to σ_β. The size of σ_β relates both to intrinsic biological variability and to *inherent* errors in estimating scaling exponents from finite size sequences as discussed in Ref. [22].

The probability distribution function of β for coding is very different from noncoding (see Fig. 2). We perform a quantitative calculation using the Kolmogorov-Smirnov D test to reject the null hypothesis that these two distributions are drawn from the same population distribution [23]. The Kolmogorov-Smirnov D value is defined as the maximal deviation between two cumulative probability distributions and ranges from 0 to 1. A large D value suggests that we can reject the null hypothesis. For our data, $D = 0.35$. Finally, we calculate the significance (p

TABLE II. The top line gives the average value of β, measured by a FFT for all eukaryotic DNA coding segments larger than 512 bp for all seven rules of mapping defined in the text. The fitting range for β is from $f = 0.012\ \text{bp}^{-1}$ to $f = 0.097\ \text{bp}^{-1}$. The bottom line gives the value of $\beta' \equiv 2\alpha - 1$, where α is computed by the DFA method (the fitting range is from $\ell = 10$ to $\ell = 100$). The standard errors of β and β' are all less than 0.01. The asterisk denotes all other mammals except primates and rodents.

			Seven rules of mapping						
Group	Segments	length (kbp)	RY	$A\bar{A}$	$T\bar{T}$	$G\bar{G}$	$C\bar{C}$	SW	KM
I. Plants	8625	12256	-0.02	-0.02	-0.06	-0.05	-0.03	-0.11	-0.02
			0.03	0.01	-0.01	-0.03	-0.02	-0.10	0.01
1. Fungi	3222	3459	-0.04	-0.04	-0.08	-0.04	-0.04	-0.10	-0.03
			0.02	0.01	-0.01	-0.03	-0.02	-0.09	0.02
2. Embryophytes	2632	3418	0.00	0.00	-0.04	-0.07	-0.03	-0.14	-0.01
			0.05	0.02	-0.01	-0.03	-0.01	-0.13	0.02
II. Invertebrates	4681	7439	-0.02	-0.03	-0.06	-0.03	-0.02	-0.08	-0.02
			0.01	-0.01	-0.02	-0.03	-0.03	-0.09	0.00
1. Insects	1887	3004	-0.03	0.02	-0.06	-0.04	-0.07	-0.09	-0.02
			-0.02	-0.02	-0.04	-0.03	0.06	-0.09	-0.01
(a) Drosophila	1187	2171	-0.02	-0.01	-0.07	-0.03	-0.07	-0.10	-0.02
			-0.01	-0.01	-0.03	-0.03	-0.06	-0.09	-0.01
2. Nematodes	866	1587	-0.05	-0.06	-0.05	-0.02	-0.01	-0.06	-0.01
			0.04	-0.01	0.00	-0.03	-0.01	-0.10	0.00
3. Protozoa	1253	1909	0.01	-0.01	-0.05	-0.02	0.02	-0.06	-0.03
			0.03	0.00	-0.01	-0.03	0.00	-0.08	0.00
III. Vertebrates	15985	23283	0.02	0.04	-0.01	0.02	0.03	0.01	0.02
			0.07	0.04	0.01	0.01	0.02	-0.05	0.02
1. Fishes and amphibians	1164	1475	0.03	0.01	-0.02	0.01	0.05	-0.04	0.03
			0.07	0.03	0.00	0.00	0.03	-0.08	0.03
2. Birds	796	1244	0.06	0.01	0.00	0.04	0.05	0.00	0.02
			0.09	0.04	0.01	0.01	0.04	-0.06	0.02
3. Mammals	14187	20773	0.02	0.04	-0.01	0.02	0.03	0.02	0.01
			0.06	0.04	0.01	0.01	0.02	-0.04	0.02
(a) Rodents	6006	8694	0.02	0.03	-0.01	0.01	0.03	0.01	0.01
			0.07	0.04	0.01	0.01	0.02	-0.05	0.02
(b) Primates	6319	9545	0.06	0.04	0.01	0.01	0.02	-0.04	0.02
			0.02	0.04	-0.01	0.01	0.02	0.02	0.01
(c) Others*	1862	2532	0.03	0.05	-0.00	0.04	0.03	0.05	0.00
			0.06	0.05	0.01	0.01	0.01	-0.03	0.01
Total	29453	43185	0.00	0.01	-0.03	-0.01	-0.00	-0.04	-0.00
			0.04	0.04	0.00	0.00	0.00	-0.08	0.02

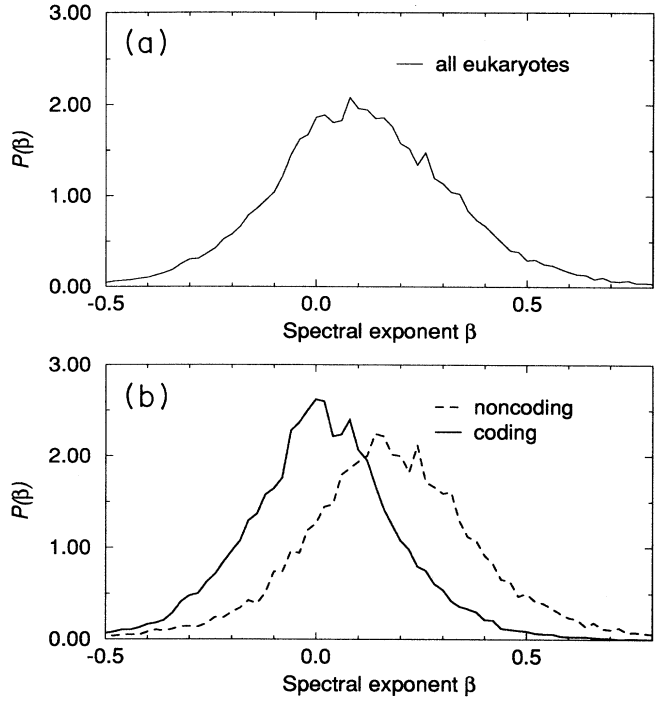

FIG. 2. Histograms of the values of β obtained for all coding and noncoding eukaryotic sequences > 512 bp of the GenBank by the FFT method. The probability density of finding a value of β between β and $\beta + d\beta$ is plotted against β (purine-pyrimidine rule of mapping). (a) Coding and noncoding sequences combined together, (b) coding and noncoding sequences analyzed separately.

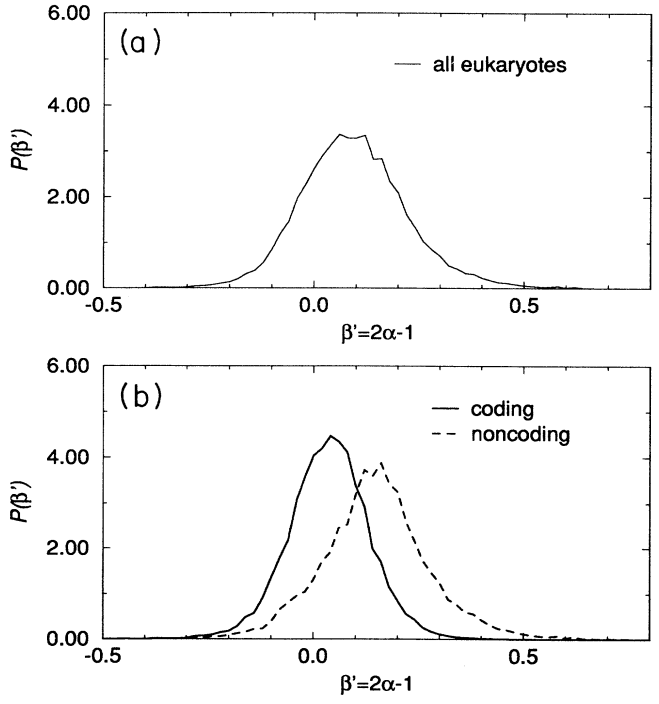

FIG. 3. This figure is the analog of Fig. 2, except $\beta' \equiv 2\alpha - 1$ is calculated by the DFA method of Ref. [12].

value), which gives the probability that these two data sets were drawn from the same distribution, and find that it is less than 10^{-10}.

B. DFA analysis

In order to obtain a completely independent check on the Fourier analysis presented above, we treat the same set of coding and noncoding sequences using the DFA method [12]. The results of both methods are presented in Tables I and II. We found that the values of β and $\beta' \equiv 2\alpha - 1$ are remarkably close to each other. For coding sequences the values of β' are usually slightly larger than the corresponding values of β. This difference emerges from slight differences in the position of the fitting range in real space and in frequency space. The standard deviation $\sigma_{\beta'}$ is also slightly smaller than corresponding values of σ_β. Hence the accuracy of both methods is almost the same. The numerical values obtained by the two methods, FFT and DFA, are remarkably close, as is clear upon examination of the bottom two lines of Tables I and II. Moreover, the histogram $P(\beta')$ of Fig. 3 closely resembles the histogram $P(\beta)$ of Fig. 2.

In order to estimate the standard error of the mean for coding and noncoding sequences (using both FFT and DFA methods), we compare the results for different groups of organisms and different rules of mapping given in Tables I and II and compute their mean and standard deviation. We conclude that for all coding sequences $\bar{\beta} = 0.00 \pm 0.04$ and for all noncoding sequences $\bar{\beta} = 0.16 \pm 0.05$ where the error bars represent 95% confidence intervals [24].

The data in Tables I and II indicate that the value of β (or β') computed by the RY rule for noncoding sequences for any species of vertebrates is significantly larger than the value of β (or β') for any species of invertebrates and plants and is largest for mammals and birds.

IV. DISCUSSION

The results of the present systematic and inclusive analysis of GenBank DNA sequences are notable for two major reasons. First, we unambiguously demonstrate that noncoding DNA, but not coding DNA, possesses long-range correlations [25]. This finding is made using two independent, complementary techniques: Fourier analysis and DFA, a modification of root-mean-square analysis of random walks. Indeed, as shown in Tables I and II, the spectral exponent β computed by both techniques for the same sequence is nearly identical. Second, we demonstrate an increase in the complexity of the noncoding DNA sequences with evolution. The value of β for vertebrates is significantly greater than that for invertebrates. This finding based on the full GenBank data set supports the suggestion based upon a systematic study of the myosin heavy gene family that there is an apparent increase in the complexity of noncoding DNA for more highly evolved species compared to less evolved ones [19]. Both of these results contradict the report of Voss, who

failed to observe any difference in the long-range correlation properties of coding and noncoding DNA and who reported a decrease in the value of the spectral exponent β with evolution.

From a practical viewpoint, the statistically significant difference in long-range power-law correlations between coding and noncoding DNA regions that we observe supports the development of gene finding algorithms based on these distinct scaling properties. A recently reported algorithm of this kind [11] is especially useful in the analysis of DNA sequences with relatively long coding regions, such as those in yeast chromosome III.

Finally, we note that although the scaling exponents α and β have potential use in quantifying changes in genome complexity with evolution (see [19]), the current GenBank database does not allow us to address the important question of whether unique values of these exponents can be assigned to different species or to related groups of organisms. At present, the GenBank data have been collected such that particular organisms tend to be represented more frequently than others. For example, about 80% of the sequences from birds are from *Gallus gallus* (the chicken) and about 2/3 of the insect sequences are from *Drosophila melanogaster*. The results of the present analysis and other recent studies [19] indicate the importance of sequencing not only coding, but also noncoding DNA from a wider variety of species.

ACKNOWLEDGMENTS

We wish to thank F. Sciortino and E. N. Trifonov for discussions, and NSF, NIH, the Mathers Charitable Foundation, and the Israel—USA Binational Science Foundation for support.

[1] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).

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

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

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

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

[6] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, and H. E. Stanley, in *Fractals in Science*, edited by A. Bunde and S. Havlin (Springer, Berlin, 1994), Chap. 2.

[7] A. Yu. Grosberg, Y. Rabin, S. Havlin, and A. Nir, Europhys. Lett. **23**, 373 (1993).

[8] E. I. Shakhnovich and A. M. Gutin, Nature **346**, 773 (1990).

[9] 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), and unpublished. Recent work suggests that the conclusions of this article may hold for the entire GenBank for plants and invertebrates, but not for higher forms of life. See H. E. Stanley *et al.*, Nuovo Cimento (to be published); S. Havlin *et al.*, Fractals (to be published).

[10] J. W. Fickett and C.-S.Tung, Nucleic Acids Res. **20**, 6441 (1992).

[11] S. M. Ossadnik, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and H. E. Stanley, Biophys. J. **67**, 64 (1994).

[12] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[13] M. Ya Azbel, Phys. Rev. Lett. **31**, 589 (1973); Biopolymers **21**, 1687 (1982). See also E. N. Trifonov, Bull. Math. Bio. **51**, 417 (1989).

[14] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, Phys. Rev. A **45**, 8902 (1992).

[15] A. S. Borovik, A. Yu. Grosberg, and M. D. Frank Kamenezki, J. Biomol. Struct. Dyn. **12**, 655 (1994).

[16] We do not tabulate our findings for prokaryotic sequences since these have a remarkably small fraction of noncoding regions (many of which may actually contain unidentified coding regions). However, our analysis of the entire prokaryotic sequence data shows $\beta = 0.00 \pm 0.05$, as expected for primarily coding regions.

[17] It can be proven that for a sequence of uncorrelated random variables $\{u_i\}$ with finite mean $\overline{u}$ and variance V, the detrended fluctuation function, defined by Eq. (6), obeys the exact relation $F_{\mathcal{D}}^2(\ell) = (\ell+3)V/15$ for any value of $\ell \geq 2$. For an artificial sequence with built-in long-range correlations [12], $F_{\mathcal{D}}(\ell)$ can be well approximated by $C(\ell+3)^{\alpha}$, where α is the exponent of long-range correlations and C is some proportionality coefficient. Thus the exponent α can be accurately measured as the slope of a double logarithmic plot of $F_{\mathcal{D}}(\ell)$ vs $\ell+3$ even for very small values of ℓ. We found that this procedure was more reliable for calculating α.

[18] It is well known that there are some preferences for the usage of a purine as the first nucleotide in the codon (32% G and 28% A) and for a weakly bonded base pair as the second (31% A and 27% T) that are quite robust across the entire phylogenetic spectrum. These preferences may create the three base-pair periodicity in the coding nucleotide sequence, which is responsible for the peak at $1/3$ bp^{-1}.

[19] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993).

[20] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[21] The white-noise subtraction procedure employed by Voss is commonly used to separate a useful signal from thermal high frequency fluctuations (e.g., in a radio device) when the white noise arises from known causes. However, for DNA sequences, there is no reason to expect white noise from such sources; indeed, the high frequencies carry biological information—such as the triplet codon ($f = 1/3$ bp^{-1})—so their subtraction is of questionable validity.

[22] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730

(1993).

[23] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, 1989).

[24] For any given species and for any given rule of mapping, the difference between the value of β for noncoding sequences and the value of β for coding sequences is always significantly larger than zero. The largest value of this difference usually corresponds to the purine-pyrimidine (RY) mapping rule.

[25] Small correlations or anticorrelations observed in coding sequences are in quantitative agreement with recent results of V. Pande, A. Yu. Grosberg, and T. Tanaka [Proc. Natl. Acad. Sci. U.S.A. **91**, 12 972 (1994)], who study amino acid sequences of proteins by a similar method.