Evolution of Long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences

Richard F. Voss

IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

(Received 18 February 1992)

A new method of quantifying correlations in symbolic sequences is applied to DNA nucleotides. Spectral density measurements of individual base positions demonstrate the ubiquity of low-frequency $1/f^{\beta}$ noise and long-range fractal correlations as well as prominent short-range periodicities. Ensemble averages over classifications in the GenBank databank (primate, invertebrate, plant, etc.) show systematic changes in spectral exponent β with evolutionary category.

PACS numbers: 87.10.+e, 05.40.+j, 06.50.−x, 72.70.+m

A wide variety of physical systems show power-law correlations in space (fractals) [1-3] or time ($1/f$ noises) [4,5]. The characterization and understanding of these correlations has become an important and surprisingly difficult problem. Although diffusion limited aggregation [2] and self-organizing criticality [6] are based on simple physical models, a detailed prediction of their power-law correlations has proven illusive. The measurements reported here demonstrate that similar power laws have evolved in DNA.

Numerous attempts have been made to characterize and graphically portray the genetic information stored in DNA nucleotide sequences [7-12]. One popular method maps the four individual bases to steps in a variant of a random walk. Unfortunately, when the number of spatial dimensions is less than 4, the mapping itself introduces correlations between the bases. Nevertheless, recent results [11] on a few sequences have shown long-range fractal or scaling correlations. Here, a new method of characterizing arbitrary symbolic sequences is described that does not introduce correlations between symbols. When applied to DNA sequences it confirms the ubiquity of low-frequency $1/f^{\beta}$ noise and the corresponding long-range fractal correlations as well as prominent short-range periodicities. Moreover, averages over $>5\times10^7$ bases from >25000 sequences in 10 classifications (primate, invertebrate, plant, etc.) of the GenBank [13] databank show systematic changes in correlations and spectral exponent β with evolutionary category.

There are a number of accepted techniques [14] for characterizing random processes $x(t)$. Both the correlation function $C_x(\tau)=\langle x(t)x(t+\tau)\rangle$ and the spectral density $S_x(f)\propto|\int x(t)e^{-2\pi ift}dt|^2/\Delta f$ quantify correlations at the time scale $\tau\approx1/f$. Generally, $S_x(f)$ and $C_x(\tau)$ are connected by the Wiener-Khintchine relations [14], $S_x(f)\propto\int C_x(\tau)\cos(2\pi\tau)d\tau$ and $C_x(\tau)\propto\int S_x(f)\times\cos(2\pi f)df$. Fast Fourier transform (FFT) algorithms allow efficient direct computation of $S_x(f)$ from sample sequences and the estimation of $C_x(\tau)$ from $S_x(f)$.

A true random process or *white noise* $w(t)$ has no correlations in time. $C_w(\tau)\propto\delta(\tau)$ and $S_w(f)\propto$ const. Its integral $X(t)=\int w(t)dt$ produces a *Brownian motion* or *random walk* with $S_X(f)\propto1/f^2$. The average distance traveled in a time T is $\Delta X(T)=\langle|X(t+T)-X(t)|^2\rangle^{1/2}\propto T^{1/2}$. Like $w(t)$, the future changes of $X(t)$ are independent of its past.

Many naturally occurring fluctuations, however, from electronic voltages and time standards to meteorological, biological, traffic, economic, and musical quantities [1,4,5,15-18] have nontrivial correlations that extend over all measured time scales. These $1/f$ *noises* exhibit $S(f)\propto1/f^{\beta}$ with $\beta\approx1$ over many decades. In most cases, the physical reason for this long-range power-law behavior remains a mystery.

The most effective mathematical model for such scaling noises has been *fractional Brownian motion* (fBm) [19,20] as the integral of *fractional Gaussian noises* $x_H(t)$. A fBm process $X_H(t)=\int x_H(t)dt$ is specified by $0<H<1$ such that

$$\Delta X_H(T)=\langle|X_H(t+T)-X_H(t)|^2\rangle^{1/2}\propto T^H,$$

and a corresponding

$$S_x(f)\propto1/f^{\beta}$$

where $\beta=2H-1$ for the increments $x_H(t)$. Thus, as $H\to0$ the fBm $X_H(t)\to1/f$ noise and $x_H(t)\to f$ noise. As $H\to1$, $X_H(t)\to1/f^3$ noise and $x_H(t)\to1/f$ noise. $H=\frac{1}{2}$ gives *normal* Brownian motion.

A definition for the product $x(t)x(t+\tau)$ is central to $C(\tau)$, $S(f)$, and $\Delta X(T)$. For non-numeric sequences $x_1,\ldots,x_N$, composed of K allowed symbols care must be taken to define procedures that do not make *a priori* assumptions about relations between the symbols. Assigning a number to each symbol, or a walk displacement [8-12] along a direction in a D-dimensional space with $D<K$, *introduces* numeric correlations.

Such problems may be avoided with *equal-symbol multiplication* defined as $x_nx_m=1$ if $x_n=x_m$, and 0 otherwise. A binary indicator function or projection operator U_k selects the elements of x_n that are equal to symbol k. $U_k[x_n]=1$ if $x_n=k$, and 0 otherwise. Consequently, $x_nx_m=\sum_kU_k[x_n]U_k[x_m]$ and

$$C(\tau)=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}U_k[x_n]U_k[x_{n+\tau}]=\sum_{k=1}^{K}C_k(\tau).$$

Here $C_k(\tau)$, the equal-symbol correlation function for symbol k, is the probability that $x_{n+\tau}=k$ if $x_n=k$. Similarly,

$$S(f)=\sum_{k=1}^{K} S_k(f)$$

where $S_k(f)$ is the equal-symbol spectral density of symbol k.

Thus, as symbolic sequence may be decomposed into the K binary sequences $U_k[x_n]$ which identify the positions of symbol k. Fourier transforms of $U_k[x_n]$ yield the frequency components, $U_k(f)$, the individual $S_k(f) \propto |U_k(f)|^2$, and $C_k(\tau)$, as well as $S_\Sigma(f)=\sum S_k(f)$ and $C_\Sigma(\tau)=\sum C_k(\tau)$. Cross-correlation spectra, $S_{j,k}(f) \propto U_j(f)U_k(f)$ between different symbols j and k may also be calculated.

This method is illustrated with the 229 354 bases in the complete genome [21] of the human Cytomegalovirus strain AD 169 (GenBank [13] accession number X17403). Figure 1 shows the four indicator functions $U_k[x_n]$ (k=A, C, G, T) at different levels of averaging. This plot graphically reveals many positive and negative correlations between different bases and illustrates the difficulties with low-D walks. For some averaging scales and positions (e.g., around 160000) the pyrimidines (C, T) are strongly anticorrelated, around 100000 the correlation is near 0 while near 115000 the correlation is positive. At all scales, however, the $U_k[x_n]$ are extremely irregular and there is little change as the averaging is increased from 10 to 100 and 1000 sites.

This *independence of scale* is confirmed by the measured equal-symbol $S_k(f)$ shown at the top of Fig. 2. The sequence was divided into independent sections of size $N_{FFT}=2^{16}$ for FFT transforms and the results from individual sections were averaged together. Although there are small differences, each of the $S_k(f)$ has the same overall behavior. At large f (small τ) $S_k(f)$ is dominated by the white noise of an uncorrelated random process. Within the white noise there is a sharp peak at $f=\frac{1}{3}$ (period $\tau=3$) that is probably related to the nucleotide triplet (or *codon*) that specifies one of 20 amino acids [7]. At low f (large τ), however, $S_k(f)$ shows power-law scaling similar to $1/f$ noise. For comparison and as a verification of the method, a superposition of the

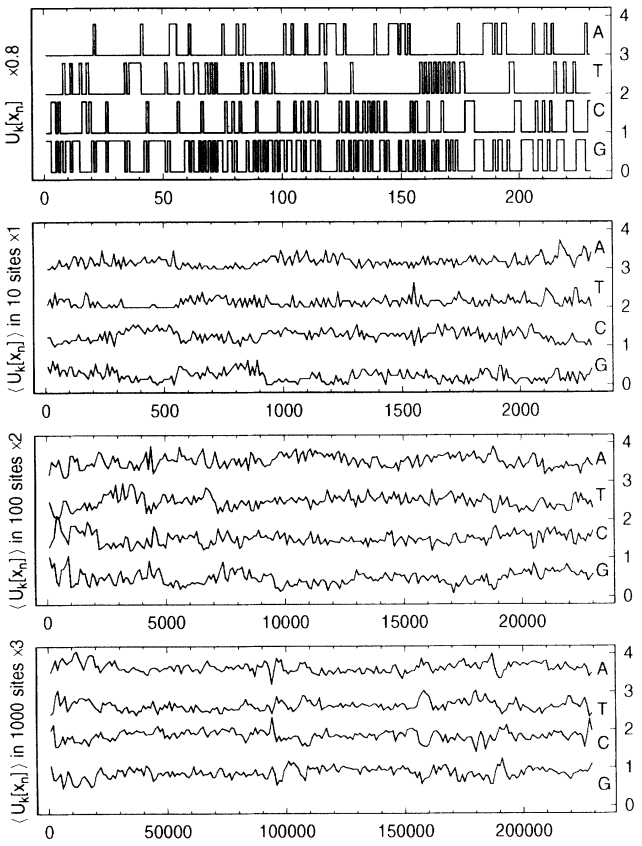

FIG. 1. The indicator functions $U_k[x_n]$ (k=A, C, G, T) for the 229 354 bases in the complete genome of the human Cytomegalovirus strain AD169. Successive curves show averages of $U_k[x_n]$ over 10, 100, and 1000 sites.

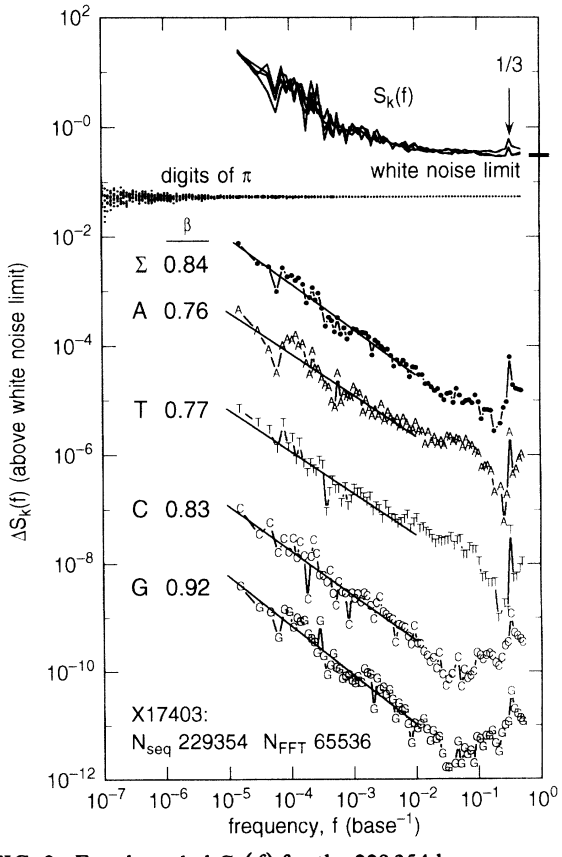

FIG. 2. Equal-symbol $S_k(f)$ for the 229 354 base sequence in Fig. 1. $S_k(f)$ for the first 1.13×10^9 decimal digits of π is shown for comparison. $\Delta S(f)=S(f)-S(\infty)$ reveals the long-range $1/f^\beta$ dependence. Least-squares estimates of β are shown as solid lines and spectra are offset for clarity.

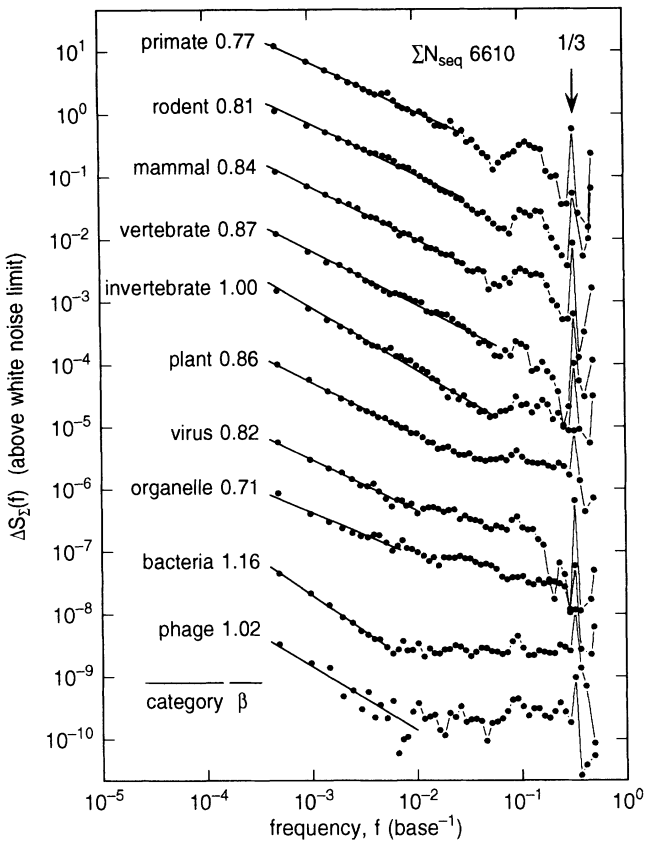

FIG. 3. Equal-symbol $\Delta S_\Sigma(f)$ for categories of DNA sequences from the GenBank Release 68 databank. Least-squares estimates of β are tabulated and the resulting fits are shown as solid lines. $\Delta S_\Sigma(f)$ offset for clarity.

10 $S_k(f)$ for the first 1.13×10^9 decimal digits of π [22] is also shown. This symbolic sequence shows no long-range correlation.

In many cases $1/f^\beta$ noises can be studied over larger ranges by subtracting the high f white-noise limit [23] $\Delta S(f) = S(f) - S(\infty)$ under the assumption that the $1/f^\beta$ and white noises are independent. The subtraction is also illustrated in Fig. 2, where the $\Delta S(f) \propto 1/f^\beta$ scaling clearly extends from around 100 bases to 100000 bases. Differences in the exponents β_k are visible. For this sample, A is most correlated with T and C with G at low f. A similar separation of the short-range and scaling components is much more difficult with $\Delta X(T)$ analysis [11].

The limited accuracy of estimating β from individual sequences can be greatly reduced by averaging over the large number of DNA sequences in the GenBank databank. The GenBank classification (mammal, invertebrate, plant, etc.), moreover, allows examination of statistical changes with evolutionary category.

Figure 3 shows log-log plots of the equal-symbol $\Delta S_\Sigma(f) = \sum \Delta S_k(f)$ averaged over specific categories in the GenBank Release 68 database with $N_{FFT} = 2048$. Sequences $\geq 2N_{FFT}$ contributed multiple samples. The average included the 6610 sequences $> N_{FFT}$ and significantly reduced the scatter about $1/f^\beta$.

The category average also clarifies the small *period* (high f) variations. Figure 4 shows $S_\Sigma(1/f)$ versus period $\tau = 1/f$ for $N_{FFT} = 512$ to emphasize short-range structure. The large narrow codon peak at *period* 3 is the most prominent feature, but the increased averaging allows other small structures to be visible. For example, the peak at *period* 9 for *primates*, *vertebrates*, and *inver-*

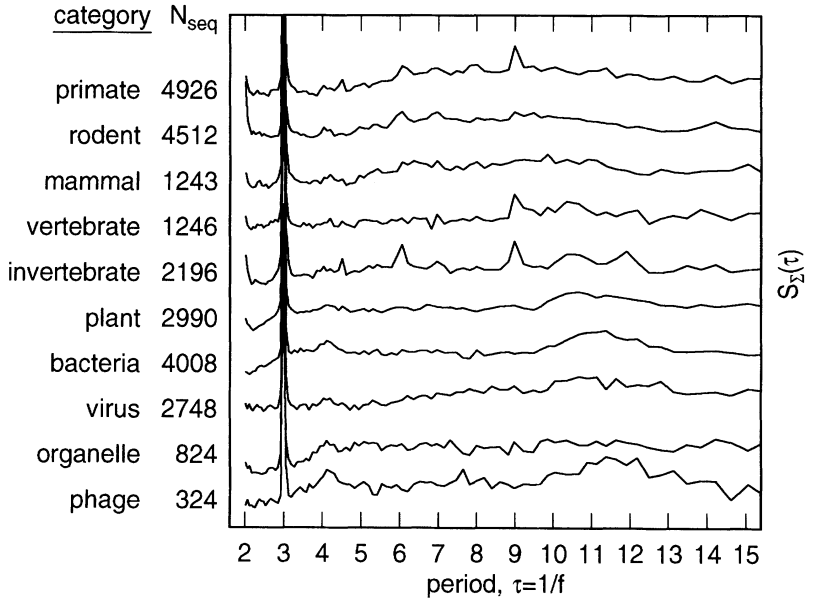

FIG. 4. Equal-symbol $S_\Sigma(f) = \sum S_k(f)$ vs period $\tau = 1/f$ for specific categories of DNA base sequences from the GenBank Release 68 database emphasizes short-range correlations. In this linear plot successive $S_\Sigma(f)$ are offset for clarity and N_{seq} is tabulated for each category.

tebrates is absent in the other categories.

Systematic changes in β are clearly seen in the category averages of Fig. 3. $S(f) \propto 1/f^{\beta}$ with $\beta \approx 1$ is the signature of fractal (scaling) correlations that extend to the size of the largest sequences (of order 100000 bases). *Bacteria* and *phage* show the smallest range of scaling. For the remaining categories, there is a systematic β increase with evolutionary status from 0.64 for *organelle* to the 1.00 (exact $1/f$ noise) of *invertebrates* followed by a decrease to 0.77 for *primates*. For $0 < \beta < 1$, increasing β increases the recovery probability from a DNA error but decreases the information content per unit length.

Earlier measurements of pitch and loudness fluctuations in music and speech [16,17,20] related $S(f) \propto 1/f$ to human communication. A white noise represents the maximum rate of information transfer, but a $1/f$ noise, with its scale-independent correlations, seems to offer the best compromise between efficient information transfer and immunity to errors on all scales.

Although the measurements presented here are quite striking, they represent only a beginning to the fractal analysis of the expanding DNA sequence databank. Equal-symbol $S_k(f)$ and category averages provide a new technique for quantifying evolutionary changes in the information content of DNA. Cross-spectra and graphic techniques are being developed to *highlight* specific correlations along the sequence.

The author wishes to thank Professor Avy Goldberger for exposure to the general problem of fractal analysis of DNA sequences and for initial samples for testing.

[1] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman and Co., New York, 1982).

[2] *Fractals in Physics*, edited by L. Pietronero and E. Toscatti (North-Holland, Amsterdam, 1986).

[3] A. Aharony and J. Feder, Physica (Amsterdam) **38D**, 1 (1989).

[4] M. B. Weissman, Rev. Mod. Phys. **60**, 537 (1988).

[5] R. F. Voss, in *Proceedings of the Thirty-Third Frequency Control Symposium*, Atlantic City (Electronics Industries Assoc., Washington, DC, 1979).

[6] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987).

[7] P. Friedland and L. H. Kedes, Commun. ACM **28**, 1164 (1985).

[8] E. Hamori and J. Ruskin, J. Biol. Chem. **258**, 1318 (1983).

[9] E. Hamori, Nature (London) **314**, 585 (1985).

[10] B. D. Silverman and R. Linsker, J. Theor. Biol. **118**, 295 (1986).

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

[12] H. J. Jeffry, Nucl. Acids Res. **18**, 2163 (1990).

[13] H. S. Bilofsky and C. Burkes, Nucl. Acids Res. **16**, 1861 (1988).

[14] See, for example, F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, New York, 1965), Chap. 15; or F. N. H. Robinson, *Noise and Fluctuations* (Clarendon, Oxford, 1974).

[15] B. B. Mandelbrot and J. R. Wallis, Water Resour. Res. **5**, 321 (1969).

[16] R. F. Voss and J. Clarke, Nature (London) **258**, 317 (1975).

[17] R. F. Voss and J. Clarke, J. Accous. Soc. Am. **63**, 258 (1978).

[18] R. F. Voss, in Proceedings of the Workshop on Fractals and Computer Graphics, Darmstadt, Germany, 1991, edited by G. Sakas (Springer-Verlag, Berlin, to be published).

[19] B. B. Mandelbrot and J. W. van Ness, SIAM Rev. **10**, 422 (1968).

[20] R. F. Voss, in *The Science of Fractal Images*, edited by H.-O. Peitgen and D. Saupe (Springer-Verlag, New York, 1988).

[21] M. S. Chee *et al.*, Curr. Top. Microbiol. Immunol. **154**, 125 (1990).

[22] First 1.13×10^9 decimal digits of π courtesy of D. Chudnovsky, Columbia University, New York, NY 10027.

[23] R. F. Voss and J. Clarke, Phys. Rev. B **13**, 556 (1976).