Scaling properties of DNA sequences and heartbeat rate

S. Havlin

Center for Polymer Studies and Department of Physics, Boston University
Boston, MA 02215, USA

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, USA

C.-K. Peng

Center for Polymer Studies and Department of Physics, Boston University
Boston, MA 02215, USA

Cardiovascular Division, Harvard Medical School, Beth Israel Hospital
Boston, MA 02215, USA

A. L. Goldberger

Cardiovascular Division, Harvard Medical School, Beth Israel Hospital
Boston, MA 02215, USA

Department of Biomedical Engineering, Boston University - Boston, MA 02215, USA

1. Long-range power law correlations

In recent years long-range power-law correlations have been discovered in a 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 [11-12]. 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.

© Società Italiana di Fisica

445
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 [13].

\[ C(t) \sim t^{-\alpha} \]

Many systems drive themselves spontaneously toward critical points [2, 14].

In the following sections, we will attempt to summarize some recent findings [15-25] 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 [22], as well as potential practical applications for distinguishing coding and noncoding regions in long nucleotide chains [24].

2. DNA

The role of genomic DNA sequences in coding for protein structure is well known [25]. The human genome contains information for approximately 100,000 different proteins, which define all inheritable features of an individual. The genome 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 molecule, 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 intergenic 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.

3. 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 [13].

In order to study the scale-invariant long-range correlations of a DNA sequence, we introduce a graphical representation of DNA sequences, which we term a fractal landscape or DNA walk [15]. For the conventional one-dimensional random walk model [37, 38], a walker moves either "up" \( \Delta n(i) = +1 \) or "down" \( \Delta n(i) = -1 \) 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 [39-41].

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 Axel [42].

There have also been attempts to map DNA sequence onto multi dimensional DNA walks [16, 43]. However, recent work [34] indicates that the original purine-pyrimidine rule provides the most robust result, 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.

4. Correlations and fluctuations

An important statistical quantity characterizing any walk [37, 38] is the root mean square fluctuation \( \bar{y}(t) \) about the average of the displacement of a quantity \( \Delta y(t) \) defined by \( \Delta y(t) = y(t) - y(0) \), where

\[ \gamma(t) = \sum_{i=1}^{t} \Delta y(i). \]
If there is no characteristic length (i.e., if the correlation between \( u(i) \) and \( u(j) \) are power law long-range correlations), then fluctuations will also be described by a power law

\[
F(l) \sim l^{-\alpha}
\]

with \( \alpha \neq 1/2 \). The case \( \alpha = 1/2 \) represents the absence of long-range correlations.

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 \( F(l) \sim l^{-\alpha} \). A least-squares fit produces a straight line with slope \( \alpha \) substantially larger than 1/2, thus providing direct experimental evidence for the presence of long-range correlations [15].

On the other hand, the dependence of \( F(l) \) 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 \( F(l) \) 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 \( F(l) \) was also studied for DNA sequences by Axel [44].

5. Detrended fluctuation analysis (DFA)

The initial report [15] on long-range (scale-invariant) correlations only in noncoding DNA sequences has generated contradicting responses. Some [16,17,20,21] support our initial finding, while some [17,22,25,27] disagree. However, the conclusions of refs. [18] and [17,22,25,27] are inconsistent with one another in that [17] and [27] doubt the existence of long-range correlations (even in noncoding sequences) while [18] and [22,25] conclude that coding regions display long-range correlations (\( \alpha > 1/2 \)). Prabhhu and Claverie [22] 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. [33] 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 \( F_i(l) \) 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
scaling properties of DNA sequences and heartbeat rate

analysis of DNA sequences with relatively long coding regions, such as those in yeast chromosome III.

Very recently Arneodo et al. [45] studied long-range correlations in DLA sequences using wavelet analysis. The wavelet transform can be made blind to "patchiness" of genomic 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. [35].

Finally, we note that although the scaling exponents \( \alpha \) and \( \beta \) 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.

7. 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 intergenicomic sequence of length \( L = 78,326 \) (GenBank name: HUMHBB). This long noncoding sequence has 50% purines (no overall strand bias) and \( \alpha = 0.7 \) (see fig. 1a). However, from nucleotide \#67,689 to \#74,229, there occurs the LINE-1 region (defined in ref. [46]). In this region of length 6639 base pairs, there is a strong strand bias with 50% purines. In this noncoding sub-region, we find power-law scaling of \( K \) with \( K^{-\beta} \), with \( \alpha = 0.55 \), quite close to that of an uncorrelated random walk.

Even more striking is another region of 6378 base pairs, from nucleotide \#23,137 to \#29,515, which has 50% pyrimidines and is noncorrelated, with remarkably good power-law scaling and correlation exponent \( \alpha = 0.49 \) (fig. 1b). This region actually consists of three subsequences, complementary to shorter parts of the LINE-1 sequence.

These features motivated us to apply a generalized Lévy walk model (see figs. 1c) (b) and (2) for the noncoding regions of DNA sequences [50]. 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 classical Lévy walk model describes a wide variety of diverse phenomena that exhibit long-range correlations [47-51]. The model is defined schematically in fig. 2a:
each of $l_i$ steps in random directions, with a fixed bias probability
\[ p_+ = \frac{1 + \epsilon_j}{2} \]
(5) to go up and
\[ p_- = \frac{1 - \epsilon_j}{2} \]
(6) to go down, where $\epsilon_j$ 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). Figure 2(b) shows such a generalized Lévy walk for the same choice of $l_i$ as in fig. 2(a).

As shown in ref. [30], the generalized Lévy walk—like the pure Lévy walk—gives rise to a landscape with a fluctuation exponent $\alpha$ that depends upon the Lévy walk parameter $\mu$ [41,48],

\[ \alpha = \begin{cases} 1 & \mu \leq 2 \\ 2 - \mu/2 & 2 < \mu < 3 \\ 1/2 & \mu \geq 3 \end{cases} \]

(7)

i.e. the non-trivial behavior of $\alpha$ corresponds to the case $2 < \mu < 3$ where the first moment of $P(l_i)$ 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. (4) that lacks a characteristic length scale. However, for $\mu \geq 3$, the distribution of $P(l_i)$ decays fast enough that an effective characteristic length scale appears. Therefore, the resulting Lévy walk behaves like a normal random walk for $\mu \geq 3$.

8. - 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 intergenic 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 [46,52]. 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 [46]. At least some of the repetitive elements are believed to be remnants of messenger RNA molecules that formerly did code for proteins [52-54]. Alternatively, these segments may represent retroviral sequences that have inserted themselves into the genome [56]. 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 $l^{-\alpha}$ [56, 57]. Theoretical estimates of $\mu$ scale by different methods [57, 58] using a self-avoiding random walk model [56] indicate that the value of $\mu$ for three-dimensional model is between 2.16 and 2.42. Our estimate made by the Rosenbluth Monte-Carlo Method [60] 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 $\alpha$ 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. (4).

Recently the size distribution of insertions and deletions in human and rodent pseudogenes has been studied experimentally by Gu and Li [61]. 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 $10^4 - 10^5$ base pair long. The asymptotic behavior of the scaling exponent $\alpha$ and whether it reaches some universal value for long DNA chains must await further data from the Human Genome Project.

9. Detecting characteristic patch sizes

Scaling methods, such as long-range correlations, may provide important information on the presence in DNA sequences consisting of large patches composed of different nucleotide concentrations and of different length scales. The DFA method allows one to identify the typical length of such elements. We study [92] the local slope of the $F_\alpha(l)$ on the double-logarithmic plot

$$\alpha(l) = \frac{d \log F_\alpha(l)}{d \log (l + 5)}$$

which is not constant but depends on the nucleotide distance $l$. It is hypothesized that

---

**Fig. 3.** (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 $-1/\nu$-type spectra, showing that studying the spectrum alone can be misleading. (b) The exponent $\beta(l)$ is calculated from the local slope of (a). (c) The DFA exponent $\alpha(l)$ for the same model. The exponent $\alpha(l)$ is found by calculating the local slope of the double-log plot of the DFA function (see eq. (8)). The exponents $\alpha(l)$ and $\beta(l)$ have 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.
the maxima of the \( a(l) \) 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. 3 we show an example of the power spectrum as well as the DFA analysis of such a sequence. The sequence in fig. 3 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

![DFA exponent plots for yeast chromosomes using (a) the RY rule and (b) the SW rule.](image)

**Fig. 4.** - DFA exponent \( a(l) \) for all yeast chromosomes using (a) the RY rule and (b) the SW rule. We note that the general slope of \( a(l) \) 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.

![Characteristic patch sizes in E. coli sequences, yeast sequences, C. elegans sequence, and human sequences estimated using DFA for the SW rule.](image)

**Fig. 5.** - 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 \( a(l) \) 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 200 bp, possibly due to the absence in yeast of HI 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, chromosomal II, III, VI, IX, and XI of *Saccharomyces cerevisiae*, the C. Elegans sequence [63, 64], the Homo sapiens sequences with accession numbers U07000, L20074, H1856, L6367, L11910, L36992.
patch size with the highest probability, and the largest patch size with the smallest probability are chosen according to the following rule: For the j-th patch,

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

2) A preliminary length \( n_j \) is computed as \( n_j = 2000^{x_j} \).

3) If \( n_j \) is less than 2000 then a patch of size 2000 bp is chosen. Otherwise if \( n_j \) is less than 20000 then a patch of size 20000 bp is chosen. Otherwise a patch of size 200000 is chosen.

It can be shown analytically that peaks should occur at scales of approximately 1.5 times the patch size. Therefore, by looking for peaks in \( R(t) \) we can estimate characteristic DNA patch sizes embedded in a sequence with an apparent 1/f power spectrum. Different binary mapping rules can detect patches of different nucleotide compositions, i.e., patches with C-G content can be detected by BY or BYM rules. For these purposes, we also apply the power spectra analysis and the wavelet analysis which was recently applied to studies of DNA sequences [15].

Having developed the techniques for detecting and examining characteristic scales of patchiness in model sequences, we apply these methods to real data. Figure 4 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. Figure 5 summarizes characteristic patch sizes in several organisms.

10. Fractal analysis of interbeat intervals

Very recently, the idea of long-range correlations has been extended to the analysis of the heat-to-beat intervals in the normal and diseased heart [28, 65]. The healthy heartbeat is generally thought to be regulated according to the classical principle of homeostasis whereby physiological systems operate to reduce variability and achieve an equilibrium-like state [65]. We find, however, that under normal conditions, heart-to-heart 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 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) [47, 50].

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 s\(^{-1}\), and plot the result, denoted by \( R(t) \), in fig. 6. We observe a more complex pattern of fluctuations for a representative healthy adult (fig. 6a) compared to the "smoother" pattern of interbeat intervals for a subject with severe heart disease (fig. 6b). 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

\[ F(n) = \frac{1}{n} \sum_{u'=(n+1)}^{B_x(u')} \left[ B_x(w') + B_y(w') \right] , \]

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

Figure 7 is a log-log plot of $F(n)$ vs. $n$ for the data in figs. 6(a) and 6(b). This plot is approximately linear over a broad physiologically-relevant time scale (200-4000 beats) implying that

\[ F(n) \sim n^{-\alpha} . \]

We find that the scaling exponent $\alpha$ is markedly different for the healthy and diseased states: for the healthy heartbeat data, $\alpha$ is close to 0, while $\alpha$ 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 $I(u) = B(u + 1) - B(u)$ are uncorrelated for $u > 200$.

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 $R(n)$ is stationary, we can apply standard spectral analysis techniques [2]. Figures 8a and 8b show the power spectra $S_l(f)$, the square of the Fourier transform amplitudes for $R(n)$, derived from the same data sets (without filtering) used in fig. 6. The fact that the log-log plot of $S_l(f)$ vs. $f$ is linear implies

\[ S_l(f) \sim \frac{1}{f^\beta} . \]

The exponent $\beta$ is related to $\alpha$ by $\beta = 2\alpha - 1$ [39]. Furthermore, $\beta$ 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 $\beta = 0$, then $R(t)$ 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") [29].

For the damped data set, we observe a flat spectrum ($\beta = 0$) in the low frequency region (fig. 8b) confirming that $R(t)$ are not correlated over long time scales (low frequencies). Therefore, $R(t)$, the first derivative of $RR(t)$, can be interpreted as being analogous to the velocity of a random walker, which is uncorrelated on long time scales, while $R(t)$ 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. 8a), we obtain $\beta = 1$, indicating non-trivial long range correlations in $RR(t)$—these correlations are not the consequence of summation over random variables or artifacts of non-stationarity. Furthermore, the anti-correlation properties of $RR(t)$ indicated by the negative $\beta$ value are consistent with a negative feedback system (closely coupled) 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 autocorrelations is a fundamental biological variable, namely the inter-beat interval increments.

II. Scaling behavior of heartbeat intervals

Time series of beat-to-beat (RR) heart rate intervals (fig. 9a(a)) obtained from digitized electrocardiograms are known to be nonstationary and exhibit extremely complex behavior[78]. A typical feature of these signals is the presence of `patchy' patterns which change over time (fig. 9b). Heterogeneous properties may be even more strongly emphasized in certain cases of abnormal heart activity. Traditional approaches—such as the power spectrum and correlation analysis—is well suited for such nonstationary (patchy) sequences, and do not carry information stored in the Fourier phase term for determining nonlinear characteristics.

To address these problems, we present an alternative method[73]—"cumulative variation magnitude analysis"—to study the global 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. 9c), because it does not require stationarity and preserves the Fourier phase information. The wavelet transform [71-76] of a time series $s(t)$ is defined as

$$\psi_i(t_n, a) = a^{-1/2} \int s(t) \psi^* \left( \frac{t - t_n}{a} \right) dt,$$

where the analyzing wavelet $\psi$ has a width of the order of the scale $a$ and is centered at $t_n$. For high frequencies (small $a$), the $\psi$ 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 $\Delta t \approx 1$ beat which smooths locally very high-frequency variations and best probes patterns of specific duration ($\approx 1/2 - 1$ min) (fig. 10). The wavelet transform is attractive because it can eliminate local polynomial behavior in the

![Graph](image_url)

Fig. 10. - (a) Probability distributions $P(x)$ of the amplitudes of heart rate variations $x = A(t)$ for a group of 18 healthy adults. Individual differences are reflected in the different average values and widths (standard deviations) of these distributions. All distributions are normalized to unit area.

(b) Same probability distributions as in (a) 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.

![Graph](image_url)

Fig. 10. - (Continued) (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).

nonstationary signal by an appropriate choice of the analyzing wavelet $\psi$ [45]. In our study we use derivatives of the Gaussian function: $\psi^{(n)} = d^n/dt^n e^{-(t/2)^2}$. 
The first derivative of the Gaussian $q^{(1)}$ 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 regions with different local average in the signal, and the smaller fluctuations represent variations of the signal within a given region. Since $q^{(1)}$ 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. $q^{(2)}$ and higher-order derivatives can eliminate the influence of linear as well as nonlinear trends in the fluctuations of the wavelet transform values.

The wavelet transform is thus a cumulative measure of the variations in the heart rate signal over a range 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 amplitude of the wavelet-filtered signal by means of an analytic signal approach $\hat{a}(t)$ which also does not require stationarity. Let $s(t)$ represent an arbitrary signal. The analytic signal, a complex function of time, is defined by $\hat{s}(t) = s(t) + j\, \hat{a}(t)$, where $\hat{a}(t)$ is the Hilbert transform $\mathcal{H}[s(t)]$ of $s(t)$. The instantaneous amplitude $\hat{a}(t)$ and the instantaneous phase of the signal $\phi(t)$ are defined as $\hat{a}(t) = \sqrt{\hat{s}(t)^2 + \hat{a}(t)^2}$ and $\phi(t) = \tan^{-1}(\hat{s}(t)/\hat{a}(t))$.

We study the distribution of the amplitudes of the beat-to-beat variations (Fig. 11) for a group of healthy subjects ($N = 18$, $5$ males, $13$ females; age $20$–$50$, mean $31$) and a group of subjects ($N = 17$) with obstructive sleep apnea ($N = 16$ males; age $22$–$56$, mean $43$). We begin by considering night phase (12 p.m.-6 a.m.) records of interbeat intervals ($\times 10^4$ 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. 10(b)). 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 examine the distributions and find for all healthy subjects that the data conform to a single scaled plot ("data collapse") (Fig. 10(b)). Such behavior is reminiscent of a wide class of well-studied physical systems with universal scaling properties [13, 81]. In contrast, the subjects with sleep apnea show individual probability distributions (Fig. 10(e)) which fail to collapse (Fig. 10(d)).

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 a Hilbert transform yields 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.

![Fig. 11.](image)

(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 $r = 1.4 \pm 0.1$ (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 [82]). (b) Data for 6 h 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 $r = 1.8 \pm 0.1$, 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 right-side 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.
We observe for the healthy group good data collapse with a stable scaling form for wavelet scales $a \geq 2$ up to $a = 61$. However, for very small scales ($a = 1, 2$) the average of the rescaled distributions of the atheroma 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 $\approx 1.2 \text{ min}$ in the time series ($a = 8, 10$). The subtle difference between day and night phases is also best seen for this scale range ($a < 6$ fig. 11).

We next analyze the distributions of the beat-to-beat variation amplitudes. For the healthy group, we find that these are well fit by the gamma form $P(x) \propto (x/x_0)^{b-1} e^{-x/x_0}$, where $b = z_0$, $P(x) \propto (x/x_0)^{b-1} e^{-x/x_0}$, where $b = z_0$, $x_0$ is the position of the peak $P = P_{\text{max}}$ (fig. 11a). 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 $u = x x_0$. 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 $P(u) \approx P(x)/x_0 b$. Thus, it is sufficient to specify only one parameter $b$ in order to characterize the heterogeneous heartbeat variabilities of each subject in the group.

We also analyzed heart rate dynamics for the healthy subjects during day-time hours (mean 6.4 pm). Our results indicate that the observed, apparently universal behavior is stable in a wide regime of time scales and holds not only for the night phase but for the day phase as well (fig. 11b).

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 [81, 82]. However, these properties, detected by Fourier and fluctuation analysis techniques, ignore information related 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 findings suggest that for healthy individuals, there may be a common structure in this nonlinear phase interaction. The scaling property cannot be accounted for by activity, since we analyzed data from subjects during nocturnal hours. Moreover, if 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 atheroma, a common and important instability of cardiovascular control, show a dramatic alteration in the scaling pattern—possibly related to pathological mode leading associated with periodic breathing dynamics [86]. 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, -red world- physiological signals.

REFERENCES

Physics investigation of financial markets

R. N. Mantegna

Istituto Nazionale di Fisica della Materia, Unità di Palermo
and Dipartimento di Energetica ed Applicazioni di Fisica, Università di Palermo
Palermo, I-90125, Italy

H. E. Stanley

Center for Polymer Studies and Department of Physics, Boston University
Boston, MA 02215, USA

"È importante, quindi, che i principi della meccanica quantistica abbiano portato a riconoscere (...) il carattere statistico delle leggi relative del processo elementare. Questa considerazione ha reso sostanzialmente l'analisi fra fisica e scienze sociali, due le quali è visibile un'identità di valore e di metodo."

— Ettore Majorana [1]

1. Introduction

During the last thirty years, physicists have achieved important results in the fields of phase transitions, statistical mechanics, nonlinear dynamics, disordered and self-organized systems. New paradigms have been developed and a range of complex systems have been carefully investigated and described. This description has sometime also been performed in the presence of noise or quenched randomness. With this, relatively recent, background the study of economics systems performed by physicists may produce results relevant for both physics and economics.

Economic systems, strictly regulated and very frequently monitored, are ideal for a study performed using tools and paradigms developed to describe physical systems. Due to strict regulation, such systems suffer only slightly from various modifications of the rules underlying the process during the time window of the investigation of the process. Moreover, due to continuous monitoring the amount of data describing the phenomenon is usually sufficient for a detailed statistical analysis.