SCALING PROPERTIES OF DNA SEQUENCES AND HEARTBEAT RATE

S. Havlin

Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, U.S.A.

Gonda-Goldschmied Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, ISRAEL

S. V. Buldyrev, P. Ch. Ivanov, M. G. Rosenblum, H. E. Stanley and G. M. Viswanathan

Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, U.S.A.

C.-K. Peng

Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, U.S.A.

Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, U.S.A.

A. L. Goldberger

Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, U.S.A.

Department of Biomedical Engineering, Boston University, Boston, MA 02215, U.S.A.

I. LONG-RANGE POWER-LAW CORRELATIONS

In recent years long-range power-law correlations have been discovered in a remarkably wide variety of systems. Such long-range power-law correlations are a physical fact that in turn gives rise to the increasingly appreciated "fractal geometry of nature"

[?,?,?,?,?,?,?,?,?,?,?,?,?]. Recognizing the ubiquity of long-range power-law correlations can help us in our efforts to understand nature, since as soon as we find power-law correlations we can quantify them with a critical exponent. Quantification of this kind of scaling behavior for apparently unrelated systems allows us to recognize similarities between different systems, leading to underlying unifications that might otherwise have gone unnoticed.

Traditionally, investigators in many fields characterize processes by assuming that correlations decay exponentially. However, there is one major exception: at the critical point, the exponential decay turns into a power law decay [?]

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

Many systems drive themselves spontaneously toward critical points [?,?].

In the following sections we will attempt to summarize some recent findings [?,?] concerning the possibility that—under suitable conditions—the sequence of base pairs or "nucleotides" in DNA also displays power-law correlations. The underlying basis of such power law correlations is not understood at present, but this discovery has intriguing implications for molecular evolution [?], as well as potential practical applications for distinguishing coding and noncoding regions in long nucleotide chains [?].

II. DNA

The role of genomic DNA sequences in coding for protein structure is well known [?]. The human genome contains information for approximately 100,000 different proteins, which define all inheritable features of an individual. The genomic sequence is likely the most sophisticated information database created by nature through the dynamic process of evolution. Equally remarkable is the precise transformation of information (duplication, decoding, etc) that occurs in a relatively short time interval.

The building blocks for coding this information are called *nucleotides.* Each nucleotide contains a phosphate group, a deoxyribose sugar moiety and either a *purine* or a *pyrimidine*

base. Two purines and two pyrimidines are found in DNA. The two purines are adenine (A) and guanine (G); the two pyrimidines are cytosine (C) and thymine (T). The nucleotides are linked end to end, by chemical bonds from the phosphate group of one nucleotide to the deoxyribose sugar group of the adjacent nucleotide, forming a long polymer (*polynucleotide*) chain. The information content is encoded in the sequential order of the bases on this chain. Therefore, as far as the information content is concerned, a DNA sequence can be most simply represented as a symbolic sequence of four letters: A, C, G and T.

In the genomes of high eukaryotic organisms only a small portion of the total genome length is used for protein coding (as low as 3% in the human genome). The segments of the chromosomal DNA that are spliced out during the formation of a mature mRNA are called *introns* (for intervening sequences). The coding sequences are called *exons* (for expressive sequences).

The role of introns and intergenomic sequences constituting large portions of the genome remains unknown. Furthermore, only a few quantitative methods are currently available for analyzing information which is possibly encrypted in the noncoding part of the genome.

III. THE "DNA WALK"

One interesting question that may be asked by statistical physicists would be whether the sequence of the nucleotides A,C,G, and T behaves like a one-dimensional "ideal gas", where the fluctuations of density of certain particles obey Gaussian law, or if there exist long range correlations in nucleotide content (as in the vicinity of a critical point). These result in domains of all size with different nucleotide concentrations. Such domains of various sizes were known for a long time but their origin and statistical properties remain unexplained. A natural language to describe heterogeneous DNA structure is long-range correlation analysis, borrowed from the theory of critical phenomena [**?**].

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* [?]. 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 [?,?,?].

One definition of the DNA walk is that the walker steps "up" if a pyrimidine (C or T) occurs at position i along the DNA chain, while the walker steps "down" if a purine (A or G) occurs at position i. The question we asked was whether such a walk displays only short-range correlations (as in an n-step Markov chain) or long-range correlations (as in critical phenomena and other scale-free "fractal" phenomena). A different kind of DNA walk was suggested by Azbel [?].

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

The DNA walk allows one to visualize directly the fluctuations of the purine-pyrimidine content in DNA sequences: Positive slopes correspond to high concentration of pyrimidines, while negative slopes correspond to high concentration of purines. Visual observation of DNA walks suggests that the coding sequences and intron-containing noncoding sequences have quite different landscapes.

IV. CORRELATIONS AND FLUCTUATIONS

An important statistical quantity characterizing any walk [?,?] is the root mean square fluctuation $F(\ell)$ about the average of the displacement of a quantity $\Delta y(\ell)$ defined by $\Delta y(\ell) \equiv y(\ell_0 + \ell) - y(\ell_0)$, where

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

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

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

with $\alpha \neq 1/2$.

Figure 1a shows a typical example of a gene that contains a significant fraction of base pairs that do *not* code for amino acids. It is immediately apparent that the DNA walk has an extremely jagged contour which corresponds to long-range correlations.

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

On the other hand, the dependence of $F(\ell)$ for coding sequences is not linear on the log-log plot: its slope undergoes a crossover from 0.5 for small ℓ to 1 for large ℓ. However, if a single patch is analyzed separately, the log-log plot of $F(\ell)$ is again a straight line with the slope close to 0.5. This suggests that within a large patch the coding sequence is almost uncorrelated. The function $F(\ell)$ was also studied for DNA sequences by Azbel [?].

V. DETRENDED FLUCTUATION ANALYSIS (DFA)

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

of alternating strand bias—and so we have non-stationarity. Hence conventional scaling analyses cannot be applied reliably to the entire sequence but only to sub-sequences.

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

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

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

VI. SYSTEMATIC ANALYSIS OF GENBANK DATABASE

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

long-range correlation properties.

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

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

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

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

Finally, we note that although the scaling exponents α and β have potential use in

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

VII. GENERALIZED LÉVY WALK MODEL

Although the correlation is long-range in the noncoding sequences, there seems to be a paradox: long *uncorrelated* regions of up to thousands of base-pairs can be found in such sequences as well. For example, consider the human beta-globin intergenomic sequence of length $L = 73,326$ (GenBank name: HUMHBB). This long noncoding sequence has 50% purines (no *overall* strand bias) and $\alpha = 0.7$ (see Fig. 1(a)). However, from nucleotide #67,089 to #73,228, there occurs the LINE-1 region (defined in Ref. [**?**]). In this region of length 6139 base pairs, there is a strong strand bias with 59% *purines.* In this noncoding sub-region, we find power-law scaling of F, with $F \sim l^{\alpha}$, with $\alpha = 0.55$, quite close to that of a random walk.

Even more striking is another region of 6378 base pairs, from nucleotide #23,137 to #29,515, which has 59% *pyrimidines* and is *uncorrelated,* with remarkably good power-law scaling and correlation exponent $\alpha = 0.49$ (Fig. 1(b)). This region actually consists of three sub-sequences, complementary to shorter parts of the LINE-1 sequence.

These features motivated us to apply a generalized Lévy walk model (see Figs. 1c, 1d and 2) for the noncoding regions of DNA sequences [**?**]. We will show in the next section how this model can explain the long-range correlation properties, since there is no characteristic scale "built into" this generalized Lévy walk. In addition, the model simultaneously accounts

for the observed large sub-regions of non-correlated sequences within these noncoding DNA chains.

The classic Lévy walk model describes a wide variety of diverse phenomena that exhibit long-range correlations [?,?,?,?]. The model is defined schematically in Fig. 2a: A random walker takes not one but l_1 steps in a given direction. Then the walker takes l_2 steps in a new randomly-chosen direction, and so forth. The lengths l_j of each string are chosen from a probability distribution, with

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

where $\sum_{i=1}^{N} l_i = L$, N is the number of sub-strings and L is the total number of steps that the random walker takes.

We consider a generalization of the Lévy walk [?] to interpret recent findings of long-range correlation in noncoding DNA sequences described above. Instead of taking l_j steps in the *same* direction as occurs in a classic Lévy walk, the walker takes each of l_j steps in *random* directions, with a fixed bias probability

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

to go up and

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

to go down, where ϵ_j gets the values $+\epsilon$ or $-\epsilon$ randomly. Here $0 \leq \epsilon \leq 1$ is a bias parameter (the case $\epsilon = 1$ reduces to the Lévy walk). Fig. 2b shows such a generalized Lévy walk for the same choice of l_j as in Fig. 2a.

As shown in Ref. [?], the generalized Lévy walk—like the pure Lévy walk—gives rise to a landscape with a fluctuation exponent α that depends upon the Lévy walk parameter μ [?,?],

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

i.e., non-trivial behavior of α corresponds to the case $2 < \mu < 3$ where the first moment of $P(l_j)$ converges while the second moment diverges. The long-range correlation property for the Lévy walk, in this case, is a consequence of the broad distribution of Eq. (??) that lacks of a characteristic length scale. However, for $\mu \geq 3$, the distribution of $P(l_j)$ decays fast enough that an effective characteristic length scale appears. Therefore, the resulting Lévy walk behaves like a normal random walk for $\mu \geq 3$.

VIII. MOSAIC NATURE OF DNA STRUCTURE

The key finding of this analysis is that a generalized Lévy walk model can account for two hitherto unexplained features of DNA nucleotides: (i) the long-range power law correlations that extend over thousands of nucleotides in sequences containing noncoding regions (e.g., genes with introns and intergenomic sequences), and (ii) the presence within these correlated sequences of sometimes large sub-regions that correspond to biased random walks. This apparent paradox is resolved by the generalized Lévy walk, a mechanism for generating long-range correlations (no characteristic length scale), that with finite (though rare) probability also generates large regions of uncorrelated strand bias. The uncorrelated sub-regions, therefore, are an anticipated feature of this mechanism for long-range correlations.

From a biological viewpoint, two questions immediately arise: (i) What is the significance of these uncorrelated sub-regions of strand bias? and (ii) What is the molecular basis underlying the power-law statistics of the Lévy walk? With respect to the first question, we note that these long uncorrelated regions at least sometimes correspond to well-described but poorly understood sequences termed "repetitive elements", such as the LINE1 region noted above [?,?]. There are at least 53 different families of such repetitive elements within the human genome. The lengths of these repetitive elements vary from 10 to 10^4 nucleotides [?]. At least some of the repetitive elements are believed to be remnants of messenger RNA molecules that formerly did code for proteins [?,?,?]. Alternatively, these segments may represent retroviral sequences that have inserted themselves into the genome [?]. Our

finding that these repetitive elements have the statistical properties of biased random walks (e.g., the same as that of active coding sequences) is consistent with both of these hypotheses.

Finally, what are the biological implications of this type of analysis? Our findings clearly support the following possible hypothesis concerning the molecular basis for the power-law distributions of elements within DNA chains. In order to be inserted into DNA, a macromolecule should form a loop of a certain length l with two ends, separated by l nucleotides along the sequence, coming close to each other in real space. The probability of finding a loop of length l inside a very long linear polymer scales as $l^{-\mu}$ [?,?]. Theoretical estimates of μ made by different methods [?,?,?] using a self-avoiding random walk model [?] indicate that the value of μ for three-dimensional model is between 2.16 and 2.42. Our estimate made by the Rosenbluth Monte-Carlo Method [?] gave $\mu = 2.22 \pm 0.05$ which yields $\alpha = 0.89$, a larger value than the effective value observed in DNA of finite length. However, the asymptotic value of the exponent α remains uncertain since the statistics of Lévy walks converge very slowly due to rare events associated with the very long strings of constant bias that may occur in the sequence according to Eq. (??).

Recently the size distribution of insertions and deletions in human and rodent pseudogenes has been studied by Gu and Li [?]. They found that both distributions are characterized by a power-law behavior. This finding supports the assumption made in this model.

In summary, it is clear that the behavior of DNA sequences cannot be satisfactorily explained in terms of only one characteristic length scale even of about $10^3 - 10^4$ base pairs long. The asymptotic behavior of the scaling exponent α and whether it reaches some universal value for long DNA chains must await further data from the Human Genome Project.

IX. DETECTING CHARACTERISTIC PATCH SIZES

Scaling methods, such as long-range correlations, may provide important information on the presence in DNA sequences of large patches composed of different nucleotide concen-

trations and of different length scales. The DFA method allows one to identify the typical length of such elements. We study the local slope of the $F_D(\ell)$ on the double logarithmic plot

$$\alpha(\ell) = \frac{d \log F_D(\ell)}{d \log(\ell + 3)}, \tag{8}$$

which is not constant but depends on the nucleotide distance ℓ. It is hypothesized that the maxima of the $\alpha(\ell)$ may correspond to patches of different nucleotide concentrations of length ℓ.

To test this hypothesis we construct artificial sequences with built-in patches of varying lengths. In Fig. **??** we show an example of the DFA analysis of such a sequence. The sequence in Fig. **??** is constructed by concatenating uncorrelated patches of 200 bp, 2000 bp, and 20000 bp. Patches with 70% of purines are randomly alternated with patches with 70% of pyrimidines. The smallest patch size with the highest probability, and the largest patch size with the smallest probability are chosen according to the following rule [**?**]. For the jth patch,

(i) A random number x_j is chosen in the interval $[0, 1]$.

(ii) A preliminary length ℓ_j is computed as $\ell_j = 200/x_j$.

(iii) If ℓ_j is less than 2000 then a patch of size 200 bp is chosen. Otherwise if ℓ_j is less than 20000 then a patch of size 2000 bp is chosen. Otherwise a patch of size 20000 is chosen.

It can be shown analytically that peaks should occur at scales of approximately 1.5 times the patch sizes. Therefore, by looking for peaks in $\alpha(\ell)$ we can estimate characteristic DNA patch sizes embedded in a sequence with an apparent $1/f$ power spectrum. During year 1 we will develop algorithms for identifying characteristic patch sizes and their nucleotide compositions in DNA sequences using control sequences with various patch sizes and nucleotide compositions as a "training set." Different binary mapping rules can detect patches of different nucleotide compositions, i.e., patches with C+G content can be detected by RY or KM rules. For these purposes, we will also apply the power spectra analysis and the wavelet analysis which was recently applied to studies of DNA sequences [**?**]. The Fourier

and wavelet analysis will be briefly discussed under Aim III of the research design methods. An information theory approach was recently used [?] to detect patches in DNA sequences. We will test their technique and develop new techniques based on information theory to detect patches.

Having developed the techniques for detecting and examining characteristic scales of patchiness in model sequences, we will apply these methods to real data. Figure ?? shows estimated characteristic patch sizes of the SW rule for several eukaryotic sequences longer than 100000 bp as well as for some *E. coli* bacterial sequences. Similar patch sizes appear in several sequences, and some even appear on sequences from different species, suggesting that the complex global structure of genomic DNA may have some universal characteristics. The patchiness in eukaryotic DNA could be partially due to the elaborate organization and folding of DNA by proteins into nucleosomes and higher-order structures of chromatin or could be due to the abundance of interspersed repeats such as LINE-1 or Alu, or due to the particular distribution of genes along chromosomes with characteristic gene sizes and intergenic distances. We will address these possibilities under Aim II. We will create a full map of patch sizes of all available eukaryotic and prokaryotic genomes in years 1 and 2 and will study how the patch size changes across the phylogenetic spectra. We will use these results under Aim II where we will test various hypotheses of patch formation and relate them to known biological phenomena.

X. FRACTAL ANALYSIS OF INTERBEAT INTERVALS

Very recently, the idea of long-range correlations has been extended to the analysis of the beat-to-beat intervals in the normal and diseased heart [?,?]. 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 [?]. 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) [**?**,**?**,**?**].

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. 9. We observe a more complex pattern of fluctuations for a representative healthy adult (Fig. 9a) compared to the "smoother" pattern of interbeat intervals for a subject with severe heart disease (Fig. 9b). 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

$$F(n) \equiv \overline{|B_L(n'+n) - B_L(n')|}, \tag{9}$$

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 10 is a log-log plot of $F(n)$ vs n for the data in Figs. 9a and 9b. This plot is approximately linear over a broad physiologically-relevant time scale ($200 - 4000$ beats) implying that

$$F(n) \sim n^{\alpha}. \tag{10}$$

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 [**?**]. Figures 11a and 11b 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. 8. The fact that the log-log plot of $S_I(f)$ vs f is linear implies

$$S_I(f) \sim \frac{1}{f^\beta}. \tag{11}$$

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

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

2. 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.

3. 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") [**?**].

For the diseased data set, we observe a flat spectrum ($\beta \approx 0$) in the low frequency region (Fig. 11b) 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. 11a), 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.

XI. SCALING BEHAVIOR OF HEARTBEAT INTERVALS

Time series of beat-to-beat (RR) heart rate intervals (Fig. 12a) obtained from digitized electrocardiograms are known to be nonstationary and exhibit extremely complex behavior [**?**]. A typical feature of these signals is the presence of "patchy" patterns which change over time (Fig. 12b). Heterogeneous properties may be even more strongly expressed in certain cases of abnormal heart activity. Traditional approaches — such as the power spectrum and correlation analysis [**?**,**?**] — are not suited for such nonstationary (patchy) sequences, and do not carry information stored in the Fourier phases (crucial for determining nonlinear characteristics).

To address these problems, we present an alternative method — "*cumulative variation magnitude analysis*" — to study the subtle structure of physiological time series. This method comprises sequential application of a set of algorithms based on wavelet and Hilbert transform analysis. First, we apply the wavelet transform (Fig. 12c), because it does not

require stationarity and preserves the Fourier phase information. The wavelet transform [?,?,?] of a time series $s(t)$ is defined as

$$T_\psi(t_0, a) \equiv \; a^{-1} \int_{-\infty}^{+\infty} s(t)\psi\left(\frac{t - t_0}{a}\right) dt, \tag{12}$$

where the analyzing wavelet ψ has a width of the order of the scale a and is centered at t_0. For high frequencies (small a), the ψ functions have good localization (being effectively non-zero only on small sub-intervals), so short-time regimes or high-frequency components can be detected by the wavelet analysis. The wavelet transform is sometimes called a "mathematical microscope" because it allows one to study properties of the signal on any chosen scale a. However, a wavelet with too large a value of scale a (low frequency) will filter out almost the entire frequency content of the time series, thus losing information about the intrinsic dynamics of the system. We focus our "microscope" on scale $a = 8$ beats which smoothes locally very high-frequency variations and best probes patterns of specific duration ($\approx \frac{1}{2} - 1$ min) (Fig. 13). The wavelet transform is attractive because it can eliminate local polynomial behavior in the nonstationary signal by an appropriate choice of the analysing wavelet ψ [?]. In our study we use derivatives of the Gaussian function: $\psi^{(n)} = d^n/dt^n e^{-\frac{1}{2}t^2}$.

The wavelet transform is thus a cumulative measure of the variations in the heart rate signal over a region proportional to the wavelet scale, so study of the behavior of the wavelet values can reveal *intrinsic properties of the dynamics* masked by nonstationarity.

The second step of the cumulative variation magnitude analysis is to extract the instantaneous variation amplitudes of the wavelet-filtered signal by means of an analytic signal approach [?,?] which also does not require stationarity. Let $s(t)$ represent an arbitrary signal. The analytic signal, a complex function of time, is defined by $S(t) = s(t) + i\tilde{s}(t) = A(t)e^{i\phi(t)}$, where $\tilde{s}(t)$ is the Hilbert transform [?] of $s(t)$. The instantaneous magnitude $A(t)$ and the instantaneous phase of the signal $\phi(t)$ are defined as $A(t) \equiv \sqrt{s^2(t) + \tilde{s}^2(t)}$ and $\phi(t) \equiv \tan^{-1}(\tilde{s}(t)/s(t))$.

We study the distribution of the amplitudes of the beat-to-beat variations (Fig. 12d) for a group of healthy subjects ($N = 18$; 5 male, 13 female; age: $20-50$, mean - 34) and a group

of subjects [?] with obstructive sleep apnea [?] ($N = 16$ males; age: $32-56$, mean - 43). We begin by considering night phase (12pm-6am) records of interbeat intervals ($\approx 10^4$ beats) for both groups to minimize nonstationarity due to changes in the level of activity. Inspection of the distribution functions of the amplitudes of the cumulative variations reveals marked differences between individuals (Fig. 13a). These discrepancies are not surprising given the underlying physiological differences among healthy subjects. To test the hypothesis that there is a hidden, possibly universal structure to these heterogeneous time series, we rescale the distributions and find for all healthy subjects that the data conform to a single scaled plot ("*data collapse*") (Fig. 13b). Such behavior is reminiscent of a wide class of well-studied physical systems with universal scaling properties [?,?]. In contrast, the subjects with *sleep apnea* show individual probability distributions which *fail* to collapse (Fig. 13d).

We next analyse the distributions of the beat-to-beat variation amplitudes. For the healthy group, we find that these are well fit by the Gamma form: $P(x) = (b^{\nu+1}/\Gamma(\nu + 1))x^{\nu}e^{-bx}$, where $b = \nu/x_0$, $\Gamma(\nu + 1)$ is the Gamma function, x_0 is the position of the peak $P = P_{max}$, and ν is the fitting parameter (Fig. 14a). Although individual distributions have different values of b, the homogeneous property of the functional form of $P(x)$ leads to reduction of the independent variable x and parameter b to a single scaled variable $u \equiv bx$. Instead of the data points falling on a family of curves, one for each value of b, we find the data points *collapse* onto a single curve given by the scaling function $\tilde{P}(u) \equiv P(x)/b$. Thus, it is sufficient to specify only one parameter b in order to characterize the heterogeneous heartbeat variations of each subject in this group.

We also analysed heart rate dynamics for the healthy subjects during day-time hours (noon — 6pm). Our results indicate that the observed, apparently universal behavior holds not only for the night phase but for the day phase as well (Fig. 14b).

This study uncovers a previously unknown nonlinear feature of healthy heart rate fluctuations. Prior reports of universal properties of the normal heart beat and other physiological signals were related to long-range correlations and power law scaling [?,?,?]. However, these properties, detected by Fourier and fluctuation analysis techniques, ignore information re-

lated to the phase interactions of component modes. The nonlinear interaction of these modes accounts for the patchy, non-homogeneous appearance of the heartbeat time series.

Our finding suggests that for healthy individuals, there may be a common structure to this nonlinear phase interaction. The scaling property cannot be accounted for by activity, since we analysed data from subjects during nocturnal hours. Moreover, it cannot be accounted for by sleep stage transitions, since we found a similar pattern during day-time hours. The basis of this robust temporal structure remains unknown and presents a new challenge to understanding nonlinear mechanisms of heartbeat control.

Additionally, we find that subjects with sleep apnea, a common and important instability of cardiopulmonary control, show a dramatic alteration in the scaling pattern—possibly related to pathologic mode locking associated with periodic breathing dynamics [**?**]. Thus, the dual use of wavelet and Hilbert transform techniques may be of practical diagnostic and prognostic value, and may also be applicable to a wide range of heterogeneous, "real world" physiological signals.

XII. ACKNOWLEDGMENTS

We are grateful to many individuals, including M. E. Matsa, S. M. Ossadnik, and F. Sciortino, for major contributions to those results reviewed here that represent collaborative research efforts. We also wish to thank C. Cantor, C. DeLisi, M. Frank-Kamenetskii, A. Yu. Grosberg, G. Huber, I. Labat, L. Liebovitch, G. S. Michaels, P. Munson, R. Nossal, R. Nussinov, R. D. Rosenberg, J. J. Schwartz, M. Schwartz, E. I. Shakhnovich, M. F. Shlesinger, N. Shworak, and E. N. Trifonov for valuable discussions. Partial support was provided by the National Science Foundation, National Institutes of Health (Human Genome Project), the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute, the National Aeronautics and Space Administration, the Israel-USA Binational Science Foundation, Israel Academy of Sciences, and (to C-KP) by an NIH/NIMH Postdoctoral NRSA Fellowship.

REFERENCES

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

[2] A. Bunde, S. Havlin, eds.: *Fractals and Disordered Systems* (Springer-Verlag, Berlin 1991) A. Bunde, S. Havlin, eds.: *Fractals in Science* (Springer-Verlag, Berlin 1994); T. Vicsek, M. Shlesinger, M. Matsushita, eds.: *Fractals in Natural Sciences* (World Scientific, Singapore, 1994)

[3] J.M. Garcia-Ruiz, E. Louis, P. Meakin, L. Sander, eds.: *Growth Patterns in Physical Sciences and Biology* [Proc. 1991 NATO Advanced Research Workshop, Granada, Spain, October 1991], (Plenum, New York, 1993)

[4] A.Yu. Grosberg, A.R. Khokhlov: *Statistical Physics of Macromolecules,* translated by Y. A. Atanov (AIP Press, New York, 1994)

[5] J.B. Bassingthwaighte, L.S. Liebovitch, B.J. West: *Fractal Physiology* (Oxford University Press, New York, 1994)

[6] A.-L. Barabási, H.E. Stanley: *Fractal Concepts in Surface Growth* (Cambridge University Press, Cambridge, 1995)

[7] B.J. West, A.L. Goldberger: J. Appl. Physiol., **60**, 189 (1986); B.J. West, A.L. Goldberger: Am. Sci., **75**, 354 (1987); A.L. Goldberger, B.J. West: Yale J. Biol. Med. **60**, 421 (1987); A.L. Goldberger, D.R. Rigney, B.J. West: Sci. Am. **262**, 42 (1990); B.J. West, M.F. Shlesinger: Am. Sci. **78**, 40 (1990); B.J. West: *Fractal Physiology and Chaos in Medicine* (World Scientific, Singapore 1990); B.J. West, W. Deering: Phys. Reports **246**, 1 (1994); S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley: in *Fractals in Science,* edited by A. Bunde and S. Havlin (Springer-Verlag, Berlin, 1994), pp. 49–83

[8] T. Vicsek: *Fractal Growth Phenomena, Second Edition* (World Scientific, Singapore 1992)

[9] J. Feder: *Fractals* (Plenum, NY, 1988)

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

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

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

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

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

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

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

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

[18] R. Voss: Phys. Rev. Lett. **68**, 3805 (1992); R. Voss: Fractals **2**, 1 (1994)

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

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

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

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

[23] P. Yam: Sci. Am. **267[3]**, 23 (1992)

[24] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley: Physica A **191**, 25 (1992); H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, J.M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino, M. Simons: Physica A **191**, 1 (1992); H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng, F. Sciortino and M. Simons, "Fractals in Biology and Medicine," in *Diffusion Processes: Experiment, Theory, Simulations Proceedings of the Vth M. Born*

Symposium, edited by A. Pekalski (Springer-Verlag, Berlin, 1994), pp. 147–178; H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng and M. Simons, "Statistical Mechanics in Biology: How Ubiquitous are Long-Range Correlations?" *Proc. International Conference on Statistical Mechanics,* Physica A **205**, 214 (1994).

[25] C.A. Chatzidimitriou-Dreismann, D. Larhammar: Nature **361**, 212 (1993); D. Larhammar, C.A. Chatzidimitriou-Dreismann: Nucleic Acids Res. **21**, 5167 (1993) C.A. Chatzidimitriou-Dreismann, R.M.F. Streffer, D. Larhammar: Biochim. Biophys. Acta **1217**, 181 (1994); C.A. Chatzidimitriou-Dreismann, R.M.F. Streffer, D. Larhammar: Eur. J. Biochem. **224**, 365 (1994)

[26] A.Yu. Grosberg, Y. Rabin, S. Havlin, A. Neer: Europhys. Lett. **23**, 373 (1993)

[27] S. Karlin, V. Brendel: Science **259**, 677 (1993)

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

[29] N. Shnerb, E. Eisenberg: Phys. Rev. E **49**, R1005 (1994)

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

[31] A. S. Borovik, A. Yu. Grosberg and M. D. Frank Kamenetski, J. Biomolec. Structure and Dynamics **12**, 655-669 (1994)

[32] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, M.H.R. Stanley, M. Simons: Biophys. J. **65**, 2673 (1993)

[33] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger: Phys. Rev. E **49**, 1685 (1994)

[34] S.M. Ossadnik, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.-K. Peng, M. Simons, H.E. Stanley: Biophys. J. **67**, 64 (1994); H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K.Peng, M. Simons: [Proceedings of Internat'l Conf. on Con-

densed Matter Physics, Bar-Ilan], Physica A **200**, 4 (1993); H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, S.M. Ossadnik, C.-K. Peng, M. Simons: Fractals **1**, 283-301 (1993); S. Havlin, S. V. Buldyrev, A. L. Goldberger, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng, M. Simons, and H. E. Stanley, Chaos, Solitons, and Fractals **6**, 171 (1995).

[35] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, and H.E. Stanley: Phys. Rev. E**51**, 5084 (1995)

[36] S. Tavaré, B.W. Giddings, in: *Mathematical Methods for DNA Sequences,* Eds. M.S. Waterman (CRC Press, Boca Raton 1989), pp. 117-132; J.D. Watson, M. Gilman, J. Witkowski, M. Zoller: *Recombinant DNA* (Scientific American Books, New York 1992).

[37] E.W. Montroll, M.F. Shlesinger: "The Wonderful World of Random Walks" in: *Non-equilibrium Phenomena II. From Stochastics to Hydrodynamics,* ed. by J.L. Lebowitz, E.W. Montroll (North-Holland, Amsterdam 1984), pp. 1–121

[38] G.H. Weiss: *Random Walks* (North-Holland, Amsterdam 1994)

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

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

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

[42] M. Y. Azbel: Phys. Rev. Lett. **31**, 589 (1973).

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

[44] M. Y. Azbel: Biopolymers **21**, 1687 (1982).

[45] A. Arneodo, E. Bacry, P. V. Graves, J. F. Muzy: Phys. Rev. Lett. **74**, 3293–3296 (1995).

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

[47] M.F. Shlesinger, J. Klafter: in *On Growth and Form: Fractal and Non-Fractal Patterns in Physics,* edited by H.E. Stanley and N. Ostrowsky (Martinus Nijhoff, Dordrecht, 1986), p. 279ff

[48] M.F. Shlesinger, J. Klafter, Y.M. Wong: J. Stat. Phys. **27**, 499 (1982)

[49] M.F. Shlesinger, J. Klafter: Phys. Rev. Lett. **54**, 2551 (1985)

[50] R.N. Mantegna: Physica A **179**, 232 (1991)

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

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

[53] E. Zuckerkandl, G. Latter, J. Jurka: J. Mol. Evol. **29**, 504 (1989)

[54] B. Levin: *Genes IV* (Oxford University Press, Oxford, 1990)

[55] P.-G. de Gennes: *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca NY, 1979)

[56] J. de Cloiseaux: J. Physique **41**, 223 (1980), p. 223

[57] S. Redner: J. Phys. A **13**, 3525 (1980)

[58] A. Baumgartner: Z. Phys. B **42**, 265 (1981)

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

[60] X. Gu, W.-H. Li: J. Mol. Evol. **40**, 464–473 (1995)

[61] A. Schenkel, J. Zhang, Y-C. Zhang: Fractals **1**, 47 (1993); M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, N. Shnerb: Fractals **2**, 7 (1994)

[62] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch [eds.], *Lévy Flights and Related Topics in Physics* (Springer-Verlag, Berlin, 1995).

[63] R. Wilson *et al*: Nature **368**, 32–37 (1994).

[64] S. G. Oliver and approx. 50 co-authors: Nature **357**, 38–46 (1992).

[65] M. Y. Leung, B. E. Blaisdell, C. Burge, S. Karlin: J. Mol. Biol. **221**, 1367–1378 (1991).

[66] S. Karlin and I. Ladunga: Proc. Natl. Acad. Sci. USA **91**(26), 12832–12836 (1994).

[67] P. Bernaola-Galvan, Ramon Roman-Roldan and Jose L. Oliver: Phys. Rev. E **53**, 5181–5189 (1996).

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

[69] W. B. Cannon: Physiol. Rev. **9** 399 (1929).

[70] 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** 220 (1981).

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

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

[73] R. I. Kitney, D. Linkens, A. C. Selman, A. H. McDonald: Automedica **4**, 141–153 (1982)

[74] D. Panter: Modulation, Noise and Spectral Analysis (McGraw-Hill, New York, 1965).

[75] S. Akselrod, *et al.*: Science **213**, 220–222 (1981)

[76] A. Grossmann, & J. Morlet: *Mathematics and Physics, Lectures on Recent Results* (World Scientific, Singapore, 1985).

[77] I. Daubechies: Comm. Pure and Appl. Math. **41**, 909–996 (1988).

[78] J. F. Muzy, E. Bacry, A. Arneodo: Int. J. Bifurc. Chaos **4**, 245–302 (1994).

[79] L. A. Vainshtein, & D. E. Vakman: *Separation of Frequencies in the Theory of Oscillations and Waves* (Nauka, Moscow, 1983).

[80] D. Gabor: J. Inst. Elect. Engrs. **93**, 429–457 (1946).

[81] *MIT-BIH Polysomnographic Database CD-ROM, second edition* (MIT-BIH Database Distribution, Cambridge, 1992)

[82] C. Guilleminault, S. Connolly, R. Winkle, K. Melvin, A. Tilkian: Lancet **1**, 126–131 (1984).

[83] T. Vicsek: *Fractal Growth Phenomena* second edition (World Scientific, Singapore, 1992);

[84] J. B. Bassingthwaighte, L. S. Liebovitch, B. J. West: *Fractal Physiology* (Oxford University Press, New York, 1994).

[85] C.-K. Peng, *et al.*: Chaos **5**, 82–87 (1995).

[86] A. A. Aghili, M. Rizwan-uddin, Pamela Griggin, J. R. Moorman: Phys. Rev. Lett. **74**, 1254–1257 (1995).

[87] L. A. Lipsitz, *et al.*: Br. Heart J. **74**, 340–396 (1995)

[88] G. L. Gerstein, B. Mandelbrot: Biophys. J. **4**, 41–68 (1964)

FIGURES

FIG. 1. DNA walk displacement $y(\ell)$ (excess of purines over pyrimidines) vs nucleotide distance ℓ for (a) HUMHBB (human beta globin chromosomal region of the total length $L = 73,239$); (b) the LINE1c region of HUMHBB starting from 23,137 to 29,515; (c) the generalized Lévy walk model of length 73,326 with $\mu = 2.45$, $l_c = 10$, $\alpha_o = 0.6$, and $\epsilon = 0.2$; and (d) a segment of a Lévy walk of exactly the same length as the LINE1c sequence from step 67,048 to the end of the sequence. This sub-segment is a Markovian random walk. Note that in all cases the overall bias was subtracted from the graph such that the beginning and ending points have the same vertical displacement ($y = 0$). This was done to make the graphs clearer and does not affect the quantitative analysis of the data.

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

FIG. 3. Displacement $y(\ell)$ vs number of steps for (a) the classical Lévy walk model consisting of 6 strings of l_j steps, each taken in alternating directions; (b) the generalized Lévy walk model consisting of 6 biased random walks of the same length with a probability of p_+ that it will go up equal to $(1 \pm \epsilon)/2$ [$\epsilon = 0.2$]; and (c) the unbiased uncorrelated random walk. Note that the vertical scale in (b) and (c) is twice that in (a).

FIG. 4. (a) Power spectrum for the artificial control DNA sequence with built-in patches of 200 bp, 2000 bp, and 20000 bp discussed in the text. The spectrum looks remarkably like "$1/f$-type" spectra, showing that studying the spectrum alone can be misleading. (b) Log-linear plot of the DFA exponent $\alpha(\ell)$ for the same model. The exponent $\alpha(\ell)$ is found by calculating the local slope of the double-log plot of the DFA function [see Eq. (??)]. The exponent $\alpha(\ell)$ peaks at three locations corresponding to the three characteristic patch sizes. As expected from theory, the peaks occur at approximately 300 bp, 3000 bp, and 30000 bp, showing that the location of the peaks is always about 1.5 multiplied by the patch sizes.

FIG. 5. DFA exponent $\alpha(\ell)$ for four yeast chromosomes using (a) the RY rule and (b) the SW rule. We note that the general shape of $\alpha(\ell)$ is very similar for all four chromosomes. In particular, the peaks and valleys (i.e. extrema) are very close to each other, showing that there are similar characteristic patch sizes present in all chromosomes.

FIG. 6. Characteristic patch sizes in *E. coli* sequences, yeast sequences, *C. elegans* sequence, and human sequences estimated using DFA for the SW rule. Only sequences larger than 100000 bp were used. The patch sizes were estimated by locating the peaks in $\alpha(\ell)$ and dividing the position of the peaks by 1.5. Similar patch sizes are found in several sequences, suggesting that the complex global structure of genomic DNA may have some universal characteristics. In eukaryotic sequences the patchiness may be a result of the elaborate organization and folding of DNA by proteins into nucleosomes and higher-order structures of chromatin. Note that the yeast sequences do not show patches on scales from 50-bp to 200bp, possibly due to the absence in yeast of H1 histones which help pack nucleosomes together. The bacterial sequences have a patch size which is absent in the other sequences. The 17 sequences longer than 100000 bp we used are the following: 6 *E. coli* sequences as indicated in the figure, chromosomes III, VI, IX, and XI of *Saccharomyces cerevisiae*, the *C. Elegans* sequence [?,?], the *Homo sapiens* sequences with accession numbers U07000, L29074, HIDSG, L05367, L11910, L36092.

FIG. 7. Histogram giving number of repetitions of n identical nucleotides A or T in the four yeast chromosomes. The coding data (filled circles) follow an exponential distribution, while the noncoding data (open circles) follow a power law—the line represents an artificial control sequence where distribution of tandem simple A and T repeats follows a power law with $\mu = 4.2$

FIG. 8. Double-logarithmic plot of the size histograms of various dimer tandem repeats in the noncoding sequences of primates (a) and invertebrates (b). The straight lines represent least square fits whose slopes give different values of the exponent μ: $\mu = -3.22$ for AA in primates, $\mu = -5.61$ for GC in primates, $\mu = -3.3$ for AA in invertebrates, $\mu = -3.26$ for CA in invertebrates and $\mu = -6.31$ for GC in invertebrates. Note that the histogram for CA repeats in primates has a plateau, and thus cannot be well fitted by a power law. Note also the fast decay of GC repeats in both types of organisms. It is thus relatively unlikely to find long stretches of GC.

FIG. 9. 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. [?].

FIG. 10. 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. [**?**].

FIG. 11. The power spectrum $S_I(f)$ for the interbeat interval increment sequences over ~ 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. [**?**].

FIG. 12. (a) Segment of electrocardiogram showing beat-to-beat (RR_i) intervals. (b) Plot of RR-time series vs. consecutive beat number for a period of 6h ($\approx 2.5 \times 10^4$ beats). Nonstationarity (patchiness) is evident over both long and short time scales. Although these patches clearly differ in the amplitude and frequency of variations, their quantitative characterization remains an open problem. (c) Wavelet transform $T_\psi(RR)$ of the RR-signal in Fig. 12b. Nonstationarities related to constants and linear trends have been filtered. The first derivative of the Gaussian $\psi^{(1)}$ is orthogonal to segments of the time series with approximately constant local average. This results in fluctuations of the wavelet transform values around zero with highest spikes at the positions where a sharp transition occurs. Thus, the larger spikes indicate the boundaries *between* regimes with different local average in the signal, and the smaller fluctuations represent variations of the signal within a given regime. Since $\psi^{(1)}$ is not orthogonal to linear (non-constant) trends, the presence of consecutive linear trends in the RR-intervals will give rise to fluctuations of the wavelet transform values around different nonzero levels corresponding to the slopes of the linear trends. $\psi^{(2)}$ and higher order derivatives can eliminate the influence of linear as well as nonlinear trends in the fluctuations of the wavelet transform values. (d) Instantaneous amplitudes $A(t)$ of the wavelet transform signal in Fig. 12c; $A(t)$ calculated using the Hilbert transform measures the cumulative variations in the interbeat intervals over an interval proportional to the wavelet scale a.

FIG. 13. (a) Probability distributions $P(x)$ of the amplitudes of heart rate variations $x \equiv A(t)$ for a group of 18 healthy adults. Individual differences are reflected in the different average value and widths (standard deviations) of these distributions. All distributions are normalized to unit area. (b) Same probability distributions as in Fig. 13a after rescaling: $P(x)$ by P_{max}, and x by $1/P_{max}$ to preserve the normalization to unit area. The data points collapse onto a single curve. (c) Probability distributions for a group of 16 subjects with obstructive sleep apnea. We note that the second (rightward) peak in the distributions for the sleep apnea subjects corresponds to the transient emergence of characteristic pathologic oscillations in the heart rate associated with periodic breathing [?,?]. (d) Distributions for the apnea group after the same rescaling as in (b). The absence of data collapse demonstrates deviation from the normal heart behavior. We note that direct analysis of interbeat interval histograms does *not* lead to data collapse or separation between the healthy and apnea group. Moreover, we find that the direct application of the Hilbert transform yielding the probability distribution of the instantaneous amplitudes of the original signal does *not* clearly distinguish healthy from abnormal cardiac dynamics. Hence the crucial feature of the wavelet transform is that it extracts dynamical properties hidden in the cumulative variations. We observe for the healthy group good data collapse with a *stable* scaling form for wavelet scales $a = 2$ up to $a = 32$. However, for very small scales ($a = 1, 2$) the average of the rescaled distributions of the apnea group is indistinguishable from the average of the rescaled distributions of the healthy group. Hence very high frequencies are equally present in the signals from both groups. Our analysis yields the most robust results when a is tuned to probe the collective properties of patterns with duration of $\approx \frac{1}{2} - 1$ min in the time series ($a = 8, 10$). The subtle difference between day and night phases is also best seen for this scale range (Fig. 14).

FIG. 14. (a) The solid line is an analytic fit of the rescaled distributions of the beat-to-beat variation amplitudes of the 18 healthy subjects during sleep hours to a stable Gamma distribution with $\nu = 1.4 \pm 0.1$ (note that stable Gamma form has been used previously in the literature to describe other processes—e.g. the spike activity of a single neuron [**?**]). (b) Data for 6h records of RR intervals for the day phase of the same control group of 18 healthy subjects demonstrate similar scaling behavior with a Gamma distribution and $\nu = 1.8 \pm 0.1$, thereby showing that the observed common structure for the healthy heart dynamics is not confined to the nocturnal phase. Semilog plots of the averaged distributions show a systematic deviation — crossover — in the tails of the night-phase distributions, whereas the day-phase distributions follow the exponential form over practically the entire range. Note that the observed crossover for the night phase indicates higher probability of larger variations in the healthy heart dynamics during sleep hours in comparison with the daytime dynamics. We find that the maximum difference between the cumulative distributions of the individual subjects and the Gamma fit in (a) evaluated with the Kolmogorov-Smirnov test can serve as a good index to separate the healthy from the apnea group. Analysis of the first and second moments of the individual distributions also shows clear separation for both groups.