ELSEVIER

Physica A 221 (1995) 180–192

PHYSICA A

Statistical properties of DNA sequences

C.-K. Peng [a,b], S.V. Buldyrev [b], A.L. Goldberger [a,c], S. Havlin [b,d], R.N. Mantegna [b,e], M. Simons [a], H.E. Stanley [b]

[a] *Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, USA*

[b] *Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA*

[c] *Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA*

[d] *Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

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

Abstract

We review evidence supporting the idea that the DNA sequence in genes containing *non-coding* regions is correlated, and that the correlation is remarkably long range – indeed, nucleotides *thousands of base pairs* distant are correlated. We do not find such a long-range correlation in the coding regions of the gene. We resolve the problem of the "non-stationarity" feature of the sequence of base pairs by applying 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 non-coding 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 non-coding) in the entire GenBank database. Finally, we describe briefly some recent work showing that the *non-coding* sequences have certain statistical features in common with natural and artificial languages. Specifically, we adapt to DNA the Zipf approach to analyzing linguistic texts. These statistical properties of non-coding sequences support the possibility that non-coding 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" [1,2]. 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.

0378-4371/95/$09.50
SSDI 0378-4371(95)00247-2

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 [3]

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

Many systems drive themselves spontaneously toward critical points [2,3]. 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 (see Refs. [4–6], and references cited in [7]) 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 [8], as well as potential practical applications for distinguishing coding and non-coding regions in long nucleotide chains [9]. It also may be related to the presence of a "language" in non-coding DNA [10].

2. DNA and the "DNA walk"

The role of genomic DNA sequences in coding for protein structure is well known [11]. The human genome contains information for approximately 100000 different proteins, which define all inheritable features of an individual. The genomic sequence is likely to be 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).

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 non-coding part of the genome.

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 a 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 sizes 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 [3].

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* [4]. For the conventional one-dimensional random walk model [12], 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.

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 (see Fig. 1). The question we asked was whether such a walk displays only short-range correlations (as in a Markov chain) or long-range correlations (as in critical phenomena and other scale-free "fractal" phenomena). A different type of DNA walk was introduced earlier by Azbel [13].

There have also been attempts to map DNA sequence onto multi-dimensional DNA walks [5,14]. However, recent work [9] indicates that the original purine-pyrimidine rule provides the most robust results, probably due to the purine-pyrimidine chemical complementarity.

3. Correlations and self-similar processes

The concept of self-similar processes was first proposed by Kolmogorov [15] in theoretical physics and later introduced into mathematics through the influential work of Mandelbrot on fractals [16]. An object is self-similar if its subsets can be rescaled to resemble (statistically) the original object itself. A scaling exponent (also called the *self-similarity parameter*) can be defined by this rescaling process. A stationary sequence with long-range correlations can be integrated, i.e. form an accumulated sum, to form a self-similar process. Therefore, measurement of the self-similarity scaling exponent of the integrated series can tell us the long-range correlation properties of the original sequence. Hurst analysis [17] and our DNA walk analysis are both based on this concept. Fig. 1d 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.

An important statistical quantity characterizing any walk 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 correlations were "infinite-range"), then fluctuations will also be described by a power law

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

with $\alpha \neq 1/2$. The exponent α is the self-similarity parameter mentioned above and therefore is directly related to long-range correlations in the sequence.

The fact that data for intron-containing and intergenic (i.e. non-coding) 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.

4. Detrended fluctuation analysis (DFA)

The initial report [4] on long-range (scale-invariant) correlations only in non-coding DNA sequences has generated contradicting responses. For details see the work of Buldyrev et al. [7]. The source of these contradicting claims may arise from the fact that, in addition to normal statistical fluctuations expected for analysis of rather short sequences, coding regions typically consist of several 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 where it is homogeneous. Fig. 1 shows a collection of DNA walks for artificial and actual DNA sequences.

Peng et al. [18] 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 basic idea underlying 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

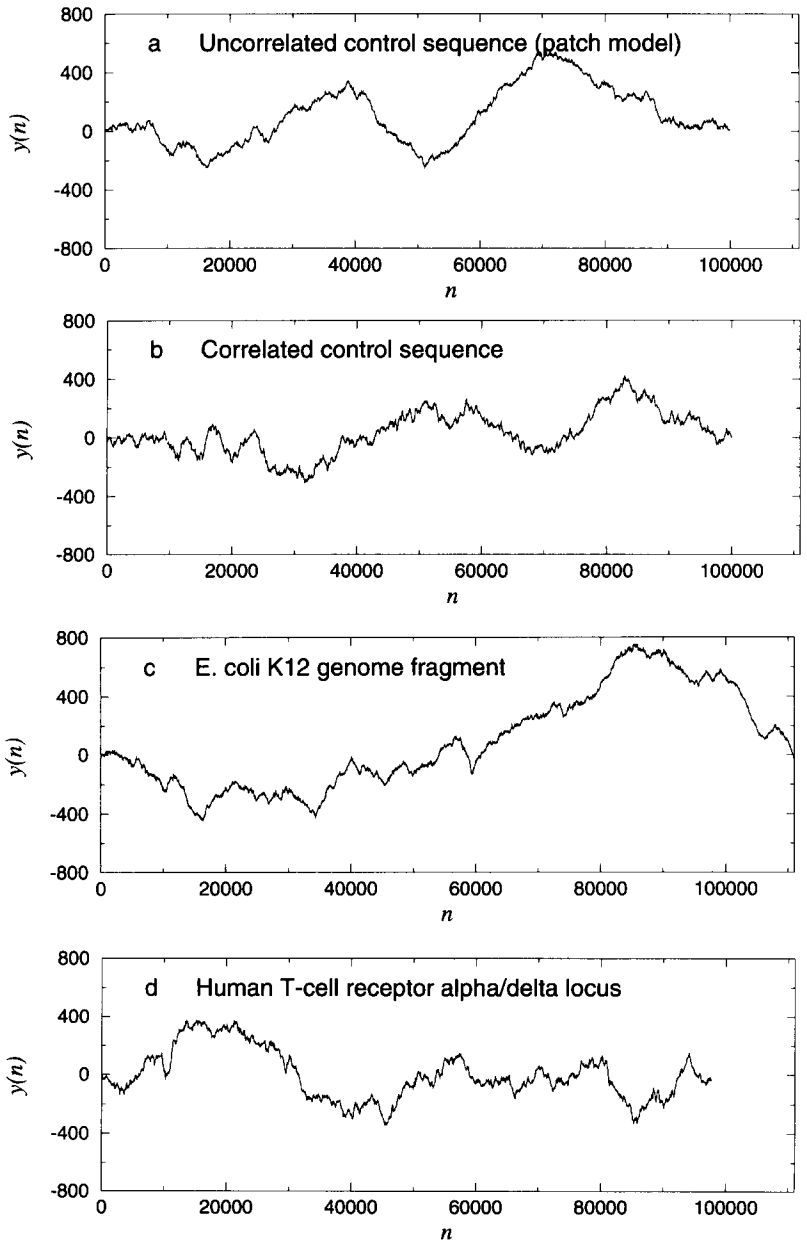

Fig. 1. (a) DNA walk for a control sequence obtained by stitching together biased random walks; the characteristic length for the patches is 2500. (b) DNA walk for a control sequence obtained from building in a long-range correlation into a set of 100,000 "nucleotides" which are correlated with power-law exponent $\alpha = 0.61$. (c) DNA walk for a genomic fragment containing mostly coding regions [*E. coli* K12 genome, 0–2.4 min. region, GenBank name: ECO110K, 111401 bp]. (d) DNA walk for a typical intron-containing chromosomal region of a comparable length (human T-cell receptor alpha/delta locus, GenBank name: HUMTCRADCV, 97634 bp). Large sub-regions ("patches") of uniform overall slope ("strand bias") reflect the mosaic structure. To facilitate the comparison of subtle fluctuations, each landscape is plotted so that the end point has the same vertical displacement as the starting point, i.e., the overall bias has been removed. After Ref. [18].

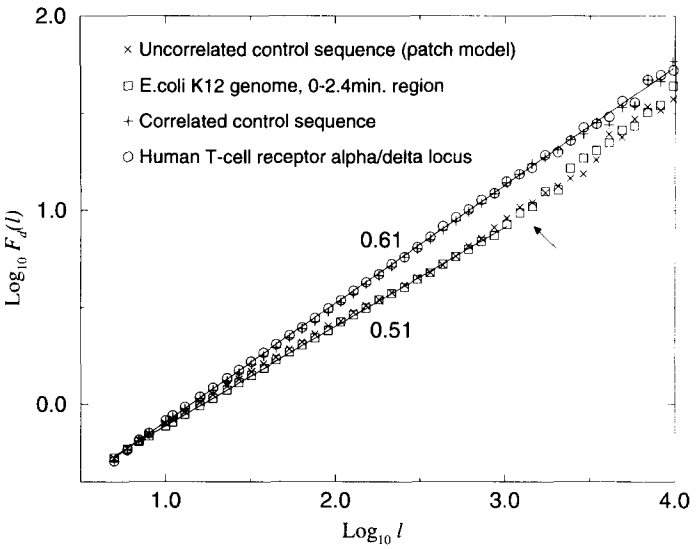

Fig. 2. DFA analysis of the four landscapes shown in Fig. 1. The uncorrelated biased random walk (a) (×) is similar to the *E. coli* genomic coding fragment (c) (□), while the correlated control sequence (b) (+) is quite similar to the highly non-coding human T-cell receptor alpha/delta locus (d) (○). The lower solid line, the best fit for *E. coli* data from $\ell = 4$ to 861, has slope 0.51. The upper solid line, the best fit for human data from $\ell = 4$ to 8192, has slope 0.61. The arrow denoting the crossover phenomenon is explained in the text. After Ref. [18].

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 change in its slope. (See Fig. 2.)

The DFA method clearly supports the difference between coding and non-coding sequences, showing that the coding sequences are less correlated than non-coding sequences for the length scales less than 1000, which is close to characteristic patch size in the coding regions.

5. Systematic analysis of the GenBank database

An open question in computational molecular biology is whether long-range correlations are present in both coding and non-coding DNA (as claimed by Voss [6]) or only in the latter (as we originally reported). To answer this question, Buldyrev et al. [20] recently analyzed all 33301 coding and all 29453 non-coding 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 $\beta \pm 2\mathrm{SD} = 0.00 \pm 0.04$). 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 *non-coding* 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 non-coding 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 non-coding) 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 non-coding sequences.

From a practical viewpoint, the statistically significant difference in long-range power-law correlations between coding and non-coding DNA regions that we observe supports the development of gene finding algorithms based on these distinct scaling properties (see Section 6).

Very recently Arneodo et al. [21] studied long-range correlation in DNA sequences using wavelet analysis. The wavelet transform can be made blind to "patchiness" of genomic qenomic sequences. They found the existence of long-range correlations in non-coding regimes, and no long-range correlations in coding regimes in excellent agreement with Buldyrev et al. [20].

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 non-coding DNA from a wider variety of species.

6. Additional application: Coding sequence finder (CSF) algorithm

To provide an "unbiased" test of the thesis that non-coding regions possess but coding regions lack long-range correlations, Ossadnik et al. [9] analyzed several artificial uncorrelated and correlated "control sequences" of size 10^5 nucleotides using the GRAIL neural net algorithm [19]. 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. 3) and find that the local minima of α as a function

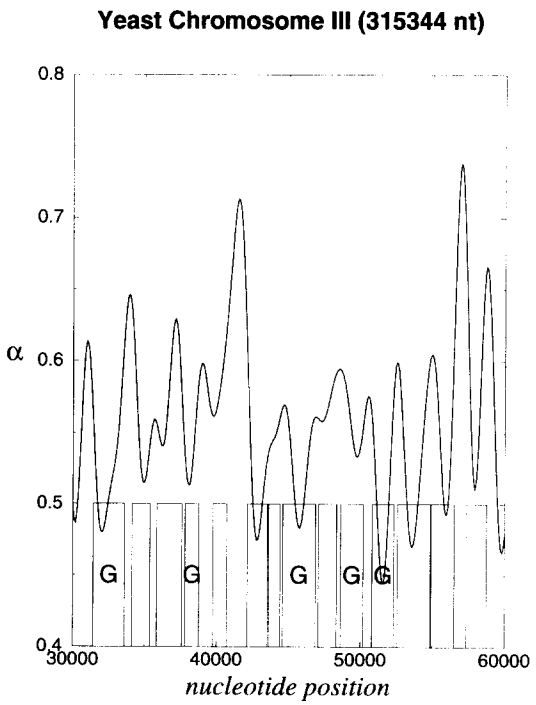

Fig. 3. Analysis of section of Yeast Chromosome III using the sliding box *Coding Sequence Finder* "CSF" algorithm. The value of the long-range correlation exponent α is shown as a function of position along the DNA chain. In this figure, the results for about 10% of the DNA are shown (from base pair #30000 to base pair #60000). 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 ($\alpha \approx 0.6$).

of a nucleotide position usually correspond to coding regions, while the local maxima correspond to non-coding regions. Statistical analysis using the DFA technique of the nucleotide sequence data for yeast chromosome III (315338 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. Linguistic analysis of DNA sequences

Long-range correlations have been found recently in human writings [22–24]. 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 by Zipf [25]. 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 non-coding regions, the words are not known. Therefore Mantegna et al. [10,26] consider the word length n as a free parameter, and perform 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.

Before we discuss the results from actual DNA sequences, let us first consider examples of artificial language. A compiled computer program and a computer data file can both be treated as sequences of binary code. Although, they both contain useful information, the structure of the information are very different. A computer program that can execute series of instructions and decisions should bear more resemblance to natural language than a binary data file which only stores information. We do not expect the binary sequence of a data file to exhibit long-range correlations or any hierarchical structure. Indeed, the DFA and Zipf analyses confirm the above assumption (see Fig. 4).

The results of the Zipf analysis for all 40 DNA sequences analyzed are summarized by Mantegna et al. [10]. 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 (Fig. 5). Moreover, the "words" used in coding and non-coding sequences appear in quite different orders. Furthermore, it is shown that the n-tuples usage is significantly different for different species. This difference may be related to the underlying evolutionary process such that a phylogenetic tree can be generated by studying the similarity and difference of n-tuples usage [27].

It is known that in different organisms (and within the same organism in different regions of the same genome) the DNA has different C+G content and different first order Markovian matrices [i.e. different probabilities $P(i,j)$] - see, e.g., Ref. [28]. A possible explanation of the difference in functional form observed in the Zipf plot could be due to the differences in the CG content and/or in the Markovian matrices characterizing the investigated sequences and their coding and non-coding regions. See Refs. [26,29] for details.

It appears that the 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 a Zipf analysis and some understanding of the implications of the Zipf analysis is now emerging [30,31].

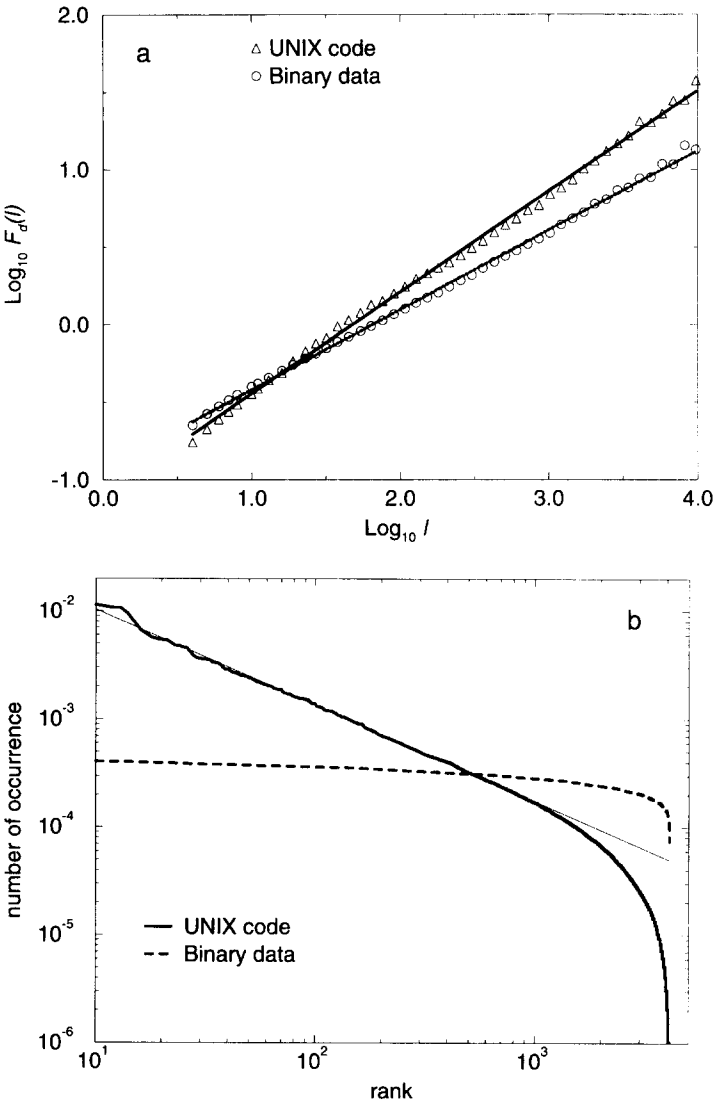

Fig. 4. (a) DFA analysis of the compiled version of the UNIX Operating System and a computer data file, both comprising $\approx 10^6$ binary bits. The best fit lines have slope 0.65 for UNIX code and 0.51 for binary data file. (b) n-tuple Zipf analysis for the same binary sequences shown in (a) with $n = 12$. For the compiled Unix code, a power-law behavior is observed for a rank interval of more than two decades. In the power-law region (rank 10 to 1000), the best linear fit of the log–log plot gives the value of $\zeta = 0.89$. Similar behavior is obtained when $n = 8, 10$ and 14. However, for the binary data file, the Zipf plot shows a very flat line, indicating, as expected, no hierarchical structure exist.

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 collabora-

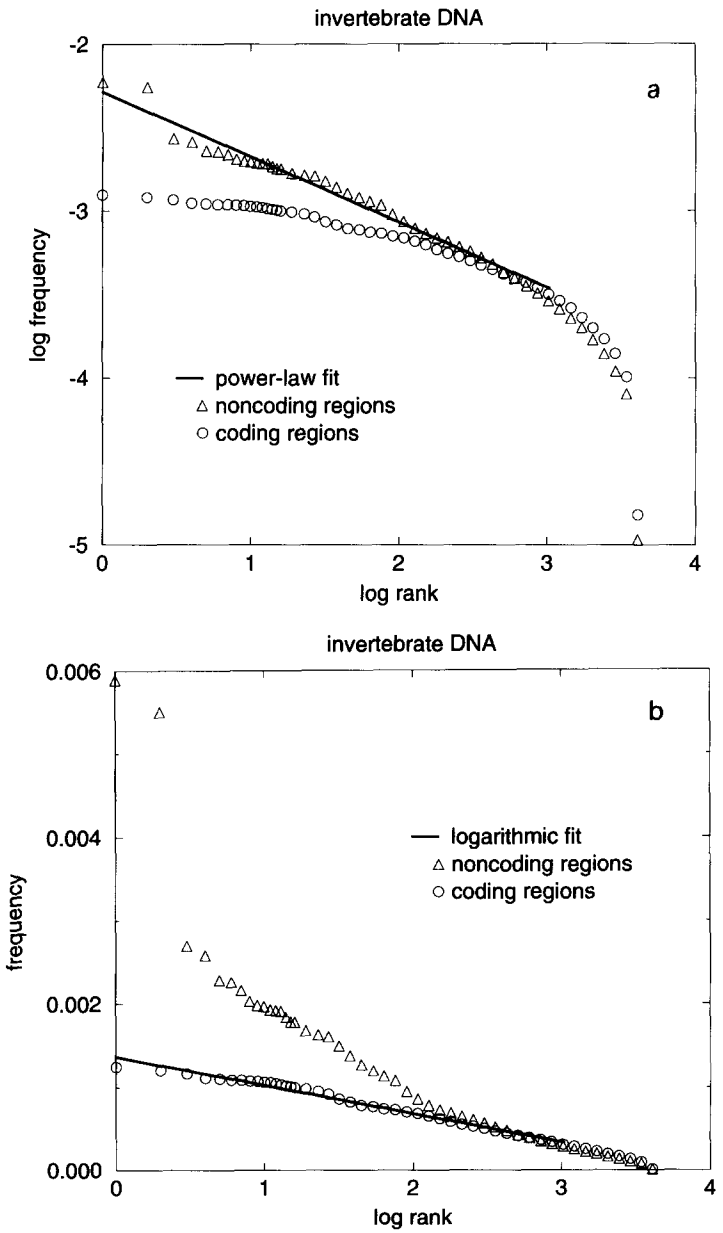

Fig. 5. Linguistic features of DNA. (a) Log–log plot of a histogram of word frequency for the coding and non-coding part from 4 invertebrate DNA sequences (CEC07A9, CELTWIMUSC, DROABDB, SCCHRIII, and SCCHRXI for a total number of 1,102,752 bp and a number of 728,998 coding DNA bp). The 6-character words 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 non-coding DNA. (b) Semilogarithmic Zipf plot of the same data shown in (a). The straight line (logarithmic behavior) is the best fit of the Zipf plot for coding DNA sequences. The fitting procedure is performed in the interval $R \leq 1000$. The curvature of the Zipf plot of the non-coding DNA is evident. After Ref. [26].

tive 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 an NIH/NIMH Postdoctoral NRSA Fellowshipthe (to C-KP), 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.

References

[1] A. Bunde and S. Havlin, eds. Fractals and Disordered Systems (Springer-Verlag, Berlin, 1991).
[2] A. Bunde and S. Havlin, eds. Fractals in Science (Springer-Verlag, Berlin, 1994).
[3] H.E. Stanley, Introduction to Phase Transitions and Critical Phenomena (Oxford University Press, London, 1971).
[4] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, Nature 356 (1992) 168.
[5] W. Li and K. Kaneko, Europhys. Lett. 17 (1992) 655.
[6] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[7] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng and H.E. Stanley, Fractals in Science, eds. A. Bunde and S. Havlin (Springer-Verlag, Berlin, 1994) 49–83.
[8] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Phys. Rev. E 47 (1993) 4514.
[9] S.M. Ossadnik, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.-K. Peng, M. Simons and H.E. Stanley, Biophys. J. 67 (1994) 64.
[10] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Phys. Rev. Lett. 73 (1994) 3169–3172.
[11] J.D. Watson, M. Gilman, J. Witkowski and M. Zoller, Recombinant DNA (Scientific American Books, New York, 1992).
[12] E.W. Montroll and M.F. Shlesinger, The wonderful world of random walks, in: Non-equilibrium Phenomena II. From Stochastics to Hydrodynamics eds. J.L. Lebowitz and E.W. Montroll (North-Holland, Amsterdam, 1984) pp. 1–121.
[13] M.Ya. Azbel, Phys. Rev. Lett. 31 (1973) 589.
[14] C.L. Berthelsen, J.A. Glazier and M.H. Skolnick, Phys. Rev. A 45 (1992) 8902.
[15] A.N. Kolmogorov, Local structure of turbulence in fluid for very large Reynolds numbers, Transl. in Turbulence, S.K. Friedlander and L. Topper, eds. (Interscience Publishers, New York, 1961) pp. 151–155.
[16] B.B. Mandelbrot, The Fractal Geometry of Nature, (Freeman, San Francisco, 1982).
[17] J. Beran, Statistics for Long-Memory Processes (Chapman and Hall, New York, 1994).
[18] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L. Goldberger, Phys. Rev. E 49 (1994) 1685.
[19] E.C. Uberbacher and R.J. Mural, Proc. Natl. Acad. Sci. USA 88 (1991) 11261.
[20] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons and H.E. Stanley, Phys. Rev. E 51 (1995) 5084.
[21] A. Arneodo, E. Bacry, P.V. Graves and J.F. Mugy, Phys. Rev. Lett. 74 (1995) 3293.
[22] S. Wolfram, Comm. Math. Phys. 96 (1984) 15.
[23] A. Schenkel, J. Zhang and Y.-C. Zhang, Fractals 1 (1993) 47.
[24] M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham and N. Shnerb, Fractals 2 (1994) 7;
W. Ebeling, Physica A 215 (1995) 233.
[25] G.K. Zipf, Human Behavior and the Principle of Least Effort (Addison-Wesley, New York, 1949).

[26] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Phys. Rev. E (1995) (in press).
[27] Y. Liu et al., to be published.
[28] E.N. Trifonov, Bull. Math. Bio. 51 (1989) 417.
[29] S.V. Buldyrev et al., to be published.
[30] A. Czirók, R.N. Mantegna, S. Havlin and H.E. Stanley, Phys. Rev. E 52 (1995) pp. 446–452.
[31] S. Havlin, Physica A 216 (1995) 148.