Finite-size effects on long-range correlations: Implications for analyzing DNA sequences

C.-K. Peng and S. V. Buldyrev
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215

A. L. Goldberger
Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215

S. Havlin
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215

M. Simons
Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215
and Biology Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

H. E. Stanley
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215
(Received 14 December 1992)

We analyze the fluctuations in the correlation exponents obtained for noncoding DNA sequences. We find prominent sample-to-sample variations as well as variations within a single sample in the scaling exponent. To determine if these fluctuations may result from finite system size, we generate correlated random sequences of comparable length and study the fluctuations in this control system. We find that the DNA exponent fluctuations are consistent with those obtained from the control sequences having long-range power-law correlations. Finally, we compare our exponents for the DNA sequences with the exponents obtained from power-spectrum analysis and correlation-function techniques, and demonstrate that the original "DNA-walk" method is intrinsically more accurate due to reduced noise.

I. INTRODUCTION

Recently, it was discovered that noncoding DNA sequences exhibit scale-invariant long-range correlations quantitatively measured by a power-law decay [1]. The exponent characterizing the power-law decay of the correlations is well defined for infinite sequences. However, for DNA sequences the accuracy of the analysis is limited by the length of the available nucleotide chains (i.e., there are only a few samples of published nucleotide sequences with length $> 10^5$ base pairs). It is therefore of importance to investigate the effect of finite length on the exponents calculated. The purpose of this report is threefold: (i) to demonstrate that there are prominent fluctuations on the exponent characterizing long-range correlations in finite-length DNA sequences; (ii) to investigate systematically the effect of finite sample size on this exponent using control sequences for comparison with actual DNA; and (iii) to compare and contrast three different methods—DNA walk, correlation function, and power-spectrum analysis—of measuring the correlation exponent. While this work is motivated by recent studies of nucleotide sequences, our findings can be generalized to other problems involving finite-length sequences.

II. FINITE-LENGTH DNA SEQUENCES

Applying the method of Ref. [1], we map a nucleotide chain to a binary sequence $u(i)$ such that $u(i) = 1$ if a pyrimidine occurs at position i and $u(i) = -1$ if a purine occurs at i [2]. We can then generate a DNA walk such that the walker will step up or down depending on the sign of $u(i)$. The trace (landscape) of the DNA walk, defined as $y(\ell) = \sum_{i=1}^{\ell} u(i)$, is plotted in Fig. 1(a) for rat embryonic skeletal myosin-heavy-chain gene (GenBank name: RATMHCG). The rms fluctuation for such a DNA walk is defined as

$$F(\ell) \equiv \sqrt{\overline{[\Delta y(\ell)]^2} - \left[\overline{\Delta y(\ell)}\right]^2}, \tag{1}$$

where

$$\Delta y(\ell) = y(\ell_0 + \ell) - y(\ell_0) \tag{2}$$

and the bars indicate an average over all position ℓ_0. Figure 1(b) shows the double-logarithmic plot of $F(\ell)$ vs ℓ, the linearity of this plot indicates that
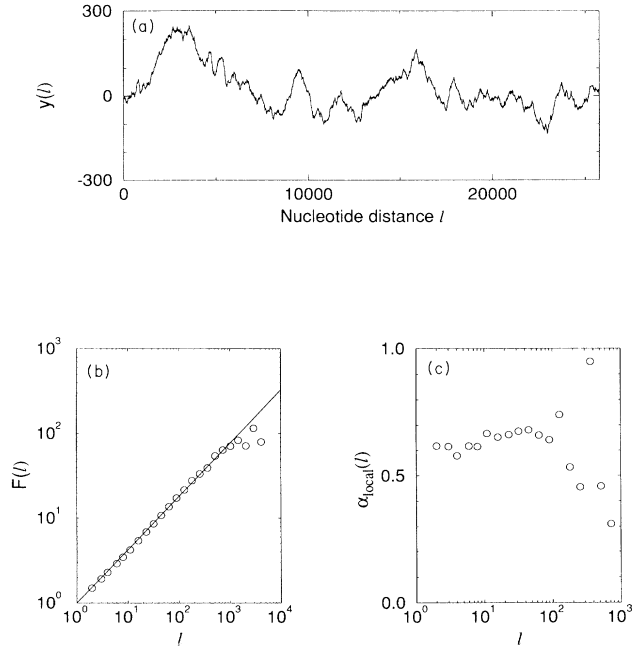
$$F(\ell) \sim \ell^{\alpha}. \tag{3}$$

FIG. 1.   (a) DNA walk for rat embryonic skeletal myosin-heavy-chain gene (GenBank name: RATMHCG). (b) Double-logarithmic plot of the rms fluctuation $F(\ell)$ vs ℓ, where $F(\ell)$ is defined in (1). The straight line, with slope $\alpha = 0.63$, represents the linear least-squares fit from $\ell = 1$ to 512. (c) The successive slopes for $\alpha_{\text{local}}(\ell)$ for the plot in (b) [7].

The linear least-squares fit (ℓ from 1 to 1000) for this plot gives $\alpha = 0.63$. The exponent $\alpha > 1/2$ indicates that $u(i)$ is not an uncorrelated random sequence or a short-range correlated sequence (such as that associated with a Markovian chain process), but instead the sequence has long-range correlation with power-law decay [3].

However, we note that the log-log plot of $F(\ell)$ vs ℓ clearly includes data points that deviate from a straight line. Furthermore, the scaling behavior in Fig. 1(b) is less than three decades (less than 1/10 of the whole sequence studied). Such deviations can be more readily visualized if one calculates the slopes of successive pairs of points in Fig. 1(b). Such a plot of "local" slope (α_{local}) is shown in Fig. 1(c). The fluctuations about the linear regression are seen most dramatically for larger values of ℓ. Qualitatively, these fluctuations could arise from the fact that we measure a finite sequence of length $N = 25\,759$ nucleotides. The main goal of this study, therefore, is to quantitatively calculate the magnitude of the expected fluctuation in α, in order to test if the fluctuations could arise solely from finite-size effects, or whether some other mechanism must be invoked.

III. FINITE-LENGTH CONTROL SYSTEMS

To see how finite-size affects the scaling exponents of DNA sequences, we next compare the scaling behavior of the DNA sequences with that of a control system—an artificial sequence with known correlation exponent α. To

generate such artificial sequences, we apply the following numerical method [3, 4]: A sequence of real numbers $u(i)$ is generated by inverse Fourier transforming a sequence of complex numbers $\tilde{u}(q)$, where

$$\tilde{u}(q) = |q|^{-\beta/2}\eta(q), \tag{4}$$

and $\eta(q)$ is Gaussian stochastic noise of amplitude A, i.e., random variables with a normal distribution of density, such that

$$\overline{\eta(q)} = 0 \tag{5}$$

and

$$\overline{\eta(q)\eta^*(q')} = A^2\delta(q - q'), \tag{6}$$

where the bars indicate an average over different realizations of the stochastic noise [5]. It is straightforward to verify that $u(i)$ thus obtained has the correct correlations described in (3) with $\alpha = (1 + \beta)/2$ [6].

To compare the finite-size effects for the control system and the DNA sequence, we perform the following steps:

(i) Generate a correlated sequence of finite length N with a given exponent α.

(ii) Calculate the successive slopes of $\log_{10} F(\ell)$ versus $\log_{10} \ell$, denoted $\alpha_{\text{local}}(\ell, N)$ [7].

(iii) Repeat the processes (i) and (ii) for many (M) times and obtain the probability distribution of $\alpha_{\text{local}}(\ell, N)$.

(iv) Calculate the average $(\overline{\alpha_{\text{local}}})$ and the standard deviation of $\alpha_{\text{local}}(\ell, N)$. For M large enough, these values are found to converge.

We find that the standard deviation of $\alpha_{\text{local}}(\ell, N)$, denoted by $\Delta\alpha_{\text{local}}(\ell, N)$, decreases with the length of N but is not sensitive to the magnitude of the stochastic noise [amplitude A of the noise $\eta(q)$ in Fourier space, see Eq. (6)].

Figure 2 shows α_{local} vs ℓ. The solid lines represent $\overline{\alpha_{\text{local}}(\ell, N)} \pm 2\Delta\alpha_{\text{local}}(\ell, N)$. Results from a typical realization of an artificial correlated sequence are also plotted (open squares) to demonstrate the large fluctuations of α_{local} even in a single sample. Of note, the comparison between the artificial control and actual DNA sequences (both with the same length N) shows good agreement, i.e., the fluctuations of α_{local} of the DNA sequence are comparable to the typical fluctuations of a finite artificial correlated sequence of the same length [8].

A simple theoretical argument can be applied to derive a scaling relation for $\Delta\alpha_{\text{local}}(\ell, N)$. A sequence of length N can be divided into N/ℓ independent subsequences of length ℓ. Hence $\alpha_{\text{local}}(\ell, N)$ of a single sequence of length N corresponds to the average (mean) value of N/ℓ independent samples. The fluctuation $\Delta\alpha_{\text{local}}(\ell, N)$ of a mean value $\alpha_{\text{local}}(\ell, N)$ is inversely proportional to the square root of the number of independent samples to be averaged. Therefore, we conclude that [9]
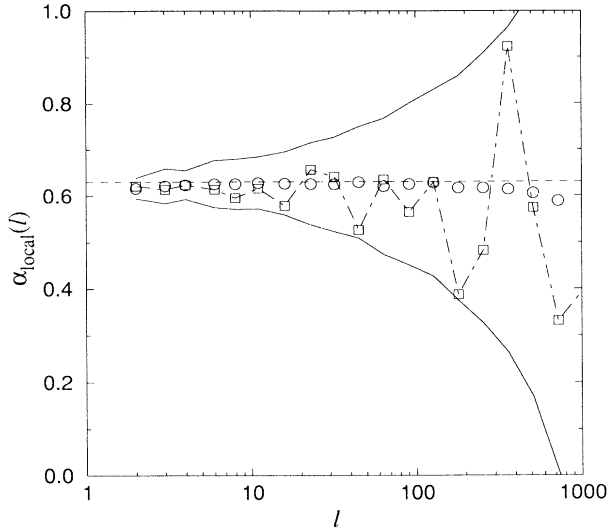
FIG. 2. The successive slope $\alpha_{\text{local}}(\ell, N)$ vs $\log_{10}\ell$ for the artificial sequence with length $N = 25\,759$. The squares represent the results from a single realization. The circles are the results of averaging over 1000 realizations, while the solid lines are twice the standard deviation $\Delta\alpha_{\text{local}}(\ell, N)$. The horizontal dashed line is the exact exponent for an infinite sequence $(N = \infty)$ [7].

$$\Delta\alpha_{\text{local}}(\ell, N) \sim \left(\frac{\ell}{N}\right)^{1/2}. \tag{7}$$

Figure 3 confirms Eq. (7) numerically for the artificial correlated sequence. Although we cannot derive the prefactor, i.e., the proportionality constant, in Eq. (7) analytically, this scaling relation is of practical importance.
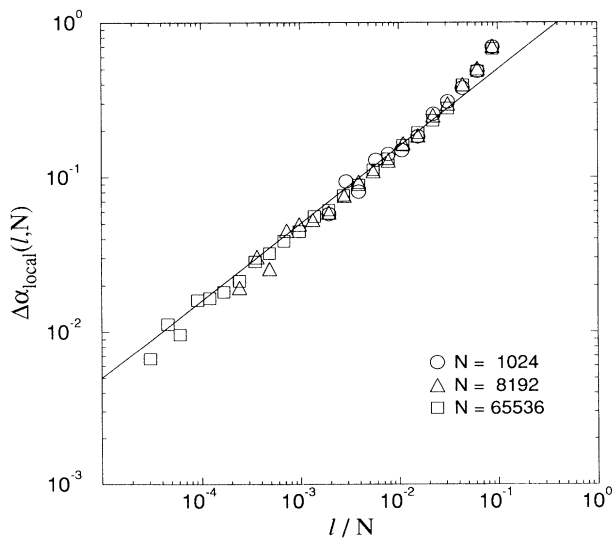

FIG. 3. Double-logarithmic plot of $\Delta\alpha_{\text{local}}(\ell, N)$ vs ℓ/N for three values of chain length, $N = 1024$ (O), 8192 ($\triangle$), and $65\,536$ ($\square$). The solid line has slope $1/2$, the value predicted by Eq. (7).

If we have only a single sequence of finite length N, we can nevertheless estimate $\Delta\alpha_{\text{local}}(\ell, N/M)$ by dividing it into M independent subsequences and calculate $\alpha_{\text{local}}(\ell, N/M)$ for each subsequence (providing that N, M, and N/M are large enough for statistical meaningful results). Using Eq. (7) we can obtain the expected fluctuations for $\alpha_{\text{local}}(\ell, N)$ and compare them with the sample we wish to study.

IV. ALTERNATIVE CORRELATION ANALYSES

To compare the fluctuations of α in our DNA walk method with those found in other methods, we also present the results from two standard methods to study the correlation property of sequences, namely the correlation function and the power spectrum. For the DNA sequences, we define the correlation function as

$$C(\ell) \equiv \overline{[u(\ell' + \ell) - \bar{u}][u(\ell') - \bar{u}]}. \tag{8}$$

The bar indicates an average over all positions ℓ'. The power spectrum density $S(q)$ is obtained by (a) Fourier transforming the sequence $\{u(i)\}$ and (b) taking the square of the Fourier component. For a stationary sequence, the power spectrum is the Fourier transform of the correlation function. If the correlation decays algebraically (not exponentially), i.e., there is no characteristic scale for the decay of the correlation, as we found in the noncoding DNA sequences, then
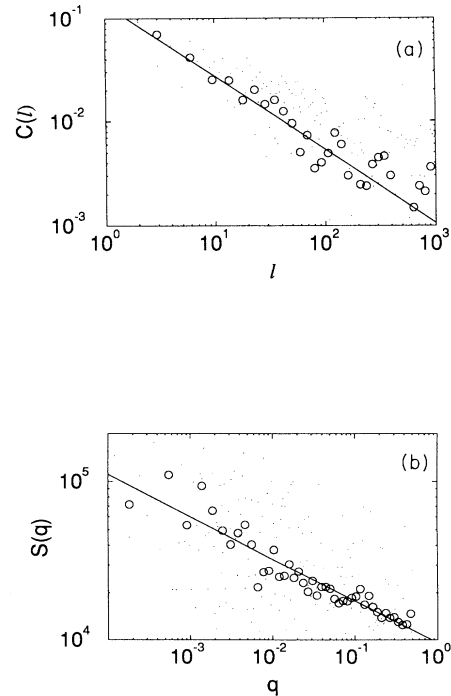

FIG. 4. Double-logarithmic plots for (a) the correlation function $C(\ell)$ vs ℓ and (b) power spectrum $S(q)$ vs q for the same DNA sequence shown in Fig. 1. The circles in (a) and (b) are data obtained by averaging neighboring points, while the dots represent raw data. The lines are a linear least-squares fit with slopes (a) $\gamma = 0.70$ and (b) $\beta = 0.27$.

$$C(\ell) \sim \ell^{-\gamma} \tag{9}$$

and

$$S(q) \sim q^{-\beta}. \tag{10}$$

The exponents α, β, and γ defined in (3), (9), and (10) are not independent [3, 4], since

$$\alpha = (1 + \beta)/2 = (2 - \gamma)/2. \tag{11}$$

Figures 4(a) and 4(b) are log-log plots of $C(\ell)$ vs ℓ and $S(q)$ vs q for the same DNA sequence studied in Fig. 1.

For a typical DNA sequence of *finite length*, both the correlation function and power spectrum are very noisy (Fig. 4). In fact, it is very difficult to get a good estimation of the scaling exponents directly from these two methods. In order to reduce the noise in Fig. 4, we have smoothed the data by simple averaging. The least-squares fits for the scaling exponents from the correlation and power spectrum are $\beta = 0.27$ and $\gamma = 0.70$, which correspond to $\alpha = 0.64$ and 0.65, respectively. We note that for most of the DNA sequences we analyzed, the discrepancies among these three methods are at least as great as the results shown here.

For an artificial correlated sequence, similar noisy fluctuations for $C(\ell)$ and $S(q)$ are observed. As is evident from the scatter of points about the regression line in Fig. 4, the local slope analyses for $C(\ell)$ and $S(q)$ will show larger fluctuations (even for the binned data) than that observed in the DNA-walk method. The reason for the smaller fluctuations of $\alpha_{\text{local}}(\ell, N)$ in the DNA-walk method [(Fig. (1b)] is due to the fact that $F^2(\ell)$ is a double summation of $C(\ell)$ [1] and, therefore, the noise is dramatically reduced.

V. CONCLUSION

In summary, we have demonstrated that the fluctuations found in estimating the correlation exponent of a finite-size sample may be quite prominent. Therefore, a careful comparison of the estimated value for the exponent (and its fluctuations) with that of a suitable control model is crucial. To carry out this sort of comparison, we provide a systematic procedure that may be relevant not only to DNA correlations but to other sequences of correlated random variables.

ACKNOWLEDGMENTS

[1] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992). Long-range correlation in a noncoding DNA sequence was reported independently by W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992), and later was confirmed by R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992). The long-range correlations were found to extend over the entire yeast chromosome III region (315 000 nucleotides) by P. J. Munson, R. C. Taylor, and G. S. Michaels, Nature **360**, 636 (1992). Very recently, the DNA walk approach has been extended to quantify the nature of long-range correlations in human writings (up to 10^6 characters) [A. Schenkel, J. Zhang, and Y.-C. Zhang, Fractals **1**, 47 (1993)] and in heart rate intervals (up to 10^4 beats) [C. K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993)].

[2] This simple rule is based on the binary classification of four DNA nucleotides; adenine and guanine are purines and cytosine and thymine are pyrimidines.

[3] S. Havlin, R. B. Selinger, M. Schwartz, H. E. Stanley, and A. Bunde, Phys. Rev. Lett. **61**, 1438 (1988).

[4] C. K. Peng, S. Havlin, M. Schwartz, and H. E. Stanley, Phys. Rev. A **44**, R2239 (1991); S. Prakash, S. Havlin, M. Schwartz, and H. E. Stanley, ibid. **46**, R1724 (1992).

[5] The property of $\eta^*(q) = \eta(-q)$ is needed to guarantee that the inverse Fourier transform gives a sequence of real numbers. Equation (6) can be derived from this criterion.

[6] This is not the only way to generate correlated variables, but the algorithm of Eq. (4) has two advantages: (i) For $\beta = 0$, we obtain Gaussian white noise. (ii) For $\beta = 2$, we obtain Brown noise (a representation of Brownian motion).

[7] The definition of α_{local} will affect its fluctuations. In this study, we measure the slopes of successive pairs of data points that are uniformly distributed on a logarithmic scale of ℓ. To be more precise, we define $\alpha_{\text{local}}(\ell_i, N) = [\log_{10} F(\ell_{i+1}) - \log_{10} F(\ell_i)]/(\log_{10} \ell_{i+1} - \log_{10} \ell_i)$, where ℓ_i is the integer part of $2^{i/2}$) and $i = 0, 1, 2, 3, \ldots$. This same definition of α_{local} has been used by many authors in the context of calculating critical point exponents; the analog of Fig. 2 was constructed for a range of genes first by A. A. Tsonis, J. B. Elsner, and P. A. Tsonis (unpublished), and by S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E (to be published).

[8] We have also checked 34 different sequences (including sequences of length ranging from 3000 to 300 000 nucleotides) and found that the fluctuations in the successive slopes α_{local} are comparable to the fluctuations in the artificial sequence.

[9] For the analogous problem of studying fluctuations in exponent α in d-dimensional systems due to finite-size effects, we expect that $\Delta\alpha_{\text{local}}(\ell, N) \sim (\ell/N)^{d/2}$. For a fractal system, d should be replaced by d_f, the fractal dimension of the system.