Binary *N*-Step Markov Chains and Long-Range Correlated Systems

O. V. Usatenko and V. A. Yampol'skii

A. Ya. Usikov Institute for Radiophysics and Electronics, Ukrainian Academy of Science, 12 Proskura Street, 61085 Kharkov, Ukraine

(Received 14 November 2002; published 20 March 2003)

A theory of systems with long-range correlations based on the consideration of *binary N-step Markov chains* is developed. In our model, the conditional probability that the ith symbol in the chain equals zero (or unity) is a linear function of the number of unities among the preceding N symbols. The correlation and distribution functions as well as the variance of number of symbols in the words of arbitrary length L are obtained analytically and numerically. If the persistent correlations are not extremely strong, the variance is shown to be nonlinearly dependent on L. A self-similarity of the studied stochastic process is revealed. The applicability of the developed theory to the coarse-grained written and DNA texts is discussed.

DOI: 10.1103/PhysRevLett.90.110601 PACS numbers: 05.40.–a, 02.50.Ga, 87.10.+e

The problem of the systems with long-range spatial and/or temporal correlations (LRCs) is one of the topics of intensive research in modern physics. As a rule, the LRC systems are characterized by a complex structure and contain a number of hierarchic objects as their subsystems. The LRC systems are the subject of study in physics, biology, economics, linguistics, sociology, geography, psychology, and others [1–4]. At the present time, there is not a generally accepted theoretical model that describes well the dynamical and statistical properties of the LRC systems. Attempts to describe the behavior of the LRCs in the framework of the Tsalis nonextensive thermodynamics [5,6] were undertaken in Ref. [7]. However, the nonextensive thermodynamics is not well grounded, and construction of the additional models could clarify the properties of the nonextensive systems.

One of the efficient methods used to investigate the correlated systems is based on a decomposition of the space of states into a finite number of parts labeled by definite symbols. This procedure referred to as coarse graining is accompanied by the loss of short-range memory between symbols but does not affect and does not damage the robust invariant statistical properties of the long-range correlated systems. In terms of the power spectrum, one loses only the short-wave part of the spectrum. The most frequently used method of the decomposition is based on the introduction of two parts of the phase space. In other words, it consists in mapping the two parts of states onto two symbols, say, 0 and 1. Thus, the problem is reduced to the investigation into the statistical properties of the binary sequences.

One of the ways to get a correct insight into the nature of correlations in a system consists in an ability of constructing a mathematical object (for example, a correlated sequence of symbols) possessing the same statistical properties as the initial system. There are many algorithms to generate long-range correlated sequences: the inverse Fourier transformation [8], the expansion-modification Li method [9], the Voss procedure of consequent random addition [10], the correlated Levy walks [11], etc. [8]. We believe that among the above-mentioned methods, using the Markov chains is one of the most important.

In this Letter, the statistical properties of the *binary N-step Markov chain* is examined. In spite of the long-time history of studying the Markov sequences (see, for example, Refs. [4,12,13]), the concrete expressions for the variance of sums of random variables in such strings have not yet been obtained. Our model operates with two parameters governing the conditional probability of the discrete Markov process, specifically with the memory length N and the correlation parameter μ. The correlation and distribution functions as well as the variance D being nonlinearly dependent on the length L of a word are derived analytically and numerically. The nonlinearity of the $D(L)$ function reflects the strong correlations in the system. The developed theory is applied to the coarse-grained written texts and to the DNA strings.

Let us consider a *stationary binary sequence* of symbols, $a_i = \{0, 1\}$. To determine the *N-step Markov chain* we have to introduce the conditional probability,

$$P(a_i = 0 \mid a_{i-N}, a_{i-N+1}, \ldots, a_{i-1}), \tag{1}$$

of following the definite symbol (for example, zero) after symbols $a_{i-N}, a_{i-N+1}, \ldots, a_{i-1}$. We choose the simplest model for the P function where the conditional probability p_k of the symbol zero occurring after an N word containing k unities depends on k only and is independent of the arrangement of previous symbols,

$$p_k = \frac{1}{2} + \mu\left(1 - \frac{2k}{N}\right). \tag{2}$$

The parameter μ defines the strength of correlations in the system. We focus our attention on the region of μ determined by the persistence inequality $0 \leq \mu < 1/2$.

Nevertheless, the major part of our results is valid for the antipersistent region $-1/2 < \mu < 0$ as well.

It is easy to see that relation (2) provides the statistical equality of the numbers of symbols zero and unity in the Markov chain. In other words, the chain is nonbiased.

In order to investigate the statistical properties of the Markov chain, we consider the distribution $W_L(k)$ of the words of definite length L by the number k of unities in them, $k_i(L) = \sum_{l=1}^{L} a_{i+l}$, and the variance $D(L) = \overline{k^2} - \overline{k}^2$. Here $f(k) = \sum_{k=0}^{L} f(k) W_L(k)$. If $\mu = 0$, one arrives at the known result for the noncorrelated Brownian diffusion, $D(L) = L/4$. At $\mu \neq 0$, the problem also allows an exact analytical solution. However, we do not describe the procedure of finding the distribution function $W_L(k)$ and present the final result alone. Readers can find all necessary details of the analytical treatment in Ref. [14].

The form of the distribution function depends strongly on the relationship between the parameters of the problem, N and μ. Under the conditions of not very strong persistence,

$$m = \frac{2\mu}{N(1-2\mu)} \ll 1, \tag{3}$$

and at large values of k, $L - k$, the function $W_L(k)$ is the Gaussian,

$$W_L(k) = \frac{1}{\sqrt{2\pi D(L)}} \exp\left\{ -\frac{(k - L/2)^2}{2D(L)} \right\}. \tag{4}$$

The distribution is centered at point $k = L/2$, which corresponds to the equal numbers of zeros and unities in the L word. The variance $D(L)$ is nonlinearly dependent upon L,

$$D(L) = \frac{L}{4}[1 + mF(L)], \tag{5}$$

with $F(L) = L$ at $L < N$ and

$$F(L) = 2(1+\gamma)N - (1+2\gamma)\frac{N^2}{L} - 2\gamma^2 \frac{N^2}{L}\left[1 - \exp\left(-\frac{L-N}{\gamma N}\right)\right] \tag{6}$$

at $L > N$. The parameter γ is determined by the equation,

$$\gamma\left[\exp\left(\frac{1}{\gamma}\right) - 1\right] = \frac{1}{2\mu}, \tag{7}$$

which has a unique solution $\gamma(\mu)$ for any value of $\mu \in (0, 1/2)$.

In the opposite limiting case where the parameter μ is very close to $1/2$, $m \gg 1$, the distribution changes its shape radically. For example, at $L = N$, instead of the bell-shaped Gaussian form, $W_N(k)$ becomes concave,

$$W_N(k) = W_N(0)\frac{2N}{mk(N-k)}, \qquad k,\ N-k \neq 0,\ N. \tag{8}$$

Point $k = N/2$ corresponds to the minimum of $W_N(k) = 2W_N(0)/mN \ll W_N(0)$. Below we deal with the most important case $m \ll 1$.

The plot of function $D(L)$ (5) for $N = 100$ and $\mu = 0.4$ is shown by the solid line in Fig. 1. For comparison, the straight line in this figure corresponds to the dependence $D(L) = L/4$ for the usual Brownian diffusion without correlations (for $\mu = 0$). It is clearly seen that the plot of variance (5) contains two qualitatively different portions. One of them, at $L \sim N$, is the superlinear curve that moves away from the line $D = L/4$ with an increase of L as a result of the persistence. The deviation of the $D(L)$ plot from the line $D = L/4$ continues even at $L > N$. The reason is the phenomenon which we refer to as *macropersistence*. The substance of this phenomenon is that the sizes of the fluctuation regions of the increased concentrations of the symbols zeros (or unities) in the chain are noticeably higher than the memory length N because of the specific retardation effect of the memory. The conditional probability (2) provides the expansion of the fluctuation region on the scale of the order of the correlation length $l_c = (1+\gamma)N$ with γ given by Eq. (7).

For $L \gg l_c$, the plot $D(L)$ achieves the linear asymptotics,

$$D(L) \cong \frac{L}{4}[1 + 2(1+\gamma)mN]. \tag{9}$$

The explanation of this phenomenon is as follows. The length L can be interpreted as the number of jumps of some particle over an integer-valued 1D lattice or as the time of the diffusion imposed by the Markov chain. For $L \gg l_c$, the wandering process is realized by using practically *independent* steps $\sim D^{1/2}(l_c)$ being traversed during the characteristic "fluctuating time" $\Delta L \sim l_c$.

We have supported the obtained analytical results by numerical simulations of the Markov chain with the conditional probability Eq. (2). The circles in Fig. 1

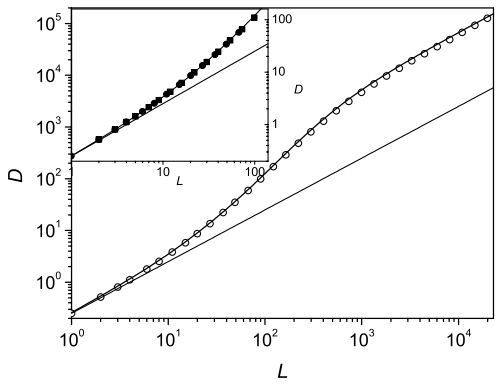

FIG. 1. The numerical simulation of the dependence $D(L)$ for the generated sequence with $N = 100$ and $\mu = 0.4$ (circles). The correspondent value of the parameter γ is 2.32. The solid line is the plot of function Eq. (7) with the same values of N and μ. The thin solid line describes the noncorrelated Brownian diffusion, $D(L) = L/4$. Squares and circles in the inset correspond to the stochastic and deterministic reduction, respectively (the parameter of decimation $\lambda = 0.5$).

represent the calculated variance $D(L)$ for the 100-step Markov chain generated at $\mu = 0.4$. A very good agreement between the analytical result Eq. (5) and the numerical simulation can be observed.

We close the discussion of the obtained theoretical results by pointing out one of the most important properties of the Markov chain, namely, its self-similarity. Let us reduce the N-step Markov sequence by regularly (or randomly) removing some symbols and introduce the decimation parameter $\lambda = N^*/N \leq 1$. Here N^* is a renormalized memory length for the reduced N^*-step Markov chain. It can be proved that the conditional probability p_k^* of the symbol zero occurring after k unities among the preceding N^* symbols is described by formula (2), where the persistence parameter μ and the memory length N should be replaced by

$$\mu^* = \mu \frac{\lambda}{1 - 2\mu(1 - \lambda)} \tag{10}$$

and N^*, respectively. This means that the reduced chain possesses the same statistical properties as the initial one but is characterized by the renormalized parameters (N^*, μ^*) instead of (N, μ). The astonishing property of the Markov chain is that *the variance $D^*(L)$ is invariant with respect to the decimation transformation* $(N, \mu) \rightarrow (N^*, \mu^*)$. Therefore, it coincides with the function $D(L)$ for the initial Markov chain:

$$D^*(L) = \frac{L}{4}(1 + m^*L) = \frac{L}{4}(1 + mL) = D(L), \tag{11}$$

i.e., $m = \text{inv}$. The invariance of the function $D(L)$ at $L < N$ is demonstrated by the inset of Fig. 1.

We have applied the developed theory to the natural objects, specifically to the coarse-grained written and DNA texts. It is well known that the statistical properties of the coarse-grained texts written in any language show a remarkable deviation from random sequences [4,15]. In order to check the applicability of the theory of the binary N-step Markov chains to literary texts we resorted to the procedure of coarse graining by the random mapping of all characters of the text onto a binary set of symbols, zeros, and unities. The statistical properties of the coarse-grained texts depend, but not significantly, on the kind of mapping. The study of different written texts has shown that all of them are featured by the pronounced persistent correlations. It is demonstrated by Fig. 2 where five curves $D(L)$ go significantly higher than the straight line $D = L/4$. However, it should be emphasized that regardless of the kind of mapping the initial portions, $L < 80$, of the curves correspond to a slight antipersistent behavior (see the lower inset). Thus, the persistence is the common property of the binary N-step Markov chains and the coarse-grained written texts at large scales. Moreover, the written texts as well as the Markov sequences possess the property of the self-similarity. Indeed, the curves in the upper inset of Fig. 2 obtained from the text

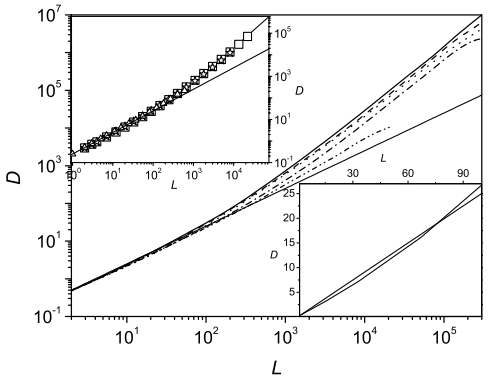

FIG. 2. The $D(L)$ dependence for the coarse-grained texts of a collection of works on computer science ($m = 2.2 \times 10^{-3}$, solid line), the Bible in Russian ($m = 1.9 \times 10^{-3}$, dashed line), the Bible in English ($m = 1.5 \times 10^{-3}$, dotted line), *History of Russians in the 20th Century,* by Oleg Platonov ($m = 6.4 \times 10^{-4}$, dash-dotted line), and A*lice's Adventures in Wonderland,* by Lewis Carroll ($m = 2.7 \times 10^{-4}$, dash-dot-dotted line). In the upper inset: the $D(L)$ dependence for the coarse-grained text of the Bible (solid line) and for decimated sequences with different parameters λ: $\lambda = 3/4$ (squares), $\lambda = 1/2$ (stars), and $\lambda = 1/256$ (triangles). In the lower inset: the antipersistent portion of the $D(L)$ plot for the Bible is magnified.

of the Bible with different levels of decimation demonstrate this property. Presumably, the self-similarity is the mathematical reflection of the well-known hierarchy in the linguistics: *letters* → *syllables* → *words* → *sentences* → *paragraphs* → *chapters* → *books* → *collected works.*

All of the above-mentioned circumstances allow us to suppose that our theory of the binary N-step Markov chains can be applied to the description of the statistical properties of the texts of natural languages. However, in contrast to the generated Markov sequence (see Fig. 1) where the full length $\mathcal{M}$ of the chain is far greater than the memory length N, the coarse-grained texts described by Fig. 2 are of relatively short length $\mathcal{M} < N$. In other words, the coarse-grained texts are similar not to the Markov chains but rather to some nonstationary short fragments. This implies that each of the written texts is correlated throughout the whole of its length. Therefore, for the written texts, it is impossible to observe the second portion of the curve $D(L)$ parallel (in the log-log scale) to the line $D(L) = L/4$, similar to that shown in Fig. 1.

Evidently, this portion can be observed only for collections containing a large number of texts that are not semantically connected to each other. Besides, the $D(L)$ dependence would not correspond to the system correlated throughout its whole length if any integral literary text is cut into a large number of pieces and then mixed in an arbitrary way.

As a result, one cannot define the values of both the parameters N and μ for the integral coarse-grained texts. The analysis of the curves in Fig. 2 can give the combination $m = 2\mu/N(1 - 2\mu)$ only. Perhaps, this particular

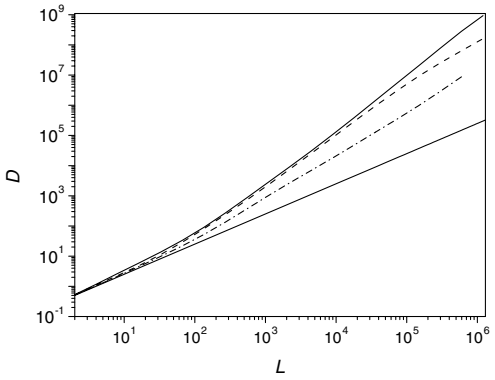

FIG. 3. The $D(L)$ dependence for the coarse-grained DNA text of *Bacillus subtilis, complete genome* for three nonequivalent kinds of the mapping. Solid, dashed, and dash-dotted lines correspond to the mappings $\{A, G\} \to 0$, $\{C, T\} \to 1$ (the parameter $m = 4.1 \times 10^{-2}$), $\{A, T\} \to 0$, $\{C, G\} \to 1$ ($m = 2.5 \times 10^{-2}$), and $\{A, C\} \to 0$, $\{G, T\} \to 1$ ($m = 1.5 \times 10^{-2}$), respectively.

combination is the real parameter governing the persistent properties of the literary texts.

We note that the origin of the long-range correlations in the literary texts is hardly related to the grammatical rules as is claimed in Ref. [4]. At short scales $L \leq 80$ where the grammatical rules are, in fact, applicable, the character of correlations is antipersistent, whereas semantic correlations lead to the global persistent behavior of the variance $D(L)$ throughout the whole of literary text.

It is known that any DNA text is written by four "characters," specifically by adenine (A), cytosine (C), guanine (G), and thymine (T). Therefore, there are three nonequivalent types of the DNA text mapping onto one-dimensional binary sequences of zeros and unities. By way of example, the variance $D(L)$ for the coarse-grained text of *Bacillus subtilis, complete genome* is displayed in Fig. 3 for all possible types of mapping. One can see that the persistent properties of DNA are more pronounced than for the written texts and the $D(L)$ curve does not contain the linear portion given by Eq. (9). This suggests that the DNA chain is not a stationary sequence. In this connection, we point out that *Bacillus subtilis* is a procaryote and the studied DNA text represents the collection of extended coding areas interrupted by small noncoding regions. According to Fig. 3, the noncoding regions do not interrupt the correlation between the coding areas, and the DNA system is fully correlated throughout its whole length. In future works we are going to apply our method to DNA of different types of organisms such as (a) higher eucaryotes (having extended noncoding and short coding areas); (b) procaryotes (having extended coding and short noncoding regions); (c) viruses (almost 100% coding areas); (d) pure coding sequences; and (e) pure noncoding sequences.

Thus, we have developed a new approach for the description of the strongly correlated one-dimensional systems. The simple exactly solvable model of the uniform binary N-step Markov chain is presented. The conditional probability Eq. (1) chosen in the simplest form (2) governs the correlation property of the chain. Usually, the correlation function $\mathcal{K}(r)$ and other moments are employed as the input characteristics for the description of the correlated random systems. Yet, the function $\mathcal{K}(r)$ describes not only the direct interconnection of the elements a_i and a_{i+r}, but also takes into account their indirect interaction via other elements. Our approach operates with the "origin" characteristics of the system, specifically with the conditional probability (1). Therefore, we believe that this way allows us to disclose the intrinsic properties of the system which provide the correlations between the elements. We have demonstrated the applicability of the suggested theory to the different kinds of correlated stochastic systems. However, there are some aspects that cannot be interpreted in terms of our two-parameter model. Obviously, more complex models of the Markov chain should be developed for an adequate description of real correlated systems.

We acknowledge Dr. S. V. Denisov who drew our attention to the exposed problem, Yu. L. Rybalko for consultations and kind assistance in the numerical simulations, and S. S. Mel'nik and M. E. Serbin for the helpful discussions.

[1] H. E. Stanley *et al.*, Physica (Amsterdam) **224A**, 302 (1996).
[2] A. Provata and Y. Almirantis, Physica (Amsterdam) **247A**, 482 (1997).
[3] R. N. Mantegna and H. E. Stanley, Nature (London) **376**, 46 (1995).
[4] I. Kanter and D. F. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).
[5] C. Tsalis, J. Stat. Phys. **52**, 479 (1988).
[6] *Nonextensive Statistical Mechanics and Its Applications*, edited by S. Abe and Yu. Okamoto (Springer, Berlin, 2001).
[7] S. Denisov, Phys. Lett. A **235**, 447 (1997).
[8] A. Czirok, R. N. Mantegna, S. Havlin, and H. E. Stanley, Phys. Rev. E **52**, 446 (1995).
[9] W. Li, Europhys. Lett. **10**, 395 (1989).
[10] R. F. Voss, in *Fundamental Algorithms in Computer Graphics*, edited by R. A. Earnshaw (Springer, Berlin, 1985), p. 805.
[11] M. F. Shlesinger, G. M. Zaslavsky, and J. Klafter, Nature (London) **363**, 31 (1993).
[12] C. V. Nagaev, Theory Probab. Its Appl. **2**, 389 (1957) (in Russian).
[13] M. I. Tribelsky, Phys. Rev. Lett. **87**, 070201 (2002).
[14] O. V. Usatenko and V. A. Yampol'skii, physics/0211053.
[15] A. Schenkel, J. Zhang, and Y. C. Zhang, Fractals **1**, 47 (1993).