Statistical Mechanics in Biology: How Ubiquitous are Long-Range Correlations?

H. E. Stanley,[a] S. V. Buldyrev,[a] A. L. Goldberger,[b] Z. D. Goldberger,[a]
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 examples of recent progress in applying statistical mechanics 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 discuss the evidence supporting the finding that for sequences containing only *coding* regions, there are no long-range correlations. We also discuss the recent finding that the exponent α characterizing the long-range correlations increases with evolution, and we discuss two related models, the *insertion model* and the *insertion-deletion model*, that may account for the presence of long-range correlations. Finally, we summarize the analysis of long-term data on human heartbeats (up to 10^4 heart beats) that supports the possibility that the successive increments in the cardiac beat-to-beat intervals of healthy subjects display scale-invariant, long-range "anti-correlations" (a tendency to beat faster is balanced by a tendency to beat slower later on). In contrast, for a group of subjects with severe heart disease, long-range correlations vanish. This finding suggests that the classical theory of homeostasis, according to which stable physiological processes seek to maintain "constancy," should be extended to account for this type of dynamical, far from equilibrium, behavior.

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." This observation explains why fractals are of great importance in biology, where surface phenomena are of crucial importance.

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 the arterial system which delivers oxygen and nutrients to all the cells of the body. For this purpose blood vessels must have fractal properties [3]. The diameter distribution of blood vessels ranging from capillaries to arteries follows a power-law distribution which is one of the main characteristics of fractals. Sernetz et al. [3] have studied the branching patterns of arterial kidney vessels. They analyzed the mass-radius relation and found that it can be characterized by fractal geometry, with fractal dimensions between 2.0 and 2.5. Similarly, the branching of trees and other plants, as well as root systems have a fractal nature. One of the most remarkable examples of a fractal object is the surface of a cauliflower, where every little head is an almost exact reduced copy of the whole head formed by intersecting Fibonacci spirals of smaller heads, which in turn consist of spirals of smaller and smaller heads, up to the fifth order of hierarchy. West and Goldberger were first to describe such a "Fibonacci fractal" in the human lung [1]. (For a general review of fractals in physiology and medicine see also Goldberger, Rigney, and West [1].)

Considerable interest in the biological community has also 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 becomes 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). The branching pattern of retinal vessels in a developed human eye is also very similar to DLA [3]. The fractal dimension was estimated to be about 1.7, in good agreement with DLA for the case of two dimensions.

The DLA-type model governing viscous fingering may also serve to resolve the age-old paradox *"Why doesn't the stomach digest itself?"* [6]. Hydrochloric acid (HCl) when released under pressure by the secretory glands 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 one is forced under pressure through it.

Yet another example of DLA-type growth is bacterial colony spread on 2d plates [7]. Vicsek et al. [8] studied bacterial colony growth on a strip geometry which results in a self-affine surface. They calculated the roughness exponent α for this surface and found $\alpha = 0.78 \pm 0.07$.

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" [9–15]. 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.

A system is said to exhibit long-range correlations when some physical properties of the system at different positions (or times) are correlated and the corresponding correlation function decays much slower than exponentially with distance or time. The mechanism for generating long-range correlations is not always obvious. One possibility is that there is an input to the system which itself has long-range correlations. However, in many cases long-range correlations are spontaneously generated, even though all the physical interactions are themselves short-range. A well-studied example is a system at or near its critical point, for which the interactions among the molecules are quite short range, being essentially nearest-neighbor only. A remarkable manifestation of the fact that the correlations extend in range to thousands of atomic distances is the phenomenon of critical opalescence, in which visible light (of wavelength *thousands of* Å) is scattered (Fig. 1). It is difficult, *a prioiri*, to imagine that a biological system is situated at its critical point, as that would require some control parameter to be tuned to just the proper value. However recently it has come to be appreciated that many systems essentially "tune themselves" spontaneously such that they become closer and closer to a critical point; such self-organized critical phenomena may occur in biological systems [16].

Systems with Characteristic Scales

In order to understand the concept of "scale-free," we start with a discussion of systems with characteristic scale, the correlation length.

Consider a physical quantity (measurement) which describes some properties of a system. This quantity may vary from one location to another location (or may fluctuate in time). If the system cannot be divided into many isolated sub-parts, the fluctuations of the physical quantity are not completely uncorrelated even though they may be driven by some stochastic process. However, if the stochastic influence is strong, the physical quantity can be viewed as an independent variable. Therefore, there is a well-defined scale for correlation length related to the distance above which the physical properties are uncorrelated (or the smallest scale of the sub-systems that can be treated as independent).

This characteristic scale has its practical role in simplifying the problem of describing the system. Several first order models for studying physical systems take advantage of the existence of this scale; they usually assume that the complete system can be divided into many sub-systems that do not communicate with each other. For example, classical gases can be viewed as a system of hard sphere particles with the radii chosen to be the characteristic length of the actual molecular interaction distance. The fact that this type of model can describe some general properties of a real system is an important justification of this characteristic scale.

For many physical systems, a similar scale in time also exists. Furthermore, we note that the characteristic scales are frequently related to the exponential decay of a specific physical quantity. For example, a half life time is a characteristic time scale for radioactive particles. The physical mechanism of exponential decay is easy to understand: Consider a function $I(t)$ that measures the amount of information (or other physical property) at time (or position) t. In many cases, it is reasonable to assume that the amount of information lost will be proportional to the amount of information that exists at that moment, i.e.,

$$\frac{dI(t)}{dt} \propto -I(t). \tag{1a}$$

The solution for this type of equation is

$$I(t) \propto \exp(-t/\tau), \tag{1b}$$

where τ is the inverse of the proportionality constant in Eq. (1). It also plays the role of characteristic scale.

In general, when the correlation of a physical quantity decays fast enough (not necessarily exponential), then the notation of a characteristic scale is meaningful and useful.

Scale-Free Systems

Not every system contains a well-defined length (or time) scale to which the system can be simplified accordingly. In the last few decades, significant attention has been directed at understanding systems that do not have any characteristic scale. A classical example is a ferromagnetic Ising spin system at the critical point. We will use this example to illustrate some important aspects of a scale-free system.

Consider a three-dimensional lattice with a spin on each lattice site. The spin can only take two directions (values), denoted by 1 and -1, and only interacts with its nearest neighbors through ferromagnetic interaction (lower energy when two spins align with each other). The system is then put in contact with a heat reservoir which provides the thermal energy for the fluctuations of the spins. If the temperature for the heat reservoir is high, then the "randomness" of the thermal noise dominates. Therefore, each spin is completely free to choose its own direction with little correlation to its neighbors. If we gradually lower the temperature, we will notice that the coupling energy becomes relatively stronger and spins are more correlated (with a short correlation length). A physically meaningful way to define the correlation length for the system is first to connect all nearest neighbor spins that are aligned in the same direction as a "cluster" and relate the correlation length to the average cluster size. Spins in different clusters know nothing about each other and, therefore, the system can be divided into independent sub-systems (clusters). For high temperature, the distribution of cluster size decays fast (it is extremely rare to find large clusters), thus a meaningful correlation scale exists. As we lower the temperature, the size distribution varies (more chance to find large clusters) and the correlation length increases. However, there is no dramatic change in the form of the distribution function. If we continue to lower the temperature, suddenly, at the critical temperature, the distribution function becomes a power law and the correlation length diverges, i.e., the size of the largest cluster one can find is comparable to the system size. At this point, the system can no longer be viewed as many independent sub-systems and any characterization of the system must involve the system as a whole: long-range order appears. In this case, the origin for generating long-range correlations is the balance between two competitions—"order" (spin coupling) and "disorder" (thermal fluctuations)—over all scales.

Systems that have long-range correlations usually have certain physical properties described by homogeneous functions. A homogeneous function f is defined as

$$f(\lambda r) \sim g(\lambda) f(r). \tag{1c}$$

The physical meaning for the homogeneous function is that the value of the function at a new scale is simply related to the value of the original scale by some constant factor. The solution to Eq. (1c) is power-law function, i.e.,

$$f(r) \sim r^{-H}. \tag{1d}$$

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 descriptions must be incomplete, since only some appropriately-chosen averages 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], selected literary compositions [22], and stock market fluctuations [23]—there has been no systematic study of a *biological* system that displays power-law correlations.

First 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, 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).

There are actually many possible rules of mapping of DNA sequence onto 1-dimensional random walk:

Correlations of Pairs of Nucleotides:

(i) $u(i) = +1$ for A or G and $u(i) = -1$ otherwise (purine-pyrimidine rule);

(ii) $u(i) = +1$ for C or G and $u(i) = -1$ otherwise; (hydrogen bond rule)

(iii) $u(i) = +1$ for A or C and $u(i) = -1$ otherwise;

Correlations of One Nucleotide With Itself:

(iv) One can assign $u(i) = +1$ if nucleotide A occurs on the $i^{\rm th}$ place and $u(i) = -1$ otherwise (in case of C,G, or T)

(v) $u(i) = +1$ for C and $u(i) = -1$ otherwise;

(vi) $u(i) = +1$ for G and $u(i) = -1$ otherwise;

(vii) $u(i) = +1$ for T and $u(i) = -1$ otherwise;

Correlations of some physical quantity (e.g., molecular mass rule):

(viii) $u(i) = 134$ for A, $u(i) = 110$ for C, $u(i) = 150$ for G, and $u(i) = 125$ for T.

This list of rules may be extended. Moreover, there have also been attempts to map the DNA sequence onto a multi-dimensional space [25,32]. Generally we find that the original purine-pyrimidine rule provides the most robust results, probably due to the purine-pyrimidine chemical complementarity.

The DNA walk allows one to visualize directly the fluctuations of the purine-pyrimidine content in DNA sequences: Positive slopes on the Fig. 2 correspond to high concentration of pyrimidines, while negative slopes correspond to high concentration of purines. Visual observation of DNA walks suggests that the coding sequences and intron-containing non-coding sequences have quite different landscapes. Figure 2a shows a typical example of a gene that contains a significant fraction of base pairs that do *not* code for amino acids. Figure 2b shows the DNA walk for a sequence formed by splicing together the coding regions of the DNA sequence of this same gene (i.e., the cDNA). Figure 2c displays the DNA walk for a typical sequence with only coding regions. Landscapes for intron-containing sequences show very jagged contours which consist of patches of all length scales, reminiscent of the disordered state of matter near critical point. On the other hand, coding sequences typically consist of a few lengthy regions of different strand bias, resembling domains in the system in the ferromagnet state. These observations can be tested by rigorous statistical analysis.

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. 2. Figure 2 naturally motivates a quantification of this fluctuation 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 $y(\ell) \equiv \sum_{i=1}^{\ell} u(i)$.

An important statistical quantity characterizing any walk 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,

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

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 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 $\ell_o = 1$ to $\ell_o = 2, \cdots$ and (c) calculating the quantity $\Delta y(\ell)$ (and its square) for each value of ℓ_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 $F(\ell) \sim \ell^{1/2}$

(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 2a shows a typical example of a gene that contains a significant fraction of base pairs that do *not* code for amino acids. It is immediately apparent that the DNA walk has an extremely jagged contour which we shall see corresponds to long-range correlations. Double logarithmic plots of the mean square fluctuation function $F(\ell)$ as a function of the linear distance ℓ along the DNA chain for intron-containing and intergenic (i.e., non-coding) sequences are linear, so $F(\ell) \sim \ell^{\alpha}$. 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.

It is known that functional proteins usually form a single compact three-dimensional conformation that corresponds to the global energetical minimum in the conformational space. Recently, Shakhnovich and Gutin [44] found that in order to have such a minimum it is sufficient that an amino acid sequence forms an uncorrelated random sequence. The finding of Peng et al. [24] of the lack of long range correlations in the coding nucleotide sequences provides more evidence for this hypothesis, since there exist almost one-to-one correspondence between amino acid sequences and their nucleotide codes. Furthermore, this finding may also indicate that the *lack* of long range correlations in the amino acid sequences is, in fact, a *necessary* condition for a functional biologically active protein.

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 α in our DNA walk method with those found in other methods, we have 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 α, β and γ are not independent, since

$$\alpha = \frac{1+\beta}{2} = \frac{2-\gamma}{2}. \tag{4d}$$

The first equality is derived in Eq. (A.13) below, while the second is derived in Eq. (11). 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. The reason for the smaller fluctuations of α 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, $S(\infty)$ [47]. 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 non-coding DNA sequences has generated contradicting responses. Some [25,27,36] support our initial finding, while some [26,31,35,37] disagree. For example, 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 [27].

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 conventional scaling analyses cannot be applied reliably to the entire sequence but only to sub-sequences.

Peng et al. [42] have recently applied the "bridge method" to DNA, and have also developed several similar methods specifically adapted to handle problems associated with non-stationary sequences which they term *detrended fluctuation analysis* (DFA).

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

The DFA method clearly confirms the difference between coding and noncoding 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 (Fig. 3).

To provide an "unbiased" test of the thesis that non-coding regions possess but coding regions lack long-range correlations, we 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.

Using the DFA method, we can measure local value of correlation exponent α along the sequence (see Fig. 4) and find that the local minima of α as a function of a nucleotide position which usually corresponds to non-coding regions, while the local maxima corresponds to non-coding regions. The statistical analysis of the nucleotide sequence data for yeast chromosome III (315,338 nucleotides) shows that the probability that the observed correspondence between the positions of minima and coding regions is due to random coincidence is less than 0.0014. Thus, this method—which we called the "beachcomber" algorithm—can be used for finding coding regions in the newly sequenced DNA, a potentially important application of DNA walk analysis.

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 [48]. 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 focused 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 the DNA walk 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 DNA landscapes show that the coding sequences of myosins remain practically unchanged with evolution, while the entire gene sequences become more heterogeneous and complex. The quantitative measurements of the exponent α by DFA method confirm this visual observation showing that for all coding sequences of MHC family $\alpha \approx 0.5$. In contrast, for entire genes of MHC family, the value of α monotonically increases from lower eukaryotes to invertebrates and from invertebrates to vertebrates.

Of note, the value of α 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 α after the exons are removed is 0.593 versus 0.586 for the complete sequence, further supporting the view that the principal source of long-range correlations in genomic sequences is the composition of non-coding elements themselves and is not just the intron versus exon alternation as was suggested by Nee [26].

The results of analysis of MHC family were confirmed by the studies of other gene families [39].

6. Insertion Model

The finding described in the previous section suggests that the source of long-range correlations could be an evolutionary process specific to non-coding sequences. In contrast, the coding sequences should preserve their uncorrelated structure in order to maintain the functional properties of the encoded proteins.

Li [49] was first to suggest a simple model of a biologically plausible process of duplication and mutation that can produce a sequence with any given value of α. His model can be used to explain certain features of highly repetitive DNA, but does not take into account many other important processes of DNA evolution like retroviral insertions and deletions, which are probably the main source of rapid evolution of DNA sequences.

An example of such a retroposone is the LINE-1 sequence which consists of 6,139 base pairs and is believed to contain a code for a functional protein [50]. In agreement with this we find that the LINE-1 sequence has value of α close to 0.5, indicating the lack long-range correlations [39]. Moreover, the LINE-1 sequence has a strong strand bias of about 59% of purines, which is also very typical for coding sequences. The total number of LINE-1 sequences and fragments in the human genome is estimated to be 107,000, while in the genome of the chimpanzee there are only 51,000 copies of the LINE-1 sequence [51]. This dramatic difference indicates that thousands of insertions or deletions of LINE-1 sequences took place over a relatively short evolutionary time scale. LINE-

1 sequences are found on both strands of DNA and therefore produce large local fluctuations of nucleotide content. Another very frequent repetitive element is the ALU sequence [52], which is also statistically similar to protein coding DNA, but, in contrast with the Line-1 sequence, is only 300 base pairs long.

The central idea of the insertion model [39] is based on the assumption that the insertion of retroelements, formed by the inverse-transcribed RNA, plays a major role in DNA evolution. The statistical properties of retroelements are similar to those of protein coding sequences. In order to be inserted into DNA, a retroelement must form a loop. The probability to find a loop of certain size l in a long polymer chain in a solvent is given [53] by the formula

$$P(l) \propto (1/l)^{\mu}, \tag{5a}$$

where μ is a critical exponent with a value close to 2.2. Thus we assume

(i) that DNA sequences are comprised of subsequences distributed according to Eq. (3.6), and

(ii) that these subsequences are statistically similar to protein coding sequences which (a) usually have a significant excess of purines over pyrimidines (or vice versa because of DNA two-strand complementarity) and (b) can be modeled by a Markovian process with short range correlations [3.55]

This biological evolution model which we developed is mathematically equivalent to the generalized Lévy walk which gives rise to a landscape with a well defined power-law long-range correlation exponent α that depends upon the Lévy walk parameter μ [39]

$$\alpha = \begin{cases} 1 & \mu \leq 2 \\ 2 - \mu/2 & 2 < \mu < 3 \\ 1/2 & \mu \geq 3, \end{cases} \tag{5b}$$

i.e., non-trivial behavior of α corresponds to the case $2 < \mu < 3$ where the first moment of $P(l)$ converges while the second moment diverges. The long-range correlation property for the Lévy walk, in this case, is a consequence of the broad distribution of Eq. (5a) that lacks a characteristic length scale. Equation (5b) is valid only asymptotically for large values of ℓ. For small ℓ the slope of the log-log plot of the function $F(\ell)$ for the generalized Lévy walk model increases monotonically from a value defined by short range Markovian correlations of the inserted subsequences to a value $\alpha = 0.9$ predicted by Eq. (5b). However, this limiting value can be achieved only for very long sequences of about 10^6 base pairs, and has a large standard error for finite sequences [38].

To test the insertion model [39], we have adjusted its parameters, to best approximate features of actual DNA sequences and found a very good agreement for the behavior of successive slopes of the $F(\ell)$ function for all sequences, that contain a substantial percentage of non-coding material.

7. Insertion-Deletion Model

In order (i) to to gain some insight into possible evolutionary mechanisms that could increase the complexity of the DNA landscapes and generate long-range correlations of DNA sequences, and (ii) to create a more realistic model of DNA evolution which includes also deletion of certain DNA subsequences, intron insertion and the exchange of genomic material between DNA strands and chromosomes, Buldyrev et al. [41] modified the generalized Lévy walk model by allowing random deletion and reinsertions of subsequences with length distribution defined by Eq. (5a). Starting with uncorrelated sequence, statistically similar to to mRNA of MHC, the model generates with each new iteration more and more heterogeneous sequences, and reproduces the monotonic increase of α with evolution observed for MHC gene family.

(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.

Of note, if the model is iterated without the insertion of random biased sequences as assumed in rule (iic), the value of α 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 [51].

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 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 α 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 α. 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 α 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 β, which is uniquely related to α) 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. 2b), one can obtain a spuriously large value of α.

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 α. 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.

8. Long-Range Correlations and DNA Spatial Structure

¿From the point of view of statistical mechanics long-range correlations cannot exist in a one-dimensional system at equilibrium. One possible explanation of their existence is the type of non-equilibrium evolutionary process described above. However, another source of these correlations could be the actual interactions between parts of DNA molecule in three dimensional space. It is known that the DNA polymer chain bound by histones folds itself into the chromosome according to a hierarchical structure of loops of many length scales. Moreover, it has an ability to unwind without forming knots. This complex behavior suggests that the sequence of DNA base pairs may contain a "code" that defines DNA packaging and unwinding. At the very least, the local nucleotide composition may be affected by the way DNA is packaged. Recently, Grosberg et al. [36] related long-range correlations in the nucleotide sequence and the spatial structure of the chromosome. Using the idea of a "crumpled globule" (a collapsed conformation of a polymer chain without knots), they predicted the universal value of $\alpha = 2/3$ for all eukaryotic DNA sequences which form chromosomes. In fact, DNA spatial structure, as well as evolutionary processes in DNA may both play a certain role in forming of long-range correlations in the nucleotide sequences. Further studies are needed to resolve this question.

9. Long-Range Correlations in Time Series

The list of systems in which power law correlations appear has grown rapidly in recent years [16,22,23]. What do we anticipate for general biological systems? As noted above, when "entropy wins over energy"—i.e., randomness dominates the behavior—we find power laws and scale invariance. The absence of characteristic length (or time) scales may confer important biological advantages, related to adaptability of response [1].

Often the data of biological experiments are presented in the form of a time series. A time series is a record of a fluctuating quantity in time, denoted as $v(t)$, with the variable t represents the index of time. We are interested in time series that is stochastic in nature, therefore, dominated by the random mechanisms that generate the time series. However, even though the time series is random, it exhibits certain predictable statistical patterns. Next, we will discuss some statistical properties of a time series and several mathematical tools to study these types of patterns.

Stationary Time Series

A time series is called "stationary" if its statistical properties do not vary with time. In more precise mathematical terms, $v(t)$ is called *completely stationary* if the joint probability distribution of $\{v(t_1), v(t_2), \ldots, v(t_n)\}$ is the same as the joint probability distribution of $\{v(t_1+m), v(t_2+m), \ldots, v(t_n+m)\}$, for any set of times $t_1, t_2, \ldots, t_n$, and any number m. Less strictly, we call a time series *weakly stationary*, if only the joint moments up to order 2 of the above joint probability distributions are identical. For practical purposes, it is impossible to obtain moments of all orders for a real data set. Therefore, usually a series is called stationary when it is in fact weakly stationary. In other words, a (weakly) stationary series indicates that its mean and variance is independent of time, and the covariance between two values measured at different times depends only on the time lag between them and not on their specific locations in time.

Self-Similar Stochastic Process

A random time series can be viewed as being generated from a stochastic process. A stochastic process $v(t)$ is self-similar with parameter H if $v(at)$ and $a^H v(t)$ have the same probability distribution for all $a > 0$.

For example, the position of a Brownian motion particle is a self-similar stochastic variable with parameter $H = 1/2$. To illustrate the concept of self-similarity defined above, consider a simplified computational version of Brownian motion: A D-dimensional random walk. The stochastic variable, $v(t)$, in this case is a D-dimensional vector denoting the position of the random walker at time t. Let us consider two simulations: In one simulation the random walker takes a step of random direction with unit length after every unit time. While in the other simulation, the random walker takes a step of two unit lengths after every four unit time. Obviously, if we record the trace of the two random walkers for every time step and play the movie of the two simulations side by side, we will see the difference between them. Now let us take one picture after every 4 unit time steps. Therefore, we have no information about the position of the random walker for the time in between. Can we still see the difference? What if our resolution (in time) is poor so we can only record the position for every 1000 time steps. Can we distinguish between them now? The fact is, we cannot distinguish between them by any statistical studies (up to a certain accuracy). In other words, there is no way to know that whether we generate the trajectories of the random walker, $v(t)$, by 1 unit step of each unit time or by two unit step of each four unit time. Mathematically speaking, the stochastic process, $v(4t)$, have the same statistical properties as $4^{1/2}v(t)$.

This example shows the easiest way to obtain a self-similar time series (by simple summation). Most self-similar time series are not so trivial, the exponent H is, nonetheless, a useful parameter for describing the self-similarity. Another interesting property demonstrated in the above example is that, in a practical sense, self-similarity is only true for a finite range of scales. A lower cut-off scale (e.g., the mean free path in Brownian motion) and an upper cut-off scale always exists such that self-similarity will break down outside this range.

Usually, self-similar time series are non-stationary. The correlation function of a stationary time series normally decays much faster than that of a self-similar time series. We will discuss this point later.

Auto-Correlation Function

The most straightforward method for analyzing a time series is to study the auto-correlation function,

$$C(\tau) \equiv \langle (v(t_0) - \overline{v})(v(t_0+\tau) - \overline{v}) \rangle_{\rm AV}, \tag{6a}$$

where $\langle \cdots \rangle_{\rm AV}$ denotes the long-time average and $\overline{v}$ is the mean for the time series.

Note that the auto-correlation function is meaningful only when the time series is stationary, i.e., $\overline{v}$ exists and $C(\tau)$ is independent of t_0.

Power Spectrum

It is useful to look at a time series in Fourier space. A Fourier transform is a representation that each basis function is localized in frequency but not in time (or space). Due to this property of the Fourier transform, it is intrinsically difficult to use Fourier analysis for non-stationary time series. However, the Fourier transform is still a useful method in most other cases.

In principle, we can transform the time series $v(t)$ to Fourier space and preserve the same amount of information. Therefore, one can characterize the Fourier spectrum according to different classes of the random process without losing any generality.

Consider a time series $v(t)$ for a very long time T. The Fourier transform of $v(t)$ is given by

$$\tilde{v}(\omega) \equiv \int_{-\infty}^{+\infty} v(t)\, e^{-i\omega t}\, dt, \tag{6b}$$

where we have assumed $v(t) = 0$ outside the interval T. Note that $\tilde{v}(\omega) = \tilde{v}^*(-\omega)$, because $v(t)$ is real. It is also useful to define the power spectrum

$$S(\omega) \equiv |\tilde{v}(\omega)|^2, \tag{7}$$

as the square of the amplitude for each frequency component.

For a pure random time series (without any correlation), the power spectrum $S(\omega)$ fluctuates randomly, usually called "white spectrum." But pure random processes hardly exist in nature. In most cases the power spectrum will not be flat but will have a well-defined underlined shape, due to the correlation in time series. The underlined form of the power spectrum and the fluctuations about this envelope will determine all properties of the stochastic process.

We call a time series long-range correlated when the power spectrum has a power law form ($1/f$ type noise):

$$\langle \tilde{v}(\omega)\tilde{v}^*(\omega')\rangle = \omega^{-\beta}\delta(\omega-\omega'), \tag{8}$$

i.e.,

$$S(\omega) \sim 1/\omega^{\beta}. \qquad \text{with} \quad -1<\beta<1, \tag{9}$$

where $\langle\rangle$ denotes ensemble average, i.e., average over different time series. Notice that for $\beta>0$ the two consecutive variables tend to be of the same sign (assuming that the average of $v(t)$ is zero), since the amplitude of oscillation for large-times (small ω) is more prominent—which we call the *ferro* case. While for $\beta<0$, they tend to be of opposite sign, because in this case the fast oscillation (large ω) dominates—which we call the *anti-ferro* case. For $\beta=0$, $v(t)$ is completely uncorrelated.

For $\beta>1$, the series is non-stationary since the integral of power spectrum from a finite frequency to zero diverges, indicating there is no well defined average for the time series. In other words, the average value of the time series increases with the length of the time series. We call $\beta=1$ ($1/f$ noise) marginally stationary since the increase is logarithmic.

A useful relation between power spectrum and the auto-correlation function $C(\tau)$ is the Weiner–Khintchine theorem: If the time series is stationary, then

$$\begin{aligned}
C(\tau) &\equiv \langle v(t)v(t+\tau)\rangle_{\mathrm{AV}} \\
&= \int\!\!\int \langle \tilde{v}(\omega)e^{i\omega t}\tilde{v}(\omega')e^{i\omega'(t+\tau)}\rangle d\omega d\omega' \\
&= \int\!\!\int \langle \tilde{v}(\omega)\tilde{v}(\omega')\rangle e^{i\omega'\tau} d\omega d\omega' \\
&\propto \int_{-\infty}^{+\infty} |F(\omega)|^2 e^{-i\omega\tau} d\omega \\
&\propto \int_{-\infty}^{+\infty} S(\omega)e^{-i\omega\tau} d\omega,
\end{aligned} \tag{10}$$

where $\langle\rangle_{\mathrm{AV}}$ denotes the long-time average and was replaced by the ensemble average in the calculation. Eq. (10) states that the auto-correlation function is related to the Fourier transform of the power spectrum $S(\omega)$.

For the case of long-range correlations [Eq. (9)], the correlation function decays as a power law in time, with an exponent that is related to β:

$$C(\tau) \sim \int_{-\infty}^{+\infty} |\omega|^{-\beta}\, e^{-i\omega\tau}\, d\omega$$

$$= 2 \int_0^{+\infty} \omega^{-\beta} \cos(\omega\tau)\, d\omega$$

$$= 2\, \tau^{-(1-\beta)} \int_0^{+\infty} u^{-\beta} \cos(u)\, du \sim \tau^{-(1-\beta)}. \qquad (11)$$

Thus $\gamma = 1 - \beta$, as stated in Eq. (4d). This result is rigorous for positive β. For $\beta < 0$, to avoid the problem of divergence, we must assume that Eq. (9) holds for small ω only, i.e., there is a cut-off frequency which corresponds to the inverse of the smallest time interval (which may be associated with the limitation of a physical measurement).

An Example

There are a wide variety of stochastic processes that have the following form of Fourier coefficients:

$$\tilde{v}(\omega) = \mathcal{D}(\omega)\xi(\omega), \qquad (12)$$

where $\xi(\omega)$ is pure random noise representing the stochastic property of the system. And the power spectrum is

$$S(\omega) = |\mathcal{D}(\omega)|^2, \qquad (13)$$

which quantifies the correlation property of the stochastic process. The form of $\mathcal{D}(\omega)$, the frequency profile, can be treated as a modification of the white spectrum, depends on the mechanism of the stochastic process.

For simplicity, consider that $\xi(\omega)$ is a Fourier transform of Gaussian white noise. Thus $\xi(\omega)$ is also Gaussian and has the following properties:

$$\langle \xi(\omega) \rangle = 0, \qquad (14)$$

$$\langle \xi(\omega)\xi^*(\omega') \rangle \sim \delta(\omega - \omega'), \qquad (15)$$

and the probability of finding the real (or imaginary) part of $\xi(\omega)$ in the interval ξ to $\xi + d\xi$, $P(\xi)\, d\xi$, is Gaussian, i.e.,

$$P(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right), \qquad (16)$$

where σ^2 is the variance of the distribution. Since $\xi(\omega)$ is a Gaussian variable, all higher order correlations can be obtained from pair correlations, i.e.,

$$\langle \tilde{v}(\omega_1)\tilde{v}(\omega_2) \cdots \tilde{v}(\omega_{2n+1}) \rangle = 0, \qquad (16)$$

and

$$\langle \tilde{v}(\omega_1)\tilde{v}(\omega_2) \cdots \tilde{v}(\omega_{2n}) \rangle = \sum\nolimits^* \langle \tilde{v}(\omega_p)\tilde{v}(\omega_q) \rangle \cdots \langle \tilde{v}(\omega_r)\tilde{v}(\omega_s) \rangle. \qquad (17)$$

The summation extends over all different ways that $1, 2, \ldots, 2n$ can be subdivided in pairs.

A well known example is the Langevin equation for Brownian motion,

$$M\frac{dv}{dt} + \gamma\, v = \xi(t), \tag{18}$$

where M is the mass of the particle, γ is the friction constant, and $\xi(t)$ is the random driven force. By Fourier transform one obtains

$$\tilde{v}(\omega) = \frac{1}{\gamma - i\omega M}\,\xi(\omega), \tag{19}$$

which is the case that

$$\mathcal{D}(\omega) = \frac{1}{\gamma - i\omega M}, \tag{20}$$

or

$$S(\omega) = \frac{1}{\gamma^2 + \omega^2 M^2}. \tag{21}$$

By substituting Eq. (21) into (10), we find that $C(\tau) \sim \exp(-\tau/\tau_0)$, with $\tau_0 = M/\gamma$. Since the correlation is short range (exponential decay), asymptotically, it leads to normal diffusion as if there were no correlation.

10. Fractal Analysis of Interbeat Intervals

Very recently, the idea of long-range correlations has baeen extended to the analysis of the beat-to-beat intervals in the normal and diseased heart [21,59]. The healthy heartbeat is generally thought to be regulated according to the classical principle of homeostasis whereby physiologic systems operate to reduce variability and achieve an equilibrium-like state [60]. We find, however, that under normal conditions, beat-to-beat fluctuations in heart rate display the kind of long-range correlations typically exhibited by physical dynamical systems far from equilibrium, such as those near a critical point. We review recently reported evidence for such power-law correlations that extend over thousands of heartbeats in healthy subjects. In contrast, heart rate time series from patients with severe congestive heart failure show a breakdown of this long-range correlation behavior, with the emergence of a characteristic short-range time scale. Similar alterations in correlation behavior may be important in modeling the transition from health to disease in a wide variety of pathologic conditions.

Clinicians describe the normal activity of the heart as "regular sinus rhythm." But in fact cardiac interbeat intervals normally fluctuate in a complex, apparently erratic manner. Much of the analysis of heart rate variability has focused on short term oscillations associated with breathing (0.15 – 0.40 Hz) and blood pressure control (0.01 – 0.15 Hz) [61–63].

To study these dynamics over large time scales, we pass the time series through a digital filter that removes fluctuations of frequencies > 0.005 beat^{-1}, and plot the result, denoted by $B_L(n)$, in Fig. 5. We observe a more complex pattern of fluctuations for a representative healthy adult (Fig. 5a) compared to the "smoother" pattern of interbeat intervals for a subject with severe heart disease (Fig. 5b). These heartbeat time series produce a contour reminiscent of the irregular landscapes that have been widely studied in physical systems.

To quantitatively characterize such a "landscape", we introduce a mean fluctuation function $F(n)$, defined as

$$F(n) \equiv \overline{|B_L(n'+n) - B_L(n')|}, \tag{22}$$

where the bar denotes an average over all values of n'. Since $F(n)$ measures the average difference between two interbeat intervals separated by a time lag n, $F(n)$ quantifies the magnitude of the fluctuation over different time scales n.

Figure 6 is a log-log plot of $F(n)$ vs n for the data in Figs. 5a and 5b. This plot is approximately linear over a broad physiologically-relevant time scale $(200 - 4000$ beats) implying that

$$F(n) \sim n^{\alpha}. \tag{23}$$

We find that the scaling exponent α is markedly different for the healthy and diseased states: for the healthy heartbeat data, α is close to 0, while α is close to 0.5 for the diseased case. It is interesting to note that $\alpha = 0.5$ corresponds to the well-studied *random walk* (Brownian motion), so the low-frequency heartbeat fluctuations for the diseased state can be interpreted as a stochastic process, in which case the interbeat increments $I(n) \equiv B(n+1) - B(n)$ are uncorrelated for $n > 200$.

To investigate these dynamical differences, it is helpful to study further the correlation properties of the time series. To this end, we choose to study $I(n)$ because it is the appropriate variable for the aforementioned reason. Since $I(n)$ is stationary, we can apply standard spectral analysis techniques [10]. Figures 7a and 7b show the power spectra $S_I(f)$, the square of the Fourier transform amplitudes for $I(n)$, derived from the same data sets (without filtering) used in Fig. 5. The fact that the log-log plot of $S_I(f)$ vs f is linear implies

$$S_I(f) \sim \frac{1}{f^{\beta}}. \tag{24}$$

The exponent β is related to α by $\beta = 2\alpha - 1$ [18]. Furthermore, β can serve as an indicator of the presence and type of correlations:

(i) If $\beta = 0$, there is no correlation in the time series $I(n)$ ("white noise").

(ii) If $0 < \beta < 1$, then $I(n)$ is correlated such that positive values of I are likely to be close (in time) to each other, and the same is true for negative I values.

(iii) If $-1 < \beta < 0$, then $I(n)$ is also correlated; however, the values of I are organized such that positive and negative values are more likely to alternate in time ("anti-correlation") [18].

For the diseased data set, we observe a flat spectrum ($\beta \approx 0$) in the low frequency region (Fig. 7b) confirming that $I(n)$ are not correlated over long time scales (low frequencies). Therefore,

$I(n)$, the first derivative of $B(n)$, can be interpreted as being analogous to the *velocity* of a random walker, which is uncorrelated on long time scales, while $B(n)$—corresponding to the *position* of the random walker—are correlated. However, this correlation is of a trivial nature since it is simply due to the summation of uncorrelated random variables.

In contrast, for the data set from the healthy subject (Fig. 7a), we obtain $\beta \approx -1$, indicating *non-trivial* long-range correlations in $B(n)$—these correlations are not the consequence of summation over random variables or artifacts of non-stationarity. Furthermore, the "anti-correlation" properties of $I(n)$ indicated by the negative β value are consistent with a nonlinear feedback system that "kicks" the heart rate away from extremes. This tendency, however, does not only operate on a beat-to-beat basis (local effect) but on a wide range of time scales. To our knowledge, this is the first explicit description of long-range anticorrelations in a fundamental biological variable, namely the interbeat interval increments.

11. Physiological Implications

Our finding of non-trivial long-range correlations in healthy heart rate dynamics is consistent with the observation of long-range correlations in other biological systems that do not have a characteristic scale of time or length. Such behavior may be adaptive for at least two reasons. (i) The long-range correlations serve as an organizing principle for highly complex, non-linear processes that generate fluctuations on a wide range of time scales. (ii) The lack of a characteristic scale helps prevent excessive *mode-locking* that would restrict the functional responsiveness of the organism. Support for these related conjectures is provided by observations from severe diseased states such as heart failure where the breakdown of long-range correlations is often accompanied by the emergence of a dominant frequency mode (e.g., the Cheyne-Stokes frequency). Analogous transitions to highly periodic regimes have been observed in a wide range of other disease states including certain malignancies, sudden cardiac death, epilepsy and fetal distress syndromes.

The complete breakdown of normal long-range correlations in any physiological system could theoretically lead to three possible diseased states: (i) a random walk (brown noise), (ii) highly periodic behavior, or (iii) completely uncorrelated behavior (white noise). Cases (i) and (ii) both indicate only "trivial" long-range correlations of the types observed in severe heart failure. Case (iii) may correspond to certain cardiac arrhythmias such as fibrillation. More subtle or intermittent degradation of long-range correlation properties may provide an early warning of incipient pathology. Finally, we note that the long-range correlations present in the healthy heartbeat indicate that the neuroautonomic control mechanism actually drives the system away from a single steady state. Therefore, the classical theory of homeostasis, according to which stable physiological processes seek to maintain "constancy" [60], should be extended to account for this dynamical, far from equilibrium, behavior.

12. Fractal Music and the Heartbeat

Fourier analysis of instantaneous fluctuations in amplitude as well as inter-note intervals for certain classical music pieces (e.g., Bach's First Brandenburg Concerto) reveals a $1/f$ distribution over a broad frequency range [64–66]. Voss and Clarke [65] used an algorithm for generating $1/f$-noise to "compose" music.

Based on the observation of Peng et al. [21] of different scaling patterns for healthy and pathologic heartbeat time series, we postulated that (i) actual biological rhythms such as the heartbeat might serve as a more natural template for musical compositions than artificially-generated noises, and (ii) audibly appreciable differences between the note series of healthy and diseased hearts could potentially serve as the basis for a clinically useful diagnostic test.

Accordingly, we devised a computer program to map heart rate fluctuations onto intervals of the diatonic musical scale. As anticipated, the *normal* ($1/f$-like) heartbeat obtained from the low pass filtered time series reported in Ref. 21 generated a more variable (complex) type of music than that generated by the *abnormal* times series (Fig. 8). Musical compositions based on these times series (composed by ZDG) were played at this conference and are available on cassette by request along with the "scores" in Fig. 8.*

The "musicality" of these transcriptions is intriguing and supports speculations about the brain's possible role as a translator/manipulator of biological $1/f$-like noise into aesthetically pleasing art works. Current investigations are aimed at extending these preliminary observations by (i) comparing the "musicality" of note sequences generated by natural (biological) vs. artificial (computer simulated) correlated and uncorrelated noises, and (ii) using heartbeat time series as the template for simultaneously generating fluctuations in musical rhythm and intensity, not only pitch.

Appendix A: Correlated Random Walks and their Description

Another useful technique to study a temporal (or spatial) series by mapping to a random walk problem. To be precise, we map the time series $v(t)$ to the velocity of a random walker. Although in practical applications of the random walk model we actually deal with discrete time, but for now we assume the time to be a continuous variable. Let $x(t)$ be the position of the particle at time t, then

$$x(t) = \int_0^t v(t')\,dt', \tag{A.1}$$

where we have set $x(0) = 0$ for simplicity.

Mean-Square Displacement

* Contact Zachary D. Goldberger (e-mail: ary@astro.bih.harvard.edu) or C. K. Peng (e-mail: peng@buphyk.bu.edu). There is a nominal charge for copying and mailing.

The mean-square displacement $\langle x^2(t)\rangle$ can be written

$$\langle x^2(t)\rangle = \int_0^t \int_0^t \langle v(t')v(t'')\rangle_C \, dt' \, dt'', \tag{A.2}$$

where $\langle\cdots\rangle_C$ denotes an average over all configurations. When the time series is long-range correlated, i.e., the power spectrum of $v(t)$ is described by (9), one can obtain

$$\langle x^2(t)\rangle \sim t^{1+\beta}. \tag{A.3}$$

Equation (A.3) tells us that the Fourier spectrum given by Eq. (9) leads to anomalous diffusion characterized by the fractal dimension $d_w = 2/(1+\beta)$, where d_w is defined by $\langle x^2(t)\rangle \sim t^{2/d_w}$. The advantage of studying mean-square displacement is that the noise is significantly reduced by the double integral of Eq. (A.2).

Probability Distribution of Displacement

One way to understand the stochastic process is to study the probability distribution, $P(x,t)$, which tells us the probability of finding the random walker with displacement x after time t. In principle, the form of $P(x,t)$ contains all information about the time series itself and, therefore, is not always obtainable. Here we will give an example of long-range correlated time series such that the form of $P(x,t)$ can be easily derived.

If the long-range correlated time series has a Fourier transform that $\mathcal{D}(\omega) \sim \omega^{-\beta/2}$ and the noise $\xi(\omega)$ is Gaussian, then the generating function of $P(x,t)$, defined as

$$\left\langle e^{ix(t)\theta}\right\rangle \equiv \int_{-\infty}^{+\infty} P(x,t)e^{ix\theta}\, dx, \tag{A.4}$$

can be calculated directly. Since $x(t)$ is linear in the Gaussian variables in which we average,

$$\left\langle e^{ix(t)\theta}\right\rangle = \exp\left(- \left\langle x^2(t)\right\rangle \theta^2\right). \tag{A.5}$$

By applying Eq. (A.3), we obtain

$$\left\langle e^{ix(t)\theta}\right\rangle \sim \exp(-c\theta^2 t^{1+\beta}). \tag{A.6}$$

The probability distribution can be calculated from the generating function,

$$P(x,t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \left\langle e^{ix(t)\theta}\right\rangle e^{-ix\theta}\, d\theta. \tag{A.7}$$

Thus, we can use (A.6) to obtain $P(x,t)$,

$$P(x,t) \sim \int_{-\infty}^{+\infty} e^{-c\theta^2 t^{1+\beta}} \cos(x\theta) d\theta \sim t^{-(1+\beta)/2)} \exp\left(-\text{const.}\, \frac{x^2}{t^{1+\beta}}\right). \tag{A.8}$$

The observation that $P(x,t)$ is Gaussian followed directly from the assumptions that $\{\xi(\omega)\}$ is Gaussian correlated noise. For non-Gaussian noise, it is much more complicated.

Before concluding, we derive (A.3) and (A.5). To derive (A.3) by using (A.2) and the inverse Fourier transform of (6b), we write

$$\langle x^2(t)\rangle = \frac{1}{4\pi^2}\int_0^t\int_0^t\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\langle\tilde{v}(\omega)\tilde{v}(\omega')\rangle e^{i(\omega t'+\omega' t'')}\,d\omega\,d\omega'\right]dt'\,dt''. \tag{A.9}$$

If $v(t)$ are long-range correlated [see Eq. (8)], i.e., if

$$\langle\tilde{v}(\omega)\tilde{v}(\omega')\rangle = |\omega\omega'|^{-\beta/2}\langle\xi(\omega)\xi(\omega')\rangle = |\omega\omega'|^{-\beta/2}\langle\xi(\omega)\xi^*(-\omega')\rangle \sim |\omega|^{-\beta}\delta(\omega+\omega'), \tag{A.10}$$

then Eq. (A.9) becomes

$$\langle x^2(t)\rangle \sim \int_0^t\int_0^t\left[\int_{-\infty}^{+\infty}|\omega|^{-\beta}e^{i\omega(t'-t'')}\,d\omega\right]dt'\,dt'' = \int_{-\infty}^{+\infty}|\omega|^{-\beta}\left|f(\omega,t)\right|^2\,d\omega, \tag{A.11}$$

where

$$\left|f(\omega,t)\right|^2 = \int_0^t\int_0^t e^{i\omega(t'-t'')}\,dt'\,dt'' = \frac{1}{\omega^2}\left|e^{i\omega t}-1\right|^2 = \frac{1-\cos(\omega t)}{\omega^2}. \tag{A.12}$$

We can relate the exponents α of Eq. (4a) and β of Eq. (4b):

$$\begin{aligned}\langle x^2(t)\rangle &\sim \int_0^{+\infty}\omega^{-2-\beta}[1-\cos(\omega t)]\,d\omega\\ &= t^{1+\beta}\int_0^{+\infty}u^{-2-\beta}[1-\cos(u)]\,du\\ &\sim t^{1+\beta}.\end{aligned} \tag{A.13}$$

Thus $\alpha=(1+\beta)/2$, as stated in Eq. (4d). This result is rigorous for positive β, since the integral over u is converged and independent of t. But for negative β, we must assume that there is a small frequency cut-off.

In order to derive (A.5), we must rephrase some results we have in a different form.

$$x(t) = \int_0^t v(t')\,dt' = \int_{t'=0}^t\frac{1}{2\pi}\int_{\omega=-\infty}^{+\infty}F(\omega)\,\xi(\omega)\,e^{i\omega t'}\,d\omega\,dt' = \int_{-\infty}^{+\infty}A(\omega,t)\,\xi(\omega)\,d\omega, \tag{A.14}$$

where ξ is Gaussian random variable and

$$A(\omega,t)\equiv\frac{1}{2\pi}F(\omega)\int_0^t e^{i\omega t'}\,dt' = |\omega|^{-\beta/2}\frac{(e^{i\omega t}-1)}{i\omega}. \tag{A.15}$$

The mean-square displacement can be expressed as

$$\langle x^2(t)\rangle \sim \int_{-\infty}^{+\infty}|A(\omega,t)|^2\,d\omega \sim t^{1+\beta}. \tag{A.16}$$

Thus we can calculate the generating function,

$$\left\langle e^{ix(t)\theta}\right\rangle = \left\langle \exp\left(i\theta \int_{-\infty}^{+\infty} A(\omega,t)\xi(\omega)\,d\omega\right)\right\rangle = \prod_{\{\omega\}} \left\langle \exp\left(i\theta A(\omega,t)\xi(\omega)\right)\right\rangle , \tag{A.17}$$

where $\prod_{\{\omega\}}$ is a product over a continuous variable ω. The average over all configurations is equivalent to the average over all possible sets of $\{\xi(\omega)\}$ with weight $P(\xi(\omega))$, which is a two-dimensional Gaussian distribution [$\xi(\omega)$ is complex]. Therefore, we have

$$\begin{aligned}\left\langle e^{ix(t)\theta}\right\rangle &= \prod_{\{\omega\}} \int\int_Z \exp\left(i\theta A(\omega,t)\xi(\omega)\right)\, P(\xi(\omega))\, d\xi(\omega)\\ &\sim \prod_{\{\omega\}} \int\int_Z \exp\left(i\theta A(\omega,t)\xi(\omega)\right)\, \exp\left(-\frac{|\xi(\omega)|^2}{2\sigma^2}\right)\, d\xi(\omega)\\ &\sim \prod_{\{\omega\}} \exp\left(-c\theta^2 |A(\omega,t)|^2\right)\\ &= \exp\left(-c\theta^2 \int_{-\infty}^{+\infty} |A(\omega,t)|^2\, d\omega\right). \end{aligned}\tag{A.18}$$

We combine Eqs. (A.16) and (A.18) to get

$$\left\langle e^{ix(t)\theta}\right\rangle \sim \exp\left(-c\theta^2 t^{1+\beta}\right). \tag{A.19}$$

Appendix B: Numerical Algorithm

There is a systematic way to generate stochastic variables with power law decay auto-correlation function [18-20,59]. Until now, we have considered the time to be a continuous variable, but for the purpose of computer simulation we want the time to be a discrete variable. We use the following assumptions to simplify the conversion of the time from continuous to discrete:

1. We assume that the time interval Δt is constant. In other words, the variable $v(t)$ is recorded at evenly spaced intervals in time. Thus, we have

$$v(t) \longrightarrow v(n) \equiv v(t = n\Delta t), \quad n = 0, 1, 2, \ldots \tag{B.1}$$

2. We assume that the time series $\{v(n)\}$ is periodic over a very large time interval $N\Delta t$.

Therefore, in the Fourier space the frequency (angular frequency divided by 2π) is discrete with interval $(N\Delta t)^{-1}$, and is also periodic with period $1/\Delta t$. Thus, the discrete Fourier transform maps N real valued variables [the $v(n)$, with $n = 0, 1, \ldots, N-1$] into N complex numbers [the $\tilde{v}(k)$, with $k = 0, 1, \ldots, N-1$],

$$\tilde{v}(k) \equiv \sum_{n=0}^{N-1} v(n)\, e^{-i2\pi kn/N}, \tag{B.2}$$

and

$$v(n) \equiv \frac{1}{N} \sum_{k=0}^{N-1} \tilde{v}(k)\, e^{i2\pi kn/N}. \tag{B.3}$$

Notice that $\tilde{v}(k) = \tilde{v}^*(-k)$, since the $\{v(n)\}$ are real. The relation between the continuous Fourier transform of $v(t)$ and the discrete Fourier transform when $\{v(n)\}$ are viewed as the discrete samples of $v(t)$ at an interval Δt is simply

$$\tilde{v}(\omega) \approx \tilde{v}(k)\Delta t, \tag{B.4}$$

where

$$\omega \equiv \frac{2\pi k}{N\Delta t}, \qquad k = 0, 1, \ldots, N-1. \tag{B.5}$$

The next step is to generate the discrete Fourier spectrum,

$$\tilde{v}(k) = \left|\frac{2\pi k}{N\Delta t}\right|^{-\beta/2} \xi(k) \qquad [-1 < \beta < 1], \tag{B.6}$$

which is the discrete version of Eq. (8). This can be achieved by first generating a sequence of pure random numbers $\{\xi(k)\}$ with Gaussian distribution, and multiply them according to (B.6) to obtain $\{\tilde{v}(k)\}$. Then, we Fourier transform $\{\tilde{v}(k)\}$, by using (B.3), to obtain $\{v(n)\}$.

The results that we derived in previous sections can be easily modified to discrete time and we notice that the long time behaviors are the same for both continuous and discrete cases. Since the time scale that we are interested in is much larger than Δt, we can assume $\Delta t \to 0$, which recovers the continuous case.

One can perform *correlated random walks* by using the velocities that generated from the above algorithm, and the displacement of the walker is

$$x(n) \equiv x(t = n\Delta t) = \sum_{i=0}^{n-1} v(i). \tag{B.7}$$

Thus one can measure the mean-square displacement $\langle x^2(n)\rangle$. The numerical data agree with (A.3) at large n, namely

$$\langle x^2(n)\rangle \sim n^{1+\beta}. \tag{B.8}$$

Note that we need the condition $N \gg n$. By varying β we can obtain anomalous diffusion from the extreme case of localization $[\langle x^2(n)\rangle \sim$ const., i.e., $\beta \to -1]$ to the other extreme of fast diffusion $[\langle x^2(n)\rangle \sim n^2$, i.e., $\beta \to 1]$.

ACKNOWLEDGEMENTS We wish to thank C. Cantor, C. DeLisi, J. M. Hausdorff, L. Liebovitch, J.Mietus, R. D. Rosenberg, J. J. Schwartz, M. Schwartz, F. Sciortino, M.F. Shlesinger, and N. Shworak for valuable interactions. 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.P by an NIH/NIMH Postdoctoral NRSA Fellowship, to MS by the American Heart Association, and to HES by the National Science Foundation.

[1] B. J. West and W. Deering, Phys. Reports **xx** (1994) xxx; 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** (1990) 42; B. J. West and M. F. Shlesinger, Am. Sci. **78** (1990) 40; B. J. West and A. L. Goldberger, Am. Sci., **75** (1987) 354; L. S. Liebovitch, J. Fischbarg and J. P. Koniarek, Math. Biosci. **89** (1987) 36; A. L. Goldberger and B. J. West, Yale J. Biol. Med. **60** (1987) 421; B. J. West and A. L. Goldberger, J. Appl. Physiol., **60** (1986) 189; L. S. Liebovitch, Biophys. J. **55** (1989) 373.

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

[3] A. A. Tsonis and P. A. Tsonis, Perspectives in Biology and Medicine **30** (1987) 355; F. Family, B. R. Masters, and D. E. Platt, Physica D **38** (1989) 98; F. Family, B. R. Masters, and D. E. Platt, Physica D **38** (1989) 98; M. Sernetz, J. Wübbeke, and P. Wlczek, Physica A **191** (1992) 13.

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

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

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

[7] T. Matsuyama, M. Sugawa, and Y. Nakagawa, FEMS Microb. Lett. **61**, 243 (1989); H. Fujikawa and M. Matsushita, J. Phys. Soc. Japan **58**, 387 (1989); Ibid **60**, 88 (1991); Physica A **168**, 498 (1990).

[8] T. Vicsek, M. Cserzö, and V. K. Horváth, Physica A **167** (1990) 315; S. Matsuura and S. Miyazima, Physica A **191** (1992) 30.

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

[10] 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).

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

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

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

[14] 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).

[15] 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).

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

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

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

[21] C.-K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, *Phys. Rev. Lett.* **70** (1993) 1343; 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** (1993) 47.

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

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

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

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

[27] R. Voss, *Phys. Rev. Lett.* **68** (1992) 3805; S. V. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, F. Sciortino, M. Simons and H. E. Stanley, Phys. Rev. Lett. **71** (1993) 1776; R. Voss, *Phys. Rev. Lett.* **71** (1992) 1775.

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

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

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

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

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

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

[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** (1992) 1.

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

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

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

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

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

[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.H.R. Stanley, and M. Simons, *Biophys. J.* **65** (1993) 2675-2681.

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

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

[44] E. I. Shakhnovich and A. M. Gutin, *Nature* **346** (1990) 773-775.

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

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

[47] In the power spectrum analysis for those sequences containing non-coding regions, subtracting of the white noise, $S(\infty)$, as performed in Ref. 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** (1993) 4, and refs. therein.

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

[49] W. Li, *International Journal of Bifurcation and Chaos* **2** (1992) 137.

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

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

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

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

[54] R. Schleif, *Science* **240** (1988) 127; C. J. Jaworski and J. Piatigorsky, *Nature* **337** (1989) 752.

[55] N. Eldredge and S. J. Gould, in *Models in Paleobiology*, edited by T. J. M. Schopf (Freeman and Cooper Inc., San Francisco, 1972), pp. 82f; *Nature* **366** (1993) 223.

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

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

[58] W. F. Doolitle, in *Intervening Sequences in Evolution and Development*, edited by E. Stone and R. Schwartz (Oxford University Press, New York, 1990), pp 42–62.

[59] C.-K. Peng, Ph.D. Thesis, Boston University, 1993.

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

[61] R. I. Kitney, and O. Rompelman, *The Study of Heart-Rate Variability* (Oxford University Press, London, 1980); S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. C. Barger and R. J. Cohen, *Science* **213** (1981) 220.

[62] M. Kobayashi and T. Musha, *IEEE Trans. Biomed. Eng.* **29** (1982) 456.

[63] A. L. Goldberger, D. R. Rigney, J. Mietus, E. M. Antman and S. Greenwald, *Experientia* **44** (1988) 983.

[64] R. V. Voss and J. Clarke, *Nature* **238** (1975) 317.

[65] R. V. Voss and J. Clarke, *J. Acous. Soc. Am.* **63** (1978) 258–263.

[66] M. Schroeder, *Fractals Chaos, Power Laws: Minutes from an Infinite Paradise* (W. H. Freeman, New York, 1991).

Fig. 1: Behavior of the binary mixture cyclohexane-aniline just below its consolute point—the fact that the correlations are long-range is manifest by the fact that visible light is strongly scattered (courtesy of R. A. Ferrell).

Fig. 2: The DNA walk representations of (a) human β-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. After Ref. 24.

Fig. 3: Comparison of the fluctuation analysis used in Ref. 37 and the DFA presented here. The DNA sequence is for the complete genome of lambda phage, whose DNA walk appears in the inset. The two parallel dotted lines have slope 0.5. A best fit straight line to $F_d(\ell)$ for the interval $\ell = 4$ to 1024 has slope $\alpha = 0.51$. After Ref. 42.

Fig. 4: 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 α 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 α for the 800 base pairs lying inside this box. Then we move the box 75 base pairs further along the chromosome, and again calculate α for the 800 base pairs lying inside this box. Iterating this procedure, we obtain $315{,}000/75 = 4186$ successive values of α, 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 α is generally small, while in between coding regions, there is frequently a peak in α. If α 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. 5: The interbeat interval $B_L(n)$ after low-pass filtering for (a) a healthy subject and (b) a patient with severe cardiac disease (dilated cardiomyopathy). The healthy heartbeat time series shows more complex fluctuations compared to the diseased heart rate fluctuation pattern that is close to random walk ("brown") noise. After Ref. 21.

Fig. 6: Log-log plot of $F(n)$ vs n. The circles represent $F(n)$ calculated from data in (a) and the triangles from data in (b). The two best-fit lines have slope $\alpha = 0.07$ and $\alpha = 0.49$ (fit from 200 to 4000 beats). The two lines with slopes $\alpha = 0$ and $\alpha = 0.5$ correspond to "$1/f$ noise" and "brown noise," respectively. We observe that $F(n)$ saturates for large n (of the order of 5000 beats), because the heartbeat interval are subjected to physiological constraints that cannot be arbitrarily large or small. The low-pass filter removes all Fourier components for $f \geq f_c$. The results shown here correspond to $f_c = 0.005$ beat^{-1}, but similar findings are obtained for other choices of $f_c \leq 0.005$. This cut-off frequency f_c is selected to remove components of heart rate variability associated with

physiologic respiration or pathologic Cheyne-Stokes breathing as well as oscillations associated with baroreflex activation (Mayer waves). After Ref. 21.

Fig. 7: The power spectrum $S_I(f)$ for the interbeat interval increment sequences over $\sim$ 24 hours for the same subjects in Fig. 1. (a) Data from a healthy adult. The best-fit line for the low frequency region has a slope $\beta = -0.93$. The heart rate spectrum is plotted as a function of "inverse beat number" (beat^{-1}) rather than frequency (time^{-1}) to obviate the need to interpolate data points. The spectral data are smoothed by averaging over 50 values. (b) Data from a patient with severe heart failure. The best-fit line has slope 0.14 for the low frequency region, $f < f_c = 0.005$ beat^{-1}. The appearance of a pathologic, characteristic time scale is associated with a spectral peak (arrow) at about 10^{-2} beat^{-1} (corresponding to Cheyne-Stokes respiration). After Ref. 21.

Fig. 8: Musical mapping of two heartbeat times series, derived from normal (top) and pathologic (bottom) data sets. The original heart beat time series were obtained from 24 hour recordings consisting of about 10^6 heartbeats. The heartbeat time series were then low-pass filtered to remove fluctuations > 0.05 (beat^{-1}), roughly equivalent to averaging every 200 beats. The pattern of fluctuations in the normal is more complex than that of the "music" generated from the abnormal data sets.