Fractional Brownian motion as a nonstationary process: An alternative paradigm for DNA sequences

Paolo Allegrini,[1] Marco Buiatti,[2] Paolo Grigolini,[1,2,3] and Bruce J. West[1]
[1]*Center for Nonlinear Science, University of North Texas, P.O. Box 5368, Denton, Texas 76203-5368*
[2]*Dipartimento di Fisica dell'Università di Pisa, Piazza Torricelli 2, 56100, Pisa, Italy*
[3]*Istituto di Biofisica del Consiglio Nazionale delle Ricerche, Via San Lorenzo 26, 56127 Pisa, Italy*

(Received 15 April 1997; revised manuscript received 15 July 1997)

The long-range correlations in DNA sequences are currently interpreted as an example of *stationary* fractional Brownian motion (FBM). First we show that the dynamics of a dichotomous stationary process with long-range correlations such as that used to model DNA sequences should correspond to Lévy statistics and not to FBM. To explain why, in spite of this, the statistical analysis of the data seems to be compatible with FBM, we notice that an initial Gaussian condition, generated by a process foreign to the mechanism establishing the long-range correlations and consequently implying a departure from the stationary condition, is maintained approximately unchanged for very long times. This is so because due to the nature itself of the long-range correlation process, it takes virtually an infinite time for the system to reach the genuine stationary state. Then we discuss a possible generator of initial Gaussian conditions, based on a folding mechanism of the nucleic acid in the cell nucleus. The model adopted is compatible with the known biological and physical constraints, namely, it is shown to be consistent with the information of current biological literature on folding as well as with the statistical analyses of DNA sequences.

[S1063-651X(98)00404-8]

PACS number(s): 87.10.+e

I. INTRODUCTION

One of the most successful models of the statistics of DNA sequences has been the DNA walk. A DNA sequence is a chain of sites, each occupied by either a purine or a pyrimidine. If we read the sequence in order and we regard the site number as a ''discrete time'' this symbolic sequence can be regarded as a dichotomous time series. Thus one can imagine a random walker whose dispacement x at the jth step increases by $+1$ if the DNA site is occupied by a purine or decreases by the same amount if the site is occupied by a pyrimidine. In spite of its success in modeling the long-range correlations observed in DNA sequences [1–5], as indicated by the power-law increase in the variance and the inverse power-law spectrum [2–5], the problem of the correct statistical interpretation of the DNA walk is still unsolved and is attracting the attention of an increasing number of investigators.

A subject of intense debate is the question of the existence of long-range correlations in exons, that part of the sequence that codes for proteins. However, problems also arise at the level of the theoretical analysis of the data; in particular, most of the analytical methods assume that the data are stationary, which in the DNA context actually means spatially homogeneous. This means that if the long-range correlations reflect some internal ''rules,'' these rules apply to the whole DNA sequence with no dependence at all from the specific position in the sequence. In addition, it is very often assumed that the statistics of the random-walk landscape variable is Gaussian. In addition to its biological implications, this question of the statistics of the process with long-range correlations has implications for other fundamental phenomena such as anomalous diffusion. In particular, one could equivalently describe the scaling property of such processes using fractional Brownian motion (FBM) or alternatively using Lévy stable processes. Although these two processes share strikingly common characteristics, such as the power-law growth of the second moment and the fractal dimension of the trajectories [6], they are fundamentally different in nature.

Different initial conditions can be realized by assigning to different sites of the DNA chain the role of departure point of the walker. This means that having available sufficiently long sequences, it is possible to realize a condition equivalent to that ordinarily adopted in statistical mechanics, namely, the observation of a large number of trajectories. This makes it possible to average over trajectories with different initial conditions and thus evaluate the second moment of the distance x traveled by the DNA random walker in a time t:

$$\langle x^2\rangle \sim t^{2H}, \tag{1}$$

with $H>1/2$. For this reason the overwhelming majority of researchers working on the statistics of DNA sequences agree that the conventional theory of Brownian motion cannot be considered as a proper paradigm to interpret the experimental data, since in that case $H=1/2$.

Which is the proper physical paradigm behind the DNA sequences then? As mentioned earlier, the current literature in the field essentially affords the following two proposals: (a) the physical paradigm of the α-stable Lévy processes [5] and (b) the physical paradigm of the FBM [7]. We think that the two proposed paradigms conflict with one another, and the main aim of this paper is to settle the problems posed by this conflict. We shall develop a model resulting in nonstationary properties for the DNA sequence. These nonstationary properties are realized through a shuffling procedure that can be interpreted as the influence of the geometrical folding of the macromolecule in the nucleus. The procedure will be discussed in detail in Sec. III.

A. The physical paradigm of the α-stable Lévy processes

Let us discuss the former paradigm first. In a recent paper Buldyrev *et al.* [8], to interpret the long-range correlation in noncoding DNA, have adopted a generalization of the Lévy walk proposed in an earlier paper by Araujio *et al.* [9]. Instead of taking l_j steps in the same directions as occurs in a classic Lévy walk, the walker takes each of l_j steps in random directions, with a fixed bias probability

$$P_+ = \frac{1+\epsilon_j}{2} \tag{2}$$

to go up and

$$P_- = \frac{1-\epsilon_j}{2} \tag{3}$$

to go down, where ϵ_j gets the value $+\epsilon$ or $-\epsilon$ randomly. Throughout this paper we shall be referring to this model as a generalized Lévy walk (GLW).

It is interesting to point out that Allegrini *et al.* [5] have used a model, called a copying mistake map (CMM), which is totally equivalent to the GLW. Let us discuss this aspect in detail. The CMM assumes that the DNA sequence results from the random joint action of two different prescriptions, one responsible for the long-range correlations and the other of a totally ''random'' nature, implying no correlations at all. The probability of running the sequence with the prescription generating correlations is p_c and the probability of running the sequence with the random law is $1-p_c$.

The CMM is equivalent to the GLW. The equivalence between the two models is made evident by noticing that if the CMM is adopted the probabilities of going up and down are

$$P_+ = \frac{1+p_c}{2} \tag{4}$$

and

$$P_- = \frac{1+p_c}{2}, \tag{5}$$

respectively, thereby implying that ϵ is identified with p_c.

The corresponding equation of motion is easily written by using the results of some recent papers [10,11]. Let us consider the simplest equation generating diffusion

$$\dot{x}(t) = \xi(t), \tag{6}$$

where x is the diffusing variable and ξ the source of fluctuation, supposed to be a dichotomous variable with the values $\xi = +1$ and $\xi = -1$. In the case of the DNA sequence this dichotomous property is dictated by the way it defines the DNA walk. The continuous time representation (6) becomes natural when the sequences studied are long enough. Let us also make the assumption that a stationary, single-time autocorrelation function

$$\Phi_\xi(t) = \frac{\langle \xi(0)\xi(t)\rangle}{\langle \xi^2\rangle} \tag{7}$$

exists and has the asymptotic property

$$\lim_{t\to\infty} \Phi_\xi(t) \propto \frac{1}{t^\beta}, \tag{8}$$

with

$$0 < \beta < 1. \tag{9}$$

Note that this is the simplest analytical expression breaking the integrability condition, which in turn is responsible for the generation of ordinary Brownian motion. Let us integrate Eq. (6) and use the resulting expression to evaluate the second moment of x. It is straightforward to show [11] that the stationary assumption and the property (8) yield Eq. (1) with

$$H = 1 - \frac{\beta}{2}. \tag{10}$$

The second moment does not exhaust all the statistical properties of this process. The important result of [10] is that a complete statistical description of this diffusion process is given by the equation of motion for the probability density $\rho(x,t)$

$$\frac{\partial \rho(x,t)}{\partial t} = \langle \xi^2\rangle \int_0^t dt' \Phi_\xi(t-t') \frac{\partial^2}{\partial x^2}\rho(x,t'). \tag{11}$$

This equation is exact and rests on the assumption that the fluctuation process ξ is stationary and dichotomous. The latter property is obviously fulfilled by the DNA sequences.

It is thus evident that both models, the GLW model and the CMM model, are described by Eq. (11). On the other hand, if we adopt the GLW model, we see that for the whole process characterized by a fixed value of ϵ_j, ϵ, the variable

$$\tilde{\xi} \equiv \xi - \epsilon \tag{12}$$

behaves like ordinary white noise with no bias. If we take into account that also the bias ϵ, on a much larger time scale, undergoes a fluctuation process, we get

$$\xi(t) \equiv \tilde{\xi}(t) + \epsilon(t). \tag{13}$$

This means that the fluctuation $\xi(t)$ is the sum of a quickly fluctuating process, with no correlation $\tilde{\xi}(t)$ and a process with long-time fluctuations $\epsilon(t)$ characterized by the autocorrelation function $\Phi_\epsilon(t)$, with the long-time property

$$\lim_{t\to\infty} \Phi_\epsilon(t) \propto \frac{1}{t^\beta}. \tag{14}$$

Thus the autocorrelation function $\Phi_\xi(t)$ determining the statistical properties of the process through Eq. (11) reads

$$\Phi_{\xi}(t) = (1-\epsilon^2)\Phi_{\tilde{\xi}}(t) + \epsilon^2 \Phi_{\epsilon}(t), \tag{15}$$

where $\Phi_{\tilde{\xi}}(t)$ denotes the autocorrelation function of the fast contribution to the fluctuation ξ. Of course, if the CMM interpretation is adopted, this prescription must be rewritten as

$$\Phi_{\xi}(t) = (1-{p_c}^2)\Phi_{\tilde{\xi}}(t) + {p_c}^2 \Phi_{\epsilon}(t). \tag{16}$$

This means that if the dynamics of the DNA sequence is determined by the joint action of a random prescription and a prescription generating correlations, and the statistical weight of the latter, p_c, is small, then the statistical effect on the correlation function of ξ is still smaller. This property was overlooked in Ref. [5].

Note that to realize an artificial sequence mimicking a real DNA sequence in [5] a deterministic mapping similar to that of Geisel, Heldstap, and Thomas [12] was used. The Geisel-Heldstap-Thomas (GHT) map will be used here, according to the prescriptions of Sec. III. However, in no way does the adoption of this map imply the assumption that the DNA sequence might be generated by a deterministic rule. This is so because, as indicated by Eq. (11), the statistical properties of the sequence are only determined by the correlation function $\Phi_{\xi}(t)$ and, once its decay properties are fixed, these statistical properties too are fixed, whatever the dynamical nature of the process driving the motion of $\Phi_{\xi}(t)$ might be. We note furthermore that, in principle, a slow motion with the same inverse power law as Eq. (8) can be generated by a random model. For instance, the random activation energy model [13] can result in slow motion with the same negative power as Eq. (8). The numerical calculations carried out in [10] are based on a random generator of the inverse power-law behavior (8). In conclusion, the adoption of either a random or a deterministic generator to produce a given slow motion does not imply that the DNA dynamics is interpreted as either random or a deterministic process. We limit ourselves to saying that Eq. (15) is a mixture of short- and a long-range correlation fluctuations.

Note that if $\epsilon=1$ and only the long-range contribution to the correlation function $\Phi_{\xi}(t)$ is present, Eq. (11) is proven [10] to generate as a diffusion distribution a truncated Lévy process, namely, a distribution with a central part given by a genuine Lévy distribution, and thus with tails with an inverse power law with the power $\beta+2$. These tails are truncated by ballistic peaks, which reflect the fact than no trajectories can exist traveling faster than those with velocity equal to $|\xi| = 1$. This kind of diffusion process implies a significant deviation from Gaussian statistics.

B. The physical paradigm of FBM

The assumption that a process with the same long-range correlation, and consequently with the same $H > 1/2$ as those discussed above, is Gaussian implies immediately that the corresponding statistical equation of motion reads

$$\frac{\partial \rho(x,t)}{\partial t} = \langle \xi^2 \rangle \left[\int_0^t dt' \, \Phi_{\xi}(t') \right] \frac{\partial^2}{\partial x^2} \rho(x,t). \tag{17}$$

This is so essentially for the following reasons. First of all, we note that the second moment obeys the same equation of motion as that of the second moment generated by Eq. (11). This means, therefore, that Eq. (17) with the same autocorrelation function $\Phi_{\xi}(t)$ as Eq. (11) leads to the same asymptotic expression (1) and to the same coefficient H. On the other hand, it is easy to prove [11] that the solution to Eq. (17) is given by a Gaussian distribution with a width proportional to second moment and, consequently, obeying the time asymptotic prescription (1).

C. The search for a different physical paradigm

All researchers in this field of investigation admit the existence of long-range correlations in the DNA sequences. This, according to Ref. [10], would imply a strong deviation from Gaussian statistics, while the investigation of Arneodo *et al.* [14] yields as an important conclusion that the DNA statistics are essentially Gaussian.

On the other hand, the CMM model does not seem to be totally satisfactory. There are two reasons why this model has to be refined. First, the ‘‘copying mistake’’ rate predicted by the model illustrated in [5] is very high and it is not clear if it is compatible with what is known from biology. Second, the stationary assumption is questionable from a biological point of view. In fact, it states that the correlation between two nucleotides depends only on their distance along the primary string and does not depend on the position of the nucleotides. This is very strange because we imagine that the origin of the long-range correlation itself is a consequence of the tertiary structure of the DNA polymer [15] or, in other words, of its self-similar folding structure. It is expected that the short-range statistical and correlation properties, namely, the local properties of a given region of the folded DNA molecule, might depend on the region considered and might vary with moving from one region of the folded molecule to another. We propose a folding model that, although schematic, contains the essential ingredient of a self-similar hierarchical structure responsible for a sort of short-range randomizing process. This model causes the breakdown of the stationary assumption, implying, therefore, that the paradigm of the FBM can be adopted provided, at the same time, the stationary assumption is rejected.

The outline of the paper is as follows. In Sec. II we illustrate the time evolution of initial Gaussian conditions due to a dynamics driven by a long-range correlated dichotomous process and we see that the statistics remain Gaussian for extended times. We shall refer to this behavior as viscosity or a viscous dependence on the initial Gaussian condition. In Sec. III we illustrate the folding model for the DNA molecule, we show how to generate through it a proper sequence, and we discuss the statistics of this sequence. In Sec. IV we discuss why in our opinion this model can be applied to DNA statistics and in Sec. V, finally, we make some concluding remarks.

II. VISCOSITY OF GAUSSIAN INITIAL CONDITIONS

In this section we discuss the diffusion effects produced by a theoretical model with a dichotomous random walker moving as a traditional random walker from $t<0$ to $t=0$ and as a dynamical generator of Lévy diffusion from $t=0$ on. This is equivalent to studying the anomalous diffusion process with an initial condition given by the Gaussian distribution

$$P(x,0)=\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}, \tag{18}$$

where σ^2 is the variance of the initial distribution. The analytic calculation of the diffusion process initiated by this condition can be easily accomplished. An analytic expression for the Green's function is now available [10] and it is given by the inverse Fourier transform of the stationary characteristic function

$$G(x,t)=\left(\frac{1}{2\pi}\int_{-\infty}^{+\infty}dk\ e^{ikx}e^{-b|k|^{\alpha}t}\right)\theta(t-|x|)+\frac{\Phi_{\xi}(t)}{2}\delta(t-|x|), \tag{19}$$

where

$$\alpha=1+\beta, \tag{20}$$

with β fulfilling condition (9). The parameter b in Eq. (19) depends only on β and on the short-time properties of the correlation function (7). If we choose an inverse–power-law generator with the short-time structure

$$\Phi_{\xi}(t)=\frac{A}{(A^{1/\beta}+t)^{\beta}}, \tag{21}$$

it is shown [10] that

$$b=\frac{\pi\beta(\beta+1)AW^{\beta+1}}{2\sin\left(\frac{\pi(\beta+1)}{2}\right)\Gamma(\beta+2)}. \tag{22}$$

Note that due to Eq. (20) the asymptotic property (8) becomes

$$\lim_{t\to\infty}\Phi_{\xi}(t)\propto\frac{1}{t^{\alpha-1}}. \tag{23}$$

The probability distribution $P(x,t)$ is obtained from the space convolution integral between Eqs. (18) and (19). We replace the Heaviside step function on the right-hand side (rhs) of Eq. (19) with 1. Then the first term on the rhs of Eq. (19) becomes an α-stable Lévy process and thus becomes responsible for the distorsion of the initial Gaussian shape and for the birth of long tails. The second term on the rhs of Eq. (19) produces two peaks that correspond to the initial Gaussian distribution shifted backward and forward by the quantity $\langle\xi^2\rangle t$. The amplitude of these two duplicates of the initial distribution decays as the correlation function (21).

All this is illustrated by Fig. 1. We see from Fig. 1 that the Lévy nature of the central part of the distribution becomes evident only after the transition from the one-mode to the three-mode shape. We can estimate this time as that necessary for the ballistic peak to travel a distance comparable to the half-width of the initial Gaussian distribution. This time can therefore be made arbitrarily large by increasing the width of the initial Gaussian distribution.

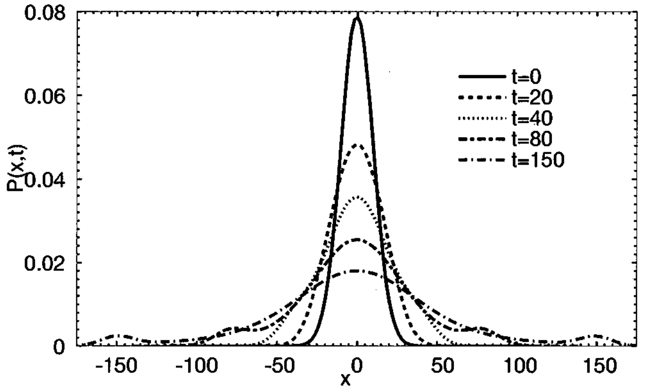

FIG. 1. From Gauss to Lévy. The numerical convolution between the initial condition (18) with $\sigma=50$, and the Green's function (19) with $\alpha=1.5$, at different times. The initial Gaussian represents the result of an earlier diffusion process generated by a dichotomous but totally uncorrelated fluctuation. The subsequent time evolution, responsible for the birth of ballistic peaks, is pursued by the stochastic generator of [10]. The parameter A of Eq. (21) is obtained by fitting the experimental results with Eq. (21) and turns out to be $\simeq 0.5$.

III. THE BETHE LATTICE

The main result of Sec. II is that the FBM can be approximately realized provided the system is not forced to fulfill the stationary condition. This is so because the decoupling of statistics, assumed to be Gaussian, from dynamics can only be realized in a nonstationary regime. The initial Gaussian condition assumed in Sec. II can be the result of uncorrelated fluctuations acting at times preceding the observation time, namely, for $t<0$. For times $t>0$ the diffusion process is determined by a dichotomous fluctuation with long-range correlation and, consequently, according to [10] should become a truncated Lévy process. However, the viscosity of the initial condition results in an approximated realization of the FBM for an extended period of time.

The purpose of this section is to build a model realizing conditions similar to these as well as effects similar to the joint action of a dichotomous fluctuation, with long-range correlations and a short-range random process. This means that the model has to account for both the same statistical properties as those simulated by the CMM and the GLW model and the Gaussian character of the resulting statistics. Note that the CMM (and of course the equivalent GLW model as well) would depart from the Gaussian statistics in the long-time limit. The Gaussian character of the model of this section is expected to be much more viscous. We shall see that both effects, correlations and Gaussian statistics, can be reproduced by a shuffling of the sequence according to certain geometrical prescriptions. We shall see in Sec. IV that these prescriptions are based on plausible assumptions on the folding of the DNA macromolecule.

We imagine a two-dimensional array of sites, each one carrying a value of either $+1$ or -1. These sites can be used to generate an ordered sequence by making a given trajectory visit them one after the other. The procedure defines a sequence ξ_i, where ξ_i is the value of the ith site visited by the trajectory. Notice that different trajectories define different numerical sequences.

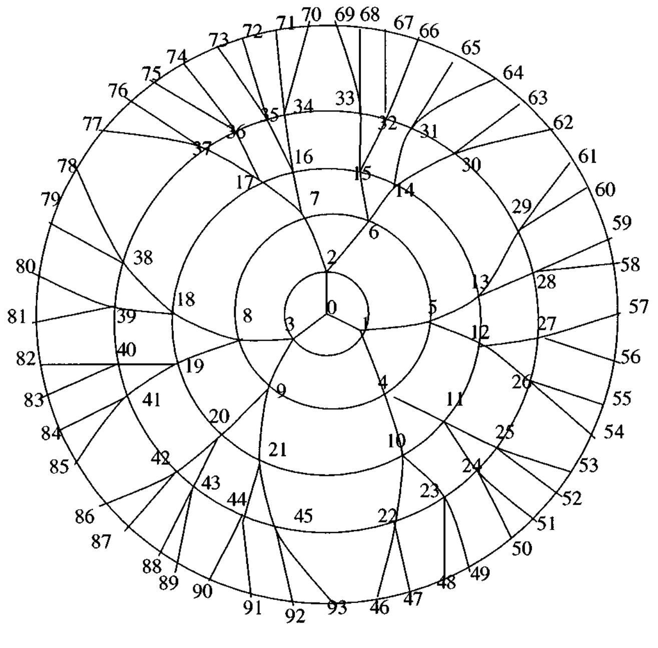

FIG. 2. The Cayley tree. Each site is connected with three other sites. Here the tree is plotted having in mind a circular symmetry and is drawn up to 5 circular shells due to the space limitations. In the numerical simulations herein we considered 17 shells.

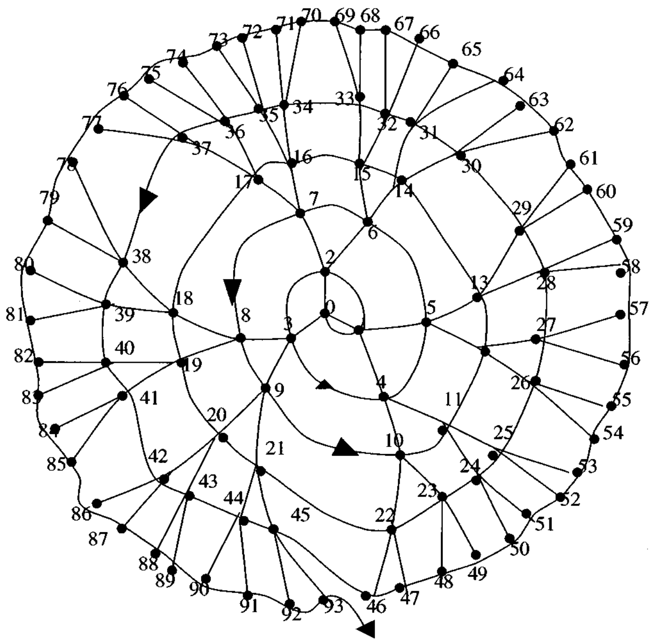

FIG. 3. The Cayley tree: The sequence is generated by the regular trajectory. We see that the numbers of the nodes have been assigned along a spiral (solid line) starting from the center of the tree.

The array of sites is generated by means of a Bethe lattice (or Cayley tree) [16,17]. This lattice is used in percolation theory [17] as an exactly solvable model sharing with lattices of arbitrarily large dimensions the property that a given site is surrounded by a given number of nearest neighbors. This number is called the coordination number and is denoted by the symbol z; it can be arbitrarily large. For this reason, Abou-Chacra, Anderson, and Thouless [18] used this model to discuss the phenomenon of localization in the case of lattices of arbitrarily large dimensions. The Cayley tree has been used recently to model the connectivity of dendrimer molecules, such as some possible configurations of biological macromolecules [19], and physical processes bearing some connections with the electron transfer in DNA [20].

We choose a coordination number $z=3$, meaning that each site has three nearest neighbors; see Fig. 2. We give the sites an ordering number, starting from a site that is assigned the number 0, and then we proceed with our ordering procedure following a path with an approximate circular symmetry around the initial site. The first layer of this circular structure around site 0 consists of three nearest neighbors, which are given the numbers from 1 to 3 going counterclockwise. The second layer has six sites that are numbered counterclockwise from 4 to 9, starting from the two nearest neighbors of 1. From now on the ordering rule is the same everywhere: We number the sites of an outer layer going counterclockwise and starting from the nearest neighbors of the site of the previous layer with the smallest ordering number. This procedures is described in Fig. 2 for the first five layers and can be easily applied to a lattice with an arbitrarily large number of sites. In our simulations, however, we have considered a finite lattice, with 17 layers, with a total number of sites given by $1+3+6+\cdots+3\times 2^{17}=393\,214$.

The ordering of the sites illustrated in Fig. 2 makes it natural to define as a trajectory the spiral-like path denoted by the solid line of Fig. 3. This regular trajectory is obtained following the numbering prescription of Fig. 2. Thus it starts from 0, makes a jump to the first layer, and rotates counterclockwise, visiting sites 1, 2, and 3, then it makes a jump to the next layer, and so on. This spiral is an imperfect but close realization of a structure with central symmetry. We make the assumption that the *stationary* long-range correlated fluctuations, generated by the GHT map, are distributed along this spiral-like path. The translational invariance implicit in the stationary condition is thus reflected into a rotational symmetry for the statistical properties of the dichotomous values carried by the sites of the Bethe lattice. The GHT map is adopted to build up these correlated fluctuations according to the prescriptions

$$y_{m+1}=f(y_m), \tag{24}$$

where

$$f(y)=\begin{cases} y+ay^{\zeta} & \text{for } 0\leq y\leq d \\ y+ay^{\zeta}-1 & \text{for } d<y<1/2 \\ y+1-a(1-y)^{\zeta} & \text{for } 1/2\leq y<1-d \\ y-a(1-y)^{\zeta} & \text{for } 1-d\leq y\leq 1 \end{cases} \tag{25}$$

and d is defined implicitly by means of $d+ad^{\zeta}=1$ and $a=2^{\zeta}$. The fluctuating variable $\xi_m^{(0)}$ takes the values $+1$ or -1, thereby resulting in a noise with intensity $\langle(\xi^{(0)})^2\rangle$, and is determined by

$$\xi_m^{(0)}=2[2y_m]-1, \tag{26}$$

where [ ] denotes the integer value. Note that Eq. (26) defines a coarse graining for the map dynamics since the interval $0<y_m<1/2$ is mapped onto the value $\xi_m^{(0)}=-1$ and the

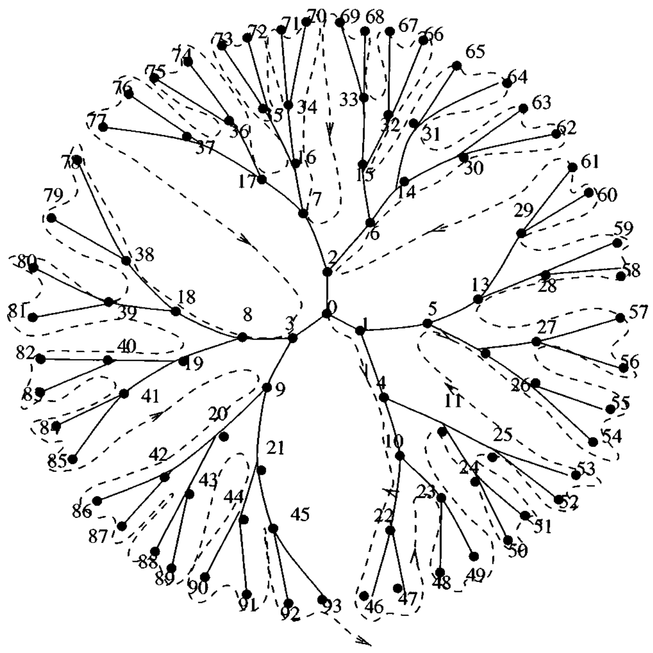

FIG. 4. The Cayley tree: The sequence is generated by the irregular trajectory. We imagine a DNA molecule folded around the tree. The Cayley tree is an extremely simplified model for the dendritic structure of a real protein matrix. The rule for the trajectory, denoted by the dotted line, is to explore the nodes of the tree without intersecting the tree, keeping the graph on the left-hand side, and avoiding the sites already explored.

interval $1/2\leqslant y_m<1/2$ is mapped onto the value $\xi_m^{(0)}=+1$. The superscript (0) is adopted to point out that this sequence corresponds to the spiral-like path of Fig. 3.

The map (24) is very similar (it has the same laminar behavior) to the map adopted in the CMM model [5] without copying mistakes, namely, in the case where $p_c=1$ in Eq. (16). The GHT map has been shown to give rise to an α-stable Lévy process, in the sense pointed out earlier, namely, with ballistic fronts [21,22], with Lévy index $\alpha=1/(\zeta-1)$. However, as can be proved with the arguments of [10], this result is independent of the details of the map and is expected to be produced by all the maps or stochastic generators resulting in the same correlation function $\Phi_\xi(t)$.

We have used the GHT map for computational simplicity and because it is well known in the literature, but the resulting statistics are virtually identical to the statistics generated with the stochastic generator of Ref. [10]. Due to the chaotic nature of the map and the crude coarse graining it is impossible to distinguish the deterministically generated sequence from the stochastic one. In the stationary case [10,21] they have the same correlation function and consequently the same statistical properties. This case was shown in Refs. [10, 21] to generate a truncated Lévy process, namely, a diffusion distribution with a central Lévy-like structure, but with the tails replaced by two ballistic peaks, corresponding to the abrupt truncation that any dynamically generated distribution must have.

Another trajectory visiting all the sites is schematically shown in Fig. 4, where again only five layers are considered. The prescription adopted to define this path is that it keeps moving with the tree on its left-hand side and skipping the sites already explored. Thus we see that the path moves from site 0, the initial site, to sites 1, 4, 10, 22, and 46. At site 46 it meets the surface of the graph, and according to the rule of keeping the tree on the left-hand side, it goes to site 47. According to the same rule, it should reenter the graph to explore sites 22 and 10. Since these sites were already visited by the trajectory, they are skipped and the trajectory goes to site 23 and so on.

Using this disordered path we can define a new sequence ξ_m. The values of this sequence are $\xi_0=\xi_0^{(0)}$, $\xi_1=\xi_1^{(0)}$, $\xi_2=\xi_4^{(0)}$, $\xi_3=\xi_{10}^{(0)}$, and so on.

Notice that with this choice of trajectory the central symmetry ordering is largely modified. If we determine the values of the sites in such a way that a certain translational invariance is fulfilled by the sequence $\xi_i^{(0)}$, generated by the spiral of Fig. 3, following the natural ordering of the plot, it is likely that this property in not fulfilled any longer by the sequence ξ_i generated by the ''folded'' trajectory discussed above and shown in Fig. 4.

An important result of this paper is that the particular reordering of the values of the natural sequence, stationary by construction and long-range correlated, into the irregular one generates a diffusion process of Gaussian nature. In Fig. 5 we show two plots with the histograms of the diffusion process for $t=25$ and $t=100$, respectively. In both cases the truncated Lévy process with ballistic peaks generated by the GHT on the natural sequence $\xi_i^{(0)}$ (upper curves) collapses to a Gaussian function, as pointed out by the linear-logarithmic nature of the plot. This fact is, in our opinion, remarkable because a numerical evaluation of the correlation function of the irregular sequence ξ_i reveals a good agreement with the theoretical inverse power law imposed on the $\xi_i^{(0)}$'s, as we can see in Fig. 6. In Fig. 6, in fact, we show the correlation function of the irregular process together with an inverse power law proportional to $t^{-1/2}$; this is a guideline corresponding to the theoretical shape of the stationary correlation function for our generator, the GHT map of Eq. (23) with $\zeta=5/3$. We see that the distinctive feature of the unfolding procedure is the emergence of a structure that is periodic with respect to $\log t$ and thus compatible with a renormalization-group approach [23]. However, no implication of this theory is explored herein.

To make complete our discussion of the statistical properties of the disordered sequence of Fig. 4, let us study the distribution $P(x,t)$ and its rescaling properties. The general property to investigate is

$$P(x,t)=\frac{1}{t^\delta}F\left(\frac{x}{t^\delta}\right), \tag{27}$$

where F is a generic function. For FBM it is known [7] that F is Gaussian and

$$\delta=H=1-\frac{\beta}{2}. \tag{28}$$

In Fig. 7(a) the distribution relative to $t=25$ is shown together with a Gaussian curve that was fitted to the data. In Fig. 7 the *same* Gaussian curve has been rescaled according to Eq. (27) with the condition (28) and compared with the

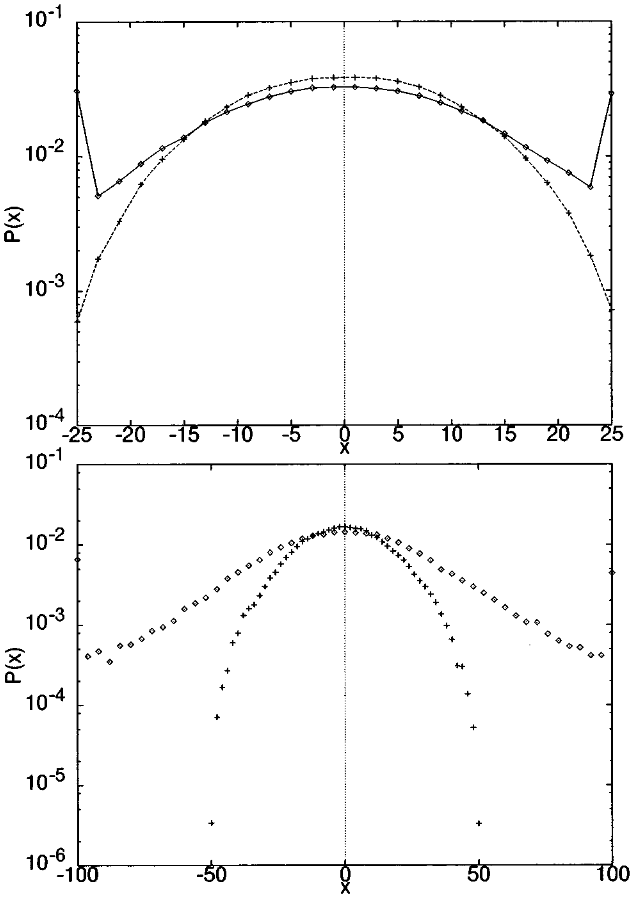

FIG. 5. Comparison of the space distributions $P(x)$ of the regular vs the irregular process at two different times. Upper figure: $t=25$, the curve with the side peaks is the GHT map generated process (through the regular trajectory), while the other curve is the ''unfolded DNA'' process, generated through the irregular trajectory of Fig. 4. Lower figure: same as before, with $t=100$. Here and in the following figures time t and space x are expressed as dimensionless quantities. The biological interpretation of t is the length of a DNA segment measured in base pairs, while the space x is the difference between the number of purines and that of pyrimidines in the segment.

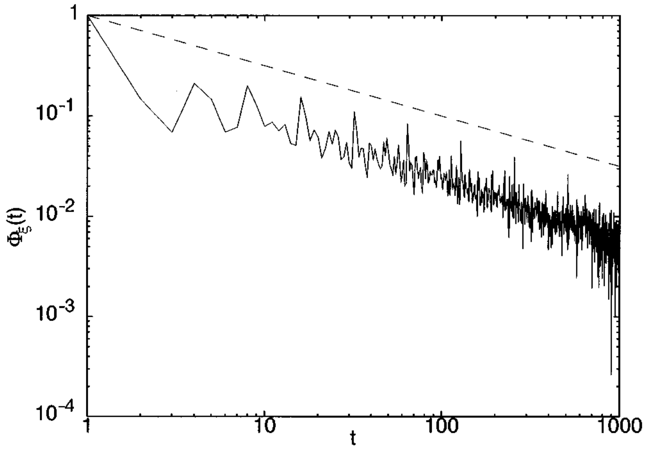

FIG. 6. Correlation function of the ''unfolded DNA'' sequence (irregular trajectory). The dashed line is a guide to the eye and corresponds to the slope -0.5, relative to the theoretical prediction for the *folded* sequence (the spiral). We see that the two curves have the same slope, but there is the emergence of a structure that is periodic with respect to $\log_{10}t$.

model data at $t=50$, 100, and 200. We see that in the regime where the Gaussian distribution is recovered, so is the rescaling of the FBM.

Note that in Ref. [10] it was proved that for a truncated Lévy process, which is the stationary solution of the process stemming from Eqs. (6) and (9), the prescription (27) is not exactly fulfilled because of the presence of ballistic peaks. The Lévy-like central part of the distribution, however, fulfills the rescaling property (27), but with a value for the index δ different from Eq. (28): It takes the value

$$\delta=\frac{1}{\beta+1}=\frac{1}{\alpha}, \tag{29}$$

where α is the Lévy index of Eq. (20). We point out that the difference between the two prescriptions (28) and (29) is numerically very small and that in Ref. [10] the difference was visible due to the adoption of numerical calculations with very accurate statistics. The curves of Fig. 7 are compared with the predictions of the rescaling (28). In the present case the accuracy of the statistical analysis is not so as to make it possible to distinguish Eq. (28) from Eq. (29). However, the Gaussian nature of the distributions seems to be so pronounced as to provide further evidence that the folding process of this section is a satisfactory realization of FBM.

In conclusion, the numerical results of this section prove that the irregular sequence produced according to the prescriptions illustrated in Fig. 4 shares the seemingly conflicting properties of Gaussian statistics and long-range correlations. In other words, this is a satisfactory dynamic realization of FBM. The price to pay to realize this physical condition, observed by Arneodo *et al.* in their statistical analysis of DNA sequences [14], is the breakdown of the stationarity assumption. From an intuitive point of view this conclusion can be drawn by comparing the regular trajectory of Fig. 3 to the irregular trajectory of Fig. 4. The stationarity property of the former is a natural reflection of its translational invariance, which in turn is generated by the almost central symmetry of its structure. All these properties are lost by the irregular trajectory of Fig. 4 and with them probably the possibility of expressing the statistical properties of the sequence by means of a single ''time'' correlation function $\Phi_\xi(t)$ is also lost.

Before concluding this section we would like to mention the possibility of applying the modeling of this section to the problem of transport in condensed matter. We have already mentioned that the Cayley tree was used by Abou-Chacra, Anderson, and Thouless [18] to discuss the problem of Anderson localization in the case of a multidimensional lattice. On the other hand, more recently an ever increasing number of researchers [25] have been studying the role that a correlated random distribution of site energies might have on the phenomenon of transport and localization [25]. Allegrini *et al.* [26] have studied the effect of creating these correlations by means of deterministic maps, the GHT, and a variation of it [26]. The interesting result of these calculations was

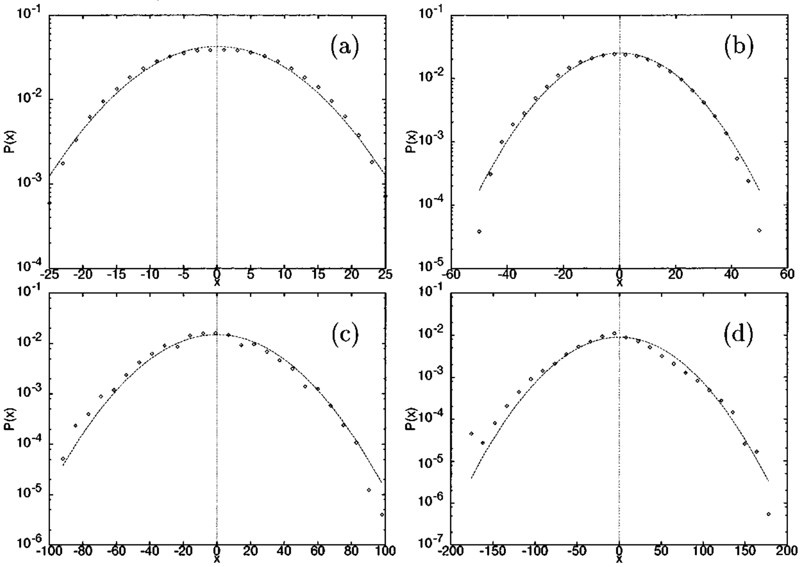

FIG. 7. Rescaling behavior for the process generated by the irregular trajectory of Fig. 4. (a) The numerical data at $t=25$ are fitted by a Gaussian function. (b) The fitting Gaussian function of (a) is rescaled according to the FBM prescription (28) and compared with the numerical data at $t=50$. (c) Same as (b) but with $t=100$. (d) Same as (b) but with $t=200$.

that the localization length might be significantly increased by the correlations among different sites. It would be interesting to adopt the model of this section as a generator of correlation and to assess the resulting transport and localization properties.

IV. APPLICATION: MODELING DNA SEQUENCES

There are several reasons to believe that the model illustrated in the preceding section may have some resemblance to a DNA molecule when it is folded in the nucleus. We have seen in Sec. III that if correlations are imposed to the geometrical structure of the Cayley tree by means of a central symmetry prescription, a self-avoiding trajectory wrapped around the hierarchical graph shows statistical features (the FBM) that are actually detected in the real DNA sequences [14]. Notice that a dynamic but stationary dichotomous model would not be compatible with the FBM condition.

In other words, we imagine the DNA molecule in the nucleus as a long knot-free polymer; it is therefore topologically equivalent to a two-dimensional self-avoiding graph, with a hierarchical folding. As pointed out by Grosberg *et al.* [15], a hierarchical folding is necessary for the DNA to be accessible to RNA and to several enzymes [15,27]. Lewin [27] also states that the highly dense packing of the DNA molecule inside the nucleus necessarily implies a hierarchical organization of the spatial structure of the macromolecule in order to function properly.

Unfortunately, only the first levels of the hierarchical folding are known [28–34]. It is known that the first level is given by the nucleosomes, the second by a helix structure of nucleosomes, and the third by the folding of this structure. We know, however, that for the chromosomes (chromatine in metaphase) one of the higher levels is constituted by a proteic scaffold (for the chromatine in interphase there is an analogous structure called a matrix) around which the chromosomes are wrapped. The folding of the DNA molecule implies constraints on the molecule flexibility properties (bending, bendability, and curvature) [33] and the capability of being anchored to the proteic matrix or to the scaffold [29,34]. These constraints obviously depend on the local nucleotides composition and therefore can be seen as statistical features of the DNA primary structure (unfolded sequence).

If the role of the matrix (for the chromatine in the interphase) [29] and of the scaffolds (for the metaphase) [34] is that of keeping a fixed three-dimensional structure, we can argue that this global constraint is very important and that the DNA sequence must obey it. The correlations necessary to keep the tertiary structure stable are therefore constraints on chemical interactions between different segments of DNA or between the DNA macromolecule and a proteic hierarchical structure. In our model, therefore, the correlations are not imposed on the sequence, but on the spiral of Fig. 3, in a way that is somewhat ‘‘perpendicular’’ to the nucleotide chain.

Our schematical model of a DNA molecule as a polymer wrapped on a Cayley tree may be imagined twisted and folded again in a complicated manner, in order to be densely packed, but saving a certain central symmetry of the complex globule. In this way we see that there are several analogies between our model and the ‘‘crumpled globule structure’’ or Ref. [15], that is, a model for DNA in eukariotes. Also in this case the interactions between the sites responsible for the stability of the structure act on hierarchical surfaces in a way that is perpendicular to the chain structure. The model of Grosberg *et al.* teaches us that the long-range correlation function is actually fundamental for the molecule to be stable; a parameter $H=2/3$ is predicted by the model in a fair agreement with the data, but nothing is actually said about the stationary properties of the corresponding sequence. Since the stationary assumption is actually a form of translational symmetry, we think that such a symmetry is not likely to be fulfilled by a hierarchically folded molecule, but we argue that this symmetry is valid along a trajectory with an approximate rotational invariance since the globule has a central symmetry. This assumption is somewhat arbitrary, but it is consistent with the constraints of Grosberg *et al.* since in our model we have seen that the ‘‘re-ordering’’ procedure does not change the inverse-power-law index of the correlation function.

V. CONCLUDING REMARKS

In this paper we have provided a practical realization of FBM using a geometrical argument. This argument has biological significance and refers to a ‘‘geometrical’’ perspective of the copying mistake mechanism introduced in earlier papers [5] on the statistical analysis of DNA sequences.

We emphasize that we do not have a rigorous mathematical basis for our conclusions. We are inclined to believe, however, on the basis of our numerical results, that the FBM condition can be realized in practice by any physical process that decouples statistics from dynamics. To explain this property we refer to the dynamical realization of Lévy statistics using a GHT map [22]. We can generate a large number of trajectories, each corresponding to a different initial condition. A possible distribution of initial conditions for these trajectories is given by the variable ‘‘velocity’’ ξ at equilibrium and the variable position x distributed according to a Gaussian distribution. This means that the initial statistics of x will be Gaussian and the ensemble will remain Gaussian for an extended period of time whose length can be predicted. When the spreading mechanism creates a new distribution, so large as to perceive the original Gaussian distribution as a Dirac δ function, the statistics will be dictated again by the dynamics according to the arguments detailed in Ref. [10]. In the case studied herein, the Gaussian nature of the initial distribution is generated by the unfolding process. This is, in this case, the ‘‘statistics’’ of the process foreign to the long-range correlations. The long-range correlations are a manifestation of a ‘‘dynamics,’’ that generates its own statistics, which, in this case, should be Lévy statistics. The unfolding process spans the whole diffusion process. Thus it sets persistent Gaussian constraints rather than only affecting the initial distribution. Remarkably, this Gaussian constraint does not affect the dependence of the ‘‘spatial’’ second moment on ‘‘time’’ and leaves unchanged the anomalous character of the diffusion process, namely, the coefficient H established by the GHT map. For the resulting diffusion process to be a genuine form of FBM, the unfolding process should also change the rescaling (29) into that of Eq. (28), a change too small to observe in experimental data.

The results of this work also shed light on why several techniques recently used to study the statistics of DNA sequences, such as the detrended analysis [35] and the Hurst analysis [24], can be adopted as proper indicators of the rescaling, in spite of the fact that there might be long-range correlations in the sequence. We know from earlier work [10] that the dichotomous condition in the presence of long-range correlations should lead to a rescaling different from that of the second moment of the distribution. According to the results of an earlier investigation [5], the detrended analysis [2,3], even if different in nature from the Hurst analysis, as a method to detect the rescaling properties of long-range correlation processes leads to results very close, if not identical, to those of the Hurst analysis. This in turn, as discussed in Ref. [24], is meaningful as a proper indicator of rescaling only if the process considered is Gaussian. Only in this case can the so-called Hurst coefficient H be identified with a parameter defining the ‘‘speed’’ of the diffusion process, as indicated by the time dependence of the second moment of the distribution. This paper shows that in practice the Gaussian statistics can be generated by the initial conditions, thereby producing the false impression that the FBM is a valid picture for the description of DNA sequences. If it is, it is for reasons that are not clearly understood in the current literature. For the same reasons the Hurst coefficient can be adopted as a proper indicator of the rescaling properties of the observed dynamical process.

As far as the specific problem of the DNA sequences is concerned, this paper has to be thought as speculative as that by Li [36]. This author proposed an expansion-modification model, where two processes compete with each other, one creating long-range correlations and the other destroying it: No real DNA sequence is then analyzed with this model. We limit ourselves to remarking that on the basis of the analysis of this paper, not even the Li model, being dichotomous and long-range correlated, can be both stationary and rigorously Gaussian. Here the problem under discussion is inspired by the conflicting results of a statistical analysis of the data. The problem is given a qualitative solution, based on numerical arguments and resting on the essential conclusion that whatever the biological origin of the process observed might be, this process cannot be stationary.

ACKNOWLEDGMENTS

We warmly thank Dr. Marcello Buiatti for his suggestions, which had the beneficial effect of making much more robust the biological significance of our folding model.

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

[2] C. K. Peng, S. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[3] H. E. Stanley, S. V. Buldyrev, A. L. Goldberg, Z. D. Goldberg, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C. K. Peng, and M. Simons, Physica A **205**, 214 (1994).

[4] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992); Fractals **2**, 1 (1994).

[5] P. Allegrini, M. Barbi, P. Grigolini, and B. J. West, Phys. Rev. E **52**, 5281 (1995); P. Allegrini, P. Grigolini, and B. J. West, Phys. Lett. A **211**, 217 (1996).

[6] M. Ding and W. Yang, Phys. Rev. E **52**, 207 (1995).

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

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

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

[10] P. Allegrini, P. Grigolini, and B. J. West, Phys. Rev. E **54**, 4760 (1996).

[11] R. Mannella, P. Grigolini, and B. J. West, Fractals **2**, 81 (1994).

[12] T. Geisel, J. Heldstab, and H. Thomas, Z. Phys. B **55**, 165 (1984).

[13] M. F. Shlesinger, Annu. Rev. Phys. Chem. **39**, 269 (1988)

[14] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[15] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Europhys. Lett. **23**, 373 (1993).

[16] H. Bethe, Proc. R. Soc. London, Ser. A **216**, 45 (1935).

[17] D. Stauffer and A. Aharony, *Percolation Theory* (Taylor and Francis, London, 1992).

[18] R. Abou-Chacra, P. W. Anderson, and D. J. Thouless, J. Phys. C **6**, 1734 (1973).

[19] S. M. Risser, D. N. Baratan, and J. N. Onuchic, J. Phys. Chem. **97**, 4523 (1993).

[20] S. M. Risser and D. N. Baratan, J. Am. Chem. Soc. **115**, 2508 (1993).

[21] G. Zumofen and J. Klafter, Phys. Rev. E **47**, 851 (1993); J. Klafter and G. Zumofen, Physica A **196**, 102 (1993).

[22] G. Trefán, E. Floriani, B. J. West, and P. Grigolini, Phys. Rev. E **50**, 2564 (1994).

[23] M. F. Shlesinger and B. J. West, Phys. Rev. Lett. **67**, 2106 (1991).

[24] H. E. Hurst, Trans. Am. Soc. Civ. Eng. **116**, 770 (1951); J. Feder, *Fractals* (Plenum, New York, 1988).

[25] D. Dunlap, H.-L. Wu, and P. Philips, Phys. Rev. Lett. **65**, 88 (1990); J. Heinrichs, Phys. Rev. B **51**, 5699 (1995); F. M. Izrailev, T. Kottos, and G. P. Tsironis, *ibid.* **52**, 3274 (1995); M. T. Béal-Monod and G. Forgacs, *ibid.* **37**, 6646 (1988); A. Crisanti, G. Paladin, and A. Vulpiani, Phys. Rev. A **39**, 6491 (1989); P. E. de Brito, C. A. A. da Silva, and H. N. Nazareno, Phys. Rev. B **51**, 6096 (1995); G. Y. Oh, C. S. Ryu, and M. H. Lee, *ibid.* **45**, 6400 (1992); C. S. Ryu, G. Y. Oh, and M. H. Lee, *ibid.* **46**, 5162 (1992).

[26] P. Allegrini, L. Bonci, P. Grigolini, and B. J. West, Phys. Rev. B **54**, 11 899 (1996).

[27] B. Lewin, *Genes VI* (Oxford University Press, New York, 1997).

[28] I. Ioshikhes, A. Bolshoy, K. Dereshteyn, M. Bovodovsky, and E. N. Trifonov, J. Mol. Biol. **262**, 129 (1996).

[29] G. B. Singh, J. A. Kramer, and S. A. Krawetz, Nucl. Acids Res. **25**, 1419 (1997).

[30] P. Baldi, S. Brunak, Yves Chauvin, and A. Krogh, J. Mol. Biol. **263**, 503 (1996).

[31] I. Brukner, R. Sanchez, D. Suck, and S. Pongor, EMBO J. **14**, 1812 (1995).

[32] D. S. Goodsell and R. E. Dickerson, Nucl. Acids. Res. **22**, 5497 (1994).

[33] A. Gabrielian and S. Pongor, FEBS Lett. **393**, 65 (1996).

[34] J. Widom, J. Mol. Biol. **259**, 579 (1996).

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

[36] W. Li, Phys. Rev. A **43**, 5240 (1991).