Quantification of DNA Patchiness Using Long-Range Correlation Measures

G. M. Viswanathan,* S. V. Buldyrev,* S. Havlin,*# and H. E. Stanley*
*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215 USA, and
#Gonda-Goldschmied Center and Department of Physics, Bar Ilan University, Ramat Gan, Israel

ABSTRACT We introduce and develop new techniques to quantify DNA patchiness, and to quantify characteristics of its mosaic structure. These techniques, which involve calculating two functions, $\alpha(\ell)$ and $\beta(\ell)$, measure correlations at length scale ℓ and detect distinct characteristic patch sizes embedded in scale-invariant patch size distributions. Using these new methods, we address a number of issues relating to the mosaic structure of genomic DNA. We find several distinct characteristic patch sizes in certain genomic sequences, and compare, contrast, and quantify the correlation properties of different sequences, including a number of yeast, human, and prokaryotic sequences. We exclude the possibility that the correlation properties and the known mosaic structure of DNA can be explained either by simple Markov processes or by tandem repeats of dinucleotides. We find that the distinct patch sizes in all 16 yeast chromosomes are similar. Furthermore, we test the hypothesis that, for yeast, patchiness is caused by the alternation of coding and noncoding regions, and the hypothesis that in human sequences patchiness is related to repetitive sequences. We find that, by themselves, neither the alternation of coding and noncoding regions, nor repetitive sequences, can fully explain the long-range correlation properties of DNA.

INTRODUCTION

It is well known that DNA polymer sequences have a mosaic structure, which is characterized by "patches" with an excess of one type of nucleotide (Bernardi et al., 1985; Churchill, 1989; Fickett et al., 1992). Patchiness is usually associated in the biological literature with the phenomenon of isochores, which are DNA regions having homogeneous base compositions and typical scales of about 1 Mbp (Bernardi, 1989). Here we extend the concept of patchiness to include nonuniformities on scales smaller than 1 Mbp. Specifically, we define a DNA sequence to be patchy if there are fluctuations in the local nucleotide concentrations that significantly depart from Gaussian statistics, i.e., if the fluctuations grow faster than $\sqrt{l}$ for a subsequence of size l. In contrast, a single patch is a subsequence in which the fluctuations stay within the limits predicted by Gaussian statistics.

Consider the following illustrative example. Let a long DNA sequence consisting of bases A (adenine), C (cytosine), T (thymine), and G (guanine) have overall probabilities for finding the nucleotides P_A, P_C, P_T, and P_G. We now take a smaller subsequence of length ℓ and define $N_A(\ell)$ as the number of occurrences of nucleotide A in the subsequence. If the nucleotides are uncorrelated, the probability distribution of each nucleotide is given by Gaussian statistics, so for A we expect $N_A(\ell) = \ell P_A \pm \sigma$, where the fluctuation is $\sigma = \sqrt{P_A(1 - P_A)\ell}$ (Azbel et al., 1982). Similar conditions hold for the other three nucleotides. DNA sequences that can be described in this way are not patchy. On the other hand, if N_A consistently differs from the expected value by more than one standard deviation, then the sequence is considered patchy, i.e., the nucleotides are not uniformly distributed throughout the sequence, but rather are organized into patches or domains of significantly higher relative concentrations.

It is found that for many DNA sequences the fluctuation σ grows not with the square root of ℓ, but as another power law: $\sigma \approx \ell^{\alpha}$, where $\alpha \neq 1/2$ is a scaling exponent that describes the "roughness" (Barabási and Stanley, 1995) of the fluctuations (Arneodo et al., 1995; Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992). In such cases the DNA sequences are not only patchy, but have patches of all length scales, i.e., there exists no characteristic patch size, because power law behavior is the signature of scale invariance (Stanley, 1971, 1995). The basic premise behind our newly developed methods is that deviations from power-law behavior can be related to characteristic scales. For example, if σ for a given sequence is characterized by two different exponents in two distinct regimes, $\ell < \ell_0$ and $\ell > \ell_0$, with a cross-over occurring at $\ell \approx \ell_0$, then for such a sequence ℓ_0 is the only characteristic scale for the system apart from factors close to unity. There may be patches on all scales, but if the sequence is scale-free in each of the two distinct regimes, $\ell < \ell_0$ and $\ell > \ell_0$, then the only meaningful scale is related to ℓ_0.

The degree to which the mosaic structure of DNA is related to such long-range correlation properties of DNA sequences has been discussed (Nee, 1992; Karlin and Brendel, 1993; Munson et al., 1992; Buldyrev et al., 1993a; Peng et al., 1994). Recently attempts have been made to identify patches using segmentation algorithms based on entropy measures (Bernaola-Galvan et al., 1996) and to study their

Received for publication 7 March 1996 and in final form 5 November 1996.
Address reprint requests to Dr. G. M. Viswanathan, Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215. Tel.: 617-353-9460; 617-353-9393; E-mail: viswan@physics.bu.edu.

0006-3495/97/02/866/10 $2.00

size distributions. However, long-range correlation measures have never been used to quantify patchiness or to identify characteristic patch sizes in DNA sequences. Fourier transform methods and random-walk analysis are two major approaches to studying fluctuation phenomena and long-range power-law correlations in statistical physics, and are well suited to the study of fluctuation phenomena in biological systems (Pickover, 1984; Peng et al., 1992; Ossadnik et al., 1994; Buldyrev et al., 1995; Viswanathan et al., 1996). We therefore adapt and extend these concepts for studying patchiness in DNA by developing techniques to quantify departures from power-law behavior and to estimate distinct characteristic DNA patch sizes embedded in power-law distributions of patch sizes. The detrended fluctuation random walk analysis that we employ here is a step forward compared to previous studies (Azbel et al., 1982; Peng et al., 1992), which are sometimes prone to producing spurious results for highly nonstationary data.

The goals of this work are:

1. To develop techniques to quantify patchiness and the correlation properties of genomic DNA. These techniques involve the calculation of two functions, $\alpha(\ell)$ and $\beta(\ell)$, which measure correlations at length scale ℓ and detect distinct characteristic patch sizes embedded in scale-invariant domain size distributions. For ideal power law correlations, the two functions are related by $\alpha(\ell) = [1 + \beta(\ell)]/2 =$ constant (Voss, 1992; Buldyrev et al., 1995).

2. To investigate the existence of a hierarchy of characteristic patch sizes in long genomic sequences, and to compare, contrast, and quantify the correlation properties of different sequences, including the yeast genome, and several long human and prokaryotic sequences.

3. To examine possible explanations for our findings, and to test a number of hypotheses concerning the origin of the known long-range correlation properties and mosaic structure of DNA.

METHODS AND CONTROLS

To apply numerical methods to a DNA sequence $\{n_i\}$ consisting of the four nucleotides A, C, T, and G, we generate a binary sequence $\{u_i\}$ for each DNA sequence (Gates, 1986; Buldyrev et al., 1995). We use the following three binary mapping rules:

1. 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 $u_i = -1$.

2. Hydrogen bond energy (SW) rule (Azbel, 1973; Azbel et al., 1982). For strongly bonded pairs (G or C), $u_i = 1$, whereas for weakly bonded pairs (A or T), $u_i = -1$.

3. Hybrid (KM) rule. For A and C $u_i = 1$, whereas for T and G $u_i = -1$.

Each of these rules probes a different aspect of the mosaic structure of DNA, e.g., the SW rule is related to the energy balance of strand separation, and the RY rule is related to strand chemical bias. Similar rules can be applied for single nucleotides (A, C, G, and T) as well as to dinucleotides such as CG or even longer nucleotide patterns (Karlin et al., 1993). We choose the above three mapping rules because they are the only three ways to map four nucleotides onto equal-sized binary bins.

First we develop techniques for detecting and examining characteristic scales of patchiness by studying a "control sequence" of +1 and −1 with patches of three different characteristic scales. The control sequence is constructed by concatenating uncorrelated patches of fixed sizes of 200 bp, 2000 bp, and 20,000 bp. For each patch j of length L_j, we randomly assign $P_j(1)$, the concentration of "+1", to be either $P_j(1) = 0.3$ or $P_j(1) = 0.7$ with equal probability, i.e., each patch has randomly assigned biases $b \equiv P_j(1) - P_j(-1) = \pm 0.40$. Then we concatenate these patches to make a sequence of length $N = 10^6$ bp or more. We generate distinct characteristic patch sizes embedded in a scale-invariant distribution of patch sizes by choosing the smallest patch size with the highest probability and the largest patch size with the smallest probability. Specifically, we use the following rule for generating long-range correlations (Shlesinger and Klafter, 1986). For the patch j,

1. A random number x_j is chosen in the interval [0, 1].

2. A preliminary length quantity ℓ_j is computed as $\ell_j = 200/x_j$.

3. If ℓ_j is less than 2000, then a patch of size $L_j = 200$ bp is chosen. Otherwise, if ℓ_j is less than 20,000 then a patch of size $L_j = 2000$ bp is chosen. Otherwise, a patch of size $L_j = 20{,}000$ is chosen.

The power spectrum $S(f)$ for this control sequence is defined as the modulus squared of the discrete Fourier transform $\tilde{u}_f$ of u_i:

$$S(f) \equiv |\tilde{u}_f|^2. \tag{1}$$

We find that $S(f)$ resembles a "1/f-type" spectrum, as shown in Fig. 1 *a*. The spectrum scales approximately as $S(f) \approx f^{-\beta}$, where $\beta \approx 1$ for this sequence. However, there are important deviations from pure power-law behavior, which indicate the presence of characteristic scales. We define the correlation exponent $\beta(\ell)$ as

$$\beta(\ell) \equiv -\left.\frac{d \log S(f)}{d \log f}\right|_{f=1/\ell}, \tag{2}$$

where $\ell = 1/f$ has dimensions of length, i.e., $\beta(\ell)$ represent successive slopes of the double-log plot of $S(f)$. We find that after additional smoothing, $\beta(\ell)$ displays three local maxima, which correspond to the three scales of patchiness of the control sequence (Fig. 1 *b*).

At the lowest frequencies, the spectrum is distorted by artifacts of the fast Fourier transform (FFT) method. Specifically, at small frequencies approaching $1/N$, where N is the FFT window size, there is a spurious contribution arising from the treatment of the data as periodic with period N (Buldyrev et al., 1995). This finite size effect shifts the peaks in $\beta(\ell)$, casting doubt on this method for estimating patch sizes that approach N. The estimation of $\beta(\ell)$ also requires an arbitrary amount of smoothing by visual inspec-

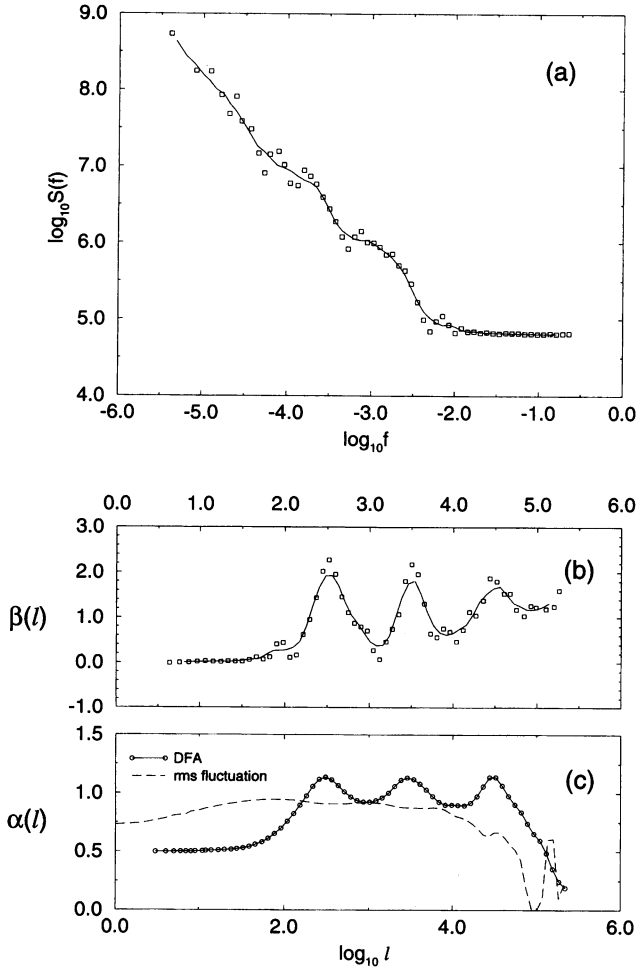

FIGURE 1 (*a*) Double log plot of the power spectrum $S(f)$ of an artificial control DNA sequence of length 2^{19} bp. The spectrum is an average over a moving window of size 2^{18} bp with shifts of 2^{16} bp. We applied logarithmic binning to smooth the spectrum by averaging over windows that grow in size as $2^{1/4}$. The solid line shows the spectrum after further smoothing using a subjective smoothing criterion. We note that the spectrum scales approximately as a power law over 3 decades. The characteristic scales are not readily discernible in the spectrum. (*b*) Log-linear plot of the power spectrum correlation exponent $\beta(\ell)$ for the same sequence, where $\ell \equiv 1/f$. The exponent $\beta(\ell)$ is estimated by taking the negative of the local slope of the solid line in *a*. The solid line is $\beta(\ell)$ after further smoothing, showing three clear maxima. We smooth $S(f)$ and $\beta(\ell)$ by averaging over a moving window of fixed size, and the degree of smoothing is to some extent arbitrary. (*c*) DFA correlation exponent $\alpha(\ell)$ for the same sequence. The exponent $\alpha(\ell)$ is found by calculating the local slope of the double-log plot of the DFA function. No smoothing or filtering is required. The exponents $\alpha(\ell)$ and $\beta(\ell)$ peak at three locations corresponding to the three characteristic patch sizes. The peaks occur at approximately 300 bp, 3000 bp, and 30,000 bp, showing that the location of the peaks is always about 1.5 multiplied by the patch sizes. Also shown is $\alpha(\ell)$ found from the "DNA walk" rms fluctuation method (*dashed line*) (Peng et al., 1992), which is unable to detect the three characteristic patch sizes.

tion, making it susceptible to human judgment. Other approaches, based on simple "DNA walk" fluctuation analysis (Azbel et al., 1982; Peng et al., 1992), are unable to detect distinct characteristic patch sizes (Fig. 1 *c*). However, detrended fluctuation analysis (DFA) (Peng et al., 1994) does not suffer from these disadvantages, and for these reasons we place greater emphasis on DFA in our studies.

We use the variant of the DFA method described by Buldyrev et al. (1995). The net displacement $y(n)$ of the sequence u is defined by $y(n) \equiv \sum_{i=1}^{n} u_i$, which can be thought of graphically as a one-dimensional random walk. The sequence $y(n)$ is then divided into a number of overlapping subsequences of length ℓ, each of which is shifted with respect to the previous subsequence by a single nucleotide. For each subsequence, linear regression is used to calculate an interpolated "detrended" walk $y'(n) \equiv a + b(n - n_0)$. Then we define the "DFA fluctuation" by $F_D(\ell) \equiv \sqrt{\langle(\delta y)^2\rangle}$, where $\delta y \equiv y(n) - y'(n)$, and the angle brackets denote averaging over all points $y(n)$. We use a moving window to obtain better statistics. The DFA exponent $\alpha(\ell)$ is defined by

$$\alpha(\ell) \equiv \frac{d \log F_D(\ell)}{d \log (\ell + 3)}, \tag{3}$$

where the +3 term is a correction that is important for small ℓ (Buldyrev et al., 1995). As we mention above, for an ideal power law $\alpha(\ell) = [1 + \beta(\ell)]/2 =$ constant. Indeed, the power spectrum $\tilde{S}(f)$ of the net displacement $y(n)$ is equal to $S(f)/f^2$ of the original sequence u_i. The integral of $\tilde{S}(f)$ for all frequencies higher than $f = 1/\ell$ measures the error of the approximation of $y(n)$ by a smooth function with resolution of length scale ℓ, which is analogous to $F_D^2(\ell)$. Thus

$$F_D^2(\ell) \sim \int_{1/\ell}^{\infty} \frac{S(f)}{f^2}\, df, \tag{4}$$

which proves the above relation for the ideal power law behavior of $S(f) \approx f^{-\beta}$.

To present both $\alpha(\ell)$ and $\beta(\ell)$ on approximately the same scale, we can plot $[1 + \beta(\ell)]/2$ instead of $\beta(\ell)$. Fig. 1 *c* shows $\alpha(\ell)$ for the artificial control model described above. (A fast DFA computer algorithm is available at http://polymer.bu.edu/dfa.)

The functions $\alpha(\ell)$ and $\beta(\ell)$ are measures of how correlated a sequence is on different length scales. Because peaks in $\alpha(\ell)$ and $\beta(\ell)$ correspond to higher correlations, therefore, by studying peaks in $\alpha(\ell)$ and $\beta(\ell)$, we can estimate distinct characteristic DNA patch sizes embedded in a sequence with an apparent $1/f$ power spectrum. We emphasize that such peaks corresponding to a given size do not imply the existence or absence of domains of that size, but rather imply an abundance of patches with that size relative to a power-law distribution of patch sizes.

We show here that the peaks should occur at scales of the patch size multiplied by a factor a, where $a = 1/\ln 2 \approx 1.44$. This is numerically close to the measured value $a = 1.5$ obtained from simulations. For sequences composed of patches of fixed size L and bias $b \equiv P(+1) - P(-1)$ that strictly alternate in sign, the average patch size is L. The power spectrum $S(f)$ for such a sequence has a large negative slope near $f = 1/L$; therefore $\beta(\ell) \equiv -d \log S(f)/d \log$

f, where $\ell = 1/f$ has a maximum near $\ell_{max} = L$. We define $a \equiv \ell_{max}/L$, and thus obtain $a = 1$. However, we assume that in real sequences the biases are not strictly alternating in sign, but can take on positive and negative values with equal probability. We find for such a sequence that the mean patch size is no longer L, but is $L/\ln 2$. We must take into account this extra factor of $1/\ln 2 \approx 1.44$, because several patches with the same bias can become joined to form larger patches. So we expect

$$a = 1/\ln 2 \approx 1.44, \tag{5}$$

which is numerically close to the measured value of $a \approx 1.5$.

RESULTS FOR KNOWN DNA SEQUENCES

We next apply these methods for detecting and examining characteristic scales of patchiness to the 16 chromosomes of *Saccharomyces cerevisiae*. Fig. 2 shows the DFA exponent $\alpha(\ell)$ using the RY rule and the SW rule. For comparison, power spectrum results are also shown. Note that the two results have common features. The greater sensitivity of the power spectrum method is offset by noise, especially for shorter sequences. For $\ell > 10^4$ there is a noticeable difference between $\alpha(\ell)$ and $\beta(\ell)$, and this is possibly due to differences in the finite size effects for the two functions. Note, however, that the extrema of the two functions occur in approximately the same positions, even for relatively large ℓ.

Fig. 3 *a* shows the DFA exponent for each of the 16 yeast chromosomes for the RY rule and Fig. 3 *b* the corresponding information for the SW rule. Note the similarity of $\alpha(\ell)$ for different chromosomes. We find that for $\ell < 10^3$ bp, the different chromosomes have almost identical $\alpha(\ell)$. This similarity indicates that the correlation properties of the different chromosomes are very similar for $\ell < 10^3$ bp. We find also that the first few maxima in $\alpha(\ell)$ roughly coincide for the different chromosomes in Fig. 3 *b*. This indicates that the 16 chromosomes have similar patch sizes, because peaks in $\alpha(\ell)$ correspond to characteristic patch sizes, although visual inspection of the concentration profiles of the chromosomes (Feldmann et al., 1995; Dujon et al., 1994) reveals no striking similarity.

Next we estimate characteristic patch sizes for several eukaryotic sequences longer than 10^5 bp, as well as for some *E. coli* bacterial sequences, as shown in Fig. 4. We used the peaks in $\alpha(\ell)$ divided by the factor $a = 1.5$ to evaluate the actual patch sizes. We find that similar patch sizes appear in several sequences, and some even appear on sequences from different species.

We now examine a number of possible explanations for these findings. To test the hypothesis that the correlation properties and patchiness in yeast chromosomes may simply be due to the alternation of coding and noncoding DNA (Nee, 1992), we study the effects of shuffling the nucleotides in each coding and noncoding region separately while preserving their respective lengths and nucleotide concentrations. If the known long-range correlation properties of DNA were due merely to the alternation of genes and intergenic sequences, then shuffling each coding and noncoding region separately would not significantly decrease or alter the presence of correlations, because the ordering of the introns and exons is not changed. We test this idea by comparing the DFA exponent $\alpha(\ell)$ for yeast chromosome III before and after the partial shuffling described above (Fig. 5). The values of $\alpha(\ell)$ thus obtained for yeast chromosome III and for the same partially shuffled sequence show that the alternation of coding and noncoding DNA may contribute to the long-range correlation properties of yeast chromosome III and explain the major patch size near and above 1000 bp. But the hypothesis cannot explain all of the correlation properties of the chromosome below 1000 bp. Specifically, as seen in Fig. 5, both for the SW rule and for the RY rule, the shuffled chromosome shows little or no

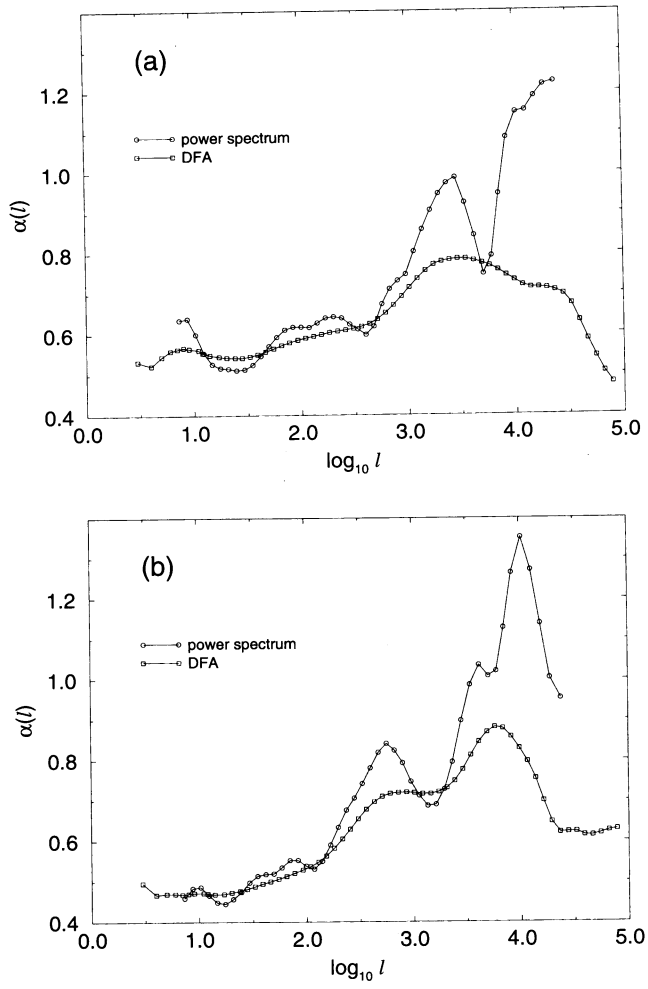

FIGURE 2 DFA exponent $\alpha(\ell)$ for the yeast chromosomes using (*a*) the RY rule and (*b*) the SW rule. The approximate relation $\alpha(\ell) = (1 + \beta(\ell))/2$ is used to plot the power spectrum correlation results on the same scale for comparison. The procedure used to calculate the power spectrum exponent $\beta(\ell)$ is described in Fig. 1. Note that the DFA and the power spectrum results have features in common.

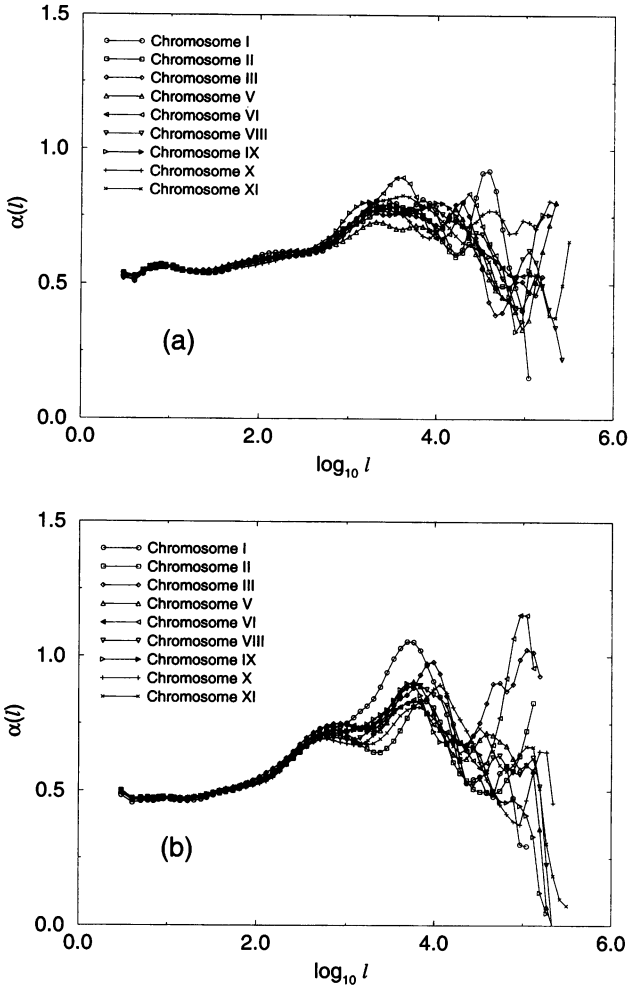

FIGURE 3 DFA exponent $\alpha(\ell)$ for yeast chromosomes using (*a*) the RY rule and (*b*) the SW rule. We note that the general shape of $\alpha(\ell)$ is similar for all 16 chromosomes. In particular, $\alpha(\ell)$ is almost identical for all 16 chromosomes for $\ell < 10^3$ bp, and the peaks and valleys (i.e., extrema) are close to each other for the SW rule, suggesting that there are similar characteristic patch sizes present in all chromosomes.

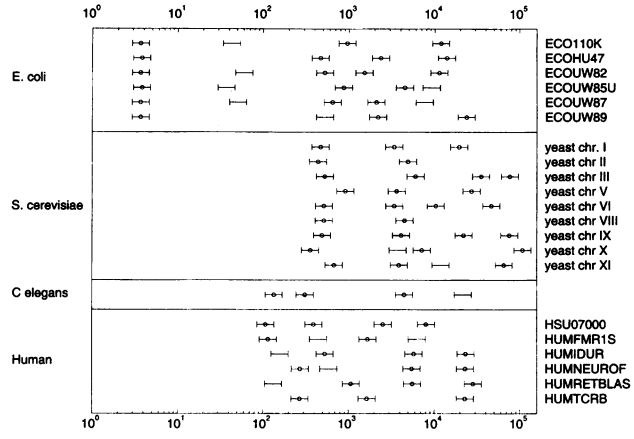

FIGURE 4 Characteristic patch sizes for the 16 yeast chromosomes estimated using the SW rule. Also shown for comparison are results for six *E. coli* sequences, one *C. elegans* sequence, and six human sequences. Only sequences larger than 10^5 bp were used. The patch sizes were estimated by locating the peaks in $\alpha(\ell)$ and dividing the position of the peaks by 1.5. Patch sizes that could only be estimated by visual inspection of the peaks are indicated by error bars without circles. The bacterial sequences have a patch size that is absent in the other sequences. The loci names of the human and *E. coli* sequences are as they appear in the figure. Except for some yeast sequences, all sequences are found in the GenBank database.

correlation on length scales $\ell < 100$ bp, whereas the unshuffled chromosome has significant levels of correlation on the same scales. Moreover, this hypothesis is not relevant for higher eukaryotes, whose genes are separated by long intergenic sequences and split by introns. Using the same method of partial shuffling of only genes or only intergenic sequences, we also find that the main contribution to the correlations on length scales below 1000 bp is made by intergenic sequences, although they constitute only about one-third of each yeast chromosome.

We next test the hypothesis that patchiness could arise from the abundance of repetitive sequences in genomic DNA (Buldyrev et al., 1993a). If this were true, a control sequence constructed from repetitive sequences would be able to reproduce the patchiness and the correlation properties of genomic DNA sequences. We use the measured concentrations of two highly repetitive sequences (Bell, 1992, 1993) found in humans to construct a control sequence. Specifically, we study the DFA exponent $\alpha(\ell)$ of a control sequence composed of 7.5% of Alu repeats and 15% of Line-1c repeats interspersed with uncorrelated sequences with average nucleotide concentrations estimated from all available human sequences larger than 50 kbp. Each uncorrelated spacer sequence has a bias of $b = P(\mathrm{AT}) - P(\mathrm{CG}) = 0.15$, and has an exponential length distribution. These parameters are typical for human DNA sequences. The results in Fig. 6 show that although repetitive sequences are able to explain some features of the patchiness found in real data, there are qualitative differences between the model and the real data. These differences are unlikely to disappear by increasing the number of types of repetitive elements because other repetitive sequences, which we found using the PYTHIA server (Jurka et al., 1992), occur in much smaller numbers than the Alu or Line-1 repeats.

We next plot $\alpha(\ell)$ for several species to examine how the correlation properties of DNA vary from species to species (Fig. 7). We find that human sequences, along with other vertebrate sequences, are anticorrelated for the SW rule and correlated for the RY rule on scales $\ell < 10$ bp. It is generally thought that in vertebrates, the methylation of C and its subsequent mutation into a T in the dinucleotide sequence CG leads to a reduced probability of finding CG and an increased probability of finding TG over evolutionary time, because the product of accidental deamination of 5-methyl C is T, which is indistinguishable from the other, nonmutant T residues in the DNA (Alberts et al., 1994). To examine the hypothesis that DNA methylation is the cause of the correlations (RY rule) and anticorrelations (SW rule) in the DNA of these organisms, we study the following

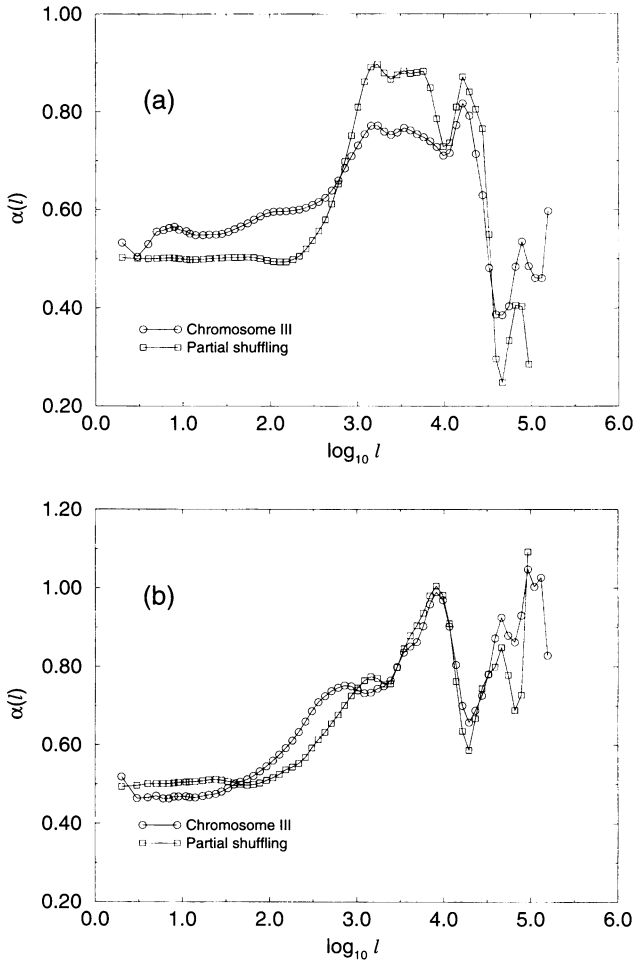

FIGURE 5 Comparison of $\alpha(\ell)$ for yeast chromosome III and the model described in the text of alternating "coding" and "noncoding" patches of uncorrelated DNA for (*a*) the RY rule and (*b*) the SW rule. The shuffling of the nucleotides in each coding and noncoding region preserves both the lengths and the ordering of the regions, as well as their nucleotide concentrations. For both rules the partial shuffling destroys correlations for $\ell < 100$ bp, showing that the alternation of coding and noncoding regions alone cannot explain the known long-range correlation properties of DNA.

Markov model (see discussion below). We measure the frequencies of occurrence of n-nucleotide "n-tuples" in real human sequences; then we generate a Markov chain $\{n_i\}$ in which the probability of finding $n_i = +1$ depends only on the previous m nucleotides $n_{i-1} \ldots n_{i-m}$ (see also Lewis et al., 1995). We find that $m = 1$ gives rise to correlations for the RY rule and anticorrelations for the SW rule. The $m = 1$ transition matrix has 16 dinucleotide concentrations whose characteristic feature is the abundance of TG relative to the scarcity of CG. To cancel the effects of DNA methylation without affecting the other dinucleotide concentrations, we alter the transition matrix by halving the TG concentration, increasing the CG concentration by the same amount, and changing other transition coefficients involving T or C accordingly. We find that the control Markov chain thus obtained lacks short-range anticorrelations for the SW rule, although the correlations for the RY rule

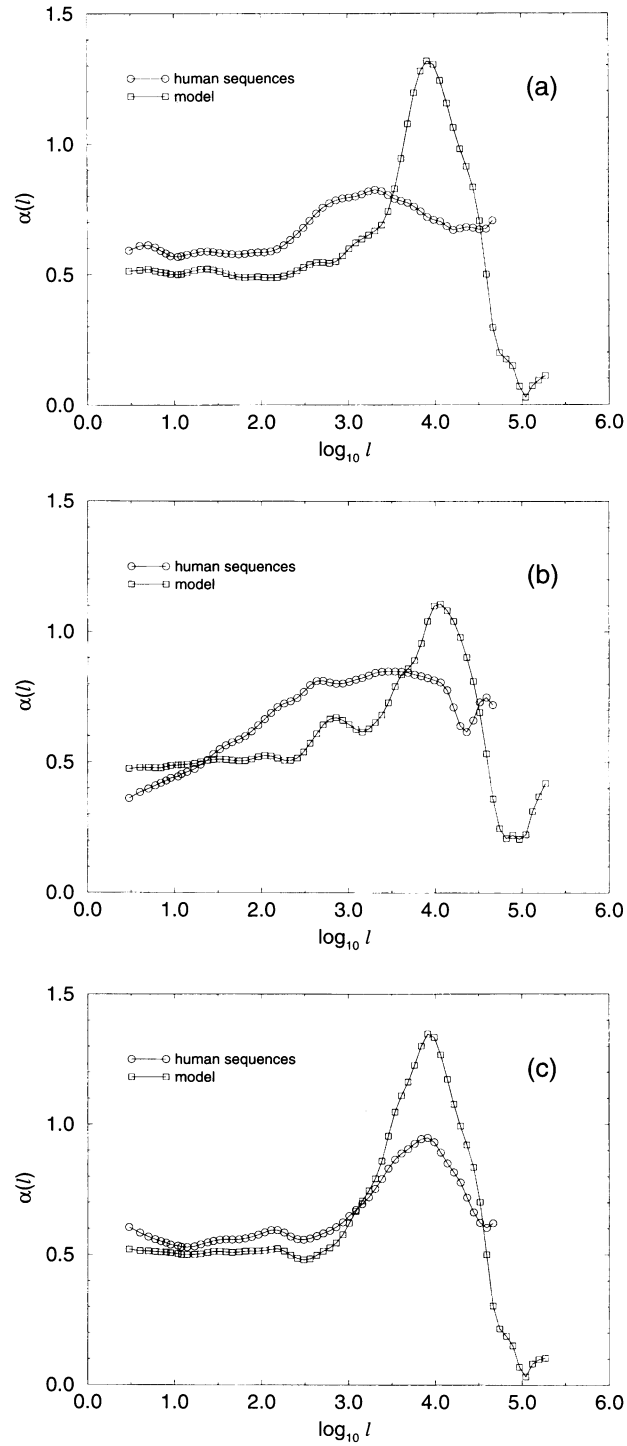

FIGURE 6 Comparison of DFA exponent $\alpha(\ell)$ for human sequences and an artificial control for (*a*) the RY rule, and (*b*) the SW rule, and (*c*) the KM rule. The artificial control sequence is composed of interspersed LINE-1c repeats, ALU repeats, and uncorrelated sequences. The maxima for the KM rule occur at the same scale for human sequences and the model, suggesting that long-range correlations may be partially due to repetitive sequences. However, we note that this artificial control sequence gives rise at most to two characteristic patch sizes, and cannot reproduce the plateau in $\alpha(\ell)$ for the SW rule. For the RY rule the model strongly disagrees with the data. We conclude that this model cannot explain the correlation properties and the patchiness found in DNA. We used the LINE-1c region in HUMHBB starting at 23137 and ending at 29515, and the ALU region starting at 66776 and ending at 67042.

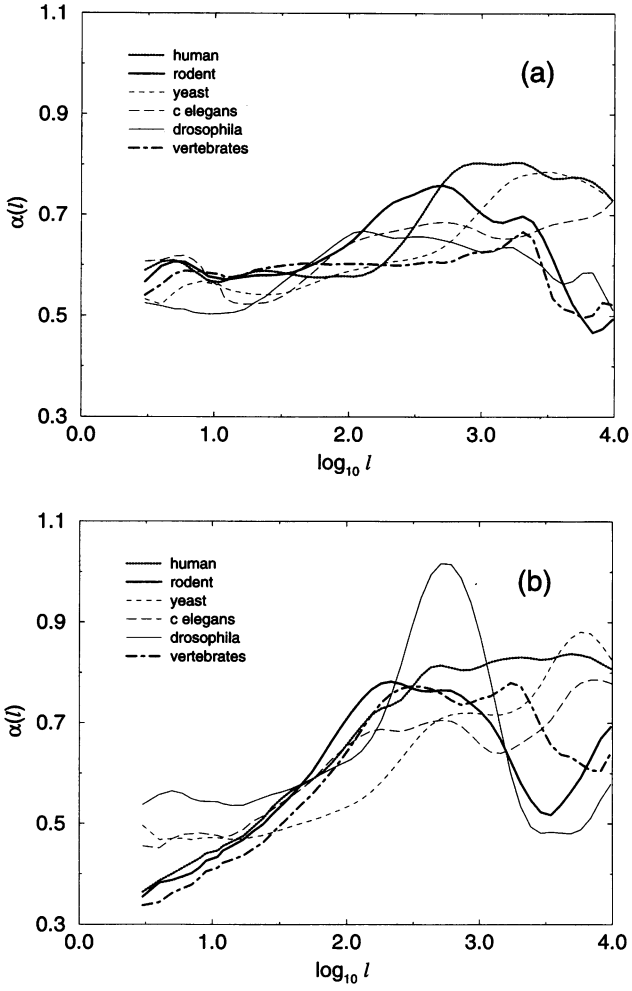

FIGURE 7 DFA exponent $\alpha(\ell)$ for several categories of DNA sequences for (*a*) the RY rule and (*b*) the SW rule. The vertebrate category excludes rodents and humans and comprises mainly chickens. Note that human and other vertebrate sequences are anticorrelated for $\ell < 30$ bp for the SW rule.

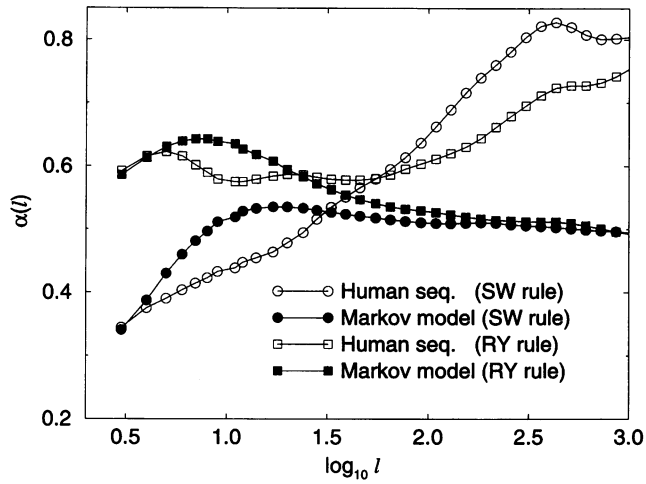

FIGURE 8 Comparison of the DFA exponent $\alpha(\ell)$ for a real human sequence (HUMTCRB) and the three-step Markov model described in the text. We find that the model is unable to account for the anticorrelations for 9 bp $< \ell <$ 30 bp.

remain unchanged, indicating that DNA methylation is likely the cause of short-range anticorrelations in the SW rule, but not of short-range correlations in the RY rule. We also find that $\alpha(\ell)$ of the $m = 1$ model sequence and the real sequence bear little resemblance, suggesting that larger m is needed to reproduce the correlation properties of DNA. Fig. 8 shows the results for $m = 3$. We find that this three-step Markov model is able to reproduce the behavior of the real sequence on length scales $\ell < 5$ bp, as expected. But for $\ell > 5$ bp, the model does not correctly describe the real sequence, and in fact, a simple Markov model of order m cannot reproduce correlations on scales much longer than m (see discussion below). Indeed, for the SW rule the three-step Markov model sequence generates anticorrelations on scales $\ell < 9$ bp as expected, but gives rise to correlations on scales 9 bp $< \ell <$ 30 bp, which bear no resemblance to the human sequence they model (Fig. 8). To understand why the three-step Markov model generates correlations for the SW rule for 9 bp $< \ell <$ 30 bp, we study the frequency of 3-tuples and 6-tuples (Fig. 9) (Mantegna et al., 1995). Whereas the 3-step Markov model and the human sequence have virtually identical frequencies for the occurrence of 3-tuples, they differ considerably for 6-tuples, as might be expected.

Finally, we examine whether tandem repeats of dinucleotide can explain the correlation properties of DNA. We calculate the distribution of lengths of the tandem repeats in human sequences. We then generate a random sequence that has the same length distributions of tandem repeats as those found from the human sequences. Fig. 10 compares $\alpha(\ell)$ computed from human sequences with $\alpha(\ell)$ computed from the tandem repeat model. We find that the model differs significantly from the real data.

DISCUSSION

The known mosaic structure of DNA (Bernardi et al., 1985; Churchill, 1989; Fickett et al., 1992) allows for the possibility that fluctuations in the nucleotide concentrations may lead to the existence of long-range correlations, and recently, such long-range power-law correlations (Arneodo et al., 1995; Munson et al., 1992; Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992) were shown to exist in some genomic DNA sequences.

Possible explanations for these long-range correlations have been put forward, including 3D structure (Grosberg et al., 1993), insertion-deletion (Buldyrev et al., 1993b), a generalized Lévy walk model of repetitive elements (Buldyrev et al., 1993a), and point mutation and duplication (Li, 1991; Li and Kaneko, 1992). There have also been several attempts to explain long-range correlations by the presence of patches of fixed size (Azbel, 1973, 1995; Karlin and Brendel, 1993), or alternation of coding and noncoding sequences of certain characteristic sizes (Nee, 1992). It has

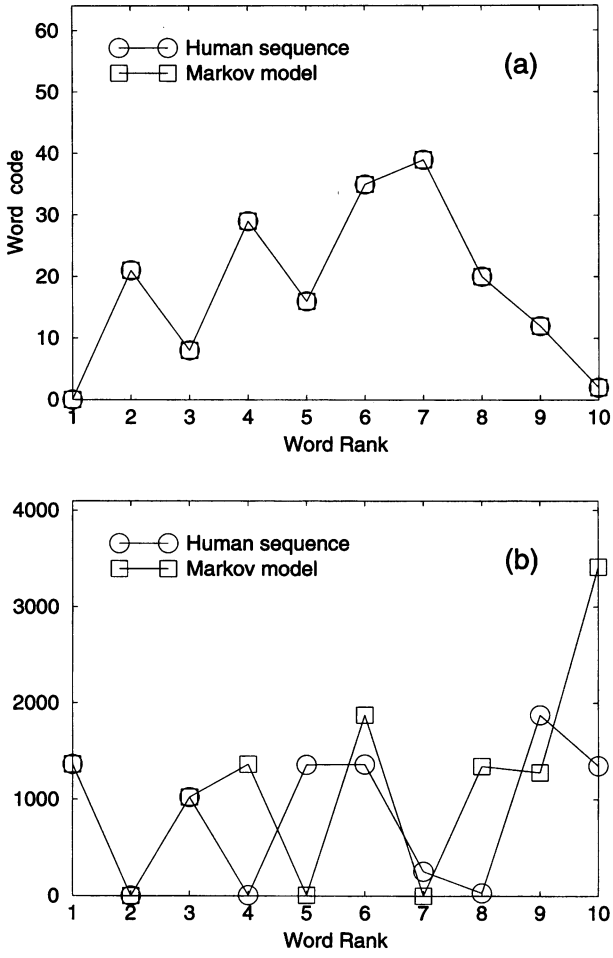

FIGURE 9 Ten most frequent 3-tuples (*a*) and 6-tuples (*b*) in a three-step Markov model of human sequence HUMTCRB. Each of the 64 3-tuples and the 4096 6-tuples has been assigned a numerical code to distinguish it from the others. We find that whereas the model reproduces *n*-tuple rankings almost exactly for 3-tuples, it is unable to do so for 6-tuples. The differences are even more noticeable for higher ranking 6-tuples.

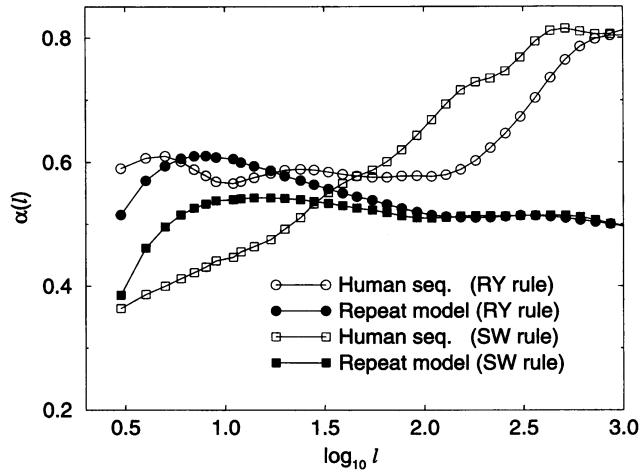

FIGURE 10 Comparison of the DFA exponent $\alpha(\ell)$ computed from all real human sequences larger than 10^5 bp and from the model of tandem dinucleotide repeats. The model does not adequately reproduce the correlation properties of the human sequences.

even been claimed that special quasiergodic Markov processes can account for correlations on very large length scales (Kanter and Kessler, 1995). Although significant progress has been made in better understanding the origin of long-range correlations in terms of expansion-modification models (Li et al., 1994), deviations from precise power-law behavior and their relation to patchiness remain open questions. Our results give insight into how a variety of biological phenomena contribute to long-range correlations, but they also suggest that none of these phenomena can provide a full explanation.

As was argued by Kanter and Kessler (1995), simple Markov processes that are quasiergodic can generate correlations on long, but finite, length scales. They construct an artificial Markov process with 2^{10} states that correspond to segments of five nucleotides, or 5-tuples. The transition probabilities from one 5-tuple to the next are defined by a transition matrix such that only two different 5-tuples can follow any given 5-tuple. However, in long enough real DNA sequences, one can find almost any 5-tuple following any other 5-tuple. It is thus doubtful whether the model proposed by Kanter and Kessler is applicable to real DNA sequences. In contrast to the quasiergodic model of Kanter and Kessler, hidden Markov models can generate long-range (but finite-scale) correlations (Pevzner, 1992; Churchill, 1992), because they involve switching between states, as in the three-patch model discussed above, which switches between states of different bias.

We comment on the finding that the yeast chromosomes have similar $\alpha(\ell)$. We find that for $\ell < 10^3$ bp, the yeast chromosomes have almost identical mosaic structure and correlation properties. This suggests that unique mechanisms organize all yeast chromosomes and that these mechanisms may be significantly different in higher eukaryotes and prokaryotes.

Our finding of distinct characteristic patch sizes in genomic DNA sequences may shed some light on this observation. As seen in Fig. 4, similar patch sizes appear in several sequences, and some even appear in sequences from different species. The patchiness in eukaryotic DNA could be due partially to the elaborate organization and folding of DNA by proteins into nucleosomes and higher-order structures of chromatin. Nucleosome structure may be responsible for strong correlations near $\ell \approx 200$ bp, whereas the packaging of DNA into higher order structures like looped domains might lead to correlations on larger length scales. Note that the yeast sequences do not show patchiness on scales from 50 bp to 200 bp. Perhaps this is due to the absence in yeast of the normal H1 histones, which help pack nucleosomes together (Thoma et al., 1993).

Although we find that the long-range correlation properties of DNA cannot be fully explained by the most abundant repetitive sequences (the Alu and the Line-1c), this does not mean that a power-law distribution of lengths of insertions

and deletions cannot in general generate long-range correlations (Buldyrev et al., 1993a). The hypothesis that the distribution of DNA patch lengths may be described by a generalized Lévy walk model has very recently received support from both biologists and physicists. Specifically, Gu and Li have studied the size distribution of insertions and deletions in human and rodent pseudogenes (Gu and Li, 1995) and found it to be consistent with a power-law distribution. Independently, Bernaola-Galvan et al. have shown that there exists a multi-length-scaled structure for many DNA sequences that is well described by a power distribution of patch lengths (Bernaola-Galvan et al., 1996).

We also comment on our finding that tandem repeats of dinucleotide sequences are unable to explain the correlation properties of human DNA. It is known that a power-law distribution $P(\ell) \equiv \ell^{-\mu}$ of repeat lengths ℓ can only generate long-range correlations if $\mu \leq 3$, i.e., if the power-law tail decays relatively slowly (Buldyrev et al., 1993a). Because we find that the length distributions of the tandem repeats of the 16 dinucleotides consistently decay with $\mu > 3$, it comes as no surprise that these tandem repeats are unable to produce correlations on scales $\ell > 100$ bp.

In summary, we find distinct characteristic DNA patch sizes embedded in scale-invariant patch size distributions. Moreover, we find that, by themselves, repetitive sequences, the alternation of coding and noncoding regions, simple Markov processes, and tandem repeats are unable to explain fully the known long-range correlation properties and patchiness of genomic DNA sequences. Tests of expansion-modification models of DNA evolution that can generate long-range correlations (Li, 1991) and hidden Markov processes (Churchill, 1992) are among the problems that may be suitably addressed in future studies using the new techniques developed here.

We wish to thank A. L. Goldberger, I. Große, P. Ivanov, C.-K. Peng, R. Mantegna, and M. Simons for significant help at the initial stages of this work. We also wish very much to thank those who have made public the newly sequenced yeast chromosomes not yet incorporated into the GenBank database (Bussey et al., 1995; Feldmann et al., 1995; Galibert et al., unpublished observations; Dietrich et al., unpublished observations). We also thank the anonymous referees.

We thank the National Institutes of Health for support.

REFERENCES

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. 1994. Molecular Biology of the Cell, 3rd Ed. Garland Publishing, New York.

Arneodo, A., E. Bacry, P. V. Graves, and J. F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* 74:3293–3296.

Azbel, M. Ya. 1973. Random two-component, one-dimensional Ising model for heteropolymer melting. *Phys. Rev. Lett.* 31:589–593.

Azbel, M. Ya. 1995. Universality in a DNA statistical structure. *Phys. Rev. Lett.* 75:168–171.

Azbel, M. Ya., Y. Kantor, L. Verkh, and A. Vilenkin. 1982. Statistical analysis of DNA sequences. *Biopolymers.* 21:1687–1690.

Barabási, A.-L., and H. E. Stanley. 1995. Fractal Concepts in Surface Growth. Cambridge University Press, Cambridge.

Bell, G. I. 1992. Roles of repetitive sequences. *Comput. Chem.* 16: 135–143.

Bell, G. I. 1993. Repetitive DNA sequences: some considerations for simple sequence repeats. *Comput. Chem.* 17:185–190.

Bernaola-Galvan, P., Ramon Roman-Roldan, and Jose L. Oliver. 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E.* 53:5181–5189.

Bernardi, G. 1989. The isochore organization of the human genome. *Annu. Rev. Genet.* 23:637–661

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. *Science.* 228:953–958.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley. 1993a. Generalized Lévy walk model for DNA nucleotide sequences. *Phys. Rev. E.* 47:4514–4523.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley. 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E.* 51:5084–5091.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, and M. Simons. 1993b. Fractal landscapes and molecular evolution: analysis of myosin heavy chain genes. *Biophys. J.* 65:2673–2679.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, and H. E. Stanley. 1994. Fractals in biology and medicine: from DNA to the heartbeat. *In* Fractals in Science. A. Bunde and S. Havlin, editors. Springer-Verlag, Berlin. 49–83.

Bussey, H., D. B. Kaback, W. Zhong, D. T. Vo, M. W. Clark, N. Fortin, J. Hall, B. F. F. Ouellette, T. Keng, A. B. Barton, Y. Su, C. K. Davies, and R. K. Storms. 1995. The nucleotide sequence of chromosome I of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA.* 92:3809–3813.

Churchill, G. A. 1989. Stochastic models for heterogenous DNA sequences. *Bull. Math. Biol.* 51:79–94.

Churchill, G. A. 1992. Hidden Markov chains and the analysis of genome structure. *Comput. Chem.* 16:107–116.

Dujon, B., D. Alexandraki, B. Andre, W. Ansorge, et al. 1994. Complete DNA sequence of yeast chromosome XI. *Nature.* 369:371–378.

Feldmann, H., M. Aigle, G. Aljinovic, B. Andre, et al. 1995. Complete DNA sequence of yeast chromosome II. *EMBO J.* 13:5795–5809.

Fickett, J. W., D. C. Torney, and D. R. Wolf. 1992. Base compositional structure of genomes. *Genomics.* 13:1056–1064.

Gates, M. 1986. A simple way to look at DNA. *J. Theor. Biol.* 119: 319–328.

Grosberg, A., Y. Rabin, S. Havlin, and A. Nir. 1993. Self-similarity in DNA structure. *Europhys. Lett.* 23:373–377.

Gu, X., and W.-H. Li. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *Mol. Biol. Evol.* 40:464–473.

Johnston, M., S. Andrews, R. Brinkman, J. Cooper, H. Ding, J. Dover, Z. Du, A. Favello, and L. Fulton. 1994. Complete nucleotide sequence of *S. cerevisiae* chromosome VIII. *Science.* 265:2077–2082.

Jurka, J., J. Walichiewics, and A. Milosavljevic. 1992. Prototypic sequences for human repetitive DNA. *J. Mol. Evol.* 35:286–291.

Kanter, I., and D. A. Kessler. 1995. Markov processes: linguistics and Zipf's law. *Phys. Rev. Lett.* 22:4559–4562.

Karlin, S., B. E. Blaisdell, R. J. Sapolsky, L. Cardon, and C. Burge. 1993. Assessments of DNA inhomogeneities in yeast chromosome III. *Nucleic Acids Res.* 21:703–711.

Karlin, S., and V. Brendel. 1993. Patchiness and correlations in DNA sequences. *Science.* 259:677–680.

Lewis, H. B., J. D. Knafels, D. R. Methog, and S. Celniker. 1995. Sequence

analysis of the *cis*-regulatory regions of the bithorax complex of *Drosophila*. *Proc. Natl. Acad. Sci. USA.* 92:8403–8407.

Li, W. 1991. Expansion-modification systems: a model for spatial $1/f$ spectra. *Phys. Rev. A.* 43:5240–5260.

Li, W., and K. Kaneko. 1992. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17:655–660.

Li, W., T. G. Marr, and K. Kaneko. 1994. Understanding long-range correlations in DNA sequences. *Phys. D.* 75:392–416.

Mantegna, R. N., S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley. 1995. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E.* 52:2939–2950.

Munson, P. J., R. C. Taylor, and G. S. Michaels. 1992. Long range DNA correlations extend over the entire chromosome. *Nature.* 360:636–636.

Nee, S. 1992. Uncorrelated DNA walks. *Nature.* 357:450–450.

Ossadnik, S. M., S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and H. E. Stanley. 1994. Correlation approach to identify coding regions in DNA sequences. *Biophys. J.* 67:64–70.

Peng, C.-K., S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature.* 356:168–171.

Peng, C. K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E.* 49:1685–1689.

Pevzner, P. A. 1992. Nucleotide sequences versus Markov models. *Comput. Chem.* 16:103–106.

Pickover, C. A. 1984. Frequency spectra of DNA sequences: application to a human bladder cancer gene. *J. Mol. Graphics.* 2:50–52.

Shlesinger, M. F., and J. Klafter. 1986. Lévy walks versus Lévy flights. *In* On Growth and Form. H. E. Stanley and N. Ostrowsky, editors. Nijhoff, Dordrecht, The Netherlands. 279–283.

Stanley, H. E. 1971. Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, London.

Stanley, H. E., 1995. Power laws and universality. *Nature.* 378:554–554.

Thoma, F., G. Cavalli, and S. Tanaka. 1993. Structural and functional organization of yeast chromatin. *In* The Eukaryotic Genome: Organization and Regulation. P. M. A. Broda, S. G. Oliver, and P. F. G. Sims, editors. Cambridge University Press, Cambridge. 43–52.

Viswanathan, G. M., V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley. 1996. Lévy flight search patterns of wandering albatrosses. *Nature.* 381:413–415.

Voss, R. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68:3805–3808.