Wavelet Based Analysis in Bio-informatics

Bharat Bhosale[1], Bouthina S. Ahmed[2], Anjan Biswas[3, 4*]

[1] S H Kelkar College of Arts, Commerce and Science, University of Mumbai, Devgad 416613 (M.S.), India
[2] Department of Mathematics, Girls' College, Ain Shams University, Cairo 11757, Egypt
[3] Department of Mathematical Sciences, Delaware State University, Dover, DE 19901-2277, USA
[4] Department of Mathematics, Faculty of Science King Abdulaziz University, Jeddah, Saudi Arabia
biswas.anjan@gmail.com

Abstract: Many biological, biochemical and biomedical phenomena exhibit fractal patterns. Moreover, these phenomena can be modeled by treating them as multiplicative random processes. These features attribute to wavelet analysis, which is mainly based on scale invariance and self-similarity properties. Wavelet methods have unique ability to reveal structural properties of the multiplicative processes resulted in such biological phenomena; that makes the wavelets a versatile tool in analyzing the bio-informatics data. Among other biological branches, molecular biology alone contributes greatly to bioinformatics. Central to many problems in molecular biology is to understand the structural organization of genomic sequences. The genomic sequences are characterized by random processes and also exhibit fractal patterns. We therefore confine our discussion to genomic sequences treating them as random processes. In our present work, we propose a wavelet based mathematical tool to analyze genomic structures in stochastic framework laying emphasis on its randomized feature. The robustness of the method is justified due to the probabilistic approach adopted throughout in the formulation of the method.
[Bharat Bhosale, Bouthina S. Ahmed, Anjan Biswas. **Wavelet Based Analysis in Bio-informatics.** *Life Sci J* 2013;10(2):853-859] (ISSN: 1097-8135). http://www.lifesciencesite.com. 120

Keyword: Wavelets; solitons; fractals; bioinformatics**.**

1. Introduction

Recently, digital signal processing application to Bioinformatics has gained much importance and much attention is received to the genomic sequence analysis; that provided a new insight in problems like detecting coding regions, periodicities, finding diverse signals and studying protein structures. With the accomplishment of considerable knowledge of human genome, a new subject-bioinformatics has been emerged as a vital discipline. The study of structure and function of membrane protein is one of the important subjects in the field of bioinformatics.

Growth and form in biology are often associated with some level of fractals; especially, anatomical growth processes lead to structures that exhibit fractal statistics. For example, vasculature of brain, where fractal like pattern can be identified in tree structures of arteries. The wavelets, which are based on scale invariance and self similarity-fractal patterns, are therefore the most suitable technique employed to study the biomedical and other texture images. Being a versatile tool especially for the analysis of quasi-chaotic signals, noisy images, wavelets have got applications in all branches of medicine, biology, computer tomology, analysis of ECG, brain wave studies. Apart from its versatility and potentiality in diverse fields, wavelet analysis can be productively applied to many different signals in bioinformatics. Recently, there has been a growing interest in employing wavelet based techniques in the analysis of biological sequences and molecular biology-related signals. Particularly, in biological systems, introducing stochastic 'noise' has been found helpful in improving the signal strength of the internal feedback loops for balance and other vestibular communication. It has been found helpful to diabetic and stroke patients with balance control.

In literature, among others, Lio [15], summarized the potential of state of the art wavelets, and in particular wavelet statistical methodology, in different areas of molecular biology such as genome sequence, protein structure and microarray data analysis. In our earlier works, we studied the wavelet interaction with solitons arising as the solutions of nonlinear partial differential equations viz. Non-linear Schrodinger Equation, Sine-Gordon equation, Korteweg-de Vries equation [6, 9]. Also, we studied extensively the strong relationships existing between wavelets, solitons and probability distributions [7]. Moreover, we studied the wavelet interaction to random processes [8]. In [11] developed a wavelet multi-component decomposition algorithm for processing data from micro-Raman spectroscopy (μ-RS) of biological tissue using data from μ-RS measurements performed *in vitro* on animal (pig and chicken) tissue samples and, in a preliminary form, on human skin and oral tissue biopsy from normal subjects. Yu and Zang in their research paper [22], carried out extensively the wavelet analysis of the

Kyte-Doolittle hydrophobicity profile of a protein signal by employing discrete wavelet transform to predict the number and location trans-membrane helical segments (TMHs). In the genome data, a large portion (about 20%-30%) of proteins in a genome encodes membrane protein, the proportion of such shows the importance of membrane protein in biology. Membrane protein, especially trans-membrane protein has very important function in organism, such as photosynthesis, respiration, neural signaling, immune response, nutrient absorption and so on, and it is also the important drug target. Of the drug target known and being researched is about 70% of the membrane protein.

One of the significant features of many biological and biochemical phenomena is randomness. These phenomena can be represented by random processes and can be analyzed in stochastic framework by employing wavelet methods. For example gene expression, has a stochastic component through the molecular collisions. In [14] investigated the role of wavelet transformation in the study of random/stochastic processes. Moreover, the Electrocardiogram (ECG) signal also represents a random process and the signal has strong cyclic recurrence of standard regions of interest named waves, complexes and segments [10]. ECG signals measure the change in electrical potential over time [4]. Moreover, ECG signals are processed to extract morphological features. The resulting time series can be eventually analyzed using wavelet transform methods.

One of the challenging problems in molecular biology is to understand the structural organization of genetic sequences which are also characterized by random processes. The several investigations pertaining to genomic sequence analysis through digital signal processing techniques using different digital representations of genomic sequences have been reported. However, it is noticed that these approaches have suffered from one or the other deficiencies, especially the one that the normalized probabilistic behavior of the randomized processes characterizing such structures has not been exploited adequately. It was shown in number of studies that the distribution of nucleotides in a DNA chain is a fractal distribution. Many such biological processes can be described by probability models such as normal distribution and its associated statistics. More importantly, the sum of random processes with arbitrary distributions results in a random variable with normal probability distribution. This feature has led us to undertake the present study in which we lay emphasis on the normalized probabilistic behavior of the biological random processes in general and genomic sequences in particular. In this work, we proposed a generalized wavelet based stochastic model for analyzing genomic sequences employing appropriate form of the Gaussian wavelet representation of the normal probability distribution of the random variate representing the digital sequence as the generic/basic function. We exploited the ability of wavelet analysis to reveal structure properties of the multiplicative processes resulted in genomic/DNA sequence, by studying the correlations of wavelet coefficients of different scales.

The paper is organized in four sections besides introduction. Introduction carries survey of literature and the relevant results reported leading to the present study. First section briefs the notations and terminologies used in sequel, second section elaborates on the different representations of the genomic sequence, the third section is devoted to the mathematical analysis comprised of formulation of genomic sequence in randomized form, selection and appropriate formulation of generic wavelet suitable for the transform and application of wavelet transform with both continuous and discrete form. Conclusions are given in the last section.

2. Notations and Terminologies

Random variable or stochastic variate:

A random variable X assigns a real number $X(s)$ to each outcome s of the experiment. Let Ω is a sample space, that is, the set of all possible outcomes of random experiment E and B is a σ-field of subsets of Ω.

A function $X: \Omega \to R$ is called a random variable if the inverse image under X of all semi-closed intervals of the form $(-\infty, x]$, where $x \in R$ are events in B, that is, $X^{-1}(-\infty, x] = \{\omega \in \Omega: X(\omega) \leq x\} \in B$.

Random process:

Random process can be thought of as a sequence of random variables. More specifically, a random process $X: X(t, \omega)$ assigns a real function of time t to each outcome ω of the random experiment. Using symbolic representation of the random process makes it easy to simulate its behavior, estimate parameters from data and compute state probabilities at different times. (Ω, B, P) denotes the probability space and $X = X(t, \omega); t \in R, \omega \in \Omega$, a random process.

Normal probability distribution:

A continuous random variable X, denoted by $X \sim N(\mu, \sigma^2)$, with parameters μ and σ, where $-\infty \leq \mu \leq \infty$ and $\sigma > 0$, is said to have a normal probability distribution if its probability density function is given by

$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, where $-\infty < x < \infty$

With the transformation $Z = \frac{X-\mu}{\sigma}$, we can obtain the standard normal variate. This implies that the standard normal variable Z has mean $\mu_Z = 0$ and variance $\sigma^2 = 1$, having, the probability density function , called standard normal density

$\varphi(Z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$, where $-\infty < z < \infty$.

The normal probability integral or the area under the normal curve that gives the probability P for the interval from the mean to the value x is,

$P = \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{0}^{z} e^{-\frac{z^2}{2}} dx$

Continuous Wavelet Transform

The continuous wavelet transform (CWT) is a decomposition of a function, $f(x)$, with respect to a basic wavelet $\psi(x)$, given by the convolution of a function with a scaled and translated version of $\psi(x)$

$W_{\psi}[f(x)](a,b) =$

$|a|^{-1/2} \int f(x)\psi^*\left(\frac{x-b}{a}\right) dx = \langle f, \frac{1}{\sqrt{|a|}} \psi\left(\frac{x-b}{a}\right)\rangle$ (1)

$= \langle f, \psi_{a,b}\rangle = \langle f, U(a,b)\psi\rangle = W_{\psi} f(a,b), \langle .,. \rangle$

is the inner product.

where, $f, \psi \in L^2$, the square integrable functions. and ψ satisfies the admissibility condition

$C_{\psi} = \int \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty.$

Subscript '*' denotes complex conjugation, 'a' is the scale parameter, $a > 0$, 'b' is the translation parameter. The term $1/\sqrt{|a|}$ is the energy conservative term that keeps energy of the scaled mother wavelet equal to the energy of the original wavelet [21].

Discrete Wavelet Transform

Discrete wavelet transform (DWT) is conveniently used for numerical implementation of the transform. Discretization is done either of the transform domain parameters- scale and translation variables or of the independent variable of the function to be transformed. In each case, DWT yields a countable set of coefficients in the transform domain that corresponds to points on a two dimensional grid or lattice $m \times n$ of discrete points in the scale-translation domain. With a and b as scale and translation parameters, taking scale a: $a = a_0^m$ and the translation b: $b = n\, b_0 a_0^m$, where a_0 and b_0 are the discrete scale and translation step sizes, respectively, the DWT is given by [21],

$W_{\psi}[f(x)](m,n) = \frac{1}{\sqrt{a_0^m}} \int_{-\infty}^{\infty} f(x)\psi\left(\frac{x - nb_0 a_0^m}{a_0^m}\right) dx$

$= \frac{1}{\sqrt{a_0^m}} \int_{-\infty}^{\infty} f(x)\psi(a_0^{-m}x - nb_0)\, dx$

$= < f, \psi_{m,n} >$

$= < f, U(a_0^m, nb_0 a_0^m)\psi >.$ (2)

3. Genomic Sequence Formulations

The analysis of complex genomic structure or the nucleotide sequence of DNA and RNA chains is one of the most important problems in molecular biology. Central to the problem is to represent the genomic information in sequence form. In literature, we find several ways of representing genomic information in sequence form. The DNA chains can be considered as an ordered symbolic sequence written in four alphabet (ATCG), representing four nucleotides A: Adenine, C: Cytosine, T: Thymine, G: Guanine. They are classified by their chemical structures: A and G are purines, T and C are pyrimidines. The DNA macromolecules consists of two complimentary strands binded to each other by hydrogen bonds between (A, T) and (C,G), respectively.

The first approach to convert genomic information in numerical sequences was given by Voss [20], as binary sequences for each base, where 1 at position k indicates the presence of the base at the position, and 0 its absence. In [3], genomic sequence is represented by a mapping in which the complex numbers and their conjugates are assigned to each base of the nucleotide sequence: $a = 1 + j, t = 1 - j, c = -1 - j, g = -1 + j$ where a, t, c, g are the numbers assigned respectively to the bases A, T, C, G. Other relevant criteria to select the numerical values to represent genomic sequence are discussed in [1].Rushdi and Tuqan [18] proposed a genomic matrix based framework that comprises most of the mappings reported in literature as special cases and can allow a number of potential new mappings is proposed.

In functional sense, the regions of the DNA sequences can be classified as coding (exons) and non-coding (introns), that is, those carry the information for protein structures, and those that do not. It was shown in recent studies that the distribution of nucleotides A, T, C, G in a real DNA chain is a fractal distribution. Thus the fractal and multi-fractal tools like wavelet analysis can be applied effectively. It is reported with relevant details that the scaling in DNA chains does exist and this scaling is of a multi-fractal nature rather than a global one [12]. Long sequences of nucleotides look like random. The long-range correlations in DNA sequences were discovered in 1992 [13]. This inspired a lot of applications of multi-fractal and wavelet analysis to the nucleic acids primary structures.

The occurrence of a certain nucleotide in a certain portion of the DNA chain, labeled by a length parameter l, can be described as a random process $X(l,.)$ [2]. Thus, for the four letter ordered sequence, we deal with a probability space (Ω, U, P), with $\Omega = \{A, T, C, G\}$ and a family of four random processes $X_z = \{X_z(l,\omega); l \epsilon R, \omega \epsilon \Omega\}$, such that

$$X_z = \begin{cases} 1 \text{ if } \omega = z, \\ 0 \text{ otherwise} \end{cases}$$

We, in our work, rather adopt the generalized approach to digitalize the genomic information into a randomized sequence that will unify all the earlier representations and develop a wavelet based mathematical tool to analyze it in stochastic framework.

4. Mathematical Analysis

We start with symbolizing a genomic sequence of DNA chain of length N_s as

$X = \{X[i]; i = 1,2, \dots, N_s\}$.

Assign the values, $x_k = +1$, if purine is present or $X_k = -1$, if pyrimidine is present, for some position k in the proposed sequence. Also, as in [5], obtain a random sequence corresponding to the DNA chain of the same length:

$S = \{s[i]; i = 1,2, \dots, N_s\}$,

where for any position k, $1 \le i \le k$,

$s[k] = \sum_{i=1}^{k} x[i]$,

is a cumulative sum of the $x[i]$.

From [16], we get to know that for a completely random sequence, the path mapping gives a Brownian motion type signal constructed as

$s(t) = \sum_{i=1}^{t} u(i)$,

where $u(i) = \begin{cases} 1 \; purine \\ -1 \; pyrimidine \end{cases}$

Note that this random sequence has equal probability of $1/4$ for all the four nucleotides positioned on DNA chain and also that the distribution of nucleotides A, T, C, G in a real DNA chain has a fractal pattern. Further, the occurrence of a certain nucleotide in a certain portion of the DNA chain, labeled by a length parameter N_s, can be described as a random process $S(N_s, .)$ defined on a probability space, say, (Ω, B, P), with $\Omega = \{A, T, C, G\}$.

In addition, we introduce a measure known to be the Entropy measure that determines the randomness of the sequence. The first definition of entropy of a discrete information source

(Producing a discrete sequence) was introduced by Shannon [19], as $(S) = -\sum_{i=1}^{N} p_i \log p_i$, where p_i are the probabilities of the set of values that can take the sequence representing the DNA chain $S: \{s_1 . s_2, \dots, s_n\}$.

Another definition frequently used is the Renzy entropy [17], given by

$H_\propto(S) = \frac{1}{1-\propto} \log \sum_{i=1}^{n} p_i^{\propto}$

Here $H_\propto(S)$ is the Renzi entropy of order $\propto$, where $\propto \geq 0$ and $\{p_i\}$ are the signal probabilities as defined before. With the help this measure, we can detect the coding regions of genomic sequence.

With these formulations, we can proceed to give a wavelet treatment to the digitalized DNA sequence characterized by a random process.

From the definition, wavelet transform is a convolution of functions $f(t) \in L^2(R)$ with certain locally supported function $\psi(t)$ shifted and dilated, called analyzing wavelet. The choice of the analyzing wavelet is of vital importance as it dictates the representation and properties of the wavelet transform. Thus, the choice depends on the factors like the kind of data to be analyzed.

We have at hand a random process/sequence representing the DNA chain. It is reported that for a stationary random process $X(t)$, for which the mean value is independent of time and the auto-correlation depends only on time difference, its probability function is given by

$$F_x(X) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{\infty} exp\left[-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right] dx \quad (3)$$

More importantly, as noted earlier, the sum of random processes with arbitrary distributions results in a random variable with normal probability distribution.

The integral at (3) represents the normal probability distribution, which determines the probability that a random value of normal variate X will fall within some specified interval, say, $\mu - \sigma < x < \mu + \sigma$:

$$P[\mu - \sigma < x < \mu + \sigma] = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{\mu-\sigma}^{\mu+\sigma} exp\left[-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right] dx \quad (4)$$

Or for the standard normal variate Z , obtained by the transformation,

$Z = \frac{x-\mu}{\sigma}$,

$$P[-1 < z < 1] = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{z^2}{2}} dz. \quad (5)$$

The corresponding probability density function for the random variable, X, will be therefore

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (6)$$

where $-\infty < x < \infty$, and the parameters μ and σ, $-\infty \le \mu \le \infty$ and $\sigma > 0$.

Consequently, the probability density function for the standard normal variate representing the given random process X, will be

$$\psi(Z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \quad (7)$$

where $-\infty < z < \infty$.

We immediately see that as $Z \to \infty$, $\psi(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2} \to 0$. This ensures that $\psi(Z)$ representing a normal distribution, quickly decays to zero, thus, meeting one of the requirements for being a wavelet. We know that the normal probability distribution itself is a Gaussian function. But, the Gaussian itself cannot be used as an analyzing wavelet. However, since the derivatives of Gaussian functions belong to the space of square- integrable functions or L^2 functions, they

can be the candidates for being employed appropriately as the generic or the analyzing function in wavelet transform.

We give further rigorous justification below. We observe that $\psi(Z)$ which represents the normal probability distribution for the standard normal variate Z, is a non-negative function satisfying the conditions

$\int_{-\infty}^{\infty} \psi(Z)dZ = 1,$

$\int_{-\infty}^{\infty} Z\psi(\eta)dZ = 0,$

$\int_{-\infty}^{\infty} Z^2\psi(Z)dZ = 1$

Moreover, $\psi(Z)$ is at least n-times differentiable $(n \geq 1)$ and its $(n-1)th$ derivative satisfies

$lim_{n\to\pm\infty}\psi^{(n-1)}(Z) = 0.$

We can therefore obtain the first n derivatives of $\psi(Z)$. In particular, the first and the second derivatives are

$$\varphi^1(Z) = \psi'(Z) = \frac{Z}{\sqrt{2\pi}}e^{-\frac{Z^2}{2}},$$

$$\varphi^2(Z) = \psi''(Z) = \frac{(Z^2-1)}{\sqrt{2\pi}}e^{-\frac{Z^2}{2}} \quad (8)$$

$\varphi^1(Z)$ and $\varphi^2(Z)$ given by the expressions at (8), belong to a family of vanishing momenta wavelets of the Gaussians:

$$\varphi^n(Z) = (-1)^{n+1}\frac{d^n}{ds^n}exp\,(-Z^2/2)$$

for $n = 1\ and\ n = 2$. These wavelets are analytically tested for being employed as analyzing wavelets.

For the family of vanishing momenta wavelets, $\varphi^n(Z)$, the condition $\int dZ\, Z^m\psi(Z) = 0$, $\forall m, 0 \leq m < n; n \in Z$, holds good.

The Fourier image of this wavelet family is $\hat{\varphi}^n(\omega) = -\sqrt{2\pi}(-i\omega)^n e^{-\omega^2/2}$ and have zeroes of order n at $\omega = 0$. The normalization/admissibility constant C_{φ^n},

$$C_{\varphi^n} = 2\int_0^{\infty}\frac{|\hat{\varphi}^n(\omega)|^2}{|\omega|}d\omega = 2\,\pi\Gamma(n) < \infty.$$

Hence, the admissibility condition for $\varphi^n(Z)$ being a wavelet holds good.

It is thus justified that either $\varphi^1(Z)$ or $\varphi^2(Z)$ wavelets engendered from normal probability distribution, $\psi(Z)$, can be judiciously employed as analyzing/basic wavelet to wavelet transform the genomic sequence represented by a random process.

Wavelet transform of Random processes

For the given random process $\psi = \psi(x,t)$ defined on the probability space (Ω, B, P), for any function $\varphi: R \to C$ satisfying the admissibility condition, from (1), the continuous wavelet transform of the random process ψ can be obtained as

$$W_\varphi[\psi(x)](a,b) = \int_R \frac{1}{\sqrt{a}}\varphi^*\left(\frac{x-b}{a}\right)\psi(x,t)dx,$$
$$b \in R, a > 0, x \in \Omega. \quad (9)$$

Thus, to obtain CWT of the DNA sequence $S = \{S_t\}_{t\in R}$ with respect to the generic wavelet φ^n, (taking φ for simplicity) , let (Ω, B, P) be a probability space and let $S = \{S(t)\}_{t\in Z}$ a second order random process representing the DNA sequence for which,

$E\mid S(t)\mid^2 = \int_S \mid S_t \mid^2 dP(x) < \infty, \forall x \in R,$

where E denotes the mathematical expectation.

Notice that, S is jointly measurable and S_t is square integrable function for each t.

Thus, the CWT of the DNA sequence $S = \{S_t\}_{t\in R}$ with respect to the generic wavelet φ will be given by

$$W_\varphi[S(t)](a,b) = \int_R \frac{1}{\sqrt{a}}\varphi^*\left(\frac{t-b}{a}\right)S_t dt,$$
$$b \in R, a > 0, t \in \Omega \quad (10)$$

To study the matter quantitatively, we need to calculate correlations between wavelet coefficients at different scales. For this, we define wavelet covariance of the covariance function for all t, $R_s(u,v) = E\varphi^*(u)\varphi(v), u, v \in R$, as

$$R_W(a,b,c,d) = EW_a(b)W_b(d)$$

$$= \int_R \frac{1}{\sqrt{ac}}\varphi\left(\frac{u-b}{a}\right)\varphi^*\left(\frac{v-d}{c}\right)R_s(u,v)dudv \quad (11)$$

provided that the condition,

$E\left\{\int_R\left|\frac{1}{\sqrt{a}}\varphi^*\left(\frac{t-b}{a}\right)S_t dt\right|\right\}^2 < \infty$, holds good.

That is,

$$R_W(a,c,b-d) = \langle W_\varphi(a,b), W_\varphi(c,d)\rangle,$$

where the curly brackets mean the covariance

$$cov(W_1, W_2) = E\left\{\frac{(W_1 - E(W_1)(W_2 - E(W_2)}{\sqrt{DW_1.DW_2}}\right\}$$

where D is the dispersion and E is the mathematical expectation. We should note that for a random sequence, wavelet coefficients correlation function will coincide with that of random signal; if no, the structure of wavelet coefficients correlation function will be different.

For numerical evaluation of the transform integrl, we need to compute the Discrete wavelet transform (DWT). This is done by discretizing the expression (10) as follows.

Given a generic wavelet φ, the DWT of the DNA sequence, $S = \{S_t\}_{t\in R}$ with respect to φ is defined to be the discrete random field $W = \{W_\varphi S(j,k)\}_{j,k\in Z}$, where $W_\varphi S(j,k)$ is defined by

$$w_{j,k} = W_\varphi S(j,k) = \int_R S_t \varphi_{j,k}(t)dt \qquad (12)$$

provided the path integral is defined with probability one.

DWT results in a multi-resolution decomposition, in which at each level, the signal is decomposed in "approximations" and "details" coefficients.

In this analysis, a signal or the DNA sequence in the instant case, can be described through a linear decomposition as

$$s(t) = \sum_j \sum_k w_{j,k}\varphi_{j,k}(t), j,k \in Z,$$

where $w_{j,k}$ are the wavelet coefficients of the expansion, and $\varphi_{j,k}$ is a set of wavelet functions of t. Here, the wavelet coefficients $w_{j,k}$ constitute a discrete set, and the values of the coefficients are calculated according to (12)

$$w_{j,k} = \langle s(t), \varphi_{j,k}\rangle = \int_{-\infty}^{\infty} s(t)\varphi_{j,k}(t)dt \qquad (13)$$

The DWT obtains the decomposition of the signals$[n]$ into a set of orthonormal wavelets and their associated scaling function $\varphi_{j,k}$ that constitute a wavelet basis. These functions can belong to different wavelet families that are expressed by the functions $\varphi_{j,k}$ which can be generated by dilation and translation of a basic wavelet. These dilations and translations are discrete, and the indexes j and k are respectively related to these processes, that can be expressed as

$$\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j t - k): j,k \in Z. \qquad (14)$$

In this expression, the functions $\varphi_{j,k}$are dilated in a dydatic form (in powers of two), when varying the value of index j, and in analogous way translated when varying the index k. In this process, translation is associated with time resolution, and dilation provides scaling. Note that the wavelet function must satisfy the condition $\lim_{n\to\infty}|\varphi_{j,k}(t)| = 0$ implying decay and $\int_{-\infty}^{\infty}\varphi_{j,k}(t)dt = 0$ implying oscillations like wave function.

Further, for being φ an analyzing wavelet, in descretized set up, it must satisfy the conditions

$\sum_{j\in Z}|\hat{\varphi}(2^j\xi)|^2 = 1, \xi \in R$

and

$\sum_{j=0}^{\infty}\hat{\varphi}(2^j\xi)\hat{\varphi}^* 2^j(\xi + 2k\pi)) = 0,$

$\xi \in R,\ k \in 2Z + 1$,and $\|\hat{\varphi}\|_2 \geq 1,$ (15)

where $\hat{\varphi}$ denotes the Fourier transform of φ:

$\hat{\varphi}(\xi) = \int_R \varphi(x)e^{-ix\xi}\ dx$

In the instant case, φ being the derivatives of normal probability function belonging to the Gaussian family of vanishing momenta wavelets, indeed satisfies the conditions at (15).

Further, the DWT of S is a random field on the probability space, (Ω, B, P) that depends on φ, we must have

$\int_R|S_t(\omega)\varphi_{j,k}(t)|dt < \infty, \quad \omega \in \Omega$ (16)

Since, S has a finite second order moments, a generic condition which ensures that (13) is well defined and is also a second order sequence such that

$$\int_R \sqrt{\rho_S(u,u)}\ |\varphi(2^j u - k)|\, du < \infty \qquad (17)$$

where $\rho_S(r,s) = E(S_r, S_s^*), r,s \in R$.

Thus, the CWT and DWT given by (10) and (12) respectively are well defined and can be utilized to obtain the wavelet transform of the DNA sequence S ; which in turn can be plotted in ab-plane. This will reveal the nucleotide pattern and help locating periodicities of these patterns in the DNA sequence. The actual values can then obtained by plotting $|W_\varphi S(a,b)|$ over the space-time plane.

5. Conclusions

Many biological phenomena can be modeled by treating them as random processes; which can in turn be analyzed by the application of wavelet transform. The goal of wavelet analysis is to extract structural information from signal in the transform domain using appropriate form of Gaussian function as analyzing wavelet.

Transformation of the signal represented by the random process yields the wavelet coefficients, $W_\varphi[s(t)](a,b)$ at a particular scale and translation which tells us how well the signal s and the scaled and translated analyzing wavelet φ match. If the signal is similar to the scaled and translated analyzing wavelet, then the wavelet coefficient will have big magnitude. The wavelet coefficient also represents the degree of correlation between two functions at a particular scale and translation. It can be inferred that if the sequence is random, the wavelet coefficient correlation function will coincide with that of random signal, otherwise,

the structure of the wavelet correlation function will be different. This feature can be used to classify the nucleotide sequences and study their functional organization.

Considering the randomness of the signal representing genomic sequence, in our wavelet treatment, we judiciously employed the generic function engendered from normal probability distribution, thus, justifies the robustness of the proposed wavelet scheme. Moreover, the proposed modulated scheme is not case specific, but, can be appropriately employed in similar formulations in bioinformatics covering wide range of biological evolutionary processes.

References

1. Akhtar M. and J. Ambikairajah, "Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction" *IEEE J Select Topics Sign Proc* **2008,** 3: 310-321
2. Altaiski M., O. Morney and R. Polozov,"Wavelet analysis of DNA sequences" J. *Genetic Analysis*, **1996,** 12: 165-169.
3. Anastassiou D., "Genomic signal processing" *IEEE Sign Proc Mag* **2001,** 18: 8-20.
4. Belgarein N., R. Fourier and F. Bereksi-Reguij, "ECG based human authentication using wavelets and random forests", J. *Int. Jourl. On Cryptography and Information Security,* **2012,** 2(2): 145-165.
5. Berger J., and S. Mitra, "Visualisation and analysis of DNA sequences using DNA walks", *A new Journal of Franklin Institute,* **2004,** 341: 37-53.
6. Bhosale B. and A. Biswas, "Wavelet analysis of optical solitons and its energy aspects", J. *Mathematics in Engineering, Science and Aerospace,* **2012**, 3(1): 15-27.
7. Bhosale B. and A. Biswas, "Wavelet analysis of soliton interaction and its relation to probability distributions", J. *Nonlinear Studies*, **2012**, 19(4): 563-572.
8. Bhosale B. and A. Biswas, "Wavelet transform of soliton solution of Sine-Gordon equation as Random process". (submitted)
9. Bhosale B., A, Yildirim and A. Biswas, "Modelling space-time-varying systems for analyzing solitons", J. *Advanced Sciences, Engineering and Medicine,* **2012**, 4: 164-170.
10. Boumbarov O., Y. Velchev and S. Sokolov, "Personal Biometric identification based on ECG features", J. *Information Technologies and Control*, **2008,** 3(4): 11-18.
11. Camerlingo C, F Zenone, G M Gaeta, R Riccio and M Lepore, "Wavelet data processing of micro-Raman spectra of biological samples", J. *Meas. Sci. Technol., **2006**,* 17: 298- 305.
12. Gale J., R. Tubey, and J. D'Anna, "Localisation of DNA sequences of a replication of origin in rhodopsin gene locus of Chinese hamster cells", J. *Molecular Biology,* **1992,** 224: 343-358.
13. Kaneko A. and W. Li,"Long-range correlations and 1/f spectrum in non-coding DNA sequences", J. *Europhys. Lett.,* **1992,**17(7): 655-660.
14. Li Z., Q. Wang, and Y. Wu, "Wavelet analysis for random processes", J. *Mod. Phys. Lett. A*, **2001,** 16: 583-589.
15. Lio P., "Histidine biosynthetic pathway and genes", J. *Bioinformatics,* **2003**, 19(1): 2-9.
16. Peng C., S. Byldyrev and H. Stanley,"Long-range correlations in nucleotide sequences" J. *Nature,* **1992,** 356: 168-171.
17. Rényi A.," On measures of information and entropy", *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* **1960,** 547-61
18. Rushdi A. and J. Tuqan, "The role of the symbolic-to-numerical mapping in the detection of DNA periodicities" *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS '08* **2008**, 10-14.
19. Shannon C.," A Mathematical Theory of Communication", J. *Bell Sys Techn,* **1948**, 27: 623-656, 1948
20. Voss R., "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", J. *Phy. Rev. Lett* **1992,** 68: 3805-3808
21. Young R. K., "Wavelet theory and its applications", *Fourth Printing, Kluwer Academic Publishers*, **1995**.
22. Yu B. and Y. Zhang, "A Simple Method for Predicting Trans-membrane Proteins Based on Wavelet Transform", *Int J Biol Sci* **2013**, 9(1): 22-33.

4/28/2013