Long-Range Correlations and Generalized Lévy Walks in DNA Sequences

H. E. Stanley,[1] S. V. Buldyrev,[1] A. L. Goldberger,[3,4] S. Havlin,[1,2]
R. N. Mantegna,[1,5] C.-K. Peng,[1,3] M. Simons[3] and M. H. R. Stanley[1]

[1]Center for Polymer Studies and Department of Physics, Boston University, Boston, MA

[2]Department of Physics, Bar Ilan University, Ramat Gan, ISRAEL

[3]Cardiovascular Div., Harvard Medical School, Beth Israel Hospital, Boston, MA

[4]Department of Biomedical Engineering, Boston University, Boston, MA

[5]Dipartimento di Energetica ed Applicazioni di Fisica, Palermo University, Palermo, I-90128, Italy

1 Long-Range Power-Law Correlations

In recent years long-range power-law correlations have been discovered in a remarkably wide variety of systems. Such long-range power-law correlations are a physical fact that in turn gives rise to the increasingly appreciated "fractal geometry of nature" [1–12]. So if fractals are indeed so widespread, it makes sense to anticipate that long-range power-law correlations may be similarly widespread. Indeed, recognizing the ubiquity of long-range power-law correlations can help us in our efforts to understand nature, since as soon as we find power-law correlations we can quantify them with a critical exponent. Quantification of this kind of scaling behavior for apparently unrelated systems allows us to recognize similarities between different systems, leading to underlying unifications that might otherwise have gone unnoticed.

Traditionally, investigators in many fields characterize processes by assuming that correlations decay exponentially. However, there is one major exception: at the critical point, the exponential decay turns into to a power law decay [13]

$$C_r \sim (1/r)^{d-2+\eta}. \tag{1}$$

Many systems drive themselves spontaneously toward critical points [2, 14]. One of the simplest models exhibiting such "self-organized criticality" is invasion percolation, a generic model that has recently found applicability to describing anomalous behavior of rough interfaces.

In the following sections we will attempt to summarize some recent findings [15–35] concerning the possibility 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 this discovery has intriguing implications for molecular evolution [32], as well as potential practical applications for distinguishing coding and noncoding regions in long nucleotide chains [34]. It also may be related to the presence of a language in noncoding DNA [36].

2 DNA

The role of genomic DNA sequences in coding for protein structure is well known [37]. The human genome contains information for approximately 100,000 different proteins, which define all inheritable features of an individual. The genomic sequence is likely the most sophisticated information database created by nature through the dynamic process of evolution. Equally remarkable is the precise transformation of information (duplication, decoding, etc) that occurs in a relatively short time interval.

The building blocks for coding this information are called *nucleotides*. Each nucleotide contains a phosphate group, a deoxyribose sugar moiety and either a *purine* or a *pyrimidine base*. Two purines and two pyrimidines are found in DNA. The two purines are adenine (A) and guanine (G); the two pyrimidines are cytosine (C) and thymine (T). The nucleotides are linked end to end, by chemical bonds from the phosphate group of one nucleotide to the deoxyribose sugar group of the adjacent nucleotide, forming a long polymer (*polynucleotide*) chain. The information content is encoded in the sequential order of the bases on this chain. Therefore, as far as the information content is concerned, a DNA sequence can be most simply represented as a symbolic sequence of four letters: A, C, G and T.

In the genomes of high eukaryotic organisms only a small portion of the total genome length is used for protein coding (as low as 3% in the human genome). The segments of the chromosomal DNA that are spliced out during the formation of a mature mRNA are called *introns* (for intervening sequences). The coding sequences are called *exons* (for expressive sequences).

The role of introns and intergenomic sequences constituting large portions of the genome remains unknown. Furthermore, only a few quantitative methods are currently available for analyzing information which is possibly encrypted in the noncoding part of the genome.

3 The "DNA Walk"

One interesting question that may be asked by statistical physicists would be whether the sequence of the nucleotides A,C,G, and T behaves like a one-dimensional "ideal gas", where the fluctuations of density of certain particles obey Gaussian law, or if there exist long range correlations in nucleotide content (as in the vicinity of a critical point). These result in domains of all size with different nucleotide concentrations. Such domains of various sizes were known for a long time but their origin and statistical properties remain unexplained. A natural language to describe heterogeneous DNA structure is long-range correlation analysis, borrowed from the theory of critical phenomena [13].

In order to study the scale-invariant long-range correlations of a DNA sequence, we first introduced a graphical representation of DNA sequences, which we term a *fractal landscape* or *DNA walk* [15]. For the conventional one-dimensional random walk model [38, 39], a walker moves either "up" $[u(i) = +1]$

or "down" [$u(i) = -1$] one unit length for each step i of the walk. 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 [40–42].

One definition of the DNA walk is that the walker steps "up" if a pyrimidine (C or T) occurs at position i along the DNA chain, while the walker steps "down" if a purine (A or G) occurs at position i. 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).

There have also been attempts to map DNA sequence onto multi-dimensional DNA walks [16, 43]. However, recent work [34] indicates that the original purine-pyrimidine rule provides the most robust results, probably due to the purine-pyrimidine chemical complementarity.

The DNA walk allows one to visualize directly the fluctuations of the purine-pyrimidine content in DNA sequences: Positive slopes correspond to high concentration of pyrimidines, while negative slopes correspond to high concentration of purines. Visual observation of DNA walks suggests that the coding sequences and intron-containing noncoding sequences have quite different landscapes.

4 Correlations and Fluctuations

An important statistical quantity characterizing any walk [38, 39] is the root mean square fluctuation $F(\ell)$ about the average of the displacement of a quantity $\Delta y(\ell)$ defined by $\Delta y(\ell) \equiv y(\ell_0+\ell) - y(\ell_0)$. If there is no characteristic length (i.e., if the correlation were "infinite-range"), then fluctuations will also be described by a power law

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

with $\alpha \neq 1/2$.

Figure 1a shows a typical example of a gene that contains a significant fraction of base pairs that do *not* code for amino acids. It is immediately apparent that the DNA walk has an extremely jagged contour which corresponds to long-range correlations.

The fact that data for intron-containing and intergenic (i.e., noncoding) sequences are linear on this double logarithmic plot confirms that $F(\ell) \sim \ell^{\alpha}$. A least-squares fit produces a straight line with slope α substantially larger than the prediction for an uncorrelated walk, $\alpha = 1/2$, thus providing direct experimental evidence for the presence of long-range correlations.

On the other hand, the dependence of $F(\ell)$ for coding sequences is not linear on the log-log plot: its slope undergoes a crossover from 0.5 for small ℓ to 1 for large ℓ. However, if a single patch is analyzed separately, the log-log plot of $F(\ell)$ is again a straight line with the slope close to 0.5. This suggests that within a large patch the coding sequence is almost uncorrelated.

Figure 1: DNA walk displacement $y(\ell)$ (excess of purines over pyrimidines) vs nucleotide distance ℓ for (a) HUMHBB (human beta globin chromosomal region of the total length $L = 73,239$); (b) the LINE1c region of HUMHBB starting from 23,137 to 29,515; (c) the generalized Lévy walk model of length 73,326 with $\mu = 2.45$, $l_c = 10$, $\alpha_o = 0.6$, and $\epsilon = 0.2$; and (d) a segment of a Lévy walk of exactly the same length as the LINE1c sequence from step 67,048 to the end of the sequence. This sub-segment is a Markovian random walk. Note that in all cases the overall bias was subtracted from the graph such that the beginning and ending points have the same vertical displacement ($y = 0$). This was done to make the graphs clearer and does not affect the quantitative analysis of the data.

5 Lévy Walk Model and its Generalization

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 $\alpha = 0.7$ (see Fig. 1(a). However, from nucleotide #67,089 to #73,228, there occurs the LINE-1 region (defined in Ref. [44]. In this region of length 6139 base pairs, there is a strong strand bias with 59% *purines*. In this non-coding sub-region, we find 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 nucleotide #23,137 to #29,515, which has 59% *pyrimidines* and is *uncorrelated*, with remarkably good power-law scaling and correlation exponent $\alpha = 0.49$ (Fig. 1(b)). This region actually consists of three sub-sequences, complementary to shorter parts of the LINE-1 sequence.

Figure 2: Displacement $y(\ell)$ 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 generalized Lévy walk 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).

These features motivated us to apply a generalized Lévy walk model (see Figs. 1c, 1d and 2) for the non-coding regions of DNA sequences [30]. We will show in the next section how this model can explain the long-range correlation properties, since there is no characteristic scale "built into" this generalized Lévy walk. 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 [45–48]. The model is defined schematically in Fig. 2a: 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

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

Figure 3: The actual DNA sequences are presented in (a) and (c): the entire HUMHBB sequence (◯) and LINE1c sequence (□). The slopes for the linear fits are 0.72 and 0.49 respectively. The Lévy model sequences are presented in (b) and (d): the entire Lévy walk sequence of Fig. 1c (◯), a segment of this walk of Fig. 1d (□). The slopes for the linear fits are 0.73 and 0.49 respectively.

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.

We consider a generalization of the Lévy walk [42] 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

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

to go up and

$$p_- = (1 - \epsilon_j)/2 \tag{5}$$

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). Fig. 2b shows such a generalized Lévy walk for the same choice of l_j as in Fig. 2a.

As shown in Ref. [30], the generalized Lévy walk—like the pure Lévy walk—gives rise to a landscape with a fluctuation exponent α that depends upon the

Figure 4: Comparison of successive slopes of the scaling exponent α for yeast chromosome III ($\square$) and (a) successive slopes of a realization of the generalized Lévy walk with parameters $L = 315,000$, $\mu = 2.5$, $l_c = 5$, $\alpha_o = 0.55$, $\epsilon = 0.16$ ($\bigcirc$); (b) average successive slopes over 15 different realization of Lévy walks with the same parameters (dashed line). The shaded area corresponds to two standard deviations of successive slopes of the model, calculated for 15 random realizations. The parameters for Markov process, α_o and ϵ, used in the model are calculated from real DNA sequence of yeast chromosome III.

Lévy walk parameter μ [42, 46],

$$\alpha = \begin{cases} 1 & \mu \leq 2 \\ 2 - \mu/2 & 2 < \mu < 3 \\ 1/2 & \mu \geq 3. \end{cases} \tag{6}$$

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 generalized L-step Lévy walk model as follows:

(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 (see Ref. [30]) with p_+ and p_- given by Eqs. (4) and (5), 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)

Figure 5: The probability distribution to find a run of certain size s of purines or pyrimidines in the coarse-grained sequence calculated using coarse-grained window size equal to 32. (a) Actual sequence of HUMHBB on log-log plot. (b) Actual sequence of HUMHBB on semi-log plot. (c) Log-log plot for the model sequence, shown in Fig. 1c. (d) Semi-log plot for the model.

(3) Iterate the process, attaching together biased random walk until the total length of the sequence reaches a given value L.

6 Comparison with DNA Data

To test the generalized Lévy walk model, we have adjusted the two parameters, μ and l_c, described in the previous section to best approximate features of an actual DNA sequence the human beta-globin DNA sequence shown in Fig. 1(a). The resulting landscape for the generalized Lévy walk model is presented in Fig. 1(c). The comparison of $F(\ell)$ for the model and DNA sequences is shown in Figs. 3a and 3b.

A more detailed scaling analysis (Figs. 3c, 3d), considers the "local slopes" of successive points in the graphs of Figs. 3a, 3b

$$\alpha(\ell_i, L) \equiv \frac{\log F(\ell_{i+1}, L) - \log F(\ell_i, L)}{\log \ell_{i+1} - \log \ell_i}, \tag{7}$$

where ℓ_{i+1} and ℓ_i are values of two subsequent data points. The local slope changes from $\alpha = 0.6$ for $\ell_1 = 1$ to $\alpha = 0.75$ for $\ell_i = 128$, and stays at this value for about two decades. It eventually drops down when ℓ_i becomes too close to L, since $F(L, L) \equiv 0$ according to Eq. (2.1). This kind of scaling behavior is general for all kinds of DNA sequences that contain non-coding material. The initial monotonic increase in α, however, does not mean that long-range correlations do not exist. Indeed, as seen in Fig. 3d, a similar type of behavior exists in the generalized Lévy walk model. Equation (6) is valid asymptotically for very large ℓ and L and the local value of $\alpha(\ell, L)$ for finite values of ℓ and L may differ considerably from its asymptotic value. The comparison of $\alpha(\ell, L)$ plots for human beta-globin chromosomal region ($L = 73326$) and a Lévy walk model of the same size is made for one of the largest available ($L = 315357$) DNA sequences [20, 49], that of Yeast chromosome III (see Fig. 4a).

For any given size L, it is possible to calculate the average value and standard deviation of $\alpha(\ell, L)$ for the Lévy walk model by calculating $\alpha(\ell, L)$ for a large number k of statistically independent realizations of the model sequence of the size L. The data for yeast chromosome III is well within a 2 standard deviation interval ($k = 15$) for the generalized Lévy walk model with $\mu = 2.5$, which corresponds to observed value of $\alpha = 0.75$ (Fig. 4b).

An alternative test of Lévy walk structure can be made if one analyzes a "coarse-grained" version of the original DNA sequence. To this end, we (i) divide the entire sequence into L/w sub-sequences of equal length w, (ii) replace each sub-sequence by 1 if there is an excess of purines or by 0 if there is an excess of pyrimidines and, (iii) calculate the distribution $P(s)$ of sizes s of long runs of ones and zeroes. These calculations for human beta-globin chromosomal region show that $P(s)$ has a scaling region of roughly one decade, where $P(s) \sim s^{-\mu}$ with $\mu \approx 2.5$. Our results are in good agreement with the value of the exponent $\alpha = 0.75$ (see Fig. 5). Unfortunately, the coarse-graining process requires a long sequence ($> 10^5$ nucleotides) in order that the statistics for the distribution be meaningful. To date, only a few documented long sequences are available, but as longer sequences become available this renormalization test should prove to be increasingly useful.

7 "Mosaic" Nature of DNA

The key finding of this analysis is that a generalized Lévy walk model can account for two hitherto unexplained features of DNA nucleotides: (i) the long-range power law correlations that extend over thousands of nucleotides in sequences containing non-coding regions (e.g., genes with introns and intergenomic sequences), and (ii) the presence within these correlated sequences of sometimes large sub-regions that correspond to biased random walks. This apparent paradox is resolved by the generalized Lévy walk, a mechanism for generating long-range correlations (no characteristic length scale), that with finite (though rare) probability also generates large regions of uncorrelated strand bias. The uncorrelated sub-regions, therefore, are an anticipated feature of this mechanism for

long-range correlations.

¿From a biological viewpoint, two questions immediately arise: (i) What is the significance of these uncorrelated sub-regions of strand bias? and (ii) What is the molecular basis underlying the power-law statistics of the Lévy walk? 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 LINE1 region noted above [44, 50]. There are at least 53 different families of such repetitive elements within the human genome. The lengths of these repetitive elements vary from 10 to 10^4 nucleotides [44]. At least some of the repetitive elements are believed to be remnants of messenger RNA molecules that formerly did code for proteins [50, 51, 52]. Alternatively, these segments may represent retroviral sequences that have inserted themselves into the genome [53]. 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.

Finally, what are the biological implications of this type of analysis? Our findings clearly support the following possible hypothesis concerning the molecular basis for the power-law distributions of elements within DNA chains. In order to be inserted into DNA, a macromolecule should form a loop of certain length l with two ends, separated by l nucleotides along the sequence, coming close to each other in real space. The probability of finding a loop of length l inside a very long linear polymer scales as $l^{-\mu}$ [54, 55]. Theoretical estimates of μ made by different methods [55–58] using a self-avoiding random walk model [54] indicate that the value of μ for three-dimensional model is between 2.16 and 2.42. Our estimate made by the Rosenbluth Monte-Carlo Method [59] gave $\mu = 2.22 \pm 0.05$ which yields according to Eq. (6) $\alpha = 0.89$, a larger value than the effective value of $\alpha(\ell, L)$, observed in DNA of finite length. However, the asymptotic value of the exponent α remains uncertain since the statistics of Lévy walks converge very slowly due to rare events associated with the very long strings of constant bias that may occur in the sequence according to Eq. (3). This results in the very large error bars for $\alpha(\ell, L)$ for large values of ℓ and finite length L (see Fig. 4). Even for the sequences of about 300K base-pairs we cannot estimate the limiting value of α with good accuracy. It is clear, however, that the behavior of DNA sequences cannot be satisfactorily explained in terms of only one characteristic length scale even of about $10^3 - 10^4$ base pairs long. The asymptotic behavior of the scaling exponent α and whether it reaches some universal value for long DNA chains must await further data from the Human Genome Project.

Recently, a report appeared that confirms the existence of long-range correlations in DNA [25]. However, where Ref. [25] might appear to disagree with Ref. [15] is in the interpretation of that finding for coding and non-coding regions. Both figures in [25] apply to the complete genome of the phage λ which does not contain non-coding sequences and consists of only three regions of different strand bias (see Fig. 1c of Ref. [15]. Each such region when analyzed separately by the DNA-walk method gives exponent $\alpha \approx 0.5$, close to that of

random walk. The combination of three such regions produces a crossover in the local values of $\alpha(l, L) \approx 0.5$ at small length scales l to $\alpha(l, L) \approx 1$ at large l. Thus, for coding sequences, there is indeed no well-defined scaling exponent α for large length-scales.

In contrast, the monotonically increasing local values of $\alpha(l, L)$ followed by a plateau at large l for non-coding sequences are completely explained by the generalized Levy walk model presented here in terms of a crossover from an uncorrelated random walk at small length scales to a Lévy walk at large length scales. The latter has well-defined scaling with an exponent α related to the exponent μ characterizing the power law distribution of steps of the Levy walk. Figure 3d of the present work clearly demonstrates that the generalized Levy walk model accounts for the upward curvature in the values of $\alpha(l, L)$, followed by a plateau with $\alpha(l, L) \approx 2 - \mu/2$ [17, 27].

8 "Linguistic" Analysis

Long-range correlations have been found recently in human writings [60]. A novel, a piece of music or a computer program can be regarded as a one-dimensional string of symbols. These strings can be mapped to a one-dimensional random walk model similar to the DNA walk allowing calculation of the correlation exponent α. Values of α between 0.6 and 0.9 were found for various texts.

An interesting hierarchical feature of languages was found in 1949 by Zipf [61]. He observed that the frequency of words as a function of the word order decays as a power law (with a power close to -1) for more than four orders of magnitude.

In order to adapt the Zipf analysis to DNA, the concept of word must first be defined. In the case of coding regions, the words are the 64 3-tuples ("triplets") which code for the amino acids, AAA, AAT, ... GGG. However for non-coding regions, the words are not known. Therefore Ref. [36] considers the word length n as a free parameter, and performs analyses not only for $n = 3$ but also for all values of n in the range 3 through 8. The different n-tuples are obtained for the DNA sequence by shifting progressively by 1 base a window of length n; hence, for a DNA sequence containing L base pairs, we obtain $L - n + 1$ different words.

The results of the Zipf analysis for all 40 DNA sequences analyzed are summarized in Ref. [36]. The averages for each category support the observation that ζ is consistently larger for the non-coding sequences, suggesting that the non-coding sequences bear more resemblance to a natural language than the coding sequences.

Related interesting statistical measures of short-range correlations in languages are the entropy and redundancy. The redundancy is a manifestation of the *flexibility* of the underlying code. To quantitatively characterize the redundancy implicit in the DNA sequence, we utilize the approach of Shannon, who provided a mathematically precise definition of redundancy [62, 63]. Shannon's redundancy is defined in terms of the entropy of a text—or, more precisely, the

Figure 6: A Zipf plot using the *Fortune 500* for 1993 [68]. On the y-axis of this log-log plot is the sales of a firm in 1985 dollars. On the x-axis is the rank of that firm. The straight line is fit to the first 100 firms. One can see that the first approximately 100 firms are well fit by a straight line, but after approximately firm rank 100, the plot is no longer a straight line. After [66].

"n-entropy"

$$H(n) = -\sum_{i=1}^{4^n} p_i log_2 p_i, \tag{8}$$

which is the entropy when the text is viewed as a collection of n-tuple words. The redundancy is defined through as $R \equiv \lim_{n\to\infty} R(n)$, where

$$R(n) \equiv 1 - H(n)/kn; \tag{9}$$

here $k = \log_2 4 = 2$.

Reference [36] calculates the Shannon n-entropy $H(n)$ for $n = 1, 2, \cdots 6$. The maximum value of n for which it is possible to determine $H(n)$ is $n = 6$—even for very long sequences (e.g., *C. elegans*, 2.2 million nucleotides)—due to the extremely slow convergence to the final value. For shorter sequences, reliable values of $H(n)$ are obtainable only up to a value of n less than 6.

For sufficiently high values of n (for example $n = 4$), we found that the redundancy is consistently larger for the primarily non-coding sequences. In fact, for most of the sequences consisting primarily of coding regions, we find

that $R(n)$ is quite close to the value $R(n) = 0$ which we find for a control sequence of random numbers.

In summary, Ref. [36] finds that *non-coding* sequences show two similar statistical properties to those of both natural and artificial languages: (a) Zipf-like scaling behavior, and (b) a non-zero value of Shannon's redundancy function $R(n)$. These results are consistent with the *possible* existence of one (or more than one) structured biological languages present in non-coding DNA sequences.

It appears that linearity of a Zipf plot is generally indicative of hierarchical ordering. For example, it is possible that a wide range of systems result in straight-line behavior when subjected to Zipf analysis [64]. An example that was the subject of some discussion at this meeting is the remarkable linearity of the Zipf plot giving the annual sales of a company as a function of its sales rank. J.P. Bouchaud [65] finds that this plot is linear for European companies, while M.H.R. Stanley [66] finds linearity for American companies (Fig. 6). Furthermore, M.H.R. Stanley et al. [67] find a significant deviation from this apparent linearity at rank ≈ 100, and relate this feature to the log-normal distribution of sales (the "Gibrat law").

9 Summary

There is a mounting body of evidence suggesting that the noncoding regions of DNA are rather special for at least two reasons:

1. They display long-range power-law correlations, as opposed to previously-believed exponentially-decaying correlations.

2. They display features common to hierarchically-structured languages—specifically, a linear Zipf plot and a non-zero redundancy.

These results are consistent with the possibility that the noncoding regions of DNA are not merely "junk" but rather have a purpose. What that purpose could be is the subject of ongoing investigation. In particular, the apparent increase of α with evolution [32] could provide insight.

In the event that the purpose is not profound, our results nonetheless may have important practical value since quantifiable differences between coding and noncoding regions of DNA can be used to help distinguish the coding regions [34].

10 Acknowledgements

We are grateful to many individuals, including M.E. Matsa, S.M. Ossadnik, M.A. Salinger, and F. Sciortino, for major contributions to those results reviewed here that represent collaborative research efforts. We also wish to thank C. Cantor, C. DeLisi, M. Frank-Kamenetskii, A.Yu. Grosberg, G. Huber, I. Labat, L. Liebovitch, G.S. Michaels, P. Munson, R. Nossal, R. Nussinov, R.D.

Rosenberg, J.J. Schwartz, M. Schwartz, E.I. Shakhnovich, M.F. Shlesinger, N. Shworak, and E.N. Trifonov for valuable discussions. Partial support was provided by the National Science Foundation, NIH, the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute, the National Aeronautics and Space Administration, the Israel-USA Binational Science Foundation, and (to C-KP) by an NIH/NIMH Postdoctoral NRSA Fellowship.

References

[1] B.B. Mandelbrot: *The Fractal Geometry of Nature* (W.H. Freeman, San Francisco 1982)

[2] A. Bunde, S. Havlin, eds.: *Fractals and Disordered Systems* (Springer-Verlag, Berlin 1991) A. Bunde, S. Havlin, eds.: *Fractals in Science* (Springer-Verlag, Berlin 1994); T. Vicsek, M. Shlesinger, M. Matsushita, eds.: *Fractals in Natural Sciences* (World Scientific, Singapore, 1994)

[3] J.M. Garcia-Ruiz, E. Louis, P. Meakin, L. Sander, eds.: *Growth Patterns in Physical Sciences and Biology* [Proc. 1991 NATO Advanced Research Workshop, Granada, Spain, October 1991], (Plenum, New York, 1993)

[4] A.Yu. Grosberg, A.R. Khokhlov: *Statistical Physics of Macromolecules*, translated by Y. A. Atanov (AIP Press, New York, 1994)

[5] J.B. Bassingthwaighte, L.S. Liebovitch, B.J. West: *Fractal Physiology* (Oxford University Press, New York, 1994)

[6] A.-L. Barabási, H.E. Stanley: *Fractal Concepts in Surface Growth* (Cambridge University Press, Cambridge, 1995)

[7] B.J. West, A.L. Goldberger: J. Appl. Physiol., **60**, 189 (1986); B.J. West, A.L. Goldberger: Am. Sci., **75**, 354 (1987); A.L. Goldberger, B.J. West: Yale J. Biol. Med. **60**, 421 (1987); A.L. Goldberger, D.R. Rigney, B.J. West: Sci. Am. **262**, 42 (1990); B.J. West, M.F. Shlesinger: Am. Sci. **78**, 40 (1990); B.J. West: *Fractal Physiology and Chaos in Medicine* (World Scientific, Singapore 1990); B.J. West, W. Deering: Phys. Reports **246**, 1 (1994); S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley: in *Fractals in Science*, edited by A. Bunde and S. Havlin (Springer-Verlag, Berlin, 1994), 49–83

[8] T. Vicsek: *Fractal Growth Phenomena, Second Edition* (World Scientific, Singapore 1992)

[9] J. Feder: *Fractals* (Plenum, NY, 1988)

[10] D. Stauffer, H.E. Stanley: *From Newton to Mandelbrot: A Primer in Theoretical Physics* (Springer-Verlag, Heidelberg & N.Y. 1990)

[11] E. Guyon, H.E. Stanley: *Les Formes Fractales* (Palais de la Découverte, Paris 1991); **English translation:** *Fractal Forms* (Elsevier North Holland, Amsterdam 1991)

[12] H.E. Stanley, N. Ostrowsky, eds.: *Random Fluctuations and Pattern Growth: Experiments and Models*, Proceedings 1988 Cargèse NATO ASI (Kluwer Academic Publishers, Dordrecht, 1988)

[13] H.E. Stanley: *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, London 1971)

[14] H.E. Stanley, N. Ostrowsky, eds.: *Correlations and Connectivity: Geometric Aspects of Physics, Chemistry and Biology*, Proceedings 1990 Cargèse Nato ASI, Series E: Applied Sciences (Kluwer, Dordrecht 1990)

[15] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley: Nature **356**, 168 (1992)

[16] W. Li, K. Kaneko: Europhys. Lett. **17**, 655 (1992)

[17] S. Nee: Nature **357**, 450 (1992)

[18] R. Voss: Phys. Rev. Lett. **68**, 3805 (1992); R. Voss: Fractals **2**, 1 (1994)

[19] J. Maddox: Nature **358**, 103 (1992)

[20] P.J. Munson, R.C. Taylor, G.S. Michaels: Nature **360**, 636 (1992)

[21] I. Amato: Science **257**, 747 (1992)

[22] V.V. Prabhu, J.-M. Claverie: Nature **357**, 782 (1992)

[23] P. Yam: Sci. Am. **267[3]**, 23 (1992)

[24] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley: Physica A **191**, 25 (1992); H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, J.M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino, M. Simons: Physica A **191**, 1 (1992)

[25] C.A. Chatzidimitriou-Dreismann, D. Larhammar: Nature **361**, 212 (1993); D. Larhammar, C.A. Chatzidimitriou-Dreismann: Nucleic Acids Res. **21**, 5167 (1993) C.A. Chatzidimitriou-Dreismann, R.M.F. Streffer, D. Larhammar: Biochim. Biophys. Acta **1217**, 181 (1994); C.A. Chatzidimitriou-Dreismann, R.M.F. Streffer, D. Larhammar: Eur. J. Biochem. **224**, 365 (1994)

[26] A.Yu. Grosberg, Y. Rabin, S. Havlin, A. Neer: Europhys. Lett. **23**, 373 (1993)

[27] S. Karlin, V. Brendel: Science **259**, 677 (1993)

[28] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, M. Simons, H.E. Stanley: Phys. Rev. E **47**, 3730 (1993)

[29] N. Shnerb, E. Eisenberg: Phys. Rev. E **49**, R1005 (1994)

[30] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley: Phys. Rev. E **47**, 4514 (1993).

[31] A. S. Borovik, A. Yu. Grosberg and M. D. Frank Kamenezki, J. Biomolec. Structure and Dynamics **xx**, xxx (1994)

[32] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, M.H.R. Stanley, M. Simons: Biophys. J. **65**, 2673 (1993)

[33] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger: Phys. Rev. E **49**, 1685 (1994)

[34] S.M. Ossadnik, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.-K. Peng, M. Simons, H.E. Stanley: Biophys. J. **67**, 64 (1994); H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K.Peng, M. Simons: [Proceedings of Internat'l Conf. on Condensed Matter Physics, Bar-Ilan], Physica A **200**, 4 (1993); H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, S.M. Ossadnik, C.-K. Peng, M. Simons: Fractals **1**, 283-301 (1993)

[35] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, and H.E. Stanley, "Long-Range Correlation Properties of Coding and Noncoding DNA Sequences," Phys. Rev. E (submitted).

[36] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley: Phys. Rev. Lett. **73**, 3169-3172 (1994); F. Flam: Science **266**, 1320 (1994); E. Pennisi: Science News **146**, 391 (1994)

[37] S. Tavaré, B.W. Giddings, in: *Mathematical Methods for DNA Sequences*, Eds. M.S. Waterman (CRC Press, Boca Raton 1989), pp. 117-132; J.D. Watson, M. Gilman, J. Witkowski, M. Zoller: *Recombinant DNA* (Scientific American Books, New York 1992).

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

[39] G.H. Weiss: *Random Walks* (North-Holland, Amsterdam 1994)

[40] S. Havlin, R. Selinger, M. Schwartz, H.E. Stanley, A. Bunde: Phys. Rev. Lett. **61**, 1438 (1988); S. Havlin, M. Schwartz, R. Blumberg Selinger, A. Bunde, H.E. Stanley: Phys. Rev. A **40**, 1717 (1989); R.B. Selinger, S. Havlin, F. Leyvraz, M. Schwartz, H.E. Stanley: Phys. Rev. A **40**, 6755 (1989)

[41] C.-K. Peng, S. Havlin, M. Schwartz, H.E. Stanley, G.H. Weiss: Physica A **178**, 401 (1991); C.-K. Peng, S. Havlin, M. Schwartz, H.E. Stanley: Phys. Rev. A **44**, 2239 (1991)

[42] M. Araujo, S. Havlin, G.H. Weiss, H.E. Stanley: Phys. Rev. A **43**, 5207 (1991); S. Havlin, S.V. Buldyrev, H.E. Stanley, G.H. Weiss: J. Phys. A **24**, L925 (1991); S. Prakash, S. Havlin, M. Schwartz, H.E. Stanley: Phys. Rev. A **46**, R1724 (1992)

[43] C.L. Berthelsen, J.A. Glazier, M.H. Skolnick: Phys. Rev. A **45**, 8902 (1992)

[44] J. Jurka, T. Walichiewicz, A. Milosavljevic: J. Mol. Evol. **35**, 286 (1992)

[45] M.F. Shlesinger, J. Klafter: in *On Growth and Form: Fractal and Non-Fractal Patterns in Physics*, edited by H.E. Stanley and N. Ostrowsky

(Martinus Nijhoff, Dordrecht, 1986), p. 279ff

[46] M.F. Shlesinger, J. Klafter, Y.M. Wong: J. Stat. Phys. **27**, 499 (1982)

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

[48] R.N. Mantegna: Physica A **179**, 232 (1991)

[49] The long-range correlations were found to extend over the entire yeast chromosome III region (315,357 nucleotides)—see Ref. [20]. The yeast chromosome III sequence was published by S.G. Oliver et al.: Nature **357**, 38 (1992)

[50] J. Jurka: J. Mol. Evol. **29**, 496 (1989)

[51] R.H. Hwu, J.W. Roberts, E.H. Davidson, R.J. Britten: Proc. Natl. Acad. Sci. USA. **83**, 3875 (1986)

[52] E. Zuckerkandl, G. Latter, J. Jurka: J. Mol. Evol. **29**, 504 (1989)

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

[54] P.-G. de Gennes: *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca NY, 1979)

[55] J. de Cloiseaux: J. Physique (Paris) **41**, 223 (1980), p. 223

[56] S. Redner: J. Phys. A **13**, 3525 (1980)

[57] A. Baumgartner: Z. Phys. B **42**, 265 (1981)

[58] H.S. Chan, K.A. Dill: J. Chem. Phys. **92**, 3118 (1990)

[59] T. M. Birshtein, S. V. Buldyrev: Polymer **32**, 3387 (1991)

[60] A. Schenkel, J. Zhang, Y-C. Zhang: Fractals **1**, 47 (1993); M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, N. Shnerb: Fractals **2**, 7 (1994)

[61] G.K. Zipf: *Human Behavior and the Principle of "Least Effort"* (Addison-Wesley, New York 1949)

[62] L. Brillouin: *Science and Information Theory* (Academic Press, New York 1956)

[63] C.E. Shannon: Bell Systems Tech. J. **80**, 50 (1951)

[64] A. Czirok, R. Mantegna, S. Havlin, H.E. Stanley: Phys. Rev. E (submitted)

[65] J.-P. Bouchaud: "More Lévy distributions in physics", These Proceedings

[66] M.H.R. Stanley: 1994 Westinghouse Report (unpublished)

[67] M.H.R. Stanley, S.V. Buldyrev, S. Havlin, R. Mantegna, M.A. Salinger, H.E. Stanley: Eco. Lett. (submitted)

[68] J. Pivinski, R. Tucksmith, A. Such, C. Haight: Fortune (18 April 1994), p. 224