NON-STATIONARY ANALYSIS OF DNA SEQUENCES

Nidhal Bouaynaya[*] *and Dan Schonfeld*

University of Illinois at Chicago, Dept. of Electrical and Computer Engineering

ABSTRACT

Previous searches for long-range correlations in DNA sequences was carried out using statistical tools for stationary signals. However, genomic signals are non-stationary as can be attested by standard statistical tests for stationarity. In this paper, we address, in the light of non-stationary time-series analysis, the questions of (i) the existence of long-range correlations in DNA sequences and (ii) whether they are present in both coding and non-coding segments or only in the latter. It turns out that the statistical differences between coding and non-coding segments are more subtle than previously claimed by the stationary analysis. Both coding and non-coding sequences exhibit long-range correlations, as asserted by an evolutionary $1/f$ spectrum (i.e., having a time-dependent spectral exponent). Moreover, the average spectral exponent of non-coding segments is higher than its counterpart for coding segments. To prove that this observation is not an artifact of the $1/f$ evolutionary spectrum, we show, using an index of randomness that we derive from the frequency-time distribution of the genomic signals, that coding sequences are "more random" (i.e., whiter) than non-coding sequences. We believe that this result is likely the source of confusion and controversy in previous work, which relied on stationary analysis of DNA correlations.

Index Terms— Non-stationary time-series analysis; evolutionary spectrum; evolutionary periodogram; Hilbert transform; empirical mode decomposition (EMD).

1. INTRODUCTION

One intriguing characteristic of the DNA of eukaryotic [1] organisms is its mosaic organization into coding sequences, called exons, interspaced by non-coding sequences, called introns. Understanding the statistical characterization of the exon-intron structure of eukaryotic genes not only help biologists discriminate between coding and non-coding sequences, but also provide valuable clues about the evolution of genes, the constraints which lead this evolution, and can be an important first step towards understanding gene-related diseases like cancer and Alzheimer disease.

In 1992, Peng et al. [1] uncovered the existence of long-range power-law correlations in the DNA of eukaryotic organisms by constructing a map of nucleotide sequences onto a walk, $u(i)$, which they termed a "DNA walk." The DNA walk is defined by the rule that the walker steps up ($u(i) = +1$) (resp., down ($u(i) = -1$)) if a pyrimidine (resp., purine) resides at position i. This long-range statistical correlation in DNA sequences means that nucleotides at any given position appeared to be related to nucleotides thousands of bases away. Moreover, they found such long-range correlations in non-coding sequences (intron-containing genes and non-transcribed regulatory sequences), but not in coding sequences (complementary DNA sequences (cDNA) or intron-less genes). The latter coding sequences appeared to be similar to white noise or exhibit at most short-range correlations (like Markov processes). Similar observations were reported independently by Li et al. in [2], who applied standard Fourier analysis to a sample of genes. What is surprising in their findings is not just the existence of long-range correlation, but also the particular form of the correlation structure: the $1/f$-like spectra [2]. This prompted a sequence of controversial papers, some affirming [3] and others disputing either the existence of long range correlations in DNA sequences or the statistical difference between coding and non-coding segments [4]. These contradictory results relied on different numerical representations of the genomic sequences [3], but all of them used standard statistical and signal processing tools for stationary time-series analysis: the root mean square fluctuation, $F(l)$, which is related to the autocorrelation function through a summation [1], the autocorrelation function, $C(l)$, and the power spectrum, $S(f)$. These quantities can distinguish between two or three types of behavior:

1. For white noise, we have $C(l) \sim \delta(l)$, $F(l) \sim l^{1/2}$ and $S(f) \sim 1$.
2. If the sequence exhibits short-range correlations, such as a Markov memory, then $C(l) \sim \exp^{-l}$, $F(l) \sim l^{1/2}$, and $S(f) \sim 1/f^2$.
3. If the sequence exhibits long-range correlations, then

[*] Nidhal Bouaynaya is currently in the Department of Systems Engineering at the University of Arkansas at Little Rock.

[1] A eukaryote is a single-celled or multicellular organism whose cells contain a distinct membrane-bound nucleus. Animals, plants, fungi, and protists are eukaryotes.

[2] We say that a process has a $1/f$ spectrum if its power spectrum is of the form $\frac{\sigma^2}{f^\alpha}$, for some $\sigma > 0$ and $\alpha > 0$ is the spectral exponent.

$C(l) \sim l^{-\gamma}$ ($\gamma > 0$), $F(l) \sim l^{-\beta}$ ($\beta \neq \frac{1}{2}$), and $S(f) \sim 1/f^{\alpha}$ ($0 < \alpha < 2$).

Karlin and Brendel [5] first questioned the implicit stationarity assumption in the analysis of Peng et al. [1], which was based on the root mean square fluctuation $F(l)$. They argued that the assumption of stochastic stationarity is absurd in view of the great degree of local and global heterogeneity in nucleotide sequences. Peng et al. retaliated by proposing the Detrended Fluctuation Analysis (DFA) technique [6]. The DFA method constructs a stationary process from the non-stationary genomic signal by dividing the entire sequence into subsequences and subtracting the trend in each subsequence. The DFA technique became a well-established method for detecting long-range correlations in DNA and other natural and man-made sequences [7]. However, the DFA is limited to the very special case of non-stationary signals consisting of stationary signals with embedded trends, i.e.,

$$X(t) = c(t) + X_0(t), \tag{1}$$

where $c(t)$ is a deterministic function and $X_0(t)$ is a stationary process. Observe that $c(t)$ can be locally approximated by a polynomial function. So, by dividing the sequence into (overlapping or non-overlapping) subsequences, estimating the linear trend in each subsequence and substracting it, we obtain the underlying stationary signal $X_0(t)$. Chen et al. [8] investigated the effects of three other types of non-stationarities (signals with segments removed, signals with random spikes and signals with different local behavior) on the DFA method. Our extensive simulations and analysis of nucleotide sequences found that DNA sequences exhibit different forms of non-stationarities that are more complex than embedded trends. Therefore, any pursuit to solve the controversy about the nature of genomic correlations should consider techniques for wider classes of non-stationary signals.

The goal of this paper is to address, in the light of non-stationary time-series analysis, the problems of (i) the existence of long-range correlations in eukaryotic DNA sequences, and (ii) whether they are present in both coding and non-coding segments or only in the latter. In our analysis, we will rely on the purine-pyrimidine mapping of the nucleotides proposed in [1] since our experiments have shown that the statistical properties remain unchanged even when we adopt a more complex multi-dimensional representation [9]. We will first prove using Prietley's statistical test for stationarity that DNA sequences are non-stationary and the nature of their non-stationarity is more complex than embedded trends. Hence, classical tools for stationary time-series analysis (e.g., stationary correlation and power spectrum) and the DFA method cannot be applied to DNA sequences. A generalization of the periodogram for estimating the power spectrum of non-stationary signals is given by the *evolutionary periodogram* (EP) [10]. We show, using the EP, that DNA sequences (both coding and non-coding) exhibit an evolutionary $1/f$ spectrum. That is, the spectral exponent is not constant but rather

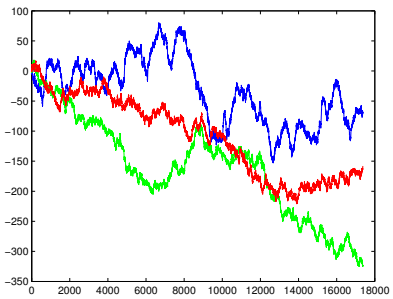


Fig. 1. DNA walks: In blue is the DNA walk for the human geneTXNDC9 (GI:89161199); In green is the DNA walk for a random gene sequence having the same nucleotide distribution as TXNDC9; In red is the DNA walk for a random sequence with uniform nucleotide distribution. Observe that the random sequences are steadily monotonic as expected for random walks with drift.

varies as a function of time (here, time denotes nucleotide position). Experimentally, we observe that the average (over time) spectral component of non-coding sequences is higher than its corresponding value for coding sequences. To demonstrate that this conclusion is not an artifact of the evolutionary $1/f$ model, we propose an index of randomness to quantify how far is a process from white noise. Our experimental results show that, indeed, coding sequences are "whiter" than non-coding sequences, confirming the evolutionary periodogram results.

2. THE EVOLUTIONARY SPECTRUM AND TEST FOR STATIONARITY

Priestley [11] proposed a method to test the overall stationarity of the complete second-order properties of a time-series. The basis of the method is to estimate its evolutionary (or time-dependent) spectrum over a discrete range of time points, and then test these spectra for uniformity over time. Priestley suggested obtaining an estimate of the evolutionary spectrum at time t_0 and frequency ω_0, $\hat{h}_{t_0}(\omega_0)$, by bandpass filtering the signal around ω_0, and then estimating the local power in a short-time window. If now we write $Y_{t,\omega} = \log(\hat{h}_t(\omega))$ and adopt the notation $Y_{i,j} = Y_{t_i,\omega_j}$, then the test for stationarity can be written in the form

$$H_0 : Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij};$$
$$H_1 : Y_{ij} = \mu + \beta_j + e_{ij},$$

where $i = 1, \cdots, I$, $j = 1, \cdots, J$, $\{e_{ij}\}$ denotes the estimation error of the evolutionary spectrum with variance σ^2, The parameters $\{\alpha_i\}$, $\{\beta_j\}$ may be interpreted as the main effects of the time and frequency factors, respectively, and the $\{\gamma_{ij}\}$ represent an interaction term between these two factors. Observe that if all the $\{\gamma_{ij}\}$ are zero, then $\log(h_t(\omega))$ is additive

Table 1. Analysis of variance for a two-factor design

Item	Degrees of freedom	Sum of squares
Between times	$I-1$	$S_T = J\sum_{i=1}^{I}(Y_{i.} - Y_{..})^2$
Between frequencies	$J-1$	$S_F = I\sum_{j=1}^{J}(Y_{.j} - Y_{..})^2$
Interaction + residual	$(I-1)(J-1)$	$S_{I+R} = \sum_{i=1}^{I}\sum_{i=1}^{I}(Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2$

in terms of time and frequency, so that $h_t(\omega)$ is multiplicative. It is then not difficult to show that $X(t)$ must be of the form $X(t) = c(t)X_0(t)$, where $c(t)$ is a deterministic function and $X_0(t)$ is a stationary process. Processes of this form are called *uniformly modulated processes*. Observe that if the exponential signal, $e^{X(t)}$, is uniformly modulated, then the signal, $X(t)$, has non-stationary trends as defined in Eq. (1). Given the computed values of Y_{ij}, we construct the standard analysis of variance table for a two-factor design, which with the usual notation, is set out in Table 1.

1. In testing for stationarity, the first step is to test for the interaction sum of squares, using the results, $S_{I+R}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$ (σ^2 is assumed to be known so that all comparisons are based on χ^2 rather than F-tests.)
2. If the interaction is not significant, we conclude that $X(t)$ is a uniformly modulated process, and proceed to test for stationarity by testing S_T using $S_T/\sigma^2 \sim \chi^2_{(I-1)}$.
3. If, however, the interaction turns out to be significant, we conclude that $X(t)$ is non-stationary and non-uniformly modulated.
4. Reversing the roles of "times" and "frequencies", the above procedure may be used to test for "complete randomness" at all times.

Figure 1 shows the DNA walk of the Human gene TXNDC9. Using the same statistical parameters in [11, Chapter 6], we applied the above test to this gene with 95% confidence. We obtain the following statistics for the exponential signal of the Human gene TXNDC9: $S_{I+R}/\sigma^2 = 1284.5 > \chi^2_{336}(0.05) = 379.74; S_T/\sigma^2 = 9.7\times 10^7 > \chi^2_{56}(0.05) = 74.46; S_F/\sigma^2 = 6912.4 > \chi^2_6(0.05) = 12.59$. The interaction, the between times sum of squares and the between frequencies sum of squares are highly significant confirming that the exponential signal is non-stationary, non-uniformly modulated and non-random. In particular, this genomic signal is non-stationary and the nature of its non-stationarity is not associated with a deterministic trend as in Eq. (1).

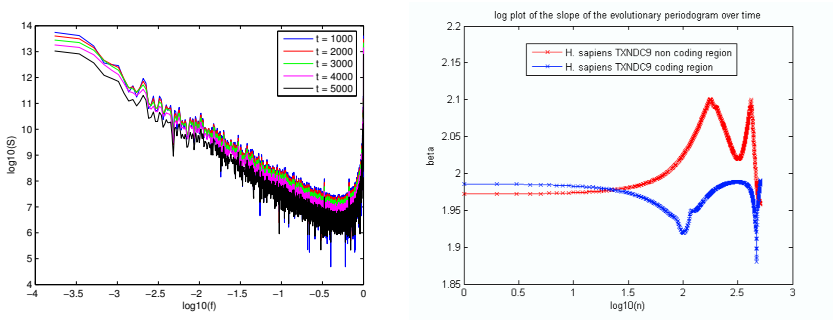
3. THE EVOLUTIONARY PERIODOGRAM AND THE EVOLUTIONARY $1/F$ PROCESS

Much of the current evidence for long-range correlations in DNA sequences stems from the experimentally observed $1/f$ spectrum [12], [13]. The $1/f$ spectrum assumes the existence of a stationary process with a fixed spectral exponent β. This assumption, however, is in contradiction to our assertion that nucleotide sequences are non-stationary. We therefore propose a new evolutionary (time-dependent) $1/f$ spectrum whose spectral exponent $\beta(n)$ varies in time. This approach also resolves the classical paradox of $1/f$ processes, namely, the variance of a $1/f$ process with a spectral exponent β, $1 < \beta < 2$, obtained by integration of the power spectral density, is infinite [14].

A generalization of the periodogram for estimating the power spectrum of non-stationary signals is given by the *evolutionary periodogram* (EP) [10]. The EP of a non-stationary signal $x(n), n = 0, \cdots, N-1$, is defined as

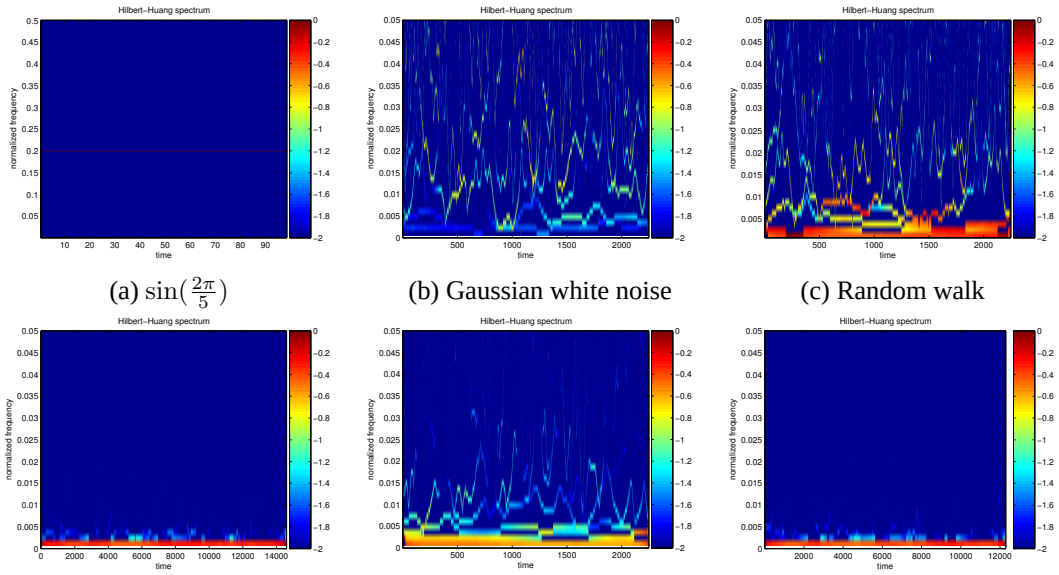
$$S(n,f) = \frac{N}{M}\Big|\sum_{i=0}^{M-1} P_i^*(n)\sum_{k=0}^{N-1} P_i(k)x(k)e^{-2\pi jfk}\Big|^2, \quad (2)$$

where * denotes complex conjugate, and $\{P_i(n)\}_{i=0}^{M-1}$ is an orthonormal basis. In our simulations, we use the discrete Legendre polynomials with $M = 3$. The EP of the coding region of the Human MHY6 gene is shown in Fig. 2(a) for $n = 1000, 2000, 3000, 4000, 5000$. Note that the two peaks, corresponding to the frequencies $1/3$ and $2/3$, are known to be related to the codon structure in DNA coding regions. Also, note that the scaling exponent β is not constant, but rather varies for different values of n. This shows that DNA correlations are much more complex than power laws with a single scaling exponent. Thus, the proposed time-varying or "evolutionary $1/f$" process, where the exponent $\beta(n)$ is a function of time, provides a far superior model of the correlation structure of DNA sequences. We estimate the function $\beta(n)$ by a linear least-squares fit of the slope of the EP at each time instant n. White noise corresponds to $\beta(n) = 0$. Figure 2(b) depicts a plot of $\beta(n)$ versus $\log_{10}(n)$ for the coding and non-coding regions of the Human gene TXNDC9. Observe that, for this gene, both the coding and non-coding regions exhibit long-range correlations. Moreover, the average exponent function of the non-coding region is higher than the corresponding value in the coding region. Next, we will demonstrate that our conclusion that (i) neither the coding nor non-coding regions are random and (ii) the "degree of randomness" of the coding regions is higher than non-coding regions, is not an artifact of the evolutionary $1/f$ model.



(a) (b)

Fig. 2. (a) Evolutionary Periodogram of the coding region of the Human MHY6 gene for $n = 1000, 2000, 3000, 4000$ and 5000. The length of the gene is $N = 5820$. (b) The scaling exponent $\beta(n)$ for the coding and non-coding regions of the Human gene TXNDC9 as a function of $\log_{10}(n)$.



(a) $\sin(\frac{2\pi}{5})$ (b) Gaussian white noise (c) Random walk

(d) Human gene NOC2L (e) coding segment (f) non-coding segment

Fig. 3. Amplitude-frequency-time distribution using the Hilbert transform (amplitudes in log).

4. EMPIRICAL MODE DECOMPOSITION AND INDEX OF RANDOMNESS

To quantify the statistical processes further, a more sensitive index is needed to give a quantitative measure of how far the process deviates from white noise; a prerequisite for such a definition is a method to present the data in the frequency-time space. There are many methods to obtain such a 3D distribution, e.g., the spectrogram, the wavelet analysis and the Wigner-Ville distribution. These techniques have been reviewed and assessed in [15], where the authors introduced a new non-linear technique, called *Empirical Mode Decomposition* (EMD), to represent non-stationary signals as sums of AM-FM components by decomposing them into mode functions and then applying the Hilbert Transform to each mode. The analytic process $Z(t)$ can then be expressed as [15]

$$Z(t) = \sum_{j=1}^{N} a_j(t) e^{i2\pi \int f_j(t)dt}. \quad (3)$$

Equation (3) enables us to represent the amplitude, $a_j(t)$, and the instantaneous frequency, $f_j(t)$, as functions of time in a three-dimensional plot, in which the amplitude can be contoured on the frequency-time plane. This frequency-time distribution of the amplitude is designated as the Hilbert spectrum. Figure 3 shows the Hilbert amplitude spectrum of a pure sine wave, a Gaussian random noise, the Human gene NOC2L and its coding and non-coding sequences. Visually, the coding segment looks "whiter" that the non-coding one. We propose to quantify the notion of "how far is a process from a white noise" by defining the index of randomness at

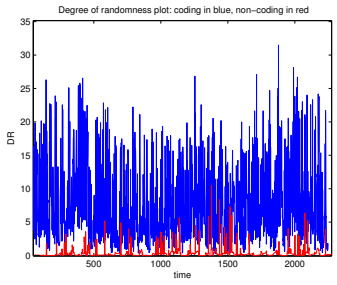


Fig. 4. Index of randomness plot for the Human gene NOC2L: the coding sequence is shown in blue and the non-coding sequence in red. The plot for the non-coding segment has been truncated to the length of the coding segment.

time instant t, IR(t), as the weighted variance or spread of the spectrum at time t. So, for a pure sine wave, the spectrum is a delta function and the variance is zero; whereas for a white noise, the spectrum is flat and the variance is infinite. Analytically,

$$\mathrm{IR}(t) = \frac{1}{N}\sum_{f=1}^{N} \frac{\mathrm{a}(f,t)}{\max_f\{\mathrm{a}(f,t)\}}(f-\mu(t))^2, \qquad (4)$$

where $a(f,t)$ is the amplitude of the Hilbert spectrum at frequency f and time t, N is the maximum number of frequency cells, and $\mu(t) = \mathrm{mean}_{f\in I(t)}\ \{f\}$, where $I(t) = \{f : a(f,t) \neq 0\}$. Figure 4 displays the index of randomness plots of the coding and non-coding regions of the Human gene NOC2L.

5. CONCLUSION

In the light of non-stationary time-series analysis, the statistical differences between coding and non-coding sequences are more subtle than previously concluded using the stationary analysis: Both coding and non-coding sequences exhibit long-range correlations as attested by an evolutionary $1/f$ spectrum. However, coding sequences are "whiter" than non-coding sequences. The results of this paper are intended to be the first step towards settling the debate of the nature of DNA correlations and setting forward the progress of understanding its origin and evolution.

6. REFERENCES

[1] C K Peng, S V Buldyrev, A L Goldberger, S Havlin, F Sciortino, M Simons, and H E Stanley, "Long-range correlations in nucleotide sequences," *Nature*, vol. 356, no. 6365, pp. 168–170, March 1992.

[2] W Li and K Kaneko, "Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence," *Europhysics Letters*, vol. 17, pp. 655, February 1992.

[3] Boris Podobnik, Jia Shao, Nikolay V. Dokholyan, Vinko Zlatic, H. Eugene Stanley, and Ivo Grosse, "Similarity and dissimilarity in correlations of genomic DNA," *Physica A*, vol. 373, pp. 497–502, 2006.

[4] C. A. Chatzidimitriou-Dreismann and D. Larhammar, "Long-range correlations in DNA," *Nature*, vol. 361, pp. 212, January 1993.

[5] S Karlin and V Brendel, "Patchiness and correlations in DNA sequences," *Science*, vol. 259, no. 5095, pp. 677–680, 1993.

[6] C K Peng, S V Buldyrev, S Havlin, M Simons, H E Stanley, and A L Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E*, vol. 49, pp. 1685 – 1689, 1994.

[7] Jan W. Kantelhardt, Eva Koscielny-Bundea, Henio H. A. Rego, Shlomo Havlinb, and Armin Bundea, "Detecting long-range correlations with detrended fluctuation analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 295, no. 15, pp. 441–454, June 2001.

[8] Zhi Chen, Plamen Ch. Ivanov, Kun Hu, and H. Eugene Stanley, "Effect of nonstationarities on detrended fluctuation analysis," *Physical Review E*, vol. 65, pp. 041107, 2002.

[9] Paul Dan Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871 – 888, April 2003.

[10] A Salim Kayhan, Amro El-Jaroudi, and Luis F Chaparro, "Evolutionary periodogram for nonstationary signals," *IEEE Transactions on Signal Processing*, vol. 42, no. 6, pp. 1527–1536, June 1994.

[11] M B Priestley, *Non-linear and Non-stationary time series analysis*, Academic Press, 1988.

[12] Richard F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, pp. 3805 – 3808, 1992.

[13] W Li and D Holste, "Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome," *Physical Review E*, vol. 71, pp. 041910, 2005.

[14] Marvin S Keshner, "1/f noise," *Proceedings of the IEEE*, vol. 70, no. 3, pp. 212–218, March 1982.

[15] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society*, vol. 454, no. 1971, pp. 903–995, March 1998.