Long-range correlations in computer diskettes

G. F. Zebende,* P. M. C. de Oliveira,† and T. J. P. Penna‡
Instituto de Física, Universidade Federal Fluminense, Avenida Litorânea s/n, 24210-340, Niterói, Rio de Janeiro, Brazil
(Received 2 May 1997; revised manuscript received 29 October 1997)

We find that successive versions of the files stored on a personal computer diskette mimic the evolution mechanism claimed to be responsible for the long-range correlations observed in DNA sequences. Starting from uncorrelated random files, long-range correlations are gradually introduced by successive editing, corresponding to point mutations, insertions, and deletions. This system has the advantage (over DNA sequences) of allowing experiments. [S1063-651X(98)04403-1]

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

One of the most studied problems in the interface between biology and statistical physics is the existence of long-range power-law correlations in certain DNA sequences [1–3]. A frequently used tool to study long-range correlations in these sequences consists of mapping each sequence onto a one-dimensional random walk. If each step is random and independent of the previous ones, we have an uncorrelated walk. For this sort of walk, the dispersion grows proportionally to the square root of time. On the other hand, if each step depends upon the history, i.e., memory effects are taken into account, we have a correlated walk. If one step depends upon only a few previous steps (short-range correlation) we still have the dispersion increasing as t^{α} with $\alpha=\frac{1}{2}$. Many other studies have appeared in the literature [4–8], showing the existence of power laws in spatial and/or time series of very different dynamical systems, each with its characteristic exponents.

One important question to be answered is why only a small fraction of DNA is used for protein coding, within highly developed organisms. Moreover, long-range correlations have been reported only in cases where noncoding regions are included in the analysis. Recently, Buldyrev *et al.* [9] studied a single gene family to test the hypothesis of growing complexity from lower to higher developed animals. Whereas the coding sequence of this family (myosin heavy chains) does not display long-range correlations for all samples studied, the complete sequence presents long-range power-law correlations, and, moreover, the higher the complexity the higher the characteristic exponent. In this way, long-range correlations found in intron-containing DNA sequences are interpreted in the literature as a consequence of biological evolution. One possible evolutionary mechanism proposed to explain these findings is the following: (i) first, one has copies of the same gene in different locations [10] along the DNA chain (gene duplication); (ii) after a suitably long evolution time, mutations can cause these copies to become distinct (but somewhat similar to each other), one preserving the original function, the others acquiring a new function.

Here we present an alternative dynamical system (successive editions of files in a floppy disk) that seems to behave just like DNA sequences. This system can be of interest because the very interesting subject (evolutionary route of DNA sequences) cannot be easily studied through experiments. The evolution history cannot be rewritten or modified in order to understand the rules and mechanisms leading to the biological world as it is nowadays. As we will show in this work, file editions in a floppy disk also give rise to long-range correlations and the corresponding exponent evolves similarly as the one for DNA sequence fluctuations [9]. The expansion-modification model, after Li [11], is closely related to this work. It provides an alternative explanation for the appearance of long-range correlations in DNA. For computer diskettes, we also found a dependence of the characteristic exponent α with the number of editions. The dynamical similarity between successive editions and storage of files in the diskette and the evolutionary mechanism for DNA sequences is that both processes correspond to insertions, deletions, and ''point mutations.'' Here we consider the edition of a file as a modification that can either randomly change a fraction of bits (''point mutations'') or include (or delete) other entire sequences of bits, therefore changing the file size (insertions and deletions).

We worked with old double-sided, double-density $5\frac{1}{4}$-in. floppy disks with 360 Kbytes of storage capacity (1 Kbyte corresponds to 8192 bits). In order to avoid correlations introduced by operational system features other than the dynamical process itself, each diskette was previously formatted by writing random bits. In our diskettes, we have eight files, occupying half the disk at the beginning. Each bit of the files is also chosen at random. In summary, we have at the beginning one floppy disk with 360 Kbytes of random bits, from which 180 Kbytes are reserved to the eight files.

We choose, at random, which one from the three possible processes will take place: insertion, deletion or point mutation. Insertions (deletions) correspond to increasing (decreasing) the file size by a random fraction *up to* 50% of its present size by introducing (removing) continuously random bits at a given starting position, also chosen at random. The storage strategy of the operational system (MS-DOS) to write files on the disk can be summarized as follows: when the file size decreases, the edited file is written at the same starting position and the free space at the end of the file becomes available for new, future texts. However, the old

*Electronic address: zebende@if.uff.br
†Electronic address: pmco@if.uff.br
‡Electronic address: tjpp@if.uff.br

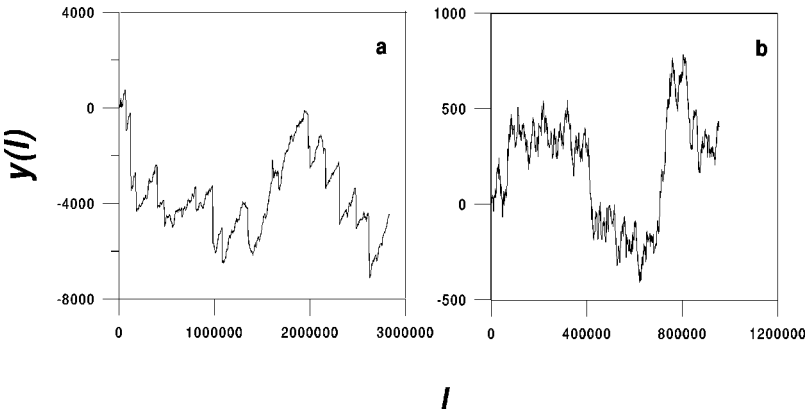

FIG. 1. Landscape for the ''bit-walk.'' We plot the walker position after l steps $y(l)$ as a function of l for (a) reading all the diskette, and (b) reading only the pieces used by files (coding sequences).

information remains stored in this (now) free space, in spite of no longer belonging to any current file. If the file size increases but the next part of the disk is already occupied by another file, the rest of the increased file is written at the next *available* position along the disk. Therefore, the file becomes fragmented into two or more pieces. However, if there is a unoccupied connected region large enough to save the file in the disk, and the file is fragmented into more than two pieces, the system will choose to save the file as a continuous sequence of bits. Point mutations correspond to selecting, at random, 10% of the bits of a given file, and changing them into other bits also selected at random.

In order to edit a file, it is first copied from the disk to the computer memory. There, our program performs random modifications (insertions, deletions, and point mutations) by inserting new sequences at random positions, deleting other parts, etc. After finishing the edition, the file is again stored on the disk, i.e., it is saved according to the MS-DOS storage rules discussed briefly above. In this sense, here we are performing real experiments on real disks, using a real operational system as an example. However, our purpose is to introduce an experimental device in order to test possible mechanisms for DNA evolution which may be proposed. In this case, one must simulate the disk as a long array of bits, modifying its contents according to the proposed evolutionary rules instead of the protocols used by the particular operational system.

In order to build the equivalent to the one-dimensional DNA walk, we consider the diskette as a sequence of bits, 0's and 1's, just as the DNA is considered as a sequence of base pairs: A, G, T, and C. The so-called purine-pyrimidine rule, for instance, considers A or G as one step in a given direction and T or C as one step in the opposite direction, along the DNA walk. Alternatively, the strong (CG) and weak (AT) rule concerning the number of hydrogen bonds (three or two bonds, respectively) binding each pair is the most used rule today. Therefore, analogously to Peng *et al.*'s recipe [1], one can map the disk contents onto a sequence of numbers $u(i)$, where $u(i)=+1$ if a ''1'' bit is found at position i along the disk or $u(i)=-1$ if the ith position is occupied by a bit set to ''0.'' The walker position $y(l)$ after l steps will be given by

$$y(l)\equiv\sum_{i=1}^{l} u(i). \tag{1}$$

In Fig. 1 we present typical curves for $y(l)$ as a function of l, after many editions, corresponding to two different situations: (a) we consider all positions along the diskette, and (b) we consider only the pieces currently used by files (coding sequences), skipping the parts which are not in use. In this test case—and for the others to be presented—files occupy half a disk. These figures are to be compared with the ones presented in Ref. [1].

A quantity of interest for long-range correlations is the root-mean-square fluctuation $F(l)$, defined as

$$F^2(l)\equiv\langle[\Delta y(l)]^2\rangle-\langle[\Delta y(l)]\rangle^2, \tag{2}$$

where $\Delta y(l)\equiv y(l_0+l)-y(l_0)$, and the average is taken over all positions l_0 [12]. For large values of l, we have

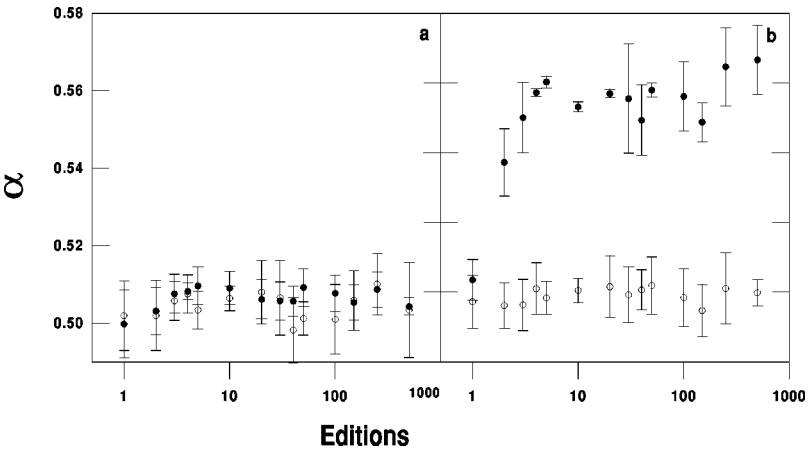

FIG. 2. Fluctuation exponent α as a function of the number of editions **E** (a) *without variation* of the file size for all the disk (●) and only files (○); (b) allowing file size variation. There is no evidence for long-range correlations in (a), whereas in (b) α_{all} deviates from 0.5. These values are taken from averages over ten different diskettes containing eight files each.

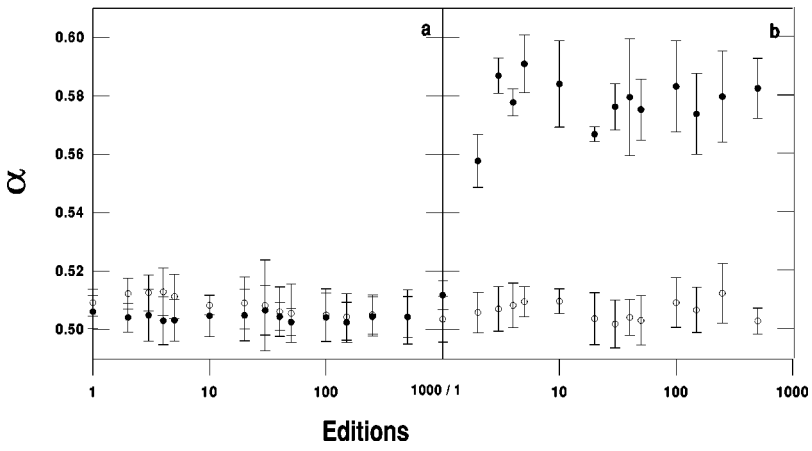

FIG. 3. The same as in Fig. 2, but introducing backup. Note the larger values of α_{all} (○) when compared to Fig. 2.

$$F(l) \sim l^{\alpha}. \tag{3}$$

Through the fluctuation-dissipation theorem we can connect this quantity to the correlation function [12]. If $\alpha = \frac{1}{2}$ we have an uncorrelated—or short range correlated—random walk, for which the correlation function decays exponentially. For $\alpha > \frac{1}{2}$ we have correlated walks (e.g., $1/f$ noise). In the case where $\alpha < \frac{1}{2}$ we have anticorrelated walks, like the cases of healthy hearts [5] and leaky faucets [8].

We investigate how the characteristic exponent α of long-range correlations changes after successive editions considering both situations: reading only the pieces used by files (α_{file}), and reading all the floppy disk (α_{all}, from now on). In our systems, the fluctuations behave like a power law, for at least three decades. These results for α as a function of the number of editions are presented in Fig. 2(a) for the case in which the file sizes are constant, i.e., considering only point mutations, and in Fig. 2(b), where insertions and deletions are considered. As we can see in Fig. 2(a), α_{all} and α_{file} are both close to 0.5, because in this case the new version of the file is always stored exactly on the same place as the old one. Since the mutations are always taken at random, we can expect random bits along the whole diskette, independently of whether they are in use by files or belong to noncoding regions. More interesting results can be seen in Fig. 2(b), where insertions and deletions took place. For the case where only the parts used by files are read, we still have an uncorrelated random walk landscape ($\alpha_{\text{file}} \approx 0.5$). On the other hand, considering all the disk, we found long-range correlations ($\alpha_{\text{all}} > 0.5$). The increasing behavior of α_{all} is analogous to that reported by Buldyrev *et al.* [9], where more highly developed organisms (corresponding to diskettes with more editions) present higher correlation exponents in their DNA sequences. A saturation value of α_{all} is evident and is due to finite-size effects of the disks (this is not supposed to occur in DNA sequences).

Another situation considered here is the edition with backup. Before each edition, we keep a copy of the old version on the diskette as a backup file. The results for backup editions are presented in Fig. 3(a) (without variation of file size), and in Fig. 3(b) (with its variation). These results are qualitatively the same as the preceding ones without backup, but with larger values for the exponent α_{all} and faster saturation. The backup files were not considered for the evaluation of α_{file}.

An additional and independent evidence for the similarity between evolutionary processes in diskettes and DNA sequences is related to the frequency of the jump sizes in the walks. Initially, for each 100 bits read, we considered $P(\Delta y)$ as the probability of finding a jump of size Δy along the diskette. In Fig. 4 we present the histogram of $P(\Delta y)/P(0)$ as a function of Δy, for reading only the parts currently used by files and for all the disk, with size file variation. We have tried to fit the histograms by Lévy distributions defined as

$$P(\Delta y, \psi, \gamma) = \frac{1}{\pi} \int_0^{\infty} \exp(-\gamma q^{\psi}) \cos(q \Delta y) \, dq, \tag{4}$$

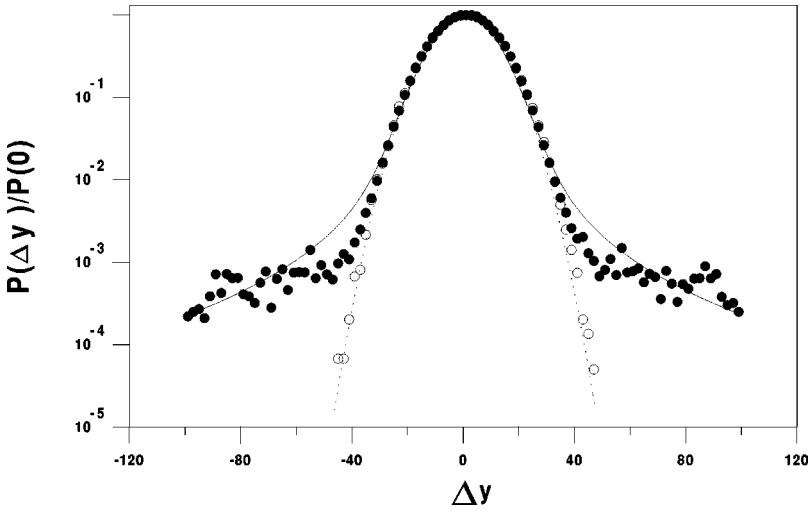

FIG. 4. $\log_{10}[P(\Delta y)/P(0)]$ as a function of Δy. The dotted line corresponds to a fit by a Gaussian distribution ($\psi = 2$) for coding regions. The solid line corresponds to a fit by a Lévy distribution with index $\psi = 1.86$, and fits data for all the diskette. In both cases $\gamma = 37$.

with indexes γ and ψ, and by Gaussian distributions. Lévy distributions are frequently associated to fractional Brownian motion and therefore are related to processes with memory effects (see Refs. [13–15]). The histogram of jumps for the whole diskette is well fitted by a Lévy distribution [16] with $\psi=1.86$, while the histogram of jump sizes corresponding only to parts used by files is well described by a Gaussian distribution ($\psi=2$). It may be argued whether a binomial distribution should be used instead of Lévy and Gaussian distributions. In order to check it, we also considered a different maximum size of jumps (we tried $\Delta y_{\max}=150$, 200, and 300, for which we can expect a better agreement between Gaussian and binomial distributions). We have found similar results. Nevertheless, we know from Ref. [17] that truncated Lévy flights present ultraslow convergence to Gaussian distributions and, therefore, we still cannot confirm the anomalous character of dispersion of jump sizes in this system.

In conclusion, it has been shown that long-range correlations in diskettes are introduced by successive editions, similarly to a previous hypothesis for the evolutionary process in DNA sequences [1]. The exponent α is presented here in two distinct forms of mapping: reading only bits corresponding to files (corresponding to the coding regions of a DNA sequence) and reading all bits of the diskette (corresponding to the complete DNA sequence). We have obtained the correlation exponent α as a function of the number of editions, starting from uncorrelated random files, simulating the precursor sequence which consists entirely of coding parts. The results for the editions in diskettes show that only the case where the file size varies generates long-range correlations. In addition, also in this system, the higher the complexity (given by the number of editions) the higher the fluctuation exponent.

This work was partially supported by Brazilian agencies CAPES, CNPq, FAPERJ, and FINEP. We are indebted to Suzana Moss, Jorge Sá Martins, and Dietrich Stauffer for critical readings of the manuscript. We also thank C. G. Carvalhaes for computational help and discussions.

[1] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992); C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992); H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng, and M. Simons, Physica A **205**, 214 (1994).
[2] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[3] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[4] A. Schenkel, J. Zhang, and Y.-C. Zhang, Fractals **1**, 47 (1993).
[5] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993).
[6] M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, and N. Shnerb, Fractals **2**, 7 (1994).
[7] W. Ebeling and T. Pöschel, Europhys. Lett. **26**, 241 (1994).
[8] T. J. P. Penna, P. M. C. de Oliveira, J. C. Sartorelli, W. M. Gonçalves, and R. D. Pinto, Phys. Rev. E **52**, 2168 (1995).
[9] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2675 (1993).
[10] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).
[11] W. Li, Europhys. Lett. **10**, 395 (1989); Phys. Rev. A **43**, 5240 (1991).
[12] J. Feder, *Fractals* (Plenum, New York, 1988).
[13] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).
[14] B. J. West and W. Deering, Phys. Rep. **246**, 1 (1994).
[15] C. Tsallis, A. M. C. Souza, S. Levy, and R. Maynard, Phys. Rev. Lett. **75**, 3589 (1995).
[16] T. J. P. Penna, Comput. Phys. **9**, 341 (1995).
[17] R. Mantegna and H. E. Stanley, Phys. Rev. Lett. **73**, 2946 (1994).