# Fractal Landscapes in Biological Systems

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

[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, ISRAEL

**Abstract**

The purpose of this opening talk is to describe an example of recent progress in applying fractal concepts to biological systems. We first briefly review several biological systems, and then focus on the fractal features characterized by the long-range correlations found recently in DNA sequences containing *non-coding* material. We also discuss the evidence supporting the finding that for sequences containing only *coding* regions, there are no long-range correlations. Finally, we discuss the finding that the exponent $\alpha$ characterizing the long-range correlations increases with evolution.

## 1. Introduction

In the last decade it was realized that many biological systems have no characteristic length scale, thus having fractal or, more generally, self-affine properties [1]. In contrast to compact objects, fractal objects are almost entirely composed of "surface". Thus, fractals have a very large surface area. This observation explains why fractals are of great importance in biology, where surface phenomena are of crucial importance. For example, matter exchange takes place across membranes, and therefore requires large contact areas of the participating systems.

Lungs exemplify this feature. The surface area of a human lung is as large as a tennis court. A lung is made up of self-similar branches with many scale lengths, which is the defining attribute of a fractal surface. The efficiency of the lung is enhanced by this fractal property, since at each breath oxygen and carbon dioxide have to be exchanged at the lung surface. The structure of the bronchial tree has been quantitatively analyzed using fractal concepts [1-2].

A second example is blood vessels. Blood must be carried to all the cells of the body. For this purpose blood vessels must have fractal properties. The diameter distribution of blood vessels ranging from capillaries to arteries follow a power-law distribution which is one of the main characteristics of fractals. Specific examples include the chick embryo circulatory system, and the human retinal circulatory system [3].

Considerable attention in the biological community has arisen from the possibility that neuron shape can be quantified using fractal concepts. For example, Smith et al. [4] studied the fractal features of vertebrate central-nervous-system neurons in culture and found that the fractal dimension is increased as the neuron is more developed. Caserta et al.[5] showed that the shapes of quasi-two-dimensional retinal neurons can be characterized by a fractal dimension $d_f$. They found for fully developed neurons *in vivo*, $d_f = 1.68 \pm 0.15$, and suggest that the growth mechanism for neurite outgrowth bears a direct analogy with the growth model called diffusion-limited-aggregation (DLA).

One of the most amusing examples where fractal concepts come into play concerns the possibility [6] that HCl when released under pressure by the secretory glands indeed crosses the viscous lining of the stomach using the principles of viscous fingering that govern the breakdown of any viscous liquid when a less viscous liquid is forced under pressure through it. The DLA-type model

governing viscous fingering thereby serves to resolve the age-old paradox *"Why doesn't the stomach digest itself?*

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" [7-12]. 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 these with a critical exponent. 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.

Our intuition tells us that correlations should decay exponentially

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

where the second equality in (1a) serves to define the correlation length $\xi \equiv -1/\log C_1$. 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 [13], 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 $\xi$ 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 molecules 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 molecule at the origin to a molecule at position $\vec{r}$ *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. [13]. Ornstein and Zernike [14-15] 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.[13]). 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 the dominant exponential decay is magically "canceled", leaving the sub-dominant but longer range power-law correlations—which are in fact observed.

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 [13]. 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 explosion 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* was awarded to Benoit Mandelbrot for the recognition of the "fractal geometry of nature." Another very prestigious Israeli prize, the *1993 Israel Prize*, has been 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.

The list of systems in which power law correlations appear has grown rapidly in recent years, including models of turbulence and even earthquakes [16]. 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" [17-20] 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 [4-5] to long-range correlations in heart beat intervals [21], human writings [22], and the stock market [23]—there has been no systematic study of a *biological* system that displays power-law correlations.

Here we will attempt to summarize the key findings of some recent work [24-43] suggesting that—under suitable conditions—the sequence of base pairs or "nucleotides" in DNA also displays power-law correlations. The underlying basis of such power law correlations is not understood at present, but it is at least possible that this reason is of as fundamental importance as it is in other systems in nature that have been found to display power-law correlations.

## 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* [24]. For the conventional one-dimensional random walk model [44], 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 [18-20].

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 $i$. 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. Figure 1 naturally motivates a quantification of this correlation by calculating the "net displacement" of the walker after $\ell$ steps, which is the sum of the unit steps $u(i)$ for each step $i$. Thus $y(\ell) \equiv \sum_{i=1}^{\ell} u(i)$.

An important statistical quantity characterizing any walk [44] is the root mean square fluctuation $F(\ell)$ about the average of the displacement; $F(\ell)$ is defined in terms of the difference between the average of the square and the square of the average,

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

of a quantity $\Delta y(\ell)$ defined by $\Delta y(\ell) \equiv y(\ell_0 + l) - y(\ell_0)$. Here the bars indicate an *average* over all positions $\ell_0$ in the gene. Operationally, this is equivalent to (a) taking a set of calipers set for

a fixed distance $\ell$, (b) moving the beginning point sequentially from $\ell_o = 1$ to $\ell_o = 2, \cdots$ and (c) calculating the quantity $\Delta y(\ell)$ (and its square) for each value of $\ell_o$, and (d) averaging all of the calculated quantities to obtain $F^2(\ell)$.

The mean square fluctuation is related to the auto-correlation function

$$C(\ell) \equiv \overline{u(\ell_0)u(\ell_0 + l)} - \overline{u(\ell_0)}^2 \tag{3a}$$

through the relation

$$F^2(\ell) = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} C(j - i). \tag{3b}$$

The calculation of $F(\ell)$ can distinguish three possible types of behavior.

(i) If the base pair sequence were random, then $C(\ell)$ would be zero on average [except $C(0) = 1$], so $F(\ell) \sim \ell^{1/2}$ (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 $C(\ell) \sim \exp(-\ell/R)$; nonetheless *the asymptotic behavior* $F(\ell) \sim \ell^{1/2}$ *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 $C(\ell)$ 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

$$F(\ell) \sim \ell^{\alpha} \tag{4a}$$

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 [45-46]. 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 $F(\ell)$ as a function of the linear distance $\ell$ along the DNA chain for a typical intron-containing gene.

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

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 [24], a point that is currently the subject of some continuing debate [27,31,37]. Figure 1b shows the DNA walk for a sequence formed by splicing together the coding regions of the DNA sequence of this same gene. Figure 1c displays the DNA walk for a typical sequence with only coding regions. In contrast to Fig. 1a, these coding sequences have less jagged contours, suggesting a shorter range correlation.

### 3. 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 $\alpha$ 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 $C(\ell)$ 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,

$$S(f) \sim (1/f)^{\beta}, \tag{4b}$$

and

$$C(\ell) \sim (1/\ell)^{\gamma}. \tag{4c}$$

The correlation exponents $\alpha$, $\beta$ and $\gamma$ are not independent, since [18,19]

$$\alpha = (1+\beta)/2 = (2-\gamma)/2. \tag{5}$$

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

Apart from the reduced noise mentioned above, one additional advantage of the DNA walk method [24] is that to find the exponent characterizing the long-range correlation one need not correct the data by subtracting the white noise, $S(\infty)$ [27]. Since there is no unambiguous method of estimating $S(\infty)$, this need to correct the data introduces an uncontrollable source of uncertainty.

**4. Difference Between Correlation Properties of Coding and Non-Coding Regions**

Our initial report [24] on long-range (scale-invariant) correlations in DNA sequences has generated contradicting responses. Some [25,27,29] support our initial finding, while some [26,31,35,37] disagree. Furthermore, the conclusions of Refs. [27] and [26,31,35] are inconsistent *with one another* in that [26] and [37] doubt the existence of long-range correlations (even in non-coding sequences) while [26] and [31,35] conclude that even coding regions display long-range correlations ($\alpha > 1/2$). Prabhu and Claverie [31] claim that their analysis of the putative *coding* regions of the yeast chromosome III [45] produce 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.

Prabhu and Claverie [31] noted that nonlinear curves were obtained for *even some intron-containing sequences*. The DNA walks for intron-containing sequences do not show any apparent length scale of strand bias; as noted above, there seems to be a broad distribution of lengths of strand bias. In principle, for this type of DNA walk, the min–max procedure is not necessary since there is no characteristic length that needs to be taken into account. For most *intron-containing* sequences, our analysis shows that there is a broad range of scaling (with $\alpha > 0.5$) when we study the entire sequence as a whole. For certain intron-containing genes, however, we find that the min–max procedure can extend the scaling region, leading to a larger range of constant-$\alpha$ behavior.

One might also wonder if the selection of "max" and "min" segments gives a bias to the calculations. The "max" and "min" operation is a systematic method to treat all DNA sequences on equal footing without applying any *a priori* knowledge of the sequence itself. Notice that although this procedure eliminates the problem of large scale variation of concentration for the coding sequences, it will not alter the result of a truly self-similar sequence. To obviate the need to perform the min-max partitioning. Peng et al [42] have recently applied the "bridge method" to DNA, and have also developed a new method specifically adapted to handle problems associated with non-stationary sequences which they term *detrended fluctuation analysis* (DFA).

To provide an "unbiased" test of the thesis that non-coding regions possess but coding regions lack long-range correlations, Ossadnik et al. [43] analyzed several uncorrelated and correlated control sequences of size $10^5$ nucleotides using the GRAIL neural net algorithm [46]. The GRAIL algorithm identified about 60 putative exons in the uncorrelated sequences, but only about 5 putative exons in the correlated sequences. We also used the beachcomber method [43], which shows pronounced dips in $\alpha$ in the region where genes are expected (Fig. 3).

Voss [27] has recently proposed that *coding* as well as non-coding DNA sequences display long-range power law correlations in their base pair (bp) sequences. This finding disagreed with our earlier analysis [24], claiming that coding DNA sequences do not display power-law correlations. However the discrepancy between [27] and [24] could have arisen because the analysis in [24] was based on partitioning the entire coding sequence into a few large subsequences of constant overall compositional bias. It is important to resolve this discrepancy, since Voss based his scientific conclusion ("immunity to errors on all scales") on his claim of power-law correlations in *coding* sequences [47].

Recently, Buldyrev et al. argued that the Voss proposal does not hold generally [48]. Specifically, they presented two counterexamples that clearly display *no* long-range correlations *when directly analyzed* (without partitioning into subsequences): (i) the complete genome of T7 bacteriophage (39936 bp), which contains *only* coding regions, and (ii) the Ti plasmid fragment (24,595 bp), which is believed to consist almost entirely of coding regions.

Fig. 4a shows the DNA walks for (i) and (ii). Fig. 4b shows $F(\ell)$, the fluctuation in rms amplitude; the slopes of the log-log plots, *fit over 3 decades*, are 0.53 and 0.49, indicating the absence of long-range correlation for both cases. Fig. 4c shows the power spectrum $S(f)$, which is almost perfectly flat (indicative of no correlation or "white noise"). The scaling behavior in Figs. 4b and 4c is markedly different than that found for genomic sequences containing substantial noncoding sub-regions, for which $F(\ell) \sim \ell^{\alpha}$ with $\alpha \approx 2/3$ and $S(f) \sim 1/f^{\beta}$ with $\beta \approx 1/3$ [24].

So why did Voss find $\beta = 1.02$ and 1.16, respectively, for phage and bacteria (which contain mostly coding regions)? A clue is apparent from comparing Voss' analysis for these cases (Fig. 3 of [27]) with his fits for the non-coding segments. We see that all except a few small-$f$ data points are well fit by a horizontal line, corresponding to $\beta = 0$ (no long-range correlation). The departure at small $f$ likely corresponds to the fact that most coding sequences contain uncorrelated subsegments—*with a characteristic length*—of alternating compositional bias. The DNA walks, therefore, resemble a spliced together string of *uncorrelated* but *biased* random walks. We confirmed this likely source of spurious low-$f$ behavior by calculations on artificial "control" sequences.

## 5. Fractal Landscapes and Molecular Evolution

Molecular evolutionary relationships are usually inferred from comparison of coding sequences, conservation of intron/exon structure of related sequences, analysis of nucleotide substitutions, and construction of phylogenetic trees [49]. The changes observed are conventionally interpreted with respect to nucleotide sequence composition (mutations, deletions, substitutions, alternative splicing, transpositions, etc.) rather than overall genomic organization.

Very recently, Buldyrev et al. [41] sought to assess the utility of DNA correlation analysis as a complementary method of studying gene evolution. In particular, they studied the changes in "fractal complexity" of nucleotide organization of a single gene family with evolution. A recent study by Voss [27] reported that the correlation exponent derived from Fourier analysis was lowest for sequences from organelles, but paradoxically higher for invertebrates than vertebrates. However, this analysis must be interpreted with caution since it was based on pooled data from different gene families rather than from the quantitative examination of any single gene family.

Buldyrev et. al [41] tested the hypothesis that the fractal complexity of genes from higher animals is greater than that of lower animals, using single gene family analysis. They focuseded

their analysis on the genome sequences from the conventional (Type II) myosin heavy chain (MHC) family. Such a choice limits potential bias that may arise secondary to non-uniform evolutionary pressures and differences in nucleotide content between unrelated genes. They also used this technique to study the MHC gene family because of the availability of completely sequenced genes from a phylogenetically diverse group of organisms, and the fact that their relatively long sequences are well-suited to statistical analysis.

The landscape produced by DNA walk analysis reveals that each MHC cDNA consists of two roughly equal parts with significant differences in nucleotide content (Fig. 5). The first part that codes for the heavy meromyosin or "head" of the protein molecule has a slight excess of purines (52% purines and 48% pyrimidines); the second part that codes for the light meromyosin or "tail" has about 63% purines and 37% pyrimidines. The *absolute nucleotide* contents are not shown in the graphical representation of Fig. 5a because we subtract the average slope from the landscape to make relative fluctuations around the average more visible. Indeed, one can easily see from Fig. 5a that the relative concentration of pyrimidines in the first part ("uphill" region) of the myosin cDNA is much higher than in the second ("downhill" region).

As previously reported [24], we find that $\alpha \approx 1/2$ for all cDNAs (corresponding to no correlations or only short-range correlations), while all MHC genes containing introns have $\alpha > 1/2$, corresponding to long-range correlations. Figure 6 shows representative scaling plots of $F_d(\ell)$ vs $\ell$, where $F_d$ denotes the fluctuation quantity of (3b) defined using the "detrended fluctuation analysis" developed in [42]. Of note, the value of $\alpha$ is not strongly related to the presence of exons since "stitching together" intron sequences (by removing exons) produces a value similar to that of the full gene, for example, for the human MHC gene the value of $\alpha$ after the exons are removed is 0.593 versus 0.586 for the complete sequence, further supporting the view that the composition of non-coding elements is the principal source of long-range correlations in genomic sequences.

## 6. Insertion-Deletion Model

To gain some insight into possible evolutionary mechanisms that could increase the complexity of the landscapes and generate long-range correlations of DNA sequences, Buldyrev et al. [41] introduce a simple model that simulates conversion of originally coding regions into noncoding introns. The model is based on the hypotheses that genetic information was originally encoded in an mRNA molecule which was subsequently converted into a DNA sequence, and that this sequence underwent modifications due to mutagenesis and insertion of non-coding genetic material (introns) [50].

(i) To simulate cDNA sequences, one starts with a biased random walk of length $L$ with an overall excess of purines over pyrimidines corresponding to that observed in the cDNA sequences.

(ii) At each time step, one "mutates" the sequence by the following procedure:

  (a) Choose a random point in the sequence and cut a sub-sequence of length $n$ starting from that point, where the length $n$ is chosen from a power law distribution $\phi(n) \sim n^{-\beta}$ with $\beta \approx 2$ (between $L_o \approx 20$ and $L/2$). The reason for this power law distribution is that the cutting of a DNA segment most likely occurs when a loop is formed, and it is known that the distribution of loop sizes in a long polymer obeys a power law [51]. Choose another random point in the sequence at which we insert this length-$n$ sub-sequence.

  (b) With probability 0.5, a *strand substitution* may occur in this sub-sequence (i.e., all purines are substituted by pyrimidines and vice versa, thereby inserting a complementary strand).

  (c) To simulate retroviral insertion occurring, with some small probability $p_i$, the subsequence to be inserted is substituted by a random sequence of equal length with the same percentage of purines and pyrimidines as in the initial cDNA sequence.

Representative purine-pyrimidine landscapes generated by this model are shown in Fig. 7. After the first few iterations (Fig. 7a), the landscape is remarkably similar to that of primitive organisms such as phages (see Fig. 1 of [24]) and *E. Coli* which have only a small percentage of introns. The scaling behavior shows $F_d(\ell) \sim \ell^{1/2}$ as anticipated. After more iterations, the landscape seems visually more complex and the $\alpha$ measured for the first two decades increases, reminiscent of the increase noted in the MHC family from yeast to invertebrates to vertebrates. After roughly 1000 iterations, the value of $\alpha$ asymptotically approaches 0.60 as shown in Fig. 6b. The landscapes for the model and for the rat MHC gene become quite similar (cf. Figs. 7b and 7c).

Of note, if the model is iterated without the insertion of random biased sequences as assumed in rule (iic), the value of $\alpha$ will return to 0.5, indicating a random sequence. Insertion of biased random regions (according to a power-law distribution) maintains the exponent $\alpha > 0.5$. The importance of rule (iic) of the model is consistent with the hypothesized role of retroviral insertions in the genomes of high animals [52].

Furthermore, without strand substitution as implemented by rule (iib), no long-range correlation will appear. This mirror-image replacement mimics molecular evolution occurring by partial gene duplication or transposition [53] and the occurrence of "extinguished exons" [54]. In order to test our assumption of strand substitution we also analyzed an alternative DNA landscape in which nucleotides cytosine (C) and guanine (G) result in an up step, while adenine (A) and thymine (T) correspond to a down step. Since such walks cannot be affected by strand substitution, our model would predict the absence of long-range correlations. Indeed, our analysis of the fluctuation $F_d(\ell)$ for this modified DNA landscape does not exhibit as robust a power law correlation as for the original purine-pyrimidine rule. Another crucial assumption is the existence of an overall bias (either of purines or of pyrimidines) in the initial sequence; it is this bias that enables strand substitution to produce differences in nucleotide content. This assumption is consistent with our observation that most coding regions exhibit overall bias in their purine-pyrimidine concentration.

The mechanism of generating power-law correlations in this insertion-deletion model is related to the competition between two countervailing "forces." The deletion and insertion of segments in rule (iia) and (iib) tends to *randomize* the sequence, while the insertion of biased segment implemented by rule (iic) tends to *organize* the system. As the iteration proceeds, the newly inserted biased segment is then broken into smaller pieces of different bias (according to a power-law distribution). After a large (but finite) number of iterations (which depends on the parameters of the model), these two competing effects will tend to balance each other. At this point the system will exhibit power-law correlation.

The observed trend of $\alpha$ to increase with evolutionary status for the MHC family is also consistent with the predictions of the model: "higher" species that appeared more recently will tend to generate long-range correlations with a larger value of the parameter $\alpha$. Thus, vertebrate myosin is likely to be more "complex" than invertebrate myosin because the former incorporated genetic material from the latter species. This view of molecular evolution is consistent with the theory of punctuated equilibrium [55] that postulates rather rapid periods of change (occurring during speciation) followed by periods of stasis.

The finding that $\alpha$ increases with evolution contradicts a recent study by Voss [27] which paradoxically reported that the strength of nucleotide correlations (quantified by the power spectral scaling exponent $\beta$, which is uniquely related to $\alpha$) increases from organelle to invertebrates but then *decreases* for primates. This apparent discrepancy is likely due to the facts that Voss [27] (i) did not analyze *single* gene families with evolution, (ii) did not distinguish intron-containing vs intron-less sequences, and (iii) did not correct for large regions of "strand bias" (unequal numbers of purines and pyrimidines). We have found that if one does not take into account the *crossover* between two large (but uncorrelated) regions of strand bias as seen in all the MHC cDNAs (corresponding to the uphill and downhill regions in Fig. 1a), one can obtain a spuriously large value of $\alpha$.

Nee [26] proposed that it is the alternation of introns and exons (regions containing different nucleotide content) which modulates the long-range correlations. This idea is somewhat similar to the proposed model, but its main conclusion—that the sequence from which all exons have been cut does not exhibit long-range correlations—appears to be incorrect. In fact, intron sequences show long-range correlations as robust as those of complete genes with approximately the same exponent $\alpha$. In contrast, our model describes not only the intron-insertion process, but also the shuffling process within non-coding sequences (introns and intergenomic sequences). This shuffling process (not just the insertion of uncorrelated introns) leads to $\alpha > 0.5$ within single introns and intergenomic sequences, a fact that cannot be explained in the framework of the Nee hypothesis. Further, our model bears potential relevance to biological evolution, by providing a possible mechanism for transformation of primordial RNA molecules (currently considered to be the first to develop) into complex DNA sequences containing noncoding elements.

Finally, two major theories have been advanced to explain the origin and evolution of introns. One suggests that precursor genes consisted entirely of coding sequences and introns were inserted later in the course of evolution to help facilitate development of new structures in response to selective pressure, perhaps, by means of "exon shuffling" [56]. The alternative theory suggests that precursor genes were highly segmented and subsequently organisms not requiring extensive adaptation or new development or, perhaps, facing the high energetic costs of replicating unnecessary sequences, lost their introns [57-58]. Support for these hypotheses has remained largely conjectural; no models have been brought forward to support either process. The landscape analysis of the MHC gene family and the stochastic model presented in this study are most consistent with the former view.

## 7. Concluding Remarks

What is the biological meaning of our finding? If two nucleotides whose positions differ by 10,000 were uncorrelated, then there might be no meaning. However, if they are correlated there must be a reason! The method we describe points out a new element of DNA structure and suggests a possible fundamental role for the non-coding regions (termed "introns"). Our work may also reveal an interesting feature of the coding regions (termed "exons").

Before concluding, we note that the long-range correlations in DNA sequences are of interest because they may be an indirect clue to its three-dimensional structure [36,40] or a reflection of certain scale-invariant properties of long polymer chains [51]. In any case, the statistically meaningful long-range "scale-invariant" (see Fig.8) correlations in non-coding regions and their absence in coding regions will need to be accounted for by future explanations of global properties in gene organization and evolution.

It is awe-inspiring that remarkably complex objects in nature can be quantitatively characterized by a single number, $\alpha$. It is equally awe-inspiring that such complex objects can be described by various models with extremely simple rules. Whether physicists can actually contribute to understanding the principles on which biological fact is based, or whether we cannot, remains to be seen.

ACKNOWLEDGEMENTS We wish to thank F. Sciortino for important contributions in the initial stages of this project, and C. Cantor, C. DeLisi, J. M. Hausdorff, L. Liebovitch, R. D. Rosenberg, J. J. Schwartz, M. Schwartz, M.F. Shlesinger, and N. Shworak for valuable discussions. Partial support was provided to ALG by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute and the National Aeronautics and Space Administration, to C.-K.Peng by an NIH/NIMH Postdoctoral NRSA Fellowship, to M. Simons by the American Heart Association, and to HES by the National Science Foundation.

[1] B.J. West and W. Deering, Phys. Reports **xx**, xxx (1994); B.J. West, *Fractal Physiology and*

*Chaos in Medicine* (World Scientific, Singapore, 1990); A.L.Goldberger, D.R.Rigney and B.J. West, Sci. Am. **262**, 42 (1990); B.J.West and M.F.Shlesinger, Am.Sci. **78**, 40 (1990); B.J. West and A.L. Goldberger, Am. Sci., **75**, 354 (1987); L. S. Liebovitch, J. Freichbarg and J.P.Koniarek, Math. Biosci. **89**, 36 (1987); A.L. Goldberger and B.J. West, Yale J. Biol. Med. **60**, 421 (1987); B.J.West and A.L. Goldberger, J. Appl. Physiol., **60**, 189 (1986).

[2] M.F.Shlesinger and B.J. West, Phys. Rev. Lett. **67**, 2106 (1991).

[3] A.A.Tsonis and P.A. Tsonis, Perspectives in Biology and Medicine, **30**, 355 (1987); F. Family, B.R.Masters, D.E.Platt, Physica D **38**, 98 (1989).

[4] T.G.Smith, W.B. Marks, G.D. Lange, W.H.Sheriff Jr., E.A.Neale, J. Neuroscience Methods **27**, 173-180 (1989).

[5] F. Caserta, H. E. Stanley, W. D. Eldred, G. Daccord, R. E. Hausman, and J. Nittmann, Phys. Rev. Lett. **64**, 95 (1990); F. Caserta, R. E. Hausman, W. D. Eldred, H. E. Stanley, and C. Kimmel, Neurosci. Letters **136**, 198-202 (1992).

[6] K. R. Bhaskar, B. S. Turner P. Garik, J. D. Bradley, R. Bansil, H. E. Stanley, and J. T. LaMont, Nature **360**, 458 (1992).

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

[8] A. Bunde and S. Havlin, eds., *Fractals and Disordered Systems* (Springer-Verlag, Berlin, 1991); A. Bunde and S. Havlin, eds., *Fractals in Science* (Springer-Verlag, Berlin, 1994).

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

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

[11] T. Vicsek, *Fractal Growth Phenomena, Second Edition* (World Scientific, Singapore, 1992); J. Feder, *Fractals* (Plenum, NY, 1988).

[12] H. E. Stanley and N. Ostrowsky, eds., *On Growth and Form: Fractal and Non-Fractal Patterns in Physics*, Proceedings 1985 Cargèse NATO ASI, Series E: Applied Sciences, VOL. 100 (Martinus Nijhoff, Dordrecht, 1985).

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

[14] H. E. Stanley and N. Ostrowsky, eds., *Correlations and Connectivity: Geometric Aspects of Physics, Chemistry and Biology*, Proceedings 1990 Cargèse Nato ASI, Series E: Applied Sciences, VOL. 188 (Kluwer, Dordrecht, 1990).

[15] F. Zernike, *Physica* **7,** 565 (1940).

[16] P. Bak and M. Creutz, in A. Bunde and S. Havlin, eds., *Fractals in Science* (Springer-Verlag, Berlin, 1994).

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

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

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

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

[21] 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).

[22] A. Schenkel, J. Zhang and Y-C. Zhang, *Fractals* **1**, 47 (1993).

[23] R. N. Mantegna, *Physica A* **179,** 23 (1991).

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

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

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

[27] R. Voss, *Phys. Rev. Lett.* **68,** 3805 (1992).

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

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

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

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

[32] P. Yam, *Sci. Am.* **267**, no. 3, p. 23 (Sept. 1992)

[33] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Physica A* **191,** 25 (1992).

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

[35] C. A. Chatzidimitriou-Dreismann and D. Larhammar, *Nature* **361,** 212 (1993).

[36] A. Yu. Grosberg, Y. Rabin, S. Havlin, and A. Neer, *Biofisika (Russia)* **26,** 1 (1993); *Europhys. Lett.* **23**, 373 (1993).

[37] S. Karlin and V. Brendel, *Science* **259,** 677 (1993).

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

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

[40] A. S. Borovik, A. Yu. Grosberg, and M. D. Frank-Kamenetskii, *Nature* (in press)

[41] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M.Stanley, and M. Simons, *Biophys. J.* **65**, xxx (1993).

[42] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, *Science.* (submitted).

[43] S. M. Ossadnik, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley (submitted).

[44] E. W. Montroll and M. F. Shlesinger, in *Nonequilibrium Phenomena II. From Stochastics to Hydrodynamics*, J. L. Lebowitz and E. W. Montroll, eds. (North-Holland, Amsterdam, 1984), pp. 1–121.

[45] S. G. Oliver et al., *Nature* **357,** 38 (1992).

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

[47] In the power spectrum analysis for those sequences containing non-coding regions, subtracting of the white noise, $S(\infty)$, as performed in [27], gives more weight to the non-coding segments (correlated) than the coding segments (uncorrelated). See also H.E.Stanley, S.V.Buldyrev, A.L.Goldberger, S. Havlin, C-K Peng and M.Simons, Physica A **200**, 4 (1993) and refs. therein.

[48] S. V. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, F. Sciortino, M. Simons and H. E. Stanley, Phys. Rev. Lett. **71,** 1776 (1993).

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

[50] G. F. Joyce, *Nature* **338**, 217 (1989).

[51] J.des Cloizeaux, *J. Physique (Paris).* **41**,223 (1980); P.G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca NY, 1979).

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

[53] R. Schleif, *Science* **240**, 127 (1988).

[54] C. J. Jaworski and J. Piatigorsky, *Nature* **337**, 752 (1989).

[55] N. Eldredge and S. J. Gould, *In* Models in paleobiology. T. J. M. Schopf, editor. Freeman and Cooper Inc., San Francisco. pp. 82 (1972).

[56] W. Gilbert, *Nature* **271**, 501 (1978).

[57] P. Hagerman, *Ann. Rev. Biochem.* **59**, 755 (1990).

[58] W. F. Doolitle, in *Intervening Sequences in Evolution and Development*, E. Stone and R. Schwartz, Eds. (Oxford University Press, NY, 1990), pp 42-62.

**Fig. 1:** The DNA walk representations of (a) human $\beta$-cardiac myosin heavy chain gene sequence, showing the coding regions as vertical golden bars, (b) the spliced together coding regions, and (c) the bacteriophage lambda DNA which contains only coding regions. Note the more complex fluctuations for (a) compared with the coding sequences (b) and (c). 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 pyrimidine 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.

**Fig. 2:** (a) Double logarithmic plots of the mean square fluctuation function $F(\ell)$ as a function of the linear distance $\ell$ along the DNA chain for the rat embryonic skeletal myosin heavy chain gene ($\circ$) and its "intron-spliced sequence" ($\bullet$). (b) The corresponding local slopes, $\alpha_{local}$, based on pairs of successive data points of part (a). We see that the values of $\alpha$ are roughly constant. For this specific gene, the sequence with exons removed has an even broader scaling regime than the DNA sequence of the entire gene, indicated by the fact that part (a) is linear up to 10,000 nucleotides.

**Fig. 3:** Beachcomber plot for a typical section containing about 10% of the Yeast Chromosome III [45]—from base pair #30,000 to #60,000. The vertical yellow bars indicate the the set of base pairs forming identified genes (GenBank Release #76), while the white bars indicate less certain "putative genes" determined from analysis of open reading frames [45]. The exponent $\alpha$ is calculated by the beachcomber method [43]: We form an observation box of length 800, place this box at the beginning of the chromosome, and calculate the long-range correlation exponent $\alpha$ for the 800 base pairs lying inside this box. Then we move the box 75 base pairs further along the chromosome, and again calculate $\alpha$ for the 800 base pairs lying inside this box. Iterating this procedure, we obtain

315,000/75 = 4186 successive values of $\alpha$, each giving a "local" measurement of the degree of long-range correlation. The red curve is obtained using rule 1—a "down" step for A or G (purines) and an "up" step for C or T (pyrimidine). We see that when the box is covering coding regions, the value of $\alpha$ is generally small, while in between coding regions, there is frequently a peak in $\alpha$. If $\alpha$ were the same for coding and non-coding regions, we would expect the peaks and dips to occur with no evident correlation in the position of genes.

**Fig. 4:** (a) "DNA walks" [24] of nucleotide sequences; a random walker moves either up or down depending on whether the nucleotide at position $\ell$ is a pyrimidine or purine. (b) The rms fluctuation, $F(\ell)$, of the DNA walk displacement $y(\ell)$. (c) Power spectrum, $S(f)$, of the binary nucleotide sequences (pyrimidine= 1, purine= $-1$). Shown are two examples that display *no* long-range correlations: (i) the complete genome of T7 bacteriophage (GenBank name: PODOT7), which contains *only* coding regions, and (ii) the Ti plasmid fragment (ATACH5), which is believed to consist almost entirely of coding regions. The solid lines in (b) and (c) have slopes $\alpha = 1/2$ and $\beta = 0$ respectively; for comparison, dashed lines of slope 2/3 and -1/3 are also shown, corresponding to the typical behavior found for sequences containing non-coding regions [24]. The data for Ti ($\circ$) shifted on the plots for visual comparison.

**Fig. 5:** The DNA walk representations of (a) 8 cDNA sequences from the MHC family and (b) the corresponding genes. DNA landscapes are plotted so that the end points have the same vertical displacement as the starting points [24]. The graphs are for yeast, amoeba, *C. elegans*, brugia, drosophila, chicken, rat and human (from top to bottom, left to right). The shaded areas in (b) denote coding regions of the genes. The DNA walks for the genes show increasing "complexity" with evolution. In contrast, the cDNA walks all show remarkably similar crossover patterns due to sequential "up-hill" and "down-hill" slopes representing different purine/pyrimidine strand biases in the regions coding for the head and tail of the MHC molecule, respectively.

**Fig. 6:** Double logarithmic plot $F_d(\ell)$ function versus $\ell$ for (a) full MHC genes (insect and human) and (b) nucleotide sequences generated by the model at different stages (number of iterations). The straight lines are linear regression fits from $\ell = 4$ to 108.

**Fig. 7:** DNA walk representations of artificial sequences generated by the stochastic model described in the text. The parameters used in this simulation are: $L = 30000$, $p_i = 0.2$, $L_o = 20$, and 62% purines in the initial sequence. (a) The early stage of "evolution" (after 800 iterations in the model simulation) shows a landscape with 2 to 3 large regions of different bias (the up-hill and down-hill regions). (b) After 1600 iterations, the landscape becomes visually more complex, resembling the actual DNA walk representation for the rat MHC sequence in (c). The values of $p_i$, $L_o$, and $L$ given here are typical of those we used in our simulation.

**Fig. 8:** The DNA walk representation for the rat embryonic skeletal myosin heavy chain gene ($\alpha = 0.63$). (a) The entire sequence. (b) The magnification of the solid box in (a). (c) The magnification of the solid box in (b). The statistical self-similarity of these plots is consistent with the existence of a scale-free or fractal phenomenon which we call a fractal landscape. Note that one must magnify the segment by different factors along the $\ell$ (horizontal) direction and the $y$ (vertical) direction; since $F$ has the same units (dimension) as $y$, these magnification factors $M_\ell$ and $M_y$ (along $\ell$ and $y$ directions respectively) are related to the scaling exponent $\alpha$ by the simple relation $\alpha = \log(M_y)/\log(M_\ell)$ [e.g., from (a) to (b), $\log(M_y)/\log(M_\ell) = \log(2.07)/\log(3.2) \cong 0.63$].