PHYSICAL REVIEW E **70**, 015104(R) (2004)

Phase transition in random walks with long-range correlations

Shahar Hod[1,2] and Uri Keshet[2]
[1]*The Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel*
[2]*Department of Condensed Matter Physics, Weizmann Institute, Rehovot 76100, Israel*

(Received 16 April 2004; published 26 July 2004)

Motivated by recent results in the theory of correlated sequences, we analyze the dynamics of random walks with long-term memory (binary chains with long-range correlations). In our model, the probability for a unit bit in a binary string depends on the *fraction* of unities preceding it. We show that the system undergoes a dynamical phase transition from normal diffusion, in which the variance D_L scales as the string's length L, into a superdiffusion phase ($D_L \sim L^{\alpha}, \alpha > 1$), when the correlation strength exceeds a critical value. We demonstrate the generality of our results with respect to alternative models, and discuss their applicability to various data, such as coarse-grained DNA sequences, written texts, and financial data.

DOI: 10.1103/PhysRevE.70.015104 PACS number(s): 02.50.−r, 05.40.Fb, 87.10.+e

Dynamical systems with long-range spatial (and/or temporal) correlations are attracting considerable interest across many disciplines. They are identified in physical, biological, social, and economic sciences (see, e.g., [1–9] and references therein). Of particular interest are situations in which the system can be mapped onto a mathematical object, such as a correlated sequence of symbols, preserving the essential statistical properties of the original system.

One of the methods most frequently used to obtain insight into the nature of correlations in a dynamical system consists of mapping the space of states onto two symbols [5]. Thus, the problem is reduced to the exploration of the statistical properties of correlated binary chains. This can also be viewed as the analysis of a history-dependent random walk. Random walk is one of the most ubiquitous concepts of statistical physics. It lends applications to numerous scientific fields (see, e.g., [10–16] and references therein).

It is well established that the statistical properties of coarse-grained DNA strings and written texts significantly deviate from those of purely random sequences [2,17]. Financial data (such as stock market quotes) are similarly far from being pure diffusive. Moreover, these systems exhibit "superdiffusive" behavior in the sense that the variance $D(L)$ grows asymptotically faster than L (where L is the length of the considered text). Specifically, $D \sim L^{\alpha}$, with $\alpha > 1$ [5]. Such a remarkable (and essentially universal) phenomenon can be attributed to long-range positive correlations. Systems with such correlations may be anticipated to exhibit a dynamical phase transition (from normal to superdiffusive behavior) at some critical correlation strength.

Thus, the problem of random walk where the jumping probabilities are history dependent is of great interest for understanding the behavior of systems with long-range correlations, such as DNA strings, written texts, and financial data. In the present paper we analyze this problem, and provide a simple yet generic analytical description of the statistical properties of these systems.

We begin by solving a simple model which incorporates long-range correlations into an otherwise random sequence. We consider a discrete binary string of symbols, $a_i = \{0,1\}$, in which the conditional probability of a given symbol (say, a unit bit) occurring at the position $L+1$ is history dependent, and given by

$$p(k,L) = \frac{1}{2}\left(1 - \mu \frac{L-2k}{L+L_0}\right), \tag{1}$$

where k is the number of such symbols (unities) appearing in the preceding L bits. The correlation parameter μ, where $-1 < \mu < 1$, determines the strength of correlations in the system. The persistence condition $\mu > 0$ implies that a given symbol in the preceding sequence promotes the birth of a new identical symbol. On the other hand, in the antipersistence region $\mu < 0$, each symbol inhibits the appearance of a new identical symbol. The parameter $L_0 > 0$ is a characteristic transient time. For $L \ll L_0$ the sequence is approximately random (uncorrelated), whereas for $L \gg L_0$ the effect of long-range correlations takes over [18].

In this model, the conditional probability $p(k,L;\mu,L_0)$ depends on the fraction of unities (or zeroes) in the preceding bits, and is independent of their arrangement. This allows one to obtain an analytical description of the system's dynamical behavior. As we shall demonstrate below, this simple model provides a good quantitative description of the observed statistical properties of various natural systems, such as coarse-grained DNA strings, written texts, and financial data.

The probability $P(k,L+1)$ of finding k identical symbols (say, unities) in a sequence of length $L+1$ follows the evolution equation

$$P(k,L+1) = [1 - p(k,L)]P(k,L) + p(k-1,L)P(k-1,L). \tag{2}$$

Crossing to the continuous limit, one obtains the Fokker-Planck diffusion equation for the correlated process,

$$\frac{\partial P}{\partial L} = \frac{1}{2}\frac{\partial^2 P}{\partial x^2} - \frac{\mu}{L+L_0}\frac{\partial (xP)}{\partial x}, \tag{3}$$

where $x \equiv 2k - L$ [19]. The evolution equation (3), along with the initial condition $P(x,L=0) = \delta(x)$, has a solution in the form of a Gaussian distribution

$$P(x,L) = \frac{1}{\sqrt{2\pi D(L)}} \exp\left[-\frac{x^2}{2D(L)}\right], \tag{4}$$

where the variance $D(L)$ is given by

$$D(L;\mu,L_0) = \frac{L+L_0}{1-2\mu}\left[1-\left(\frac{L_0}{L+L_0}\right)^{1-2\mu}\right]. \tag{5}$$

Equation (5) breaks down at the special case $\mu=1/2$, in which case the variance is given by

$$D(L;\mu_c,L_0) = (L+L_0)\ln\left(\frac{L+L_0}{L_0}\right). \tag{6}$$

Remarkably, one finds that the correlated system undergoes a dynamical phase transition at the critical correlation strength $\mu_c \equiv 1/2$. The variance $D(L)$ of the correlated sequence has three qualitatively different asymptotic behaviors (in the $L \gg L_0$ limit)

$$D(L) \simeq \begin{cases} (1-2\mu)^{-1}L & \mu<\mu_c, \\ L\ln(L/L_0) & \mu=\mu_c, \\ (2\mu-1)^{-1}L_{01-2\mu}L^{2\mu} & \mu>\mu_c. \end{cases} \tag{7}$$

Thus, for $\mu<\mu_c$ the asymptotic variance scales linearly with the string length, whereas for a history-dependent chain with strong positive correlations ($\mu>\mu_c$) the system is characterized by a superdiffusion phase, in which case $D(L) \propto L^{\alpha}$, where $\alpha=2\mu(\alpha>1)$ [20].

The analytical model can readily be extended to encompass situations in which the binary sequence is biased. Let

$$p(k,L) = \frac{1}{2}\left(1+q-\mu\frac{L-2k}{L+L_0}\right), \tag{8}$$

with $-1<q<1$. The distribution $P(x,L)$ corresponding to this conditional probability is given by a Gaussian function, centered about the position

$$x_c(L) = \frac{q}{1-\mu\left(\frac{L}{L+L_0}\right)}L. \tag{9}$$

Thus, the drift velocity approaches an asymptotically constant value $q/(1-\mu)$. The variance $D(L)$, unaltered by the bias is given by Eqs. (5) and (6).

In order to confirm the analytical results, we perform numerical simulations of (discrete) binary sequences. Figure 1 displays the resulting scaled variance $L^{-1}D(L)$ of correlated strings with various different values of the correlation parameter μ. We find an excellent agreement between the analytically predicted results [see Eqs. (5) and (6)] and the numerical ones.

Robustness of the linear model. In order to show the generality of the model discussed above, we consider arbitrary (history-dependent) jump probabilities which are odd functions [21] of the fraction $\xi \equiv x/(L+L_0)$ of unities (zeros) that appeared in the previous L symbols

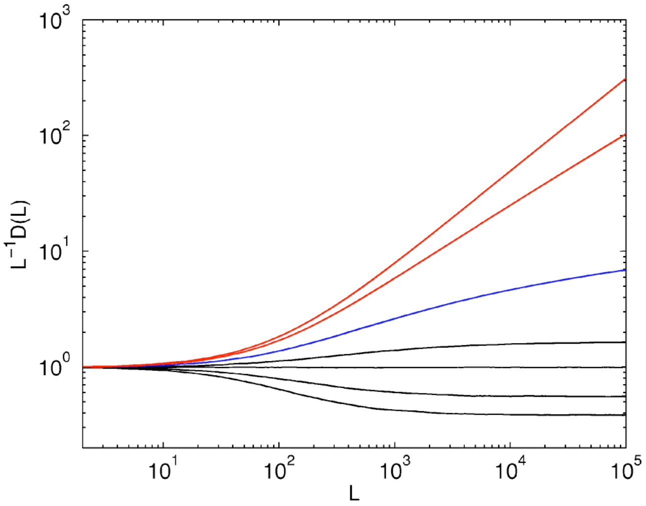

FIG. 1. The scaled variance $L^{-1}D(L)$ as a function of the string length L. We present results for $\mu=-0.8$, -0.4, 0, 0.2, 0.5, 0.8, and 0.9 (from bottom to top), with $L_0=100$. The numerically computed asymptotic slopes agree with the analytical predictions [see Eqs. (5) and (6)] to within less than 1%.

$$p(x,L) = \frac{1}{2}[1+\mu F(\xi)]. \tag{10}$$

For asymptotically large L, one always finds $\xi \to 0$ for nonballistic diffusion, justifying a power-law expansion of $F(\xi)$. As long as this expansion includes a linear term, the original differential equation (3) is recovered for large L. We therefore expect the previous analytical results [Eqs. (5) and (6)] to hold true for generic (nonlinear) models as well. The generality of the model is illustrated in Fig. 2, in which we depict results for various choices of the probability function $F(\xi)$. As predicted, the results are found to agree with the linear model.

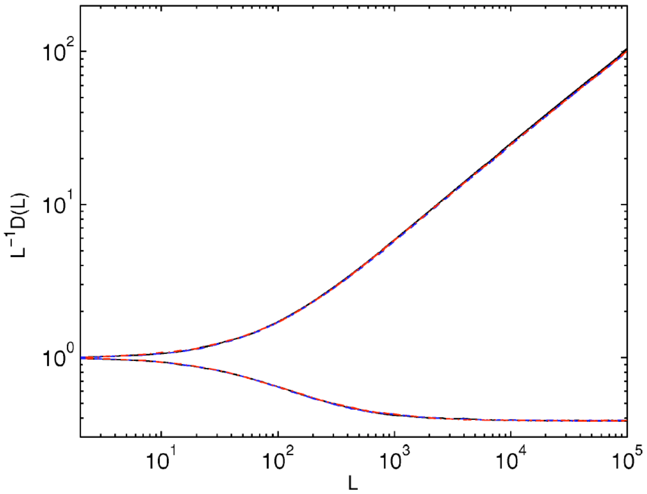

FIG. 2. The scaled variance $L^{-1}D(L)$ for three different forms of the function $F(\xi)$: ξ, $(2/\pi)\sin[(\pi/2)\xi]$, and $\tanh(\xi)$. We present results for $\mu=-0.8$ and $\mu=0.8$, with $L_0=100$. The different curves are almost indistinguishable.

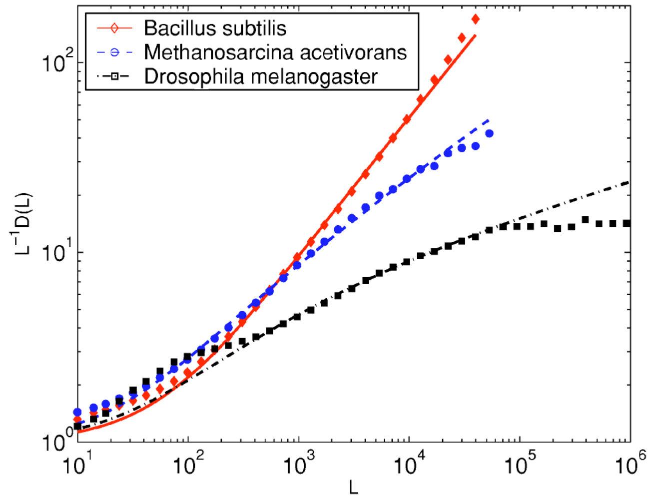

FIG. 3. The scaled variance $L^{-1}D(L)$ as a function of the string length L, for coarse-grained DNA sequences of various organisms. The mapping and parameters used are given in Table I. Theoretical results [see Eq. (5)] are represented by curves.

Applications. The robustness of the linear model (see Fig. 2) suggests that it may capture the essence of the underlying correlations in a diversity of systems in nature. We therefore examine the use of the results derived in the present work as an analytical explanation for the observed statistical properties of natural systems, such as DNA strings, written texts, and financial data [22].

As mentioned, these systems often exhibit a significant deviation from random sequences [2,17], and are characterized by a superdiffusive behavior in which $D \sim L^{\alpha}$, with $\alpha > 1$ [5]. In such systems, superdiffusion may be attributed to long-range (positive) correlations. In fact, the analytical model allows one to determine the correlation strength of these chains.

Figure 3 depicts the scaled variance $L^{-1}D(L)$ calculated from DNA sequences of various organisms, as a function of the string length L. It is of considerable interest to examine in such methods the statistical properties characterizing the DNA of organisms in various evolutionary levels: Bacillus subtilis (*Bacteria*), Methanosarcina acetivorans (*Archaea*), and Drosophila melanogaster (*Eukarya*) [5,23]. The theoretical model provides a good description of the empirical data [24], attributing different correlation strengths μ to different organisms, as summarized in Table I.

The superdiffusive behavior, shown in Fig. 3 to persist across very long sequences, is highly suggestive of long-range correlation extending over more than one gene (e.g., $\sim 5 \times 10^4$ base pairs in Drosophila). This result is in accord with recent studies [25].

TABLE I. The correlation strength parameter μ for various binary strings. We use the following mappings: $\{A,G\} \rightarrow 0$, $\{C,T\} \rightarrow 1$ for DNA sequences [5,23]; $(a \text{ to } m) \rightarrow 0$, $(n \text{ to } z) \rightarrow 1$ for written texts [5]; and daily fall $\rightarrow 0$, daily rise $\rightarrow 1$ for stock market quotes [26].

Data type	String source	μ
DNA sequences	Drosophila melanogaster	0.57
	Methanosarcina acetivorans	0.70
	Bacillus subtilis	0.86
Written texts	Alice's Adventures in Wonderland	0.58
	The Holy Bible in English	0.84
	Works on Computer Science	0.88
Stock markets	NASDAQ	0.39
	DJIA	0.76

In Fig. 4 we show the scaled variance of coarse-grained financial data [daily quotes of the Dow Jones industrial average (DJIA), and the NASDAQ [26]]. We note that the linear model underestimates the empirical variance at short time scales. This fact can be traced back to short-term correlations in the markets. (It is interesting to note that the DJIA maintains an approximately normal diffusive behavior for a period of about one month.) However, this short-term memory is washed out at longer time scales, in which case the analytical model provides a good description of the empirical results, as evident from Fig. 4. The corresponding values of the correlation parameter μ are summarized in Table I. Table I also presents the results of the analytical model applied to various coarse-grained written texts [2,17,5]. The parameter L_0 can be read directly from the figures.

In summary, in this paper we have analyzed the dynamics of random walks with *history-dependent* jump probabilities. Our work was motivated not only by the intrinsic interest in such dynamical processes, but also by the flurry of activity in the field of long-range correlated systems, and by some universal statistical features observed in many different natural systems.

We have broadened the study of binary strings to include long-range correlations, extending throughout the length of the chain. Using a simple and exactly solvable model, we identify a dynamical phase transition, from normal diffusion

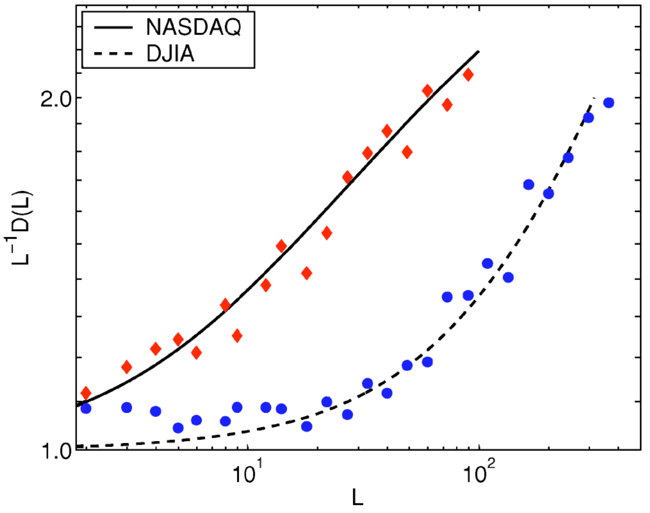

FIG. 4. The scaled variance $L^{-1}D(L)$ as a function of the sequence length L, for coarse-grained financial data: DJIA and NASDAQ daily quotes [26]. The mapping and parameters used are given in Table I. Theoretical results [see Eq. (5)] are represented by curves.

$[D(L) \sim L]$ to superdiffusive behavior $[D(L) \sim L^{2\mu}]$, taking place as the correlation parameter μ exceeds its critical value. We show that in spite of the simplicity of the model, it is robust, and can easily be extended to describe various features (such as a biased history-dependent random walk or subdiffusion).

Next, we have applied the analytical results of the model to various binary strings, extracted from very different natural systems, such as coarse-grained DNA sequences, written texts, and financial data. We find that the model adequately describes the long-term behavior of these systems. Furthermore, the model provides a straightforward method to measure the correlation strength of these systems. Our results can be applied to various natural systems, and may shed light on the underlying rules governing their dynamics. For example, the superdiffusive behavior of DNA sequences (see Fig. 3) suggests long-range correlations extending across more than one gene. The model attributes different correlation strengths to different organisms.

The research of S.H. was supported by the GIF Foundation and by the Dr. Robert G. Picard Fund in Physics. We would like to thank Oded Agam, Yitzhak Pilpel, Eli Keshet, Ilana Keshet, Clovis Hopman, Eros Mariani, Assaf Pe'er, Oded Hod, and Ehud Nakar for helpful discussions. We thank O. V. Usatenko and V. A. Yampol'skii for interesting discussions and for providing us with their data. This research was supported by Grant No. 159/99-3 from the Israel Science Foundation.

---

[1] R. N. Mantegna and H. E. Stanley, Nature (London) **376**, 46 (1995).

[2] I. Kanter and D. F. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[3] H. E. Stanley *et al.*, Physica (Amsterdam) **224**A, 302 (1996).

[4] A. Provata and Y. Almirantis, Physica (Amsterdam) **247**A, 482 (1997).

[5] O. V. Usatenko and V. A. Yampol'skii, Phys. Rev. Lett. **90**, 110601 (2003).

[6] A. C. C. Yang, S. S. Hseu, H. W. Yien, A. L. Goldberger, and C. K. Peng, Phys. Rev. Lett. **90**, 108103 (2003).

[7] C. Tsalis, J. Stat. Phys. **52**, 479 (1988).

[8] *Nonextensive Statistical Mechanics and its Applications*, edited by S. Abe and Yu. Okamoto (Springer, Berlin, 2001).

[9] S. Denisov, Phys. Lett. A **235**, 447 (1997).

[10] M. N. Barber and B. W. Ninham, *Random and Restricted Walks* (Gordon and Breach, New York, 1970).

[11] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992).

[12] R. Fernandez, J. Frohlich, and A. D. Sokal, *Random Walks, Critical Phenomena, and Triviality in Quantum Field Theory* (Springer, Berlin, 1992).

[13] G. H. Weiss, *Aspects and Applications of the Random Walk* (North-Holland, Amsterdam, 1994).

[14] D. ben-Avraham and S. Havlin, *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge University Press, Cambridge, 2000).

[15] R. Dickman and D. ben-Avraham, Phys. Rev. E **64**, 020102(R) (2001).

[16] S. Hod, Phys. Rev. Lett. **90**, 128701 (2003).

[17] A. Schenkel, J. Zhang, and Y. C. Zhang, Fractals **1**, 47 (1993).

[18] The introduction of the parameter L_0 is mainly motivated by the observed behavior of the variance of DNA sequences, written texts, and financial data. These systems are characterized by normal diffusion $[D(L) \sim L]$ for small L values, and by a superdiffusive behavior $[D(L) \sim L^{\alpha}$, with $\alpha > 1]$ for large L values.

[19] We have neglected higher-order terms, but these do not change the results in the $1 \ll L_0 \ll L$ case.

[20] The model may be broadened to describe sub-diffusive behavior as well, by considering the conditional probability $p(k,L) = f(\frac{1}{2}1-\mu[(L-2k)/(L+L_0)^{1-m}])$, where $f(u) \equiv u\Theta(u)-(u-1)\Theta(u-1)$ and $\Theta(u)$ is the Heaviside step function. This yields, for $L \gg L_0$, $m > 0$, and $\mu < 0$, a Gaussian distribution of variance $D(L) \sim L^{1-m}$.

[21] For the probability distribution $P(x,L)$ to be an even function of x (and thus $\langle x \rangle = 0$), the function $F(\xi)$ should be an odd function of its argument.

[22] The results of our model represent ensemble averaging (time averaging would produce different results). Since ensemble averages cannot be obtained for the natural systems of interest, we use their time averages.

[23] DNA sequences of various organisms were obtained from ftp://ftp.ncbi.nih.gov/genomes.

[24] We have verified that for the DNA mapping used ($\{A,G\} \rightarrow 0$, $\{C,T\} \rightarrow 1$), the distribution $P(x, L=\text{const})$ is well approximated by a Gaussian. The alternative mappings yield a broader distribution ($\{T,G\} \rightarrow 0$) or a large asymmetry ($\{C,G\} \rightarrow 0$).

[25] A. Provata and Y. Almirantis, J. Stat. Phys. **106**, 23 (2002);A. Provata and Y. Almirantis, Fractals **8**, 15 (2000);Y. Almirantis and A. Provata, J. Stat. Phys. **97**, 233 (1999).

[26] Financial data for the DJIA and NASDAQ stock markets are quoted from http://finance.yahoo.com.