EMERGENCE OF NEW STRUCTURE FROM NON-STATIONARY ANALYSIS OF GENOMIC SEQUENCES

Nidhal Bouaynaya[1] *and Dan Schonfeld*[2]

[1]Department of Systems Engineering, University of Arkansas at Little Rock
[2]Department of Electrical and Computer Engineering, University of Illinois at Chicago
nxbouaynaya@ualr.edu, dans@uic.edu

ABSTRACT

In this paper, we will bring to bear new tools to analyze non-stationary signals that have emerged in the statistical and signal processing community over the past few years. The emergence of these new methods will be used to shed new light and help resolve the issues 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 much more subtle than previously thought using stationary analysis. In particular, both coding and non-coding sequences exhibit long-range correlations, as asserted by a $\mathbf{1}/\mathbf{f}^{\beta(\mathbf{n})}$ evolutionary (i.e., time-dependent) spectrum. However, we will use an index of randomness, which we derive from the Hilbert-Huang Transform, to demonstrate that coding sequences, although not random as previously suspected, are often "more random" (i.e., more white) than non-coding sequences. Moreover, the study of the evolution of the rate of change of these time-dependent parameters in homologous gene families shows a sudden jump around the rat, which might be related to the well-known supercharged evolution of this rodent.

1. INTRODUCTION

In 1992, Peng et al.[1] studied the stochastic properties of DNA sequences by constructing a map of the 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. In our analysis, we will rely on the same mapping of the nucleotides. Peng et al. found that non-coding sequences exhibit long-range correlations; whereas coding sequences behave like random sequences or sustain at most short-range correlations. Similar observations were reported independently by Li et al. [2] This prompted a sequence of controversial papers, some affirming [3] and others disputing [4] the existence of long-range correlations in DNA sequences or the statistical difference between coding and non-coding segments. This debate continues till today and consequently impedes further progress to explain the origins and functions of these correlations and their effect on the evolution of the DNA. We believe that such contradictory results are an artifact of using stationary signal processing and statistics tools to study non-stationary genomic signals. The Detrended Fluctuation Analysis (DFA) technique [5] constructs a stationary process from the non-stationary DNA walk by subtracting the non-stationary trend from the sequence. However, the DFA method is limited to the very special case of non-stationary signals consisting of stationary signals with embedded (polynomial) trends, i.e.,

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

where $c(t)$ is a deterministic (usually assumed polynomial) function and $X_0(t)$ is a stationary process. Moreover, even if the data were embedded in some trend, then (i) one has to estimate the form of the trend (polynomial, logarithmic, exponential, sinusoidal, etc) in order to subtract it, and (ii) one has to guarantee that the window size adopted in the DFA always coincides with the local stationary time scale. As we will show below, genomic sequences are quite complex and exhibit different forms of non-stationarities that are more heterogeneous than embedded trends. Therefore, in the hope to resolve the issue of long range correlations in genomic sequences, one should apply techniques for a wider class of non-stationary signals.

2. NON-STATIONARY MODEL

Through our extensive simulations and analysis of different nucleotide sequences, we found that genomic sequences exhibit different forms of non-stationarity. Priestley [6] proposed a statistical test for stationarity. The basis of the method is to estimate the evolutionary (or time-dependent) spectrum of the process over a discrete range of time points, and then test these spectra for uniformity over time. Figure 1(a) shows the DNA walk of the Human gene TXNDC9. Applying Priestley's test on this gene reveals, with 95% confidence, that its DNA walk is non-stationary and its non-stationarity is not associated with a deterministic trend. Therefore, the DFA is not an appropriate tool to study this gene.

3. THE EVOLUTIONARY $1/F$ SPECTRUM

Much of the current evidence for long-range correlations in DNA sequences stems from the experimentally observed

1-4244-2372-9/08/$20.00 ©2008 IEEE

$1/f$ spectrum [4]. 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 [7].

A generalization of the periodogram for estimating the power spectrum of non-stationary signals is given by a powerful new method called the *evolutionary periodogram* (EP) [8]. The EP of a non-stationary signal $x(n), n = 0, \cdots, N-1$, is defined as

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

where * denotes complex conjugate, and $\{P_i(n)\}_{i=0}^{M-1}$ is an orthonormal basis, and $M \leq N$. In our simulations, we use the discrete Legendre polynomials with $M = 3$. Observe that Eq. (2) can be interpreted as the magnitude squared of the Fourier transform of $x(k)$ windowed by the sequence $v(n,k) = \sum_{i=0}^{M-1} \beta_i^*(n)\beta_i(k)$. The EP of the coding region of the Human MHY6 gene is shown in Fig. 1(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 1(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 or 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.

4. INDEX OF RANDOMNESS

A prerequisite for a quantitative definition of a"degree of randomness" is a method to represent the data in the frequency-time space. The Fourier transform represents a signal as a composition of stationary sinusoidal components with constant amplitude and frequency, and so is not appropriate for the analysis of non-stationary signals. An emerging method for the representation of non-stationary signals relies on the AM-FM model and often uses the Hilbert Transform for demodulation. We use the powerful new method of *Empirical Mode Decomposition* (EMD) [9] to decompose the genetic process into a finite number of basis functions admitting well-behaved Hilbert transforms. We then apply the Hilbert transform to each basis function and construct the energy-frequency-time distribution, designated as the Hilbert spectrum [9]. The process $\{X(t)\}$ can then be expressed as

$$X(t) = \sum_{j=1}^{n} a_j(t) \exp(i \int \omega_j(t)dt). \qquad (2)$$

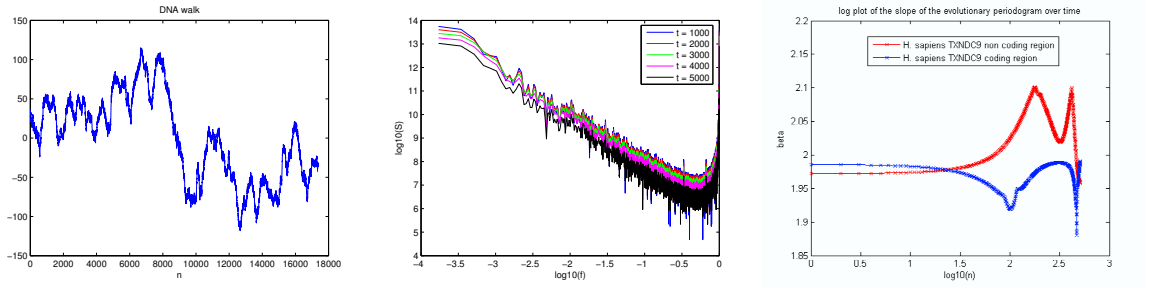
We define the index of randomness, $IR(t)$, of a signal at instant t, as the weighted variance or spread of the spectrum at time t. Therefore, for a pure sine wave, the spectrum is a delta function and the variance is zero; whereas for white noise, the spectrum is flat and the variance is infinite. Analytically,

$$\mathrm{IR}(t) = \frac{1}{N} \sum_{f=1}^{N} \frac{\mathrm{a}(f,t)}{\max_f\{\mathrm{a}(f,t)\}}(f - \mu(t))^2, \qquad (3)$$

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 $\mu(t) = \mathrm{mean}_{1 \leq f \leq N}\{a(f,t)\}$.

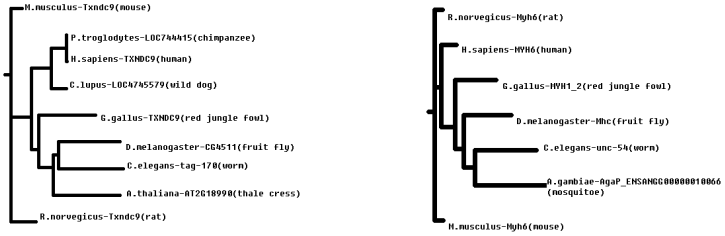
5. EVOLUTIONARY TRENDS

We now apply the non-stationary tools presented to two homologous gene families: the myosin heavy cardiac muscle gene and the thioredoxin domain containing 9 gene. Both homologous groups were identified using the online NCBI HomoloGene system for automated detection of homologs among annotated genes (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene`). We plotted the inferred phylogenies of both families in Fig. 2 using the PHYLIP package developed at the University of Washington (`http://evolution.genetics.washington.edu/phylip.html`). The exponent curves $\beta(n)$ of the coding and non-coding segments of each gene are displayed in Fig. 3, along with the average index of randomness of the coding and non-coding segments of each gene group. Notice that the exponent curve $\beta(n)$ is more conserved across evolution in exons than in introns. This result is consistent with the findings that functional DNA sequences tend to undergo mutation at a slower rate than nonfunctional sequences [10]. For example, the coding sequence of a human protein-coding gene is typically about 80% identical to its mouse ortholog, while their genomes as a whole are much more widely divergent. Moreover, the average index of randomness in coding sequences is higher than its counterpart in non-coding sequences. Finally, even though the exponent curves $\beta(n)$ do not seem to follow a particular evolutionary trend, we will show that some statistical features derived from $\beta(n)$ exhibit very interesting evolutionary patterns. For each gene, we consider the average exponent β_a given by the mean of



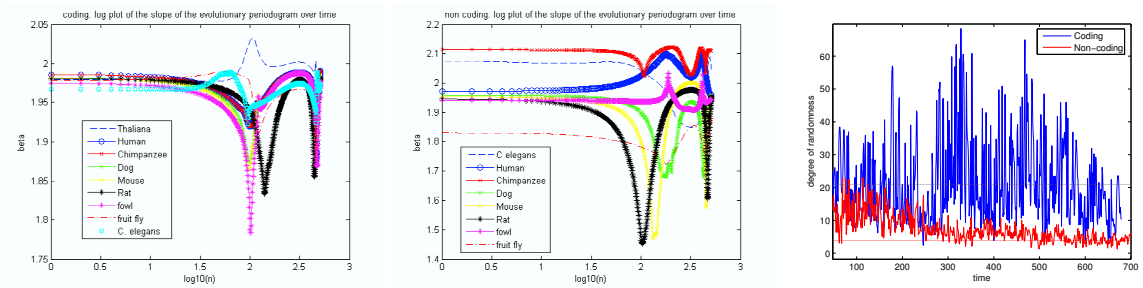
(a) (b) (c)

Figure 1. (a) DNA walk of the Human gene MHY6 using the purine-pyrimidine rule; (b) 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 $N = 5820$; (c) The scaling exponent $\beta(n)$ for the coding and non-coding regions of the Human gene TXNDC9 as a function of $\log_{10}(n)$.

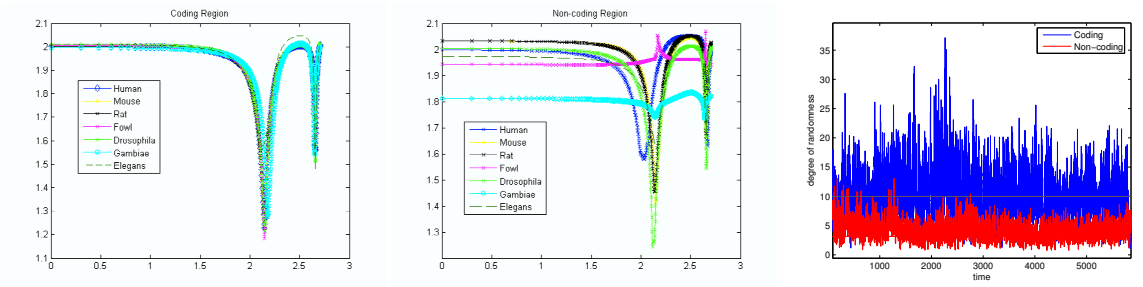


(a) Homologous TXNDC9 genes (b) Homologous MYH6 genes

Figure 2. The Phylogenetic trees of the gene groups: TXNDC9 and MYH6.



(a) Coding region (b) Non-coding region (c) Index of randomness



(d) Coding region (e) Non-coding region (f) Index of randomness

Figure 3. Exponent curves and index of randomness. Row one: Gene TXNDC9 (a) The exponent curves $\beta(n)$ of the coding region of gene TXNDC9; (b) The exponent curves $\beta(n)$ of the non coding region of gene TXNDC9; (c) Index of randomness of the coding (blue) and non-coding (red) segments of the TXNDC9 gene group. The plot of the non-coding graph was truncated to the length of the coding segment. The lower (upper) horizontal line is the average index of randomness of the non-coding (coding) regions. Row 2: same as Row 1 for gene MYH6.

Table 1. Evolutionary rates and their variances

Gene TXNDC9	Evolutionary Rate	Variance	Gene MYH6	Evolutionary Rate	Variance
Thaliana			Gambiae		
Elegans	-0.04	0.00	Elegans	0.08	0.02
Drosophila	0.00	0.00	Drosophila	-0.09	0.01
Fowl	-0.06	0.00	Fowl	0.00	0.03
Rat	0.03	0.01	Rat	0.26	0.58
Mouse	0.12	0.18	Mouse	-1.16	0.58
Dog	-0.67	0.21	Human	0.00	
Chimpanzee	0.15	0.19			
Human	0.00				

the coding curve. We define the evolutionary rate, r_g, at a node gene g as the derivative of β_a along the tree branch between the gene, g, and its ancestor G, i.e., $r_g = \frac{\beta_a(g)-\beta_a(G)}{t_g-t_G}$, where $\beta_a(g)$, $\beta_a(G)$ are the values of β_a for the genes g and G, respectively; and t_g, t_G are the relative evolutionary times of genes g and G, respectively. The evolutionary distance $t_g - t_G$ was computed as the distance between the aligned gene sequences g and G, provided by the PHYLIP package. Table 1 provides the evolutionary rates of both gene groups and shows a clear jump in the evolutionary rate around the mouse in both gene groups. This observation is quite remarkable given the well-known explosive evolution of this rodent. Furthermore, the variance of the evolutionary rates, using a window of size 3, shows an increasing trend throughout evolution. The evolutionary rate could therefore possibly be used to observe and predict the dynamics of change in a lineage.

6. CONCLUSION

We have introduced new non-stationary methods to study the correlation properties in nucleotide sequences, and defined a quantitative measure of the degree of randomness. We find that coding and non-coding DNA sequences exhibit long range correlations, as attested by an evolutionary $1/f$ spectrum. So, DNA correlation are much more complex than power laws with a single scaling exponent. Furthermore, to quantify the statistical processes further, an index is introduced to give a quantitative measure of how far the process deviates from a random white noise. The higher the index value, the more random is the process. We find that coding segments are "closer", on average, to random sequences than non-coding segments. This observation might have been the source of confusion and controversy in previous work related to DNA correlations. Finally, we showed that the evolutionary rate, which is the derivative of the average power law scaling exponent, can be used to observe and possibly predict the dynamics of change in a lineage.

7. 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] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg, and J. L. Oliver, "Identifying chracteristic scales in the human genome," *Physical Review E*, vol. 75, pp. 032903, 2007.

[4] S. Guharay, B. R. Hunt, J. A. York, and O. R. White, "Correlations in dna sequences across the three domains of life," *Physica D*, vol. 146, no. 1-4, pp. 388–396, 2000.

[5] 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.

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

[7] V. Solo, "Intrinsic random functions and the paradox of 1/f noise," *SIAM Journal on Applied Mathematics*, vol. 52, no. 1, pp. 270–291, February 1992.

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

[9] 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.

[10] M. G. S. Consortium, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, no. 6915, pp. 520–62, December 2002.