Statistical and Linguistic Features of Noncoding DNA: A Heterogeneous "Complex System"

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] and M. Simons[3]

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

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

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

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

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

Abstract

We present evidence supporting the idea that the DNA sequence in genes containing *noncoding* regions is correlated, and that the correlation is remarkably long range—indeed, base pairs *thousands of base pairs* distant are correlated. We do not find such a long-range correlation in the coding regions of the gene; we utilize this fact to build a *Coding Sequence Finder Algorithm*, which uses statistical ideas to locate the coding regions of an unknown DNA sequence. We resolve the problem of the "non-stationarity" feature of the sequence of base pairs (that the relative concentration of purines and pyrmidines changes in different regions of the mosaic-like chain) by describing a new algorithm called *Detrended Fluctuation Analysis (DFA)*. We address the claim of Voss that there is no difference in the statistical properties of coding and noncoding regions of DNA by systematically applying the DFA algorithm, as well as standard FFT analysis, to every DNA sequence (33 301 coding and 29 453 noncoding) in the entire GenBank database. We describe a simple model to account for the presence of long-range power-law correlations (and the systematic variation of the scaling exponent α with evolution) which is based upon a generalization of the classic Lévy walk. Finally, we describe briefly some recent work showing that the *noncoding* sequences have certain statistical features in common with natural languages. Specifically, we adapt to DNA the Zipf approach to analyzing linguistic texts, and the Shannon approach to quantifying the "redundancy" of a linguistic text in terms of a measurable entropy function. We demonstrate that noncoding regions in eukaryotes display a smaller entropy and larger redundancy than coding regions, further supporting the possibility that noncoding regions of DNA may carry biological information.

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 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 $\Delta y(\ell) \equiv y(\ell_0 + \ell) - y(\ell_0)$, where

$$y(\ell) \equiv \sum_{i=1}^{\ell} u(i). \tag{2}$$

If there is no characteristic length (i.e., if the correlation were "infinite-range"), then fluctuations will also be described by a power law

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

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.

5 Detrended Fluctuation Analysis (DFA)

The initial report [15] on long-range (scale-invariant) correlations only in noncoding DNA sequences has generated contradicting responses. Some [16, 17, 20, 21] support our initial finding, while some [17, 22, 25, 27] disagree. However, the conclusions of Refs. [18] and [17, 22, 25, 27] are inconsistent *with one another* in that [17] and [27] doubt the existence of long-range correlations (even in noncoding sequences) while [18] and [22, 25] conclude that even coding regions display long-range correlations ($\alpha > 1/2$). Prabhu and Claverie [22] claim that their analysis of the putative *coding* regions of the yeast chromosome III produces a *wide range of exponent values*, some larger than 0.5. The source of these contradicting claims may arise from the fact that, in addition to normal statistical

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.

fluctuations expected for analysis of rather short sequences, coding regions typically consist of only a few lengthy regions of alternating strand bias—and so we have non-stationarity. Hence conventional scaling analyses cannot be applied reliably to the entire sequence but only to sub-sequences.

Peng et al. [33] have recently applied the "bridge method" to DNA, and have also developed a similar method specifically adapted to handle problems associated with non-stationary sequences which they term *detrended fluctuation analysis* (DFA).

The idea of the DFA method is to compute the dependence of the standard error of a linear interpolation of a DNA walk $F_d(\ell)$ on the size of the interpolation segment ℓ. The method takes into account differences in local nucleotide content and may be applied to the entire sequence which has lengthy patches. In contrast with the original $F(\ell)$ function, which has spurious crossovers even for ℓ much smaller than a typical patch size, the detrended function $F_d(\ell)$ shows linear behavior on the log-log plot for all length scales up to the characteristic patch size, which is of the order of a thousand nucleotides in the coding sequences. For ℓ close to the characteristic patch size the log-log plot of $F_d(\ell)$ has an abrupt

Figure 2: Analysis of section of Yeast Chromosome III using the sliding box *Coding Sequence Finder* "CSF" algorithm. The value of the long-range correlation exponent α is shown as a function of position along the DNA chain. In this figure, the results for about 10% of the DNA are shown (from base pair #30,000 to base pair #60,000). Shown as vertical bars are the putative genes and open reading frames; denoted by the letter "G" are those genes that have been more firmly identified (March 1993 version of *GenBank*). Note that the local value of α displays **minima** where genes are suspected, while between the genes α displays **maxima**. This behavior corresponds to the fact that the DNA sequence of genes lacks long-range correlations ($\alpha = 0.5$ in the idealized limit), while the DNA sequence in between genes possesses long-range correlations ($\alpha \approx 0.6$).

change in its slope.

The DFA method clearly supports the difference between coding and noncoding sequences, showing that the coding sequences are less correlated than noncoding sequences for the length scales less than 1000, which is close to characteristic patch size in the coding regions. One source of this difference is the tandem repeats (sequences such as AAAAAA...), which are quite frequent in noncoding sequences and absent in the coding sequences.

6 Coding Sequence Finder (CSF) Algorithm

To provide an "unbiased" test of the thesis that noncoding regions possess but coding regions lack long-range correlations, Ossadnik et al. [34] analyzed several artificial uncorrelated and correlated "control sequences" of size 10^5 nucleotides using the GRAIL neural net algorithm [49]. The GRAIL algorithm identified about 60 putative exons in the uncorrelated sequences, but only about 5 putative exons in the correlated sequences.

Using the DFA method, we can measure the local value of the correlation exponent α along the sequence (see Fig. 2) and find that the local minima of α as a function of a nucleotide position usually correspond to noncoding regions, while the local maxima correspond to noncoding regions. Statistical analysis using the DFA technique of the nucleotide sequence data for yeast chromosome III (315,338 nucleotides) shows that the probability that the observed correspondence between the positions of minima and coding regions is due to random coincidence is less than 0.0014. Thus, this method—which we called the "coding sequence finder" (CSF) algorithm—can be used for finding coding regions in the newly sequenced DNA, a potentially important application of DNA walk analysis.

7 Systematic Analysis of GenBank Database

An open question in computational molecular biology is whether long-range correlations are present in both coding and noncoding DNA or only in the latter. To answer this question, Buldyrev et al. [35] recently analyzed all 33 301 coding and all 29 453 noncoding eukaryotic sequences—each of length larger than 512 base pairs (bp)—in the present release of the GenBank to determine whether there is any statistically significant distinction in their long-range correlation properties.

Buldyrev et al. find that standard fast Fourier transform (FFT) analysis indicates that *coding* sequences have practically no correlations in the range from 10 bp to 100 bp (spectral exponent $\beta \pm 2SD = 0.00 \pm 0.04$). Here β is defined through the relation $S(f) \sim 1/f^{\beta}$, where $S(f)$ is the Fourier transform of the correlation function, and β is related to the long-range correlation exponent α by $\beta = 2\alpha - 1$ so that $\alpha = 1/2$ corresponds to $\beta = 0$ (white noise).

In contrast, for *noncoding* sequences, the average value of the spectral exponent β is positive (0.16 ± 0.05), which unambiguously shows the presence of long-range correlations. They also separately analyzed the 874 coding and 1157 noncoding sequences which have more than 4096 bp, and found a larger region of power law behavior. Buldyrev et al. calculated the probability that these two data sets (coding and noncoding) were drawn from the same distribution, and found that it is less than 10^{-10}. Buldyrev et al. also obtained independent confirmation of these findings using the DFA method, which is designed to treat sequences with statistical heterogeneity such as DNA's known mosaic structure ("patchiness") arising from non-stationarity of nucleotide concentration. The near-perfect agreement between the two independent analysis methods, FFT and DFA, increases the confidence in the reliability of the conclusion that long-range correlation properties of coding and noncoding sequences.

¿From a practical viewpoint, the statistically significant difference in long-range power law correlations between coding and noncoding DNA regions that we observe supports the development of gene finding algorithms based on these distinct scaling properties. A recently reported algorithm of this kind [34] is especially useful in the analysis of DNA sequences with relatively long coding

regions, such as those in yeast chromosome III.

Finally, we note that although the scaling exponents α and β have potential use in quantifying changes in genome complexity with evolution, the current GenBank database does not allow us to address the important question of whether unique values of these exponents can be assigned to different species or to related groups of organisms. At present, the GenBank data have been collected such that particular organisms tend to be represented more frequently than others. For example, about 80% of the sequences from birds are from *Gallus gallus* (the chicken) and about 2/3 of the insect sequences are from *Drosophila melanogaster*. The results indicate the importance of sequencing not only coding but also noncoding DNA from a wider variety of species.

8 Generalized Lévy Walk Model

Although the correlation is long-range in the noncoding 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 noncoding 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 noncoding 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.

These features motivated us to apply a generalized Lévy walk model (see Figs. 1c, 1d and 2) for the noncoding 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 subregions of non-correlated sequences within these noncoding 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

$$P(l_j) \propto (1/l_j)^{\mu}, \tag{4}$$

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.

Figure 3: 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).

We consider a generalization of the Lévy walk [42] to interpret recent findings of long-range correlation in noncoding 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

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

to go up and

$$p_- = (1 - \epsilon_j)/2 \tag{6}$$

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

Lévy walk parameter μ [42, 46],

$$\alpha = \begin{cases} 1 & \mu \leq 2 \\ 2 - \mu/2 & 2 < \mu < 3 \\ 1/2 & \mu \geq 3. \end{cases} \tag{7}$$

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. (4) 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 $\mu \geq 3$.

9 Mosaic Nature of DNA Structure

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 noncoding 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 a 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 [58] gave $\mu = 2.22 \pm 0.05$ which yields $\alpha = 0.89$, a larger value than the effective value 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. (4).

In summary, it is clear 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.

10 Linguistic Analysis of Noncoding and Coding DNA

Long-range correlations have been found recently in human writings [59]. 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 [60]. He observed that the frequency of words as a function of the word order ("rank") decays as a power law (with a power ζ 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 noncoding regions, the words are not known. Therefore Mantegna et al. [36] consider 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 noncoding sequences, suggesting that the noncoding sequences bear more resemblance to a natural language than the coding sequences. Moreover, the "words" used in coding and noncoding sequences appear in quite different orders (Fig. 4).

Related interesting statistical measures of short-range correlations in lan-

Figure 4: Linguistic features of noncoding DNA. (a) Log-log plot of a histogram of word frequency for the noncoding part of Yeast Chromosome III ($\approx$ 315,000 bp). The 6-character words are placed in rank order, where rank 1 corresponds to the most frequently used word, rank 2 to the second most frequently used word, and so forth. The straight line behavior provides evidence for a structured language in noncoding DNA. Rainbow color code corresponds to the rank of words in the language of this sequence, which is used as a "reference language" below. (b) Linear-log plot of word frequency histogram for the *coding* part of the same chromosome. The straight line behavior shows that the coding part lacks the statistical properties of a structured language. The colors are re-arranged, corresponding to the re-arrangements of their rank with respect to the reference language. (c) Same as part (a), except for Yeast Chromosome XI ($\approx$ 666,000 bp), demonstrating the striking parallels between the language in these two chromosomes. (d) Same as part (b), except for Yeast Chromosome XI.

guages 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 [61, 62]. Shannon's redundancy is defined in terms of the entropy of a text—or, more precisely, the "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.

Here p_i is the normalized frequency of occurrence of n-tuple i. 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 $k = \log_2 4 = 2$.

Mantegna et al. [36] also calculate 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 noncoding 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 *noncoding* 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 noncoding 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 and some understanding of the implications of Zipf analysis is now emerging [63]. An example that was the subject of some discussion 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 [64] finds that this plot is linear for European companies, while M.H.R. Stanley [65] finds linearity for American companies. Furthermore, M.H.R. Stanley et al. [66] 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").

11 Outlook for the Future

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].

The results of the systematic and inclusive analysis of GenBank DNA sequences are notable for two major reasons.

(i) The GenBank data unambiguously demonstrate that noncoding DNA, but not coding DNA, possesses long-range correlations. This finding is made using two independent, complementary techniques: Fourier analysis and DFA, a modification of root-mean-square analysis of random walks. Indeed, as shown in Tables I and II of [35], the spectral exponent β computed by both techniques for the same sequence, is nearly identical.

(ii) The GenBank data demonstrate an increase in the complexity of the noncoding DNA sequences with evolution. The value of β for vertebrates is significantly greater than that for invertebrates. This finding based on the full GenBank data set supports the suggestion based upon a systematic study of the myosin heavy gene family that there is an apparent increase in the complexity of noncoding DNA for more highly evolved species compared to less evolved ones [32].

Both of these results contradict the report of Voss [18], who failed to observe any difference in the long-range correlation properties of coding and noncoding DNA and who reported a decrease in the value of the spectral exponent β with evolution.

The ultimate meaning of long-range correlations is still not clear. Its is possible that long-range correlations exist also in other systems of biological interest. For example, the idea of long-range correlations has been extended to the analysis of the beat-to-beat intervals in the normal and diseased heart [68, 69], and to human gait [70]. The healthy heartbeat is generally thought to be regulated according to the classical principle of homeostasis whereby physiologic systems operate to reduce variability and achieve an equilibrium-like state [71]. We find, however, that under normal conditions, beat-to-beat fluctuations in heart rate display the kind of long-range correlations typically exhibited by physical dynamical systems far from equilibrium, such as those near a critical point. Specifically, we find evidence for such power-law correlations that extend over thousands of heartbeats in healthy subjects. In contrast, heart rate time series from patients with severe congestive heart failure show a breakdown of this long-range correlation behavior, with the emergence of a characteristic short-range time scale. Similar alterations in correlation behavior may be important in modeling the transition from health to disease in a wide variety of pathologic conditions.

12 Acknowledgements

We are grateful to many individuals, including M.E. Matsa, S.M. Ossadnik, 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, National Institutes of Health (Human Genome Project), 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, Israel Academy of Sciences, 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); H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng, F. Sciortino and M. Simons, "Fractals in Biology and Medicine," in *Diffusion Processes: Experiment, Theory, Simulations Proceedings of the Vth M. Born Symposium]*, edited by A. Pekalski (Springer-Verlag, Berlin, 1994), pp. 147–178; H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng and M. Simons, "Statistical Mechanics in Biology: How Ubiquitous are Long-Range Correlations?" *Proc. International Conference on Statistical Mechanics*, Physica A **205**, 214–253 (1994).

[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 **12**, 655-669 (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); S. Havlin, S. V. Buldyrev, A. L. Goldberger, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng, M. Simons, and H. E. Stanley, Chaos, Solitons, and Fractals **6**, 171-201 (1995).

[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); P. Yam: Scientific American (March 1995)

[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] E.C. Uberbacher, R.J. Mural: Proc. Natl. Acad. Sci. USA **88**, 11261 (1991)

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

[59] 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)

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

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

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

[63] A. Czirók, R. N. Mantegna, S. Havlin and H. E. Stanley, "Correlations in Binary Sequences and Generalized Zipf Analysis," Phys. Rev. E (submitted).

[64] J.-P. Bouchaud: "More Lévy distributions in physics", in *Proc. 1993 International Conf. on Lévy Flights*, edited by U. Frisch, M. F. Shlesinger, and G. Zaslavsky (Springer, Berlin, 1995).

[65] M.H.R. Stanley: 1994 Westinghouse Report (unpublished); H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and M. H. R. Stanley, "Long-Range Correlatins and Generalized Lévy Walks in DNA Sequences," in *Proc. 1993 International Conf. on Lévy Flights*, edited by U. Frisch, M. F. Shlesinger, and G. Zaslavsky (Springer, Berlin, 1995).

[66] M.H.R. Stanley, S.V. Buldyrev, S. Havlin, R. Mantegna, M.A. Salinger, H.E. Stanley: "Zipf plots and the size distribution of Firms" ,Eco. Lett. (submitted). See also R. N. Mantegna and H. E. Stanley, "Ultra-Slow Convergence to a Gaussian: The Truncated Lévy Flight," in *Proc. 1993 International Conf. on Lévy Flights*, edited by U. Frisch, M. F. Shlesinger, and G. Zaslavsky (Springer, Berlin, 1995); R. N. Mantegna and H. E. Stanley, "Scaling and Intermittency in the Mesoscopic Dynamics of an Economic Index," Nature (submitted).

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

[68] C.-K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, *Phys. Rev. Lett.* **70**, 1343 (1993); C. K. Peng, S. V. Buldyrev, J. M. Hausdorff, S. Havlin, J. E. Mietus, M. Simons, H. E. Stanley, and A. L. Goldberger, in *Fractals in Biology and Medicine*, G. A. Losa, T. F. Nonnenmacher and E. R. Weibel, eds. (Birkhauser Verlag, Boston, 1994).

[69] C. K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time Series," [*Proc. NATO Dynamical Disease Conference*], edited by L. Glass, Chaos **5**, xxx (1995); C. K. Peng, J. M. Hausdorff, J. E. Mietus, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Fractals in Physiological Control: From Heartbeat to Gait," in *Proc. 1993 International Conf. on Lévy Flights*, edited by U. Frisch, M. F. Shlesinger, and G. Zaslavsky (Springer, Berlin, 1995).

[70] J. M. Hausdorff, C.-K. Peng, Z. Ladin. J. Y. Wei, and A. L. Goldberger, "Is Walking a Random Walk? Evidence for Long-Range Correlations in the Stride Interval of Human Gait," J. Appl. Physiol. **78**, 349-358 (1995).

[71] W. B. Cannon, *Physiol. Rev.* **9**, 399 (1929).