Digital Signal Processing in the Analysis of Genomic Sequences

Juan V. Lorenzo-Ginori*[,1], Aníbal Rodríguez-Fuentes[1], Ricardo Grau Ábalo[2] and Robersy Sánchez Rodríguez[3]

[1]Centro de Estudios de Electrónica y Tecnologías de la Información, Facultad de Ingeniería Eléctrica, Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuaní Km. 5 ½, 54830 Santa Clara, Villa Clara, Cuba; [2]Centro de Estudios de Informática, Facultad de Ingeniería Eléctrica, Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuaní Km. 5 ½, 54830 Santa Clara, Villa Clara, Cuba; [3]Instituto Nacional de Investigaciones en Viandas Tropicales, (INIVIT), Biotechnology Group, Santo Domingo, Villa Clara, Cuba

Abstract: Digital Signal Processing (DSP) applications in Bioinformatics have received great attention in recent years, where new effective methods for genomic sequence analysis, such as the detection of coding regions, have been developed. The use of DSP principles to analyze genomic sequences requires defining an adequate representation of the nucleotide bases by numerical values, converting the nucleotide sequences into time series. Once this has been done, all the mathematical tools usually employed in DSP are used in solving tasks such as identification of protein coding DNA regions, identification of reading frames, and others. In this article we present an overview of the most relevant applications of DSP algorithms in the analysis of genomic sequences, showing the main results obtained by using these techniques, analyzing their relative advantages and drawbacks, and providing relevant examples. We finally analyze some perspectives of DSP in Bioinformatics, considering recent research results on algebraic structures of the genetic code, which suggest other new DSP applications in this field, as well as the new field of Genomic Signal Processing.

Keywords: Digital Signal Processing, genomic sequences, coding regions.

INTRODUCTION

Digital Signal Processing (DSP) is an area of science and engineering that has developed during the past 40 years as a result of the constant evolution of computer science and technology. DSP comprehends the representation, transformation and manipulation of digital signals as well as the information associated to them. In this context, signals are usually physical magnitudes that vary in time or space, and digital signals are those represented as sequences of numbers, as in the case of time series.

The discipline of DSP uses a set of mathematical tools to analyze and process signals, among them can be mentioned the Discrete Fourier Transform, the Z transform, Digital Filters, Parametric Models, the Wavelet Transform, Correlation Functions and others. When considering the informational content of signals, other concepts from Information Theory such as entropy and mutual information are also used.

A key concept in DSP is the possibility of representing the signals in the frequency domain making use of the Discrete Fourier Transform. This representation leads to some important signal properties that are not revealed in the time domain, which are associated to their frequency spectrum.

In the case of the genomic sequences, these have been represented mathematically by character strings of symbols from a size-4 alphabet consisting of the letters A, T, G and C, which represent each one of the nucleotide bases. In the case of proteins, the alphabet size is 20, corresponding to the possible amino acids. The possibility of finding a wide application of DSP techniques to the analysis of genomic sequences arises when these are converted appropriately into numerical sequences, for which several rules have been developed. Notice that genomic signals do not have time or space as the independent variable, as occur with most physical signals.

This paper is organized in the following way. Firstly an overview of the main DSP algorithms used in applications to genomic sequence analysis is shown: digital filters, the Discrete Fourier Transform (DFT), the Short-Time Fourier Transform (STFT), parametric models (AR, MA, ARMA), Wavelet Transform and the Information Theory concept of entropy. Hidden Markov Models can be considered also as a DSP tool, but this topic will not be covered, as there is a recent comprehensive review article by De Fonzo *et al.* [1]. Then the numerical representation of genomic sequences is presented. This allows the application of DSP tools to study genomic sequences. After this, a review of the major applications of DSP to the analysis of genomic sequences is realized, such as identification of protein coding DNA regions, identification of reading frames, location of splice sites and others. We finally review the perspectives of DSP in this field, considering recent research results on algebraic structures of the genetic code and the new field of Genomic Signal Processing.

MAIN DSP ALGORITHMS EMPLOYED IN THE ANALYSIS OF GENOMIC SEQUENCES

In this section a synthetic overview of the main DSP algorithms that have been used in the analysis of genomic sequences is presented. There are excellent books on DSP theory by Oppenheim and Schafer [2] and Proakis and Manolakis [3].

*Address correspondence to this author at the Centro de Estudios de Electrónica y Tecnologías de la Información, Facultad de Ingeniería Eléctrica, Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuaní Km. 5 ½, 54830 Santa Clara, Villa Clara, Cuba;
E-mail: juanl@uclv.edu.cu

1574-8936/09 $55.00+.00

A) Digital Filters

A digital filter is a particular class of discrete system capable of realizing some transformation to an input discrete numerical sequence. There are different classes of digital filters according to the properties of their input-output relationships, as for example linear, nonlinear, time-invariant or adaptive. The basic, frequency selective digital filters, are linear and time-invariant (LTI) discrete systems.

Digital filters are characterized by numerical algorithms that can be implemented in any class of digital processors. In particular, LTI digital filters can pertain to one of two categories, according to the duration of their response to the impulse, or Dirac delta function, when it is used as the input signal: infinite (IIR) or finite (FIR) impulse response. The input-output relationships for IIR digital filters are characterized and implemented algorithmically through a finite difference equation of the form

$$\sum_{k=0}^{N} a_k y[n-k] = \sum_{k=0}^{M} b_k x[n-k], \quad (1)$$

where $x[n]$ and $y[n]$ are the input and output numerical sequences respectively, a_k and b_k are numerical coefficients, n is the sample index, and k is an integer delay with maximum values N and M for the output and input sequences respectively. On the other hand FIR digital filters are characterized by a discrete convolution operation of the form

$$y[n] = \sum_{m=0}^{N-1} h[m]x[n-m] \quad (2)$$

In this equation, $h[m]$ is the impulse response of the filter, which has a length of N samples. The bilateral Z transform operator is defined as

$$Z\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n]z^{-n} \quad (3)$$

where z is a complex variable. When this operator is applied to equations (1) or (2), the system transfer function in the Z-transform domain is obtained. The system transfer function relates the input and output sequences $x[n]$ and $y[n]$, through their respective Z transforms $X[z]$ and $Y[z]$. The transfer function has the general form

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} \quad (4)$$

The transfer function $H(z)$ for this class of systems is a ratio of polynomials in the complex variable z and has a convergence region associated to it, which is closely related to the positions of its poles in the complex Z plane. A property of the transfer function of LTI systems is that the complex exponential sequences of the form

$$x[n] = e^{i\omega n}$$

where i is the imaginary unit, are eigenfunctions of these systems, and this lead to the concept that these systems have an associated *frequency response*, which can be obtained by equating $z = e^{i\omega}$ in equation (4), i.e.

$$H(e^{i\omega}) = H(z)\Big]_{z=e^{i\omega}} \quad (5)$$

The presence of the imaginary unit in the exponent implies that $H(e^{i\omega})$ is a complex function in the frequency domain, whose frequency response is usually expressed as a magnitude response together with a phase, or angle response. The system transfer function is periodic in ω (emphasizing this periodicity is the reason for using $e^{i\omega}$, instead of simply ω, as the argument of H), and it is usually plotted for its values in the main interval $-\pi \leq \omega < \pi$. An example of a sharp resonance peak in the magnitude response of an IIR filter is shown in Fig. (**1**), together with the corresponding phase response. The sharp magnitude peak means a high selectivity in frequency. The phase response of this filter is highly nonlinear (lower graph) and this nonlinearity tends to produce a high signal distortion.

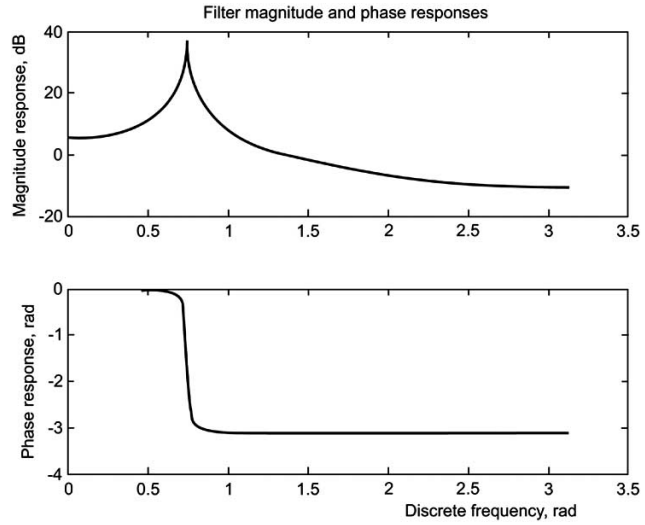

Fig. (1). Frequency response in magnitude and phase of an IIR system exhibiting a sharp peak in the magnitude response.

A variety of digital filter design techniques allow to obtain any desired magnitude response with frequency selectivity properties, whereas it is desired that the phase response be a linear function of ω, in order to have low distortion. According to the frequency interval (band) transmitted, the magnitude of the basic ideal prototype filter frequency responses, can be *lowpass*, *highpass*, *bandpass* and *bandstop.* A combination of these responses leads to a *multiband* filter. The typical ideal frequency responses (in magnitude) of the prototype filters are shown in Fig. (**2**). These ideal responses can be only approximated in practical filters, where better approximations in general are obtained by increasing the order of $H(z)$, which means a higher computational complexity of the digital filters.

Constant magnitude response together with perfect linearity in the phase response is the condition for signal trans-

mission without distortion through a filter in the desired frequency band. IIR digital filters have in general a nonlinear phase response, that depends on the design method employed. On the other hand, a property of FIR digital filters is that they can exhibit a perfect linear phase response under certain conditions of symmetry in their impulse response. This has been a motivation for the use of digital FIR filters in many applications.

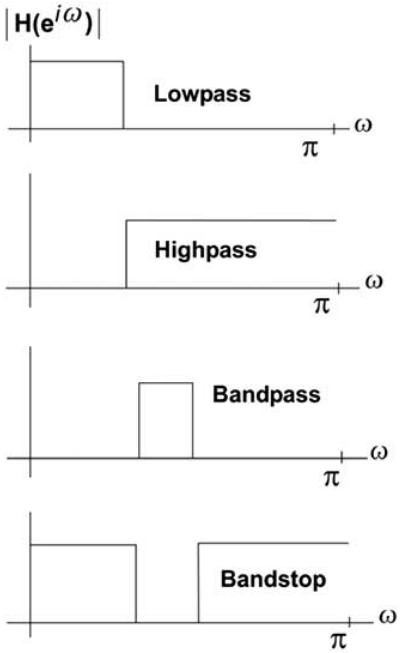

Fig. (2). Frequency response in magnitude for the prototype ideal filters: lowpass, highpass, bandpass and bandstop.

B) Discrete Fourier Transform

The Discrete Fourier Transform is a mathematical operation that transforms one discrete, limited (*finite*) N duration function into another function, according to

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}nk}, \quad 0 \le n, k \le N\text{-}1 \qquad (6)$$

The function $X[k]$ is the Discrete Fourier Transform (*DFT)* of the sequence $x[n]$ and constitutes the *frequency domain* representation of $x[n]$, which is usually (or conventionally considered) a function in the time domain. The Discrete Fourier Transform only evaluates the frequency components required to reconstruct the finite segment of the sequence that was analyzed. In general, the DFT is a function in the complex domain as a result of the complex exponential in the right side of equation (6), and for the particular case of real sequences, it will be a sequence of complex numbers of the same length as $x[n]$. The DFT is usually represented in terms of the corresponding magnitude and phase functions that constitute the *frequency spectrum* of the sequence $x[n]$.

The Discrete Fourier transform is a very useful tool, because it can reveal periodicities in the input data as well as the relative intensities of these periodic components. An example of the magnitude and phase graphs of the 64-points DFT for a sum of two pure sinusoids at discrete frequencies $2\pi/14$ and $4\pi/15$ is shown in Fig. (**2**). Each discrete value of the DFT is usually called a DFT coefficient.

The DFT, however, suffer from three important drawbacks as a tool for spectral analysis: a) Spectral leakage, which means the presence of energy in zones where the spectrum should be zero (this is clearly seen in Fig. (**3**): two pure frequencies are analyzed while many nonzero samples are obtained in the spectrum at other frequencies); b) the frequency response of the DFT coefficients is not constant with frequency ("picket-fence" effect), and c) the spectral resolution, or ability to separate frequency lines that are close in frequency, depends inversely upon the length of the sequence in the time domain. This means that the DFT cannot distinguish appropriately close spectral components for time signals of short duration. Multiplying the time signals by special weighting functions called windows, and controlling the signal length, can help in overcoming these limitations in some extent.

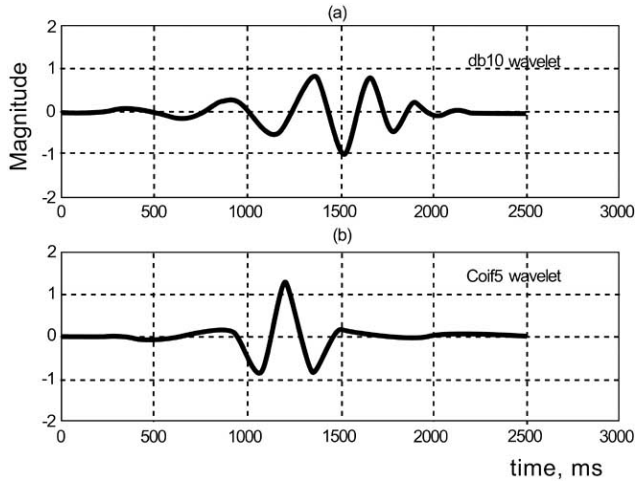

Fig. (3). Example of DFT frequency spectrum (magnitude and phase) for two sinusoids closely spaced in frequency. Frequency axis is normalized to f_s/N, where f_s is the sampling frequency and N the number of samples in the sequence (64 in this example).

Using the DFT for spectral analysis of random signals (or *stochastic processes*) require certain considerations to obtain a statistically valid result.

For stationary random signals, a commonly employed procedure to obtain a power spectral density (PSD) function in the frequency domain is the Welch's modified periodograms method. The PSD function is obtained in this case by calculating the mean value of the squared DFT coefficients at each frequency value, for adjacent and usually overlapping *windowed* signal segments. The measure obtained in this way is a consistent estimate of the power spectrum. A typical spectrum obtained by the Welch's method, for a pure sinusoid embedded in white Gaussian noise, is shown in Fig. (**4**). Notice the peak that corresponds to the sinusoid, whose magnitude is significantly greater than the noisy background.

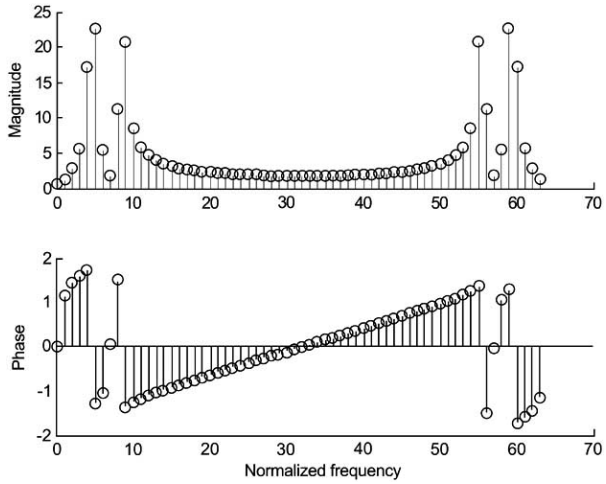

Fig. (4). An example of PSD spectrum obtained through Welch's method, for a sinusoid embedded in white, Gaussian noise.

In the case of non-stationary signals, The Short Time Fourier Transform (STFT) is an algorithm frequently used for the DFT-based spectral analysis. In the STFT, the time signal is divided into short segments (usually overlapped) and a DFT is calculated for each one of these segments. A three dimensional graph called *spectrogram* is obtained by plotting the squared magnitude of the DFT coefficients as a function of time. This squared magnitude is usually represented by the brightness of the graph, as shown in Fig. (**5**).

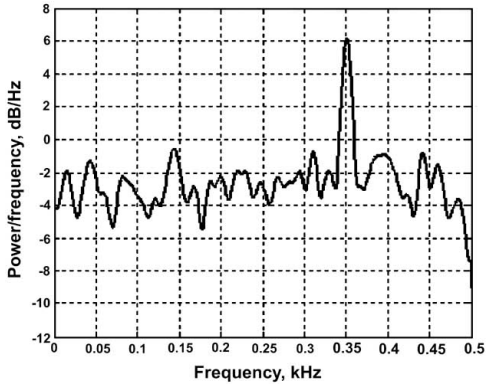

Fig. (5). Spectrogram of a harmonic signal whose frequency varies linearly with time ("linear chirp").

An important special case of the STFT is the Gabor Transform, in which a Gaussian weighting window is applied to the analyzed time sequence. This procedure allows obtaining a better simultaneous resolution in time and frequency.

C) Spectral Analysis Using Parametric Models

Parametric spectral analysis is a method that can be used in many cases with some advantages over the non-parametric methods. Its advantages rely in that it is possible to obtain a parametric description of the second-order statistics of a random sequence, by assuming a certain production model for it. A comprehensive analysis of such methods is given in Stoica and Moses [4].

Spectral analysis using parametric methods does not suffer from the limitations in spectral resolution that characterize the DFT-based methods, because they do not imply a windowing (segment selection) process.

The mathematical expression of the PSD function of a random sequence is described in this case in terms of the model parameters, and the variance of a white (constant PSD) random noise process used as the input signal of the model. In consequence, the values to be computed in this method are the parameters of the model and the variance of the input process.

The general expression for the transfer function of the model in parametric spectral analysis is analogous to that of a digital filter as shown in equation (3), which is expressed as the ratio of polynomials in the complex variable *z*

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^{q} b_k z^{-k}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{7}$$

to which corresponds the equation in finite differences

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + \sum_{k=0}^{q} b_k w[n-k] \tag{8}$$

in which $w[n]$ is the input sequence and the observed data $x[n]$ represent the model's output. Equations (7) and (8) are related through the *Z* transform operator shown in equation (3). The PSD function is obtained from (7) using (5) to obtain the model's frequency response, and is given by

$$\Gamma_{xx}(\omega) = \left|H(e^{i\omega})\right|^2 \Gamma_{ww}(\omega) \tag{9}$$

In equation (9) $H(e^{i\omega})$ is the frequency response of the model, while Γ_{ww} and Γ_{xx} are respectively the PSD functions of the corresponding input and output signals. For a white-noise input,

$$\Gamma_{xx}(\omega) = \left|H(e^{i\omega})\right|^2 \sigma_w^2 \tag{10}$$

where σ_w^2 is the input noise variance.

According to the characteristics of the PSD for the analyzed random sequence there are three types of parametric models:

- Autoregressive (AR) models, corresponding to the particular case $\{b_k = 0\}$ for $k > 0$, resulting in an all-pole transfer function.
- Moving average (MA) models, which correspond to $\{a_k = 0\}$, resulting in an all-zero transfer function.
- Autoregressive, moving average (ARMA) models, which is the general case in which there are poles and zeros in the model's transfer function.

There is equivalence between the three types of models if the order is selected appropriately, i. e., a process which is inherently AR of a certain order, can be described by an MA model of higher order. However, AR models are more used because of the relative simplicity in calculating the model's parameters through the Yule-Walker equations. Fig. (**6**) shows the PSD curve for a typical AR spectrum.

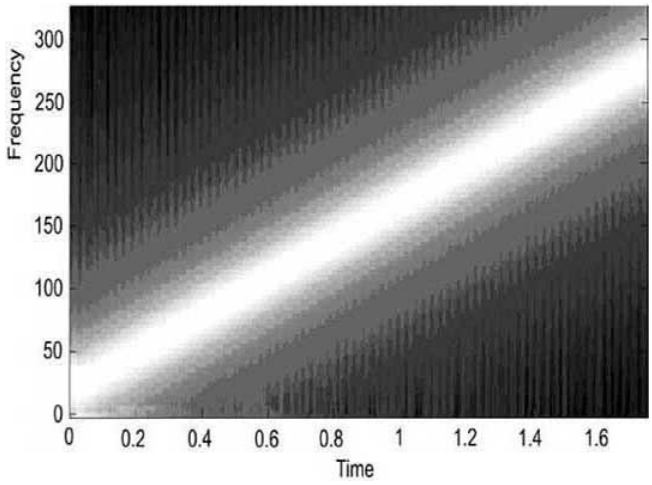

Fig. (6). A typical PSD function obtained for an AR model, exhibiting two peaks corresponding to two pairs of complex conjugate poles in the model's transfer function.

D) Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a mathematical tool that can be used very effectively for non-stationary signal analysis. There is a great amount of literature on DWT, see for example Burrus *et al.* [5].

In DWT analysis, a signal $x(t)$ can be described through a linear decomposition as

$$x(t) = \sum_k \sum_j a_{j,k} \psi_{j,k}(t) \tag{11}$$

In this equation $j,k \in Z$ are integer indexes, $a_{j,k}$ are the wavelet coefficients of the expansion, and $\psi_{j,k}$ is a set of wavelet functions in *t.* Notice that the wavelet coefficients $a_{j,k}$ constitute a discrete set, and that the coefficient's values are calculated according to

$$a_{j,k} = < x(t)\psi_{j,k}(t) > = \int_{-\infty}^{+\infty} x(t)\psi_{j,k}(t)\, dt \tag{12}$$

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

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right), j,k \in Z \tag{13}$$

In Eq. (13) the functions $\psi_{j,k}$ are dilated in a dyadic form (in powers of two), when varying the values of the 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, a concept closely related here to frequency resolution.

Wavelet functions must satisfy the conditions

$$\lim_{t\to\infty} \left|\psi_{i,j}(t)\right| = 0 \tag{14}$$

and

$$\int_{-\infty}^{\infty} \psi_{i,j}(t)dt = 0 . \tag{15}$$

In these conditions, (14) implies decay, and (15) implies oscillations like a wave function. Fig. (**7**) shows examples of wavelets functions that are well described in the literature.

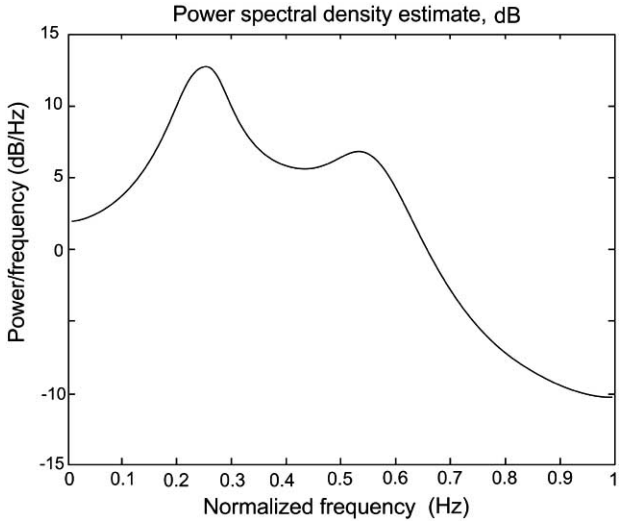

Fig. (7). Examples of wavelets: (a) Daubechies *Db10*, (b) Coiflet *Coif5*.

The DWT, for which an algorithm called Fast Wavelet Transforms (FWT) allows a very efficient calculation, plays currently a central role in many DSP applications. The result of the DWT is a multi-resolution decomposition, in which at each level the signal is decomposed in "approximation" and "detail" coefficients. This decomposition is realized through a process that is equivalent to lowpass and highpass filtering for the approximation and for the details respectively, using special digital filters called "Quadrature Mirror Filters" (QMF.) There are two types of QMF filters: the lowpass *scaling* filter *h*, and the highpass *wavelet* filter *g*. The *g* filter is equivalent to the *h* filter reversed in time and alternating the signs of its coefficients. DWT decompositions can be depicted by a tree structure as shown in Fig. (**8**), where approximation and detail coefficients are represented. Each one of the *J* decomposition levels corresponds to a certain dilation *j*, whereas the index *k* determines the corresponding translations. The DWT can be also extended to non-orthogonal decompositions.

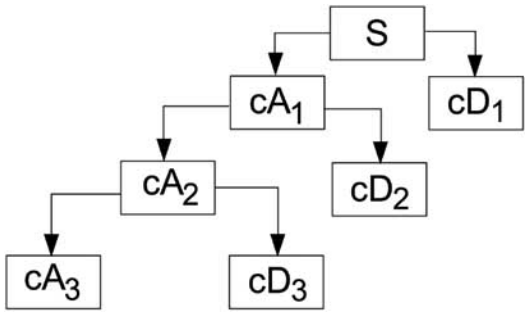

Fig. (8). Approximation and Detail coefficients in a tree structure for a DWT three-level decomposition. S is the original signal, cD_i and cA_i stand respectively for detail and approximation coefficients at level *i*.

E) Entropy Measures

Entropy measures are another example of a signal processing concept that has been used in genomic sequence analysis.

The concept of entropy is used in signal analysis as a measure of randomness. The first definition of the entropy of a discrete information source (producing a discrete sequence) was introduced by Shannon [6] as

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i \tag{16}$$

where p_i are the probabilities of the set of values that can take the sequence X, $\{x_1, x_2, \ldots, x_n\}$.

Another definition frequently used is the Rényi entropy [7], given by

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_i^\alpha \tag{17}$$

Here $H_\alpha(X)$ is the Rényi entropy of order α, where $\alpha \geq 0$, and $\{p_i\}$ are the signal probabilities as defined before.

F) Final Remarks

Although in this section the more frequently used DSP techniques were overviewed, it is important to notice that there are other various important techniques that in some cases have been used in the Bioinformatics field, such as different transforms (Cosine, Sine, Walsh-Hadamard, Hilbert), fractal analysis, and others.

NUMERICAL REPRESENTATION OF GENOMIC SEQUENCES

The first approach to convert genomic information in numerical sequences was given by Voss [8] with the definition of the indicator sequences, defined as binary sequences for each base, where 1 at position *k* indicates the presence of the base at that position, and 0 its absence. For example, given the DNA sequence

ACTTAGCTACAGA…

The binary indicator sequences X for each base A, T, C and G are respectively:

$X_A[k]$ = 1000100010101…

$X_T[k]$ = 0011000100000…

$X_C[k]$ = 0100001001000…

$X_G[k]$ = 0000010000010… (18)

The main advantages of the indicator sequences are their simplicity, and the fact that they can provide a four-dimensional representation of the frequency spectrum of a character string, by means of computing the DFT of each one of the indicator sequences. This dimensionality can be reduced to three through the *Z* curves [9, 10] and the tetrahedron [11] methods.

Another relevant numerical representation of genomic sequences is a mapping in which a complex number is assigned to each base of the nucleotide sequence. In this case, these complex numbers are appropriately selected to provide useful properties of the numerical sequences. One of such properties is obtained by assigning complex conjugate complex numbers to the base pairs A, T and C, G. In this case all palindromes will have conjugate symmetric numerical sequences. This lead to the generalized linear phase described by Anastassiou [12]. A simple example of such mapping, used in this reference is

$$a = 1 + j,\ t = 1 - j,\ c = -1 - j,\ g = -1 + j \tag{19}$$

where a, t, c and g are the numbers assigned respectively to the bases A, T, C and G.

A more complete mapping that gives the representation of all IUPAC nucleotide classes comprising single nucleotides, doublets, triplets and quadruplets is given by Cristea *et al.* in [13] and applied in [14] to analyze the variability of pathogens' genomes.

Other relevant criteria to select the numerical values to represent genomic sequences are discussed by Akhtar *et al.* [15]: equal magnitudes, equidistance, compactness of the representation and easiness to use various mathematical tools. Other examples of representations that have been used are

$t = 0,\ c = 1,\ a = 2,\ g = 3$ in [16], which correspond to a Galois field assignment, and

$a = 1.5,\ t = -1.5,\quad c = 0.5,\ g = -0.5$

used in [15]. Notice that the latter shows the complementary property, in the same way as in the complex assignment (19). Rushdi and Tuqan [17] proposed a generic matrix based framework that comprises most of the mappings reported in the literature as special cases and can allow a number of potential new mappings.

A representation of genomic sequences by means of quaternions was introduced by Brodzik and Peters in [18], which allows using the quaternionic Fourier Transform for pattern detection in DNA sequences.

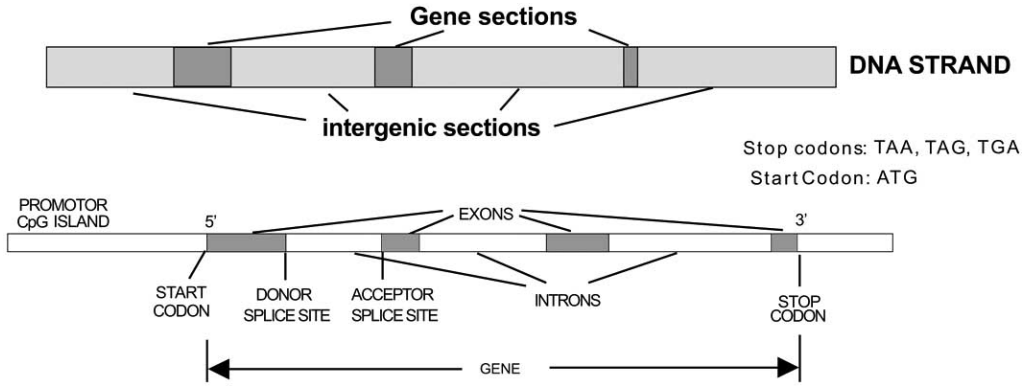

Fig. (9). Diagram of a protein-coding DNA region and of a gene from an eukaryotic DNA, showing different characteristic points whose detection is a source of applications of DSP techniques.

A relationship between the numerical assignment to the nucleotides and to the amino acids has been established through FIR digital filtering in [12].

APPLICATIONS OF DSP IN THE ANALYSIS GENOMIC SEQUENCES

Digital Signal Processing applications to Bioinformatics started in recent years in which great attention was put to the problem of genomic sequence analysis. Fig. (**9**) depicts a protein-coding DNA region and, in particular, a gene from an eukaryotic genome, indicating the introns and exons and the points where the gene begins (start codon), its end (stop codon), donor splice sites (transition from an exon to an intron), donor splice sites (transition from intron to exon) and a CpG island (a region rich in CG pairs that may promote gene function). Detecting all these places in a genomic sequence is a source of application for DSP techniques.

One of the main motivations to introduce DSP in this field was the find of hidden periodicities or oscillating patterns in the genomic sequences, which were described by Trifonov in [19] as 3, 10.5, 200 and 400-base periodicities. Among them, the three-base periodicity was found to be a characteristic of the protein-coding regions in both prokaryotic and eukaryotic sequences.

The 3-periodicity is explained in more detail by Tuqan and Rushdi [20] as related to the *codon bias*. Consider a genomic sequence analyzed through a rectangular window with three-base length, that is displaced along the entire sequence in three-base length intervals. The relative number of occurrences of base l in the k^{th} (k=0, 1, 2) position of the codon in the specific window positions, reveals that there is an unbalance of the abundance of base l in codon position k with respect to the average frequency of occurrence of base l in the three possible codon positions. This phenomenon is reflected in the frequency spectrum of the DNA sequence as a spectral line exactly at $N/3$ in the DFT, N being the DFT length. Another contribution to explain the three-base periodicity was made by Sánchez and López-Villaseñor [21] through the concept of *same-phase triplet clustering*, a condition in which a triplet appears several times in one phase with no interruptions by the two other possible phases.

Detection of Protein-Coding Regions Through Spectral Analysis and the 3-Periodicity Property

A number of authors have devised algorithms to detect protein coding regions in genomic sequences by finding regions exhibiting a three-periodicity. Vaidyanathan and Yoon [22] applied to the indicator sequences an anti-notch IIR digital filter with a sharp narrow band centred at $\omega_0 = 2\pi/3$, with the purpose of detecting the period 3 component. They showed also lattice and multistage implementations, as well as an equivalent DFT approach to this problem. The concept that DNA sequences have an $1/f$ power spectrum that can be considered as a noisy background, is used to argue that the window length used to calculate the DFT should be long enough, typically a few hundreds bp as 351, to a few thousands, in order that the 3-periodicity dominates the noise background. A typical result is given in Fig. (**10**), where comparison to a threshold is usually employed to determine the detected regions.

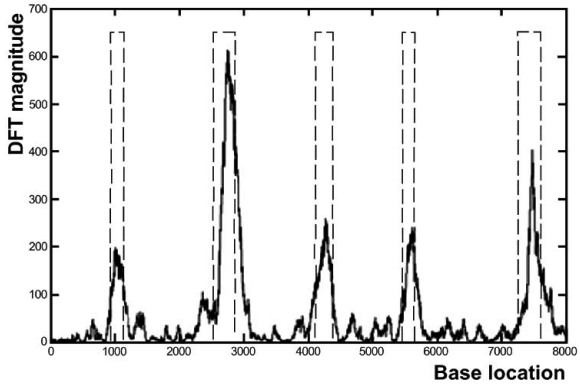

Fig. (10). Detection of 3-periodicity regions using DSP. Typical plot in which noticeable peaks correspond to coding regions.

Another digital filtering approach, the polyphase Filtered DNA spectrum, was presented by Tuqan and Rushdi [23].

Fox and Carreira [24] introduced a method in which only one digital filter operation is required, followed by a quadratic windowing operation which produces a signal that has almost zero energy in the non-coding regions, improving the effectiveness of the method.

The DFT approach to find the 3-periodicity regions in genomic sequences has been used by various authors. Afreixo *et al.* [25] analyze several methods for the Fourier analysis of symbolic data oriented to DNA sequences, considering different approaches as the indicator sequences, vector and symbolic correlation sequences and spectral envelope.

Tiwari *et al.* [26] presented an early study of the application of DFT analysis for gene prediction, where an experimental study for a variety of genomic sequences was performed. Another early example can be found in Yan *et al.* [27], based in the format of the Z curve. Anastassiou [12] used the DFT and the STFT spectrograms to analyze the indicator sequences and introduced an optimized spectral content measure to improve the discriminating properties of the method. Datta *et al.* [28] used the DFT to find the 3-periodicity regions and formalized mathematically some properties of the DNA sequences. A fast DFT based gene prediction algorithm and a DFT based splicing algorithms are presented by these authors in [29, 30].

Isaac *et al.* [31] showed FTG, a web server to predict genes based on DFT techniques, which allows rapid visualization by providing an output in GIF format. Stoffer *et al.* [32] presented a study on the local spectral envelope used together with a dyadic-tree based adaptive segmentation for gene detection. This work considers DNA as a piecewise stationary series, and provide a thorough mathematical foundation for its analysis.

Epps *et al.* [33] developed an integer period DFT for biological sequence processing that has some advantages in detecting DNA periodicities. Rushdi and Tuqan [34] analyzed other trigonometric transforms as the discrete cosine transform (DCT), the discrete sine transform (DST) and the discrete Hartley transform (DHT), to find periodicities in DNA sequences. They showed also a unified multirate DSP model based on these transforms.

Berger *et al.* [35] analyzed the power spectrum of the genomic sequences using the Warped DFT and the Walsh Hadamard Transform to improve the effectiveness in detecting periodicities. Rodríguez-Fuentes *et al.* [36] introduced computational improvements in using the STFT to analyze genomic sequences.

The phase of the DFT has been used as well in detecting coding regions. Kotlar and Lavner [37] introduced the *Spectral Rotation Measure*, deriving a method in which the DFT phase is computed at the 1/3 frequency for the binary sequences for A, T, C, and G. Experimental analysis of the genes of *S. cerevisiae* and other organisms showed a distribution of the phase in a bell-like curve around a central value in all four nucleotides, and a nearly uniform distribution in the non-coding regions, allowing to define measures to identify coding regions based on this phase property. Rushdi and Tuqan [38] derived the *filtered spectral rotation measure* based on the polyphase filtered DNA spectrum introduced in [23], as an alternative measure to detect coding regions.

Yin and Yau [39] introduced an algorithm called Exon Prediction *via* Nucleotide Distributions (EPND), which combines the information from the peak at the N/3 frequency in the DFT and the frequencies of occurrence of the nucleotides in the three codon positions (position asymmetry measure) obtaining an improvement of the effectiveness in the detection of coding regions.

Akhtar *et al.* [40] showed an optimization of the period-3 methods taking into account both computational complexity and the relative accuracy of gene prediction. In this work, a paired and weighted spectral rotation (PWSR) measure previously defined by the authors was employed. This study used as additional information the statistical property of eukaryotic sequences by which introns are rich in nucleotides 'A' and 'T' whereas exons are rich in nucleotides 'C' and 'G'.

At this point, it is worth to mention that other studies like that of Xing *et al.* [41] reveal that the PSD itself does not provide sufficient resolving power to detect periodic signals in short coding sequences, and consequently other approaches in addition to the DFT have been used for this purpose.

Autoregressive modeling of DNA sequences was addressed by Chakravarthy *et al.* [42] who presented a model in which AR parameters are used as features. The AR residual error analysis shows a high specificity of coding DNA sequences, and the analysis based in AR features was useful in distinguishing between coding and non-coding DNA sequences. The AR model was very specific to the coding DNA sequences, and its specificity increased with increasing model orders. Rao and Shepherd [43] addressed the problem of detecting 3-periodicity in short genomic sequences based on the AR technique, in an effort to take advantage of the inherent improved frequency resolution of the AR models.

Akhtar *et al.* [44] presented an autoregressive modelling for the classification of genomic sequences, that provides a compact multi-dimensional feature that characterize the short term spectrum. The AR feature was also combined with a time-frequency hybrid (TFH) feature composed by the PWSR measure and the time-domain average magnitude difference function (AMDF). A Gaussian mixture model classifier was employed and showed improved recognition capabilities. Another approach based on Singular Value Decomposition was presented by the same authors in [45]. Akhtar [46] also presents a comparison between time and frequency domain techniques to detect short coding regions and show some advantages of the former.

Cristea *et al.* [47] address the detection of nucleotide sequences using a two step procedure comprising a Principal Components Analysis (PCA) stage, which retains only the high variance components of the input signal, and a feed-forward Artificial Neural Network (ANN), which performs the prediction. It is shown that the PCA stage performs an approximate DFT, passing from the time (space) domain to the frequency domain, and the ANN implements the inverse DFT, generating the estimate of the next sample of the sequence in the time (space) domain. Rodríguez-Fuentes *et al.* [48] used a combination of DSP approaches to detect coding regions in genomic sequences and showed the advantages of the combined method over the individual ones. Gunawan *et al.* [49] introduced a signal boosting technique to enhance

the coding region and improve the likelihood of its correct identification. The authors claim that when using this method together with ANN classification, the ratio of coding to non-coding energy is almost doubled.

Reading frame identification is an important issue in the detection of coding regions. This topic has also received attention from the DSP point of view. Anastassiou [12] and Kotlar and Lavner [37] presented algorithms for this purpose, which make use of the phase properties of the weighted transformed indicator sequences and showed good results.

Detection of Coding Regions and Other Applications Using an Information Theory Approach

The concept of entropy as it is used in Information Theory has been employed as well to detect coding regions. Román-Roldán *et al.* [50] defined a complexity measure, based on the entropic segmentation of DNA sequences into homogeneous domains. Bernaola-Galván *et al.* [51] introduced a computational approach to finding borders between coding and non-coding DNA, in which the sequences are described by a 12-letter alphabet, capable of representing the differential base composition at each codon position, and the borders are searched by means of an entropic segmentation through the Jensen-Shannon measure. The method showed to be very accurate and does not require prior training.

Nicorici and Astola [52] extended this approach by applying recursively an entropic segmentation method on DNA sequences using 12 and 18-symbol alphabets to capture the differential nucleotide composition in codons as well as the differential stop-codon in all phases of both strands. The method uses the Jensen-Rényi divergence measure, nucleotide statistics and stop codon statistics in the two DNA strands in order to find the borders between the coding and non-coding regions. This method does not require prior training and showed good results.

Multihac *et al.* [53] used a more theoretical information theory perspective to interpret the amount of information carried by the binding site patterns in the DNA molecules, using maximum entropy methods. Benson [54] defined a new distance measure for comparing sequence profiles by estimating path lengths along an entropy surface and used it to analyze similarities within families of tandem repeats in the *C. elegans* genome. May *et al.* [55] reviewed the existing coding (both source and channel) theoretic methods for modelling genetic systems, and present research results for *Escherichia coli* K-12. As a last reference to be cited in this area, Hussinia *et al.* [56] analyzed in a formalized mathematical framework the properties of the languages used in DNA computations.

Relative Merits of Different Approaches to Detect Coding Regions in Genomic Sequences

The methods to detect coding regions in genomic sequences based in finding regions with a remarkable period-3 component in the frequency spectrum, constitute a qualitatively different approach that is independent from other methods (for example statistical) applied so far to solve this task. Among the methods based in spectral analysis, the DFT-based Spectral Rotation Measure, the Paired and Weighted Spectral Rotation (PSWR) measure, as well as the paired spectral content (PSC) outperforms the conventional 1-D frequency-domain methods (i. e. the simple detection of the period-3 spectral component in its various forms), producing higher values of specificity. By comparison with other period-3 based measures, [15] reports that the DFT-based PWSR measure method showed significant improvements, respectively, over the Spectral Content and Spectral Rotation measures in the detection of exonic nucleotides at a fixed false positive rate.

Other classical methods based in the period-3 detection like the antinotch filter and the autoregressive (AR) models showed lower coding region detection capabilities. Formal evaluations made in [15] revealed that the more recent AMDF time domain method performs better in terms of exonic nucleotide detection rates than the classical period-3 methods. The limitations of the classical methods in this case have been attributed to their relatively large window size, which reduces the time resolution. It has been suggested that the optimum window length for period-3 based methods depends on the length of the exon regions and that further improvements over the previously discussed methods are obtained using the time-frequency hybrid method (TFH). The authors consider that a promising line of development is the use of combined methods in which the detection capabilities of the combination outperforms that of the individual methods included, an approach that was used in [48].

Other Studies on Genomic Sequences Using DSP Techniques

There are other characteristics of the genomic sequences that have been studied using DSP techniques. One example is the general analysis of latent periodicities in genomic sequences which appears in Arora *et al.* [57], where sequential averaging is used when the data exhibits cyclostationarity properties.

Cristea [58, 59] studied the behaviour of the phase for complex representations of the bases in genomic sequences. These papers report the existence of a global helicoidal wrapping of the complex representations of the bases along the sequences. This is considered as a large scale trend of genomic signals. Here other properties are analyzed as well, related to the cumulated and unwrapped phase. These theoretical concepts were applied by Cristea *et al*. [60] to identify HIV Protease (PR) and Reverse Transcriptase (RT) mutations leading to multiple drug resistance to PR and RT inhibitors.

Bouaynaya and Schonfeld [61, 62] analyze the long-range power-law correlations detected in eukaryotic DNA, introducing new non-stationary methods to study the correlation properties in genomic sequences. They defined a quantitative measure of the degree of randomness (deviation from a white Gaussian process) derived from the Hilbert transform spectrum. It was shown there that DNA sequences exhibit long range correlations and that DNA correlations are much more complex than power laws with a single scaling exponent.

The Discrete Wavelet Transform has been used to analyze genomic sequences. A general perspective on the use of Wavelets and the DWT in Bioinformatics is presented by Liò [63]. An introductory analysis of genomic sequences using the DWT was presented by Ning *et al.* [64], and an approach to visualize regular patterns in DNA was introduced by Dodin *et al.* [65].

Referring to other various applications, Buchner and Janjarasjitt [66] introduced an algorithm based on processing a DNA sequence with the short-time periodicity transform, to detect and visualize tandem repeats in DNA sequences, Cristea *et al.* [67] use DSP methods for trend extraction from sets of genomic signals and apply their methodology to study the mutations in pathogen genomes, and Akhtar [15] evaluated different DSP methods to detect splice sites.

Sharma *et al.* [68] studied the repetitive DNA sequences using the DFT to identify significant periodicities present and providing a complete detection of repeats together with interactive and detailed visualization of the spectral analysis.

Dasgupta *et al*. [69] combined wavelet transform and Hidden Markov Models to identify the location of CpG islands in Human Genome. Another DSP approach for the same purpose was introduced by Rushdi and Tuqan [70]. Gupta *et al.* [71] devised an efficient algorithm to detect palindromes in DNA sequences using a signal processing operation called periodicity transform. Providence [72] applied time-varying cellular automata to the problem of finding signals in DNA sequences. Zhang and Kinsner [73] employed a multifractal analysis to DNA feature extraction, using the Rényi and Mandelbrot fractal dimension spectra for extracting the information contained in the DNA sequences.

Su *et al.* [74] applied the matched filter algorithm to analyze the structure of genomic sequences, in particular to locate and align similar segments between two sequences. Andrade and Manolakos [75] addressed the application of DSP to the electrophoresis process used in DNA sequencing and developed algorithms for signal background estimation and baseline correction.

Other DSP applications related to studies on proteins can be found in Hong and Tewfik [76], Aydin and Altunbasak [77], Lazovic [78], Ramachandran and Antoniou [79] and D'Avenio *et al.* [80].

New Perspectives of DSP Applications Based on the Algebraic Structures of the Genetic Code

The numerical representation of the genetic code and consequently of genomic sequences as has been presented in the various references cited in this article are not unique and extraordinary. In fact, the genetic codification systems that have been used so far, could be non-optimum. The nature of the genetic code is now fairly well known and there are trends to improve predictions. From the second half of 20^{th} century, many attempts have been made to understand the internal regularity of the genetic code, based on several mathematical or geometrical points of view, by Bashford and Jarvis [81], Bashford *et al.* [82], Beland and Allen [83], Crick [84], Eck [85], Epstein [86], Jimenez-Montaño [87], Jukes [88] and Hornos and Hornos [89]. In any case the Code represents an extension of the four-letter alphabet of deoxyribonucleic (DNA) bases: A, G, C, T (U in RNA).

In recent years, the genetic code algebraic structures have been introduced by Sánchez *et al.* [90-92]. It has been shown that this code constitutes a more fundamental concept than a "conventional codification system", as a consequence of its biological meaning. Depending on the algebraic operation defined in the base set, different structures were obtained. If the Watson-Crick base pairing (G:C and A:T) is expressed by the classical logical operations with "OR" ($\vee$) and "AND" ($\wedge$) in such a way that the following expressions hold: $G \vee C=C$, $T \vee A=C$, $G \wedge C=G$ and $T \wedge A=G$ then a Boolean algebra is obtained which is isomorphic to the Boolean algebra defined on the set $\{0,1\}^2$: $G \leftrightarrow 00$, $A \leftrightarrow 01$, $T \leftrightarrow 10$ and $C \leftrightarrow 11$ [90]. This leads to a binary representation of DNA sequences. On the other hand, if the Watson-Crick base pairing is expressed by the sum "+": G+C=C and U+A=C then this requirement leads to define an additive group on the DNA base set, isomorphic to the complex representation: $G \leftrightarrow 1$, $A \leftrightarrow exp(\pi i/2)$, $T \leftrightarrow exp(\pi i)$ and $C \leftrightarrow exp(3\pi i/2)$ [92].

Notice that here a numerical representation of DNA bases refer to their algebraic representation, which means the existence of an isomorphism between an algebraic structure with a biological meaning defined in the base or codon sets, and another one defined in some numerical set. We point out that the numerical representations mentioned before in this paper are codification (ad hoc) but not algebraic representations because of the absence of algebraic operations. These new models lead to go beyond the genetic code limits to deal with the quantitative relationship between DNA genomic sequences.

In particular, the extension of the four DNA base set with a dummy variable (D) leads to analogous algebraic structures, useful to deal with the multiple sequence alignments of genomic regions where the gaps are replaced by the symbol D [93]. For instance, the additive group defined in the set {D, G, A, T, C} is isomorphic to the complex representation: $D \rightarrow 1$, $G \rightarrow exp(2\pi i/5)$, $A \rightarrow exp(4\pi i/5)$, $T \rightarrow exp(6\pi i/5)$ and $C \rightarrow exp(8\pi i/5)$. The 3-periodicity was detected this way in the power spectra of the complex representations of multiple aligned genomes from HIV-1 [94]. These results showed the theoretical possibilities of using generalized DSP techniques in the comparative genomics.

CONCLUSION

The application of Digital Signal Processing in Genomic Sequence Analysis has received great attention in the last few years, providing a new insight in the solution of various problems like

- Detection of coding regions in genomic sequences based on spectral analysis.
- Reading frame identification.
- Detection of periodicities in genomic sequences.
- Detection of CpG islands.
- Detection of palindromes.

- Finding diverse signals and features in genomic sequences.
- Studies on proteins.

On the other hand, the main DSP tools that have found application in this field are

- Digital filters (IIR, FIR).
- Discrete transforms (Fourier, Cosine, Walsh Hadamard, Wavelet).
- Parametric models (mainly autoregressive).
- Information Theory concepts (entropy).
- Fractals.

Other algorithmic tools that have been applied in Bioinformatics although not addressed in this paper are considered usually as neighbouring areas. This is the case of Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Fuzzy Sets and Genetic Algorithms.

A recent development closely related to the impact of DSP on Bioinformatics is the new field of Genomic Signal Processing (GSP). An early survey on this can be found in Zhang *et al.* [95]. A formal definition of GSP was given by Dougherty *et al.* [96] as "the analysis, processing, and use of genomic signals for gaining biological knowledge and the translation of that knowledge into systems-based applications." Schonfeld *et al.* [97] remark the current interest in using DSP methods to obtain information from genomic and proteomic data to build models of molecular biological systems. This would allow obtaining a deeper understanding of the structure and functions of living systems and will help in developing new diagnostic tools, therapeutic procedures and pharmacological drugs. An application example in cancer classification and prediction can be seen in Qiu *et al.* [98].

Finally, it is interesting to notice that Bioinformatics is also having an influence on new developments, as can be seen in [99, 100].

ACKNOWLEDGEMENTS

The authors wish to acknowledge the constructive comments and critical reading of the manuscript made by the anonymous reviewers.

This research was partially funded by the Canadian International Development Agency Project Tier II-394-TT02-00 and by the Flemish VLIR-UOS Programme for Institutional University Co-operation (IUC).

REFERENCES

[1] De Fonzo V, Aluffi-Pentini F, Parisi V. Hidden Markov Models in Bioinformatics. *Curr Bioinform* **2007**; 2: 49-61.

[2] Oppenheim AV, Schafer R. Discrete-Time Signal Processing (3rd Edition), Prentice-Hall, NY **2009**.

[3] Proakis JG, Manolakis DK, Digital Signal Processing (4th Edition), Prentice Hall, NY **2006**.

[4] Stoica P, Moses RL. Spectral Analysis of Signals, Prentice-Hall, NY **2005**.

[5] Burrus CS, Gopinath RA, Guo H. Introduction to Wavelets and Wavelet Transforms: A Primer, Prentice-Hall, NY **1997**.

[6] Shannon CE. A Mathematical Theory of Communication. *he Bell Sys Techn J* **1948**; 27: 379-23, 623-56.

[7] Rényi A. On measures of information and entropy. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* **1960**: 547-61.

[8] Voss RF. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phy. Rev. Lett* **1992**; 68: 3805-08.

[9] Zhang R, Zhang CT. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.* **1994**; 11: 767-82.

[10] Rushdi A, Tuqan J. Gene Identification Using the Z-Curve Representation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP* **2006**; 2: II-

[11] Cristea PD . Genomic signal analysis: Study of pathogen variability. *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics GENSIPS '06* **2006**: 51-52.

[12] Anastassiou D. Genomic signal processing**.** *IEEE Sign Proc Mag* **2001**; 18: 8-20.

[13] Cristea P, Deklerck R, Cornelis J, Tuduce R, Nastac I, Andrei M. Signal Representation and Processing of Nucleotide Sequences. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE* **2007**: 1214-19.

[14] Cristea PD, Tuduce R, Cornelis J. Signal Analysis of Pathogens Genomic Sequences. *Frontiers in the Convergence of Bioscience and Information Technologies, FBIT* **2007**: 245- 50.

[15] Akhtar M, Epps J, Ambikairajah E. Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. *IEEE J Select Topics Sign Proc* **2008**; 3: 310-21.

[16] Cristea PD. Genetic Signal Analysis. *Proc Int Symp Sign Proc Appl (ISSPA)* **2001**; 2: 703-08.

[17] Rushdi A, Tuqan J. 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**: 1-4.

[18] Brodzik AK, Peters O. Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. *Proceedings IEEE International Conference onAcoustics, Speech, and Signal Processing, (ICASSP '05)* **2005**; 5: 373-76.

[19] Trifonov EN. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A* **1998**; 249: 511-16.

[20] Tuqan J, Rushdi A. A DSP Approach for Finding the Codon Bias in DNA Sequences. *IEEE J Select Topics Sign Proc* **2008**; 2: 343-56.

[21] Sánchez J, López-Villaseñor I. A simple model to explain three-base periodicity in coding DNA. *FEBS Lett* **2006**; 580: 6413-22.

[22] Vaidyanathan PP, Yoon BJ. The role of signal-processing concepts in genomics and proteomics. *J Franklin Inst* **2004**; 341: 111-35.

[23] Tuqan J, Rushdi A. A DSP perspective to the period-3 detection problem**.** *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS '06* **2006**: 53-54.

[24] Fox TW, Carreira A. A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression. *EURASIP J Appl Sign Proc* **2004**; 1: 108-11.

[25] Afreixo V, Ferreira PJSG, Santos D. Fourier analysis of symbolic data: A brief review. *Digit Sign Proc* **2004**; 14: 523-30.

[26] Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS* **1997**; 13: 263-70.

[27] Yan M, Lin ZS, Zhang CT. A new Fourier Transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics* **1998**; 14: 685-90.

[28] Datta S, Asif A, Wang H. Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics. *Proceedings of the IEEE Sixth International Symposium on Multimedia Software* **2004**: 160-63.

[29] Datta S, Asif A. A fast DFT based gene prediction algorithm for identification of protein coding regions. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '05)* **2005**; 5: 653-56.

[30] Datta S, Asif A. DFT based DNA splicing algorithms for prediction of protein coding regions. *Proceedings of the IEEE Thirty-Eighth Asilomar Conference on Signals, Systems and Computers* **2004**; 1: 45-49.

[31] Isaac B, Singh H, Kaur H, Raghava GPS. Locating probable genes using Fourier Transform approach. *Bioinform Appl Note* **2002**; 18: 196-97.

[32] Stoffer D, Ombao HC, Tyler DE. Local spectral envelope: an approach using dyadic tree-based adaptive segmentation. *Ann Inst Statist Math* **2002.** 54: 201-23.

[33] Epps J, Ambikairajah E, Akhtar M. An integer period DFT for biological sequence processing. *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics GENSIPS* **2008**: 1-4.

[34] Rushdi A, Tuqan J. Trigonometric transforms for finding repeats in DNA sequences. *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS '08* **2008**: 1-4.

[35] Berger JA, Mitra SK, Astola J. Power spectrum analysis for DNA sequences. *Proc IEEE Seventh Int Symp Sign Proc Appl* **2003**; 2: 29-32.

[36] Rodríguez-Fuentes A, Lorenzo-Ginori JV, Grau-Ábalo R. Detection of coding regions in large DNA sequences using the short time Fourier Transform. *Lect Notes Comput Sci* **2006**. 4225: 902-909.

[37] Kotlar D, Lavner Y. Gene Prediction by Spectral Rotation Measure: A New Method for identifying Protein-Coding Regions. *Genome Res* **2003**; 13: 1930-1937.

[38] Rushdi A, Tuqan J. The Filtered Spectral Rotation Measure. *Proc the IEEE Fortieth Asilomar Conf Sign Sys Comput ACSSC '06* **2006**: 1875-79.

[39] Yin C, Yau SS-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J Theor Biol* **2007**; 247: 687-94.

[40] Akhtar M, Ambikairajah E, Epps J. Optimizing period-3 methods for eukaryotic gene prediction. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP* **2008**: 621-24.

[41] Xing C, Bitzer DL, Alexander WE, Stomp AM, Vouk MA. Free Energy Analysis on the Coding Region of the Individual Genes of Saccharomyces cerevisiae. *Proceedings of the 28th IEEE EMBS Annual International Conference* **2006**: 4225-28.

[42] Chakravarthy N, Spanias A, IasemidisLD, Tsakalis K. Autoregressive Modeling and Feature Analysis of DNA Sequences. *EURASIP J Appl Sign Proc* **2004**; 1: 13-28.

[43] Rao N, Shepherd SJ. Detection of 3-periodicity for small genomic sequences based on AR technique. *Proceedings of the IEEE International Conference on Communications, Circuits and Systems, ICCCAS* **2004**; 2: 1032-36.

[44] Akhtar M, Ambikairajah E, Epps J. Comprehensive autoregressive modeling for classification of genomic sequences. *Proceedings of the IEEE 6th International Conference on Information, Communications & Signal Processing* **2007**: 1-5.

[45] Akhtar M, Ambikairajah E, Epps J. Detection of Period-3 Behavior in Genomic Sequences Using Singular Value Decomposition. *Proceedings of the IEEE International Conference on Emerging Technologies* **2005**: 13-17.

[46] Akhtar M. Comparison of Gene and Exon Prediction Techniques for Detection of Short Coding Regions. *Int J Inform Technol* **2005**; 8: 26-35.

[47] Cristea P, Deklerck R, Cornelis J, Tuduce R, Nastac I, Andrei M. ANN Prediction of Nucleotide Sequences Link of Principal Component Analysis to Fourier Transform. *Proceedings of the 14th International Workshop on Systems, Signals and Image Processing, and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services* **2007**: 69-73.

[48] Rodríguez-Fuentes A, Lorenzo-Ginori JV, Grau-Ábalo R. Coding Region Prediction in Genomic Sequences Using a Combination of Digital Signal Processing Approaches. *Lect Notes Comput Sci* 2007; 4756: 635-642.

[49] Gunawan TS, Ambikairajah E, Epps J. Boosting approach to exon detection in DNA sequences. *Electron Lett* **2008**; 44: 323-24.

[50] Román-Roldán R, Bernaola-Galván P, Oliver JL. Sequence Compositional Complexity of DNA through an Entropic Segmentation Method. *Phys Rev Lett* **1998**. 80: 1344-47.

[51] Bernaola-Galván P, Grosse I, Carpena P, Oliver JL, Román-Roldán R, Stanley HE. Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method. *Phys Rev Lett* **2000**; 85: 1342-45.

[52] Nicorici D, Astola J. Segmentation of DNA into Coding and Noncoding Regions Based on Recursive Entropic Segmentation and Stop-Codon Statistics. *EURASIP J Appl Sign Proc* **2004**; 1: 81-91.

[53] Mutihac R, Cicuttin A, Mutihac RC. Entropic approach to information coding in DNA molecules. *Mat Sci Eng C* **2001**; 18: 51-60.

[54] Benson G. A new distance measure for comparing sequence profiles based on path lengths along an entropy surface. *Bioinformatics* **2002**; 18 (Suppl 2): S44-S53.

[55] May EE, Vouk MA, Bitzer DL, Rosnick DI. An error-correcting code framework for genetic sequence analysis. *J Franklin Inst* **2004**; 341: 89-109.

[56] Hussinia S, Karib L, Konstantinidisa S. Coding properties of DNA languages. *Theor Comput Sci* **2003**; 290: 1557-79.

[57] Arora R, Sethares WA, Bucklew JA. Latent Periodicities in Genome Sequences. *IEEE J Select Topics Sign Proc* **2008**; 3: 332-42.

[58] Cristea PD. Phase analysis of DNA genomic signals**.** *Proceedings of the 2003 Int Symp Circuits Sys ISCAS '03* **2003**; 5: 25-28.

[59] Cristea PD. Multiresolution phase analysis of genomic signals. *Proceedings of the First International Symposium on Control, Communications and Signal Processing* **2004**: 743-46.

[60] Cristea PD, Tuduce RA. Otelea D. Study of HIV Variability based on Genomic Signal Analysis of Protease and Reverse Transcriptase Genes. *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2005**: 4795-98.

[61] Bouaynaya N, Schonfeld D. Emergence of new structure from nonstationary analysis of genomic sequences. *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics GENSiPS* **2008**: 1-4.

[62] Bouaynaya N, Schonfeld D. Nonstationary Analysis of Coding and Noncoding Regions in Nucleotide Sequences. *IEEE J Select Topics in Sign Proc* **2008**; 3: 357-64.

[63] Liò P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinform Rev* **2003**; 19: 2-9.

[64] Ning J, Moore CN, Nelson JC. Preliminary wavelet analysis of genomicsequences. *Proceedings of the IEEE Bioinformatics Conference* **2003**; 509-10.

[65] Dodin G, Vandergheynst P, Levoir P, Cordier C, Marcourt L. Fourier and Wavelet Transform Analysis, a Tool for Visualizing Regular Patterns in DNA Sequences. *J Theor Biol* **2000**; 206: 323-26.

[66] Buchner M, Janjarasjitt S. Detection and Visualization of Tandem Repeats in DNA Sequences. *IEEE Transac Sign Proc* **2003**; 51: 2280-87.

[67] Cristea P, Tuduce R, Monteanu A, Cornelis J. Common Trend Extraction from Sets of Genomic Signals. *Proc IEEE Int Symp Commun Control Sign Proc ISCCSP* **2008**: 1205-10.

[68] Sharma D, Issac B, Raghava GPS, Ramaswamy R. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* **2005**; 20: 1405-12.

[69] Dasgupta N, Lin S, Carin L. Sequential Modeling for Identifying CpG Island Locations in Human Genome. *IEEE Sign Proc Lett* **2002**; 9: 407-09.

[70] Rushdi A, Tuqan J. A New DSP-Based Measure for CPG Islands Detection. *Proceedings of the IEEE 12th Signal Processing Education Workshop, 4th Digital Signal Processing Workshop* **2006**: 561-65.

[71] Gupta R, Mittal A, Gupta S. An efficient algorithm to detect palindromes in DNA sequences using periodicity transform. *Signal Processing* **2006**; 86: 2067-73.

[72] Providence SV. Utilization of Cellular Automata in the DNA Signal Search Problem. *Proc IEEE Southeast Conf* **2004**; 325-29.

[73] Zhang H, Kinsner W. Feature extraction from DNA sequences by multifractal analysis. *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2001**; 1567-72.

[74] Su SC, Yeh CH, Kuo CCJ. Structural analysis of genomic sequences with matched filtering. *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2003**; 3: 2893-96.

[75] Andrade L, Manolakos ES. Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing. *J VLSI Sign Proc* **2003**; 35: 229-43.

[76] Hong C, Tewfik AH. Efficient Updating of Biological Sequence Analyses. *IEEE J Select Topics Sign Proc* **2008**; 2: 365-77.

[77] Aydin Z, Altunbasak Y. A signal processing application in genomic research: protein secondary structure prediction. *IEEE Sign Proc Mag* **2006**; 23: 128-31.
[78] Lazovic J. Selection of amino acid parameters for Fourier transform-based analysis of proteins. *CABIOS COMMUNICATION* **1996**; 12: 553-62.
[79] Ramachandran P, Antoniou A. Identification of Hot-Spot Locations in Proteins Using Digital Filters. *IEEE J Select Topics Sign Proc* **2008**; 2: 378-89.
[80] D'Avenio G, Grigioni M, Orefici G, Creti R. SWIFT (sequence-wide investigation with Fourier transform): a software tool for identifying proteins of a given class from the unannotated genome sequence. *Bioinform* **2005**; 21: 2943-49.
[81] Bashford JD, Jarvis PD. The genetic code as a periodic table. *Biosystems* **2000**; 57: 147-161.
[82] Bashford JD, Tsohantjis I, Jarvis PD. A supersymmetric model for the evolution of the genetic code. *Proc Natl Acad Sci USA* **1998**; 95: 987-992.
[83] Beland P, Allen TF. The origin and evolution of the genetic code. *J Theor Biol* **1994**; 170: 359-365.
[84] Crick FHC. The origin of the genetic code. *J Mol Biol* **1968**; 38: 367-379.
[85] Eck RV. Genetic code – Emergence of a symmetrical pattern. *Science* **1963**; 140: 477-481.
[86] Epstein CJ. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature* **1966**; 210: 25-28.
[87] Jimenez-Montaño MA. The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions in vivo and in vitro. *Biosystems* **1996**; 39: 117-125.
[88] Jukes TH. The amino acid code. In A. Neuberger, Comprehensive Biochemistry. Elsevier Amsterdam **1977**, pp. 235-293.
[89] Hornos JE, Hornos YM. Algebraic model for the evolution of the genetic code. *Phys Rev Lett* **1993;** 71: 4401-4404.
[90] Sánchez R, Morgado E, Grau R. A genetic code boolean structure I. The meaning of boolean deductions. *Bull Math Biol* **2005**; 67: 1-14.
[91] Sánchez R, Morgado E, Grau R. Gene algebra from a genetic code algebraic structure. *J Math Biol* **2005**; 51, 431-457,
[92] Sánchez R, Grau R. A novel algebraic structure of the Genetic Code over the Galois Fields of four DNA Bases. *Acta Biotheoretica* **2006**; 54: 27-42.
[93] Sánchez R, Grau R, Morgado E. A Novel DNA Sequence Vector Space over an extended Genetic Code Galois Field. *MATCH Commun Math Comput Chem* **2006**; 56: 5-20.
[94] Sanchez R., Grau R. An algebraic hypothesis about the primeval genetic code. Second international workshop Cuba/Flanders IWOBI 08. ISBN 978-959-250-394-6. arXiv:0805.1128v3
[95] Zhang XY, Chen F, Zhang YT, *et al*. Signal Processing Techniques in Genomic Engineering. *Proc IEEE* **2002**; 90: 1822-33.
[96] Dougherty ER, Datta A, Sima C. Research issues in genomic signal processing. *IEEE Sign Proc Mag* **2005**; 22: 46-48.
[97] Schonfeld D, Goutsias J, Shmulevich I, Tabus I, Tewfik AH. Introduction to the Issue on Genomic and Proteomic Signal Processing. *IEEE J Select Topics Signl Proc* **2008**; 2: 257-59.
[98] Qiu P, Wang Z J, Liu KJR. Genomic Processing for Cancer Classification and Prediction. *IEEE Sign Proc Mag* **2007**; 24: 100-10.
[99] Chen J, Li H, Sun K, Kim B. How will bioinformatics impact signal processing research? *IEEE Sign Proc Mag* **2003**; 6: 106-26.
[100] Tsaftaris SA, Katsaggelos AK, Pappas TN, Papoutsakis ET. How Can DNA Computing be Applied to Digital Signal Processing? *IEEE Sign Proc Mag* **2004**; 21: 57-61.

Received: September 29, 2008 Revised: October 14, 2008 Accepted: October 28, 2008