

Fractal landscapes in biological systems: Long-range correlations in DNA and interbeat heart intervals

H.E. Stanley^a, S.V. Buldyrev^a, A.L. Goldberger^b,
J.M. Hausdorff^b, S. Havlin^{a,c}, J. Mietus^b, C.-K. Peng^a,
F. Sciortino^a and M. Simons^{b,d}

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

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

^c*Department of Physics, Bar-Ilan University, Ramat-Gan 52100, Israel*

^d*Biology Department, MIT, Cambridge, MA 02139, USA*

Here we discuss recent advances in applying ideas of fractals and disordered systems to two topics of biological interest, both topics having in common the appearance of scale-free phenomena, i.e., correlations that have no characteristic length scale, typically exhibited by physical systems near a critical point and dynamical systems far from equilibrium. (i) DNA nucleotide sequences have traditionally been analyzed using models which incorporate the possibility of *short-range* nucleotide correlations. We found, instead, a remarkably *long-range* power law correlation. We found such long-range correlations in intron-containing genes and in non-transcribed regulatory DNA sequences as well as intragenomic DNA, but not in cDNA sequences or intron-less genes. We also found that the myosin heavy chain family gene evolution increases the fractal complexity of the DNA landscapes, consistent with the intron-late hypothesis of gene evolution. (ii) The healthy heartbeat is traditionally 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 found, however, that under normal conditions, beat-to-beat fluctuations in heart rate display long-range power law correlations.

1. Introduction

The purpose of this paper is to describe some recent progress in fractals and disordered systems [1] that overlap two topics from biology. The first of these concerns a discovery about the difference in the long-range correlations between base pairs in the coding and non-coding parts of DNA [2,3]. The second concerns long-range correlations in the intervals between heartbeats [4]. For both examples, we analyzed the data by producing contours reminiscent of irregular “fractal” (self-affine) landscapes that are widely studied in

physical systems [1]. Both examples are found to be similar, in that their fractal landscapes display long-range correlations of the sort recently studied in various abstract physical models of porous media [5]. This work is being carried out by a remarkable team of interdisciplinary workers, and I am pleased that they have consented to join me as co-authors.

2. DNA

DNA nucleotide sequences have been previously analyzed using models – such as an n -step Markov chain – which incorporate the possibility of *short-range* nucleotide correlations [6]. To study the possibility of long-range correlations, we first introduced a novel method for studying the stochastic properties of nucleotide sequences by constructing a 1:1 map of the nucleotide sequence onto a walk – which we term a DNA walk [2]. We used this mapping to provide a quantitative measure of the correlation between nucleotides over long distances along the DNA chain. We discovered in the nucleotide sequence a remarkably *long-range* power law correlation that is significant because it implies a new scale invariant property of DNA [2]. Most intriguingly, we found such long-range correlations in intron-containing genes and in non-transcribed regulatory DNA sequences, but not in cDNA sequences or intronless genes (see also the independent work of ref. [7]).

In the 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. The DNA walk is defined by the rule that the walker steps up ($u(i) = +1$) if a pyrimidine occurs at position a linear distance i along the DNA chain, while the walker steps down ($u(i) = -1$) if a purine occurs at position i . For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated 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. The question we ask is 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).

This DNA walk provides a novel graphical representation for each gene and permits the degree of correlation in the nucleotide sequence to be directly visualized, by constructing a one-dimensional landscape. The “altitude” at each coordinate point l is given by the “net displacement” $y(l)$ of the walker after l steps, which is the sum of the unit steps $u(i)$ for each step i , $y(l) \equiv \sum_{i=1}^l u(i) - l\bar{u}$, where \bar{u} is the average value of u over the entire sequence.

An important statistical quantity characterizing any landscape (or surface) is the root mean square fluctuation $F(l)$ about the average altitude. A calculation of $F(l)$ can distinguish three possible types of behavior.

(i) If the nucleotide sequence were random, then $F(l) \sim l^{1/2}$ – as expected for a *normal* random walk or any system where the correlation function decreases exponentially – cf. eq. (8.23) of ref. [8].

(ii) If there were a local correlation extending up to a characteristic range (such as in Markov chains), then *the asymptotic behavior* $F(l) \sim l^{1/2}$ *would be unchanged from the purely random case.*

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

$$F(l) \sim l^\alpha \quad (\alpha \neq 1/2). \quad (1)$$

We made double-logarithmic plots of the mean square fluctuation $F(l)$ as a function of the linear distance l along the DNA chain for representative genomic and cDNA sequences across the phylogenetic spectrum [2]. In addition, we analyzed other sequences encoding a variety of other proteins as well as regulatory DNA sequences – 84 in total. We discovered that remarkably long-range correlations ($\alpha > 1/2$) are characteristic of intron-containing genes and non-transcribed genomic regulatory elements [2]. In contrast, for cDNA sequences and genes without introns, we find that $\alpha \cong 1/2$ indicating the absence of long-range correlation. Thus, the calculation of $F(l)$ for the DNA walk representation provides a new, *quantitative* method to distinguish genes with multiple introns from intron-less genes and cDNAs based solely on their statistical properties. The finding of long-range correlations in intron-containing genes appears to be independent of the particular gene or the encoded protein – it is observed in genes as disparate as myosin heavy chain, beta globin and adenovirus. The functional (and structural) role of introns remains uncertain, and although our discovery does not resolve the “intron-late” vs. “intron-early” controversy about gene evolution [9], it does reveal intriguing fractal properties of genome organization that need to be accounted for by any such theory.

More recently, Buldyrev et al. [3] studied the evolution of the myosin heavy chain (MHC) gene family using the same technique (figs. 1–3). They found that α seems to increase with evolution (fig. 4). The monotonic increase of α with evolution does not support a recent study of Voss [10]. This apparent discrepancy is due to the fact that ref. [10]: (i) did not analyze single gene families with evolution, (ii) did not distinguish intron-containing vs intron-less sequences, and (iii) did not account for “strand bias” (unequal number of purines and pyrimidines over some regions, see fig. 2).

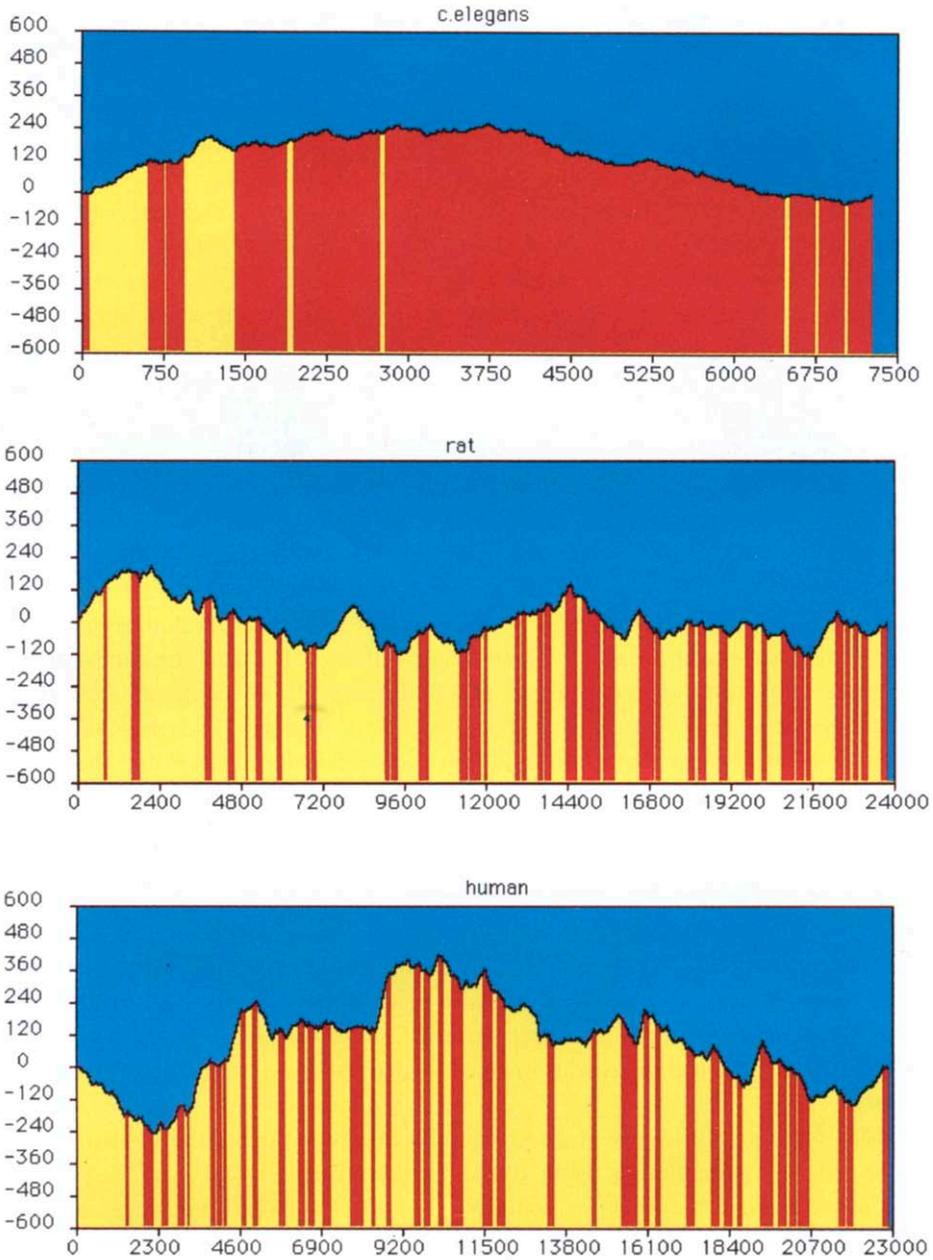


Fig. 1. Fractal landscape representations of three members of the MHC family: worm (*C. elegans*), rat (*Ratus norvegicus*), and human (*Homo sapiens*). The red bars denote coding regions (“exons”), while the yellow denotes the non-coding regions (“introns”). As the number of introns increase, the average exon length decreases and the average intron length increases.

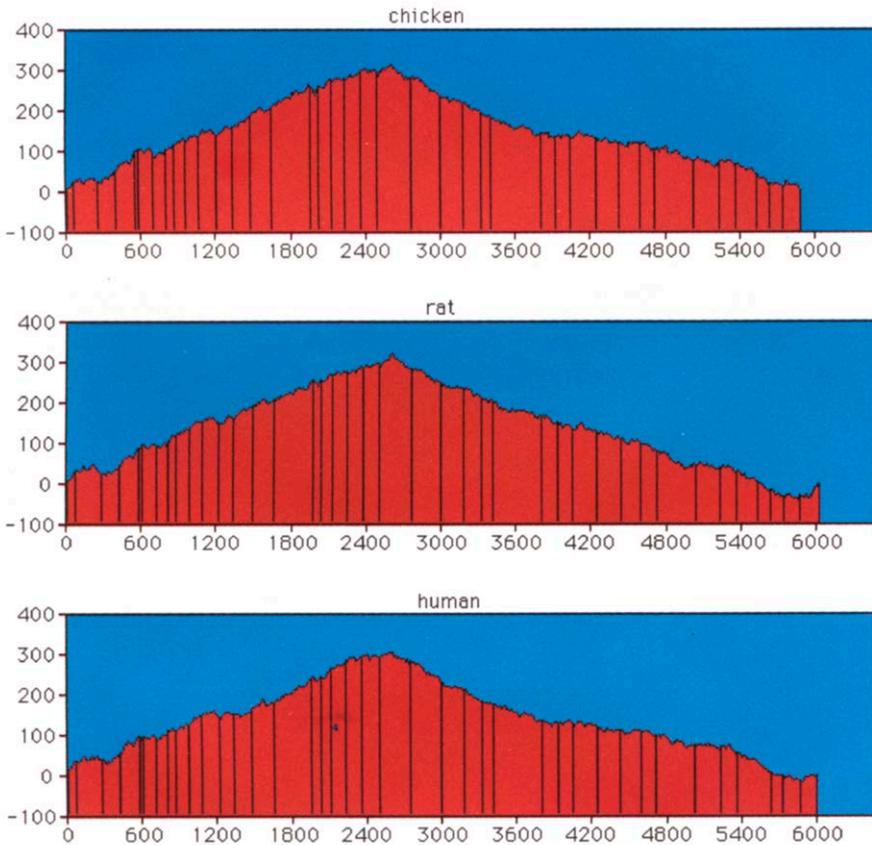


Fig. 2. Fractal landscape formed by stitching together the coding regions (“exons”) in chicken (*Gallus gallus*), rat, and human. Note that the total exon size remains approximately constant, and even the general shape of the “exon landscape” is similar.

Nee [11] has recently raised the possibility that exon regions correspond to uncorrelated *biased* random walks, while intron regions correspond to uncorrelated *unbiased* random walks. Since exons and introns interdigitate, the overall behavior could appear to be a correlated random walk. To rule out this possibility, we studied genes with the exons removed (“spliced together introns”). If Nee’s hypothesis were correct, we would have observed $\alpha = 0.5$, as for an uncorrelated unbiased random walk. Instead, we observed roughly the same value of α as in the full gene (fig. 5a) [11]. A more detailed comparison of the two cases is obtained by plotting the “local slope” (fig. 5b). For example: for the rat MHC gene (X04267) $\alpha = 0.64$ for the entire sequence, and $\alpha = 0.67$ for the “stitched together” intron sequence.

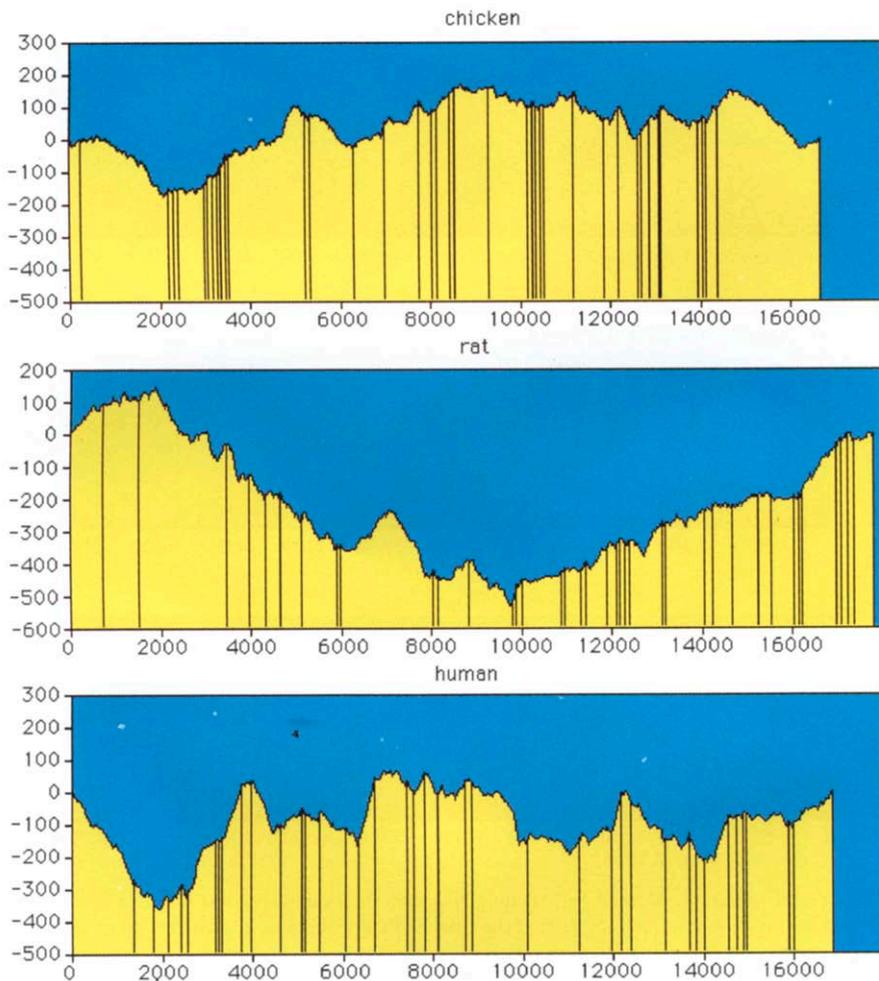


Fig. 3. Fractal landscape formed by stitching together the non-coding regions (“introns”) in chicken, rat, and human. Note that the general features of the intron landscape differ from one organism to another.

To understand possible evolutionary mechanisms [12] that increase the complexity of the fractal landscape of genomic DNA sequences, we also developed a model that utilizes intron insertion into pre-existing coding sequences [3]. The model is based on the hypotheses that originally the information was encoded in an mRNA molecule that later was converted into a DNA sequence, and that this sequence underwent modifications due to mutagenesis and insertion of non-coding genetic material (introns). Iterating this model produces sequences of increasing fractal complexity fully analogous

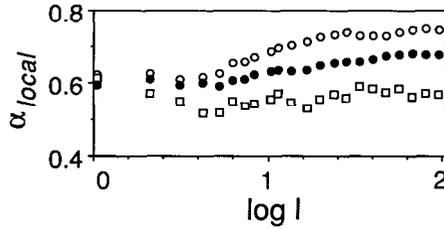


Fig. 4. Dependence on l of the local slopes for three genes in ascending order of evolution: worm (\square), rat (\bullet), and human (\circ). Each local slope, α_{local} , is the slope of each pair of successive data points in a log-log plot of the fluctuation $F(l)$ vs. l . Note that the long-range correlation parameter α is increasing monotonically with evolution.

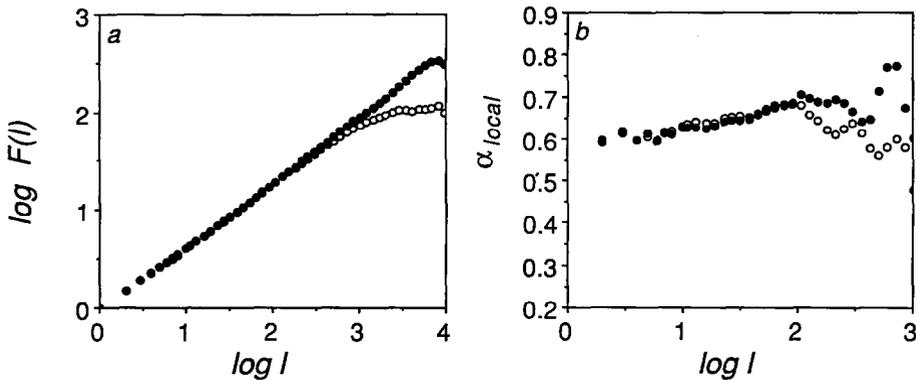


Fig. 5. (a) Double-logarithmic plots of the mean square fluctuation function $F(l)$ as a function of the linear distance l along the DNA chain for the rat embryonic skeletal myosin heavy chain gene (\circ) and its "intron-spliced sequence" (\bullet). (b) The corresponding local slopes, α_{local} , based on pairs of successive data points of part (a). We see that the values of α are roughly constant. For this specific gene, the sequence with exons removed has an even broader scaling regime than the DNA sequence of the entire gene, indicated by the fact that part (a) is linear up to 10 000 nucleotides.

to a Lévy walk (consisting of patches of different bias with lengths distributed according to a power law).

3. Interbeat heart intervals

Very recently, the idea of long-range correlations has been extended to the analysis of the beat-to-beat intervals in the heart [4]. 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 [13]. We found [4], 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 [14] – in fact, we found evidence that such power-law correlations extend over thousands of heartbeats in healthy subjects.

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 [15]. Fourier analysis of lengthy heart rate data sets from healthy individuals typically reveals a $1/f$ -like spectrum for lower frequencies (<0.01 Hz), but the long-term correlation properties of heart rate time series remain unstudied [16].

Our work was based on the analysis of the beat-to-beat heart rate fluctuations recorded for long time intervals (up to 24 hours) with an ambulatory electrocardiographic (Holter) monitor [4]. Fig. 6, the interbeat interval time series $B(n)$ of a representative healthy adult, reveals a complex type of variability. Such variations could correspond to three possible cases: (i)

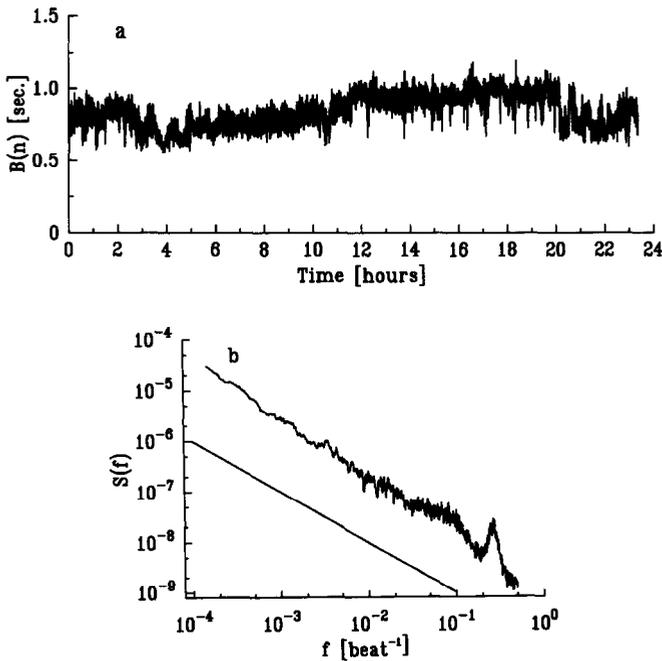


Fig. 6. (a) $B(n)$, the beat-to-beat time interval between beat n and beat $n + 1$, for a period of 24 hours in a healthy adult. Note that the trend apparent in (a) corresponds to the fact that $B(n)$ is non-stationary. (b) The Fourier transform $S(f)$ of $B(n)$. The best-fit line of this double-logarithmic plot for over 3 decades has slope $\beta = -1.07$. The solid line has slope -1 . The data in (b) are smoothed by averaging over 50 values. The peak at about 10^{-1} corresponds to physiologic heart rate fluctuations associated with breathing (respiratory sinus arrhythmia).

random fluctuations in which the duration of each interbeat interval is entirely uncorrelated from the previous one, (ii) *short-range correlations* (e.g., those due to respiration) in which the duration of any interbeat interval is correlated only with previous beats “locally” and the correlation decays rapidly (exponentially), or (iii) *long-range correlations* that decay very slowly (as a power-law function of time), extending over hundreds to thousands of beats. The difference between (ii) and (iii) is not only in the range of correlations but in their different functional form: a characteristic time scale exists in (ii) but not in (iii).

The question we address, therefore, is which type of correlation behavior the normal heartbeat displays. A standard time series measurement is the autocorrelation function $C(t) \equiv \langle (B(n) - \bar{B})(B(n+t) - \bar{B}) \rangle$, which measures the statistical influence of how one interbeat interval deviated from the average will affect the deviation of other interbeat intervals after a time lag t . Alternatively, we can compute the Fourier power spectrum $S(f)$. If the time series is stationary, i.e., the average and the variance are independent of the sampling length of the time series, $S(f)$ is simply the Fourier transform of $C(t)$ [17]. For a random uncorrelated process (*case (i)*), $C(t)$ must fall to zero after one beat so $S(f)$ is flat for a broad range of frequencies (white noise). For a time series with only short-range correlations (*case (ii)*), $C(t)$ must decay exponentially so $S(f)$ will be flat or almost constant for the low frequency region and fall off for the high frequencies. For long-range correlations (*case (iii)*), both $C(t)$ and $S(f)$ must have a power-law behavior, so a double-logarithmic plot must be linear.

A requirement for calculating $C(t)$ is that the time series be stationary. Visual inspection of the trends in fig. 6a reveals that the healthy interbeat intervals do not meet this criterion. As a first step toward obviating the difficulties arising from nonstationarity due to circadian rhythms, endocrine cycles and other factors, we compute $S(f)$ to obtain some general idea of what the correlation may look like. Fig. 6b is a log-log plot of $S(f)$ for the healthy heartbeat time series of fig. 6a. The best fit line in fig. 6b has slope -1.07 and corresponds to the $1/f$ spectrum consistent with previous reports [16].

In principle, applying the Fourier transform method assumes that the time series under study is much longer than the largest period of the sequence, so that longer sequences contain no additional information. However, this is not the case for long-range correlated time series with power law behavior,

$$S(f) \sim f^\beta \quad (\beta < 0), \quad (2a)$$

because the diverging Fourier components for the low frequencies indicate that the longest oscillatory component is always comparable to the length of the

entire sequence under study. In fact, the time series is always non-stationary when $\beta \leq -1$ since if we integrate the power spectrum from some finite frequency to zero, the total power will diverge, implying no well-defined average values over long times. It is necessary, therefore, to test whether the long-range correlation behavior suggested by the power law scaling in fig. 6a is simply an artifact of a non-stationary time series with finite length. To this end, we measure the difference between two consecutive interbeat time intervals and define the *interbeat increment function* ($I(n) \equiv B(n+1) - B(n)$). We find that this new function, shown in fig. 7a, is stationary by calculating the means and variances of several subsequences of $I(n)$. Since the increment is equivalent to the *first derivative* of the time series, it follows that if $S(f)$ is not affected by non-stationarity, then $S_1(f)$, the power spectrum of the increment series, will behave as

$$S_1(f) \sim f^{\beta_1} \quad (\beta_1 = \beta + 2). \quad (2b)$$

The extra f^2 factor for the increment series arises from the fact that the Fourier

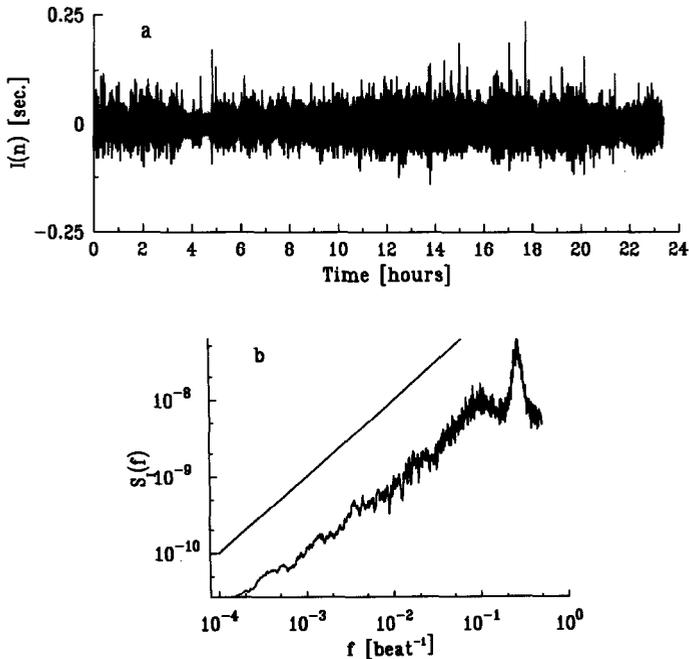


Fig. 7. (a) The increment function $I(n) \equiv B(n+1) - B(n)$. Note that the non-stationarity is no longer present. (b) The power spectrum $S_1(f)$ for the increment sequence. The best-fit line has a slope $\beta_1 = 0.93$, while the solid line has slope $+1$ (cf. eq. (2b)). The data in (b) are smoothed by averaging over 50 values.

transform of a derivative will yield an extra factor of f for the amplitude of the Fourier components, and the power spectrum is the square of the amplitudes. Fig. 7b is a double-logarithmic plot of $S_1(f)$, the power spectrum for the increment data of fig. 7a. The slope gives $\beta_1 = 0.93$, consistent with eq. (2b) since $\beta = -1.07$ for the original time series in fig. 6a. Thus this measurement provides an important consistency check that the scaling exponent β is in fact due to long-range correlations (extending over 10^4 heartbeats), and not to non-stationarity. For a group of 12 adult subjects without clinical evidence of heart disease (age range: 20–61 years, mean 34), we found that $\beta = -1.02 \pm 0.16$ and $\beta_1 = 0.98 \pm 0.17$ [4]. Furthermore, we obtained similar results when we divided the time series into three subsets (approximately 8 hour periods) and repeated the above analysis; thus our findings are not simply related to the different levels of daily activities.

Our demonstration of long range correlations without a characteristic time scale in healthy heart rate dynamics is consistent with observation of fractal properties of other physiologic processes and anatomic structures that do not have characteristic scale of time or length [16,18–20]. This fractal-like behavior may be adaptive for at least two reasons. The long-range correlations serve as an organizing principle for highly complex, non-linear processes that may generate fluctuations on time scales ranging from less than milliseconds to greater than months. At the same time, the lack of a characteristic scale helps prevent excessive *mode-locking* that would restrict the functional responsiveness of the organism. Finally, we note that the scale-invariant properties of the healthy heartbeat indicate that the control mechanisms are operating far from equilibrium [14], in violation of the classical concept of primarily “homestatic” regulation [13].

Acknowledgements

We wish to thank C. DeLisi, L. Liebovitch, R.D. Rosenberg, and M. Schwartz for valuable discussions. Partial support was provided to ALG by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute and the National Aeronautics and Space Administration, to MS by the American Heart Association, and to HES by the NSF and ONR.

Note added in proof

Very recently, Munson and Michaels [21] have reported long-range correlations extending up to almost 10^5 nucleotides in an entire chromosome, yeast

chromosome III, consisting of many genes as well as about 50% non-coding regions between the genes. Munson and Michaels report $\alpha = 0.7$, well above the value $\alpha = 0.5$ which would be found if there were no long-range correlations.

References

- [1] A. Bunde and S. Havlin, eds., *Fractals and Disordered Systems* (Springer, Berlin, 1991).
- [2] C.K. Peng, S.V. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, *Nature* 356 (1992) 168.
- [3] S.V. Buldyrev, C.K. Peng, A. Goldberger, S. Havlin, H.E. Stanley and M. Simons, *Fractal landscapes and molecular evolution: analysis of myosin heavy chain genes*, preprint.
- [4] C.K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H.E. Stanley and A.L. Goldberger, *Long-range anti-correlations and non-Gaussian behavior of the heartbeat*, preprint.
- [5] S. Havlin, R. Selinger, M. Schwartz, H.E. Stanley and A. Bunde, *Phys. Rev. Lett.* 61 (1988) 1438; R.B. Selinger, S. Havlin, F. Leyvraz, M. Schwartz and H.E. Stanley, *Phys. Rev. A* 40 (1989) 6755; C.K. Peng, S. Havlin, M. Schwartz and H.E. Stanley, *Phys. Rev. A* 44 (1991) 2239; S. Prakash, S. Havlin, M. Schwartz and H.E. Stanley, *Phys. Rev. A* 46 (1992) R-1724; S.V. Buldyrev, S. Havlin, J. Kertész, H.E. Stanley, and T. Vicsek, *Phys. Rev. A* 43 (1991) 7113.
- [6] S. Tavaré and B.W. Giddings, in: *Mathematical Methods for DNA Sequences*, M.S. Waterman, ed. (CRC Press, Boca Raton, 1989) p. 117.
- [7] W. Li and K. Kaneko, *Europhys. Lett.* 17 (1992) 655; W. Li, *Int. J. Bifurcation Chaos* 2 (1992) 1.
- [8] H.E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford Univ. Press, Oxford, New York, 1971).
- [9] W. Gilbert, *Nature* 271 (1978) 501; J.E. Darnell, Jr., *Science* 202 (1978) 1257; W.F. Doolittle, *Nature* 272 (1978) 581.
- [10] R. Voss, *Phys. Rev. Lett.* 68 (1992) 3805.
- [11] S. Nee, *Nature* 357 (1992) 450; C.K. Peng, S.V. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, *Nature*, in press; P.J. Munson, R.C. Taylor and G.S. Michaels, *Statistically significant long-range DNA correlations over an entire chromosome*, preprint.
- [12] W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1991); J.D. Watson, N.H. Hopkins, J.W. Roberts, J.A. Steiz and A.M. Weiner, *Molecular Biology of the Gene* (Benjamin/Cummings, Menlo Park, CA, 1987).
- [13] W.B. Cannon, *Physiol. Rev.* 9 (1929) 399.
- [14] P. Bak, C. Tang and K. Wiesenfeld, *Phys. Rev. Lett.* 59 (1987) 381.
- [15] A.L. Goldberger, D.R. Rigney and B.J. West, *Sci. Am.* 262 (1990) 42.
- [16] A.L. Goldberger and B.J. West, *Yale J. Biol. Med.* 60 (1987) 421, and references therein.
- [17] A.M. Yaglom, *Correlation Theory of Stationary and Related Random Functions*, vols. I, II (Springer, New York, 1987).
- [18] B.J. West, V. Bhargava and A.L. Goldberger, *J. Appl. Physiol.* 60 (1986) 1089.
- [19] L. Liebovitch, in: *Advanced Methods of Physiological Systems Modeling*, vol. II, V.A. Marmarelis, ed. (Plenum, New York) in press.
- [20] F. Caserta, H.E. Stanley, W. Eldred, G. Daccord, R.E. Hausman and J. Nittman, *Phys. Rev. Lett.* 64 (1990) 95.
- [21] P.J. Munson and G.S. Michaels, *Nature*, submitted.