Noisy Lévy walk analog of two-dimensional DNA walks for chromosomes of S. cerevisiae

Guillermo Abramson,[1,*] Pablo A. Alemany,[1,2] and Hilda A. Cerdeira[1]
[1] International Centre for Theoretical Physics, P.O. Box 586, 34100 Trieste, Italy
[2] Theoretische Polymerphysik, Hermann-Herder-Strasse 3, 79104 Freiburg, Germany

(Received 11 September 1997)

The DNA sequences of all the chromosomes of *Saccharomyces cerevisiae* are mapped onto a $d=2$ space. The resulting patterns are interpreted as a two-dimensional walk. Their mean square displacement shows a superdiffusive behavior. We address the question if this behavior can be understood in terms of a random walk model. We found that it can be modeled as a superposition of a Lévy walk and white noise.
[S1063-651X(98)12407-8]

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

There has been interest in the physics community in studying DNA sequences from a physical or mathematical point of view. The findings of some recent works suggest that the sequence of base pairs or nucleotides in DNA displays power-law correlations [1,2] and several controversial points have been discussed in recent years. Most of these works are based on a one-dimensional mapping of the sequence. Here we study a two-dimensional mapping that will be described immediately below.

In particular, in this work we analyze the two-dimensional mapping of DNA sequences of the organism *Saccharomyces cerevisiae* (*S.c.*) (baker's yeast), the first eukaryote whose complete genome has been sequenced [3]. This provides us with several very large sequences corresponding to the same organism. Such sequences consist of a succession of four symbols: A (adenine), G (guanine), T (thymine) and C (cytosine). Typically, the frequency of A and of T is around 0.30 and that of G and C is about 0.20. The frequency of both $A+G$ (purines) and $C+T$ (pyrimidines) is very close to 0.5. We will envisage these sequences as realizations of a stochastic process. For its analysis, we introduce a mapping into a two-dimensional walk. With each C (G) symbol we associate one step in the positive (negative) direction along the vertical y axis. With each T (A) symbol we associate one step to the right (left) along the horizontal x axis. In this way we obtain a roughly unbiased walk. A typical resulting pattern is displayed by chromosome *II*; see Fig. 1.

Among the many quantities useful to characterize a walk, the mean square displacement (MSD) is one of the most important, as it is closely related to the correlations [4]. For a standard random walk (RW), the MSD $\langle \mathbf{r}^2(s)\rangle$ is a linear function of the number of steps s: $\langle \mathbf{r}^2(s)\rangle = 2dDs$. The proportionality constant $2dD$ (where d is the dimension of the space) defines the diffusion coefficient D. This purely linear behavior is an immediate consequence of the fact that a RW is a sum of identically distributed *independent* random variables: The total displacement after s steps $\mathbf{r}(s)$ is the sum of s single displacements, each equally distributed with variance σ^2. Then, if there is no bias, the MSD corresponds to the variance of the whole displacement after s steps. As the variance of a sum of independent random variables is the sum of their variances, we get a linear dependence on s. Deviations from the pure linearity is characteristic of correlations between steps. These correlations can be so strong or so long ranged that even for the asymptotics $s \to \infty$ a linear regime is never reached. This lack of linearity makes it impossible to define a diffusion coefficient and one speaks of *anomalous diffusion*, in most of the cases characterized by a power law $\langle \mathbf{r}^2(s)\rangle \sim s^{\alpha}$. If $0<\alpha<1$ one speaks of *subdiffusion* and the case $1<\alpha$ is called *superdiffusion* or *enhanced diffusion*. While subdiffusion is typical for transport of charge carriers in disordered media and amorphous materials [5], enhanced diffusion is characteristic in turbulent transport [6,7], chaotic [7–10], polymer [13], and biological [14] systems, and generalized statistical thermodynamics [15].

The MSD of a RW $\langle \mathbf{r}^2(s)\rangle$ is understood as an average over many realizations of the walk, each one performed under the same conditions. On the other hand, for DNA se-

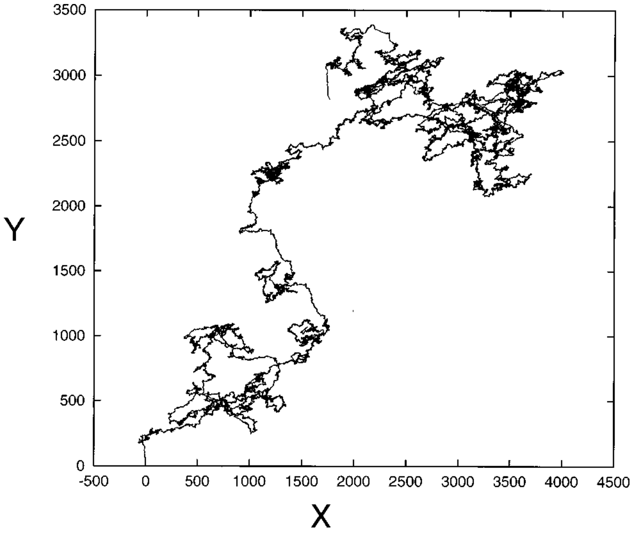

FIG. 1. DNA walk corresponding to the chromosome *II* of *S. cerevisiae*.

[*] Present address: Max-Plank-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 38, 01187 Dresden, Germany.

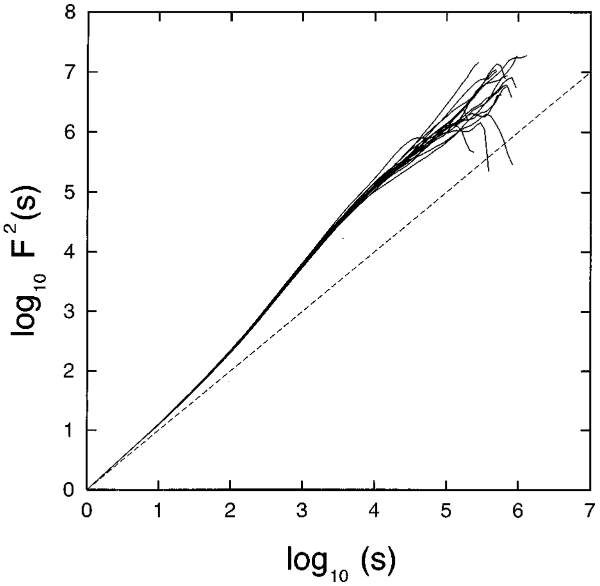

FIG. 2. Mean square fluctuation $F^2(s)$ ($\log_{10}-\log_{10}$ plot) for the 16 chromosomes of *S.c.* As a reference, the dashed line corresponds to the pure linear behavior $F^2(s)=s$.

quences the equivalent quantity is the mean square deviation (or mean square fluctuation) of the walk [1,4]. This is defined as

$$F^2(s)=\langle \Delta \mathbf{r}^2\rangle - \langle \Delta \mathbf{r}\rangle^2, \quad (1)$$

where $\Delta \mathbf{r}=\mathbf{r}(s_0+s)-\mathbf{r}(s_0)$ is the difference in position between the walker at step s and the walker at the (initial) step s_0 and the averages are taken over initial positions s_0.

In Fig. 2 the mean square fluctuation $F^2(s)$ of the DNA walks corresponding to the 16 chromosomes of *S.c.* is shown as a function of the step number s in a $\log_{10}$-$\log_{10}$ plot. It is evident that the walk is superdiffusive. Two superdiffusive regimes seem to appear: at short and at long distances, with a transition at $s\sim 10^2$. Moreover, at $s\sim 10^4$ there seems to occur a transition to a linear regime, following which, for some chromosomes, a negative slope appears.

In this work we will consider the problem of finding a RW model with a MSD similar to this DNA walk. That is, we address the question whether there exists a simple RW analog whose MSD $\langle \mathbf{r}^2(s)\rangle$ behaves similarly to $F^2(s)$. As explained above, we could think of a random walk with correlated steps. Instead of this, we choose a simpler model that consists of a renormalized walk. A single step of the renormalized walk corresponds to s (now a random variable) steps of the original walk. Then the renormalized step displacement (corresponding to the sum of s single displacements of the original walk) will be a random variable $\mathbf{r}$ correlated with s. Therefore, their joint probability distribution $\psi(\mathbf{r},s)$ does not factorize (otherwise, if the variance of each renormalized step is finite, we would obtain pure diffusion).

The coupled scheme that we propose to investigate has the following joint probability distribution for each renormalized single-step displacement:

$$\psi(r,s)\sim \; r^{-1-\gamma}\;\delta(r-s). \quad (2)$$

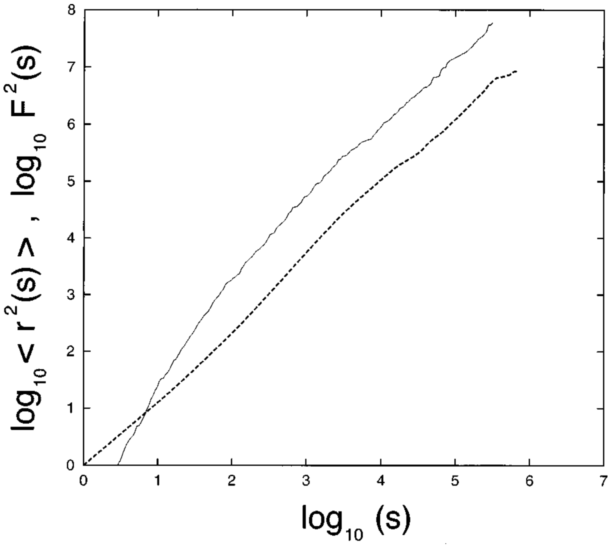

FIG. 3. Shown in a $\log_{10}-\log_{10}$ plot are $\langle \mathbf{r}^2(s)\rangle$ (full line) of the simulated LW [Eq. (2)], with $\gamma=1.537$, and $F^2(s)$ of chromosome *II* (dashed line). Note the deviation at middle and short distances.

Here r is the length of the displacement, $r=|\mathbf{r}|=\sqrt{x^2+y^2}$, while the direction of each displacement (the angle of the vector $\mathbf{r}$) is uniformly distributed in $[0,2\pi)$. We will call this a Lévy walk (LW) model. The MSD can be computed (see the Appendix) and we have

$$\langle \mathbf{r}^2(s)\rangle \sim \begin{cases} s^2 & \text{if } 0<\gamma<1 \\ s^{3-\gamma} & \text{if } 1<\gamma<2 \\ s & \text{if } 2<\gamma. \end{cases} \quad (3)$$

As a first example let us take chromosome *II* of *S.c.* From Fig. 2 we can measure the corresponding slope, that is, the exponent $3-\gamma$ in Eq. (3). We get $3-\gamma=1.463$. Then, with $\gamma=1.537$, we simulate a walk with the here proposed coupled transition probability (2). In Fig. 3 we show the MSD $\langle \mathbf{r}^2(s)\rangle$ obtained from simulations in comparison with the mean square fluctuation $F^2(s)$ of this chromosome.

We see that only the exponent of $F^2(s)$ at large distances can be reproduced by $\langle \mathbf{r}^2(s)\rangle$. In order to account for the whole range of distances, we have found that it is necessary to incorporate a noisy component into the model. This can easily be achieved in the following way:

$$\psi(r,s)\sim p r^{-1-\gamma}\;\delta(r-s)+(1-p)\;\;\delta(r-1)\delta(s-1). \quad (4)$$

The first term, weighted with probability p, corresponds to the already mentioned coupled distribution (2). The second term, with probability $1-p$, is a decoupled probability distribution corresponding to a standard RW. In other words, with probability p the walk is of the coupled form just mentioned, while with probability $1-p$ it is a standard RW. This model has only two parameters to be fitted: $1\leq\gamma\leq 2$ and $0\leq p\leq 1$. In Fig. 4 we show the same case as in Fig. 3, but with this new distribution. The agreement with chromosome *II* of *S.c.* is satisfactory up to the deviation due to finite size effects in the sequence.

We see that the evolution of $F^2(s)$ can be well reproduced by the MSD of this noisy LW, up to the transition to

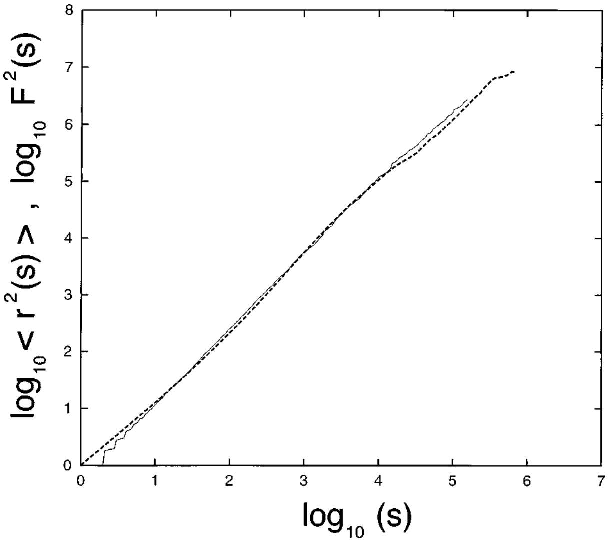

FIG. 4. Same as Fig. 3, using Eq. (5), for chromosome *II*. The analogous noisy Lévy walk (full line) has the parameters $\gamma = 1.758$ and $p = 0.02$.

the linear regime, which occurs at $s \sim 10^4$. We can further ask if this transition can also be understood in terms of a more elaborate RW analog, which in turn could provide a better understanding of its origin. In the following we show that it is enough to include the fact that there exists a cutoff L for the maximum possible step size. The cutoff L is necessary, since in this model the maximum step length cannot be longer than the size of the sequence, i.e., $L < L_{max}$. This is achieved with the model

$$\psi(r,s) = \begin{cases} pc[r^{-1-\gamma} - L^{-1-\gamma}]\ \ \delta(r-s) & \text{for } 2 \leqslant s \leqslant L \\ (1-p)\ \ \delta(r-s) & \text{for } s=1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Here c is a factor that ensures the normalization

$$\int_2^L dr[r^{-1-\gamma} - L^{-1-\gamma}] = 1/c.$$

In other words, at each (renormalized) step of this analogous walk, with probability p the displacement corresponds to a LW and with probability $1-p$ to a standard random walk.

Let us now define the marginal distributions

$$R(r) \equiv \sum_{s=1}^{\infty} \psi(r,s). \tag{6}$$

Now, due to the cutoff, the variance of each single step σ^2,

$$\sigma^2 = \int_0^{\infty} r^2 R(r) dr,$$

and the mean value of the renormalized step $\langle s \rangle$,

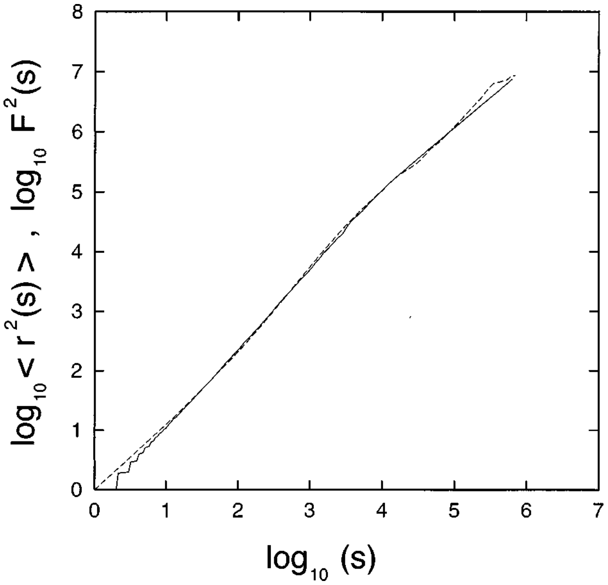

FIG. 5. Same as Fig. 3, using Eq. (5), for chromosome *II*. The analogous noisy Lévy walk (full line) has the parameters $\gamma = 1.537$, $L = 5381$ ($L_{max}/L = 150$), and $p = 0.035$.

$$\langle s \rangle = \sum_{s=1}^{\infty} sS(s),$$

are finite for any value $0 < \gamma$. Therefore, at long distances the diffusive behavior

$$\langle \mathbf{r}^2(s) \rangle = \sigma^2 s / \langle s \rangle \tag{7}$$

is to be expected. The factor $\sigma^2/\langle s \rangle$ can be obtained by measuring the prefactor in the long-s linear regime of $F^2(s)$. The

TABLE I. The 16 chromosomes of *S.c.* are shown. The columns display the length L_{max} of the DNA sequence, the exponent in the proposed power-law form $F^2(s) \sim s^{3-\gamma}$ [see Eq. (3)], the exponent γ, the maximum step size of the LW analog L, the ratio L_{max}/L, and the fraction of LW, p.

Sequence	L_{max}	$3-\gamma$	γ	L	L_{max}/L	p
chr *I*	226646	1.534	1.466	9065	25	0.035
chr *II*	807188	1.463	1.537	5381	150	0.035
chr *III*	315341	1.455	1.545	26278	12	0.040
chr *IV*	1531974	1.504	1.496	5106	300	0.040
chr *V*	574393	1.495	1.505	6660	230	0.037
chr *VI*	270149	1.516	1.484	2785	550	0.041
chr *VII*	1090935	1.496	1.504	7272	150	0.050
chr *VIII*	562638	1.474	1.526	3750	150	0.038
chr *IX*	439885	1.496	1.504	1760	250	0.038
chr *X*	745442	1.467	1.533	8282	90	0.036
chr *XI*	666448	1.485	1.515	2897	230	0.039
chr *XII*	1066141	1.496	1.504	6663	160	0.035
chr *XIII*	924430	1.477	1.523	3555	260	0.039
chr *XIV*	784328	1.492	1.508	3826	205	0.039
chr *XV*	1091282	1.473	1.527	3637	300	0.039
chr *XVI*	948061	1.484	1.516	2370	400	0.039
average	752830	1.488	1.512	6205	216	0.039

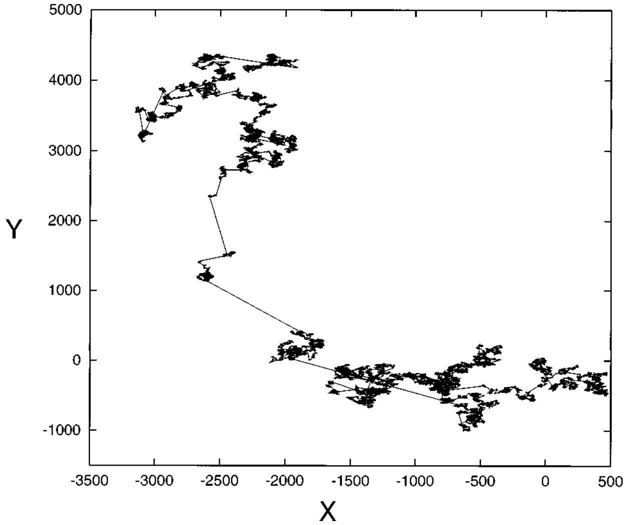

FIG. 6. Noisy Lévy walk analog of the DNA walk corresponding to chromosome *II* of *S.c.* Compare with Fig. 1.

exponent γ is obtained by measuring the slope of the log-log plots of $F^2(s)$ in the (now transitory) superdiffusive regime.

In Fig. 5 we show the resulting $\langle \mathbf{r}^2(s) \rangle$ compared with $F^2(s)$ for chromosome *II*. In this way we obtained the fitted parameters γ, L, and p shown in Table I.

In Figs. 6 and 7 we show a simulation of the noisy Lévy walk defined by Eq. (5). It is a walk with the same parameters as for chromosome *II*. Compare with Fig. 1.

In this work we presented a RW analog of DNA sequences, as exemplified by the chromosomes of the organism *S.c.* We found that this two-dimensional walk can be modeled or simulated as a noisy Lévy walk, as long as one focuses on the MSD and on the resulting pattern of the visited points. The resemblance of this modified LW to the DNA walk of the *S.c.* is apparent. Note that the main proportion of the analog is noise (standard RW), while the LW component amounts to only 4%. The main parameter of the LW is the exponent γ of the step-size distribution (2). For all chromosomes it is $1<\gamma<2$ with a mean value $\langle \gamma \rangle = 1.51 \pm 0.02$. This leads to a *superdiffusive* MSD $\langle \mathbf{r}^2(s) \rangle \sim s^{3-\langle \gamma \rangle} = s^{1.49}$ [see Eq. (3)]. The resulting pattern is a random fractal, whose fractal dimension cannot be easily related theoretically to γ. This finding raises the question about why a simple LW model behaves so similarly to a DNA sequence, both qualitatively and quantitatively. If there is an underlying conceptual reason for this analogy a different pathway in the mathematical study of DNA sequences could be opened.

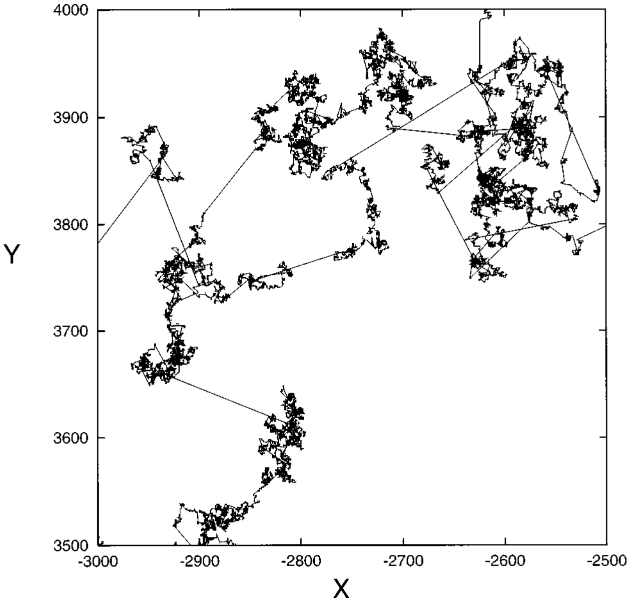

FIG. 7. Amplification of Fig. 6.

Recently, our attention was directed to the works of West and co-workers [16,17]. As in the previously cited works of Stanley *et al.*, these authors consider a one-dimensional RW mapping, while we are introducing here a two-dimensional RW mapping. A second difference is that they develop a ‘‘dynamical’’ (deterministic) method that mimics an α-stable Lévy process with $1<\alpha<2$. The generator of the deterministic evolution is a nonlinear map belonging to a class of maps recently tailored to mimic the process of weak chaos responsible for the birth of anomalous diffusion. As a similar conceptual idea, these authors consider this process to be superposed to another random one and to be δ-function correlated. They call this prescription to generate statistical sequences the copy mistake map (CMM). On the contrary, in our model we consider no deterministic dynamics at all, but the superposition of a Lévy walk, responsible for the emergence of correlations, with a standard RW (i.e., a random noise as in the CMM model). In the CMM model the correlation effect of the deterministic dynamics is canceled on the short-range scale, but shows up in the long-range one. In our model this effect also appears, as one can see from a comparison of Fig. 3 with Fig. 4 or 5. Both models cannot be quantitatively compared since they are essentially different. However, besides this difference, there remains the conceptual similarity of describing intronless (coding) and intron-containing (noncoding) sequences in a unified way, interpreting the DNA sequences as the superposition of two processes, one responsible for the long-ranged correlations and the other essentially a noise. The latter can be interpreted as uncorrelated random mutations that destroy short-range correlations. In fact, in real DNA sequences no large patches of consecutive sites (straight displacements in the RW mapping) are observed.

We acknowledge fruitful discussions with Carlo Bruschi. P.A.A. gratefully acknowledges partial support of the Deutsche Forschungsgemeinschaft (Grant No. SFB 428) and the kind hospitality of Theoretische Polymerphysik at Freiburg University.

APPENDIX

The so-called Lévy walk model was studied by Blumen, Klafter, and Zumofen (BKZ) [7,11,12]. While in a Lévy flight [5,7,13,14] the single-step displacement of the walker has a divergent variance and therefore the mean square displacement is not defined (and, strictly speaking, is also infinite) in the LW, this trouble is solved by introducing a spatiotemporal coupling. The LW is defined as a continuous-time random walk with a coupled single-step displacement and waiting time between the steps of the form

$$\psi(\mathbf{r},t) \sim r^{-1-\gamma} \delta(r-t^{\nu}). \tag{A1}$$

The model has two parameters γ and ν. For $0<\gamma<2$ the variance of each single step is infinite. On the other hand, the δ function penalizes large steps by requiring longer times for them. The resulting MSD was computed by BKZ. In terms of

$$\mu\equiv\gamma+1$$

and for $\gamma>1$ and $\nu>1/2$ it is

$$\langle \mathbf{r}^2(t)\rangle\sim\begin{cases} t^{2\nu} & \text{if} \quad 1<\nu\mu<2 \\ t^{2-\nu\mu+2\nu} & \text{if} \quad 2<\nu\mu<1+2\nu \\ t & \text{if} \quad 1+2\nu<\nu\mu. \end{cases} \tag{A2}$$

Then our noisy LW analog takes $\nu=1$ and we view the DNA walk as a LW in which the waiting time between consecutive displacements t is given by the number of steps s required to reach the distance $\mathbf{r}(s)$.

[1] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, S. M. Ossadnik, C. K. Peng, and M. Simons, Fractals **1**, 283 (1993).

[2] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, S. R. N. Mantenga, M. Simons, and H. E. Stanley, Physica A **221**, 180 (1995).

[3] D. Benson *et al.*, Nucleic Acids Res. **24**, 1 (1996). Genbank can be accessed through the WWW at the National Center for Biotechnology information (USA) (http://www.ncbi.nlm.nih.gov).

[4] C. K. Peng *et al.*, Nature **356**, 168 (1992).

[5] E. W. Montroll and B. J. West, in *Fluctuation Phenomena*, edited by E. W. Montroll and J. L. Lebowitz (Elsevier, Amsterdam, 1979).

[6] G. Zumofen, A. Blumen, J. Klafter, and M. F. Shlesinger, J. Stat. Phys. **54**, 1519 (1989).

[7] J. Klafter, M. F. Shlesinger, and G. Zumofen, Phys. Today **49** (2), 33 (1996).

[8] T. Geisel, J. Nierwetberg, and A. Zacherl, Phys. Rev. Lett. **54**, 616 (1985).

[9] M. F. Shlesinger and J. Klafter, Phys. Rev. Lett. **54**, 2551 (1985).

[10] M. F. Shlesinger, B. J. West, and J. Klafter, Phys. Rev. Lett. **58**, 1100 (1987).

[11] A. Blumen, G. Zumofen, and J. Klafter, Phys. Rev. A **40**, 3964 (1989).

[12] A. Blumen, J. Klafter, and G. Zumofen, Europhys. Lett. **13**, 223 (1990).

[13] A. Ott, J. P. Bouchaud, D. Langevin, and W. Urbach, Phys. Rev. Lett. **65**, 2201 (1990).

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

[15] P. A. Alemany and D. H. Zanette, Phys. Rev. E **49**, R956 (1994); D. H. Zanette and P. A. Alemany, Phys. Rev. Lett. **75**, 366 (1995); P. A. Alemany, Phys. Lett. A **235**, 452 (1997).

[16] P. Allegrini, M. Barbi, P. Grigolini, and B. J. West, Phys. Rev. E **52**, 5281 (1995).

[17] P. Allegrini, P. Grigolini, and B. J. West, Phys. Lett. A **211**, 217 (1996).