PHYSICAL REVIEW E VOLUME 52, NUMBER 5 NOVEMBER 1995

Dynamical model for DNA sequences

P. Allegrini,[1] M. Barbi,[2] P. Grigolini,[1,2,3] and B. J. West [1]
[1] *Center for Nonlinear Science, University of North Texas, P.O. Box 5368, Denton, Texas 76203*
[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 28, 56127 Pisa, Italy*

(Received 4 May 1995)

We address the problem of DNA sequences, developing a "dynamical" method based on the assumption that the statistical properties of DNA paths are determined by the joint action of two processes, one deterministic with long-range correlations and the other random and δ-function correlated. The generator of the deterministic evolution is a nonlinear map belonging to a class of maps recently tailored to mimic the processes of weak chaos responsible for the birth of anomalous diffusion. It is assumed that the deterministic process corresponds to unknown biological rules that determine the DNA path, whereas the noise mimics the influence of an infinite-dimensional environment on the biological process under study. We prove that the resulting diffusion process, if the effect of the random process is neglected, is an α-stable Lévy process with $1 < \alpha < 2$. We also show that, if the diffusion process is determined by the joint action of the deterministic and the random process, the correlation effects of the "deterministic dynamics" are canceled on the short-range scale, but show up in the long-range one. We denote our prescription to generate statistical sequences as the copying mistake map (CMM). We carry out our analysis of several DNA sequences and their CMM realizations with a variety of techniques and we especially focus on a method of regression to equilibrium, which we call the Onsager analysis. With these techniques we establish the statistical equivalence of the real DNA sequences with their CMM realizations. We show that long-range correlations are present in exons as well as in introns, but are difficult to detect, since the exon "dynamics" is shown to be determined by the entanglement of three distinct and independent CMM's.

PACS number(s): 87.10.+e, 05.40.+j, 05.70.Ln

I. INTRODUCTION

In the past decade or so there has been a ground swell of interest in unraveling the mysteries of DNA. One approach that has, in just a few years, proven to be particularly fruitful in this regard is the statistical analysis of DNA sequences [1–9] using modern statistical measures. One focus of this analysis has been on the distribution of the four bases adenine, cytosine, guanine, and thymine (A, C, G, and T) in order to shed light on the following fundamental problems: (i) establishing the role of the noncoding regions in DNA sequences (introns) in the hierarchy of biological functions [2,5], (ii) finding simple methods of statistical analysis of such sequences to distinguish the noncoding from the coding regions (exons) [6], (iii) discovering the constraints and regularities behind DNA evolution and their connections to the Darwin theory of selection and more generally to contemporary evolution theories [7,8], (iv) extracting new global information on DNA and its function [2,3,5], and (v) establishing the roles of chance and determinism in genetic evolution and coding regarded as being the "program" underlying the development and life of every organism [7].

A familiar kind of analysis of DNA sequences is that used by Voss [3] based on the equal-symbol correlation. He uses a binary indicator function $U_k(x_n)$ that is equal to 1 if a letter k occurs at the position x_n and to 0 otherwise. The letter k is defined by the four nucleotides $k = A, C, T, G$. The indicator functions are used to construct the correlation function and its Fourier transform, the spectral density

$$S(f) = \sum_{k=A,C,T,G} S_k(f), \tag{1}$$

from which he removed the white noise floor. The details of this technique are reviewed in Sec. III D.

The analysis of the spectrum $S(f)$ led Voss to the following two major observations regarding the general properties of DNA spectra: (a) the spectra have a peak at $f = 1/3$, $t = 3$ (for coding sequences) and (b) the DNA sequences have long-range correlations as indicated by the slope of the spectrum, when plotted on a log-log graph paper.

We shall discuss the 1/3 peak subsequently. Here we stress that the long-range correlation means that

$$\lim_{f\to 0} S(f) \propto \frac{1}{f^{\nu}}, \tag{2}$$

with $1 \geq \nu > 0$ (the case $\nu = 0$ corresponds to a completely random distribution, with no correlation). The result for $S(f)$ given by (2) is equivalent to the corresponding result for the correlation function [10]

$$\lim_{\tau\to\infty} C(\tau) \propto \frac{1}{\tau^{\beta}}, \quad (3)$$

with $\beta = 1 - \nu$, which is obtained using a Tauberian theorem. This consequently implies the condition $1 > \beta \geq 0$, so that it is not possible to define a length scale for the correlation function, e.g., the correlation is nonzero at all the distances τ separating elements in the sequence. This interpretation is the reason why this inverse-power-law behavior is called long ranged (see also [11]).

Voss finds these inverse-power-law spectra for the sequences studied regardless of the percent of intron content. This is where the results of Voss disagree with those of Stanley *et al.* [5], who, on the contrary, focus their attention on the different degrees of correlation in intronless and intron-containing sequences and find no correlation in cDNA sequences.

Let us now briefly review some of the main results of the research work of Stanley *et al.* [5]. They find long-range correlations in the noncoding regions and no correlation at all in the coding regions of DNA sequences. They use methods of analysis different from that of Voss and are related to the dynamical treatment that we illustrate in Sec. II. They study the landscape variable, which, adopting the notation of this paper, reads

$$x(\ell) = \sum_{i=1}^{\ell} \xi_i. \quad (4)$$

Here i represents the position in a sequence and ℓ the distance along a DNA sequence (ℓ is an integer between 1 and N, the length of the sequence) and ξ_i is a variable that assumes the value $+1$ if a purine occurs and -1 if a pyrimidine occurs at the position i. Thus the cumulative variable $x(\ell)$ with the increase of "time" ℓ executes a trajectory similar to that of diffusional one-dimensional motion, called, by Stanley and co-workers, a DNA walk. This trajectory has a fractal structure (like a mountain) and is therefore called a "landscape." Stanley *et al.* [5] focus their attention on the second-order properties of the landscape, such as the mean square deviation from the mean

$$F^2(\ell) \equiv \langle(\Delta x - \langle\Delta x\rangle_{\ell_0})^2\rangle_{\ell_0}, \quad (5)$$

where ℓ_0 is the initial point of the walk, the ℓ_0 subscript on the bracket means an average over initial positions, and

$$\Delta x \equiv x(\ell_0 + \ell) - x(\ell_0). \quad (6)$$

In Sec. II we introduce the statistical arguments appropriate for correlation fluctuations and show that with a correlation function of the form (3) with $1 > \beta > 0$ the asymptotic form of the second moment (5) becomes

$$\lim_{\ell\to\infty} F^2(\ell) \propto \ell^{2H}, \quad (7)$$

where $H = 1 - \beta/2$, so that $1 \geq H > 0.5$ (again, the case of complete randomness corresponds to the extreme value $H = 0.5$). In the case of introns Stanley *et al.* [5] actually find $H > 1/2$, in agreement with the results of Voss. In the case of the intronless sequences (where the introns were removed), on the contrary, Stanley *et al.* [5] find $H = 0.5$ in the case of sufficiently short ℓ. In the analysis of some coding sequences Stanley *et al.* [5] noticed that some coding DNA landscapes were a juxtaposition of patches of different biases whose lengths were distributed around a typical length scale; they noticed further that the dispersion they measured within such subsequences was normal. To avoid the subjectiveness of selecting such subsequences, they developed the detrended fluctuation analysis: they generalized the function $F^2(\ell)$ and adopted the function $F_d^2(\ell)$, allowing them to distinguish the cases where the inverse-power-law behavior exists at all length scales from those where the correlation only appears on a typical length L. This length scale is identified with the typical length of the random subsequences that they find in the studied intronless sequences. Thus Stanley *et al.* [5] attribute the presence of correlations at large ℓ in the exons to a crossover effect among the subsequences, thereby implying that no substantial long-range correlation exists in the exons.

Stanley *et al.* [5] also proposed some models of evolution and reported the results of analyses of sequence coding for the same protein belonging to organisms in different positions in the evolutionary tree. They find an interesting increase of the coefficient H with biological complexity, i.e., H is a function of the position in the tree.

The differences in the findings of the two groups—long-range correlations being ubiquitous in DNA sequences by Voss [3] and such correlations being absent in exons by Stanley *et al.* [5]—has motivated us to develop a phenomenological dynamical model that might not only mitigate these differences, but also suggests the dynamical origins of the observed statistical properties. The proposed model is an application of nonlinear mappings to the understanding of the statistics of DNA sequences. We also believe that this affords a completely different strategy for determining the biological mechanisms underlying DNA structure and thereby indirectly biological functions.

Within this context we must mention the work of Grosberg *et al.* [2], who suggest that an intrinsic constraint might lead DNA evolution towards a given statistical conformation. In fact, Grosberg *et al.* studied the statistical properties of a polymer confined within a minimum volume but constrained to remain essentially knot-free. Under these conditions the sequence must result in long-range correlations with $H = 2/3$. This kind of packing (crumpled globule structure) shows up in the complete sequence of the DNA of the eukariots, i.e., the living beings whose DNA is contained in a nucleus and are characterized by the presence of introns, that is, DNA sequences that do not code for proteins. The nuclear DNA must keep the capability of unfolding itself for the purpose of transcription and duplication. The complete sequence consists mainly of noncoding DNA. For all these reasons Grosberg *et al.* argue that the role of introns might be that of producing the needed long-range correlation so as to rigorously maintain the convenient spatial configuration for the whole genome and, consequently, the correct function. According to this interpretation of the work of Grosberg *et al.* the lack of correlation in the coding sequences would be justified by the fact that in addition to the exons (which are responsible for the other fundamental function, the code) a further "structure" responsible for the function should exist.

It must be added that Lió *et al.* [8] also stress that the statistical properties of the DNA sequences imply either a series of internal causes or relations with the cellular environment. These are constraints concerning the proper function of the DNA code and the complex mechanisms needed for the cell life. These authors apply the mutual information function to the pairs of bases AT and CG, distinguishing between weak and strong bonds, and find a period-3 correlation for the pair CG (strong) in organism living in limiting life conditions. They also mention a sort of internal natural selection that should account for these properties. This is additional evidence that the statistical properties of the DNA sequences may be related to internal and external constraints.

We plan to approach the discussion of all these issues by adopting a dynamical model from the point of view that the different positions of the sequence can be regarded as distinct values of a discrete time and the landscape variable (4) can be regarded as the collection of all the fluctuations that the statistical variable ξ, the "velocity" of our "Brownian particle," undergoes throughout the observed time interval. Our modeling is based on the assumption that this diffusion process rests on the joint action of two distinct statistical sources, the former being a statistical process with *long-range correlations* and the latter being a *noise*, namely, a random process with *no correlations*. The generator of the process with long-range correlations is assumed to be a *deterministic* nonlinear map, mimicking a state of weak chaos, and is thought of as expressing the rules determining the dynamics of the biological process under study. This biological process interacts with an infinite-dimensional environment and, according to traditional wisdom, this interaction is mimicked by a δ-function correlated random process. The weight of these two distinct statistical sources is determined by a fitting procedure of the experimental data. This results in a special map, which we term the copying mistake map (CMM) and is our proposed model to interpret the DNA sequences.

Thus we see that our model balances the two major sources of randomness in statistical mechanics: noise, the traditional process introduced to model the infinite number of degrees of freedom of a complex mechanical system, and chaos, the paradigm of deterministic randomness from nonlinear dynamics. This choice of including both noise and chaos is dictated by a criterion of efficiency as well as by a "philososophical" perspective on DNA sequences, in which the sequence is perceived as the result of a compromise between chance and necessity. In fact, as will become transparent from the content itself of this paper, the DNA sequences are a biological case of anomalous diffusion and anomalous diffusion is determined by waiting time distributions in each of the states 1 or -1 of the velocity ξ with an inverse power law. The deterministic map used in this paper is one of several possible generators of inverse-power-law distributions, another well know one being, for instance, a hierarchical model [12]. The choice of the deterministic map is also dictated by a criterion of efficiency, which makes the CMM an especially simple way of generating sequences statistically indistinguishable from the real DNA sequences. Furthermore, we shall see that the adoption of the deterministic map makes it possible to realize a variety of different conditions, including the oscillations detected by Voss [3] and Lió *et al.* [8] using a single approach.

The outline of the paper is as follows. In Sec. II we review the dynamical basis of anomalous diffusion, with emphasis on nonlinear deterministic maps. Here the statistics of anomalous diffusive processes are shown to be Lévy stable and the waiting time distribution functions to be inverse power laws. These considerations lead to the formation of a map, which results in the CMM when the enviromental perturbations are properly taken into account. The CMM generates sequences by a chaotic map with long-range correlation and uncorrelated random mutations that destroy short-range correlations. In Sec. III we review the traditional methods of analysis of DNA sequences: diffusion analysis, detrended analysis, Hurst analysis, and spectral analysis. In addition we introduce into this context a procedure inspired by Onsager [13] in which the regression of a perturbed system back to equilibrium is used to determine the equilibrium correlation function. In Sec. IV we apply the standard methods of analysis to real DNA sequences to compare and contrast the result with those generated by the CMM. We summarize our results and draw some conclusions in Sec. V, which also illustrates the research directions suggested by the results of this paper.

II. DYNAMICAL THEORY OF ANOMALOUS DIFFUSION

The purpose of this section is to present a dynamical approach to the generation of the statistical behavior of DNA sequences and the theoretical motivation behind it. First of all, we show that a dynamical approach to the diffusion of a variable x is due to its velocity ξ fluctuating between two values 1 and -1, naturally resulting in a Lévy process if the fluctuations are stationary, and the waiting time distribution of the velocity ξ is an inverse power law with a finite first moment. Second, we propose a deterministic map, which is probably the most convenient generator of this inverse-power-law distribution of sojourn times to model the fluctuating velocity. Finally, we build up a CMM, namely, a process resulting from the joint use of a deterministic map, responsible for the birth of correlations, and a δ-function-correlated random process. This δ-function-correlated process mimics random pointlike mutations and has the effect of destroying these map-generated correlations on a short-time scale. The biophysical and biochemical sources of the inverse-power-law distributions of waiting times, necessary to produce anomalous diffusion in the form of α-stable Lévy processes, remain unexplained. However, our approach demonstrates that the resulting diffusion is remarkably similar to that generated in Hamiltonian systems, where the source of a waiting time distribution with an inverse power law is known to be the fractal nature of the border between stable islands and the chaotic sea [14–18].

A. General dynamical remarks

A diffusion process in the one-dimensional case stems from the remarkably simple equation

$$\dot{x}(t) = \xi(t), \tag{8}$$

where x is the diffusing variable and ξ is the stochastic process generating diffusion. We assume that the stochastic variable x is independent of ξ. We focus on the special case where this is a statistical process, with only two possible values $\xi = 1$ and $\xi = -1$. We also assume that these two states are equally weighted, thereby resulting in

$$\langle \xi(t) \rangle_{\rm eq} = 0. \tag{9}$$

In the case of DNA sequences, as pointed out in the Introduction, the two values 1 and -1 denote different molecular groups and the time t corresponds to the distance of the molecular site considered from a given origin. Note that for large sequence distances (and therefore large times) one can safely adopt the continuous time representation, which makes it easier to establish a formal connection with typical diffusion processes such as ink in water. It must be remarked that in physical systems the averages are made on a Gibbs ensemble of identical systems and the assumption is made that the velocity ξ is in a state of statistical equilibrium. In principle, one might instead use a single system and replace the Gibbs averages with averages over long times. The connection between the two pictures involves the ergodic assumption. In the case of DNA sequences we have available, so to speak, only single trajectories or realizations. Consequently, the connections with the dynamical approach outlined here is made possible by assuming that the DNA sequence can be dealt with as being a single realization of an ergodic process. We also make the assumption that the time averages on the variable x would correspond to a stationary, or equilibrium, condition. The integration of (8) allows us to construct the second moment

$$\langle x^2(t) \rangle = \langle x^2(0) \rangle + 2\langle \xi^2 \rangle_{\rm eq} \int_0^t dt' \int_0^{t'} dt'' \Phi_\xi(t''), \tag{10}$$

where $\Phi_\xi(t)$ denotes the equilibrium correlation function defined by

$$\Phi_\xi(t) = \frac{\langle \xi(0)\xi(t) \rangle}{\langle \xi^2 \rangle}. \tag{11}$$

It has to be stressed that (10) implies that $\langle \xi(t')\xi(t'') \rangle$ depends on the time difference $|t' - t''|$, as suggested by the assumption that the process is stationary, thereby making $\langle x^2(t) \rangle$ depend on the one-time correlation function (11). The investigation carried out herein also rests on this assumption.

Normal diffusion is a natural consequence of the existence of the microscopic time scale, defined by

$$\tau = \int_0^\infty \Phi_\xi(t) dt. \tag{12}$$

If the correlation function $\Phi_\xi(t)$ decays quickly enough to make τ finite, we can explore the process for times t very large compared to τ, thereby making the long-time limit of (10) indistinguishable from

$$\langle x^2(t) \rangle = \langle x^2(0) \rangle + 2Dt, \tag{13}$$

where the diffusion coefficient for the process is defined by

$$D \equiv \langle \xi^2 \rangle_{\rm eq} \tau. \tag{14}$$

The time scale separation between the diffusion process and the velocity fluctuations allows the central limit theorem to work, thereby realizing a Gaussian diffusion process.

What about the case where the definition of the microscopic time scale is impossible ($\tau \to \infty$)? A natural way of realizing this unusual condition would be given by an autocorrelation function $\Phi_\xi(t)$ with the asymptotic property

$$\lim_{t\to\infty} \Phi_\xi(t) \propto \frac{1}{t^\beta}, \tag{15}$$

with

$$0 < \beta < 1. \tag{16}$$

It is evident indeed that in this unusual condition the correlation time τ (12) diverges and the time scale separation between the macroscopic (diffusion) and the microscopic process (fluctuations of the velocity variable ξ) would not be possible. To shed light on this question one might use the connection, established by Geisel *et al.* [19], between the stationary correlation function $\Phi_\xi(t)$ and another important statistical function, the waiting time distribution $\psi(t)$. This function determines the probability that $\xi(t)$ has made a transition between states in a time t. In the specific case where the variable ξ is a dichotomous process, as in the case of the DNA sequences we are exploring in this paper, this connection between $\Phi_\xi(t)$ and $\psi(t)$ is exact and reads

$$\Phi_\xi(t) = \frac{\int_t^\infty (T - t)\psi(T)dT}{\int_0^\infty T\psi(T)dT}. \tag{17}$$

From this exact relation we see that the condition (16) is realized provided

$$\lim_{t\to\infty} \psi(t) \propto \frac{1}{t^\mu}, \tag{18}$$

with

$$2 < \mu < 3. \tag{19}$$

This restriction on the index μ arises from (16) since from (17) it is evident that

$$\beta = \mu - 2. \tag{20}$$

In conclusion, we see that the functional form of $\psi(t)$ (18) with the power μ in the range (19) generates the inverse-power-law behavior of $\Phi_\xi(t)$ and hence the breakdown of the condition of a finite τ (12) for normal diffusion.

It is easy to prove that in the case where (19) applies the asymptotic behavior of the second moment of the diffusing variable x is given by

$$\langle x^2 \rangle \propto t^{2H}, \tag{21}$$

with

$$H = 2 - \frac{\mu}{2}, \tag{22}$$

which therefore ranges from 1/2 to 1. The relation between the indices (22) can be easily obtained by twice differentiating (21) and (11) and equating the resulting expressions. Much more exciting is the fact that the distribution of x is not Gaussian and it is characterized by long-range tails. These tails cannot result in diverging moments, a fact that would be incompatible with the dynamical realization of the process, where the diffusing particle cannot travel with a velocity faster than that of the limiting trajectory $|x| = t$. However, if this unavoidable truncation is ignored, the distribution is indistinguishable from that of a Lévy process [20,21]. Let us denote by $\hat{P}(k,t)$ the Fourier transform of the distribution $P(x,t)$ and let us focus our attention on it. It is shown [20,21] that if the dynamical truncation of the distribution is neglected, the diffusion generated by the fluctuating variable ξ with the waiting time distribution $\psi(t)$ fulfilling (18) and (19) results in the characteristic function for a symmetric Lévy stable process

$$\hat{P}(k,t) = \exp\left(-b|k|^{\alpha} t\right), \tag{23}$$

with the Lévy index

$$\alpha = \mu - 1. \tag{24}$$

This means that we are observing an α-stable Lévy process with an index in the interval $1 < \alpha < 2$. Notice that in principle the α-stable Lévy processes concerns the wider range $0 < \alpha < 2$. However, the condition $\alpha < 1$ refers to processes faster than ballistic diffusion and so is incompatible with the dynamical nature of the process described by (8), with the further assumption, obvious in the case of the DNA sequences, that the fluctuating variable ξ is independent of x. The condition $\mu \leq 2$ does not lead to a Lévy process, but it is proved [20,21] to result, throughout the whole range $1 < \mu < 2$, in a process with $H = 1$. In fact, in this case the integration of (17) leads to a time-independent Φ_ξ to which (20) does not apply.

Are there physical and biophysical systems that fulfill the conditions leading to (21), with (22) and (19)? It has been noticed recently by several investigators [14,15] that the momentum of a kicked rotator in the so-called accelerator state is driven by a variable, playing the role of the velocity ξ, which fluctuates between two distinct values. The corresponding waiting time distribution has the same structure as (18) with the condition, (19). It is curious that Hamiltonian processes realize the condition (19), which corresponds precisely to the overlap between the Lévy processes and their dynamical realization. In the case of Hamiltonian systems, it is well understood how the inverse power law (18) is generated, this being due to the fractal nature of the border between chaotic and ordered regions in the phase space for the dynamical system [14–18]. However, so far there are no theories establishing the range of the index μ. The fact that μ is located in the range (19) is the result of a numerical "observation" not yet substantiated by a theory. It must be pointed out that this observation, supplemented by the theory of Refs. [20,21], is equivalent to proving that Hamiltonian systems in a stationary condition realize anomalous diffusion in the specific form of Lévy processes, where these processes are compatible with a dynamical realization ($2 < \mu < 3$).

What about DNA sequences? The DNA sequences, when the association between letters and numbers that we use is adopted, are a biological realization of the equation of motion (8), with the velocity ξ fluctuating between the two values 1 and -1. According to the results illustrated here, the simplest possible way for these sequences to produce anomalous diffusion is by realizing a waiting time distribution with the structure of (18) and with $\mu < 3$. The case $\mu \geq 3$ would result in trivial Brownian motion. We note that the prescription of Grosberg *et al.* resulting in $H = 2/3$ corresponds precisely to the dynamical condition (19), with $\mu \sim 2.7$, and is compatible with a dynamical derivation based on the stationary correlation function $\Phi_\xi(t)$. Is the condition $2 < \mu < 3$ fulfilled by the rule of Grosberg *et al.*, a general property of DNA sequences? Actually, there are no compelling reasons why, if the anomalous character of diffusion is accepted, DNA sequences should also fulfill the condition $\mu > 2$, in addition to $\mu < 3$. However, it is attractive to imagine that DNA sequences share the same "dynamical" property as the Hamiltonian systems and realize anomalous diffusion only under the form of an α-stable Lévy process with $2 < \mu < 3$. Note indeed that $\mu < 2$ would lead to a ballistic process distinct from a Lévy diffusion [21]. Thus the analysis of this paper on the DNA sequences is made having in mind the condition (19), namely, the same range as that of Hamiltonian systems in the so-called weak chaos state [14–16].

B. Deterministic approach to an inverse power law for the waiting time distribution $\psi(t)$

Here we illustrate the deterministic map we adopt to generate the waiting time distribution (19). This map is very similar to that originally derived by Geisel *et al.* [19] and more recently studied by Zumofen and Klafter [21] with the help of generalized versions of the continuous time random walk and generates a waiting time distribution $\psi(t)$ with an inverse power law and an index μ ranging from 1 to ∞ and thus in principle also including the nonstationary region $1 < \mu < 2$.

The explicit form of the map is

$$y_{i+1} = f(y_i), \tag{25}$$

where i is the iteration number. The functional form of the map is

$$f(y) = \begin{cases} y + ay^z & \text{for } 0 \le y \le d \\ \frac{1-d-y}{1-2d} & \text{for } d < y < 1-d \\ y - a(1-y)^z & \text{for } 1-d \le y \le 1, \end{cases} \tag{26}$$

where $a = (1-2d)d^{1-z}$ (this choice is made to fulfill ergodicity and to avoid oscillating trajectories). The map is shown in Fig. 1 to be piecewise continuous. The variable that takes the values $+1$ or -1 is

$$\zeta_i = [2y_i] - 1, \tag{27}$$

where the square brackets denote the integer part of the quantity inclosed. Note that in the case $z = 1$ and $z = 2$ this map becomes identical to that recently developed by Leibovitch and Tóth [22] to study the dynamics of ionic channels.

A unidimensional map is a dynamical description of a variable y_i, where the subscript i indicates a discrete time, according to the recursive rule (25). A class of maps fulfilling the property of having motion inside a certain region, with an inverse power law for the waiting time distribution $\psi(t)$, are those containing hyperbolic fixed points where the mapping function $z = f(y)$ is tangential to the straight line $z = y$. In our mapping we have two such points, one at $y = 0$ and the other at $y = 1$ (see Fig. 1). In this case the particle undergoes a very slow motion out of the region near the fixed point (weak repeller) and then eventually exits the laminar region giving rise to the so-called intermittent behavior. The typical form of this kind of map near the fixed point is

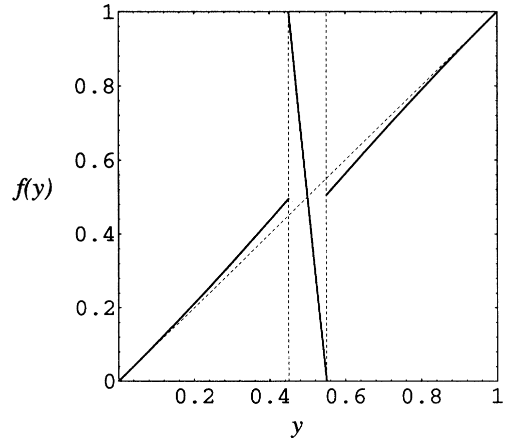

FIG. 1. Solid lines represent the nonlinear map of (26) with $z = 5/3$ and $d = 0.45$. We can see that the 45° diagonal is tangential to the curve at 0 and 1. The side regions (called "laminar regions") are separated by a switching region. The three regions are clearly distinguishable as we plotted vertical lines between them.

$$f(y) = y + ay^z, \tag{28}$$

where $z > 1$. Since the dynamics are slow, that is, the particle undergoes many recursions near the fixed point, we can express (25), together with (28) in the continuous time limit, in the form

$$\dot{y} = ay^z, \tag{29}$$

which can be solved by quadrature. In particular, if we imagine our particle to start from a position y_0, within one laminar region (let us say the left one), and to exit in an escape time t_{esc}, when the particle reaches the position $y = d$, we can write

$$\int_{y_0}^{d} \frac{dy}{y^z} = a \int_0^{t_{\text{esc}}} dt. \tag{30}$$

It is easy to cast the solution of this equation in the form $y_0(t_{\text{esc}})$; if $\rho(y_0)$ is the distribution of initial condition (we can take it to be flat), we can connect the two distributions through the Jacobian, i.e.,

$$\psi(t_{\text{esc}}) = \rho(y_0) \left| \frac{dy_0}{dt_{\text{esc}}} \right|, \tag{31}$$

so that it is easy to prove (18) where

$$\mu = \frac{z}{z-1}. \tag{32}$$

C. DNA sequences as CMM's

The map (26) does not describe all the dominant properties of DNA sequences. Although the map (26) gives rise to the dynamical long-range correlations observed in introns, it does not have the uncorrelated diffusive properties observed in exons. To achieve this latter property we introduce [7] a kind of *a posteriori* noise, so as not to interfere with the long-range correlations. This noise is intended to mimic pointlike mutations. In real DNA sequences indeed no large patches of consecutive sites with the same sign, like the one predicted by the map, are observed. The noise we introduce is called copying mistake noise and the resulting map is given by the variable ξ_i,

$$\xi_i = \begin{cases} \zeta_i & \text{with probability } \epsilon \\ \text{random } \{-1,+1\} & \text{with probability } (1-\epsilon), \end{cases} \tag{33}$$

where $0 \le \epsilon \le 1$ and (33) is called the copying mistake map.

The CMM is based on the independent action of the deterministic one-dimensional nonlinear dynamical process (27) and a stochastic uncorrelated one. The independence of the two processes from one another results in a second moment for the landscape variable given by

$$\langle x^2(\ell) \rangle = A\ell^{2H} + B\ell, \tag{34}$$

where $H = 2 - z/2(z-1)$. Exploiting the independence of the two processes, it is easy to prove that the ratio A/B is proportional to $\epsilon/(1-\epsilon)$.

In Sec. IV we show that the CMM model describes intronless and intron-containing sequences in a unified way, the differences arising from different values of the copying mistake probability $(1-\epsilon)$. However, (34) makes it clear that it is very difficult to detect long-range correlations within a short distance scale if such probability is large, namely, if $B \gg A$. This is indeed what happens in the case of coding sequences, i.e., exons.

The structure of (34) might suggest that the role of chance in exons is much stronger than in introns, when one at first would expect the opposite to be true. Stanley *et al.* [5] made the hypothesis that introns (and introns only) carry long-range information. The work of Grosberg *et al.* [2] shows that this information could be simply related to global properties of DNA, such as the tertiary structure. We show how this correlation information shows up (with some difficulty) in exons as well as in introns and in the Conclusion attempt to give some possible biological explanations. Here we mention that in part the copying mistake disorder in exons is indeed a consequence of the physical and biological constraints on the protein coding, which do not produce long-range effects and may be perceived as an uncorrelated noise at the level of DNA base-base correlation.

III. METHODS OF ANALYSIS

The development of techniques to analyze the statistical properties of DNA sequences has become a very active field of research. A frequently used technique [1,5,8] is the method of information entropy, which is thought to be free from the somewhat arbitrary identification of symbols with real numbers [9]. We are aware of that problem and consequently of the potential importance of adopting the information entropy and other such bias-free methods. However, since our aim here is to apply our dynamical model of DNA sequences, we prefer to focus on those traditional methods of analysis that associate letters with numbers in the assessment of the utility of the CMM. On the other hand, our statistical method leads us to conclusions consistent with those reached by Lió *et al.* [8], on the basis of the entropy information method.

In this section we present a brief illustration of other methods of analysis that can be related to the dynamical approach, which we therefore apply in this paper.

A. Diffusion analysis

Diffusion analysis is the most direct way to detect the diffusion exponent H. The method consists of transforming the symbolic data sequences into a series of +1's and −1's, by substituting nucleotides A and G (purines) with a +1 and C and T (pyrimidines) with a −1. Although this choice, as pointed out by Stanley *et al.* [5], is arbitrary, it has some merit compared to other possible choices. Let us consider, for instance, the choice of AT vs CG. This is biologically correct since it groups the pairs according to the hydrogen bond in the doubled stranded DNA. However, this choice is proved to be affected by biases that have to do with the local tertiary structure (pieces rich in AT are more flexible than pieces rich in CG) and that constitute a kind of external noise in our search of long-range correlations. The numerical treatment of the data is essentially performed according to (5). The evaluation of H comes from the equation for the second moment (10), where the averages are taken over all the possible initial conditions. It is also possible to evaluate the distribution probabilities $P(x,t)$ that describe the probability of having traveled a "displacement" distance x in a "time" t. Such distributions are Gaussians for uncorrelated processes and, for the reasons pointed out in Sec. II, are Lévy functions for the long-range correlated processes supplemented by the stationary assumption. This kind of analysis was applied to the study of the DNA sequences by Li and Kaneko [1]. They studied a three-dimensional pseudorandom walk, of which each of the four bases represented a velocity vector relative to the four angles of a regular tetraedron. Our one-dimensional approach is a simplified version of this method since it can be regarded as being a geometric projection on a straight line connecting the middle points of two opposite sides of this tetraedron.

B. Detrended fluctuation analysis

The detrended fluctuation analysis (DFA) was originally introduced by Stanley and co-workers [5] for the purpose of distinguishing, within a reasonably short time scale, if the dynamical process stemming from a DNA sequence is dominated by correlated or uncorrelated fluctuations. This method was shown to be successful in distinguishing introns from exons in a yeast chromosome [6]. It rests on the rule that short correlations are destroyed for cDNA sequences; a rule that has still to be shown to be valid in general. When applied to CMM, the DFA gives, for short and intermediate time regimes, values of H around 1/2 if the probability of mutation $1-\epsilon$ is large.

The DNA sequence $y(n)$ (+1's and −1's) is divided into N/l sequences of length l, each subsequence being confined in a box. The boxes are labeled by the index s. The total bias in the sth box is

$$M_l^s = \frac{1}{l}\left(\sum_{n=(s-1)l+1}^{sl} y(n)\right). \tag{35}$$

Then the detrended walk is defined

$$Y_l^s(n) = y(n) - nM_l^s \text{ for } (s-1)l+1 \le n \le sl, \tag{36}$$

with the variance in the box given by

$$\sigma_2^{s,l} = \frac{1}{l}\sum_{n=(s-1)l+1}^{sl} [Y_l^s(n)]^2. \tag{37}$$

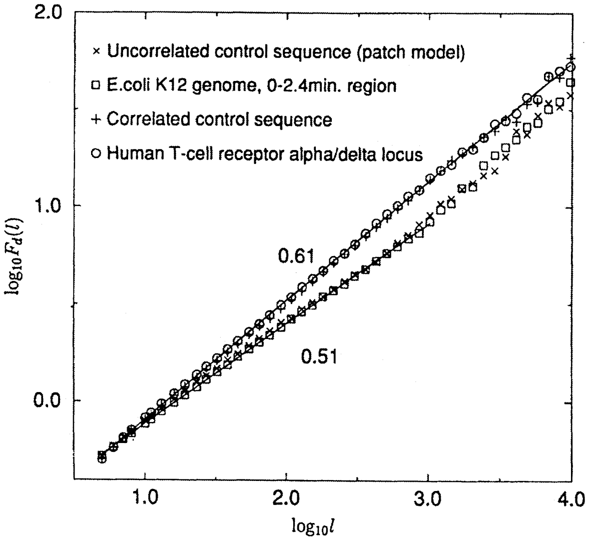

FIG. 2. DFA analysis performed over four sets of data. The squares represent a sequence from the *E. coli* K12 genome. This sequence is composed of exons and is fitted with an uncorrelated control sequence (×'s). The resulting H is ~ 0.5. The circles represent a human intron containing sequence, which is in turn fitted with a correalated control sequence. Both sequences turn out to have $H > 0.5$ (adapted from [32]).

The function $F_d^2(l)$ is defined as the average of $\sigma_2^{s,l}$ over the N/l boxes and is a funcion of l (the width the boxes into which the sequence is divided). If the dynamical process is a fractional Brownian motion (and so a Gaussian distributions is assumed) characterized by a certain H, it is possible [5] to show that

$$\sqrt{F_d^2(l)} \propto l^H. \tag{38}$$

Stanley *et al.* [5] find $H > 0.5$ for intron-containing sequences, while for intronless sequences they find $H \approx 0.5$ under a certain characteristic length and $H > 0.5$ for l over that length. This can be seen in Fig. 2, which shows some results of Stanley *et al.* [5] for intronless sequences using the DFA analysis.

C. Hurst analysis

Hurst analysis [23] is a classic procedure for detecting anomalous diffusion behavior. It is based on the following steps. First, one defines the span of the DNA walk

$$S(\tau) \equiv \max_{1\leq t\leq\tau} \sum_{i=1}^{t} (\xi_i - \langle\xi\rangle_\tau) - \min_{1\leq t\leq\tau} \sum_{i=1}^{t} (\xi_i - \langle\xi\rangle_\tau), \tag{39}$$

where the angular brackets denote the ensemble average up to the time $\tau \geq t$. Second, the variance of the walk is constructed

$$V(\tau) \equiv \left[\frac{1}{\tau}\sum_{i=1}^{\tau} (\xi_i - \langle\xi\rangle_\tau)^2\right]^{\frac{1}{2}}. \tag{40}$$

Finally, we construct the rescaled range variable

$$R(\tau) \equiv \frac{S(\tau)}{V(\tau)} = \left(\frac{\tau}{2}\right)^{H_H}, \tag{41}$$

where H_H is the Hurst exponent usually denoted as H. In the specific case of fractional Brownian motion $H = H_H$, but in general the two exponents are not equal [24].

This analysis was recently adopted to study anomalous diffusion produced by either deterministic maps or stochastic processes [24]. We performed Hurst analysis on DNA sequences and found the same long-time behavior as that stemming from the DFA and the same value of H. This is so because these two kinds of analyses are essentially equivalent. It must be pointed out that this value of H might be different from that provided by the diffusion method (study of the second moment of x) if the process is not Gaussian [24].

D. Spectral analysis

Spectral methods rest on the numerical evaluation of the equilibrium correlation function (11) followed by the application of a fast Fourier transform. The long-time correlations are then related to the low-frequency region of the spectrum due to the complementary relation between the Fourier transform and its inverse. As we mentioned in the Introduction, Voss defined a symbolic correlation function introducing a binary indicator function $U_k(x_n)$ that is equal to 1 if a letter k occurs at the position x_n and to 0 otherwise. The letter k is defined by $k = A, C, T, G$. In this way the four-symbol indicator correlation function can be defined as

$$\begin{aligned} C(\tau) &\equiv \frac{1}{N}\sum_{n=1}^{N} \sum_{k=A,C,G,T} U_k(x_n)U_k(x_{n+\tau}) \\ &\equiv \sum_{k=A,C,G,T} C_k(\tau), \end{aligned} \tag{42}$$

while $C_k(\tau)$ is defined as an equal-symbol correlation function and N is the total length of the considered sequence. As already stated, if the Fourier spectrum is

$$S(f) \sim \frac{1}{f^\nu} \tag{43}$$

with $1 \geq \nu > 0$, then the correlation function is

$$C(\tau) \sim \frac{1}{\tau^\beta} \tag{44}$$

with $\beta = 1 - \nu$, which consequently fulfills the condition $1 > \beta \geq 0$. This method is implemented in a purely symbolic way so as to avoid cross-correlation effects due to projections in spaces with dimension smaller than 4.

However, we can directly apply the Fourier spectrum evaluation to the dichotomous sequence generated by the purine-pyrimidine random walk rule (the evaluation in the one-dimensional numeric subspace, however, seems to recover the same value for the correlation function, when the purine-pyrimidine rule is chosen). The advantage of the spectral method is that it is possible to subtract the white noise background for the spectrum so as to make easier the detection of anomalous behavior, even within relatively short-time scales. One can compare the spectrum with that of a really uncorrelated sequence (e.g., from the decimal figures of π) of the same length. The white noise subtraction is carried out in a subjective way since one cannot determine its level unambiguously and for this reason the spectral method has been criticized by Buldyrev *et al.* [25]. Yet this method might be used to distinguish coding from noncoding sequences since it is able to reveal a peak of frequency 1/3, namely, a harmonic component of period three in the correlation function, suggesting the presence of "codons," the nucleotide triplets responsible for the codification of a single aminoacid (or a "stop" signal for the coding procedure). To be precise, we have to mention that pseudogenes, a kind of noncoding sequences recently (in the evolution time scale) evolved from exons, also reveal period-3 oscillations and other statistical properties that are typical of cDNA; therefore one has to be careful when using algorithms based on statistical rather than biological properties.

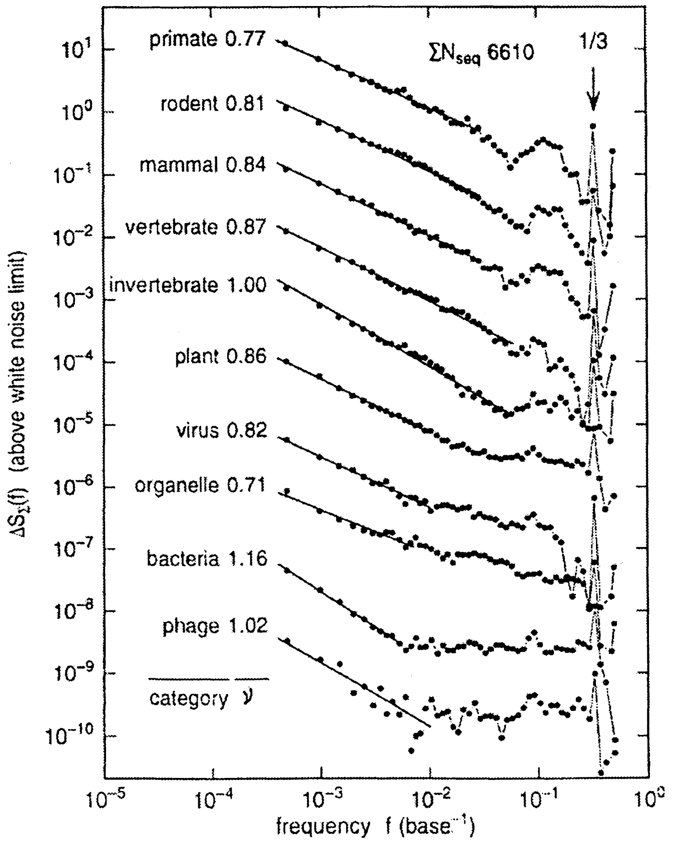

FIG. 3. Equal symbol spectrum analysis for several DNA sequences. The white noise background has been subtracted and data are offset for clarity. We notice that the measured ν's lead to values of $H > 0.5$ for all categories. In particular for prokariotes a behavior near the ballistic regime is observed. We remark that this symbolic measure is compatible with the dynamical interpretation of (10) only if $\nu < 1$, since $H = (\nu + 1)/2$ [adapted from 3(a)].

The inverse-power-law behavior of the spectra of the base-base correlation function is shown in Fig. 3. The 1/3 frequency peaks can be observed for cDNA sequences.

E. Onsager experiment

A method we have found particularly useful in the determination of the statistics of DNA sequences is the Onsager regression technique [13], according to which an initial macroscopic fluctuation $\xi(0)$ far from equilibrium regresses to equilibrium in a manner proportional to the equilibrium correlation function (11). Thus, if one views the DNA sequence as being a dynamical system whose statistical properties are unknown (in particular the sequence might not be long enough to guarantee the stationarity condition), then one can proceed as follows: create an initial state corresponding to a nonvanishing macroscopic fluctuation (this is done choosing as an initial condition the first L sites with value $+1$) and observe its regression to equilibrium.

We now make the assumption that the equilibrium exists, even if it is reached very slowly with an inverse-power-law decay rather than with an exponential regression, as in ordinary statistical mechanics. In this case the regression to equilibrium of the macroscopic fluctuation is proportional to the equilibrium correlation function. It has to be remarked that the Onsager method is eventually equivalent to a direct calculation of the correlation function, but this equivalence requires the stationary property on which the definition of $\Phi_\xi(t)$ itself rests. However, we can perform the Onsager experiment even without knowing if the sequence is long enough to guarantee the attainment of the correct equilibrium with averages in time. We see that in cDNA sequences the presence of oscillations helps us to deal with this problem in an unambiguous way. These oscillations help us to assess whether the system satisfies the stationarity condition necessary to properly define the equilibrium correlation function (11). In future work we plan to investigate how the Onsager experiment is able to draw information out of nonstationary sequences.

We stress that in the case where the system is able to reach a stationary condition, the Onsager experiment is an efficient way of determining the correlation function. The only disadvantage is that it requires fairly long sequences, also because of the effects produced by the finite length of the time series: the lower L is, the worse the statistics are [3,26].

As proved by work in progress, a beneficial aspect of the Onsager analysis method is that in addition to detecting the presence of inverse-power-law correlation functions, it is an effective and accurate method for the detection of short-range correlational features (e.g., first neighbors anticorrelations and features introduced by repeated sequences and by codon usage statistics).

IV. DATA ANALYSIS AND RESULTS

In this section we present results given by the methods discussed in the previous sections, when applied to real DNA sequences and to the CMM generated sequences. Figures 2 and 3 are extracted from the papers of Stanley *et al.* [5] and of Voss [3], respectively. We notice that while there is substantial agreement on the presence of long-range correlation in intron containing sequences, there is substantial disagreement for cDNA sequences. In particular, Voss [3] finds for viruses values of H near the ballistic regime, while for Stanley *et al.* [5] in viruses, like in the other prokariotes, there are no long-range correlations at all.

In our dynamical model, intronless and intron-containing sequences are generated with the same map. However, we have to choose the parameters in order to adapt the map to the DNA sequence explored. To generate cDNA sequences with the CMM we need parameter values such that $B \gg A$ in (34) and this is achieved by choosing a large copying mistake rate $1-\epsilon$. This superposition of anomalous and normal diffusion explains why the detection of the long-range correlation is so difficult and why the short-time dynamics are essentially dominated by the properties of standard diffusion ($H = 1/2$), as pointed out by the detrended analysis of Stanley *et al.* [5]. We emphasize that their major discovery, namely, a difference in correlation for the two kinds of sequences at short-time scale, holds true. However, we are not satisfied with their explanation that the patches of biases present in intronless sequences are unimportant and do not contribute to the asymptotic correlational properties.

Figure 4 shows landscapes generated by the random walk prescription applied to the human Cytomegalovirus strain AD169, compared to that of a CMM, with the parameters indicated. The qualitative similarity between the two landscapes is quite impressive. However, we find that quantitative measures applied to the two landscapes are even more impressive. We apply three kinds of analysis to the two data sets of Fig. 4, namely, the determination of H, H_H (stemming from Hurst analysis), and H_d (stemming from DFA). The results are indicated in Fig. 5 and lead us to the conclusion that the CMM generates sequences that are virtually indistinguishable from those of real DNA, using these measures. Notice that $H_H = H_d = H$ is also justified by the histograms for the probability density $P(x,t)$ shown in Fig. 6 whose widths increase with time and that are characterized by the following properties: (i) at short times, when the sequence is sufficiently long to provide adequate statistics, the histograms are essentially indistinguishable from Gaussian distributions; (ii) at later times, the lack of statistics makes it impossible to assess whether or not the densities develop long-range tails and consequently prove that the Gaussian assumption is incorrect. If DNA were really well described by our simple model, the theoretical distribution could not be Gaussian, since the Gaussian and the Lévy process act independently, and so we expect a linear superposition for $P(x,t)$. However, the Gaussian process is strong enough to destroy possible tails in the short-time regime, while in long-time regime the tails are destroyed by the errors generated by the finite length of the sequences.

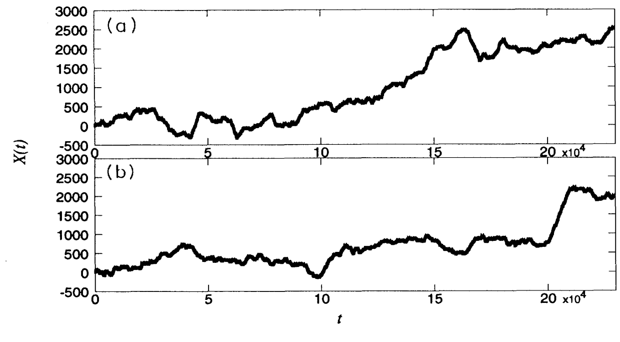

FIG. 4. (a) Landscape generated with the CMM with $z = 5/3$, $d = 0.45$, and $\epsilon = 1/9$. (b) Landscape generated by the Cytomegalovirus strain AD169. For both the landscapes the number of base pairs (BP) is 229 354.

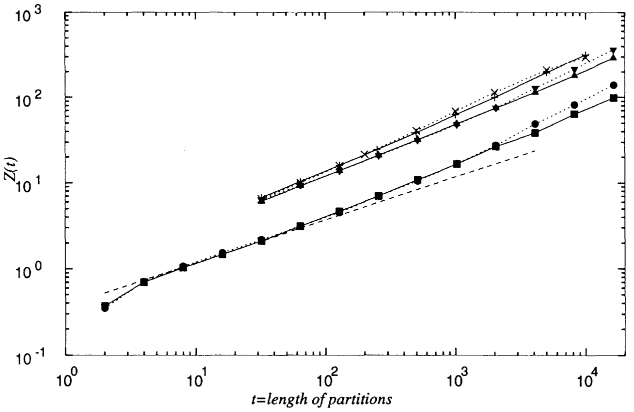

FIG. 5. Three analyses (from top to bottom: diffusion, rescaled hurst, and detrended) applied to the Cytomegalovirus strain AD169 sequence (solid curves) and to the CMM (dotted curves) with $z = 5/3$, $d = 0.45$, and $\epsilon = 1/9$. The function $Z(t)$ is defined as $\sigma(t)$, $R(t)$, and $F_d(t)$ [$F_d(t)$ as denoted in the work of Stanley *et al.* [5]] for the three analyses, respectively. The theoretical prediction for the CMM is $H_D = 3/4$, slightly larger than the slope of this curve. Notice that the slopes of the detrended curves for both the map and the virus change from a slope ≈ 0.5 (dashed line) for a short "time" partition to a slope > 0.5 for longer partitions.

Long-range correlations can also be detected by observing the eventual asymptotic inverse power law of the correlation function (11). As explained in Sec. III, we use the Onsager experiment to determine $\Phi_\xi(t)$. This method is equivalent to a direct evaluation of $\Phi_\xi(t)$ through the definition (11), but it has some technical advantages, since the ensemble over which the averages are performed is chosen throughout the whole sequence. So it is possible to detect an inverse power law with good statistics. In fact, it has to be pointed out that although $\Phi_\xi(t)$ implies the existence of a stationary condition, the direct observation of its slow relaxation to zero (with an

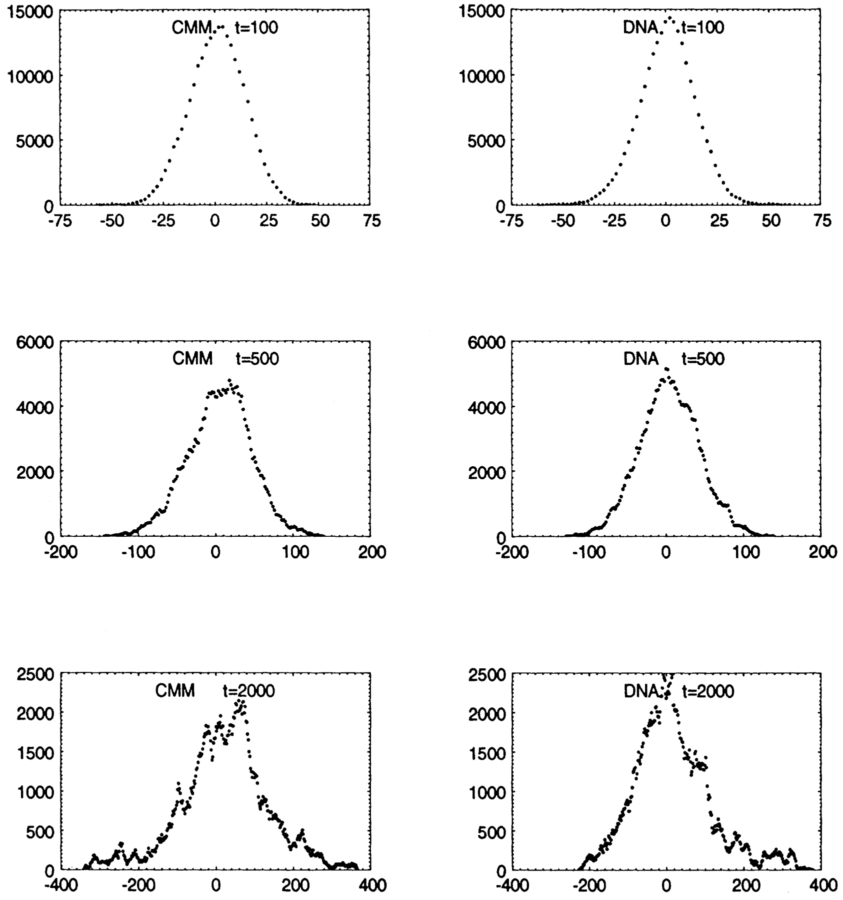

FIG. 6. Histograms of the probability distribution $P(x,t)$ relative to the CMM with $z = 5/3$, $d = 0.45$, $\epsilon = 1/9$, and to the Cytomegalovirus strain AD169 sequence (labeled as DNA).

inverse power-law form) might involve technical difficulties. The observation of the regression to equilibrium of a macroscopic fluctuation is a very efficient way of calculating $\Phi_\xi(t)$.

In Fig. 7(a) the regression to equilibrium of the map with no copying mistakes is shown. A log-log plot of the long-time regression, shown in Fig. 7(b), yields a straight line, indicating an inverse-power-law correlation function with index $\alpha = 0.5$, agreeing perfectly with the theoretical predictions of Trefán *et al.* [20] and others [19,21], for the map with $z = 5/3$. Notice that the agreement is good in the asymptotic limit, where the continuous treatment of the map dynamics is valid, and fulfills a stationary scaling relation according to the treatment from (29) to (32). The results of the Onsager regression experiment on the CMM and the DNA sequence are also shown in Fig. 7(a). We find that there is good qualitative agreement between the two. In both cases we notice a rapid regression to the white noise background at the first time step. This is in line with the analysis that led to (34). We see, however, that there is also a striking discrepancy: in the case of the DNA sequence, after the short-time regression to the level of background noise we find regular oscillations with a time period of 3. This property is the time counterpart of the peak found by Voss at $f = 1/3$ in the frequency spectra shown in Fig. 3 [3]. It is important to stress that the Onsager experiment makes it possible for us to establish a connection between such oscillations and anomalous diffusion, in a clearer way than using spectral analysis alone.

In Fig. 8(a) we see indeed that the maxima of the oscillations regress to equilibrium with the power law $\alpha \simeq 0.5$ that is in perfect agreement with the theoretical predictions for the CMM shown in Figs. 4 and 5. This makes it clear that the detection of the anomalous diffusion is extremely delicate (for $B \gg A$) and is strongly affected by the coarse graining. The analysis that led us to the results of Fig. 5 is based on the adoption of time repartitions much larger than the period 3, thereby making it impossible to observe them.

The Onsager analysis of Fig. 8(a) implies $H = 3/4$, which is slightly larger than the value given by the analysis of the finite sequence (Fig. 5); this prediction is, however, in very good agreement with the CMM studied in Fig. 4. Note that the analysis of this paper (Fig. 5) proves this CMM to be equivalent to the DNA sequence and that the exact H of the CMM is known [19–21]. This is so because using (22) and (32) H is directly derived from the parameter z of the CMM ($z = 5/3$ leading to $H = 3/4$). The discrepancy between the standard equilibrium analysis and the "correct" prediction of the nonstationary Onsager analysis might become dramatically important when $H \sim 1$.

In Fig. 8(b) we plot the results of the Onsager experiment in a way that we can distinguish between two sets of points. The curve formed by the points in position $3n+1$ (the "peaks" curve), where n is an integer number, and the curve formed by the others points are separated in the short-time regime and they merge only at later times. The points that do not belong to the peaks curve form a kind of white noise background under the peaks curve itself. The absence of large drifts in the lower uncorrelalated curve tells us that the stationary assumption is fulfilled by our sequence. We can actually take the level of this curve as the true base line for computing the correlation function. We notice that for the splitting to become detectable we need fairly long sequences. However, when the splitting becomes visible, this represents an unambiguous method to remove part of the white noise background. We also stress that this behavior is typical of viruses of the same kind and we plot another example (Varicella Zooster) in Fig. 8(c).

As stressed by Voss [3], the period-3 oscillations correspond to the number of bases present in a codon. To shed light on these oscillations we have built up three subsequences relative to the position (1, 2, or 3) of the bases within their codons. We have established that the three separate subsequences are characterized by a regression

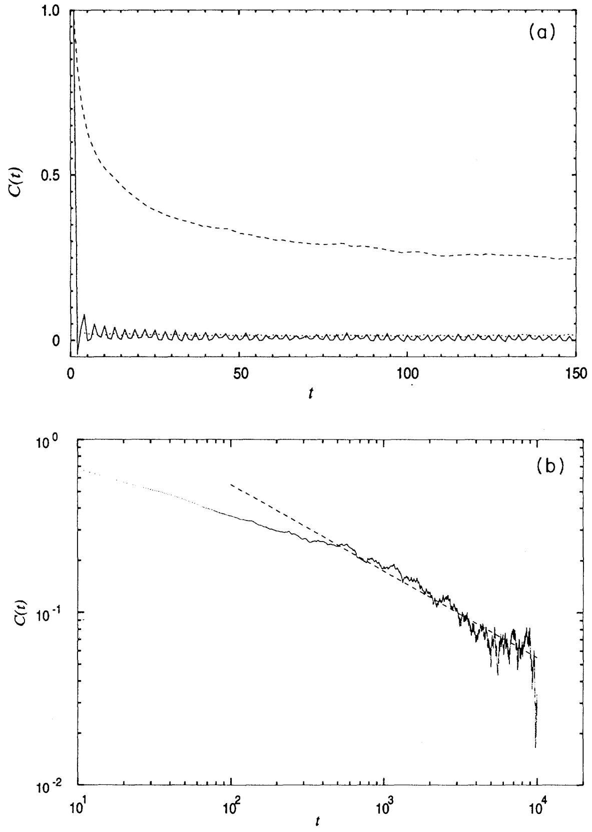

FIG. 7. (a) Onsager regression function $C(t)$ (where t denotes the discrete time) relative to the Liebovitch-Tóth map (dashed line), to the Cytomegalovirus strain AD169 sequence (solid line), and to the CMM with $z = 5/3$, $d = 0.45$, and $\epsilon = 1/9$ (dotted line). The number of initial conditions is $L = 10^5$. (b) $C(t)$ relative to the Liebovitch-Tóth map. The dashed line represent a fit curve relative to $\beta = 1/2$ and therefore $H = 3/4$.

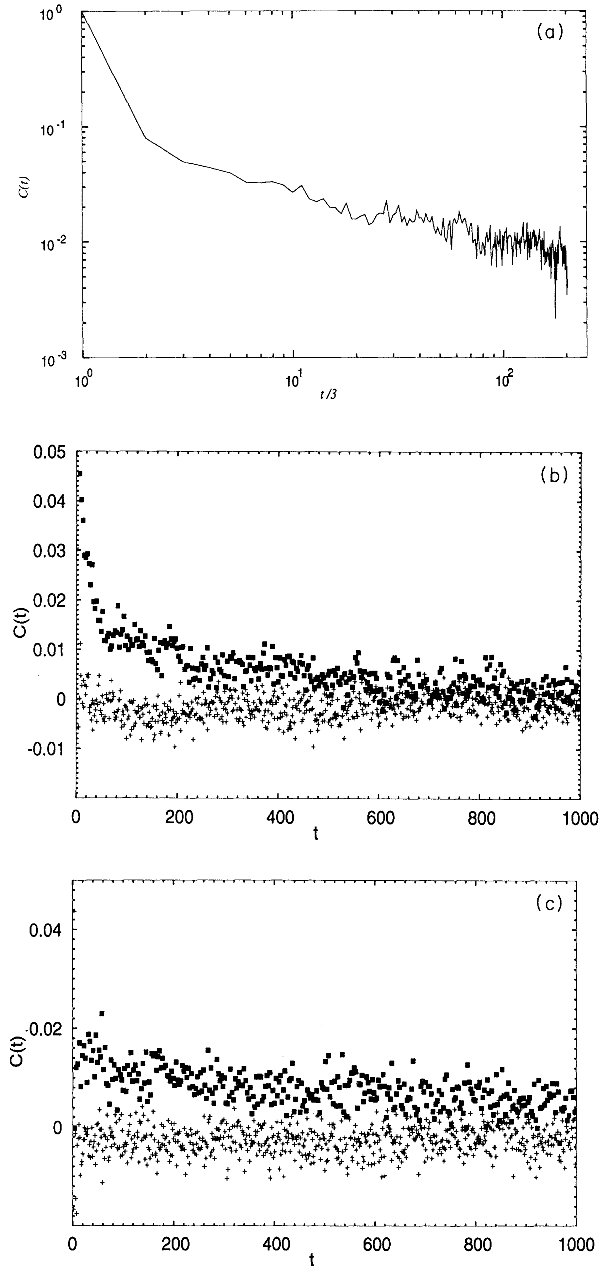

FIG. 8. (a) First relative maxima of the Onsager regression function $C(t)$ (with $L = 10^5$) concerning the Cytomegalovirus strain AD169 sequence. All these maxima are located in position $3t+1$, where t denotes the discrete time. (b) Same sequence as in (a), with all points plotted. Those corresponding to the position $3n+1$ are denoted by solid squares, all the remaining points are denoted by $+$'s. (c) Same as (b), for the Varicella Zoster Virus genoma. Here $L = 60\,000$; the length of the sequence is 124 884 BP.

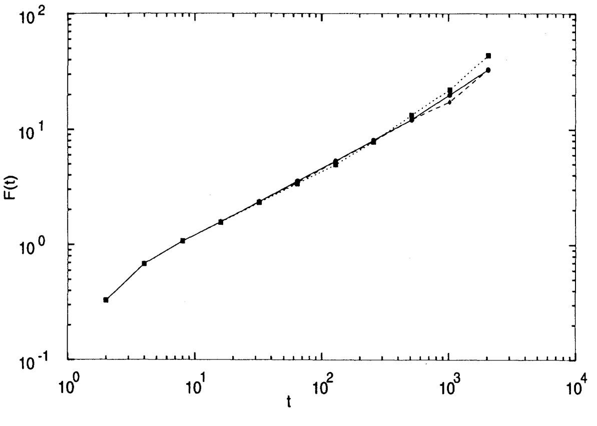

FIG. 9. DFA performed on the three subsequences of Cytomegalovirus Strain AD169. The solid line (circles) is relative to the subsequence of the bases on the first position in codons; the dashed line (diamonds) represents the analog for the second position; the dotted line (squares) is the analog for the third position in codons.

with a short-time inverse power law and no regular oscillation, in full agreement with (34). The regression of the three independent subsequences agrees quite well with the predictions of CMM with the right amount of copying mistakes (reduced by 1/3 as expected on the basis of simple arguments). In Figs. 9 and 10 we show the behavior of these sequences under the DFA and the Onsager analyses, respectively. On the basis of these results we are forced to rule out the attractive biological conjecture [27] that the correlation in cDNA sequences could be attained through the degeneracy (presence of synonyms) of the code. Since most of the synonyms are due to the change of the third base in the codon, one would expect the subsequence relative to the third position to be much more correlated than the other two. This is only partially true, as we can see in Fig. 10.

The presence of the period-3 oscillations in Fig. 8 suggests that the three subsequences are mutually independent, i.e., there is no correlation among them. This means that in the context of our model cDNA sequences are described by three independent CMM maps, entangled in the following way. One map describes the values

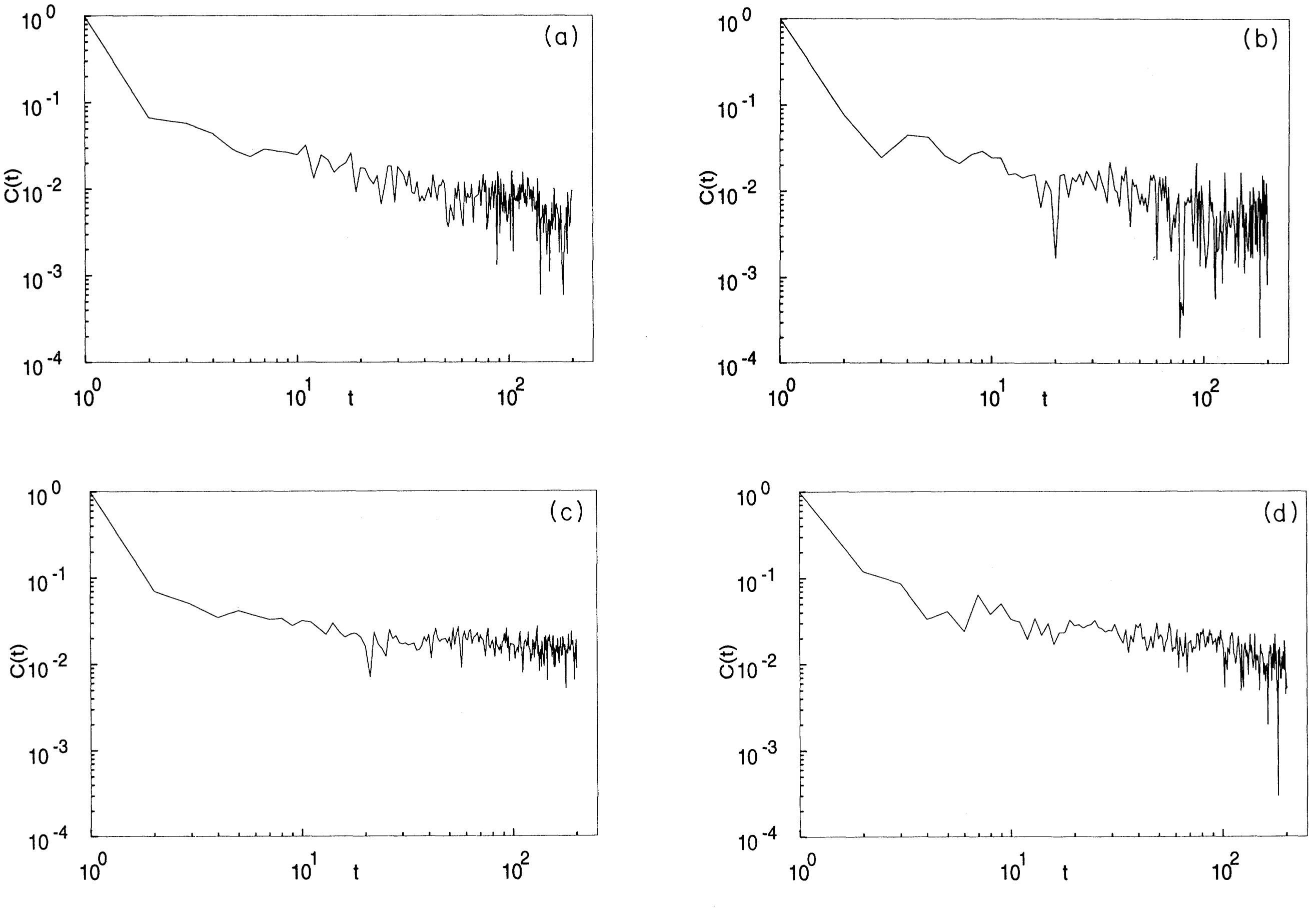

FIG. 10. Onsager experiment with the same $L = 30\,000$ for (a) the subsequence relative to the first position in the codon, (b) the subsequence relative to the second position in the codon, (c) the subsequence relative to the third position in the codon, and (d) a human intron containing the sequence HUMHBB (human beta-globin chromosomal region) of total length 73 239 BP. Comparing (d) with, e.g., (c), we notice that this human intron containing the sequence HUMHBB presents the same degree of correlation as the intronless subsequences of the Cytomegalovirus, a virus affecting human beings.

of the sequence for positions $3n$, a second for positions $3n + 1$, and a third for positions $3n + 2$. In future work we plan to study the formal statistical properties for this kind of entanglement. For the time being we limit ourselves to predict, on the basis of Sec. II, that the asymptotic behavior of the total waiting time distribution is an inverse power law if the three subsequences have the same power μ.

A significant result of the work reported herein is that long-range correlations show up in cDNA sequences, but this property becomes difficult to detect since the correlation is divided among three independent subsequences. We note that the adoption of a coarse-graining procedure, masking the fast oscillations, would generate sequences statistically equivalent to an unusually large copying mistake probability. This explains why in Sec. II C we have been compelled to adopt the condition $A \gg B$.

We must point out that such subsequences present a degree of correlation that is comparable to that found by Stanley *et al.* [5] in human intron containing sequences. Figure 10(d) shows the Onsager experiment on the gene HUMHBB, a human gene that contains mostly introns. A comparison between the subsequence relative to the third position in codons of the cytomegalovirus strand AD 169 and the sequence HUMHBB shows that the two sequences have the same degree of correlation and essentially the same H.

V. CONCLUSIONS

On the bases of the results obtained herein we are led to make the following final remarks.

(a) We have introduced a method of analysis of statistical sequences. The Onsager method is able to detect the principal correlational features of DNA sequences. No coarse-graining procedure is involved and no stationarity assumption is necessary. We will determine in future work whether it is possible for this method to extract information from nonstationary CMM and if there are real DNA sequences corresponding to this condition. Further work will be devoted to determine if the Onsager method is able to detect important short-range correlations due to the presence of repeated sequences, codon usage properties, and other biological features.

(b) We are now in a position to draw some interesting biological conclusions. The relative role of chance and necessity is one of the key issues in the present debate on evolutionary processes. According to the classical neodarwinist theory [28], mutations, the basis of genetic diversity, are the result of random discrete changes in the genetic material. Chance plays a fundamental role also in the dynamics of allele populations. Natural selection, on the other hand, acts on phenotypes, as an external constraint. More recently, structural internal constraints, due to self-organization rules, have been considered as an important factor in evolution, particularly within the framework of the somewhat extremist view of autoevolutionists [29].

The CMM can be interpreted as a compromise between chance and determinism since it clearly affords a dynamical picture that is the proper combination of the two, giving rise to a random fractal process. It must be stressed, however, that in this case the distinction between order and randomness is not so unambiguous, as it might seem at first sight. In fact, we must emphasize that the dynamics that gives rise to the long-range correlations is chaotic. On the other hand, the lack of short-range correlations (which in our picture is a consequence of the copying mistake fluctuations) is partly due to the fact that the ordered sequences of amino acids of the coded proteins are not long-range correlated [30] and thus the corresponding DNA sequences may result in a random distribution. We demonstrated that long-range correlations can also be seen in exons as well as in introns. A possible biological explanation could be the following. There are reasons why DNA sequences develop a global order. This means that all types of sequences obey the same structural constraints (*internal selection*), while the coding sequences must also obey the functional constraint of the coded proteins, and thus the relative mutations are affected by Darwinian natural selection acting on phenotype (*external selection*). Nevertheless, the latter are able to develop long-range correlations through a period-3 periodicity that is naturally connected with the codon structure and also partly via the degeneracy of the code [27]. This property can be detected only by looking at very long distance scales.

Lió *et al.* [8], who have observed another statistical property emerging via the code degeneracy, namely, a relative abundance of G and C nucleotides in the third position of the codons in several coding regions, have interpreted this periodicity as a signal for an internal selection constraint. The long-range correlation, modulated with the same periodicity, of the virus genomas of Fig. 8 might thus be the evolutionary response to certain internal constraints or to constraints of the host cell. Along this line, the map dynamics responsible for the correlated motion [first term of right-hand side of (34)] can be associated with the *internal selection*.

(c) Once the theoretical arguments of Sec. II A are established and the reason why we restrict our analysis to the region (19) is understood, the nature of the generator of a specific μ seems to become essentially irrelevant. However, it should be clear by now why a specific choice of a given generator has to be made. It is relatively simple to associate anomalous diffusion in the range with $1/2 < H < 1$ to the waiting time distribution (18) with the index μ in the range (19) and it is possible to prove that this is a dynamical realization of the Lévy processes [20,21]. However, as we have seen, the DNA sequences realize anomalous diffusion in a more complex way. First of all, the long-range properties are partially hidden by the interaction with the infinite-dimensional environment, thereby requiring the introduction of the CMM, which already implies a more complicated expression for the function $\psi(t)$. Furthermore, even in the long-time regime they show distinct high-frequency oscillations. Accounting for all these properties by means of a suitable choice of a waiting time distribution $\psi(t)$ would have turned out to be extremely complicated and would not have illuminated these biological processes. This suggested that we choose a specific generator of an inverse power law and we deal with the intricacies of DNA sequences through the joint use of this generator and random noise or the joint use of three distinct generators.

Which is the most convenient generator of a distribution of waiting times with an inverse power law? The choice of a nonlinear deterministic map as a generator of an inverse power law is suggested by the philosophical arguments we illustrate in (d). In addition to this, as pointed out in the Introduction, this choice is also dictated by the criterion of efficiency. To make a convincing example, let us refer to the work of Liebovitch. This author made two distinct proposals to account for the fractal nature of the dynamics of ion channels. One is a deterministic map, of which (26) is a special version. The other is a model where the inverse power law of the distribution of waiting times in the closed channel state [31] is determined by the fact that, due to the thermal activation, which in turn implies the interaction between the system and its environment, the particle jumps into wells of increasing depth. Although the resulting $\psi(t)$ in the latter case is the same as that given by the deterministic map, the numerical realization of it would require much more computer time and would be less efficient than the deterministic map. Thus we see that our choice is dictated by a criterion of efficiency, elegance, and conceptual clarity. The adoption of the model where the waiting time distribution with an inverse time distribution is already determined by thermal fluctuation and thus by the interaction with the environment would have made the distinction between determinism and randomness very confusing.

In future work we plan to investigate if it is possible to distinguish between these two kinds of processes and to shed light on the chance-necessity dichotomy in evolution.

(d) Herein we established that if DNA sequences result, as they do, in long-range correlations, and if they are stationary processes, then they must be α-stable Lévy processes. This conclusion opens up avenues for further interesting work of both biological and statistical research relevance. To make this aspect transparent let us consider (8). The power μ of the waiting time distribution of the variable ξ in one of the two states 1 or -1, in principle, can range from $\mu = 1$ up to infinity. The region $\mu > 3$ is where the usual form of the central limit theorem is recovered since it is possible to define the finite time scale τ of (12). The region with $\mu < 2$ is incompatible with equilibrium. We thus reach the conclusion that a complex system, either physical or biological, "aiming" at realizing long-range correlations with no conflict with the requirement of taking place at equilibrium, has to locate itself in the region $2 < \mu < 3$. In the case of Hamiltonian systems, the reasons why the waiting time distribution $\psi(t)$ must have an inverse-power-law structure of (18) are known [16], and are related to the fractal nature of the region at the border between the chaotic sea and deterministic islands. However, it is not yet known how to derive μ from the Hamiltonian and consequently it is not yet understood why the power μ so far has always been found, as a result of numerical observation, in the range (19). We are tempted to say that it must be so if we have to fulfill the tenet that thermodynamical equilibrium is compatible with a microscopic derivation, but more direct evidence would be welcome.

In the case of DNA sequences, the situation is similar: the implication of DNA tertiary structure in eukariotes [2] represent a biophysical argument, which can be considered the counterpart of the Hamiltonian arguments mentioned earlier, establishing that the distribution of waiting times must be an inverse power law. Moreover, from the work of Grosberg *et al.* we can obtain a prediction for what concerns the value of μ, that is, 2/3. However, this prediction is not valid for prokariotes and we have shown that we have long-range correlation at least in long-DNA-sequence viruses affecting human beings. What has this to do with a "Darwinian" selection for viruses? In other words, for an external DNA segment to be efficient in a host cell, is it necessary for it to obey certain constraints, one of them being a long-range correlation with a certain power? This kind of conjecture has to be tested in future work for the important genetical implication that it may have.

However, we think that the analysis made herein indirectly proves that, within a coarse graining at least, the states 1 and -1 are characterized by an inverse-power-law distribution of waiting times. Where is the power μ expected to be located in the case of DNA sequences? The analyses of this paper show that in the case of DNA sequences the condition (19) corresponding to the dynamical realization of Lévy processes is fulfilled. However, we cannot rule out the possibility that in some cases $\mu < 2$, which would correspond to ballistic motion, with $H = 1$, studied by Zumofen and Klafter [21]. We cannot rule out the possibility that $H > 1$ either. This would be in conflict with either the assumption that the stationary correlation function (11) can be defined or that the fluctuations of ξ are independent of x. We have adopted the strategy of first investigating the consequences of a dynamical model compatible with (10). In the future we shall focus on the possible violations of this picture and, if they exist, on their effects.

We must also point out that according to the dynamical modeling for the DNA sequences proposed in this paper, such sequences are not derived only from the deterministic map but they result from the joint dynamics of the deterministic map and noise (CMM) or from the parallel run of three independent CMM's. This makes it difficult, if not impossible, to establish all the dynamical properties in terms of a simple analytical expression of $\psi(t)$. However, it also has the attractive effect of disclosing a channel for the short-time dynamics to show up in the long-time regime, which according to traditional wisdom is expected to be dominated by an inverse power-law decay, if this exists. We think that the statistical dynamics of these processes is worth a more careful analysis.

ACKNOWLEDGMENTS

We thank M. Buiatti and L. Galleni for helpful discussions and suggestions. B. J. W. and P. A. would like to thank the Office of Naval Research for partial support of this work.

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

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

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

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

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

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

[7] P. Allegrini, M. Barbi, P. Grigolini, and B. J. West (unpublished).

[8] P. Lió, S. Ruffo, and M. Buiatti, J. Theor. Biol. **171**, 215 (1994).

[9] H. Herzel, W. Ebeling, and A.O. Schmitt, Phys. Rev. E **50**, 5061 (1994).

[10] Note that throughout this paper we denote by $C(t)$ the correlation function as a "dynamical" property derived from the analysis of experimental data, adopting either the Voss prescription [3] or the Onsager regression method of Sec. III E. We shall adopt the symbol Φ_ξ to stress the connection with the dynamical theory of Sec. II A. In the case of analyses with no error $C(t)$ should concide with $\Phi_\xi(t)$.

[11] B. J. West and W. Deering, Phys. Rep. **246**, 1 (1994).

[12] L. S. Liebovitch, Math. Biosc. **93**, 97 (1989).

[13] L. Onsager, Phys. Rev. **37**, 405 (1931); **38**, 2265 (1931).

[14] R. Ishizaki, H. Hata, T. Horita, and H. Mori, Prog. Theor. Phys. **84**, 179 (1990); R. Ishizaki, T. Horita, T. Kobayashi, and H. Mori, *ibid.* **85**, 1013 (1991).

[15] J. Klafter, G. Zumofen, and M. F. Shlesinger, Fractals **1**, 389 (1993); G. Zumofen and J. Klafter, Europhys. Lett. **25**, 565 (1994).

[16] J. D. Meiss and E. Ott, Phys. Rev. Lett. **55**, 2741 (1985); Physica D **20**, 387 (1986).

[17] J. D. Hanson, J. R. Cary, and J. D. Meiss, J. Stat. Phys. **39**, 27 (1985).

[18] C. F. F. Kearney, Physica D **8**, 360 (1983).

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

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

[21] G. Zumofen and J. Klafter, Phys. Rev. E **47**, 851 (1993).

[22] L. S. Liebovitch and T. I. Tóth, J. Theor. Biol. **148**, 243 (1991).

[23] J. Feder, *Fractals* (Plenum, New York, 1988).

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

[25] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, Phys Rev. E **51**, 5084 (1995).

[26] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993).

[27] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).

[28] *Genetic Constraints on Adaptive Evolution*, edited by V. Loeschcke (Springer-Verlag, Berlin, 1987).

[29] A. Lima-de-Faria, *Evolution without Selection: Form and Function by Autoevolution* (Elsevier, Amsterdam, 1988).

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

[31] In the original model of Liebovitch only the closed channel state is assumed to have a dynamics leading to an inverse-power-law distribution. It would straightforward, however, to use this modeling for both states, thereby leading to a model totally equivalent to that illustrated in Sec. II A.

[32] C.-K. Perg *et al.*, Phys. Rev. E **49**, 1685 (1994).