Nonstationary Analysis of Coding and Noncoding Regions in Nucleotide Sequences

Nidhal Bouaynaya, *Member, IEEE*, and Dan Schonfeld, *Senior Member, IEEE*

***Abstract*—Previous statistical analysis efforts of DNA sequences revealed that noncoding regions exhibit long-range power law correlations, whereas coding regions behave like random sequences or sustain short-range correlations. A great deal of debate on the presence or absence of long-range correlations in nucleotide sequences, and more specifically in coding regions, has ensued. These results were obtained using signal processing techniques for stationary signals and statistical tools for signals with slowly varying trends superimposed on stationary signals. However, it can be verified using statistical tests that genomic sequences are nonstationary and the nature of their nonstationarity varies and is often much more complex than a simple trend. In this paper, we will bring to bear new tools to analyze nonstationary 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 noncoding segments or only in the latter. It turns out that the statistical differences between coding and noncoding segments are much more subtle than previously thought using stationary analysis. In particular, both coding and noncoding sequences exhibit long-range correlations, as asserted by a $1/f^{\beta(n)}$ evolutionary (i.e., time-dependent) spectrum. However, we will use an index of randomness, which we derive from the Hilbert transform, to demonstrate that coding segments, although not random as previously suspected, are often "closer" to random sequences than noncoding segments. Moreover, we analytically justify the use of the Hilbert spectrum by proving that narrowband nonstationary signals result in a small demodulation error using the Hilbert transform.**

***Index Terms*—AM-FM signals, empirical mode decomposition, evolutionary periodogram, Hilbert transform, long-range correlations, nonstationary time-series analysis.**

I. Introduction

OVER the past decade, there has been a flow of conflicting papers about the long-range power-law correlations detected in eukaryotic DNA [1]–[21]. The controversy is generated by conflicting views that either advocate that noncoding DNA sustains long-range correlations whereas coding DNA behaves like random sequences [1]–[12] or maintains that both coding and noncoding DNA exhibit long-range power-law correlations [13]–[21]. Disturbingly, not only their results were contradictory for different gene data but also for the same set of genes [17]. This issue seems to hamper further progress towards explaining the origins of such correlations and their role in gene evolution, which can help understand gene-related diseases like cancer and Alzheimer disease. Indeed, understanding the correlation structure of genes helps recreate the dynamical processes that led to the current DNA sequences. For instance, based on the first view of long-range correlations in noncoding DNA, Li [22], [23], proposed a simple iterative model of gene evolution. The model incorporates the basic features of DNA evolution, that is, sequence elongation due to gene duplication and mutations. Dodin *et al.* [24] studied the correlations in DNA sequences to revisit the intriguing question of triplet repeats with the aim of providing an estimate of the propensity of genes to lead to possible aberrant structures. Their work examined Djian *et al.* [25] finding that long-range correlations due to the repeat of identical triplets has proved to be linked with severe neurological diseases in humans and primates.

The investigation of long-range correlations in DNA sequences started with the pioneering works of Peng *et al.* [1], Li and Kanenko [26], and Voss [13] in 1992. Peng *et al.* constructed a 1-D random walk model of the DNA sequence generated by an incremental variable that associates to position i the value $u(i) = 1$ if the i^{th} nucleotide of the sequence is a pyrimidine and $u(i) = -1$ if it is a purine. An important statistical quantity characterizing any walk is the root mean square fluctuation about the average of the displacement, $F(l)$. $F(l)$ is related to the auto-covariance function, $C(l)$, through the relation $F(l) = \sum_{i=1}^{N}\sum_{j=1}^{N} C(j-i)$, where $C(l) = E[u(l_0)u(l_0+l)] - E[u(l_0)]^2$, and $E[.]$ denotes an average over all positions l_0 in the gene. Independently, Li and Kineko [26] and Voss [13] took a different approach to investigate the nature of correlations in DNA sequences by studying their power spectrum. Both the root mean square fluctuation, $F(l)$, and the power spectrum, $S(f)$ 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 $C(l) \sim l^{-\gamma}$ $(\gamma > 0)$, $F(l) \sim l^{-\beta}$ $(\beta \neq (1/2))$, and $S(f) \sim 1/f^{\alpha}$ $(0 < \alpha < 2)$.

Processes whose sample power spectrum is of the form σ^2/f^{β}, for some finite nonzero σ and $\beta > 0$ are called "$1/f$ processes."

Manuscript received September 7, 2007; revised March 8, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ioan Tabus.

N. Bouaynaya is with the Department of Systems Engineering, University of Arkansas at Little Rock, Little Rock, AR 72204 USA (e-mail: nxbouaynaya@ualr.edu).

D. Schonfeld is with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607-7053 USA (e-mail: dans@uic.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2008.923852

Such processes have been empirically observed in a wide variety of physical systems such as electronic devices [27], geophysical records [28], the flood levels of the river Nile [29], sunspot activity, financial data [30], network flow [31], image texture [32], heart-rate variability [33], and many more. Peng *et al.* [1] and Li and Kaneko [26] found that noncoding sequences (intron-rich genes and nontranscribed regulatory regions) exhibit long-range power law correlations, whereas coding sequences (cDNA or intron-less genes) sustain short range correlations or behave like random sequences. Voss [13], however, found that all DNA regions exhibit long-range correlations. Analyzing the same data set as Peng *et al.*, Chatzidimitriou-Dreismann and Larhammar [17] found no difference between intron-containing versus intronless sequences. Since then, researchers from different backgrounds have been searching for long-range correlations in various DNA sequences using either the fluctuation analysis or the power spectrum. The availability of large DNA sequences and complete genomes in public databases did not help resolve the debate, but rather accentuated it [12]. This raises the question: What is the cause of all these conflicting results?

Both the fluctuation analysis and the power spectrum techniques implicitly assume that the stochastic process underlying the distribution of nucleotides is stationary. This assumption is problematic due to the complex mosaic nature of DNA sequences, with structures such as isochores, intergenic sequences, long and short interspaced repeats, tandom repeats, exons, introns, etc. Each structure has its own size distribution, nucleotide frequency and seem to have followed a different evolutionary path than other structures. Stationary studies inherently involve averaging over large portions of a sequence, and so they smear the fine details where important information might be concealed. Karlin and Brendel [16] first questioned the stationarity assumption in Peng's fluctuation analysis. They argued that most DNA sequence variation can be explained by compositional patchiness and does not involve the higher order organization implied by the long-range correlations. Peng's group disputed their argument by refining the fluctuation analysis to determine patchiness arising from the heterogenous purine-pyrimidine content in the DNA walk [2]. The main idea consists of an attempt to construct a stationary time-series from the nonstationary DNA process as follows:

1) divide the DNA sequence into subsequences;
2) estimate and remove the linear trend in each subsequence;
3) compute the variance or fluctuation of each "detrended" subsequence;
4) the overall fluctuation is then given by the average of these variances over all subsequences.

The relationship between the detrended root mean square fluctuation, $F_d(l)$, and the window length l is identical to the standard fluctuation analysis technique. That is, if only short-range correlations (or no correlations) exist in the nucleotide sequence, then $F_d(l) \sim l^{1/2}$, and if there is long-range power-law correlations, then $F_d(l) \sim l^{\alpha}$ with $\alpha \neq (1/2)$. This technique became known as the detrended fluctuation analysis (DFA). The DFA method became then the "standard" technique in analyzing correlations in DNA sequences and other time-series [12], [34]. However, this method has two drawbacks. 1) The scaling behavior is only approached asymptotically and so deviations from a straight line are often observed in the log-log plot for small scales. In particular, these deviations limit the capability of the DFA to determine the correct correlation behavior in short records. 2) The DFA is limited to the very special case of nonstationary signals consisting of stationary signals with embedded trends, i.e., signals of the form

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

where $X_0(t)$ is a stationary process and $c(t)$ is a deterministic function. If the trend $c(t)$ is continuous on a closed interval $0 \leq t \leq T$, then by the Weierstrass theorem, it can be uniformly approximated by polynomials. Therefore, one can assume that the trend is polynomial, provided its degree is properly chosen. So, by dividing the sequence into subsequences, estimating the polynomial trend in each subsequence and subtracting it, we obtain the underlying stationary process $X_0(t)$. One can then apply stationary analysis tools like the fluctuation analysis or the power spectrum to assess the correlation structure of the underlying sequence. However, our extensive simulations of DNA sequences showed that they exhibit different forms of nonstationarities that are more complex than embedded trends. Therefore, any quest to resolve the nature of DNA correlations should consider techniques for a wider class of nonstationary signals.

This paper is organized as follows. In Section II, we adopt a 2-D numerical representation of DNA sequences and show, using Priestley's test for stationarity, that genomic sequences are nonstationary, and the nature of their nonstationarity is more complex than a simple trend. In Section III, we show, using the evolutionary periodogram, which is a generalization of the periodogram to nonstationary signals, that genomic sequences exhibit an evolutionary $1/f$ spectrum, i.e., a $1/f$ spectrum with time-dependent spectral component $\beta(n)$. Moreover, our experimental results show that the spectral curve of noncoding DNA sequences is usually higher than the corresponding curve for coding sequences. In Section IV, we affirm this finding and show that it is not an artifact of the evolutionary $1/f$ model by proposing an index of randomness based on the Hilbert spectrum, and showing that coding segments are "closer" to random sequences than noncoding segments. The Hilbert spectrum is constructed using the empirical mode decomposition (EMD) to decompose nonstationary signals into a sum of AM-FM signals, then estimate their instantaneous amplitude and frequency using the Hilbert transform. We analytically justify the use of the Hilbert transform by analyzing AM-FM signals and proving that the magnitude of the approximation error tends to zero for narrowband AM-FM signals. Finally, in Section V, we conclude and outline some extensions of our work.

II. Nonstationarity of DNA Sequences

A. Two-Dimensional Numerical Representation of DNA Sequences

In this section, we consider the problem of testing a given DNA sequence for stationarity. Before doing so, we first need to map the DNA string into a numerical sequence or a series of vectors x_k representing the four base types. This might seem an easy task, but it is actually a critical step, on which the test for stationarity and the correlation analysis depend heavily. It fact,

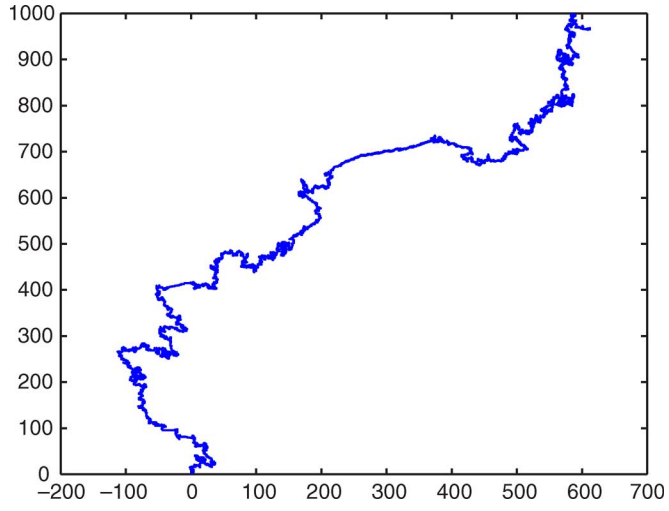

Fig. 1. Two-dimensional walk of the Human gene NOC2L.

embedding any high-dimensional data (here the dimension of the DNA is equal to 4) into a smaller space introduces correlations. On the other hand, minimizing the embedding dimension minimizes the computation time. The complementary base pairing of A with T and C with G suggests a natural embedding of a sequence into a 2-D space. Berthelsen *et al.* [35] proposed a 2-D representation of DNA sequences, characterized by a Hausdorff dimension (also called fractal dimension) that is invariant under 1) complementarity, 2) reflection symmetry, 3) compatibility, and 4) substitution symmetry of A $\leftrightarrow$ T and C $\leftrightarrow$ G. The corresponding embedding assignment is given by A $= (-1, 0)$, T $= (1, 0)$, C $= (0, -1)$, and G $= (0, 1)$. The DNA walk is defined as the sequence $\{y_i\}$, where

$$y_i = \sum_{k=1}^{i} x_k. \tag{2}$$

Fig. 1 shows the 2-D walk of the Human gene NOC2L using this 2-D embedding scheme.

B. Test Procedure

Priestley and Rao [36], [37] developed a method to test the stationarity of a given time-series. Their approach is based on evolutionary (or time-dependent) spectral analysis, and consists essentially in testing the "homogeneity" of a set of evolutionary spectra evaluated at different instants in time. An estimate of the evolutionary spectrum at time t_0 and frequency ω_0, $\hat{h}_{t_0}(\omega_0)$, is obtained by bandpass filtering the signal around ω_0, and then estimating the local power in a short-time window. The length of the time window must be long enough so that fairly stable estimates are obtainable for a reasonable number of spectral components but not too long so that the occurrence of a fundamental change will not be lost in averaging. Suppose now that we have evaluated the estimated evolutionary spectra over a set of times $t_1, t_2, \cdots, t_I$ and a set of frequencies $\omega_1, \omega_2, \cdots, \omega_J$. Considering

$$Y_{t,\omega} = \log\left\{\hat{h}_t(\omega)\right\} \tag{3}$$

and adopting the notation $Y_{i,j} = Y(t_i, \omega_j)$, the test for stationarity can be formulated as follows:

$$H_0 : Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij} \tag{4}$$

$$H_1 : Y_{ij} = \mu + \beta_j + e_{ij} \tag{5}$$

where $\{e_{i,j}\}$ is 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. If all the $\{\gamma_{ij}\}$ are zero, then $log\{h_t(\omega)\}$ is additive in terms of time and frequency, so that $h_t(\omega)$ is multiplicative, i.e., may be written in the form $h_t(\omega) = c^2(t)h(\omega)$, for some functions $c(t), h(\omega)$. It is then not difficult to show that $\{X(t)\}$ must be of the form [36]

$$X(t) = c(t)X_0(t) \tag{6}$$

where $\{X_0(t)\}$ is a stationary process with spectral density function $h(\omega)$. Processes of the form of (6) are called *uniformly modulated processes*. Thus, a test for the presence of interaction is equivalent to testing whether or not $\{X(t)\}$ is a uniformly modulated process. It is interesting to observe that if the exponential signal $e^{X(t)}$ is uniformly modulated, then the signal $X(t)$ has nonstationary trends as defined in (1). The test, formulated in (4) and (5), is equivalent to a two-factor analysis of variance procedure. The standard analysis of variance table for a two-factor design, with the usual notation, is set up in Table I. The test procedure follows the steps enumerated below.

1) In testing for stationarity, the first step is to test for the interaction sum of squares, using the result, $S_{I+R}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$ (since we are assuming that σ^2 is known, 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 nonstationary, and nonuniformly modulated.
4) Reversing the roles of "times" and "frequencies," the above procedure may be used in exactly the same way to test for "complete randomness."

Using the same statistical parameters in [37, Chapter 6], we applied the above test to the gene in Fig. 1 with 95% confidence. We obtain the following statistics for the exponential signal $S_{I+R}/\sigma^2 = 1404.6 > \chi^2_{336}(0.05) = 379.74$; $S_T/\sigma^2 = 1.2\times10^8 > \chi^2_{56}(0.05) = 74.46$; $S_F/\sigma^2 = 6152.6 > \chi^2_6(0.05) = 12.59$ The interaction, the between times sum of squares and the between frequencies sum of squares are highly significant confirming that the exponential signal is nonstationary, nonuniformly modulated and nonrandom. Therefore, this genomic signal is nonstationary and the nature of its nonstationarity is not associated with a deterministic trend as described in (1).

III. Evolutionary 1/f Process

Much of the current evidence for long-range correlations in DNA sequences stems from the experimentally observed $1/f$ spectrum [9], [13]. Although the power spectrum is valid under

TABLE I
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	$(I-1)(J-1)$	$S_{I+R} = \sum_{i=1}^{I}\sum_{i=1}^{I}(Y_{ij}-Y_{i.}-Y_{.j}+Y_{..})^2$

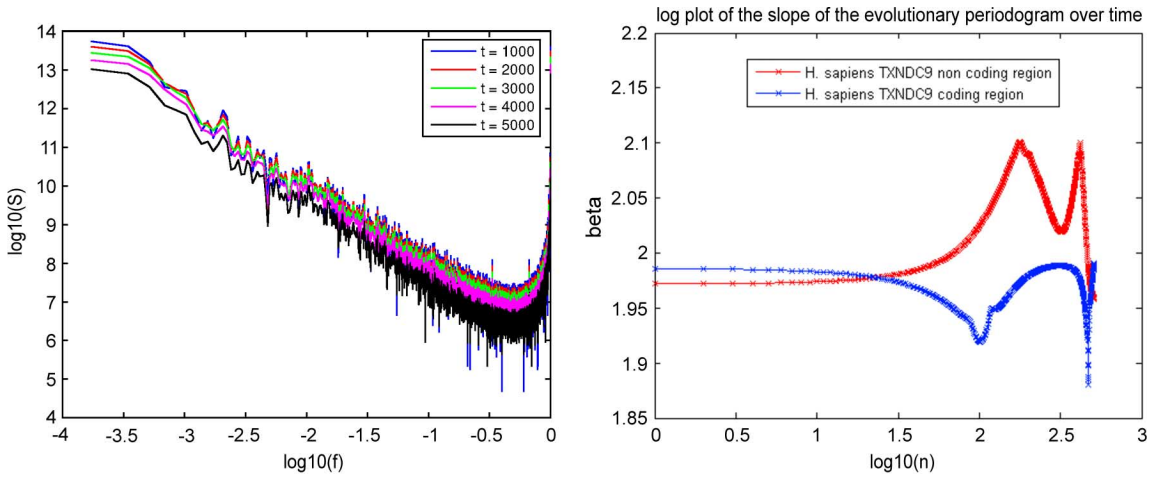

Fig. 2. (a) Evolutionary periodogram of the coding region of the Human MHY6 gene for $n = 1000, 2000, 3000, 4000, 5000$. The length of the gene is $N = 5820$. (b) The scaling exponent $\beta(n)$ for the coding and noncoding regions of the Human gene TXNDC9 as a function of $\log_{10}(n)$.

extremely general conditions [38], there are some crucial restrictions: the system must be linear and the data must be strictly periodic or stationary. The periodogram gives an estimate of the spectrum of a linear and stationary signal, which can be efficiently implemented using the fast Fourier transform (FFT) algorithm. Kayhan *et al.* [39] defined the evolutionary spectrum of a nonstationary signal using the Wold–Cramer decomposition of nonstationary processes as the output of a causal, linear and time-varying system. They subsequently extended the derivation of the stationary periodogram in [40] to the nonstationary case. The evolutionary or time-varying periodogram of a nonstationary discrete process $x(k), k = 0, \cdots, N-1$, is given by

$$S(n,f) = \frac{N}{M}\left|A(n,f)\right|^2 = \frac{N}{M}\left|\sum_{i=0}^{M-1} P_i^*(n)\sum_{k=0}^{N-1} P_i(k)x(k)e^{-2\pi jfk}\right|^2 \quad (7)$$

where * denotes complex conjugate, and $\{P_i(n)\}_{i=0}^{M-1}$ is an orthonormal basis. The number $M(\leq N)$ indicates the degree to which $A(n,f)$ varies with time. For small values of M, $A(n,f)$ is slowly varying, and for large values of M, it is rapidly varying. In our simulations, we use the discrete Legendre polynomials with $M = 3$. Observe that (7) can be interpreted as the magnitude squared of the Fourier transform of $x(k)$ windowed by the sequence $v(n,k) = \sum_{i=0}^{M-1}\beta_i^*(n)\beta_i(k)$. Therefore, the evolutionary periodogram can be efficiently implemented using the FFT algorithm.

Fig. 2(a) shows a log-log plot of the evolutionary periodogram of the Human gene MHY6 for $n = 1000, 2000, 3000, 4000, 5000$. Interestingly, we observe a $1/f$-type structure over time; that is the evolutionary periodogram of the nonstationary process underlying the DNA sequence is of the form $1/f^{\beta(n)}$, where $\beta(n)$ is the time-dependent spectral exponent. Three important observations are drawn here: 1) the evolutionary $1/f$ structure is manifested in both the coding and noncoding segments, and, hence, both regions exhibit long-range correlations; 2) DNA correlations are much more complex than power laws with a constant scaling exponent as the stationary analysis suggested; 3) the spectral exponent curve is, on average, higher in noncoding regions than coding regions. We estimate the function $\beta(n)$ by a linear least-squares fit of the slope of the evolutionary periodogram. For a Brownian motion, it is known that $\beta(n) = 2$, $\forall n$ [41]. Fig. 2(b) depicts the plots of $\beta(n)$ versus $log10(n)$ for the coding and noncoding regions of the TXNDC9 gene. Notice that the spectral exponent of the noncoding segment is higher, on average, than its corresponding value for the coding segment. Next, we will demonstrate that our conclusion that 1) neither the coding nor noncoding regions are random and 2) the "degree of randomness" of the coding regions is higher than noncoding regions, is not an artifact of the evolutionary $1/f$ model.

IV. MULTICOMPONENT AM-FM MODEL AND INDEX OF RANDOMNESS

In the light of observation (3) and to quantify the statistical processes further, a more sensitive index is needed to give a quantitative measure of how far the process deviates from a random walk.

A. Multicomponent AM-FM Model

A prerequisite for such a definition is a method to represent the data in the frequency-time space. The Fourier transform represents a signal as a composition of stationary sinusoidal com-

ponents with constant amplitude and frequency, and so is not appropriate for the analysis of nonstationary signals. An emerging method for the representation of nonstationary signals relies on the AM-FM model and often uses the Hilbert Transform for demodulation. We use the new method of *empirical mode decomposition* (EMD) [42] to decompose the genetic process into a finite number of adaptive basis functions admitting "well-behaved" Hilbert transforms (i.e., resulting in a small demodulation error)

$$X(t) = \sum_{n=1}^{N} c_n(t) + r_n(t) \tag{8}$$

where $\{c_n(t)\}_{n=1}^{N}$ are called intrinsic mode functions (IMF) and $r_n(t)$ is the residual. We then apply the Hilbert transform to each IMF (excluding the residual) and construct the energy-frequency-time distribution, designated as the Hilbert spectrum [42]. The analytic process $\{Z(t)\}$ can then be expressed as

$$Z(t) = \sum_{j=1}^{n} a_j(t) e^{2\pi i \int f_j(t)dt}. \tag{9}$$

Equation (9) can be considered as a generalization of the Fourier transform, which corresponds to constant amplitude and frequency components. Thus, it enables us to represent the amplitude $\{a_j(t)\}_{j=1}^{n}$ and the instantaneous frequency $\{f_j(t)\}_{j=1}^{n}$ as functions of time t in a 3-D 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, $H(f,t)$.

The EMD decomposition, essentially defined by an algorithm, has received little work to assess its performance and limitations [43]. In what follows, we show that the intuition behind the construction of the intrinsic mode functions is to decompose the signal into AM-FM components with narrowband amplitude, which will result in a small demodulation error using the Hilbert transform.

A nonstationary real signal $x(t)$ can be represented as a sum of multicomponent AM-FM signals

$$x(t) = \sum_{k=1}^{N} a_k(t) \cos\left(\phi_k(t)\right) \tag{10}$$

where $a_k(t)$ and $\phi_k(t)$ are the amplitude, assumed to be bandlimited, and phase of the k^{th} component, respectively. Equation (10) can be easily verified by letting $N = 1$, $\phi_1(t) = 0$ and $a_1(t) = x(t)$. We denote by $A_k(\omega)$ the Fourier transform of $a_k(t)$ and $\mu_{a_k} = (1/\pi)\int_0^{\infty} |A_k(\omega)| d\omega$ the spectral absolute moment of $a_k(t)$, assumed to be finite. The Hilbert transform of $x(t)$ is $\mathcal{H}[x(t)] = \hat{x}(t) = x(t) * (1/\pi t)$. The analytic signal of $x(t)$ is then given by

$$z(t) = x(t) + j\hat{x}(t) = r(t)e^{j\theta(t)}. \tag{11}$$

Let the quadrature signal of the monocomponent $x(t) = a(t)\cos(\phi(t))$ be defined as

$$x_q(t) = a(t)\sin\left(\phi(t)\right). \tag{12}$$

Clearly, if the Hilbert transform of $x(t)$ is equal to its quadrature signal, then the Hilbert transform estimates of $r(t)$ and $\theta(t)$ are equal to the actual information signals $|a(t)|$ and $\phi(t)$. However, from the Bedrosian theorem,[1] we know that $\mathcal{H}[a(t)cos[\phi(t))] = a(t)\mathcal{H}[cos[\phi(t)]]$ only if the amplitude is varying so slowly that the frequency spectra of the envelope and the carrier waves are disjoint. This is not true in practice. Thus, there is, in general, a nonzero error between the Hilbert transform signal and the quadrature signal. Let $e(t) = x_q(t) - \hat{x}(t)$. The next proposition provides an upper bound for the approximation error in the case of a sinusoidal FM modulation.

Proposition 1: (Monocomponent AM-FM Signal With Sinusoidal Phase): Consider a nonstationary monocomponent AM-FM signal $x(t) = a(t)\cos[\phi(t)]$, where $a(t)$ is bandlimited with bandwidth W, and finite spectral absolute moment μ_a, $\phi(t) = \omega_c t + \beta\sin(\omega_m t)$, with $\omega_c \gg \omega_m$, W, and $\beta > 0$ is the FM index. Let M be the highest integer such that $\omega_c \geq M\omega_m + W$. Then, we have

$$\lim_{M\to\infty} e(t) = 0 \tag{13}$$

$$|e(t)| \leq 2\mu_a \sum_{n=M+1}^{\infty} J_n(\beta) \leq 2\mu_a e^{\frac{\beta}{2}\left(1-\frac{\beta}{2}\right)} \tag{14}$$

where J_n is the n^{th}-order Bessel function of the first kind.

Observe from (13) that an amplitude signal $a(t)$ with a narrow bandwidth corresponds to a higher value of M and, therefore, a smaller error.

Corollary 1: (Multicomponent AM-FM Signal With Sinusoidal Phase): Consider a nonstationary multicomponent AM-FM signal $x(t) = \sum_{k=1}^{N} a_k(t)\cos[\phi_k(t)]$, where $a_k(t)$ is bandlimited with bandwidth W_k, and finite spectral absolute moment μ_{a_k}, $\phi_k(t) = \omega_{c_k} t + \beta_k \sin(\omega_{m_k} t)$, with $\omega_{c_k} \gg \omega_{m_k}$, W_k, and $\beta_k > 0$ is the FM index of the k^{th} component. Let $W = \max_{1\leq k\leq N} W_k$ and M be the highest integer such that $\omega_{c_k} \geq M\omega_{m_k} + W$, for all $1 \leq k \leq N$. Then, we have

$$\lim_{M\to\infty} e(t) = 0, \tag{15}$$

$$\begin{aligned} |e(t)| &\leq 2\sum_{k=1}^{N} \mu_{a_k} \sum_{n=M+1}^{\infty} J_n(\beta_k) \\ &\leq 2\sum_{k=1}^{N} \mu_{a_k} e^{\frac{\beta_k}{2}\left(1-\frac{\beta_k}{2}\right)} \end{aligned} \tag{16}$$

where J_n is the n^{th}-order Bessel function of the first kind.

Corollary 2: (Multicomponent AM-FM Signal With a Sum of Harmonically Independent Periodic Functions Phase): Consider a nonstationary multicomponent AM-FM signal $x(t) = \sum_{k=1}^{N} a_k(t)\cos[\phi_k(t)]$, where $a_k(t)$ is bandlimited with bandwidth W_k, and finite spectral absolute moment μ_{a_k}; $\phi_k(t) = \omega_{c_k} t + \sum_{l=1}^{L} \beta_l^k m_l^k(t)$, where $m_l^k(t)$ is periodic with angular frequency $\omega_{m_l}^k$. Assume that $\omega_{c_k} \gg \omega_{m_k}^l$, W_k, for all $1 \leq l \leq L$, and $\beta_l^k > 0$. Let $W = \max_{1\leq k\leq N} W_k$ and M be the highest integer such that $\omega_{c_k} \geq M\sum_{l=1}^{L} \omega_{m_l}^k + W$. Then, we have

$$\lim_{M\to\infty} e(t) = 0$$

$$|e(t)| \leq 2\sum_{k=1}^{N} \mu_{a_k} \sum_{n_1=M+1}^{\infty} c_{n_1}^k\left(\beta_1^k\right) \cdots \sum_{n_L=M+1}^{\infty} c_{n_L}^k\left(\beta_L^k\right) \tag{17}$$

[1]The Bedrosian theorem states that the Hilbert transform for the product of two functions $f(t)$ and $h(t)$ can be written as $\mathcal{H}[f(t)h(t)] = f(t)\mathcal{H}[h(t)]$ only if the Fourier spectra for $f(t)$ and $h(t)$ are disjoint in frequency domain, and the frequency range of the spectrum for $h(t)$ is higher than that of $f(t)$.

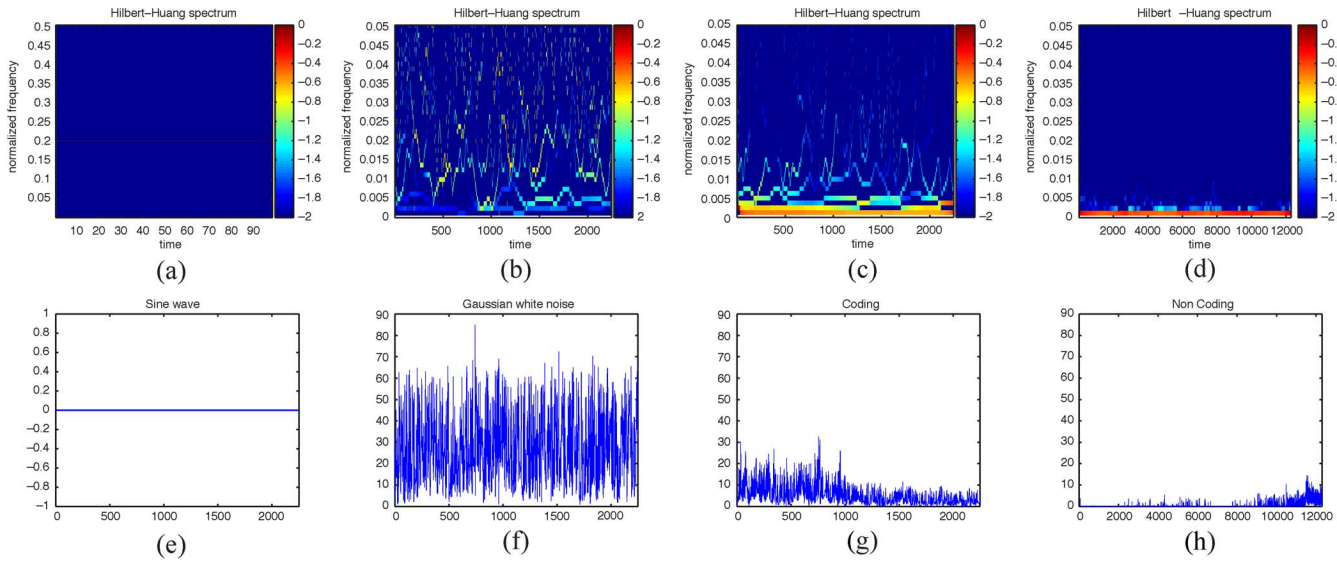

Fig. 3. Row 1: Amplitude-frequency-time distribution using the Hilbert transform (amplitudes depicted in a logarithmic scale). Row 2: Index of randomness of the signals in row 1. (c) and (d) display the Hilbert spectrum of the coding and noncoding segments of the Human gene NOC2L (GI:89161185), respectively. (a) $\sin(2\pi/5)$; (b) Gaussian white noise; (c) coding sequence; (d) noncoding sequence; (e) IR of (a); (f) IR of (b); (g) IR of (c); (h) IR of (d).

where $c_{n_l}^k(\beta_l^k)$ are the Fourier coefficients of $e^{j\beta_k m_l^k(t)}$, for $1 \leq k \leq N$ and $1 \leq l \leq L$.

Since every nonperiodic signal can be asymptotically approximated by a linear combination of harmonically independent periodic signals, Corollary 2 can be applied to a wide class of nonperiodic phases. Proposition 1 and Corollaries 1 and 2 show that narrowband multicomponent AM-FM signals have an asymptotically small demodulation error, which decreases exponentially with the FM index. Therefore, narrowband nonstationary signals can be reliably represented in the frequency-time space using the Hilbert spectrum. Having established the theoretical ground for the EMD decomposition, we use the Hilbert spectrum to define the index of randomness.

B. Index of Randomness

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

$$\mathrm{IR}(t) = \frac{1}{N}\sum_{f=1}^{N} \frac{\mathrm{a}(f,t)}{\max_f \{\mathrm{a}(f,t)\}} (f-\mu(t))^2 \qquad (18)$$

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 $I(t) = \{f : a(f,t) \neq 0\}$. Fig. 3 depicts the Hilbert spectrum and the index of randomness for different signals. We once again observe that 1) the coding and noncoding regions are not random, and 2) the coding regions are more random than the noncoding regions. The confusion and controversy about the randomness of coding DNA could be due to the fact that coding segments, though exhibit long-range correlations, are in fact closer on average to random sequences than noncoding segments. The stationary analysis conducted thus far in the literature was unable to fine-tune this complex nature of the correlations of coding and noncoding DNA.

V. Conclusion

We have introduced new nonstationary methods to study the correlation properties in genomic sequences, and defined a quantitative measure of the degree of randomness. We find that coding and noncoding DNA sequences exhibit long range correlations, as attested by an evolutionary $1/f$ spectrum. So, DNA correlations are much more complex than power laws with a single scaling exponent: actually the exponent of such power laws are different for different scales; thus, a clear scaling does not seem to exist at all. 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 noncoding segments. We have also investigated the theoretical foundations of the empirical mode decomposition, which became an effective tool for frequency-time analysis.

Appendix

Proof of Proposition 1

$$\begin{aligned}
x(t) &= a(t)\cos(\phi(t)) \\
&= a(t)\cos[\omega_c t + \beta\sin(\omega_m t)] \\
&= a(t)\mathcal{R}\left[e^{j\omega_c t}e^{j\beta\sin(\omega_m t)}\right] \\
&= a(t)\mathcal{R}\left[e^{j\omega_c t}\sum_{n=-\infty}^{+\infty} J_n(\beta)e^{jn\omega_m t}\right] \\
&= a(t)\sum_{n=-\infty}^{+\infty} J_n(\beta)\cos[(\omega_c + n\omega_m)t] \\
&= \left[\frac{1}{2\pi}\int_{-W}^{W} A(\omega)e^{j\omega t}d\omega\right] \\
&\quad \times \left[\sum_{n=-\infty}^{+\infty} J_n(\beta)\cos[(\omega_c + n\omega_m)t]\right]
\end{aligned}$$

where $\mathcal{R}(z)$ denotes the real part of z. After developing the above expression using Euler formula and exhanging the order of the summation and the integral, we obtain

$$\begin{aligned}\mathcal{H}[x(t)] &= a(t)\sum_{n=-\infty}^{\infty} J_n(\beta)\sin[(\omega_c+n\omega_m)t] \\ &\quad - 2a(t)\sum_{n=-\infty}^{-(M+1)} J_n(\beta)\sin[(\omega_c+n\omega_m)t] \\ &= a(t)\sin(\phi(t)) \\ &\quad - 2a(t)\sum_{n=-\infty}^{-(M+1)} J_n(\beta)\sin[(\omega_c+n\omega_m)t].\end{aligned}$$

Therefore

$$\begin{aligned}|e(t)| &\le 2a(t)\sum_{n=-\infty}^{-(M+1)} |J_n(\beta)| \\ &= 2a(t)\sum_{n=M+1}^{+\infty} |J_n(\beta)| \\ &\le 2\mu_a \sum_{n=M+1}^{\infty} |J_n(\beta)|.\end{aligned}$$

Since the series $\sum_{n=-\infty}^{+\infty} J_n(\beta)$ converges, we obtain (13). Equation (14) can be easily verified as follows:

$$\begin{aligned}\sum_{n=M+1}^{+\infty} J_n(\beta) &= \sum_{n=M+1}^{+\infty}\sum_{k=n}^{\infty} \frac{(-1)^{k-n}}{(k-n)!}\frac{\left(\frac{\beta}{2}\right)^{2k-n}}{k!} \\ &\le \sum_{n=M+1}^{+\infty}\sum_{k=0}^{+\infty} \frac{(-1)^{k}}{k!}\frac{\left(\frac{\beta}{2}\right)^{2k+n}}{n!} \\ &= \sum_{n=M+1}^{+\infty} e^{-\frac{\beta^2}{4}}\frac{\left(\frac{\beta}{2}\right)^{n}}{n!} \\ &\le e^{-\frac{\beta^2}{4}} e^{\frac{\beta}{2}}.\end{aligned}$$

■

The proofs of Corollaries 1 and 2 can be obtained directly from the proof of Proposition 1.

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, Mar. 1992.

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

[3] S. M. Ossadnik, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C. K. Peng, M. Simons, and H. E. Stanley, "Correlation approach to identify coding regions in DNA sequences," *Biophys. J.*, vol. 67, no. 1, pp. 64–70, 1994.

[4] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis," *Phys. Rev. E*, vol. 51, pp. 5084–5091, 1995.

[5] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Phys. Rev. Lett.*, vol. 74, no. 16, pp. 3293–3296, Apr. 1995.

[6] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, "Quantification of DNA patchiness using long-range correlation measures," *Biophys. J.*, vol. 72, pp. 866–875, Feb. 1997.

[7] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, and M. Simons, "Scaling features of noncoding DNA," *Phys. A*, vol. 273, no. 1, pp. 1–18, 1999.

[8] J. Gao, Y. Qi, Y. Cao, and W.-W. Tung, "Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences," *J. Biomed. Biotechnol.*, vol. 2, pp. 139–146, 2005.

[9] 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," *Phys. Rev. E*, vol. 71, p. 041910, 2005.

[10] A. D. Haimovich, B. Byrne, R. Ramaswamy, and W. J. Welsh, "Wavelet analysis of DNA walks," *J. Comput. Biol.*, vol. 13, no. 7, pp. 1289–1298, 2006.

[11] B. Podobnik, J. Shao, N. V. Dokholyan, V. Zlatic, H. E. Stanley, and I. Grosse, "Similarity and dissimilarity in correlations of genomic DNA," *Phys. A*, vol. 373, pp. 497–502, 2006.

[12] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg, and J. L. Oliver, "Identifying chracteristic scales in the human genome," *Phys. Rev. E*, vol. 75, p. 032903, 2007.

[13] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Phys. Rev. Lett.*, vol. 68, pp. 3805–3808, 1992.

[14] V. V. Prabhu and J. M. Claverie, "Correlations in intronless DNA," *Nature*, pp. 359–782, 1992.

[15] S. Nee, "Uncorrelated DNA walks," *Nature*, vol. 357, p. 450, 1992.

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

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

[18] V. S. Pande, A. Y. Grosberg, and T. Tanaka, "Nonrandomness in protein sequences—Evidence for a physically driven stage of evolution," *Proc. Nat. Acad. Sci.*, vol. 91, no. 26, pp. 12972–12975, 1994.

[19] M. Y. Azbel, "Universality in a DNA statistical structure," *Phys. Rev. Lett.*, vol. 75, no. 1, pp. 168–171, Jul. 1995.

[20] G. Abramson, H. A. Cerdeira, and C. Bruschi, "Fractal properties of DNA walks," *Biosystems*, vol. 49, no. 1, pp. 63–70, 1999.

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

[22] W. Li, "Spatial 1/f spectra in open dynamical systems," *Europhys. Lett.*, vol. 10, no. 5, pp. 395–400, 1989.

[23] W. Li, "Expansion-modification systems: A model for spatial 1/f spectra," *Phys. Rev. A*, vol. 43, no. 10, pp. 5240–5260, 1991.

[24] G. Dodin, P. Levoir, and C. Cordier, "Triplet correlation in DNA sequences and stability of heteroduplexes," *J. Theoret. Biol.*, vol. 183, pp. 341–343, 1996.

[25] P. Djian, J. M. Hancock, and H. Chana, "Codon repeats in genes associated with human diseases: Fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration," *Proc. Nat. Acad. Sci.*, vol. 93, pp. 306–310, 1996.

[26] W. Li and K. Kaneko, "Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence," *Europhys. Lett.*, vol. 17, p. 655, Feb. 1992.

[27] E. Rubiola and V. Giordano, "On the 1/f frequency noise in ultra-stable quartz oscillators," *IEEE Trans. Ultrason., Ferroelectron., Freq. Control*, vol. 54, no. 1, pp. 15–22, Jan. 2007.

[28] L. Telesca, V. Cuomo, and V. Lapenna, "$1/f^{\alpha}$ fluctuations of seismic sequences," *Fluctuation Noise Lett.*, vol. 2, no. 4, pp. 357–367, 2002.

[29] A. Montanari, R. Rosso, and M. S. Taqqu, "A seasonal fractionally differenced ARIMA model: An application to the Nile River monthly flows at Aswan," *Water Resources Res.*, vol. 36, pp. 1249–1259, 2000.

[30] V. Gontis and B. Kaulakys, "Long-range memory model of trading activity and volatility," *J. Statist. Mech.*, p. 10016, 2006.

[31] F. Liu, X. Shan, Y. Ren, and J. Zhang, "Phase transition and 1/f noise in a computer network model," *Phys. A*, vol. 328, no. 3, pp. 341–350, 2003.

[32] T. Lundahl, W. Ohley, S. Kay, and R. Siffert, "Fractional Brownian motion: A maximum likelihood estimator and its application to image texture," *IEEE Trans. Med. Imag.*, vol. MI-5, pp. 152–161, May 1986.

[33] Z. R. Struzik, J. Hayano, R. Soma, S. Kwak, and Y. Yamamoto, "Aging of complex heart rate dynamics," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 1, pp. 89–94, Jan. 2006.

[34] M. Jospin, P. Caminal, E. W. Jensen, H. Litvan, M. Vallverdu, M. M. R. F. Struys, H. E. M. Vereecke, and D. T. Kaplan, "Detrended fluctuation analysis of EEG as a measure of depth of anesthesia," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 840–846, May 2007.

[35] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, "Global fractal dimension of human DNA sequences treated as pseudorandom walks," *Phys. Rev. A*, vol. 45, no. 12, pp. 8902–8913, Jun. 1992.

[36] M. B. Priestley and T. S. Rao, "A test for nonstationarity of time-series," *J. Roy. Statist. Soc.*, vol. 31, no. 1, pp. 140–149, 1969.
[37] M. B. Priestley, *Non-Linear and Non-Stationary Time Series Analysis*. New York: Academic, 1988.
[38] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*. Oxford, U.K.: Oxford Univ. Press, 1948.
[39] A. S. Kayhan, A. El-Jaroudi, and L. F. Chaparro, "Evolutionary periodogram for nonstationary signals," *IEEE Trans. Signal Process.*, vol. 42, no. 6, pp. 1527–1536, Jun. 1994.
[40] S. M. Kay, *Modern Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[41] V. Solo, "Intrinsic random functions and the paradox of 1/f noise," *SIAM J. Appl. Math.*, vol. 52, no. 1, pp. 270–291, Feb. 1992.
[42] 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," *Proc. Roy. Soc.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.
[43] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb. 2004.

Nidhal Bouaynaya (M'07) received the B.S. degree in electrical and computer engineering from the Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA), France, the M.S. degree in electrical and computer engineering from the Illinois Institute of Technology, Chicago, in 2002, the Diplôme d'Etudes Approfondies in signal and image processing from ENSEA, France, in 2003, and the M.S. degree in mathematics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Chicago in 2007.

In Fall 2007, she joined the University of Arkansas, Little Rock, where she is currently an Assistant Professor in the Department of Systems Engineering. Her research interests are in signal, image, and video processing, mathematical morphology and genomic signal processing.

Dr. Bouaynaya won the Best Student Paper Award in Visual Communications and Image Processing 2006 and was a finalist in the Best Student Paper Award in Image and Video Communication and Processing 2005.

Dan Schonfeld (M'90–SM'05) was born in Westchester, PA, on June 11, 1964. He received the B.S. degree in electrical engineering and computer science from the University of California, Berkeley, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Johns Hopkins University, Baltimore, MD, in 1986, 1988, and 1990, respectively.

In August 1990, he joined the Department of Electrical Engineering and Computer Science, University of Illinois at Chicago, where he is currently a Professor in the Departments of Electrical and Computer Engineering, Computer Science, and Bioengineering, and Co-Director of the Multimedia Communications Laboratory (MCL) and member of the Signal and Image Research Laboratory (SIRL). He has authored over 100 technical papers in various journals and conferences. His current research interests are in signal, image, and video processing; video communications, retrieval, and networks; image analysis and computer vision; and genomic signal processing.

Dr. Schonfeld was coauthor (with W. Qu) of a paper that won the Best Student Paper Award at the IEEE International Conference on Image Processing 2006. He was also coauthor (with N. Bouaynaya) of a paper that won the Best Student Paper Award in Visual Communication and Image Processing in 2006. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING in Image and Video Storage, Retrieval, and Analysis and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in Video Analysis. He has also served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING in Multidimensional Signal Processing and Multimedia Signal Processing, as well as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING in Nonlinear Filtering. He is currently serving on the IEEE Image and Multidimensional Signal Processing Technical Committee. He is currently also serving as Chair of the SPIE Conference on Visual Communication and Image Processing 2007. He was a member of the organizing committees of the IEEE International Conference on Image Processing 1998 and IEEE Workshop on Nonlinear Signal and Image Processing 1997. He was also the plenary speaker at the INPT/ASME International Conference on Communications, Signals, and Systems in 1995 and 2001.