Physica A 200 (1993) 4–24
North-Holland

SDI: 0378-4371(93)E0225-4

PHYSICA A

Long-range power-law correlations in condensed matter physics and biophysics

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

[a] *Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA*
[b] *Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, USA*
[c] *Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*
[d] *Department of Biology, MIT, Cambridge, MA 02139, USA*

We discuss the appearance of long-range power-law correlations in various systems of interest to condensed matter physicists and biophysicists, with emphasis on the recent discovery of long-range correlations in DNA sequences that contain non-coding regions.

1. Introduction

For what basic physics advances will the twentieth century be remembered? Certainly the first half will be known principally for the discovery of quantum mechanics. The second half witnessed the developed of a myriad of applications of quantum mechanics, without which much of everyday life would not be recognizable. But what are the *basic* advances in fundamental understanding of the workings of nature?

Here, we shall exemplify one such basic advance – the discovery of long-range power-law correlations 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" [1–6]. 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 quantity these with a critical exponent (called α in this paper). Quantification of different behavior allows us to recognize similarities between different systems, thereby eventually leading to recognizing underlying unifications that might otherwise have gone unnoticed. For example, as soon as phenomena occurring near critical points were quantified with critical exponents, it was recognized that

Elsevier Science Publishers B.V.

the entire "zoo" of critical phenomena partitioned itself neatly into a relatively small number of distinct "universality classes".

Our intuition tells us that correlations should decay exponentially, not as power laws. Consider, e.g., a set of two-state Ising spins in dimension one ($d = 1$) with interactions J between neighbors. If C_1 denotes the correlation function for two spins that are nearest neighbors, then intuition tells us that the correlation function for any two spins separated by a distance r is [7]

$$C_r = (C_1)^r = e^{-r/\xi} , \tag{1a}$$

where the second equality in (1a) serves to define the correlation length $\xi \equiv -1/\log C_1$.

For $d = 1$ site percolation, C_r denotes the pair connectedness, the probability that a site at position r is both occupied and also connected by a string of occupied sites to an occupied site at the origin [8]. Again, (1a) holds, but now with $C_1 = p$, the probability that a site is occupied.

Our simple intuition, that correlations decay exponentially because of the fashion in which order is "propagated", seems to always work – except at the critical point, where the exponential decay of (1a) gives over to a power law decay

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

The difference between (1a) and (1b) is profound: (1a) states that there is a characteristic length ξ fixed by the strength of the nearest-neighbor correlation C_1, while (1b) states that there is no characteristic length at all.

Can we intuitively understand how it is possible to find a *non-exponential* decay of correlations? At first sight, it might appear that whenever we increment the distance between two spins by one lattice constant, the correlation should decrease by roughly the same factor, but this intuition leads immediately to exponential decay. A possible resolution to this paradox stems from the fact that near a critical point, "information" propagates from a spin at the origin to a spin at position r *not via a single path* (as for $d = 1$), but *rather via an infinite number of paths*; some of these paths are explicitly enumerated in fig. 9.4 of ref. [7]. Ornstein and Zernike [9] recognized this fact, but approximated the fashion in which "order is propagated" and so obtained predictions that today we call "classical" (fig. 7.5 of ref. [7]). Exact enumeration methods, such as high-temperature series expansions, take into account exactly such paths up to a certain length k_{max}, where k_{max} is typically 20. To obtain power-law correlations, the exact results for $k < 20$ are extrapolated to obtain an estimate of the behavior for all k. In some sense, although the correlation along each path *decreases* exponentially with the length of the path,

the number of such path *increases* exponentially. Therefore, the net effect is that there arise longer range power-law correlations.

At one time, it was imagined that the "scale-free" case of (1b) was relevant to only a fairly narrow slice of physical phenomena – only to systems that had been "tuned" by exceedingly painstaking experimental work to be exactly at a critical point [7]. Now we appreciate the ubiquity of systems displaying scale-invariant behavior. First of all, any system examined on length scales smaller than the correlation length is likely to display power-law behavior (because all paths between the origin and r are relevant up to the correlation length, and these cancel out the exponential decay for $r < \xi$). Moreover, the number and nature of systems displaying power-law correlations has increased dramatically, including systems that no one might ever have suspected as falling under the umbrella of "critical phenomena". The latter part of the century has witnessed a veritable expulsion in the study, both experimental and theoretical, of such systems. The 1991 Nobel Prize was awarded to P.-G. de Gennes in part for his recognition that polymer systems behave analogously to systems near their critical points. The 1993 Wolf Prize will be awarded to Benoit Mandelbrot for the recognition of the "fractal geometry of nature". Another very prestigious Israeli prize, the 1993 Israel Prize, is being awarded this year to Shlomo Alexander, in large part for his discoveries that under appropriate conditions a wide range of systems obey scaling or scale invariance.

Indeed, many systems drive themselves spontaneously toward critical points. One of the simplest systems exhibiting such "self-organized criticality" [10] is invasion percolation, a generic model that has recently found applicability to describing anomalous behavior of rough interfaces [11]. Instead of occupying all sites with random numbers below a pre-set parameter p, in invasion percolation one "grows" the incipient infinite cluster right at the percolation threshold by the trick of occupying always the perimeter site whose random number is smallest. Thus small clusters are certainly not scale-invariant and in fact contain sites with a wide distribution of random numbers. As the mass of the clusters increases, the cluster becomes closer and closer to being scale invariant or "fractal". One says that such a system drives itself to a "self-organized critical state' [10].

The list of systems in which power law correlations appear has grown rapidly in recent years, including models of turbulence and even earthquakes [12]. What do we anticipate for biological systems? Generally speaking, when "entropy wins over energy" – i.e., randomness dominates the behavior – we find power laws and scale invariance. Biological systems sometimes are described in language that makes one think of a Swiss watch. Mechanistic or

"Rube Goldberg" descriptions must in some sense be incomplete, since it is only some appropriately chosen averages that appear to behave in a regular fashion. The trajectory of each individual biological molecule is of necessity random – albeit correlated. Thus one might hope that recent advances in understanding "correlated randomness" [13–16] could be relevant to biological phenomena. While there have been reports of scale invariant phenomena in isolated biological systems – ranging from the fractal shapes of neurons [17] to long-range correlations in heart beat intervals [18], human writings [19], and the stock market [20] – there has been no systematic study of *biological* system that displays power-law correlations.

Here we will attempt to summarize the key findings of some recent work [21–40] suggesting that under suitable conditions – the sequence of base pairs or "nucleotides" in DNA also displays power-law correlations. The underlying basis of such power-law correlations is not understood at present, but it is least possible that this reason is of as fundamental importance as it is in other systems in nature that have been found to display power-law correlations.

2. Discovery of long-range correlations in DNA

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* [21]. For the conventional one-dimensional random walk model [41], a walker moves either "up" $[u(i) = +1]$ or "down" $[u(i) = -1]$ one unit length for each step i of the walk [1]. 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 [14–16].

One definition of the DNA walk is that the walker steps "up" $[u(i) = +1]$ if a pyrimidine (C or T) occurs at position a linear distance i along the DNA chain, while the walker steps "down" $[u(i) = -1]$ if a purine (A or G) occurs at position i. Other definitions are discussed in the caption to fig. 7. 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).

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

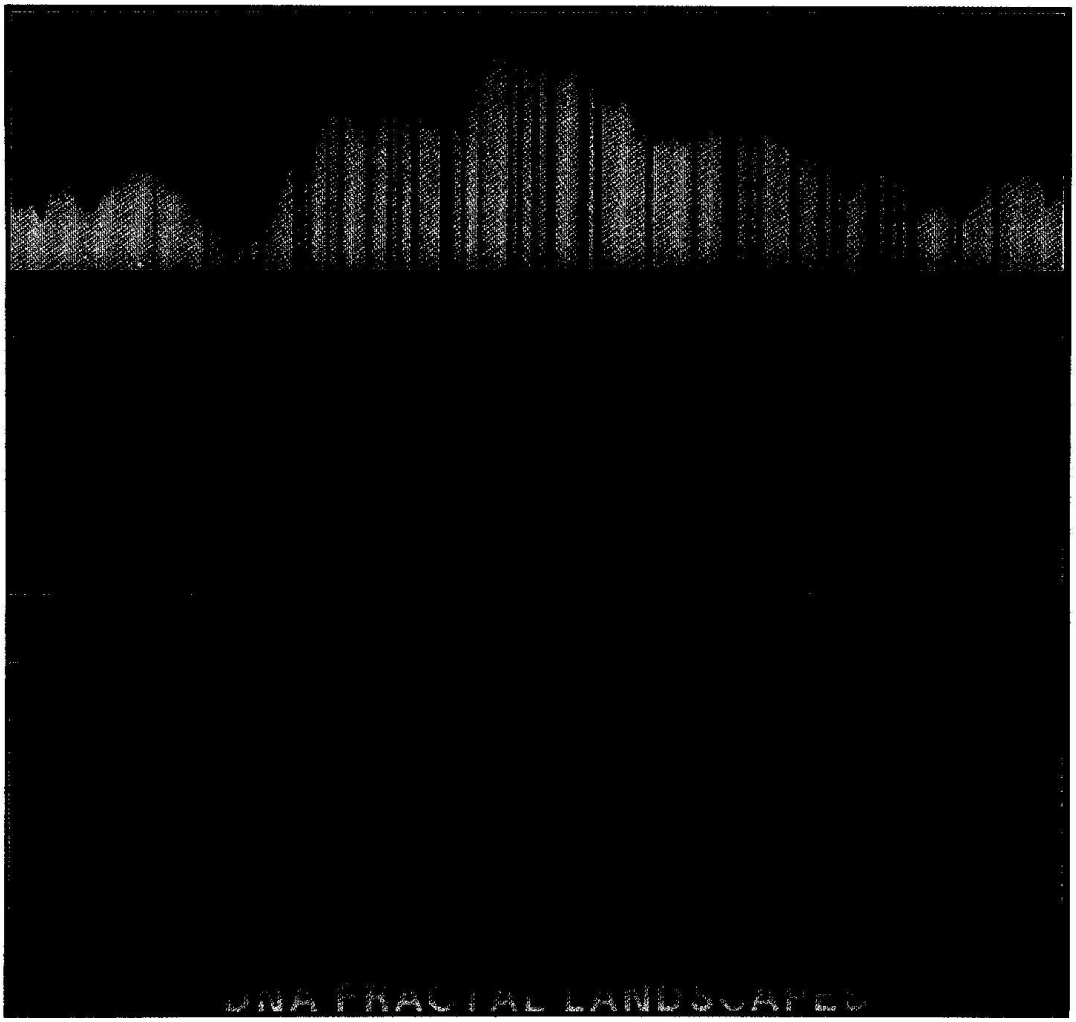

Fig. 1. The DNA walk representations of (top) human β-cardiac myosin heavy chain gene sequence, showing the coding regions as vertical dark bars, (middle) the spliced together coding regions, and (bottom) the bacteriophage lambda DNA which contains only coding regions. Note the more complex fluctuations for (top) compared with the coding sequences (middle) and (bottom). We found that for almost all coding sequences studied that there appear regions with one strand bias, followed by regions of a different strand bias. The fluctuation on either side of the overall strand bias we found to be random, a fact that is plausible by visual inspection of the DNA walk representations. We used different step heights for purine and pyrimide in order to align the end point with the starting point. This procedure is for graphical display purposes only (to allow one to visualize the fluctuations more easily) and is not used in any analytic calculations.

An important statistical quantity characterizing any walk [41] is the root mean square fluctuation $F(l)$ about the average of the displacement; $F(l)$ is defined in terms of the difference between the average of the square and the square of the average,

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

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

The mean square fluctuation is related to the auto-correlation function $C(l) \equiv \overline{u(l_0)\, u(l_0 + l)} - \overline{u(l_0)}^2$ through the relation: $F^2(l) = \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} C(j - i)$. The calculation of $F(l)$ can distinguish three possible types of behavior. (i) If the base pair sequence were random, then $C(l)$ would be zero on average [except $C(0) = 1$], so $F(l) \sim l^{1/2}$ (as expected for a *normal* random walk). (ii) If there were a local correlation extending up to a characteristic range R (such as in Markov chains), then $C(l) \sim \exp(-l/R)$; nonetheless *the asymptotic behavior* $F(l) \sim l^{1/2}$ *would be unchanged from the purely random case.* (iii) If there is no characteristic length (i.e., if the correlation were "infinite-range"), then the scaling property of $C(l)$ would not be exponential, but would most likely to be a power-law function, and the fluctuations will also be described by a power law

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

with $\alpha \neq 1/2$. Fig 1(top) shows a typical example of a gene that contains a significant fraction of base pairs that do *not* code for amino acids [42–44]. It is immediately apparent that the DNA walk has an extremely jagged contour, which we shall see corresponds to long-range correlations. Fig. 2 shows double logarithmic plots of the mean square fluctuation function $F(l)$ as a function of the linear distance l along the DNA chain for the three *randomly chosen* sub-sequences (1000 base pairs of each) from fig. 1(top).

The fact that the data are linear on this double logarithmic plot confirms that $F(l) \sim l^{\alpha}$. 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 result that there exists long-range correlation.

Peng et al. also addressed the question of whether the long-range correlation properties are different for coding and non-coding regions of a DNA sequence [21], a point that is currently the subject of some continuing debate [24,28]. Fig. 1(middle) shows the DNA walk for a sequence formed by splicing together the coding regions of the DNA sequence of this same gene. Fig. 1(bottom) displays the DNA walk for a typical sequence with only coding regions. In contrast to fig. 1(top), these coding sequences have less jagged contours, suggesting a shorter-range correlation.

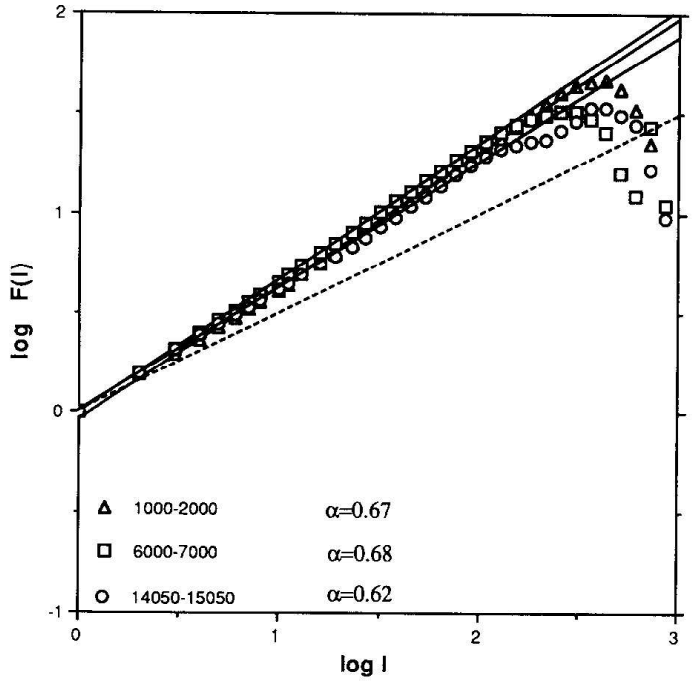

Fig. 2. Double logarithmic plots of the mean square fluctuation function $F(l)$ as a function of the linear distance l along the DNA chain for three *randomly chosen* sub-sequences (1000 base pairs of each) from fig. 1a. The dashed line has slope 0.5, corresponding to the expectation if the correlations were only short-range.

To analyze the middle and bottom parts, we first observe that for almost all sequence that we studied, purine-rich regions (compared to the average concentration over the entire strand) alternate with pyrimidine-rich regions, corresponding to the "up-hill" and "down-hill" portions of the DNA walk. To take into account the fact that the concentrations of purines and pyrimidines are not constant throughout the single strand base pair sequence, each DNA walk representation is partitioned into three segments demarcated by the global maximum ("max") and minimum ("min") displacements. Then we analyze the fluctuation within each segment separately. We found that for the middle and bottom parts that $\alpha = 1/2$ to within the level of fluctuation associated with the finite length of chain analyzed.

3. Possible artifacts

Naturally, we worried constantly that there was some possible "artifact" in the analysis that would invalidate our finding that spliced together coding regions as well as sequences containing only coding regions are uncorrelated, while sequences containing non-coding "junk" possess long-range power-law

correlations. Hence we carried out numerous tests, some of which are reported on below. Since the calculation of $F(l)$ for the DNA walk representation thus has the potential of providing a new, *quantitative* method to distinguish coding and non-coding regions, it is particularly important to be certain that there are no artifacts of this method.

3.1. Sampling statistics

In order to see if this scaling behavior is "universal", we first applied our analysis to more than 100 representative DNA sequences across the phylogenetic spectrum (comprising altogether some 10^7 base pairs analyzed – by contrast, Voss [24] has confirmed our findings using 25 000 DNA sequences). The result of some of this analysis is provided in table 1 of ref. [21]. The results confirm that long-range correlations ($\alpha > 1/2$) are characteristic of DNA containing non-coding base pairs but for coding sequences, $\alpha \cong 0.50 \pm 0.05$.

3.2. Biased but uncorrelated walks

One of the first concerns that we met in presenting our work to others was the confusion that for a biased but uncorrelated random walk, α would also be larger than 1/2. Much DNA material contains regions that have more purines (or pyrimidines) than 50%; this phenomenon is termed "strand bias". Therefore, we studied a variety of "artificial" base sequences in which we deliberately introduce a controlled measure of strand bias. These artificial sequences nonetheless all have $\alpha = 1/2$.

To demonstrate this fact explicitly, we constructed fig. 3. Fig. 3a shows an *unbiased* random walk with exactly the same number of steps, 2941, as in the case of the same gene analyzed in fig. 4. The data clearly corroborate the expected result, $\alpha = 1/2$. Fig. 3b shows a 2941-step *biased* random walk, and again the data clearly corroborate the expected result, $\alpha = 1/2$. Fig. 3c shows a 2941-step *correlated* random walk, with correlation parameter 0.61, and now the data corroborate the expected result, $\alpha = 0.61$. Thus, long-range correlations bear no relation whatsoever to strand bias: *The exponent is determined by the correlations, not by the bias.* What can in fact produce artifacts in estimating α is the abrupt change of bias, which is discussed in section 4 below.

3.3. Effect of finite sequence length

To demonstrate how the finite size sample affects the statistical quantity F, we show in fig. 4 plots of F for human metallothionein for three different sizes. First we *randomly* choose a sub-sequence of 300 nucleotides from the entire

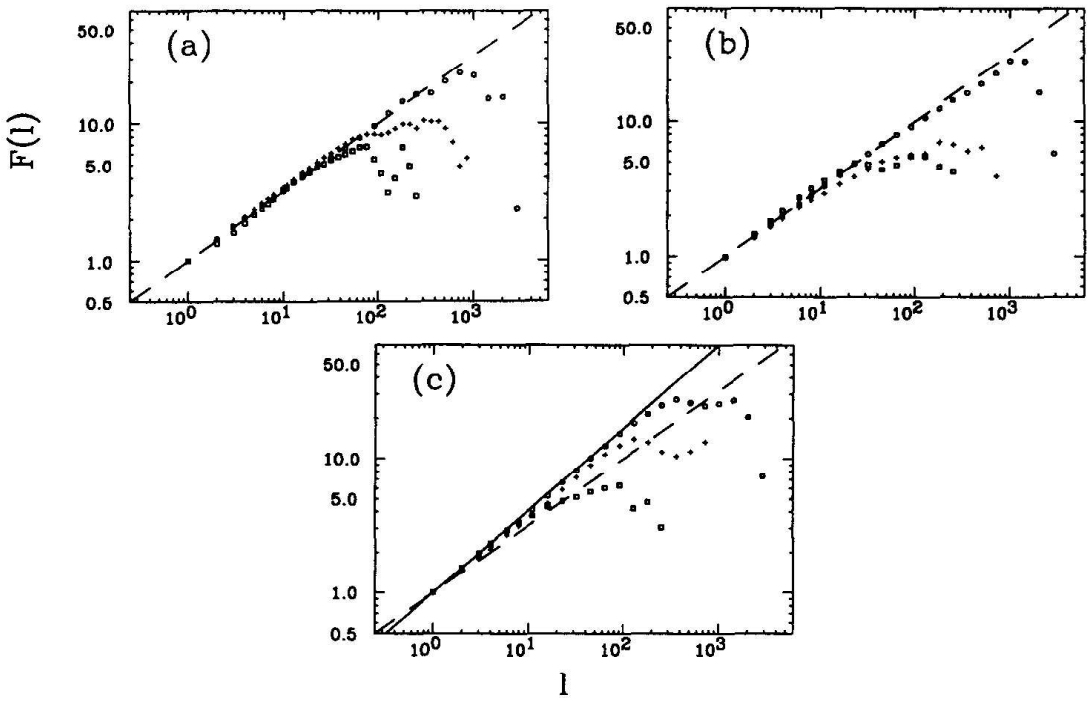

Fig. 3. $F(l)$ versus l for three different type of artificial sequences: (a) *Unbiased* random uncorrelated sequence (i.e., 50% purines); (b) *biased* but still uncorrelated sequence (with 60% purines); and (c) *correlated* sequence with correlation parameter 0.61 (with 50% purines). The dashed line has slope 0.5, while the solid line has slope 0.61. The different symbols represent different size of the sequences: The entire sequence of 2941 nucleotides (circles), the sub-sequences of 1000 nucleotides (crosses) and 300 nucleotides (open squares).

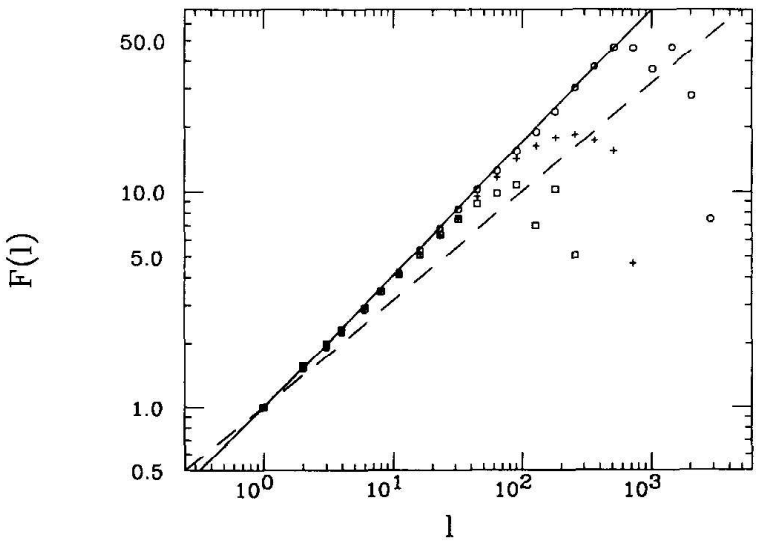

Fig. 4. $F(l)$ versus l for three different sizes of samples: The whole sequence of 2941 nucleotides (circles), the sub-sequences of 1000 nucleotides (crosses) and 300 nucleotides (open squares). The dashed line has slope 0.5, while the solid line has slope 0.61. Note that the linearity in all cases extends up to a fraction of about 1/10 the sample size, a fact familiar to workers involved in statistical analysis of this sort.

gene sequence to calculate the quantity F within this small sample (open square). Second, we choose (again randomly) a piece of the sequence of 1000 nucleotides from the same gene, repeat the analysis and plot the results (crosses). Finally, we analyze the entire sequence (2941 nucleotides) and plot the results (circles). The "fall-off" in the straight line behavior after the distance l reaches approximately one tenth of the size of the sample is typical of all fractal analyses. It is also found in sequences of correlated and random numbers (fig. 3). The trend of obtaining longer regions of straight line behavior for larger size samples is what one expects on statistical grounds.

One can always worry that the long-range features will disappear for longer DNA sequences. As evidence that the long-range feature does not disappear for larger samples, ref. [21] analyzed the entire human beta globin region (73326 bases) and found linearity up to $l \approx 7000$. Recently, Munson et al. [26] analyzed the entire yeast III chromosome (315 000 bases) [45] and found linearity up to $l \approx 31\,500$. In general, the data are linear over a range that is about a factor of ten less than the range of the data. This increase in statistical error when there are less than roughly 10 independent data sets is usually found for analyses of this sort. Thus, e.g., if a gene has 10 000 nucleotides, then there are only 10 *independent* sets of data obtained when the calipers are separated by a distance 1000.

One can worry about the apparent lack of consistency between values of α measured for different genes, or even for different regions within the same gene. Peng et al. have recently carried out a systematic study of the fluctuations in the correlation exponents obtained [35]. They indeed find prominent sample-to-sample variations in the scaling exponent, as well as variations within a single sample. To determine if these fluctuations may result from finite system size, they generate correlated random sequences of comparable length and study the fluctuations in this control system. Peng et al. find that the DNA exponent fluctuations are consistent with those obtained from the control sequences having long-range power law correlations.

3.4. Other methods of measuring long-range correlations

One can also worry that the apparent long-range correlation is some artifact of the DNA walk method itself. To compare the fluctuations of α in our DNA walk method with those found in other methods, Peng et al. used two standard methods to study the correlation property of sequences, namely the correlation function $C(l)$ and the power spectrum $S(f)$. The power spectrum density, $S(f)$, is obtained by (a) Fourier transforming the sequence $\{u(i)\}$ and (b) taking the square of the Fourier component. For a stationary sequence, the power spectrum is the Fourier transform of the correlation function. If the

correlation decays algebraically (not exponentially), i.e., there is no characteristic scale for the decay of the correlation, as we found in the non-coding DNA sequences, then we expect power-law behavior for both the power spectrum and the correlation function,

$$S(f) \sim (1/f)^{\beta} \tag{2b}$$

and

$$C(l) \sim (1/l)^{\gamma} . \tag{2c}$$

The correlation exponents α, β and γ are not independent, since [14,15]

$$\alpha = (1+\beta)/2 = (2-\gamma)/2 . \tag{3}$$

For a typical DNA sequence of finite length, both the correlation function and power spectrum are fairly noisy, but the estimates of β and γ obtained are consistent with those calculated from the DNA walk method (see fig. 5). The reason for the smaller fluctuations of α in the DNA walk method is due to the fact that $F^2(l)$ is a double summation of $C(l)$. Thus it would seem that the original DNA walk method is more useful due to reduced noise.

4. Difference between correlation properties of coding and non-coding regions

Our initial report [21] on long-range (scale-invariant) correlations in DNA sequences has generated contradicting responses. Some [22,24,26] support our initial finding, while some [23,28,32,34] disagree. Furthermore, the conclusions of refs. [24] and [23,28,32] are inconsistent *with one another* in that [23] and [34] doubt the existence of long-range correlations (even in non-coding sequences) while [23] and [28,32] conclude that even coding regions display long-range correlations ($\alpha > 1/2$). Prabhu and Claverie [28] claim that their analysis of the putative *coding* regions of the yeast chromosome III [45] 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 fluctuations expected for analysis of rather short sequences coding regions typically consist of only a few lengthy regions of alternating strand bias. Hence scaling analysis cannot be applied reliably to the entire sequence but only to sub-sequences.

Figure 6a displays our analysis of a typical coding sequence, consisting of two large sub-regions, each with different strand bias; the first (roughly 22 000 nucleotides) is G rich (compared to the average concentration of the entire

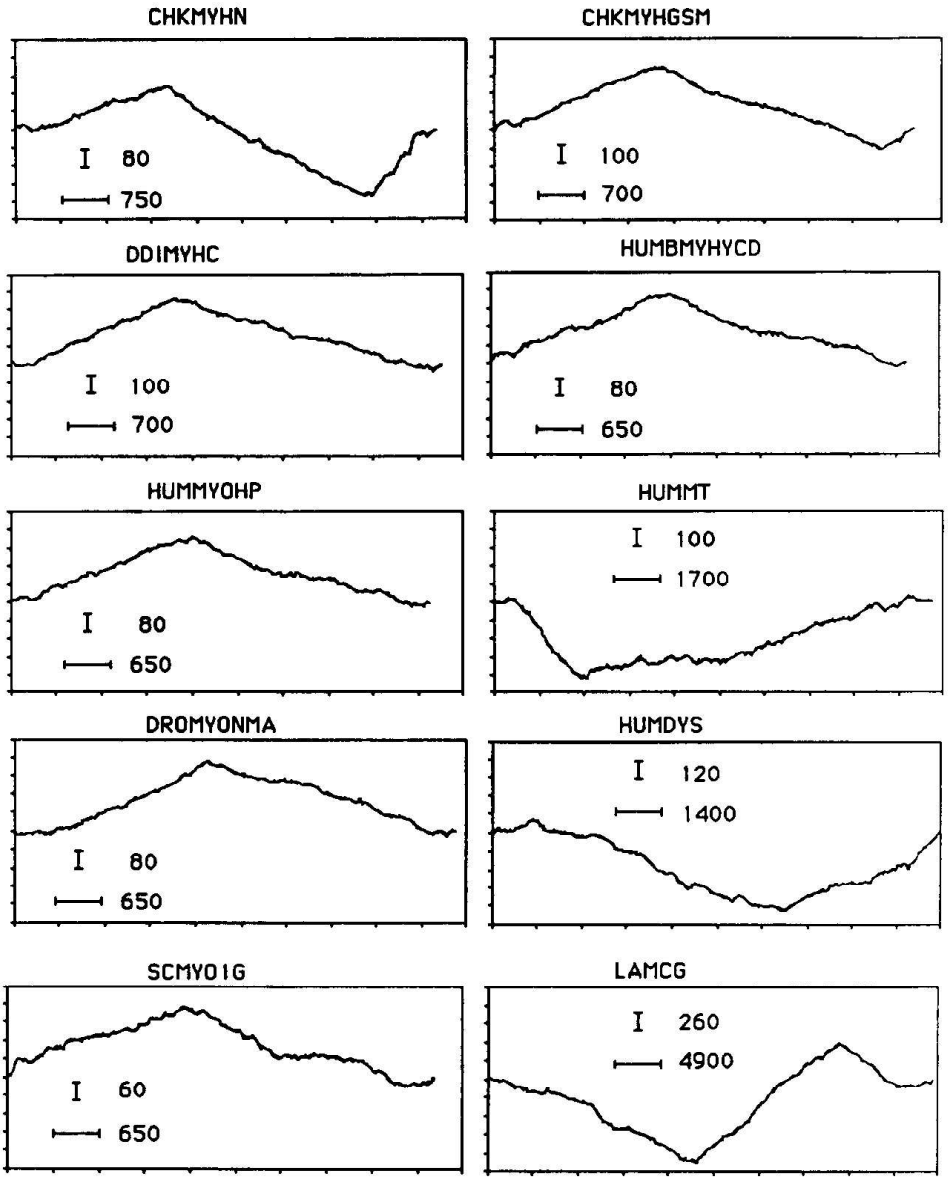

Fig. 5. Ten different coding sequences, plotted in the same form as fig. 1. Note that for all ten cases, there appear regions with one strand bias, followed by regions of a different strand bias. The fluctuation on either side of the overall strand bias is found to be random, a fact that is plausible by looking at these DNA walk representations. The bias introduced by the change in concentration of purine and pyrimidine would not be eliminated by the average term in eq. (2) of the manuscript if pieces of different bias would be analyzed together.

sequence), the second G poor. The scaling analysis of $F(l)$ (fig. 6c) on the *entire* sequence shows a *crossover* behavior, i.e., the log–log plot of the $F(l)$ versus l line has an initial slope 0.5 and curving toward 1 at larger values of l. This crossover behavior is typical of many physical systems having a characteristic scale. In this case, this characteristic length scale is associated with the

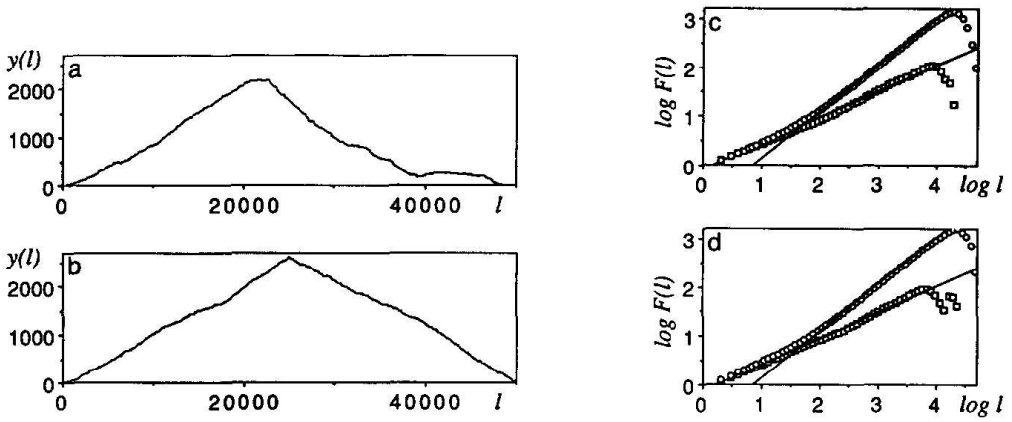

Fig. 6. (a) Landscape representation of the intronless sequence of the complete genome for lambda phage (GenBank: LAMCG, 48502 base pairs). Each "up" step corresponds to a guanine (G) and a "down" step corresponds to any one of the other three nucleotides (A, C or T). For graphical representation we plot the DNA walk such that the end point has the same vertical displacement as the starting point (for the statistical analysis, we use the original definitions). (b) Landscape for a biased random walk, where the bias is similar to that of the DNA sequence: in the first half, $p_{up} = 0.3$ (>1/4, the value expected in the absence of strand bias) followed by $p_{up} = 0.2$ (<1/4) in the second half. (c) The rms fluctuation in landscale altitude [2], $F(l)$, for the full genome (○, slope $\alpha \equiv (\beta + 1)/2 = 0.95$ for $l > 100$) and for the first sub-region (□, slope 0.54). (d) The rms fluctuation for the biased random walk (slope 0.96) and for the first sub-region (slope 0.54). Similar behavior was observed for the DNA walk walk with the purine–pyrimidine rule (step "up" for C and T; step "down" for A and G).

length of the two regions of strand bias. However, when the effects of this strand bias are first removed by *separately analyzing the sub-regions*, then we find α close to 0.5, indicating no long-range correlations.

We also calculate $F(l)$ for an artificial "control" sequence consisting of a 50 000-step biased random walk with similar strand biases as in the two regions of the DNA sequence (fig. 6b). We observe the same crossover behavior when the *entire* sequence was analyzed, but obtain the correct exponent $\alpha = 0.5$ when *each sub-region is separately analyzed* (fig. 6d). Figs. 6c, d also show that failure to correct for the crossover due to alternating regions of strand bias gives rise to a larger slope (upper curves) at larger values of l and hence misleadingly large values for the correlation exponents.

The power spectrum $S(f)$ for the *entire* sequence has an initial region of $1/f^{\beta}$ behavior at low frequency (that could be misinterpreted as indicative of long-range correlation, $\beta \neq 0$) followed by a flat region (indicative of no correlation or "white noise"). However, if the effects of strand bias are first removed by *separately analyzing the sub-regions*, then we find a flat $S(f)$ – and hence the correct correlation exponent $\beta = 0$. We also calculated $S(f)$ for a "control" consisting of a 50 000-step biased random walk with similar strand biases as in the two regions of the gene (fig. 6b), and found a misleading