Generalized Lévy-walk model for DNA nucleotide sequences

Sergey V. Buldyrev
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215

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

Shlomo Havlin
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215 and Department of Physics, Bar Ilan University, Ramat-Gan, Israel

Chung-Kang Peng
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215

Michael Simons
Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215

H. Eugene Stanley
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215
(Received 6 January 1993)

We propose a generalized Lévy walk to model fractal landscapes observed in noncoding DNA sequences. We find that this model provides a very close approximation to the empirical data and explains a number of statistical properties of genomic DNA sequences such as the distribution of strand-biased regions (those with an excess of one type of nucleotide) as well as local changes in the slope of the correlation exponent α. The generalized Lévy-walk model simultaneously accounts for the long-range correlations in noncoding DNA sequences and for the apparently paradoxical finding of long subregions of biased random walks (length l_j) within these correlated sequences. In the generalized Lévy-walk model, the l_j are chosen from a power-law distribution $P(l_j) \propto l_j^{-\mu}$. The correlation exponent α is related to μ through $\alpha = 2 - \mu/2$ if $2 < \mu < 3$. The model is consistent with the finding of "repetitive elements" of variable length interspersed within noncoding DNA.

PACS number(s): 87.10.+e

I. INTRODUCTION

Recently there has been considerable interest in the finding of long-range (power-law) correlations in certain genomic DNA sequences [1–4]. While several tentative explanations have been proposed regarding the origin, function, or biological significance of this observation [5, 6], this question can be regarded as open. In this paper, we offer a straightforward mechanism for generating such long-range correlations in DNA sequences based on a generalization of a Lévy walk. In Sec. II, we discuss the motivation for a model describing DNA sequences and in Sec. III, we define this generalized Lévy-walk model. In Sec. IV we qualitatively compare the DNA landscape for noncoding sequences with the landscape of the generalized Lévy walk, and then quantitatively analyze the fluctuations in these landscapes by estimating the correlation exponent α. We also consider the distribution of regions of distinct "strand bias" in DNA, and compare this to the predictions of the generalized Lévy walk. Finally, in Sec. V we summarize our results and discuss the possible biological implications of this type of analysis.

Some of the technical details are presented in the Appendixes, including the treatment of an alternative model which contains a single characteristic length scale ("correlation length"). We show that this type of model cannot adequately describe the observed long-range correlation properties.

II. MOTIVATION

The method of DNA walks, introduced by Peng *et al.* [1], allows graphical representation of the fluctuations of the nucleotide content [see Fig. 1(a)]. A "DNA walk" is initiated from the first nucleotide of the sequence and continued to the last nucleotide. For each pyrimidine at position i, the walker takes a step up $[u(i) = +1]$, and for each purine, a step down $[u(i) = -1]$. This procedure generates an irregular graph resembling a fractal landscape [Fig. 1(a)]. The defining feature of such a landscape is the statistical self-similarity (self-affinity) of the plots obtained at various magnifications.

We analyze fluctuations of the actual data as described in Ref. [1]. Specifically, we focus on the standard devia-

tion in the nucleotide content:

$$F^2(l,L) \equiv \frac{1}{L-l}\sum_{l_0=1}^{L-l}\left(\Delta y(l_0,l) - \overline{\Delta y(l)}\right)^2, \qquad (2.1)$$

where L is the number of the nucleotides in the entire sequence and

$$\overline{\Delta y(l)} \equiv \frac{1}{L-l}\sum_{l_0=1}^{L-l}\Delta y(l_0,l) \qquad (2.2)$$

is the average value of $\Delta y(l_0,l)$ over entire sequence. Here $\Delta y(l_0,l)$ is defined by

$$\Delta y(l_0,l) \equiv y(l_0+l) - y(l_0), \qquad (2.3)$$

and

$$y(l) \equiv \sum_{i=1}^{l} u(i), \qquad (2.4)$$

where $u(i) = 1$ for pyrimidines [cytosine (C) or thymine (T)], and $u(i) = -1$ for purines [adenine (A) or guanine (G)]. It was found [1] that the fluctuations can be approximated by

$$F(l,L) \sim l^{\alpha}, \qquad (2.5)$$

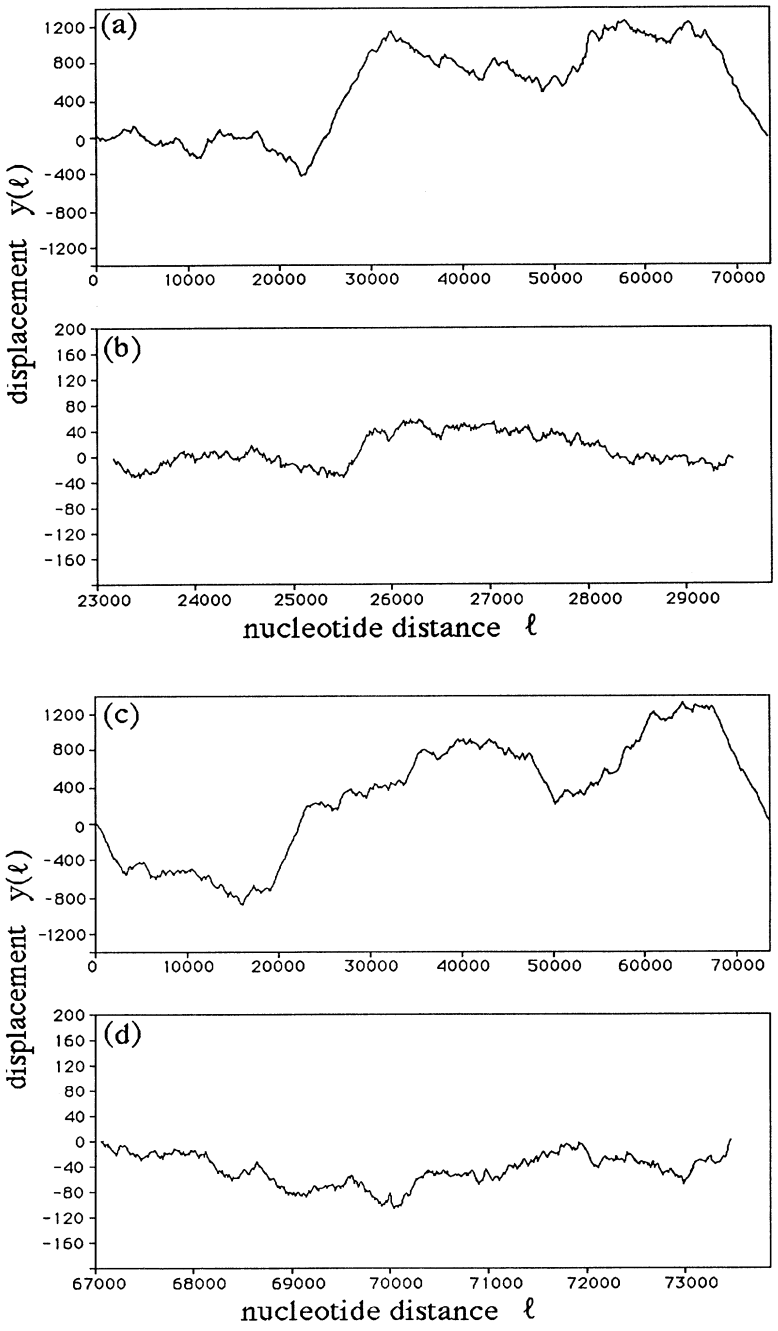

FIG. 1. DNA-walk displacement $y(l)$ (excess of purines over pyrimidines) vs nucleotide distance l for (a) HUMHBB (human beta-globin chromosomal region of the total length $L = 73\,239$); (b) the LINE-1c region of HUMHBB starting from 23 137 to 29 515; (c) the generalized Lévy-walk model of length 73 326 with $\mu = 2.45$, $l_c = 10$, $\alpha_o = 0.6$, and $\epsilon = 0.2$ (see Appendix B); and (d) a segment of a Lévy walk of exactly the same length as the LINE-1c sequence from step 67 048 to the end of the sequence. This subsegment is a Markovian random walk. Note that in all cases the overall bias was subtracted from the graph such that the beginning and ending points have the same vertical displacement ($y = 0$). This was done to make the graphs clearer and does not affect the quantitative analysis of the data.

where α is the correlation exponent. For α close to 0.5, there is no correlation or only short-range correlation in the sequence. If α is significantly deviated from 0.5, it indicates long-range correlation [7].

The DNA-walk analysis demonstrates a striking difference between coding and noncoding sequences: the coding sequences usually consist of few lengthy alternating regions of different nucleotide content ("strand bias"), corresponding to up-hill and down-hill regions of the DNA walk, while noncoding sequences consist of very many such regions of a wide range of length scales (see Fig. 1 of Ref. [1]). The coding sequences can, therefore, be easily divided into few subsequences (by eye or by simple computer routine [1]) of different nucleotide concentration. The average value of α for those subsequences is close to 0.5 which indicates the absence of long-range correlations within such subsequences. By contrast, noncoding sequences cannot be divided into a small number of such regions and their correlation exponents are significantly greater than 0.5.

Although the correlation is long range in the noncoding sequences, there seems to be a paradox: *long uncorrelated regions of up to thousands of base pairs can be found in such sequences as well.* For example, consider the human beta-globin intergenomic sequence of length $L = 73\,326$ (GenBank name: HUMHBB). This long noncoding sequence has 50% purines (no *overall* strand bias) and $\alpha = 0.7$ [see Fig. 1(a)]. However, from nucleotide No. 67 089 to 73 228, there occurs the LINE-1 region (defined in Ref. [8]). In this region of length 6139 base pairs, there is a strong strand bias with 59% *purines.* In this noncoding subregion, we find power-law scaling of F, with $F \sim l^{\alpha}$, with $\alpha = 0.55$, quite close to that of a random walk [7].

Even more striking is another region of 6378 base pairs, from nucleotide No. 23 137 to 29 515, which has 59% *pyrimidines* and is *uncorrelated*, with remarkably good power-law scaling and correlation exponent $\alpha = 0.49$ [Fig. 1(b)]. This region actually consists of three subsequences, complementary to shorter parts of the LINE-1 sequence.

These features motivate us to apply a generalized Lévy-walk model [see Figs. 1(c), 1(d), and 2] for the noncoding regions of DNA sequences. We will show in the next section how this model can explain the long-range correlation properties, since there is no characteristic scale "built into" this generalized Lévy walk. In addition, the model simultaneously accounts for the observed large subregions of noncorrelated sequences within these noncoding DNA chains.

III. LÉVY-WALK MODEL AND ITS GENERALIZATION

The classic Lévy-walk model describes a wide variety of diverse phenomena that exhibit long-range correlations [9–15]. The model is defined schematically in Fig. 2(a): A random walker takes not one but l_1 steps in a given direction. Then the walker takes l_2 steps in a new randomly-chosen direction, and so forth. The lengths l_j of each string are chosen from a probability distribution, with

$$P(l_j) \propto (1/l_j)^{\mu}, \tag{3.1}$$

where $\sum_{i=1}^{N} l_i = L$, N is the number of substrings and L is the total number of steps that the random walker takes.

We consider a generalization of the Lévy walk [14] to interpret recent findings of long-range correlation in noncoding DNA sequences described above. Instead of taking l_j steps in the *same* direction as occurs in a classic Lévy walk, the walker takes each of l_j steps in *random* directions, with a fixed bias probability

$$p_+ = (1 + \epsilon_j)/2 \tag{3.2a}$$

to go up and

$$p_- = (1 - \epsilon_j)/2 \tag{3.2b}$$

to go down, where ϵ_j gets the values $+\epsilon$ or $-\epsilon$ randomly. Here $0 \le \epsilon \le 1$ is a bias parameter (the case $\epsilon = 1$ reduces to the Lévy walk). Figure 2(b) shows such a generalized Lévy walk for the same choice of l_j as in Fig. 2(a).

As shown in Appendix A, the generalized Lévy walk—like the pure Lévy walk—gives rise to a landscape with a fluctuation exponent α that depends upon the Lévy walk parameter μ [10, 14],

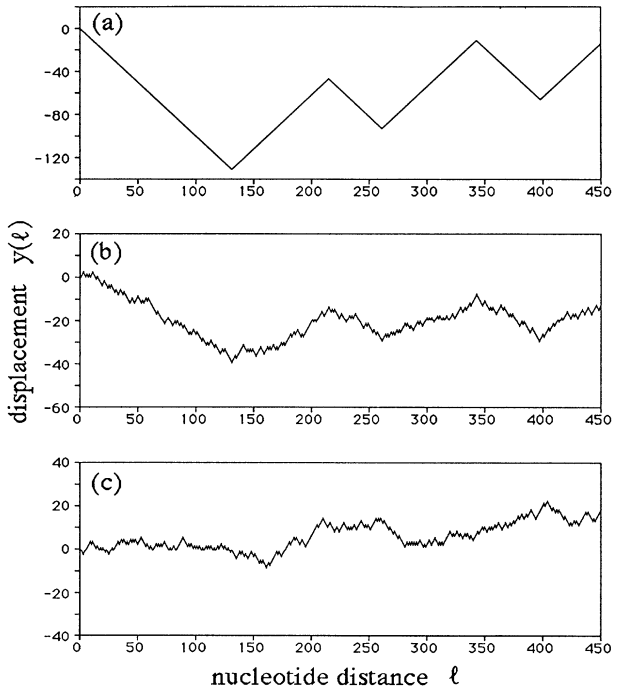

FIG. 2. Displacement $y(l)$ vs number of steps for (a) the classical Lévy-walk model consisting of six strings of l_j steps, each taken in alternating directions; (b) the generalized Lévy-walk model consisting of six biased random walks of the same length with a probability of p_+ that it will go up equal to $(1 \pm \epsilon)/2$ ($\epsilon = 0.2$); and (c) the unbiased uncorrelated random walk. Note that the vertical scale in (b) and (c) is twice that in (a).

$$\alpha = \begin{cases} 1, & \mu \leq 2 \\ 2-\mu/2, & 2 < \mu < 3 \\ 1/2, & \mu \geq 3, \end{cases} \tag{3.3}$$

i.e., nontrivial behavior of α corresponds to the case $2 < \mu < 3$ where the first moment of $P(l_j)$ converges while the second moment diverges. The long-range correlation property for the Lévy walk, in this case, is a consequence of the broad distribution of Eq. (3.1) that lacks a characteristic length scale. However, for $\mu \geq 3$, the distribution of $P(l_j)$ decays fast enough that an effective characteristic length scale appears. Therefore, the resulting Lévy-walk behaves like a normal random walk for $\mu \geq 3$.

To be precise, we define our generalized L-step Lévy-walk model as follows.

(1) Choose a random number u which is uniformly distributed between 0 and 1, and define $l_j \equiv l_c u^{\mu-1}$ where l_c is some lower cutoff characteristic length. The number (l_j) thus generated will obey the distribution of Eq. (3.1).

(2) Produce a biased random walk of length l_j (see Appendix B) with p_+ and p_- given by Eq. (3.2), where ϵ_j takes on the value $+\epsilon$ or $-\epsilon$ randomly and ϵ is a fixed value close to 0.2 (corresponding to the percentage of purines vs pyrimidines in real DNA sequences).

(3) Iterate the process, attaching together biased random walks until the total length of the sequence reaches a given value L.

IV. COMPARISON WITH DNA DATA

To test the generalized Lévy-walk model, we have adjusted the two parameters μ and l_c [16] described in the previous section to best approximate features of an actual DNA sequence [the human beta-globin DNA sequence shown in Fig. 1(a)]. The resulting landscape for the generalized Lévy-walk model is presented in Fig. 1(c). The comparison of $F(l)$ for the model and DNA sequences is shown in Figs. 3(a) and 3(b).

A more detailed scaling analysis [Figs. 3(c) and 3(d)], considers the "local slopes" of successive points in the graphs of Figs. 3(a) and 3(b):

$$\alpha(l_i, L) \equiv \frac{\log_{10} F(l_{i+1}, L) - \log_{10} F(l_i, L)}{\log_{10} l_{i+1} - \log_{10} l_i}, \tag{4.1}$$

where l_{i+1} and l_i are values of two subsequent data points. The local slope changes from $\alpha = 0.6$ for $l_1 = 1$ to $\alpha = 0.75$ for $l_i = 128$, and stays at this value for about two decades. It eventually drops down when l_i becomes too close to L, since $F(L, L) \equiv 0$ according to Eq. (2.1). This kind of scaling behavior is general for all kinds of DNA sequences that contain noncoding material. The initial monotonic increase in α, however, does not mean that long-range correlations do not exist. Indeed, as seen in Fig. 3(d), a similar type of behavior exists in the generalized Lévy-walk model. Equation (3.3) is valid asymptotically for very large l and L and the local value of $\alpha(l, L)$ for finite values of l and L may differ considerably from its asymptotic value. The comparison of $\alpha(l, L)$ plots for human beta-globin chromosomal region ($L = 73\,326$) and a Lévy-walk model of the same size is

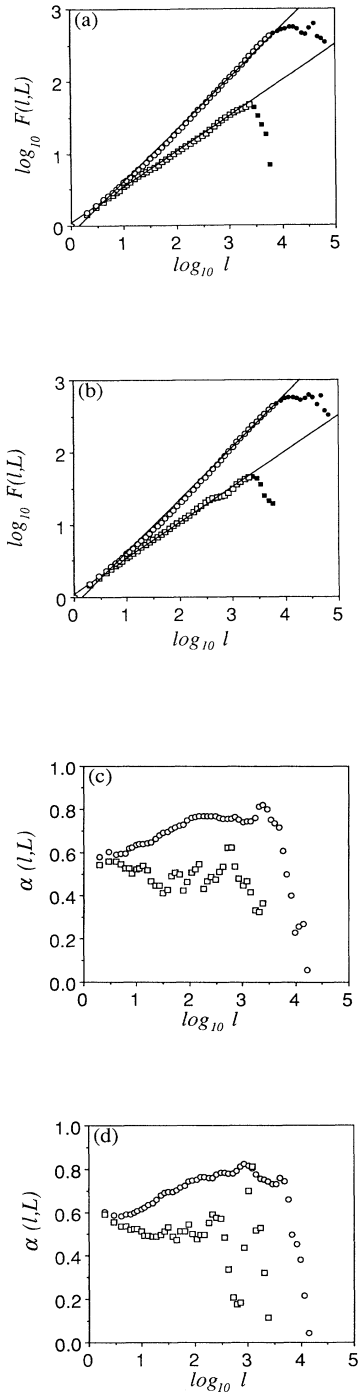

FIG. 3. Double-logarithmic plots of fluctuation $F(l, L)$ vs nucleotide distance l [(a) and (b)] for the sequences presented in Fig. 1 and the successive slopes of these plots $\alpha(l, L)$ vs $\log_{10} l$ [(c) and (d)]. The actual DNA sequences are presented in (a) and (c): the entire HUMHBB sequence (○) and LINE-1c sequence (□). The slopes for the linear fits are 0.72 and 0.49, respectively. The Lévy model sequences are presented in (b) and (d): the entire Lévy-walk sequence of Fig. 1(c) (○), a segment of this walk of Fig. 1(d) (□). The slopes for the linear fits are 0.73 and 0.49, respectively. The solid circles and solid squares in (a) and (b) are data omitted in linear regression fit.

shown in Figs. 3(c) and 3(d). In our model we use the value of $\mu = 2.45$ which corresponds to the asymptotic value of $\alpha = 0.775$, observed for the human beta-globin chromosomal region. A similar comparison is made for the largest available ($L = 315357$) DNA sequence [4], that of yeast chromosome III [see Fig. 4(a)].

For any given size L, it is possible to calculate the average value and standard deviation of $\alpha(l, L)$ for the Lévy-walk model by calculating $\alpha(l, L)$ for a large number k of statistically independent realizations of the model sequence of the size L. The data for yeast chromosome III are well within a two standard deviation interval ($k = 15$) for the generalized Lévy-walk model with $\mu = 2.5$, which corresponds to observed value of $\alpha = 0.75$ [Fig. 4(b)].

An alternative test of Lévy-walk (see also Appendix C) structure can be made if one analyzes a "coarse-grained" version of the original DNA sequence. To this end, we (i) divide the entire sequence into L/w subsequences of equal length w, (ii) replace each subsequence by 1 if there is an excess of purines or by 0 if there is an excess of pyrimidines, and (iii) calculate the distribution $P(s)$ of sizes s of long runs of 1's and 0's. These calculations for human beta-globin chromosomal region show that $P(s)$ has a scaling region of roughly one decade, where $P(s) \sim s^{-\mu}$

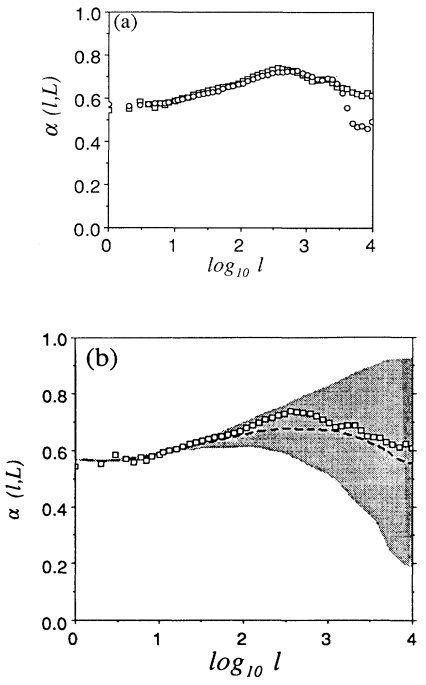

FIG. 4. Comparison of successive slopes of the scaling exponent α for yeast chromosome III (□) and (a) successive slopes of a realization of the generalized Lévy walk with parameters $L = 315000$, $\mu = 2.5$, $l_c = 5$, $\alpha_0 = 0.55$, $\epsilon = 0.16$ (○); (b) average successive slopes over 15 different realization of Lévy walks with the same parameters (dashed line). The shaded area corresponds to two standard deviations of successive slopes of the model, calculated for 15 random realizations. The parameters for Markov process, α_0 and ϵ, used in the model are calculated from real DNA sequence of yeast chromosome III (see Appendix B).

with $\mu \approx 2.5$. Our results are in good agreement with the value of the exponent $\alpha = 0.75$ (see Fig. 5). Unfortunately, the coarse-graining process requires a long sequence ($> 10^5$ nucleotides) in order that the statistics for the distribution be meaningful. To date, only a few documented long sequences are available, but as longer sequences become available this renormalization test should prove to be increasingly useful.

V. DISCUSSION

The key finding of this analysis is that a generalized Lévy-walk model can account for two hitherto unexplained features of DNA nucleotides: (i) the long-range power-law correlations that extend over thousands of nucleotides in sequences containing noncoding regions (e.g., genes with introns and intergenomic sequences), and (ii) the presence within these correlated sequences of sometimes large subregions that correspond to biased random walks. This apparent paradox is resolved by the generalized Lévy walk, a mechanism for generating long-range correlations (no characteristic length scale), that with finite (though rare) probability also generates large regions of uncorrelated strand bias. The uncorrelated subregions, therefore, are an anticipated feature of this mechanism for long-range correlations.

From a biological viewpoint, two questions immediately arise: (i) What is the significance of these uncorrelated subregions of strand bias? and (ii) What is the molecular basis underlying the power-law statistics of the Lévy walk? With respect to the first question, we note that these long uncorrelated regions at least sometimes correspond to well described but poorly understood sequences termed "repetitive elements," such as the LINE-1 region noted above [8, 17, 18]. There are at least 53 different families of such repetitive elements within the human genome. The lengths of these repetitive elements vary from 10 to 10^4 nucleotides [8, 19]. At least some of the repetitive elements are believed to be remnants of messenger RNA molecules that formerly did code for proteins [17, 18, 20, 21]. Alternatively, these segments may represent retroviral sequences that have inserted themselves into the genome [22]. Our finding that these repetitive elements have the statistical properties of biased random walks (e.g., the same as that of active coding sequences) is consistent with these hypotheses.

Finally, what are the biological implications of this type of analysis? Our findings clearly support the following possible hypothesis concerning the molecular basis for the power-law distributions of elements within DNA chains. In order to be inserted into DNA, a macromolecule should form a loop of certain length l with two ends, separated by l nucleotides along the sequence, coming close to each other in real space. The probability of finding a loop of length l inside a very long linear polymer scales as $l^{-\mu}$ [23]. Theoretical estimates of μ made by different methods [24–27] using a self-avoiding random-walk model [23] indicate that the value of μ for three-dimensional model is between 2.16 and 2.42. Our estimate made by the Rosenbluth Monte-Carlo method [28–30] gives $\mu = 2.22 \pm 0.05$ which yields according to Eq.

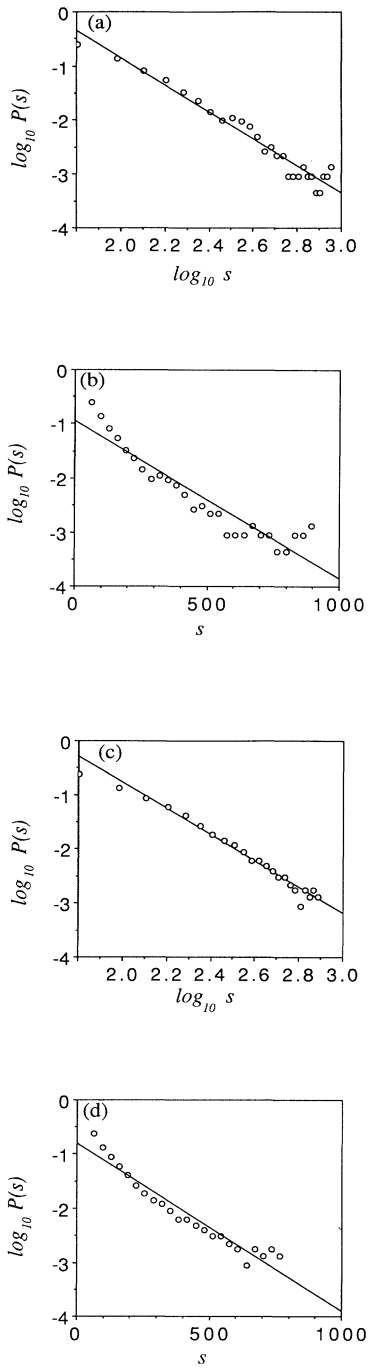

FIG. 5. The probability distribution to find a run of certain size s of purines or pyrimidines in the coarse-grained sequence calculated using coarse-grained window size equal to 32 (window size 32 is the minimal length in which the bias of random walk with $p = 0.6$ becomes larger than standard deviation). (a) Actual sequence of HUMHBB on log-log plot straight line has slope -2.50 and regression coefficient 0.957. (b) Actual sequence of HUMHBB on semilogarithmic plot straight line has slope -0.0029 and regression coefficient 0.882. (c) Log-log plot for the model sequence, shown in Fig. 1(c); the slope is -2.42, the regression coefficient 0.978. (d) Semilogarithmic plot for the model; the slope is -0.0031, the regression coefficient 0.933.

(3.3) $\alpha = 0.89$, a larger value than the effective value of $\alpha(l, L)$, observed in DNA of finite length. However, the asymptotic value of the exponent α remains uncertain since the statistics of Lévy walks converge very slowly due to rare events associated with the very long strings of constant bias that may occur in the sequence according to Eq. (3.1). This results in the very large error bars for $\alpha(l, L)$ for large values of l and finite length L (see Fig. 4). Even for the sequences of about 300×10^3 base pairs we cannot estimate the limiting value of α with good accuracy.

It is clear, however, that the behavior of DNA sequences cannot be satisfactorily explained in terms of only one characteristic length scale even of about $10^3 - 10^4$ base pairs long (see Appendix D). The asymptotic behavior of the scaling exponent α and whether it reaches some universal value for long DNA chains must await further data from the human genome project.

Note added. After this work was submitted, a report appeared that confirms the existence of long-range correlations in DNA [35]. However, where Ref. [35] might appear to disagree with Ref. [1] is in the interpretation of that finding for coding and noncoding regions. Both figures in [35] apply to the complete genome of the phage λ which does not contain noncoding sequences and consists of only three regions of different strand bias (see Fig. 1c of Ref. [1]). Each such region when analyzed separately by the DNA-walk method gives exponent $\alpha \approx 0.5$, close to that of random walk. The combination of three such regions produces a crossover in the local values of $\alpha(l, L) \approx 0.5$ at small length scales l to $\alpha(l, L) \approx 1$ at large l. Thus, for coding sequences, there is indeed no well-defined scaling exponent α for large length scales.

In contrast, the monotonically increasing local values of $\alpha(l, L)$ followed by a plateau at large l for noncoding sequences are completely explained by the generalized Levy-walk model presented here in terms of a crossover from an uncorrelated random walk at small length scales to a Lévy walk at large length scales. The latter has well-defined scaling with an exponent α related to the exponent μ characterizing the power-law distribution of steps of the Levy walk. Figure 3(d) of the present work clearly demonstrates that the generalized Levy-walk model accounts for the upward curvature in the values of $\alpha(l, L)$, followed by a plateau with $\alpha(l, L) \approx 2 - \mu/2$ [36].

ACKNOWLEDGMENTS

We wish to thank A. Yu. Grosberg, I. Labat, P. J. Munson, G. S. Michaels, E. I. Shakhnovich, M. F. Shlesinger, E. N. Trifonov, and G. H. Weiss for helpful discussions. Partial support was provided to S.V.B. by NSF, to A.L.G. by the G. Harold and Leila Y. Mathers Charitable Foundation, NHLBI, NIDA, and NASA, to C.K.P. by NIH, to M.S. by AHA, and to S.H. and H.E.S. by NSF.

APPENDIX A: DERIVATION OF EQ. (3.3)

The derivation of Eq. (3.3) is based on the fact that the distribution of differences of altitudes of the landscape

$\Delta y(l_0,l)$ separated by horizontal distance l is governed, for large values of $\Delta y(l_0,l)$, by a single largest string $l_{\max}$. The total number of strings n in the interval l for $\mu > 2$ (with converging first moment) is proportional to l. Since the probability for a single event $l_j > x$ is $x^{-\mu+1}$, the probability for $l_{\max}$ larger than x in n events is equal to $1-(1-x^{-\mu+1})^n$. When n is large, this probability can be simplified as $1-\exp(-nx^{-\mu+1})$. For large $x \gg n^{1/(\mu-1)}$ it is asymptotic to $nx^{-\mu+1}$. For $x \ll n^{1/(\mu-1)}$ it is close to 1, with $n^{1/(\mu-1)}$ being a characteristic value of $l_{\max}$. The value $n^{1/(\mu-1)} \sim l^{1/(\mu-1)}$ serves as a natural cutoff for the string distribution in the finite interval l. The second moment of this truncated distribution converges even for $\mu \leq 3$:

$$\langle {l_j}^2 \rangle \sim \int^{l^{1/(\mu-1)}} dx\, x^{-\mu+2} \sim l^{(3-\mu)/(\mu-1)}. \tag{A1}$$

Thus for $\Delta y(l_0,l) \ll l^{1/(\mu-1)}$, according to the central limit theorem, the distribution of $\Delta y(l_0,l) = \sum_{i=1}^{n} \epsilon_i l_i$, the sum of n independent variables $\epsilon_i l_i$, is Gaussian with variance $n\langle {l_j}^2 \rangle \sim l^{2/(\mu-1)}$:

$$P(\Delta y(l_0,l)) \sim \exp\left(-C[\Delta y(l_0,l)/l^{1/(\mu-1)}]^2\right), \tag{A2a}$$

for

$$\Delta y(l_0,l) \ll l^{1/(\mu-1)}. \tag{A2b}$$

Hence, in almost all cases $|\Delta y(l_0,l)|$ is of the order $l^{1/(\mu-1)}$.

However, when a rare event $l_{\max} \gg l^{1/(\mu-1)}$ happens, then $\Delta y(l_0,l) \approx l_{\max}$. (The contribution of all the rest of the strings is usually of the order $l^{1/(\mu-1)}$, as shown above, and can be neglected.) Thus for $\Delta y(l_0,l) \gg l^{1/(\mu-1)}$ the probability density of $\Delta y(l_0,l)$ coincides with the probability density of $l_{\max}$ which, as shown above, is proportional to $nx^{-\mu} \sim lx^{-\mu}$:

$$P(\Delta y(l_0,l)) \sim l\Delta y(l_0,l)^{-\mu} \tag{A3a}$$

for

$$\Delta y(l_0,l) \gg l^{1/(\mu-1)}. \tag{A3b}$$

It is clear, however, that $\Delta y(l_0,l) \leq l$. Thus ,

$$\langle \Delta y(l_0,l)^2 \rangle \sim l \int^{l} dx\, x^{-\mu+2}. \tag{A4}$$

For $\mu > 3$ integral in Eq. (A4) is finite and $\langle \Delta y(l_0,l)^2 \rangle \propto l$. For $2 < \mu \leq 3$ Eq. (A4) yields $\langle \Delta y(l_0,l)^2 \rangle \propto l^{4-\mu}$. For $\mu \leq 2$ we have $l_{\max} \sim l$ and $\langle \Delta y(l_0,l)^2 \rangle \propto l^2$. A rigorous derivation of Eqs. (A1), (A2), and (A3) can be found in Ref. [14].

Equation (3.3) can also be derived by studying the correlation function

$$\mathcal{C}(l) \equiv \overline{u_i u_{i+l}}, \tag{A5}$$

where the bar indicates average over all i, and u_i is the step (± 1) of the Lévy walk.

The nonvanishing contribution to $\mathcal{C}(l)$ is that u_i and u_{i+l} both belong to the same string (the expectation value for $u_i u_{i+l}$ is zero when they belong to different strings, since there is no correlation between one string and the others). Thus

$$\mathcal{C}(l) \sim \int_l^{\infty} (m-l)P(m)dm, \tag{A6}$$

where $P(m) \sim m^{-\mu}$ is the probability to find a string with length m, and $m-l$ is the number of configurations that both u_i and u_{i+l} belong to the same string. Therefore, the leading term of this integral is

$$\mathcal{C}(l) \sim \int_l^{\infty} m^{-\mu+1} dm \sim l^{-\mu+2} \tag{A7}$$

for $\mu > 2$.

Since the mean-square fluctuation is a double summation of the correlation function [1], i.e.,

$$F^2(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} \mathcal{C}(i-j), \tag{A8}$$

we obtain

$$F^2(l) \sim l^{-\mu+4}, \tag{A9}$$

and therefore $\alpha = 2 - \mu/2$.

APPENDIX B: MARKOV RANDOM WALKS

Previous studies have shown that DNA sequences exhibit short-range correlations that can be well described by a Markov chain [31]. The short-range correlation can be found both in the coding regions (where $\alpha = 0.5$ for large l) as well as in the noncoding regions. This short-range correlation affects the behavior of $F(l,L)$ for the range $l < 10$ and manifests itself through the changes of the initial slope $\alpha_0 = \alpha(l = 1, L)$ of the log-log plot of $F(l,L)$ versus l.

As we discuss in the main text, a generalized Lévy walk is an ensemble of many uncorrelated biased random walks that are spliced together, where the length of these biased random walks follows a power-law distribution. To take into account short-range correlations such as those found in DNA, we can use a biased Markov random walk instead of the pure uncorrelated biased random walk. We will discuss the procedure of generating a biased Markov random walk in this appendix.

The probability of finding a certain type of nucleotide at position i is represented by a state vector

$$X(i) = \begin{pmatrix} P_Y(i) \\ P_R(i) \end{pmatrix}, \tag{B1}$$

where $P_Y(i)$ and $P_R(i)$ are the probabilities of finding pyrimidine (Y) and purine (R) at position i, respectively. Of course,

$$P_Y(i) + P_R(i) = 1. \tag{B2}$$

For a first-order Markov chain, the evolution of a state vector can be described by a master equation [32]

$$X(i+1) = AX(i), \tag{B3}$$

where A denotes the transition matrix, i.e.,

$$A = \begin{pmatrix} P_{YY} & P_{RY} \\ P_{YR} & P_{RR} \end{pmatrix}, \tag{B4}$$

where P_{AB} is the conditional probability of finding the next nucleotide (along the chain) to be a B (either Y or R) given that the present nucleotide is an A. It is obvious that

$$P_{YY} + P_{YR} = 1, \tag{B5a}$$

$$P_{RY} + P_{RR} = 1. \tag{B5b}$$

Since the sequence is biased in nucleotide concentration, therefore, the state vector X should approach a steady state

$$X_s = \begin{pmatrix} p_+ \\ p_- \end{pmatrix}, \tag{B6}$$

where p_+ is the concentration of pyrimidine for the whole sequence and $p_- = 1 - p_+$ is the concentration of purine. Mathematically speaking, X_s is the eigenvector of the transition matrix A (with an eigenvalue 1). It is straightforward to show that, with the constraints of Eqs. (B2)–(B6),

$$A = \begin{pmatrix} 1 - p_-\lambda & p_+\lambda \\ p_-\lambda & 1 - p_+\lambda \end{pmatrix}, \tag{B7}$$

where $1 - \lambda$ is another eigenvalue, corresponding to the eigenvector $\binom{1}{-1}$, of matrix A. The physical meaning of this eigenvalue is that it defines short-range correlation length $\xi = -1/\ln|1 - \lambda|$. It can also be expressed in terms of $\alpha_0 \equiv \alpha(1, L)$, the initial slope of the $F(l, L)$ log-log plot for large values of L,

$$1 - \lambda = \frac{F^2(2, L)}{2F^2(1, L)} - 1 = 2^{2\alpha_0 - 1} - 1. \tag{B8}$$

Therefore, to model an actual DNA sequence by a Markov chain, first we need to calculate the bias for the concentration, p_+, from the actual data. Then we need to determine the parameter λ from the actual measurement of α_0 [see Fig. 3(c)]. The transition matrix A is completely determined by p_+ and λ according to Eq. (B7). Note that the case $\lambda = 1$, corresponding to $\alpha_0 = 1/2$, reduces the Markov chain to an uncorrelated biased random walk.

APPENDIX C: OTHER STATISTICAL PROPERTIES OF LÉVY WALKS

The Lévy-walk model is consistent with our previous finding [33] of Gaussian distribution of values of $\Delta y(l_0, l)$ for small $\Delta y(l_0, l) < l^{1/(\mu-1)}$. Since it was shown that for Lévy walks this quantity has a large region of Gaussian behavior near its maximum [Eq. (A2a)] and power-law tails [Eq. (A3a)] for large $\Delta y(l_0, l) > l^{1/(\mu-1)}$ (See Fig. 6). However, sufficiently long data sets are not yet available to test the tails of this distribution directly. It should be mentioned that a model of DNA evolution proposed by Li [5] does not obey Gaussian distribution of Δy and, hence, fails to describe important features of real DNA.

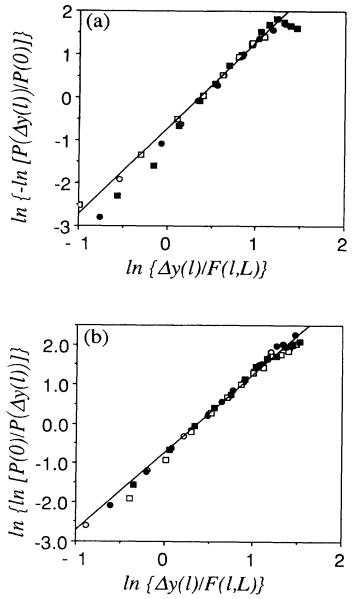

FIG. 6. Comparison of the distributions of $\Delta y(l_0, l)$ for (a) HUMHBB sequence and (b) generalized Lévy-walk model, presented in Fig. 1(c). The data presented as double-logarithmic scaling plots of the logarithm of normalized distribution vs scaling variable $\Delta y(l_0, l)/F(l, L)$, for different values of $l = 8$ (○), $l = 16$ (□), $l = 32$ (●), $l = 64$ (■). Note that for small values of l, $F(l, L)$ scales like $l^{\alpha(l,L)}$, where $\alpha(l, L) \approx 0.7$ as can be seen from Figs. 3(c) and 3(d), while from Eq. (A2) one can expect a scaling variable of the distribution to be $\Delta y(l_0, l)/l^{1/(\mu-1)}$, but for $\mu = 2.45$ $1/(\mu - 1) = 0.69$. Thus, our scaling variable approximately corresponds to the theoretical one. The degree to which the rescaled data fall on a single curve ("data collapse") shows rather good agreement with Eq. (A2) both for real DNA and model sequence. The slopes of the straight lines are equal to 2 for Gaussian distribution.

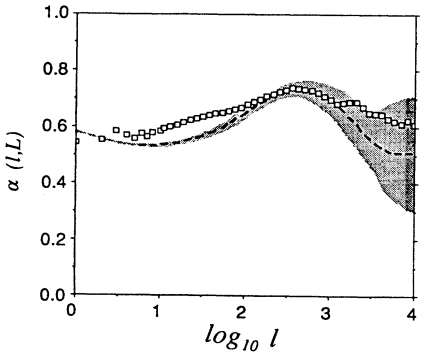

FIG. 7. Comparison of the successive slopes of the yeast chromosome III (□) with the average data for 15 realizations of the finite scale model with $l_c = 600$, $\alpha_0 = 0.58$, $\epsilon = 0.074$ (dashed line); the shaded area corresponds to two standard deviations of the data for the model [compare to Fig. 4(b)].

APPENDIX D: ALTERNATIVE MODEL WITH FINITE CORRELATION LENGTH

In order to determine whether the behavior of long DNA sequences can be explained using simple finite range correlations we change the distribution of lengths of biased random walks in our model to that of a Poisson distribution with a very large characteristic length l_c and replace rule 1 of our model by $l_j = -l_c \ln(u)$. We have systematically varied l_c between 100 and 1000 with different values of p_+ and λ but are unable to find parameter ranges that adequately fit the actual data for yeast chromosome III. The best fit is shown in Fig. 7. We found that the same is true for other long sequences that we have studied: human and rabbit beta-globin chromosomal regions, and the complete genome of human cytomegalovirus. These data are well fit by the Lévy-walk model but not by the finite correlation length model. However, for some shorter sequences such as myosin genes [33], one can fit the data for $F(l, L)$ equally well with the Lévy walk model as with the finite range model with length scale $l_c = 200$. Of interest, this length corresponds to nucleosome size. This maybe related to the complex exon-intron structure which has the property that the total length of successive intron and exon is usually a multiple of the nucleosome size [34].

For other types of DNA walks, e.g., for the bonding energy classification [with $u_i = 1$ in Eq. (2.4) for cytosine or guanine which are strongly bonded together and $u_i = -1$ for adenine or thymine which are weakly bonded], we also found that the data may be in some cases very well fitted by finite range model with $l_c = 200$ or larger l_c (see Fig. 8). However, the scaling of $F(l, L)$ for this rule of yeast chromosome III is even better than the scaling for the purine-pyrimidine rule and can be fitted only by a Lévy-walk model.

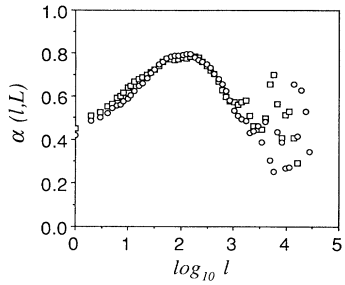

FIG. 8. Comparison of the successive slopes of the log-log plot of fluctuations of CG-AT content for HUMHBB sequence (□) and finite scale model with $l_c = 200$, $\alpha_0 = 0.4$, $p_+ = 0.57$ (○). The fluctuations in the CG-AT content were calculated according to Eqs. (2.1)–(2.4), where $u(i) = 1$ for C or G and $u(i) = -1$ for A or T.

[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] Long-range correlations in a noncoding DNA sequence were reported independently by W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[3] Long-range correlations were later confirmed by R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[4] The long-range correlations were found to extend over the entire yeast chromosome III region (315 357 nucleotides) by P. J. Munson, R. C. Taylor, and G. S. Michaels, Nature **360**, 636 (1992). The yeast chromosome III sequence was published by S. G. Oliver *et al.*, *ibid.* **357**, 38 (1992).

[5] W. Li and K. Kaneko, Nature **360**, 635 (1992).

[6] A. Yu. Grosberg, Y. Rabin, S. Havlin, and A. Nir, Biofiz. **38**, 75 (1993); Europhys. Lett. (to be published).

[7] For an ideal sequence of infinite length, $\alpha = 0.5$ indicates absence of long-range correlation, and $\alpha \neq 0.5$ corresponds to long-range correlation. But for a sample of finite length, we have to take into account the statistical fluctuations due to finite size. Therefore, we consider a DNA sequence to exhibit long-range correlation only if the value of α is significantly deviated from 0.5. For an uncorrelated sequence, it is easy to show that the statistical fluctuations, i.e., standard deviation $\delta\alpha$, of the exponent α, estimated by linear regression fit to Eq. (2.5) for data $1 \leq l \leq x$, follows a simple relation: $\delta\alpha = C\sqrt{x}/(\sqrt{L}\ln x)$. Numerical estimation shows $C \approx 0.6$. Thus for a random walk of length $L = 6000$, the standard deviation is $\delta\alpha \approx 0.035$ for α estimated from $l = 1$ to 1000. In other words, roughly one out of ten realizations of random walks of length 6000 will give the exponent $\alpha \geq 0.55$ [provided that α is estimated from the first three decades of data in $F(l)$ plot]. See also C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3729 (1993); Physica **191**, 25 (1992).

[8] J. Jurka, T. Walichiewicz, and A. Milosavljevic, J. Mol. Evol. **35**, 286 (1992).

[9] M. F. Shlesinger and J. Klafter, in *On Growth and Form: Fractal and Non-Fractal Pattern in Physics*, edited by H. E. Stanley and N. Ostrowsky (Martinus Nijhoff, Dordrecht, 1986), p. 279.

[10] M. F. Shlesinger, J. Klafter, and Y. M. Wong, J. Stat. Phys. **27**, 499 (1982).

[11] M. F. Shlesinger and J. Klafter, Phys. Rev. Lett. **54**, 2551 (1985).

[12] S. Havlin, S. Buldyrev, H. E. Stanley, and G. H. Weiss, J. Phys. A **24**, L925 (1991).

[13] R. N. Mantegna, Physica A **179**, 232 (1991).

[14] M. Araujo, S. Havlin, G. H. Weiss, and H. E. Stanley, Phys. Rev. A **43**, 5207 (1991).

[15] C. K. Peng, J. Meitus, J. M. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993).

[16] We will focus on the adjustment of the two new parameters that were introduced in the Lévy-walk model: l_c and μ. The other two parameters ϵ and α_0 (related to the parameters for the Markov chain: p_+ and λ, see Appendix B) can be precisely determined from the strand-biased regions in the actual DNA sequences. The parameter l_c can be estimated from the length of the largest strand-biased region in an actual DNA sequence of length L. We find that l_c is around 10 for almost all sequences we stud-

ied. The value of μ has to be chosen slightly larger than it is estimated, according to Eq. (3.3), from the plateau region (maximum) of $\alpha(l, L)$ in the actual data. This is because in generalized Lévy walks, the average value of $\alpha(l, L)$ for finite L is always smaller than the asymptotic value of $\alpha = 2 - \mu/2$.

[17] J. Jurka, J. Mol. Evol. **29**, 496 (1989).

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

[19] We removed all known repetitive regions and their fragments (constituting about $\frac{1}{3}$ of the total DNA chain length) in the human beta-globin intergenomic sequence. When the remaining nucleotides are spliced together they still exhibit long-range (scale-invariant) correlations. This test suggests that (i) the repetitive sequences are not the only source for the long-range correlations (e.g., the so-called minisatellite repeats [18] may also contribute to the long-range correlations), or (ii) our knowledge of the repetitive sequences is not complete.

[20] H. R. Hwu, J. W. Roberts, E. H. Davidson, and R. J. Britten, Proc. Natl. Acad. Sci. U.S.A. **83**, 3875 (1986).

[21] E. Zuckerkandl, G. Latter, and J. Jurka, J. Mol. Evol. **29**, 504 (1989).

[22] B. Levin, *Genes IV* (Oxford University Press, Oxford, 1990).

[23] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979).

[24] J. de Cloiseaux, J. Phys. (Paris) **41**, 223 (1980).

[25] S. Redner, J. Phys. A **13**, 3525 (1980).

[26] A. Baumgartner, Z. Phys. B **42**, 265 (1981).

[27] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).

[28] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).

[29] F. McCrackin, J. Mazur, and C. M. Guttman, Macromolecules **6**, 859 (1973).

[30] T. M. Birshtein and S. V. Buldyrev, Polymer **32**, 3387 (1991).

[31] S. Tavaré and B. W. Giddings, in *Mathematical Methods for DNA Sequences*, edited by M. S. Waterman (CRC, Boca Raton, 1989), pp. 117–132 and references therein.

[32] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1970-1971), Vols. 1 and 2.

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

[34] I. Ioshikhes, A. Bolshoy, and E. N. Trifonov, J. Biomolec. Struct. Dyn. **9**, 1111 (1992); E. Trifonov (private communication).

[35] C.A. Chatzidimitriou-Dreismann and D. Larhammar, Nature **361**, 212 (1993).

[36] The Lévy-walk model also accounts for the DNA "patchiness," discussed by S. Nee, Nature **357**, 450 (1992) and S. Karlin and V. Brendel, Science **259**, 677 (1993).

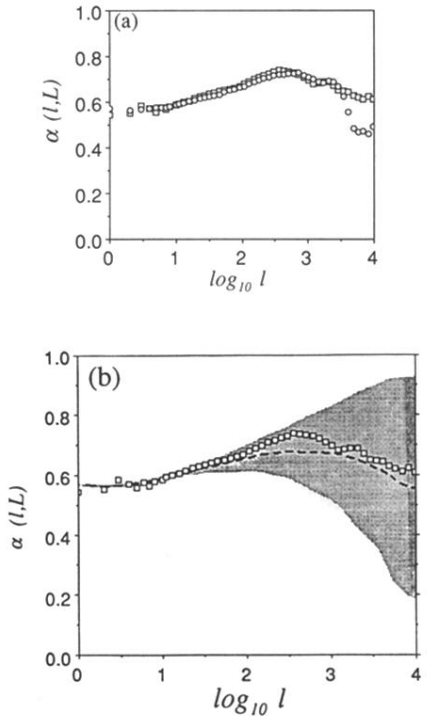

FIG. 4. Comparison of successive slopes of the scaling exponent α for yeast chromosome III (□) and (a) successive slopes of a realization of the generalized Lévy walk with parameters $L = 315\,000$, $\mu = 2.5$, $l_c = 5$, $\alpha_0 = 0.55$, $\epsilon = 0.16$ (○); (b) average successive slopes over 15 different realization of Lévy walks with the same parameters (dashed line). The shaded area corresponds to two standard deviations of successive slopes of the model, calculated for 15 random realizations. The parameters for Markov process, α_0 and ϵ, used in the model are calculated from real DNA sequence of yeast chromosome III (see Appendix B).

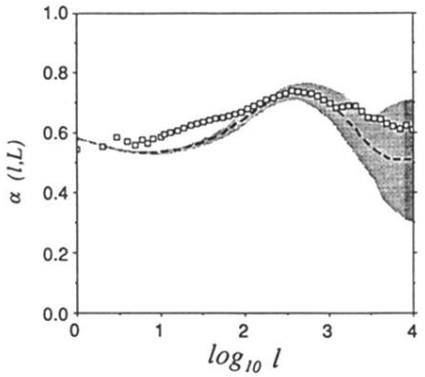

FIG. 7. Comparison of the successive slopes of the yeast chromosome III (□) with the average data for 15 realizations of the finite scale model with $l_c = 600$, $\alpha_0 = 0.58$, $\epsilon = 0.074$ (dashed line); the shaded area corresponds to two standard deviations of the data for the model [compare to Fig. 4(b)].