Non-normal Statistics of DNA Sequences of Prokaryotes

BRUCE J. WEST¹, PAOLO ALLEGRINI², M. BUIATTI⁴ and
PAOLO GRIGOLINI^{2–4}

¹*Army Research Office, Research Triangle Park, NC, USA*

²*Center for Nonlinear Science, University of North Texas, Denton, Texas, 76203 USA*

³*Dipartimento di Fisica dell' Università di Pisa, Piazza, Torricelli 2, and 56127 Pisa, Italy*

⁴*Istituto di Biofisica del Consiglio Nazionale delle Ricerche, Via San and Lorenzo 26, 56127 Pisa, Italy*

Abstract. The $\sqrt{n}$ -rule of Schrödinger in his discussion of DNA is based on normal statistics and equilibrium physics. Herein the kurtosis is used to measure the deviation from normality of the statistics of non-equilibrium DNA sequences. A pattern for this deviation from normality is identified and this signature is found in prokaryotes. The signature is explained by a theory of DNA sequences that involves finite length DNA walks with dynamically generated long-range correlations.

Key words: DNA walk, inverse power law, Lévy statistics, nonlinear dynamics

1. Introduction

In his seminal book, ‘What is Life?’, Erwin Schrödinger [1], one of the founders of quantum mechanics, repeatedly makes the point that all the physical and chemical laws that play important roles in the life of organisms are statistical. At the time it was quite remarkable to have one of the leaders of the physical sciences point out that the descriptive paradigm of natural order, the laws of physics, are inaccurate within a probable relative error of order $1/\sqrt{n}$, where n is the number of molecules involved in the phenomenon to which the law is applied. The example he gave concerned a gas under certain conditions of pressure and temperature that has a certain density. He emphasized that if a container confines n molecules of gas then at any moment in time the relation among the thermodynamic variables could be tested and would be found to be inaccurate by departures of order $\sqrt{n}$. According to Schrödinger the $\sqrt{n}$ -rule concerns the degree of inaccuracy to be expected in any physical law.

Although he does not discuss this fact, buried in his $\sqrt{n}$ -rule, is the legacy of Gauss and the law of errors, which subsequently evolved into the Central Limit Theorem (CLT) and the Gaussian or Normal distribution of probability. Schrödinger did, however, discuss the idea that the macroscopic laws we observe in natural phe-

nomena are the consequence of the interaction of a large number of particles, and therefore can only be described by means of statistical physics. He further went on to say that because of the regularity in biology, in particular in heredity, that while still conforming to the laws of physics, biology might, in fact, have additional laws which had not up to that time been identified. The idea of formulating a new law of physics did not intimidate Schrödinger, since he, after all, had been one of the architects of the quantum revolution.

Much has changed in the fifty years since Schrödinger first passed along his insights regarding the application of physical law to biology, in particular the way in which we understand the statistics of natural phenomena in the context of the life sciences. Herein we do not propose a new law in Schrödinger's sense, but we do present evidence for the deviation from his $\sqrt{n}$ -rule, which suggests that the statistical basis for his understanding of biology has changed. One of the most successful models of the statistics of DNA sequences during this time has been the DNA walk. A DNA walk is a chain of sites, each occupied by either a purine (adenine and guanine) or a pyrimidine (cytosine and thymine). One imagines a random walker whose displacement x at the j th site increases by +1 if the DNA site is occupied by a purine, or decreases by the same amount if the site is occupied by a pyrimidine, so that

$$x(n) = \sum_{j=1}^n \xi_j, \quad \xi_j = \pm 1. \quad (1)$$

In spite of its success in modeling the long-range correlations observed in DNA sequences [2–6], as indicated by the power-law increase in the variance and the inverse power-law spectrum [3–6], the problem of the correct statistical interpretation of the DNA sequence is still unsolved.

2. Normal Statistics

For these statistical arguments we turn to the discussion of the statistics of DNA sequences presented by Peng et al. [3]. In their DNA walk different initial conditions are realized by assigning to different sites of the DNA chain the role of departure point of the walker. This means that having available a sufficiently long DNA sequence, it is possible to realize a condition equivalent to that ordinarily adopted in statistical physics, namely, the observation of a large number of realizations of the walk. This makes it possible to average over a collection of realizations, each with a different initial condition, and thus evaluate the second moment of the distance $x(n)$ traveled by the DNA random walker in a 'time' n :

$$\langle x(n)^2 \rangle \propto n^{2H}. \quad (2)$$

In the case of noncoding regions in DNA sequences (introns) Stanley et al. [4] find $H > 1/2$, in agreement with the results of Voss [5]. In the language of

random walks this corresponds to a walker who once having taken a step in a given direction has a greater likelihood of continuing in that direction than in changing direction. This is persistence and is interpreted as a long-range correlation in the walk, one which is superdiffusive. On the other hand, in the case of intronless sequences (where the introns were removed leaving only the coding regions of DNA sequences called exons) Stanley et al. [4] find $H \approx 0.5$ for sufficiently small n . Again in the language of random walks this value of H corresponds to a walker whose steps are independent of one another and leads to a normal diffusive process. This dual aspect of the DNA sequence data was explained by Allegrini et al. [6] using the copying mistake map (CMM) model which is a statistical superposition of two distinct processes. One dominated by an uncorrelated random walk at small n and by a long-range correlated stochastic process (intermittent chaotic dynamical process) at large n .

The probability of constructing the sequence with the correlation-generating dynamical prescription is p_c , and the probability of constructing the sequence with the random law is $1 - p_c$. it is then a simple matter to show that the correlation function is given by [6]

$$\Phi_x(t) = (1 - p_c^2) \Phi_\varepsilon(t) + p_c^2 \Phi(t), \quad (3)$$

where $\Phi_\varepsilon(t)$ is the correlation function of the fast fluctuations (delta function) and $\Phi(t)$ is the long-range correlation of the underlying process (inverse power law). It is equally straightforward to show that (3) yields an expression for the second moment of the landscape variable given by

$$\langle x(n)^2 \rangle = An + Bn^{2H}, \quad (4)$$

which is the superposition of a term linear in time, corresponding to ordinary Brownian motion, and of a term faster than linear, corresponding to anomalous diffusion ($H > 1/2$). Thus, the CMM model results in diffusion processes that are indistinguishable from standard Brownian motion at short times (short sequences) and are expected to exhibit anomalous diffusion at long times (long sequences). To mimic the statistical properties of exons the parameter values fit to the data in the CMM model requires that $A \gg B$ in (4), explaining why the detection of the long-range correlations is so difficult for these kinds of sequences, and why the short-time dynamics is essentially dominated by the properties of standard diffusion. However, even though the correlation properties of the DNA sequences can be understood in terms of the CMM model, the expected statistics of the DNA sequence differs among the various groups.

The consensus is that an uncorrelated random walk cannot be a proper paradigm to interpret the experimental data for introns since $H = 1/2$ in that case. On the other hand, some of these groups [4, 7] do interpret DNA sequences using the theory of fractional Brownian motion (*fBm*). The latter is equivalent to assuming that the statistical distribution of random walkers is normal, but with a power-law correlation function. Alternatively, it has been shown [8–10] that a dichotomous

random process with an inverse power-law correlation function produces, in the stationary regime, a truncated Lévy process description of superdiffusion. This exact solution to a two-state model with memory can be used to describe the DNA walk, resulting in the DNA statistics being Lévy stable rather than normal.

The Lévy stable distribution is the general solution to the CLT. We recall that the Central Limit Theorem, in addition to being dependent on the sum of a large number of identically distributed random variables, is essentially based on two facts: (a) the finiteness of the second moment of the underlying variables, and (b) the mutual statistical independence of the variables themselves. When these conditions are satisfied the normal distribution can be proved and Schrödinger's $\sqrt{n}$ -rule follows. If condition (a) is relaxed but (b) is satisfied, the generalization of the CLT predicts a Lévy distribution. In a complementary way, when condition (b) is relaxed and (a) is satisfied, as in the two-state model with memory, the resulting distribution is not Lévy stable because there are ballistic propagation fronts. The probability distribution at distances $|x| > Wt$ vanishes, where W is the 'velocity' of the walker at each step, and the population of the missing tail concentrates on the front thereby resulting in two peaks. A very accurate representation of the probability distribution $P(x, t)$ is given by [10]

$$P(x, t) = P_L(x, t) \Theta(|x| - Wt) + \delta(|x| - Wt) \frac{\Phi(t)}{2}, \quad (5)$$

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

$$\phi_\mu(k, t) = e^{-bt|k|^\mu}, \quad (6)$$

where $b > 0$ is a known parameter and the power-law index is in the interval $0 < \mu \leq 2$. The first term on the right-hand side of (5) is a truncated Lévy distribution. The second term is the correlation function of the two-step process driving the random walk. When the correlation function is an inverse power law with index β , the Lévy index is given by $\mu = \beta + 1$. Thus, the probability density $P(x, t)$ is not Gaussian as a consequence of the lack of a finite microscopic time scale, that is, $\beta < 1$. The long, inverse power-law tails of the Lévy distribution are truncated, but their weight is included in the propagating peaks, so that asymptotically Lévy statistics ensue [10].

3. Measure of non-normality

Thus, we have a number of theoretical analyses of the statistics of the DNA sequences, and some of them converge on a single description, those being the CMM model of Allegrini et al. [6, 10] and the generalized Lévy walk of Araujo et al. [11] and Buldyrev et al. [12], which has been shown to be equivalent by Allegrini et al. [13]. However, we want a more direct measure of the non-normal character of DNA sequence statistics; one that is independent of any and all of the above models. Therefore we introduce the kurtosis $\eta(n)$ defined by

Figure 1. The kurtosis is plotted as a function of t , the number of steps, for different realizations of the CMM model. The realizations are determined by the relative probability of the uncorrelated random process and the dynamically generated random process with long-range correlations: $p_c = 1$ (+), $p_c = 0.5$ (x), $p_c = 0.2$ (*), $p_c = 0.002$ (□). The total length of the sequence is 10^6 and the index of the power law correlation function is 0.67.

$$\eta(n) \equiv 1 - 3 \frac{\langle x^2(n) \rangle^2}{\langle x^4(n) \rangle}, \quad (7)$$

where the brackets denote an average over an ensemble of realizations of a DNA walk of length n . The second and fourth moments of steps in the underlying process are denoted by $\langle \varepsilon^2 \rangle$ and $\langle \varepsilon^4 \rangle$, respectively. In a dichotomous process the moments factor, so that $\langle \varepsilon^4 \rangle = \langle \varepsilon^2 \rangle^2$ and therefore the kurtosis is -2 for $n=1$ as depicted in Figure 1.

The fourth moment, for uncorrelated, identically distributed, random variables in an n -step walk, can be written as

$$\langle x^4(n) \rangle = n \langle \varepsilon^4 \rangle + 3n(n-1) \langle \varepsilon^2 \rangle^2, \quad (8)$$

and the second moment in the n -step walk is

$$\langle x^2(n) \rangle = n \langle \varepsilon^2 \rangle, \quad (9)$$

so that the kurtosis Eq. 7 becomes

$$\eta(n) \equiv 1 - 3 \frac{\langle \varepsilon^2 \rangle^2}{\frac{1}{n} \langle \varepsilon^4 \rangle + 3 \left(1 - \frac{1}{n}\right) \langle \varepsilon^2 \rangle^2}, \quad (10)$$

which vanishes as $n \rightarrow \infty$. Since Eq. 8 and Eq. 9, with $\langle \varepsilon^2 \rangle < \infty$, are the conditions for the landscape variable to be normal as $n \rightarrow \infty$ according to the CLT, we see that the kurtosis is zero for a normal random variable. Thus, for a simple random walk (no memory) the kurtosis regresses from -2 to 0 as the number of steps becomes large, as is seen in the numerical realization depicted in Figure 1. Therefore, deviations of the kurtosis from zero provide a quantitative measure

of the deviation of the landscape process from normal statistics. In addition, one source of non-normal statistics is the finite length of the sequence even when all the other conditions of the CLT are fulfilled. A second source of non-normal statistics is the correlation between the steps in the walk. In particular the influence of the observed inverse power-law correlation function on the statistical properties of the DNA sequences.

4. Results and Conclusions

In Figure 1 the dependence of the kurtosis on the size of the sequence, evaluated using Eq. 7, is depicted for an artificial DNA walk with and without correlations. In the absence of correlations the kurtosis is shown to relax from -2 to zero after approximately ten terms in the sequence, after which it fluctuates around zero as the sequence is made longer and longer. When an inverse power-law correlation among the steps of the random walker is included, using the CMM model to generate the DNA sequence, there is an apparently universal effect that is only modestly dependent on the inverse power-law index. When the correlation of steps in the two-state random walk is a delta function, the evolution of the probability density is determined by a Fokker-Planck equation whose solution is a Gaussian distribution. The variance of the random variable in this case increases linearly with ‘time’ (n). However, when the correlation of steps in the two-state random walk is an inverse power law of the form $n^{-\beta}$, with $0 < \beta < 1$, the evolution of the probability density is given by a fractional diffusion equation in which the second derivative of the Fokker-Planck equation is replaced with a fractional spatial derivation, as observed by West and Seshadri in 1982 [16]. The equation of evolution for the characteristic function is given by

$$\frac{\partial \phi_\mu(k, t)}{\partial t} = -b |k|^\mu \phi_\mu(k, t), \quad (11)$$

which has the solution given by (6) [14]. Note that for $\mu = 2$ the Fourier transform of (11) yields the Fokker-Planck equation, but for $0 < \mu < 2$ the right-hand side of (11) yields a fractional derivative of the Riesz-potential form [15]. The solution to this two-state random walk model is given by the truncated Lévy distribution given by (5).

The central moments of the truncated Lévy distribution (5) are finite and the kurtosis decays as

$$\eta(n) = 1 - C_\beta n^{-\beta}, \quad (12)$$

where C_β is a known constant. Therefore, the kurtosis overshoots zero in the same way for all values of the power-law index after about 10^3 terms in the sequence, attempting to reach an equilibrium value of one as would be appropriate for a truncated Lévy process. However, because of its finite length the random walk never reaches the equilibrium value of one, but instead forms a bump. The bump

Figure 2. The kurtosis is plotted as a function of the number of steps t . (a) the complete genome of the Human Cytomegalovirus (DEDCMVCG) (+) and the CMM sequence (×). The two sequences share the length 229,354. The CMM sequence is given the parameter $p_c = 1/9$, which is the same value as that adopted in Allegrini et al. [6] 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 91,430. The CMM sequence is given the parameter $p_c = 0.1$.

has a maximum at approximately 3×10^4 terms after which the kurtosis again apparently begins relaxing back to zero. The CMM model therefore predicts a signature kurtosis curve as a function of n , a bump on the plateau of normality.

Arneodo et al. [7] observed that the diffusion statistics in eukaryotes are essentially normal in spite of the existence of long-range correlations. To explain this effect the present authors developed a folding model which has the effect of decoupling statistics from dynamics [13]. This means that statistics can be produced by sources distinct from that responsible for the long-range fluctuations, thereby resolving the apparent conflict between normal statistics and long-range correlations. However, as far as prokaryotes are concerned, different conditions apply, and these make it possible to detect non-normal statistics using the kurtosis.

What about real DNA sequences of prokaryotes? To apply the kurtosis to experimental DNA sequence data, the second and fourth moments used in the definition (2) are calculated using the method of averaging over the initial conditions described in Peng et al. [3]. In Figure 2 we see for two classes of prokaryotes that the signature of the kurtosis obtained using the CMM model-generated data depicted in Figure 1 is obtained using data. The agreement is good at both the qualitative and quantitative levels. The bump is a clear indication of the non-normal character of the statistics of the prokaryotes, but even more strongly it is a deviation from normality that is produced by long-range correlations in the DNA sequences of prokaryotes and which is eventually suppressed by the finite lengths of the sequences.

This deviation from normality can be used to address Schrödinger's question regarding the failure of the $\sqrt{n}$ -rule to apply in heredity. It is not that the laws of physics do not apply to biology, but that they apply in a stronger form than they do in the equilibrium physical phenomena discussed by Schrödinger. It is the nonequilibrium nature of live matter discussed by Schrödinger that accounts for the inapplicability of the $\sqrt{n}$ -rule. In equilibrium phenomena there is no memory of the initial state of the system so that all phenomena in equilibrium are equivalent. Nonequilibrium phenomena, however, remember previous states and are guided in their evolution by that knowledge. It is this memory that inhibits live matter from relaxing into the equilibrium statistics of normality, and it is this same memory that produces the bump in the prokaryote data in Figure 2. In fact the non-Gaussian effects observed here in prokaryotes are even more intense than those found in quenched glasses.

References

1. Schrödinger, E.: 1995. *What is Life?*. Cambridge: Cambridge University Press; first published 1944.
2. Li, W. and Kaneko, K.: *Europhys. Lett.* **17** (1992), 655.
3. Peng, C.K., Buldyrev, S., Goldberger, A.L., Havlin, S., Sciotino, F., Simons, M. and Stanley, H.E.: Long-range correlations in nucleotide sequences. *Nature* **356** (1992), 168–172.
4. Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Goldberg, Z.D., Havlin, S., Mantegna, R.N., Ossadnik, S.M., Peng, C.K. and Simon, S.M.: Statistical mechanics in biology: how ubiquitous are long-range correlations. *Physica A* **205** (1994), 214–235.
5. Voss, R.: Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **68** (1992), 3805–3808.
6. Allegrini, P., Barbi, M., Grigolini, P. and West, B.J.: Dynamical model for DNA sequences. *Phys. Rev. E* **52** (1995), 5281–5286.
7. Arneodo, A., Bacry, E., Graves, P.V. and Muzy, J.F.: *Phys. Rev. Lett.* **74** (1995), 3293–3296.
8. Klafter, J. and Zumofen, G. *Physica A* **196** (1993), 102–111.
9. Trefán, G., Florian, E., West, B.J. and Grigolini, P.: Dynamical approach to anomalous diffusion: Response of Lévy processes to a perturbation. *Phys. Rev. E* **50** (1994), 2564–2579.
10. Allegrini, P., Grigolini, P. and West, B.J.: Dynamical approach to Lévy processes. *Phys. Rev. E* **54** (1996), 4760–4767.

11. Araujo, M., Havlin, S., Weiss, G.H. and Stanley, H.E.: Diffusion of walkers with persistent velocities. *Phys. Rev.* **43** (1991), 5240–5247.
12. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E.: Generalized Lévy-walk model for DNA nucleotide sequences. *Phys. Rev. E* **47** (1993), 4514–4521.
13. Allegrini, P., Buiatti, M., Grigolini, P. and West, B.J.: Non-Gaussian statistics of anomalous diffusion: the DNA sequences of prokaryotes. *Phys. Rev. E* **57** (1998), 1–11.
14. Montroll, E.W. and West, B.J.: An enriched collection of stochastic processes, In: E.W. Montroll and J.L. Lebowitz (eds.), *Fluctuation Phenomena*, 2nd ed., *Studies in Statistical Mechanics* Vol. **7**, North-Holland, Amsterdam, 1987.
15. B.J. West and V. Seshadri: Linear systems with Lévy fluctuations, *Physica A* **113** (1982), 203–216.
16. B.J. West, *Physiology, Promiscuity and Prophecy at the Millennium: A Tale of Tails*, World Scientific, Singapore, 1999.