EUROPHYSICS LETTERS 7 February 1992

Europhys. Lett., 17 (7), pp. 655-660 (1992) *published in January 1992*

Long-Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence.

W. LI(*)(§) and K. KANEKO(**)

(*) *Santa Fe Institute - 1660 Old Pecos Trail, Suite A, Santa Fe, NM 87501, USA*
(**) *Department of Pure and Applied Sciences, University of Tokyo*
Komaba, Meguro, Tokyo 153, Japan

(received 6 August 1991; accepted in final form 25 November 1991)

PACS. 87.10 – General, theoretical, and mathematical biophysics.
PACS. 02.50 – Probability theory, stochastic processes, and statistics.

Abstract. – Mutual information function, which is an alternative to correlation function for symbolic sequences, and a «symbolic spectrum» are calculated for a human DNA sequence containing mostly intron segments, those that do not code for proteins. It is observed that the mutual information function of this sequence decays very slowly, and the correlation length is extremely long (at least 800 bases). The symbolic spectrum of the sequence at very low frequencies can be approximated by $1/f^{\alpha}$, where f is the frequency and α ranges from 0.5 to 0.85. It is suggested that the existence of the repetitive patterns in the sequence is mainly responsible for the observed long-range correlation. A possible connection between this long-range correlation and those in music notes is also briefly discussed.

There are many systems which are considered to be complex because they have structures at different length scales. The existence of structures at very large scales results in long-range correlations. These long-range correlations can be detected by examining the two-point correlation function to see whether it decays slower than an exponential function, or, whether it reaches the zero value at a very large distance (though the two-point correlation function is not the only, and in some cases, the best way to measure the long-range correlation, due to the fact that it does not take into account the structures in between the two points, as well as that it does not measure the correlation between two units larger than a point).

If the two-point correlation function decays as a power law, we have a «scaling phenomenon». The power spectrum $P(f)$, which is the Fourier transformation of the correlation function, will also be a power law function: $P(f) \sim 1/f^{\alpha}$, where f is the frequency and α is the scaling exponent. If the two-point correlation function decays even slower than a power law, such as the case of logarithmic function, the power spectrum is then exactly inversely proportional to the frequency, *i.e.* $P(f) \sim 1/f$, or $\alpha = 1$. For these two cases, we have an interesting connection with the $1/f^{\alpha}$ noise—time series with $1/f^{\alpha}$ power spec-

(§) Address after September 1, 1991: Physics Department, Rockefeller University, 1230 York Avenue, New York, NY 10021.

tra—which are quite common in nature [1]. The only difference is that it is the spatial power spectrum instead of the temporal spectrum that is calculated.

One of the simplest ways to generate spatial long-range correlations is by elongation, or expansion of the space. In particular, expansion with a small amount of error leads to spatial scaling and $1/f^\alpha$ spectra [2]. Another very simple way to generate long-range correlations is by repetition of the same structure with the repeated structures being separated by some arbitrarily long distances. Though this mechanism sounds trivial, we will see that it is actually very important for our discussion.

The basic fact that the lengths of the present-day nucleotide sequences (including DNA and RNA sequences) are much longer than those of the prebiotic sequences indicates that elongation played a role in the evolution of nucleotide sequences. In fact, gene duplication is such an essential feature of life, that the elongation has been accomplished by the duplication of the original sequence, and then the copied sequence is added to the original sequence. The duplication is typically not perfect, but with a small amount of error. The repetitive structure in DNA sequences is very common, and has been experimentally observed in many different biological systems [3]. It is even suggested that the elongation of sequences by gene duplication is an important mechanism for evolution [4].

The similarity between the model of spatial $1/f^\alpha$ spectra with elongation and error [2], on the one hand, and the basic evolutionary mechanism of nucleotide sequences with duplication and error, on the other, lead us to suspect that there might exist long-range correlations and perhaps $1/f^\alpha$ spectra in DNA sequences.

Our previous search for long-range correlations and spatial $1/f^\alpha$ spectra in DNA sequences was mainly carried out in protein-coding sequences (called *exons*), and the result was negative [5, 6]. Our observation is consistent with the known results that models with short-range memories such as the Markov chain can approximate the coding sequences quite well, and that even random sequences are not a bad approximation of the coding sequences after all [7]. It seems that in a typical exon sequence the correlation between two nucleotides (in DNA sequences, nucleotides consist of adenine A, cytosine C, guanine G, and thymine T) decays to zero at around only 10 bases according to the study of some sample sequences [6].

The absence of long-range correlation in exon sequences does not imply that there is no long-range correlation in DNA sequences in general. DNA sequences consist of genes and segments between genes («junk genes»). Genes consist of protein-coding segments and noncoding segments (called *introns*). It is not clear what the typical statistical properties are for junk genes, due to the lack of the data in DNA sequence database. Nevertheless, it is well known that intron segments have quite different statistical features as compared with those of exon segments [8].

Promisingly, we have found that the correlation decays more slowly in intron segments than exon segments, with a typical correlation length of 20 bases in the sample sequences studied [6]. These correlation lengths, though longer, are not long enough to increase the low-frequency spectra, and certainly not enough to produce a $1/f^\alpha$ spectra. It is the purpose of this paper to point out that there exist intron sequences whose correlation lengths are much longer than other intron sequences, and whose low-frequency power spectra behave as $1/f^\alpha$, with α between 0.5 and 1.

The sequence to be analyzed is the human-blood coagulation factor VII gene [9]. This sequence is given the name HUMCFVII in the GenBank (for information about the GenBank, see [10]). The sequence contains 12850 bases. The exons are located at sites 522-585, 1654-1719, 4294-4454, 6383-6407, 6478-6591, 8307-8447, 9419-9528, 10124-10247, and 11064-11659 (total 1401 bases), and introns are located at sites 586-1653, 1720-4293, 4455-6382, 6408-6477, 6592-8306, 8448-9418, 9529-10123, and 10248-11063 (total 9737 bases). Most of the sequence are introns (76 percent, as compared with the 11 percent for exons).

The first statistical quantity we are calculating is one similar to the two-point correlation function and applicable to symbolic sequences. The natural choice is the two-symbol mutual information function [11, 12], which is defined as

$$M(d) \equiv \sum_{\alpha,\beta=A,C,G,T} P_{\alpha\beta}(d) \log_2 \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta}, \tag{1}$$

where P_α is the density for symbol α and $P_{\alpha\beta}(d)$ is the joint probability for the two-symbol pair: symbol α and symbol β which is a distance d away from the symbol α. Both P's and $P_{\alpha\beta}$'s are determined by taking the statistics along the sequence.

The $M(d)$ for the sequence HUMCFVII is shown in fig. 1 (in log-log scale). Since P_α's and $P_{\alpha\beta}(d)$'s are calculated by counting statistics on the sequence, the finite sequence length introduces a fluctuation to the mutual information. It is known that this fluctuation generally overestimates the value of $M(d)$ [12]. In order to see this, in fig. 1, we also plot the $M(d)$ of a random sequence with the same sequence length and same composition of A, C, G, T as in HUMCFVII (this random sequence is also called a *scrambled sequence*).

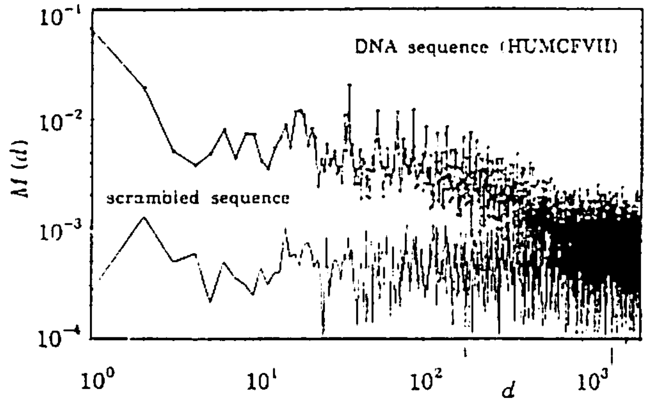

Fig. 1. – Mutual information function $M(d)$ of the sequence HUMCFVII (in log-log scale). The $M(d)$ of the scrambled sequence is also shown.

As can be seen in fig. 1, $M(d)$ is not a straight line in the log-log plot, so it is not a power law function. Nevertheless, it decays to small value at extremely large distances, somewhere around $d \approx 800$. Beyond this distance, the values of mutual information are still slightly larger than those of the scrambled sequence. It should be pointed out that a correlation length of the order of $d_c \approx 800$ is considerably larger than that of a typical intron sequence, which is about $d_c \approx 20$ [6]. Also note that there are peaks at distances $d = 17i$ ($i = 1, 2, \ldots$) clearly visible, due to the repetition of the pattern with period 17 between site ≈ 2050 and ≈ 2820 (this pattern is CCCGGGGGCGTGGGTGT).

Besides the mutual information function, we would also like to calculate a statistical quantity similar to the power spectrum *and* applicable to symbolic sequences (the standard definition of power spectrum is only defined for numerical sequences). One such «symbolic spectrum» is proposed by Silverman and Linsker [13]. In this approach, each symbol is represented by a vertex in the $(n-1)$-simplex, where n is the total number of symbols (for example, for nucleotide sequences, $n = 4$, and we have four vertices on the 3-simplex, *i.e.* the tetrathedron), and a symbolic sequence becomes a vector sequence (for nucleotide sequences, this vector sequence contains three component sequences). The spectrum for symbol sequences proposed by Silverman and Linsker is simply the sum of the power spectrum of each component sequence:

$$P(f) \equiv \sum_{c=1}^{3} \frac{1}{N} \left| \sum_{j=0}^{N-1} x(j)_c \exp[-i2\pi(f/N)j] \right|^2 \qquad (f = 0, \ldots, N-1), \tag{2}$$

where $x(j)_c$ is the c'th component of the vector pointing to the vertex which represents the symbol located at site j. We use the following vector representation of the four nucleotides in which the three components for each symbol are

symbol:	x-component	y-component	z-component	
A:	$\{1,$	$0,$	$0\}$	
C:	$\{-0.333333,$	$0,$	$0.942809\}$	(3)
G:	$\{-0.333333,$	$-0.816497,$	$-0.471405\}$	
T:	$\{-0.333333,$	$0.816497,$	$-0.471405\}.$	

It is suggested that the symbolic spectrum defined by eq. (0.2) is invariant with respect to different projections of the tetrahedron to the three axes as well as different labeling of the vertex with the four symbols [13].

Figure 2*a*) shows the symbolic spectrum (in log-log scale) defined by eq. (0.2) for the first 8192 ($= 2^{13}$) bases of the sequence HUMCFVII (it represents 64 percent of the whole sequence). The dots represent the spectral component, and the solid line represents the average of two neighboring components. The unit of frequency f is 1 cycle/8192 sites. A peak at $f = 481$, for example, indicates a cycle with periodicity 17. Besides the almost flat spectrum at high frequencies, the increase of the power at low frequencies is discernible. By fitting the solid line from $f = 1$ to $f = 100$ (that corresponds to the length scale from 81.92 to 8192 bases) by a power law function $1/f^\alpha$, we get $\alpha \approx 0.84$. We will call this a *partial* $1/f^\alpha$ *spectrum* to distinguish it from the true $1/f^\alpha$ spectrum which can be fit by a power law function for all spectral components, not just the low-frequency components.

The first 8192 bases cover mostly intron segments. As we move further upstream of the sequence, due to the nonstationarity of the sequence, the spectrum is actually changed. Figure 2*b*) shows the symbolic spectrum of the 8192 bases from site 2873 to 11064. This sub-

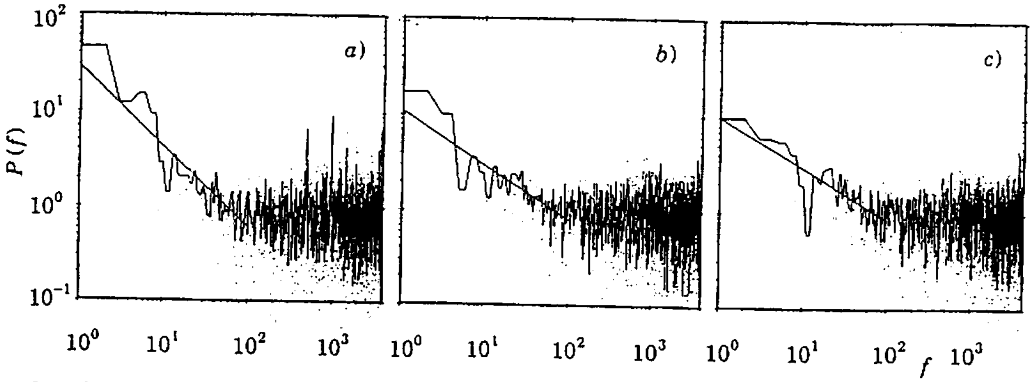

Fig. 2. – Symbolic spectra $P(f)$ of the sequence HUMCFVII (in log-log scale). Also given are the scaling exponents α for the first 100 spectral components (out of total 4096 components) in $P(f) \sim 1/f^\alpha$. *a*) For the first 8129 bases (from site 1 to site 8129), with $\alpha \approx 0.84$; *b*) for some 8129 bases in the middle of the sequence (from site 2873 to site 11064), with $\alpha \approx 0.57$; *c*) for the last 8129 bases (from site 4659 to site 12850), with $\alpha \approx 0.53$.

sequence does not contain the repetitive patterns with the periodicity 17. If the low-frequency spectrum is approximated by the $1/f^\alpha$, the $\alpha \approx 0.57$ from $f = 1$ to $f = 100$.

Figure 2*c*) shows the symbolic spectrum of the last 8192 bases of the sequence (from site 4659 to 12850). This segment contains an exon from site 11064 to 11659 (with 596 bases). If we again try to approximate the low-frequency spectrum by $1/f^\alpha$, then $\alpha \approx 0.53$. Comparing fig. 2*b*) and 2*c*) with fig. 2*a*), we suggest that the highly repetitive patterns in the sequence help to increase the correlation at longer distances, and consequently increase the exponent α in the $1/f^\alpha$ spectrum from around 0.5 to around 0.8.

Since different chunks of the sequence have different α values, the sequence as a whole is nonstationary. We then calculated the «average spectrum» of the whole sequence as shown in fig. 3. The sequence is partitioned into 12 pieces, each with 1024 bases. The symbolic spectrum is calculated for each subsequence, and fig. 3 is the average of these spectra. The fitting of the first 50 spectral components (out of total 512 spectral components) gives $\alpha \approx 0.54$. We have also calculated the average spectrum by partitioning the sequence into 6 pieces, each containing 2048 bases. The spectrum (not shown here) is almost the same with fig. 3.

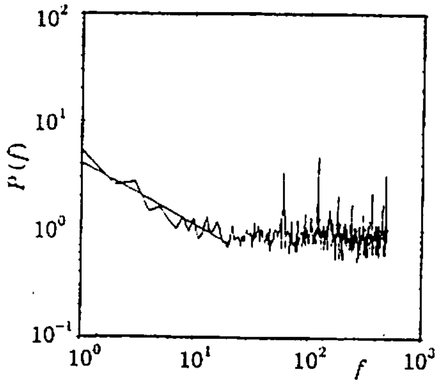

Fig. 3. – The symbolic spectrum (in log-log scale) averaging over 12 sub-sequences, each of them containing 1024 bases. The best-fit line, fitting the first 50 spectral components (out of total 512 components) gives $\alpha \approx 0.57$, or $P(f) \sim 1/f^{0.57}$.

The result presented in this paper firmly establishes that there exist long-range correlations in DNA sequences. These long-range correlations are due to the existence of repetitive structures. If there are many different repetitive structures with different lengths, we might have a scaling phenomenon. It is now believed that many repetitive structures in the DNA sequences do not have any biological functions, and it explains why these structures are better preserved in the intron sequences that do not code for proteins, instead of exon sequences [4].

There is a possible connection between the long-range correlation in DNA sequences reported here and the long-range correlation observed in music, in particular, the classical music. The time series taken from the music (loudness and pitch series) are shown to exhibit $1/f^\alpha$ power spectra [14]. Although it has not been checked, we believe that the long-range correlation in musical time series results from some long-range correlation in the music notes. An interesting perspective for the $1/f^\alpha$ noise in music, the partial $1/f^\alpha$ spectrum in the DNA sequence, and the spatial $1/f^\alpha$ spectra generated by the expansion-modification systems proposed by one of the authors [2] is that all of them have some copy-with-error mechanisms (for example, the repetition of the music theme and its variation in music notes) [15]. It will be interesting to see whether some models for music composition with the copy-and-error mechanism can generate music notes that can be played to have a $1/f^{\alpha}$ power spectrum.

* * *

WL would like to thank the financial support from the Center for Nonlinear Studies of Los Alamos National Laboratory during the summer of 1989 when the ideas presented in this paper were first discussed, and Aspen Center for Physics where fruitful interaction with other participants of the Recognizing Gene Workshop (in June 1990) lead to this paper. The work at Santa Fe Institute is funded by grants from MacArthur Foundation, National Science Foundation (PHY-87-14918), and the Department of Energy (DE-FG05-88ER25054).

REFERENCES

[1] *E.g.*, PRESS W. H., *Comm. Astron.*, 7 (1978) 103.

[2] LI W., *Europhys. Lett.*, 10 (1989) 395; *Phys. Rev. A*, 43 (1991) 5240.

[3] BRITTEN R. and KOHNE D., *Science*, 161 (1968) 529; *Sci. Am.*, 222 (1970) 24; LONG E. and DAWID I., *Ann. Rev. Biochem.*, 49 (1980) 727; JELINEK W. and SCHMID C., *Ann. Rev. Biochem.*, 51 (1982) 813.

[4] OHNO S., *Evolution of Gene Duplication* (Springer-Verlag) 1970; *The Origin and Evolution of Life* (University of Tokyo Press) 1988 (in Japanese).

[5] LI W. and KANEKO K., unpublished results (1989).

[6] LI W., *Generating non-trivial long-range correlations and 1/f spectra by replication and mutation*, SFI-91-002 (Santa Fe Institute) 1991, submitted.

[7] *E.g.*, TAVARÈ S. and GIDDINGS B. W., *Some statistical aspects of the primary structure of nucleotide sequences*, in *Mathematical Methods for DNA Sequences*, edited by M. S. WATERMAN (CRC Press) 1989.

[8] FICKETT J. W., *Nucleic Acids Res.*, 10 (1982) 5303; STORMO G. D., *Identifying coding sequences*, in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*, edited by M. J. BISHOP and C. J. RAWLINGS (IRL Press) 1987.

[9] O'HARA P. J. *et al.*, *Nucleotide sequence of the gene coding for human factor VII, a vitamin K-dependent protein participating in blood coagulation*, in *Proc. Nat. Acad. Sci. USA*, 84 (1987) 5158; O'HARA P. J. and GRANT F. J., *Gene*, 66 (1988) 147.

[10] BURKS C. *et al.*, *Methods Enzymol.*, 183 (1989) 3; BURKS C. *et al.*, *Nucleic Acids Res.*, 10 (1991) 2221; CINKOSKY M. J. *et al.*, *Science*, 252 (1991) 1273.

[11] SHANNON C. E., *Bell Syst. Techn. J.*, 27 (1948) 379; SHAW R., *The Dripping Faucet as a Model Chaotic System* (Aerial Press) 1984.

[12] LI W., *J. Stat. Phys.*, 60 (1990) 823.

[13] SILVERMAN B. and LINSKER R., *J. Theor. Biol.*, 118 (1986) 295.

[14] VOSS R. and CLARKE J., *Nature*, 258 (1975) 317; *J. Acous. Soc. Amer.*, 63 (1978) 258; MUSHA T., *The World of Fluctuation* (Kodansha) 1980 (in Japanese).

[15] OHNO S. and OHNO M., *Immunogenetics*, 24 (1986) 71.