Non-Gaussian statistics of anomalous diffusion: The DNA sequences of prokaryotes

Paolo Allegrini,[1,2] Marco Buiatti,[2,3] Paolo Grigolini,[2,3,4] and Bruce J. West[4]
[1]*Istituto Nazionale di Fisica della Materia, Unitá di Ricerca di Pisa, Piazza Torricelli, 2-56100 Pisa, Italy*
[2]*Dipartimento di Fisica dell'Universitá di Pisa, Piazza Torricelli 2, 56100 Pisa, Italy*
[3]*Istituto di Biofisica del Consiglio Nazionale delle Ricerche, Via S. Lorenzo 26, 56127 Pisa, Italy*
[4]*Center for Nonlinear Science, University of North Texas, P.O. Box 305370, Denton, Texas 76203-5370*
(Received 26 January 1998; revised manuscript received 5 May 1998)

We adopt a non-Gaussian indicator to measure the deviation from Gaussian statistics of a diffusion process generated by dichotomous fluctuations with infinite memory. We also make analytical predictions on the transient behavior of the non-Gaussian indicator as well as on its stationary value. We then apply this non-Gaussian analysis to the DNA sequences of prokaryotes adopting a theoretical model where the ''DNA dynamics'' are assumed to be determined by the statistical superposition of two independent generators of fluctuations: a generator of fluctuations with no correlation and a generator of fluctuations with infinite correlation ''time.'' We study also the influence that the finite length of the observed sequences has on the non-Gaussian statistics of diffusion. We find that these non-Gaussian effects are blurred by the joint action of short-range fluctuation and sequence truncation. Nevertheless, under proper conditions, fulfilled by all the DNA sequences of prokaryotes that have been examined, a non-Gaussian signature remains to signal the correlated nature of the driving process. [S1063-651X(98)11009-7]

PACS number(s): 87.10.+e, 05.40.+j, 33.15.Vb

I. INTRODUCTION

The pioneer work of Rahman [1] on liquid argon contains, among many interesting statistical properties, the numerical evaluation of a non-Gaussian indicator, whose one-dimensional version would read

$$\sigma(t) \equiv \frac{\langle x^4(t) \rangle}{3\langle x^2 \rangle^2} - 1. \tag{1}$$

Rahman found that the intensity of this non-Gaussian indicator vanishes on the initial condition, grows with increasing time, reaches a maximum, and then makes a slow regression to zero for t tending to infinity.

Is the result found by Rahman universal? To address this issue we consider another example, derived again from molecular-dynamics simulation, but concerning the much more complex case of binary alloys quenched into glassy states. This is the more recent work by Miyagawa and Hiwatari [2]. These authors find that in the glassy state the non-Gaussian indicator $\sigma(t)$ after reaching a maximum shows no sign of regressing to zero. There is a striking difference, therefore, with the earlier result of Rahman [1]. Notice that the former case [1] is an example of non-Gaussian behavior of a dynamical system with a time scale separation between the macroscopic and the microscopic regime, whereas the latter [2] is the non-Gaussian behavior of a dynamical system with no time scale separation. In this paper we explore the consequence of this lack of time scale separation using a simple model and the results obtained are shown to result in non-Gaussian effects even more intense than those found in quenched glasses [2].

What are the origins of non-Gaussian properties? According to the perspective afforded by the general discussion of Ref. [3], non-Gaussianicity is a consequence of microscopic nonlinearity and, at the same time, of memory, namely, of an incomplete time scale separation between the ''macroscopic'' diffusing variable and microscopic dynamics. According to Ref. [3], we must imagine that there exist two levels: the macroscopic and the microscopic. For any random phenomenon of interest there exists a variable responsible for its fluctuations that is closer to the microscopic level than the variable being measured. For instance, the microscopic variable related to the measured position of a Brownian particle is the velocity. The microscopic variable corresponding to the measured velocity is the acceleration and so on.

For any step closer to the microscopic level the nonlinear nature of dynamics becomes more significant. Moving in the reverse direction, from the microscopic to the macroscopic level, we see a suppression of the effects of nonlinearity. Thus, if we move from the level of acceleration to that of velocity, the suppression of microscopic nonlinearity is quantified by the formula [3]

$$G \propto \Omega W g^3. \tag{2}$$

Here G represents the strength of the nonlinearity on the velocity level, Ω denotes the frequency of oscillation of the tagged particle in the cage of the surrounding particles, and W defines the strength of the microscopic nonlinearity, namely, that of the level next to the velocity level, which is the temperature-dependent harmonic strength of the potential within which the particle of interest oscillates. The parameter g defines the strength of the memory

$$g \equiv \frac{\Omega}{\Gamma}, \tag{3}$$

where $1/\Gamma$ denotes the relaxation time of the center of the cage.

This theoretical prediction suggests that the strength of nonlinearity, and thus of non-Gaussian behavior, transmitted

from a given level to the next level closer to the macroscopic world, becomes larger and larger with increasing memory. In the case of an infinitely long memory, the non-Gaussian indicator given by Eq. (1), as we shall see, becomes infinitely large, thereby implying technical problems in recording it. For this reason, we adopt another form of non-Gaussian indicator

$$\eta(t) \equiv 1 - 3\frac{\langle x^2(t)\rangle^2}{\langle x^4(t)\rangle}, \tag{4}$$

which is related to Rahman's indicator by

$$\sigma(t) = \frac{\eta(t)}{1-\eta(t)}. \tag{5}$$

In Sec. II we shall see that in the case of infinite memory $\eta(t)$ tends to the value of 1 and so $\sigma(t)$ tends to infinity. Thus we find it more convenient to use the kurtosis $\eta(t)$ rather than Rahman's measure of non-Gaussianicity $\sigma(t)$.

The meaning of this result is that the infinite memory of a dichotomous fluctuation results in an infinite non-Gaussian strength (if Rahman's indicator is adopted). The DNA sequences offer an interesting example of dichotomous fluctuations with infinite memory. However, the DNA sequences are *truncated* single trajectories and this is a reason why in Sec. III we discuss the influence exerted on non-Gaussian statistics by the finite length of the sequences under investigation. The long-range correlations of prokaryotes are also perturbed by uncorrelated fluctuations and for this reason we devote Sec. IV to the study of the effects produced on non-Gaussian statistics by the joint action of short-range fluctuations and of the finite length of the sequences under study.

Section V is devoted to establishing the statistical significance of the results obtained in this paper. The content of this section aims at the very important purpose of explaining why the generalization of the earlier work of Ref. [3], which in turn is an extension of that of Rahman, finds a useful application in the field of DNA sequences.

II. THEORY FOR THE CASE OF INFINITE MEMORY

In this paper we go beyond the limits of the time-scale separation with a specific picture in mind where the macroscopic variable is the *position* and the corresponding microscopic variable is the *velocity*. In the case of DNA sequences this means position and velocity in the sense specified in Sec. IV. Let us examine a microscopic condition where the departure from Gaussianicity is as large as possible and study how these statistics are transmitted to the next level, closer to the macroscopic world, in the specific case when there is no time-scale separation between levels. With this ideal condition in mind we study the diffusion process

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

where x is the position variable and ξ is a dichotomous stochastic process, namely, a stochastic velocity variable with only two possible values $\pm W$. The two values are set equally probable to make the diffusion process unbiased. The choice of a dichotomous fluctuation means that the ''microscopic'' dynamics is strongly non-Gaussian and the equilibrium value of the corresponding kurtosis is

$$\eta_\xi^{eq} = 1 - 3\frac{\langle \xi^2\rangle^2}{\langle \xi^4\rangle} = 1 - 3\frac{W^4}{W^4} = -2. \tag{7}$$

This is the measure of the microscopic statistics (MICROS) of the system. We want to explore the case of infinite memory, so we must make a proper choice of the autocorrelation function

$$\Phi_\xi(t) \equiv \frac{\langle \xi(0)\xi(t)\rangle}{\langle \xi^2\rangle}. \tag{8}$$

A convenient choice is

$$\Phi_\xi(t) = \frac{A}{(A^{1/\beta}+t)^\beta}, \tag{9}$$

with the power-law index in the interval

$$0 < \beta < 1. \tag{10}$$

This choice makes the autocorrelation function nonintegrable or, equivalently, the microscopic time scale become infinite while fulfilling the normalization constraint $\Phi_\xi(0) = 1$.

We now assess the statistical properties of the variable x. It has been shown [4] that the corresponding diffusion process becomes equivalent to a truncated Lévy process, namely, a process whose distribution function is described by a Lévy distribution whose tails have been eliminated. This is a diffusion process with a finite propagation front. The probability distribution at distances $|x| > Wt$ vanishes and the population of the missing tail concentrates on the front thereby results in two peaks. A very accurate representation of the probability distribution $P(x,t)$ is given by

$$P(x,t) = P_L(x,t)\Theta(Wt - |x|) + \frac{\Phi_\xi(t)}{2}\delta(Wt - |x|), \tag{11}$$

where Θ is the Heaviside step function, δ is the Dirac delta function, $P_L(x,t)$ is the inverse Fourier transform of

$$\hat{P}_L(k,t) = e^{-b|k|^{\beta+1}t}, \tag{12}$$

and the parameter b is defined in terms of the parameters of the fluctuation process as

$$b \equiv \frac{\pi\beta(\beta+1)AW^{\beta+1}}{2\sin\left(\frac{\pi(\beta+1)}{2}\right)\Gamma(\beta+2)}. \tag{13}$$

Equation (11) is the macroscopic statistics (MACROS) of the system: $P(x,t)$ is not Gaussian as a consequence of the lack of a finite microscopic time scale.

This picture makes it possible to solve the problem of the time evolution of the kurtosis $\eta(t)$. From Eq. (11) we immediately find

$$\langle x^{2n}(t)\rangle = \int_{-Wt}^{+Wt} x^{2n} P_L(x,t)dx + t^{2n}\Phi_\xi(t). \tag{14}$$

In the time asymptotic limit, from Eq. (9), we obtain

$$\Phi_\xi(t) \simeq At^{-\beta}. \tag{15}$$

Using the time asymptotic properties of the Lévy distributions [5], we obtain, using Eqs. (12) and (13) and setting $W=1$,

$$P_L(x,t) \simeq \frac{ct}{|x|^{\beta+2}}, \tag{16}$$

with c given by

$$c = \frac{A\beta(\beta+1)}{2}. \tag{17}$$

Thus, from Eq. (14) we obtain the approximated expressions

$$\langle x^2(t)\rangle \simeq \left(A + \frac{2c}{1-\beta}\right) t^{2-\beta} \tag{18}$$

and

$$\langle x^4(t)\rangle \simeq \left(A + \frac{2c}{3-\beta}\right) t^{4-\beta}. \tag{19}$$

As a consequence, we make the following prediction for the kurtosis in the time-asymptotic limit:

$$\eta(t) \simeq 1 - c_\eta t^{-\beta}, \tag{20}$$

with

$$c_\eta \equiv \frac{3\left(A + \frac{2c}{1-\beta}\right)^2}{\left(A + \frac{2c}{3-\beta}\right)}. \tag{21}$$

In conclusion, using the theory of [4], we predict the long-time limit kurtosis when the physical condition (10) applies, namely, in the case of infinite memory. On the other hand, we know that the initial value of the kurtosis must be -2, because at very short times the statistics of x are dictated by ξ and this variable in turn must fulfill the condition (7). Assuming that the time evolution of the kurtosis does not undergo any abrupt change, we conclude that the kurtosis increases from the initial value of -2 and, monotonically increasing, tends to the time asymptotic value of 1. We note that the resulting behavior is reminiscent of Ref. [2], thereby suggesting that the lack of a time scale separation provokes the breakdown of the regression of the kurtosis to 0, namely, to the Gaussian behavior, and that, eventually, with increasing time, stationary non-Gaussian statistics are reached.

We note from Eq. (20) that the process of transition to the non-Gaussian regime predicted by the model of Ref. [4] becomes slower and slower as β approaches 0. This slowing down implies that the strongly non-Gaussian MICROS change slowly into the MACROS and, coming closer to the limiting case of $\beta=0$, the transition becomes so slow as to never depart from the initial dichotomous statistics. This is expected on the basis of the following simple argument: We know that at $\beta=0$ the resulting diffusion process is essentially ballistic, thereby implying

$$\langle x^2(t)\rangle = W^2t^2 \tag{22}$$

and

$$\langle x^4(t)\rangle = W^4t^4, \tag{23}$$

so that, consequently,

$$\eta(t) \simeq -2. \tag{24}$$

What about the case $\beta>1$? In this case a finite microscopic time scale is recovered since the time integral of the correlation function is finite. Consequently, we expect that the standard condition detected by Rahman years ago [1] is recovered. After a transient regime, corresponding to a time scale where the process of molecular collisions can be perceived, the system must reach the regime of conventional diffusion, thereby implying that the Gaussian statistics is recovered, with kurtosis $\eta(t)=0$.

To check the validity of these theoretical predictions we made a numerical calculation based on the average over a large number of trajectories derived from a very extended single trajectory, which in turn is generated by means of a stochastic generator [4]. The method used in [4] is essentially a method to produce a dichotomous fluctuation ξ with a stationary correlation function corresponding to that of Eq. (9). The adoption of a deterministic generator [6–8] producing the same stationary correlation function would lead to the same result: The advantage of using the stochastic generator is mainly due to the higher computational accuracy and speed. Note that the distinct trajectories, on which the averaging is carried out, are derived from the single but extended trajectory, shifting the initial condition. This method, in principle, aims at realizing an ensemble virtually equivalent to an equilibrium Gibbs ensemble [7].

The results of this numerical calculation are illustrated in Fig. 1, which refers to the time evolution of $\psi(t) \equiv 1-\eta(t)$ as well as of $\eta(t)$. Note that the results of the cases (a), (b), and (c) of Fig. 1, corresponding to three distinct values of $\beta<1$, are in a good qualitative agreement with the prediction of Eq. (20). The results of Fig. 1(d), with $\beta=1.5$, show the regression to the Gaussian statistics, corresponding to the prediction of the pioneer work of Rahman [1]. We see that before regressing to the prediction of Gaussian statistics the kurtosis overshoots the value $\eta(t)=0$. This overshooting results in a pronounced maximum, and the birth of a maximum, as we shall see, is an interesting property produced also by the finite length of the explored sequences. To establish the significance of this latter property, however, it is necessary to go through a deeper discussion of the method adopted to derive a Gibbs ensemble from a single sequence. This will be discussed in Sec. III.

What about the quantitative agreement between theory and numerical experiments? We point out that at short times and at low values of β, typically for values up to the order of 0.5, the quantitative agreement between theory and numeri-

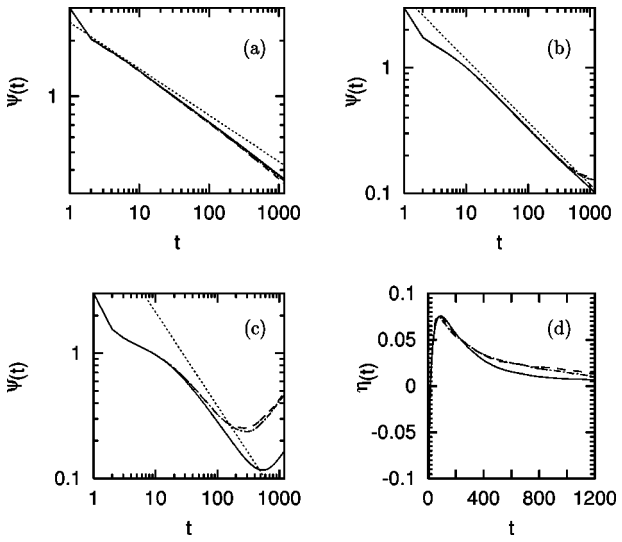

FIG. 1. Time evolution of the non-Gaussian indicator. (a) The value of $\psi(t) \equiv 1 - \eta(t)$. η and therefore ψ are dimensionless variables. The calculation was made by adopting a stochastic generator [4] (t is the number of iterations of the generator) resulting in the correlation function (9) with $A \approx 0.5$. The dotted straight line is a guideline indicating the theoretical slope $-\beta$, with $\beta = 0.25$. The sequence lengths T are as follows: $T = 3 \times 10^6$ (dashed line), $T = 10^7$ (dot-dashed line), and $T = 2 \times 10^7$ (solid line). (b) Same as (a), but with $\beta = 0.5$, $A \approx 0.25$, and $T = 10^6$ (dashed line), $T = 3 \times 10^6$ (dot-dashed line), and $T = 10^7$ (solid line). (c) Same as (a), but with $\beta = 0.75$, $A \approx 0.2$, and $T = 10^6$ (dashed line), $T = 2 \times 10^6$ (dot-dashed line), and $T = 10^7$ (solid line). (d) Time evolution of $\eta(t)$. Here $\beta = 1.5$, $A \approx 0.1$, and the sequence lengths are as follows: $T = 10^7$ (dashed line), $T = 2 \times 10^7$ (dot-dashed line), and $T = 6 \times 10^7$ (solid line). Notice that in all the cases the dot-dashed lines, clearly visible only in cases (c) and (d), refer to an intermediate case, with a statistics of intermediate accuracy. In cases (a) and (b) the dot-dashed lines virtually overlap the dashed lines.

cal calculations is very good. The agreement becomes worse at large values of β, typically those of the order of 0.75. This discrepancy is caused by the fact that the theoretical prediction is asymptotic in time, whereas it takes a certain amount of time for the system to reach the asymptotic regime. This time is estimated as the time it takes the correlation function (9) to reach a time regime where its dependence on time is a genuine inverse power law. The resulting value

$$t_{tr} \propto A^{1/\beta} \beta^{1/(\beta+1)} \tag{25}$$

quantifies how long it takes the correlation function to become an inverse power law. Consequently, in the range $0 < \beta < 1$, the decrease of β produces a faster transition to the time-asymptotic regime, thereby making it possible to fulfill the prediction (20) within the explored time range.

It has to be pointed out that the rule according to which the agreement between theoretical prediction and numerical results is improved with the smaller values of β is not completely true. In fact, with decreasing β another important property has to be taken into account. The lower the value of β, the more persistent the presence of ballistic peaks. A satisfactory numerical treatment would imply a significant increase of the number of systems in the Gibbs ensemble. If this is kept fixed, the statistical inaccuracy becomes larger and larger with the decrease of β. As a consequence, the best agreement between theory and numerical treatment is reached at intermediate values of β. In fact, we see that the case $\beta = 0.5$ of Fig. 1(b) yields an agreement between theory and numerical treatment much better than at $\beta = 0.25$ [Fig. 1(a)] as well as at $\beta = 0.75$ [Fig. 1(c)]. Notice that the statistical analysis is made on sequences of finite length according to the procedure described in Sec. III. The corresponding numerical results, illustrated in Fig. 1, provide significant signs of this tendency: The longer the sequence length T and consequently the more accurate the statistics available, the better the agreement between theoretical predictions and numerical results.

III. FINITE-SIZE-INDUCED REGRESSION TO GAUSSIAN BEHAVIOR

The qualitative discussion of Sec. IV refers to the statistical analysis of DNA sequences. As we shall see, a DNA sequence can be imagined as a time series and the length of this sequence is finite. To prepare the ground for the discussion of Sec. IV let us consider the case when the fluctuating variable $\xi(t)$ is observed at the discrete times t_i and the time interval is $t_{i+1} - t_i = 1$. It must be pointed out that the calculation illustrated earlier refers precisely to this condition and that the adoption of the continuous-time representation has been an idealization made possible by the fact that we are interested in the long-time limit. To make the earlier statistical analysis we used the possibility of computer generating practically infinitely many and infinitely extended sequences $\{\xi^{(r)}(t_i)\}$. Ideally a single sequence is obtained keeping r fixed and moving i from 1 to infinity. Changing from one given r to a different r' is equivalent to moving from a given system to another system of the Gibbs ensemble. Therefore, $\{\xi^{(r)}(t_i)\}$ can be interpreted as a mathematical notation defining this Gibbs ensemble. Since this set consists of infinitely many sequences and the length of any sequence is infinite, we shall refer to it as ideal Gibbs ensemble (IGE).

Now let us select one of these infinitely many sequences and truncate it at a given time T. We denote this single, truncated sequence with the name of sample sequence (SS). The challenging problem is now that of deducing the statistics of the IGE from the analysis of the SS. In principle, it is possible to derive from the SS a sort of simulation of the IGE. This is done as follows. The first system of the Gibbs ensemble is the SS itself. The second is a new sequence derived from the SS by shifting the time origin from t_1 to t_2. The nth trajectory is obtained shifting the time origin from t_1 to t_n and so on. Let us refer to this as the effective Gibbs ensemble (EGE).

It is evident that for $T \to \infty$ there should be no essential difference between the EGE and the IGE. This would make redundant the adoption of the superscript r to define the Gibbs ensemble: A single sequence would contain the same statistical information as that afforded by any other trajectory of the ensemble. Furthermore, it is evident that in the case of ordinary diffusion the adoption of the EGE rather than of the IGE would not affect the statistical analysis. In fact, when the autocorrelation of Eq. (9) has a finite lifetime τ [9] and $T \gg \tau$ the adoption of the EGE rather than the IGE does not produce any significant deviation from the original statistical

behavior. This is no longer expected to be the case when the power-law index is in the interval (10) where $\tau=\infty$. In this case, although the assumption of an equilibrium invariant measure is made, it takes an infinitely long time for the system to regress to equilibrium. Consequently, we expect that the adoption of the EGE might produce a significant departure from the statistics of the IGE for any finite length of the SS: This length, in fact, cannot be as large as the correlation time τ, which is infinite in this case.

We are not aware of any theoretical treatment of this difficult issue, except for a work of Peng *et al.* [10] in which a detailed discussion is made of the uncertainty affecting the local anomalous rescaling index $H(t)$ as a finite-size effect. We notice that that work rests on assuming the variable ξ to be a correlated Gaussian noise, thereby implying no deviation from a Gaussian MACROS. Therefore, we cannot apply that analysis to the case under discussion in this paper.

As regards the discussion of this issue in the present case, we essentially rest on computer simulation. A DNA sequence is a single truncated trajectory. However, within the theoretical perspective adopted in Ref. [11] this single trajectory is assumed to have the same statistical properties as those of a trajectory generated by a set of deterministic nonlinear equations corresponding to a condition of weak chaos (or, equivalently, to a stochastic generator of long-range correlations [4]). Thus we can produce as many SS's as we need to establish to what extent the departure of the EGE statistics from the IGE statistics depends on the SS considered. We consider 1000 independent truncated trajectories. We associate each SS with its own EGE and thus to its own $\eta(t)$. We find a distribution of these curves $\eta(t)$ and consequently we are led to define the mean value $\langle \eta(t)\rangle$. In addition to the mean value we also evaluate the standard deviation $\Delta\eta(t) \equiv[\langle \eta^2(t)\rangle-\langle \eta(t)\rangle^2]^{1/2}$, which measures the spreading about the mean value and thus the ''error'' affecting the evaluation of $\eta(t)$.

On the basis of the calculations illustrated in Fig. 2, as well as of others that are not reported here for the sake of brevity, we reach the following conclusion. For any finite length T we generate a kind of bent ''sausage'' of the same type as that illustrated in Fig. 2. The width of the sausage $2\Delta\eta(t)$ and the time at which the sausage reaches its maximum level depend on T. The larger the T the thinner the resulting sausage and the longer the time at which the maximum value is reached. In Fig. 2 we denote by the dashed line the result of the same analysis applied to a sequence of so large length as to provide results virtually equivalent to those corresponding to the IGE and so a sausage with virtually a vanishing width and a maximum at infinite time. We see from Fig. 2 that at short times the sausage width is very thin and the mean value $\langle \eta(t)\rangle$ coincides with the dashed curve. At later times the sausage width increases, thereby making it possible for the single constituents of the sausage to significantly depart from the mean value $\langle \eta(t)\rangle$. However, it is also evident that the single constituents of the average value $\langle \eta(t)\rangle$ with high probability move within the sausage. We note from Figs. 2(b) and 2(c) that the whole sausage, after crossing the real axis at short times, and not only the mean value $\langle \eta(t)\rangle$, departs from the abscissa axis. After a finite time interval, which we denote as a *non-Gaussian window*, the sausage tends to include also the abscissa axis. On the basis of these properties we are led to conclude that most of the single constituents of the sausage are expected to share the same behavior, namely, a growth to a maximum value above the Gaussian level (the abscissa axis), followed by a regression to zero: a behavior reminiscent of that corresponding to $\beta>1$ [see Fig. 1(d)].

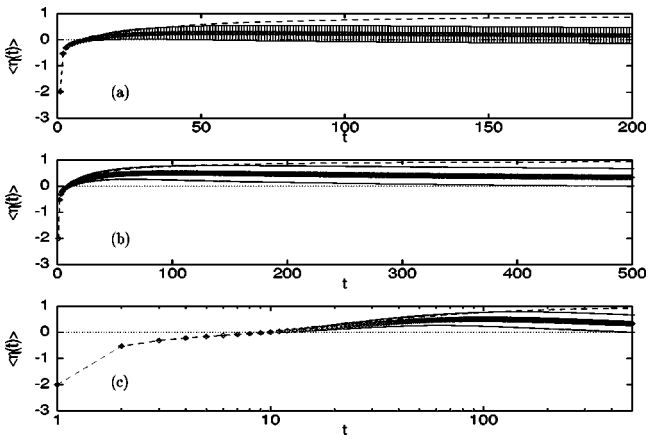

FIG. 2. Mean kurtosis $\langle \eta\rangle$ as a function of time, over an ensemble of 1000 EGE's resulting from 1000 independent single sequences of finite length T. The dotted line denotes the Gaussian level $\langle \eta(t)\rangle=0$. (a) $T=10\,000$. In addition to the mean kurtosis, denoted by the thick line resulting from a dense sequence of diamonds, we also plot the error bars (defining the width of the sausage mentioned in the text). The dashed line is the mean kurtosis of a single sequence with $T=10^7$. This case is expected to be a good realization of the IGE, with a virtually vanishing error bar. (b) Same as (a), but with $T=50\,000$. For the sake of clarity we plot only the upper and bottom values of the error bars, thus producing the two solid lines around the mean kurtosis. (c) Same as (b), but with a logarithmic rather than linear time axis. The calculation was made by adopting a stochastic generator [4] resulting in the correlation function (9) with $A\approx0.025$ and $\beta=0.5$.

To some extent the effect found numerically in this paper, on the influence of the truncation at the time T on the statistics of diffusion process, is similar to the influence that an external fluctuation of intensity D has on the statistics of dichotomous fluctuations with the correlation function (9). This problem has been discussed in earlier work of our group [12,13]. It has been shown that, as an effect of the perturbing noise, at a given time t_c, a crossover takes place from the slow decay regime to a faster, exponential-like, decay regime. The crossover time t_c is proportional to $D^{-\alpha}$, where α is an index of the order of unity. If the kurtosis of these processes were observed, the resulting behavior would be similar to that produced by $\beta>1$ [see Fig. 1(d)]. On the basis of the numerical results we are inclined to believe that the effect of using a finite sequence may be equivalent to introducing a disturbance of intensity $D\propto f(1/T)$, where the function $f(1/T)$ is a slowly increasing function of its argument. In other words, increasing T might have the effect of making the disturbance weaker. Note that in [12,13] an average over the natural Gibbs system was made, thereby producing single mean values rather than a set of distinct mean values.

It is evident that if the sausage is not thin enough it is not possible to conclude that all the SS's will produce a kurtosis with a maximum and a regression to Gaussian statistics within the observation time. The bump is produced by the

kurtosis overshooting the Gaussian plateau and regressing to it after a given time. Thus it corresponds to a window of finite size within which the non-Gaussian nature of the observed process becomes ostensible. In other words, if the non-Gaussian window is not ostensible, the observation of truncated sequences might generate the false impression that the statistics are Gaussian, in conflict with the detection of long-range correlation and with the observation [4] that in such a case the statistics of the resulting diffusion process cannot be Gaussian. It is also convenient to notice that the length of the DNA sequences that are analyzed in the next section is of the same order as that corresponding to the birth of the non-Gaussian window in Fig. 2.

IV. NON-GAUSSIAN STATISTICS OF DNA SEQUENCES OF PROKARYOTES

The statistical analysis of DNA sequences is carried out assigning the value $\xi=-1$ to purines and $\xi=1$ to pyrimidines [14,15] or vice versa [11]. Then the DNA sequences are conceived as the generators of fluctuations of the dichotomous variable ξ. After adopting this prescription the ''dynamics'' of DNA sequences become equivalent to the diffusion process generated by

$$x(N)=\sum_{n=1}^{N}\xi_i . \tag{26}$$

The ith position along the sequence can be thought of as a discrete time in a random walk process. Consequently, $x(N)$ can be regarded as being the position of a random walker at the discrete time N. If the sequence is very long, we adopt the continuous-time representation of Eq. (6), and consequently we can use the results of Sec. II to analyze the non-Gaussian properties of DNA sequences.

The results of some earlier investigations in this field [14,15,11,16–19] established the existence of long-range correlations, excluding the possibility that the paradigm of ordinary Brownian motions is the correct one to account for DNA statistics. Neither is the paradigm of fractional Brownian motion. In fact, as shown in [4], a dichotomous fluctuation with long-range correlations results in a distinctly non-Gaussian diffusion process: a truncated Lévy process [4]. The observation made by Arneodo *et al.* [20] that the diffusion statistics in eukaryotes are essentially Gaussian in spite of the existence of long-range correlations forced the authors of this paper to develop a folding model that has the effect of decoupling statistics from dynamics [21]. This means that statistics can be produced by a source distinct from that responsible for the long-range fluctuations, thereby explaining why the diffusion process can be approximately Gaussian in spite of the existence of long-range correlations.

As far as the prokaryotes are concerned, different conditions apply and these, as we shall see, make it possible to detect non-Gaussian statistics. Before addressing this issue, let us summarize the conclusions reached in literature on the DNA sequences of prokaryotes. Two groups have independently developed two seemingly distinct models, which, nevertheless turn out to be equivalent from a statistical viewpoint. Let us mention the earlier model first. In a recent paper, to interpret the long-range correlation in noncoding DNA, Buldyrev *et al.* [22] have adopted a generalization of the Lévy walk proposed in an earlier paper by Araujio *et al.* [23]. The process is realized as follows: At the jth step a random walker, in the case of an ordinary Lévy walk, makes a jump of size l_j forward or backward. Essentially the same result, except for the birth of a propagation front signaled by the presence of peaks [4], is obtained by assuming that the walker makes, in a time t_j, l_j steps in the same direction. Both of these assumptions would conflict with the idea of having no correlations at short distances. For this reason, in a recent paper Buldyrev *et al.* [22] assumed that a walker takes each of l_j steps in random directions, with a fixed bias probability

$$P_+=\frac{1+\epsilon_j}{2} \tag{27}$$

to go forward and

$$P_-=\frac{1-\epsilon_j}{2} \tag{28}$$

to go backward, where ε_j assumes the value $+\varepsilon$ or $-\varepsilon$ randomly. We shall refer to this as the generalized Lévy walk (GLW).

Allegrini *et al.* [11,19] have used a model that they called copying mistake map (CMM). The CMM assumes that the DNA sequence results from the randomly joint action of two different prescriptions, one responsible for the long-range correlations and the other of an uncorrelated random nature. The probability of constructing the sequence with the correlations-generating prescription is p_c and the probability of constructing the sequence with the random law is $1-p_c$. The equivalence of the CMM and GLW is made evident by noticing that if the CMM is adopted the probabilities of going forward and backward are

$$P_+=\frac{1\pm p_c}{2}, \tag{29}$$

$$P_-=\frac{1\mp p_c}{2}, \tag{30}$$

respectively, thereby implying that ϵ is identified with p_c.

It is straightforward to show [21] that

$$\Phi_\xi(t)=(1-p_c^2)\Phi_{\tilde{\xi}}(t)+p_c^2\Phi_\epsilon(t), \tag{31}$$

where $\Phi_{\tilde{\xi}}(t)$ is the correlation function of the fast fluctuations (a δ function) and $\Phi_\epsilon(t)$ is the long-range correlation of the bias $\epsilon(t)$. It is equally straightforward [11] to show that Eq. (31) yields an expression of the second-moment time evolution, which is the superposition of a term linear in time, corresponding to the prediction of ordinary Brownian motion, and of a term faster than linear, corresponding to the anomalous diffusion. This means that the short-time dynamics is dominated by ordinary diffusion, while the long-time dynamics, if the DNA sequence is sufficiently large as to make this observation possible, is dominated by anomalous

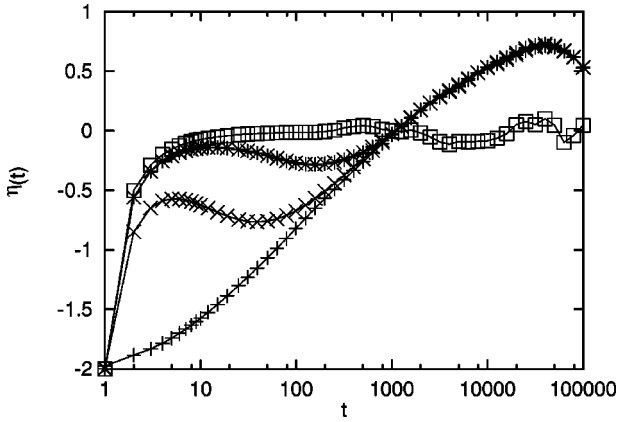

FIG. 3. Function $\eta(t)$ for different realizations of the CMM, with different values of p_c: $p_c=1$ (+), $p_c=0.5$ (×), $p_c=0.2$ (*), and $p_c=0.002$ (□). Here $T=10^6$. The parameters A and β are the same as those of [11]. This means that $\beta=0.67$. A can be derived from the numerical approach to the correlation function of Eq. (9) and turns out to be ~ 10.

diffusion [11]. In other words, this model results in diffusion processes that are indistinguishable from standard Brownian motion at short times and are expected to exhibit anomalous diffusion at long times.

What about the time evolution of the kurtosis $\eta(t)$ in this case? Figure 3 gives a satisfactory answer to this question. We see, in fact, that when p_c is so large as to have a DNA sequence dominated by the prescription with long-range correlation, $\eta(t)$ steadily increases from $\eta(1)=-2$ towards the value 1. However, before reaching this maximum non-Gaussian value, the finite-sequence effect emerges under the form of a regression to the Gaussian statistics. At smaller values of p_c, another interesting effect appears. The short-time increase of $\eta(t)$ becomes much faster, a sort of temporary Gaussian plateau is reached, and then the kurtosis recovers the same behavior as that corresponding to higher values of p_c.

The qualitative explanation of this seemingly complex behavior is straightforward: At short times the time evolution of the kurtosis is dominated by ordinary Brownian diffusion. Consequently, a fast transition to the Gaussian level takes place. Upon further passage of time, as noticed earlier, the effect of long-range correlations becomes predominant and the kurtosis leaves the Gaussian plateau and ''tries'' to reach the plateau corresponding to truncated Lévy statistics. This is prevented from occurring by the fact that the sequence is finite and in fact, at later times, a regression to a Gaussian MACROS takes place. This behavior leaves a signature (a bump on the Gaussian level) of the kurtosis evolution curve. Of course, in the limiting case $p_c=0$, only ordinary Brownian diffusion would be present and no departure from Gaussian statistics would take place.

In other words, at short times the non-Gaussian properties are overcome by the uncorrelated fluctuations and at long times by the finite-sequence effects. The bump represents a different kind of non-Gaussian window [24], through which one can perceive the statistics that would show up in the ideal case of no perturbation.

What about real DNA sequences of prokaryotes? The answer to this question is given by Fig. 4, which shows two

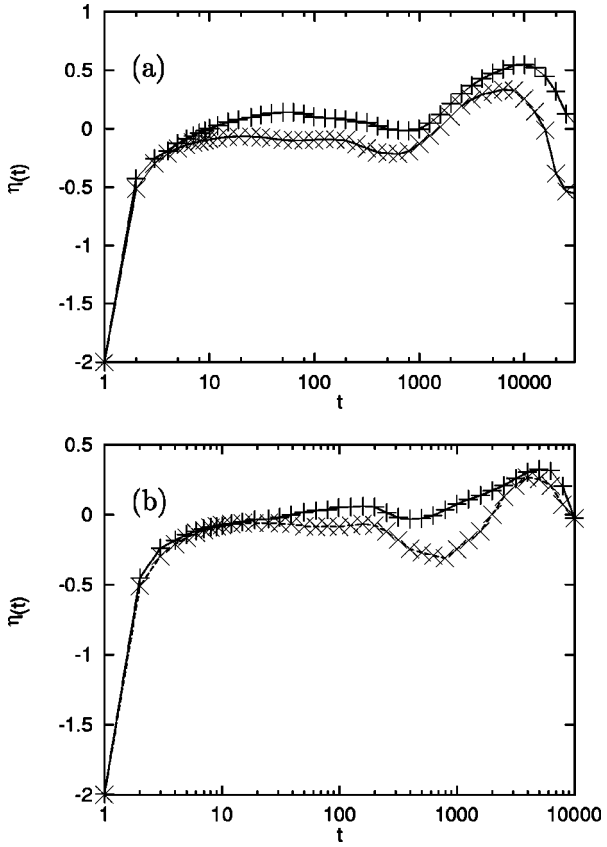

FIG. 4. Time evolution of the non-Gaussian indicator $\eta(t)$. (a) The complete genome of the human Cytomegalovirus (HEHCMVCG) (+) and the CMM sequence (×). The two sequences share the length $T=229\,354$. The length T of the DNA sequence is the number of nucleotides in each of the two strands [base pairs (bps)]. The CMM sequence is given the parameter $p_c=1/9$, which is the same value as that adopted in Ref. [11] to fit different statistical properties of the same DNA sequence. (b) Segment of Escherichia Coli (ECOTSF) (+) and the CMM sequence (×). The two sequences share the length $T=91\,430$ bps. The CMM sequence is given the parameter $p_c=1/10$.

examples of real DNA sequences. The striking discovery is that of an impressive similarity between the non-Gaussian window of the real DNA sequences and that of the artificial DNA sequences generated by the CMM. The agreement is good at both the qualitative and quantitative levels. On the basis of the theoretical discussion of Sec. II, we can also express this result in a slightly different form. The debate on the existence or nonexistence of long-range correlations in DNA sequences is obscured in part by the fact that the non-Gaussian statistics of the diffusion process resulting from the long-range correlations is blurred by the truncation of the sequences. As we have seen, if the width of the sausage is too large, a large portion of its constituents might not result in a ''plateau plus bump'' signature. The fact that the majority of the DNA sequences of prokaryotes examined by us (the longest ones available) results in this effect suggests that the DNA sequences of prokaryotes can be regarded as the single constituents of an ''ideal sausage'' whose width is small enough to force all its constituents to exhibit the same plateau plus bump signature.

V. STATISTICAL, PHYSICAL, AND BIOLOGICAL MEANING AND IMPORTANCE OF THE RESULTS OBTAINED

This paper yields some interesting additions to the recent discoveries on the long-range correlations in DNA sequences by solving a problem of interest for diffusion processes in condensed matter [3]. Let us see why. First of all, we note that according to Ref. [3], a close connection exists between memory and non-Gaussian statistics at the macroscopic level. We know from [3] that if a finite microscopic time scale exists, the non-Gaussian effects at the macroscopic level can only be temporary and Gaussian statistics are recovered when the stationary regime is reached. We can probably persuade the reader, who may not have the time to go through the theoretical arguments of Ref. [3], to accept the golden rule of Eq. (2) by remarking that this is nothing but another manifestation of the celebrated central limit theorem [25].

This perspective is challenged by the phenomenon of superdiffusion, namely, a diffusion faster than ordinary Brownian diffusion, invoked to account for the statistical properties of the DNA sequences with long-range correlations [10,11,14–22]. This is so because a diffusion process where the second moment of the diffusing variable increases in time more quickly than in the case of ordinary Brownian motion (see, for example, Ref. [4]) implies a fluctuation with an infinite correlation time. According to the perspective of Ref. [3], the resulting non-Gaussian effect should be infinite as well. This paper establishes in Sec. II that it is precisely so. This is so much so that the conventional Rahman non-Gaussian indicator, applied to the diffusion process generated by a dichotomous fluctuation with the correlation function of Eq. (9), would diverge in the time asymptotic limit. For this reason we have been forced to adopt the non-Gaussian indicator of Eq. (4).

Not only do we establish that the non-Gaussianicity becomes infinite, we also discover how this unusual condition is reached in time. This is given by the analytical prediction of Eq. (20), which is qualitatively corroborated by numerical results. We also explain why there are significant quantitative deviations of the numerical results from the theoretical prediction.

All this has to be regarded as the first important result of this paper. It has been made possible by the adoption of the theory of Ref. [4], which in turn addresses the important issue of how to derive Lévy statistics, a diffusion process with infinite moments, from within a treatment based on the numerical determination of moments. The wise use of the information provided by the Lévy statistics is made possible by the fact that the process under observation is essentially a truncated Lévy process.

The difficulty for the reader to overcome to appreciate the significance of the second result of this paper is that of looking at the DNA sequences as dichotomous time fluctuation, according to the perspective established by Ref. [15]. The position of a site along the DNA has to be identified with time and the value assigned to a given site (either -1 or $+1$, according to whether the site is occupied by a purine or a pyrimidine) is the value of the fluctuation at that time.

Once this correspondence is established and the due reference to the literature establishing that this is a superdiffusional process [10,11,14–22] is made, the reader can easily understand that the natural question is raised of whether or not this diffusion process is also non-Gaussian. Unfortunately, even if there is a widely accepted conviction that the study of DNA sequences is equivalent to that of anomalous diffusion processes of condensed matter, a link between this subject and that of Ref. [3], it is well known that these sequences are finite. Furthermore, any DNA sequence is a single trajectory and no recourse can be made to the concept of Gibbs ensemble. It is possible to derive a sort of effective Gibbs ensemble from a DNA sequence, conceived as a single trajectory, by using different sites as departure points of a set of new trajectories. This way of generating a Gibbs ensemble, however, produces results that are strongly influenced by the finite length of the DNA sequence if the DNA sequence is characterized by an infinite correlation ''time.'' This is the main reason why the non-Gaussian indicator cannot reach its maximum value and a sort of regression to Gaussian statistics is produced.

This is the second important result of this paper. To convince the reader that it is really important, we have to establish to what extent the detection of this maximum as a method of statistical analysis of a single and finite sequence is reliable. To do that we generated 1000 independent sequences by means of the stochastic generator that, according to the earlier work [4], results in superdiffusion. It is a plausible conjecture to imagine the kurtosis time evolution to be a fluctuating function of the finite sequence considered. This raises a legitimate doubt on the general validity of the second result of this paper, expressed by the following question: Do these fluctuations allow a single and finite sequence to produce a kurtosis time evolution without the non-Gaussian bump? It is evident that an affirmative answer to this question would make questionable the adoption of our method of stastistical analysis. The second result of this paper is therefore made really important by the discovery, made in Sec. III, that each and every sequence, with a sufficiently large length, must be characterized by a non-Gaussian window. This is proved by Figs. 2(b) and 2(c), showing that the whole ''sausage,'' and so all the single trajectories contributing to the sausage width, distinctly departs from the Gaussian plateau within a finite time interval. This means that each and every single sequence, produced by the same stochastic generator, with the same length as those studied in Sec. III (the DNA sequences examined in Sec. IV have the same length) are expected to result in a non-Gaussian window if their correlation length is infinite.

Finally, Sec. IV illustrates the third result of this paper. This last result, as far as the project of a technique of statistical analysis of DNA sequences is concerned, is the main result of this paper. Adopting as an artificial DNA sequence that generated by the theoretical model developed by us in our earlier work [11,19], the CMM model, we make about the time evolution of the non-Gaussian indicator the following predictions. The non-Gaussian indicator is expected to increase very quickly from the initial value of -2 to the vanishing value of the Gaussian plateau. We also expect that after a temporary stay in this Gaussian plateau a transition to the bump regime occurs. This means a further increase of the non-Gaussian indicator with a regression to a vanishing

value that prevents the non-Gaussian indicator from reaching the top value of 1. All these expectations have been satisfactorily confirmed by our statistical analysis of real DNA sequences of prokaryotes, with a quantitative as well as a qualitative agreement between the theoretical predictions and the results of the analysis of real data. Therefore, we conclude this paper by pointing out that the original conviction [11] that the CMM is a good model for prokaryotes is confirmed beyond the original expectation. To reach this important result we had to solve many other subsidiary problems, each of which is of interest by itself, concerning the generalization of the Rahman's pioneer work [1] to the case of infinite memory.

ACKNOWLEDGMENTS

P.A. thanks the INFM for partial support of this work and B.J.W. thanks the Office of Naval Research. We thank Dr. Bruno Zambon for useful comments on an earlier version of this manuscript. We are also grateful to Professor Marcello Buiatti for his help and guidance throughout the biological aspects of this research work.

[1] A. Rahman, Phys. Rev. A **136**, 405 (1964).
[2] H. Miyagawa and Y. Hiwatari, Phys. Rev. A **40**, 6007 (1989).
[3] M. Ferrario, P. Grigolini, A. Tani, R. Vallauri, and B. Zambon, Adv. Chem. Phys. **62**, 225 (1985).
[4] P. Allegrini, P. Grigolini, and B. J. West, Phys. Rev. E **54**, 4760 (1996).
[5] E. W. Montroll and B. J. West, in *Fluctuation Phenomena*, 2nd ed., edited by E. W. Montroll and J. L. Lebowitz, Studies in Statistical Mechanics Vol. 7 (North-Holland, Amsterdam, 1987).
[6] T. Geisel, J. Heldstab, and H. Thomas, Z. Phys. B **55**, 165 (1984).
[7] J. Klafter and G. Zumofen, Physica A **196**, 102 (1993).
[8] G. Zumofen and J. Klafter, Physica D **69**, 436 (1993).
[9] Note that this happens when $\beta>1$, a condition yielding $\tau \propto A^{1/\beta}$.
[10] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993).
[11] P. Allegrini, M. Barbi, P. Grigolini and B. J. West, Phys. Rev. A **52**, 5281 (1995).
[12] E. Floriani, R. Mannella, and P. Grigolini, Phys. Rev. A **52**, 5910 (1995).
[13] R. Bettin, R. Mannella, B. J. West, and P. G. Grigolini, Phys. Rev. E **51**, 212 (1995).
[14] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[15] C. K. Peng, S. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[16] 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).
[17] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[18] R. Voss, Fractals **2**, 1 (1994).
[19] P. Allegrini, P. Grigolini, and B. J. West, Phys. Lett. A **211**, 217 (1996).
[20] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).
[21] P. Allegrini, M. Buiatti, P. Grigolini, and B. J. West, Phys. Rev. E **57**, 4558 (1998).
[22] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).
[23] M. Araujio, S. Havlin, G. H. Weiss, and H. E. Stanley, Phys. Rev. A **43**, 5240 (1991).
[24] Note that there is a subtle difference between this kind of non-Gaussian window and that discussed in Sec. III. In Sec. III we defined the non-Gaussian window as the interval of time between the uncertainty sausage overshooting the condition $\langle \eta(t) \rangle = 0$ and the approximate time at which the uncertainty sausage starts steadily including this Gaussian condition. Here the non-Gaussian window is a property of the single trajectory, rather than a property of the average $\langle \eta(t) \rangle$, as enforced by the nature of the DNA sequences. The non-Gaussian window could have been determined in the same way as in Sec. III only in the case of the computer generated DNA sequences of Figs. 3 and 4. However, this would have required an exceedingly long computational time. The computer time necessary to evaluate the uncertainty sausage depends essentially on the position of the maximum of the function $\langle \eta(t) \rangle$. The larger the time at which this function gets its maximum, the larger the corresponding computational time. On the other hand, the time position of this maximum is determined by the value of the parameter A of Eq. (9). As regards the results of cases (a) and (b) of Fig. 2, the time position of the maximum, before $t = 200$ and $t = 500$, respectively, was determined by adopting the value $A \approx 0.025$. This choice turned out to be compatible with a reasonably short computational time, thereby making it possible to produce the uncertainty sausage of Fig. 2. The CMM sequences resulting in the non-Gaussian indicator illustrated in Figs. 3 and 4 rest on larger values of T and A. This choice generates maxima between $t = 1000$ and $t = 100\,000$, namely, conditions comparable to those of the real DNA sequences of Fig. 4. Thus this choice of A and T made it impossible to evaluate the uncertainty sausage within a reasonably short computational time. In conclusion, to save computational time, we have examined only a single sequence also in the case when this can be computer generated. As a consequence, we do not know whether or not the real DNA sequences examined in Fig. 4 produce a non-Gaussian indicator $\eta(t)$ lying within the uncertainty sausage. Further computational work should be done to establish this important property.
[25] S.-K. Ma, *Statistical Mechanics* (World Scientific, Philadelphia, 1985), p. 198.