JOURNAL DE PHYSIQUE IV
Colloque C1, supplément au Journal de Physique II, Volume 3, mai 1993

Scaling concepts and complex fluids: long-range power-law correlations in DNA

H.E. STANLEY, S.V. BULDYREV, A.L. GOLDBERGER*, S. HAVLIN, C.-K. PENG, F. SCIORTINO and M. SIMONS*,**

Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, U.S.A.
* *Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, U.S.A.*
** *Department of Biology, MIT, Cambridge, MA 02139, U.S.A.*

We review recent evidence supporting the discovery of long-range power-law correlations in DNA sequences that contain non-coding regions. The possible interpretation of this finding is also discussed.

Nous présentons les récentes évidences supportant la découverte de la corrélation de la loi puissance sur des séquences à longue distance d'ADN qui contiennent des régions non-codantes. L'interprétation de cette découverte est également discutée.

Scaling concepts have played a key role in our understanding of phenomena occurring near critical points. A scale invariant function $f(x)$ has the remarkable property that each time x is doubled, the function $f(x)$ changes by the same factor. There is thus no way to set a characteristic scale for such a function.

Stated mathematically, if the variable x is increased by an arbitrary factor λ, then the function is changed by a factor λ^p which is independent of the value of x,

$$f(\lambda x) = \lambda^p f(x) \tag{1a}$$

for all λ. An algebraic equation for x, such as $x^2 = 4$, constrains the values of x to be ± 2. Similarly, a functional equation, such as (1a), constrains the set of possible functional forms of $f(x)$: any function $f(x)$ satisfying (1a) must be a power law, as may be seen by substituting the choice $\lambda = 1/x$ in (1a),

$$f(x) = Ax^p \tag{1b}$$

We say that scale invariance [Eq. (1a)] implies power-law behavior [Eq. (1b)]. Conversely, power-law behavior implies scale invariance, since any function $f(x)$ obeying (1b) also obeys (1a)—one can verify this by substitution. Thus scale invariance is mathematically equivalent to power law behavior.

Power laws are found to describe various functions in the vicinity of critical points. Such systems include not only materials with Hamiltonians (such as the Ising and Heisenberg models) but also purely geometric systems, such as percolation. Scaling is also found to hold for polymeric systems, including both linear and branched polymers. Here power law correlations develop in the asymptotic limit in which the number of monomers approaches infinity. The list of systems in which power law correlations appear has grown rapidly in recent years, including models of rough surfaces, turbulence, and earthquakes. In this talk, I will present recent work suggesting that—under suitable conditions—the sequence of base pairs or "nucleotides" in DNA also displays

power law correlations. The underlying basis of such power law correlations is not understood at present, but it is at least possible that this reason is of as fundamental importance as it is in other systems in nature that have been found to display power-law correlations.

1. Information coding in DNA

Genomic sequences contain numerous "layers" of information. These include specifications for mRNA sequences responsible for protein structure, identification of coding and non-coding parts of the sequence, information necessary for specification of regulatory (promoter, enhancer) sequences, information directing protein-DNA interactions, directions for DNA packaging and unwinding. The genomic sequence is likely the most sophisticated and efficient information database created by nature through the dynamic process of evolution. Equally remarkable is the precise transformation of different layers of information (replication, decoding, etc) that occurs in a short time interval. While means of encoding some of this information is understood (for example, the genetic code directing amino acid assembly, sequences directing intron/exon splicing, etc.), relatively little is known about other layers of information encrypted in a DNA molecule. In the genomes of high eukaryotic organisms, only a small portion of the total genome length is used for protein coding. The role of introns and intergenomic sequences constituting a large portion of the genome remains unknown. Furthermore, only a few quantitative methods are currently available for analyzing such information.

2. Conventional statistical analysis of DNA sequences

DNA sequences have been analyzed using a variety of models that can basically be considered in two categories. The first types are "local" analyses; they take into account the fact that DNA sequences are produced in sequential order, so the neighboring base pairs will affect the next attaching base pair. This type of analysis, such as n-step Markov models, can indeed describe some observed short-range correlations in DNA sequences. The second category of analyses is more "global" in nature; they concentrate on the presence of repeated patterns (such as periodic repeats and interspersed base sequence repeats) that can be found mostly in eukaryotic genomic sequences. A typical example of analysis in this category is the Fourier transform analysis which can identify repeats of certain segments of the same length in base pair sequences [1].

However, DNA sequences are more complicated than these two standard types of analysis can describe. Therefore it is crucial to develop new tools for analysis with a view toward uncovering the mechanisms used to code other types of information in DNA sequences.

3. Scale-invariant (fractal) analysis of DNA sequences

In the last decade, scaling analysis (fractal) techniques have been developed for detecting scale-invariant statistical patterns and study physical properties in complex fluids and other random systems. These methods have been successfully applied in a number of disciplines and to a number of problems including stochastic growth processes in physics and chemistry, polymer physics, as well as other problems [2–4]. Since DNA sequences are long polymer chains, some general scale-invariant properties found in polymer physics [6,7] may appear in DNA, and alterations of those general properties may serve for characterization of DNA sequences.

A new approach to studying stochastic properties of DNA involves the construction of a 1:1 map of the base pair sequence projected onto a walk—which we term a "DNA walk" [5]. The mapping is then used to obtain a quantitative measure of the *correlation* between base pairs over long distances along the DNA chain. In addition, the technique provides a novel graphical "fingerprint" representation of DNA structures. Since DNA sequences are long polymer chains, some general scale-invariant properties found in polymer physics [6,7] may appear in DNA, and variations of those general properties may serve for characterization of DNA sequences.

In this fashion we uncovered in the base pair sequence a remarkably *long-range* power law correlation that is significant because it implies a new scale invariant (fractal) property of DNA. Such long-range correlations are limited to non-coding sequences (introns, regulatory untranscribed gene elements and intergenomic sequenees) and occur in organisms as diverse as hepatitis delta agent, cytomegalovirus, yeast chromosome and a large number of eukaryotic genes encoding a variety of proteins (see [5]).

The power-law decay correlations are of interest because they cannot be accounted for by the standard Markov chain model or other short-range correlations models (which will only give rise to an exponential decay in correlation). On the other hand, unlike the standard Fourier transform analysis [1] that detects the periodical repeats described by a few characteristic length scales, our analysis shows that there exist statistically self-similar patterns on all length scales.

4. The "DNA walk" or "fractal landscape" representation

In order to study the scale-invariant long-range correlations of the DNA sequences, we first introduced a graphical representation of DNA sequences, which we term a "fractal landscape" or "DNA walk". For the conventional one-dimensional random walk model, a walker moves either up $[u(i) = +1]$ or down $[u(i) = -1]$ one unit length for each step i of the walk [2]. For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated random walk, the direction of each step depends on the history ("memory") of the walker. The DNA walk is defined by the rule that the walker steps up $[u(i) = +1]$ if a pyrimidine occurs at position a linear distance i along the DNA chain, while the walker steps down $[u(i) = -1]$ if a purine occurs at position i (Fig. 1). The question we asked was whether such a walk displays only short-range correlations (as in an n-step Markov chain) or long-range correlations (as in critical phenomena and other scale-free "fractal" phenomena).

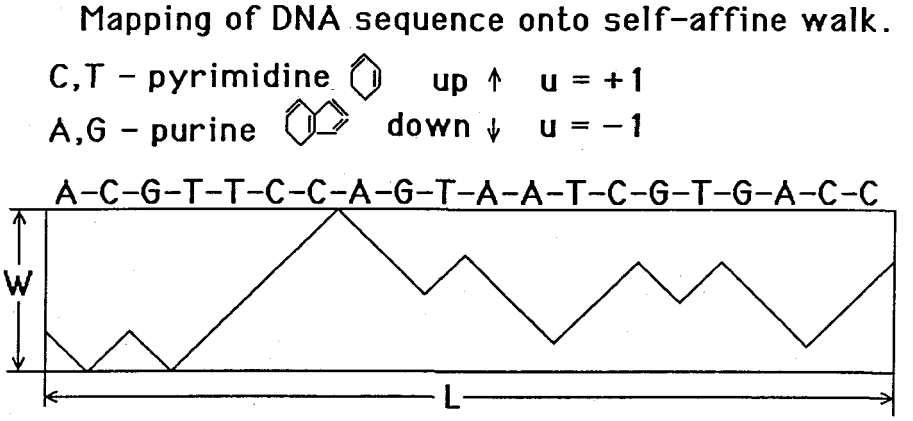

Fig. 1: Schematic illustration showing the definition of the "DNA walk".

The DNA walk provides a graphical representation for each gene and permits the degree of correlation in the base pair sequence to be directly visualized, as in Fig. 2. Figure 2 naturally motivates a quantification of this correlation by calculating the "net displacement" of the walker after l steps, which is the sum of the unit steps $u(i)$ for each step i. Thus $y(l) \equiv \sum_{i=1}^{l} u(i)$.

An important statistical quantity characterizing any walk [2] is the root mean square fluctuation $F(l)$ about the average of the displacement; $F(l)$ is defined in terms of the difference between the average of the square and the square of the average,

$$F^2(l) \equiv \overline{[\Delta y(l) - \overline{\Delta y(l)}]^2} = \overline{[\Delta y(l)]^2} - \overline{\Delta y(l)}^2, \tag{2a}$$

of a quantity $\Delta y(l)$ defined by $\Delta y(l) \equiv y(l_0 + l) - y(l_0)$. Here the bars indicate an *average* over all positions l_0 in the gene. Operationally, this is equivalent to (a) taking a set of calipers set for a fixed distance l, (b) moving the beginning point sequentially from $l_o = 1$ to $l_o = 2, \cdots$ and (c) calculating the quantity $\Delta y(l)$ (and its square) for each value of l_o, and (d) averaging all of the calculated quantities to obtain $F^2(\ell)$.

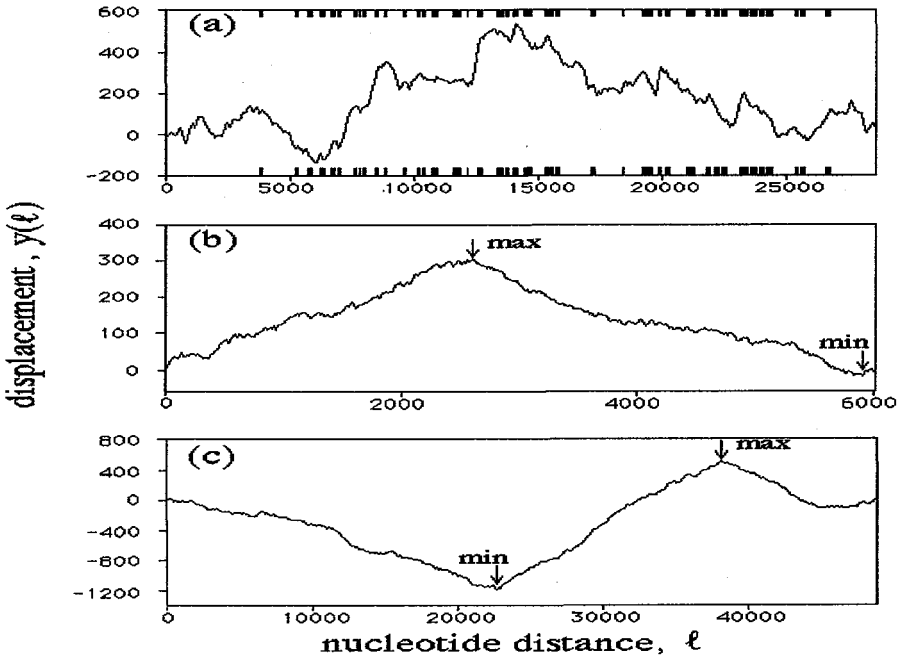

Fig. 2: The DNA walk representations of (a) intron-rich human β-cardiac myosin heavy chain gene sequence, (b) its cDNA, and (c) the intron-less bacteriophage lambda DNA sequence. Note the more complex fluctuations for the intron-containing gene in (a) compared with the intron-less sequences (b) and (c). The heavy bars in (a) correspond to the coding regions of the gene. In order that the graphical representation not be affected by the global differences in concentration between purines and pyrimidines, we plot the DNA walk representations such that the end point has the same vertical displacement as the starting point (for the statistical analysis, we used the original definition, without any adjustment of vertical displacement). The minimum (min) and maximum (max) points on the landscape are denoted by arrows. We found that for almost all intron-less genes and cDNA sequences studied that there appear regions with one strand bias, followed by regions of a different strand bias. The fluctuation on either side of the overall strand bias we found to be random, a fact that is plausible by visual inspection of the DNA walk representations.

The mean square fluctuation is related to the auto-correlation function $C(l) \equiv \overline{u(l_0)u(l_0 + l)} - \overline{u(l_0)}^2$ through the relation: $F^2(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} C(j - i)$. The calculation of $F(l)$ can distinguish three possible types of behavior. (i) If the base pair sequence were random, then $C(l)$ would be zero on average [except $C(0) = 1$], so $F(l) \sim l^{1/2}$ (as expected for a *normal* random walk). (ii) If there were a local correlation extending up to a characteristic range R (such as in Markov chains), then $C(l) \sim \exp(-l/R)$; nonetheless *the asymptotic behavior* $F(l) \sim l^{1/2}$ *would be unchanged from the*

purely random case. (iii) If there is no characteristic length (i.e., if the correlation were "infinite-range"), then the scaling property of $C(l)$ would not be exponential, but would most likely to be a power law function, and the fluctuations will also be described by a power law

$$F(l) \sim l^{\alpha} \tag{2b}$$

with $\alpha \neq 1/2$. Figure 2a shows a typical example of an intron-containing gene. It is immediately apparent that the DNA walk has a very jagged contour which we shall see corresponds to long-range correlations. Calculation of $F(l)$ for this gene is shown in Fig. 3a. The fact that the data are linear over three decades on this double logarithmic plot confirms that $F(l) \sim l^{\alpha}$. A least-squares fit produces a straight line with slope $\alpha = 0.67 \pm 0.01$.

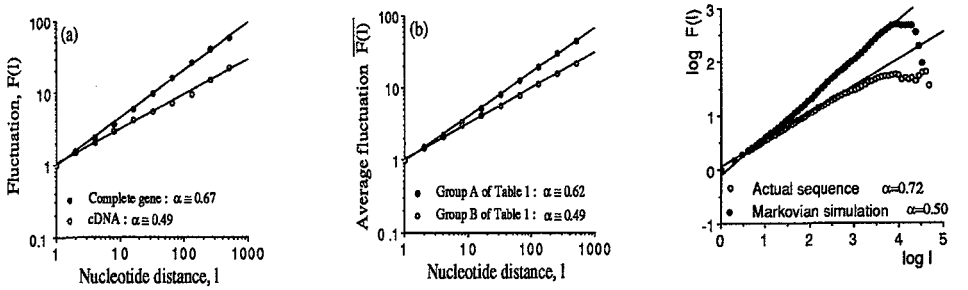

Fig. 3: Double logarithmic plots of (a) the mean square fluctuation function $F(l)$ as a function of the linear distance l along the DNA chain for the human β-cardiac myosin heavy chain gene and its cDNA, and (b) the average of $F(l)$ over the entries in Groups A and B of Table 1 of Ref. 5. The difference in the slopes is statistically significant, consistent with the possibility of long-range correlations in Group A and short-range correlations in Group B. For clarity, the data points shown are separated by intervals of 2^k. Part (c) demonstrates a correlation of Opeven longer range for the intron-containing human beta globin chromosomal region (73,376 base pairs). Shown for comparison is the "1-step" standard Markov chain analysis of the *same* base pair sequence, displaying the expected exponent $\alpha = 1/2$.

Figure 2b shows the DNA walk for the cDNA sequence of this same gene, while Fig. 2c displays the data for a typical intron-less sequence. In contrast to Fig. 2a, these intron-free sequences have less jagged contours, suggesting a shorter range correlation. To analyze parts (b) and (c), we first observe that for almost all intron-less sequences that we studied, purine-rich regions (compared to the average concentration over the entire strand) alternate with pyrimidine-rich regions, corresponding to the "up-hill" and "down-hill" portions of the DNA walk. To take into account the fact that the concentrations of purines and pyrimidines are not constant throughout the single strand base pair sequence, each DNA walk representation is partitioned into three segments demarcated by the global maximum ("max") and minimum ("min") displacements and then we analyze the fluctuation within each segment. Figure 3a also shows the data for the cDNA, and a least-squares fit gives a straight line with slope $\alpha = 0.49 \pm 0.01$.

5. Universality of Long Range Correlations in Base Pair Sequences.

In order to see if this scaling behavior is universal, we applied our analysis to 120 representative genomic and cDNA sequences across the phylogenetic spectrum, some of which are shown in Table 1 of Ref. 5. The myosin heavy chain family is particularly useful because it encompasses a number of long genomic and cDNA sequences for species ranging from yeast to humans. In

addition, we analyzed other sequences encoding a variety of other proteins as well as regulatory DNA sequences. The results show that long-range correlations ($\alpha > 1/2$) are characteristic of intron-containing genes and non-transcribed genomic regulatory elements (Group A). Thus, the average value ($\pm$2SEM) of α for the first 14 entries of Table 1 is 0.61 ± 0.03. In contrast, for cDNA sequences and genes without introns, $\alpha \cong 0.50 \pm 0.05$ indicating no long-range correlation (Group B). In keeping with this, the average value of α for the last 10 entries of Table 1 is 0.50 ± 0.01. This significant difference in the value of α for the two groups of base pair sequences is further shown in Fig. 3b, where the actual fluctuations over the two sets of entries in Table 1 are averaged. Thus, the calculation of $F(l)$ for the DNA walk representation provides a new, *quantitative* method to distinguish genes with multiple introns from intron-less genes and cDNAs based solely on their statistical properties.

To confirm that the base pair correlations are truly long-range, we now discuss the breakdown of linearity that must ultimately occur in graphs such as Figs. 3a and 3b. The reason we do not show the graphs for distances larger than 1000 is that the statistical error in $F(l)$ increases. Indeed, the "fall-off" in the straight line behavior is typical of all fractal analysis and is also found for artificial sequences of correlated numbers. We found that sequences with more base pairs possess still longer regions of linear (power law) behavior than the three decades of linearity shown in Figs. 3a and 3b. For example, Fig. 3c is the correlation graph for the human beta globin chromosomal region, which has 73,376 base pairs; note that the linearity extends to roughly 7000 base pairs. For comparison, Fig. 3c also shows the standard Markov chain analysis, which corroborates the classical result $\alpha = 1/2$. The long-range correlations were found by an NIH team [8] to extend over the entire yeast chromosome III region (315,357 base pairs). Finite size effects on the correlation exponent α are discussed in Ref. 9.

Thus we have introduced a new method to display correlations in the sequence of base pairs, and define a quantitative measure of the degree of correlation which is derived from a random walk representation. We found that the base pair sequence in intron-containing genes is highly correlated, and that the correlation is remarkably long range—indeed, base pairs *thousands of base pairs* distant are correlated. Moreover, the quantitative scaling of the correlation is of the power law form observed in numerous phenomena having a self-similar or "fractal" origin. Independent reports of long-range correlations in DNA have supported our findings [8,10,11]. Finally, we note that such long-range correlations are generally associated with the existence of a *non-equilibrium* dynamic process [2,12–14]. Interestingly, cDNAs do not exhibit this property and appear to exist in an *equilibrium* state. Our finding of long-range correlations in intron-containing genes appears to be independent of the particular gene or the encoded protein. It is observed in genes as disparate as myosin heavy chain, beta globin and adenovirus (Table 1 of Ref. 5).

6. Insertion model of genome organization

Although the correlation is long-range in the non-coding sequences, there seems to be a paradox: *long uncorrelated regions of up to thousands of base-pairs can be found in such sequences as well.* For example, consider the human beta-globin intergenomic sequence of length $L = 73,326$ (GenBank name: HUMHBB). This long non-coding sequence has 50% purines (no *overall* strand bias) and $\alpha = 0.7$. However, from base pair #67,089 to #73,228, there occurs the LINE-1 region [13]. In this region of length 6139 base pairs, there is a strong strand bias with 59% *purines*. In this non-coding sub-region, we found power-law scaling of F, with $F \sim l^{\alpha}$, with $\alpha = 0.55$, quite close to that of a random walk.

Even more striking is another region of 6378 base pairs, from base pair #23,137 to #29,515, which has 59% *pyrimidines* and is *uncorrelated*, with remarkably good power-law scaling and correlation exponent $\alpha = 0.49$. This region actually consists of three sub-sequences, complementary to shorter parts of the LINE-1 sequence.

These features motivated us to develop an "insertion model" based on the generalized Lévy

walk model for the non-coding regions of DNA sequences [15]. We show in the following paragraphs how this model can explain the long-range correlation properties, since there is no characteristic scale "built into" this insertion model. In addition, the model simultaneously accounts for the observed large sub-regions of non-correlated sequences within these non-coding DNA chains.

The classic Lévy walk model describes a wide variety of diverse phenomena that exhibit long-range correlations [3,16–22]. The model is defined schematically in Fig. 4a: A random walker takes not one but l_1 steps in a given direction. Then the walker takes l_2 steps in a new randomly-chosen direction, and so forth. The lengths l_j of each string are chosen from a probability distribution, with

$$P(l_j) \propto (1/l_j)^{\mu}, \tag{3}$$

where $\sum_{i=1}^{N} l_i = L$, N is the number of sub-strings and L is the total number of steps that the random walker takes.

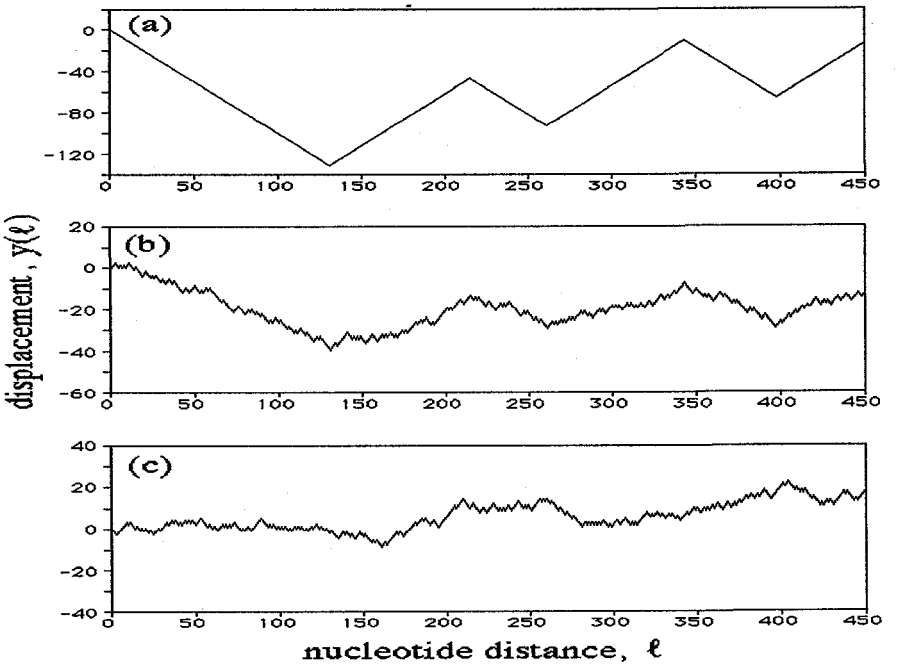

Fig. 4: Displacement $y(l)$ vs. number of steps for (a) the classical Lévy walk model consisting of 6 strings of l_j steps, each taken in alternating directions; (b) the insertion model consisting of 6 biased random walks of the same length with a probability of p_+ that it will go up equal to $(1 \pm \epsilon)/2$ [$\epsilon = 0.2$]; and (c) the unbiased uncorrelated random walk. Note that the vertical scale in (b) and (c) is twice that in (a).

We consider a generalization of the Lévy walk [22] to interpret recent findings of long-range correlation in non-coding DNA sequences described above. Instead of taking l_j steps in the *same* direction as occurs in a classic Lévy walk, the walker takes each of l_j steps in *random* directions, with a fixed bias probability

$$p_+ = (1 + \epsilon_j)/2 \tag{4a}$$

to go up and

$$p_- = (1 - \epsilon_j)/2 \tag{4b}$$

to go down, where ϵ_j gets the values $+\epsilon$ or $-\epsilon$ randomly. Here $0 \leq \epsilon \leq 1$ is a bias parameter (the case $\epsilon = 1$ reduces to the Lévy walk). Figure 4b shows such a generalized Lévy walk for the same choice of l_j as in Fig. 4a.

The generalized Lévy walk—like the pure Lévy walk—gives rise to a landscape with a fluctuation exponent α that depends upon the Lévy walk parameter μ [21,22],

$$\alpha = \begin{cases} 1 & \mu \leq 2 \\ 2 - \mu/2 & 2 < \mu < 3 \\ 1/2 & \mu \geq 3, \end{cases} \tag{5}$$

i.e., non-trivial behavior of α corresponds to the case $2 < \mu < 3$ where the first moment of $P(l_j)$ converges while the second moment diverges. The long-range correlation property for the Lévy walk, in this case, is a consequence of the broad distribution of Eq. (3) that lacks of a characteristic length scale. However, for $\mu \geq 3$, the distribution of $P(l_j)$ decays fast enough that an effective characteristic length scale appears. Therefore, the resulting Lévy walk behaves like a normal random walk for $\mu \geq 3$.

To be precise, we define our insertion walk model as follows [15]:

1. Choose a random number u which is uniformly distributed between 0 and 1, and define $l_j \equiv l_c u^{\mu-1}$ where l_c is some lower cutoff characteristic length. The number (l_j) thus generated will obey the distribution of Eq. (3).
2. Produce a biased random walk of length l_j with p_+ and p_- given by Eq. (4), where ϵ_j takes on the value $+\epsilon$ or $-\epsilon$ randomly and ϵ is a fixed value close to 0.2 (corresponding to the percentage of purines vs. pyrimidines in real DNA sequences)
3. Iterate the process, attaching together biased random walk until the total length of the sequence reaches a given value L.

To test the insertion model [15], we have adjusted the two parameters, μ and l_c, to best approximate features of an actual DNA sequence (the human beta-globin DNA sequence). The parameter ϵ can be determined from the strand biased regions in the actual DNA sequences, while l_c can be estimated from the length of the largest strand biased region in an actual DNA sequence of length L.

7. Relation of Long-Range Correlations to Genome Organization and Structure

Of interest, DNA walks of a variety of genes and intergenomic sequences with well expressed scale-invariant correlations reveal regions of "strand bias," i.e., regions with an excess of one kind of base pairs over the other [4,21,22]. These sub-regions (when analyzed separately from the rest of the sequence) show no long-range correlations and thus resemble coding sequences of cDNA.

From a biological viewpoint, two questions immediately arise: (a) *What is the significance of these uncorrelated sub-regions of strand bias?* and (b) *What is the molecular basis underlying the power-law statistics of the insertion model?*

With respect to the first question, we note that these long uncorrelated regions at least sometimes correspond to well-described but poorly understood sequences termed "repetitive elements," such as the LINE-1 region [23–25]. There are at least 53 different families of such repetitive elements within the human genome [26]. The lengths of these repetitive elements vary from 10 to 10^4 base pairs [26]. At least some of the repetitive elements are believed to be remnants of messenger RNA molecules that formerly did code for proteins [25–27]. Alternatively, these segments may represent retroviral sequences that have inserted themselves into the genome [28]. Our finding

that these repetitive elements have the statistical properties of biased random walks (e.g., the same as that of active coding sequences) is consistent with these hypotheses.

With respect to the second question, we note that from the point of view of modern polymer theory [6,7] the power law distribution of these uncorrelated subregions can be related to the known power-law statistics of the loops formed by a long polymer chain in a solvent. The pair of monomers from two distant parts of the polymer chain may sometimes approach to each other when the chain bends in the solvent forming a loop of a certain length. The probability of such an event is known to be approximately inverse-proportional to a square of this loop length [7,29]. Any splicing or insertion in the polymer chain can happen—without breaking its connectivity—if it is proceeded by formation of a loop that later can be removed or inserted. This underlying physical principle may govern the probability distribution of the observed heterogeneous regions of DNA that were inserted into genome in the course of evolution [30–38].

8. General significance

Our observations suggest that the application of fractal analysis to DNA sequences reveals a property that may be responsible for a variety of additional "informational codes" in the DNA sequence. Furthermore, discovering the origin of long-range correlations may lead to better understanding of the role of introns and intergenomic sequences. Our results may shed some light on the following aspects of DNA sequences.

(i) *Dynamical DNA Processes:*

One possible explanation of the origin of the long range correlations is the evolutionary changes in DNA associated with insertion of introns, transpositions, exon shuffling and gene duplication. The current stage of genomic sequences is a product of a dynamical process which involves mutation, deletion, insertion and other rearrangements [30–34]. This dynamical process is presumably driven by the need for more complex functionality and more efficient utilization of the sequences. Our preliminary results suggest that analysis of long-range correlations may provide new understanding of how this dynamical process proceeds.

(ii) *New organization principles for gene/chromosome:*

The long-range correlations may also come from some physical interaction in the DNA chain. This interaction could be local in nature and manifest itself in long-range correlations via 3-d packaging. In this case, our understanding of long-range correlations may give useful clues to 3-d structure of genomic sequences. Power-law behavior reflects a *scale-invariant property* of DNA which might be related to hierarchical order DNA/chromatin structure, DNA bending or looping [37,38].

(iii) *Relationship of coding/non-coding sequences and automated criteria for distinguishing them:*

To date there is little knowledge about the information content of introns and intergenomic sequences. However, our analysis shows a systematic statistical difference between coding and non-coding sequences. Further investigation may lead to understanding the role of these non-coding sequences. As a practical matter, our analysis can contribute to the development of software (complementary to other existing programs) and a database for better identification of coding and non-coding sequences. Such a database could help to develop a neural network system that will learn to identify different sequences in DNA, such as coding and non-coding parts and repetitive elements.

9. Acknowledgements

We wish to thank C. Cantor, C. DeLisi, J. Hausdorff, R. D. Rosenberg, J. J. Schwartz, M. Schwartz, and N. Shworak for valuable discussions. Partial support was provided to ALG by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute and the National Aeronautics and Space Administration, to MS by the American Heart Association, and to HES by the National Science Foundation and Office of Naval Research.

[1] Tavaré, S. and Giddings, B. W., "Some Statistical Aspects of the Primary Structure of Nucleotide Sequences," in *Mathematical Methods for DNA Sequences*, eds. M. S. Waterman (CRC Press, Boca Raton, 1989), pp. 117–132, and refs. therein.

[2] Montroll, E. W. and Shlesinger, M. F., "The Wonderful World of Random Walks," in *Nonequilibrium Phenomena II. From Stochastics to Hydrodynamics*, eds. J. L. Lebowitz and E. W. Montroll (North-Holland, Amsterdam, 1984), pp. 1-121.

[3] Stanley, H. E. and Ostrowsky, N., eds., *On Growth and Form: Fractal and Non-Fractal Pattern in Physics* (Martinus Nijhoff Publishers, Dordrecht, 1986).

[4] Bunde, A. and Havlin, S., eds., *Fractals and Disordered Systems* (Springer-Verlag, Berlin, 1991).

[5] Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E., "Long-Range Correlations in Nucleotide Sequences," *Nature* **356** (1992) 168.

[6] de Gennes, P. G., *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca NY, 1979).

[7] des Cloiseaux, J., "Short-Range Correlation between Elements of a Long Polymer in a Good Solvent," *J. Physique (Paris)* **41** (1980) 223.

[8] Munson, P. J., Taylor, R. C., and Michaels, G. S., "DNA Correlations," *Nature* **360** (1992) 636.

[9] Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Simons, M., and Stanley, H. E., "Finite Size Effects on Long-Range Correlations: Implications for Analyzing DNA Sequences," Phys. Rev. E **47** (1993) 3729.

[10] Li, W. and Kaneko, K., "Long-Range Correlations and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence," *Europhys. Lett.* **17** (1992) 655.

[11] Voss, R., "Evolution of Long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences" *Phys. Rev. Lett.* **68** (1992) 3805.

[12] Dutta, P. and Horn, P. M., "Low-Frequency Fluctuations in Solids: 1/f Noise," *Rev. Mod. Phys.* **53** (1981) 497.

[13] Bak, P., Tang, C., and Wiesenfeld, K., "Self-Organized Criticality: An Explanation of 1/f Noise," *Phys. Rev. Lett.* **59** (1987) 381.

[14] Shlesinger, M. F., "Fractal Time in Condensed Matter," *Ann. Rev. Phys. Chem.* **39** (1988) 269.

[15] 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) xxx.

[16] Havlin, S., Buldyrev, S., Stanley, H. E., and Weiss, G. H., "Probability Distribution of the Interface Width in Surface Roughening: Analogy with a Lévy Flight," *J. Phys. A* **24** (1991) L925.

[17] Mantegna, R. N., "Lévy Walks and Enhanced Diffusion in Milan Stock Exchange," *Physica A* **179** (1991) 232.

[18] Peng, C. K., Mietus, J., Hausdorff, J., Havlin, S., Stanley, H. E., and Goldberger, A. L., "Long-Range Anti-Correlations and Non-Gaussian Behavior of the Heartbeat," Phys. Rev. Lett. **70** (1993) 1343.

[19] Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E., "Fractal Landscape Analysis of DNA Walks," Physica A **191** (1992) 25.

[20] Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Hausdorff, J. M., Havlin, S., Mietus, J., Peng, C. K., Sciortino, F., and Simons, M., "Fractal Landscapes in Biological Systems: Long-range Correlations in DNA and interbeat heart intervals," Physica A **191** (1992) 1.

[21] Shlesinger, M. F., Klafter, J., and Wong, Y. M., "Random Walks with Infinite Spatial and Temporal Moments," *J. Stat. Phys.* **27** (1982) 499.

[22] Araujo, M., Havlin, S., Weiss, G. H., and Stanley, H. E., "Diffusion of Walkers with Persistent Velocities," *Phys. Rev. A* **43** (1991) 5207.

[23] Jurka, J., "Subfamily Structure and Evolution of the Human L1 Family of Repetitive Sequences," *J. Mol. Evol.* **29** (1989) 496.

[24] Watson, J. D., Gilman, M., Witkowski, J., and Zoller, M., *Recombinant DNA* (Scientific American Books, New York, 1992).

[25] Hwu, H. R., Roberts, J. W., Davidson, E. H., and Britten, R. J., "Insertion and/or Deletion of Many Repeated DNA Sequences in Human and Higher Ape Evolution," *Proc. Nat'l Acad. Sci. USA* **83** (1986) 3875.

[26] Jurka, J., Walichiewicz, T., and Milosavljevic, A., "Prototypic Sequences for Human Repetitive DNA," *J. Mol. Evol.* **35** (1992) 286.

[27] Zuckerkandl, E., Latter, G., and Jurka, J., "Maintenance of Function without Selection: *Alu* Sequences as 'Cheap Genes'," *J. Mol. Evol.* **29** (1989) 504.

[28] Levin, B., *Genes IV* (Oxford University Press, Oxford, 1990).

[29] Redner, S., "Distribution Functions in the Interior of Polymer Chains," *J. Phys. A* **13** (1980) 3525.

[30] Doolittle, W. F., "Understanding Introns: Origins and Functions" in *Intervening Sequences in Evolution and Development*, eds. E. Stone and R. Schwartz (Oxford University Press, New York, 1990), pp. 42–62.

[31] Gilbert, W., "Why Genes in Pieces?" *Nature* **271** (1978) 501.

[32] Darnell, J. E., Jr., "Implications of RNA: RNA Splicing in Evolution of Eukaryotic Cells," *Science* **202** (1978) 1257.

[33] Doolittle, W. F., "Genes in Pieces: Were They Ever Together?" *Nature* **272** (1978) 581.

[34] Li, W.-H. and Graur, D., *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland MA, 1991).

[35] Beckmann, J. S. and Trifonov, E. N., "Splice Junctions Follow a 250-Base Ladder," *Proc. Natl. Acad. Sci. USA* **88** (1991) 2380.

[36] Ioshikhes, I., Bolshoy, A., and Trifonov, E. N., "Preferred Positions of AA-Dinucleotides and TT-Dinucleotides in Aligned Nucleosomal DNA Sequences," *J. Biomolec. Struct. & Dynamics* **9** (1992) 1111.

[37] Schleif, R., "DNA-Looping Perspective," *Science* **240** (1988) 127.

[38] Hagerman, P. J., "Sequence-Directed Curvature of DNA," *Ann. Rev. Biochem.* **59** (1990) 755.