# 8

# Spectral Analysis of Global Behaviour of C. Elegans Chromosomes

Afef Elloumi Oueslati[1], Imen Messaoudi[1],
Zied Lachiri[2] and Noureddine Ellouze[1]
*Unité Signal, Image et Reconnaissance de Formes, Département de Génie Electrique,*
*[1]Ecole Nationale d'Ingénieurs de Tunis, BP 37, Campus Universitaire,*
*Le Belvédère, 1002, Tunis,*
*[2]Département de Génie Physique et Instrumentation*
*Institut National des Sciences Appliquées et de Technologie, BP 676,*
*Centre Urbain Cedex, 1080, Tunis,*
*Tunisie*

## 1. Introduction

Fourier analysis is one of the most useful decomposition into frequency bands to provide a signal's variations and irregularities measure. DNA spectral analysis based on Fourier Transform contributes in the systematic search of special DNA patterns which may correspond to biological important markers. For example, the Fourier harmonic analysis of the occurrence of a base "A" can give us the corresponding frequency with amplitude and a phase without being able to locate it in time. However it is interesting to also detect the moments of "silence" of base "A" i.e. the moments when this base does not exist. Such a representation of Fourier is thus limited with signals which contain transitory elements or evolutions in their spectral contents. For these non stationary signals, the DNA sequences, to highlight the frequency behavior, it becomes necessary to give the frequency the possibility of changes over time. It's the time frequency analysis aim assured by the Short Time Fourier Transform. In fact, the punctual aspect is very important to localize particular regions in chromosomes, to characterize the beginning of a protein coding regions or a nucleosome or its end. By depicting the frequencies by a smoothed STFT, a 2D or 3D spectrogram representation, specific regions appear distinctly. In this paper, we are concerned with the periodicities 3, 6, 9 and 10.5. The periodicity 3 discussed in (Anastassiou, 2001; Berger et al, 2003; Cohanim et al, 2005; Kornberg, 1977; Segal et al, 2006; Susillo et al 2003; Trifonov & Sussman, 1980; Trifonov, 1998; Vaidyanathan & Yoon, 2004) is related with protein coding regions (called exons) in the gene. The periodicity 10.5 is related with nucleosome's positions in the DNA sequence and the degree of deformability of the sequence in the DNA helix (Hayes et al, 1990; Trifonov & Sussman, 1980; Widom, 1996; Worcel et al 1981). The periodicity 6 and 9 are specific to C. Elegans organism.

This chapter is divided in five parts. First, we expose an introduction for relevant regions on chromosomes. In part three, we detailed the DNA's sequence analysis approach, related to sequence global behavior problem. It exposes the spectral analysis, which follows a certain

methodology that generates results to highlight the periodicities studied. This analysis is based on organisms translated into signals by three coding techniques. The algorithm steps of this technique are detailed to mention the generation method of spectrums and spectrograms. The fourth part deals with the study of the frequencies' evolution. It present results for smoothed STFT, as a 1D, 2D or 3D spectrogram representation. Part five concludes this chapter.

## 2. The relevant regions in chromosomes

The specific succession in the bases (A, G, C, and T) constitutes the hereditary message. Each DNA fragment involves a specific protein synthesis process. Proteins are synthesized from a set composed of 20 different amino acids, which are determined by three bases occurring in subsequent order. A group of three consecutive nucleotides with desoxyribose and phosphoric group is called a codon and a total of 64 different combinations specify 20 amino acids and three stop codons, namely TAA, TAG, and TGA. The protein synthesis (Fig.1) is realized in two steps: (1) the transcription within which the hereditary information is copied into the messenger RNA and, (2) the translation in which the messenger RNA is exploited by the ribosome to form the amino acid chain. To obtain numerical data from this succession of symbolic bases of a DNA sequence, we use binary indicator coding techniques.

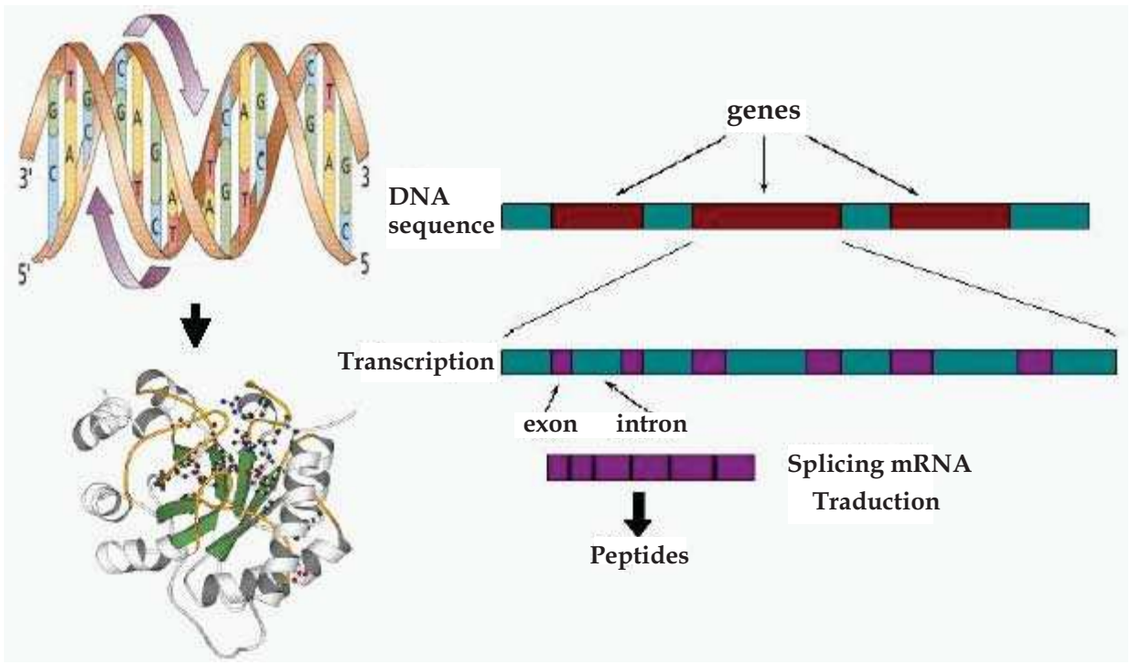



Fig. 1. The protein's synthesis steps

In a DNA sequence, electron microscopy and biochemical studies have established that the bulk of the chromatin DNA is compacting into repeating structural units, named nucleosomes. A model of this DNA structure in such regions is proposed by Kornberg in (Kornberg, 1974, 1977). The chromatin is a dynamic structure, oscillating between the nucleosome and open structures depending on the environmental conditions (Kornberg, 1974, 1977; Oudet et al, 1978). And each nucleosome is formed by two molecules of each histone (protein) H2A,

H2B, H3 and H4. Each nucleosome has a diameter of 12.5±1 nm and contains about 200 base pairs of DNA. This number is varying according to the chromatin's origin (Hayes et al 1990; Kornberg, 1977; Oudet et al, 1978; Worcel et al 1981). In contrast a particle named 'nucleosome core' is invariant in its DNA content about 146 base pairs. Interesting electron microscopic evidence elaborated in (Oudet et al, 1978) suggests that under appropriate conditions a nucleosome could open up into two separate half nucleosomes of diameter 9.3±1 nm. The finding of each type of histones in the nucleosome has suggested that a nucleosome could be made up of two symmetrical halves (Altenburger, 1976).

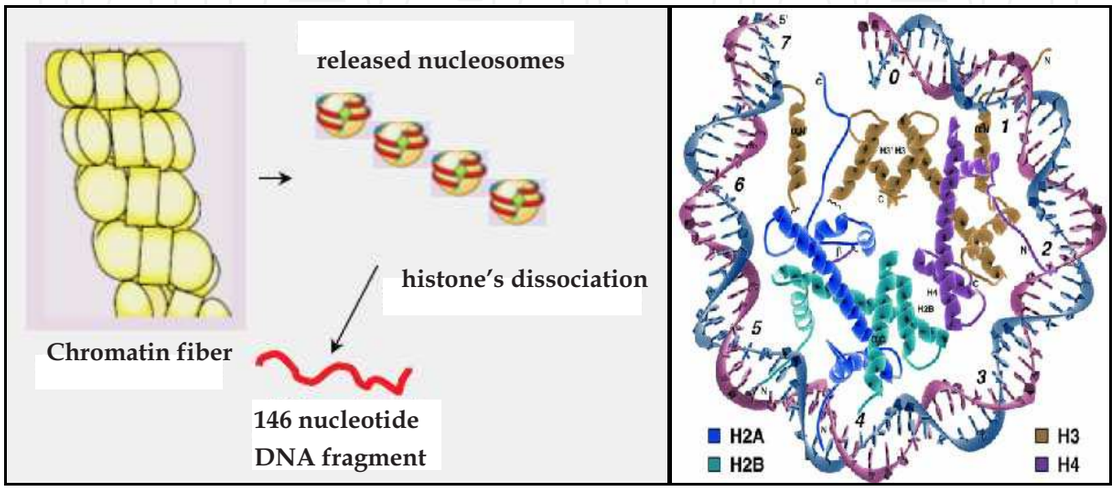



Fig. 2. Chromatine's and nucleosome's structure

In order to study the protein coding regions signals and the nucleosome regions ones, the DNA symbolic data must be converted to DNA signals.

## 3. Genomic sequence analysis based on Short Fourier Transform

In order to give frequencies more precise location in time, Gabor proposes to use a Fourier local analyze with windows. The technique consists in segmenting signal by multiplication by sliding window of fixed length (Mallat, 1999). Each part is analyzed independently with a classic Fourier transform to enhance frequencies behavior. The totality of these transforms forms the short Fourier transform and precise the frequencies location in time.

Applying coding process, the numerical signals are obtained by base's succession description as follows:

$$x[n] = \{x(i), i \in [1,..,N]\} \tag{1}$$

The classic discrete Fourier transform related to numerical sequence is expressed as:

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} \tag{2}$$

In order to locate the signal frequencies in time, the analysis is applied to sequence's parts generated by multiplication with a sliding analysis window.

For this purpose, the numerical signal x[n] is divided into frames of N length. The expression become

$$x_w[n,i] = x[n].\omega[i - \Delta n] \tag{3}$$

When based on the binary indicator 'A', the equation becomes:

$$x_{Aw}[n,i] = U_A[n].\omega[i - \Delta n] \tag{4}$$

With i is the window's order and the Δn is the adopted sliding value. The window's length must be chosen to have an appropriate number of samples to guarantee the best frequency resolution. On each block $x_w$ [n], is applied a Fourier transform to determine Xw [k], k∈[0:N-1], k represents the frequency index. The FT expression associated with each frame is as follows:

$$X_\omega^i[k] = \sum_{n=0}^{N-1} x_\omega[n,i] e^{-j\frac{2\pi}{N}nk} \tag{5}$$

With binary indicator 'A' coding, the equation is:

$$X_A[k] = \sum_{n=0}^{N-1} x_{A\omega}[n,i] e^{-j\frac{2\pi}{N}nk} \tag{6}$$

On the basis of this expression, many representations can be obtained. The sequence is associated to chromosome, the first analyze consists in studying the frequency global behavior. To enhance the frequencies, we used a mean smoothed spectrum. The principle consists in calculating the mean of the obtained spectrum of equation.

$$\bar{X}_\omega^j[k] = \frac{1}{N} \sum_{i=0}^{N-1} X_\omega^i[k] \tag{7}$$

The chromosomes are generally constituted by more than 10 Mbp, so the obtained spectrum needs to be smoothed. A second mean of the mean spectrums is applied. The converted DNA sequence x[n] is divided into frames of M length with an overlap Δm. Each of these frames is also divided into N frames by multiplication with a sliding analysis window w[n]. On each part, a mean smoothed spectrum is generated. Finally, the mean of the spectrum for all the parts is calculated. The final expression of the spectrum is:

$$\bar{X}[k] = \frac{1}{M} \sum_{j=0}^{M-1} \bar{X}_\omega^j[k] \tag{8}$$

### 3.1 The chromosomes coding techniques

This analysis aims to study the chromosome's frequency global behaviour. For this purpose, it is important to enhance particularly the signals generated by the protein coding regions and the nucleosome regions. That's why, three types of coding techniques are considered:

- A linear coding based on Binary indicators which is related to the base succession,
- A Structural coding with Pnuc, which is an experimental coding based on the helix deformability
- A two-dimensional coding based on Frequency Chaos Game Representation which has submerged from the field of physics known as 'chaotic dynamical systems'

### 3.1.1 Binary indicator's techniques

The linear coding consists in attributing a binary value for each unit of the all indicators. Which are included in {'A','T','C','G', 'TT', 'TA', 'GC'', 'AAA'… 'GGG'}. The marker associated takes the value of either 1 or 0 at location n for the first character, depending on whether or not the corresponding character group exists from the location n.

Base's binary indicator:

$$S[n] = \sum_{b \in B} U_b[n] \tag{9}$$

Where:

$$U_b[n] = \begin{cases} 1 \; if \; base \; b \; is \; in \; position \; n \\ 0 \; else \end{cases} \tag{10}$$

is the binary indicator of the base B={A,T,C,G}

Considering sequence $S_{DNA}$ and $U_A[n]$ the associated base's binary indicator
$S_{DNA}$ = 'AATCGCGACACTCATTCGG'
$U_A[n]$ = 1 1 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0

Two Base's binary indicator:

$$S[n] = \sum_{bb \in BB} U_{bb}[n] \tag{11}$$

where

$$U_{bb}[n] = \begin{cases} 1 \; if \; base \; bb \; is \; in \; position \; n \\ 0 \; else \end{cases} \tag{12}$$

is the binary indicator of the base B

B={AA,AT,AC,AG,TA,TT,TC,TG,CA,CT,CC,CG,GA,GT,GC,GG}

Considering sequence $S_{DNA}$ and $U_{CG}[n]$ the associated base's binary indicator
$S_{DNA}$ = 'AAT**CGCG**ACACTCATT**CG**G'
$U_{CG}[n]$ = 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0

Some dinucleotides as 'AA', 'TT', 'TA' are enhancing the ADN flexibility around histones to constitute nucleosomes (Fig. 3).

Codon's binary indicator: the three bases association called triplet or codon have a fundamental role in the process of amino acids fabrication. For these reasons, a coding

technique based on these base's association is used. We adopt binary indicators to each of the 64 codons (Table 1)

$$S[n] = \{U_{cod}[i], i = 1 \ldots N_s\} \tag{13}$$

where:

$$U_{cod}[i] = \begin{cases} 1 \; if \; the \; codon \; cod \; starts \; at \; position \; n \\ 0 \; else \end{cases} \tag{14}$$

is the binary indicator of the codon cod and Ns is the sequence's length. This marker takes the value of either 1 or 0 at location n for the first character depending on whether the corresponding character exists from the location n. Let's consider the codon binary indicator $U_{TCG}[n]$.

Considering sequence $S_{DNA}$ and the associated codon's binary indicator
$S_{DNA}$ = 'AAT**CGCG**ACACTCATT**CG**G'
$U_{TCG}[n]$ = 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0

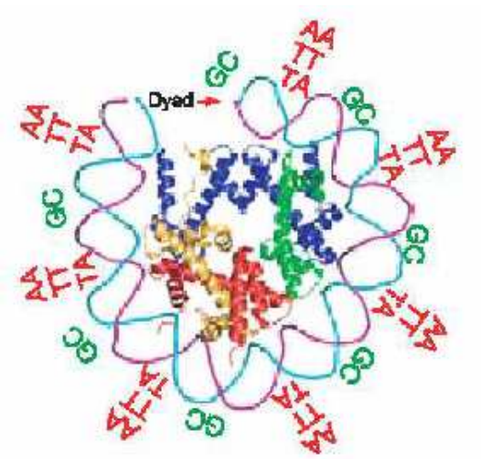



Fig. 3. DNA flexibility around histones is enhanced by dinucleotide as 'AA', 'TT', 'TA'

| BASE | Codon associated |
|---|---|
| A | AAA, AAT, AAC, AAG, ATA, ATT, ATC, ATG<br>ACA, ACT, ACC, ACG, AGA, AGT, AGC, AGG |
| T | TAA, TAT, TAC, TAG,TTA, TTT, TTC, TTG<br>TCA, TCT, TCC, TCG, TGA, TGT, TGC, TGG |
| C | CAA, CAT, CAC, CAG, CTA, CTT, CTC, CTG<br>CGT, CGC, CGG, CCA, CCT, CCC, CCG, CGA |
| G | GAA, GAT, GAC, GAG, GTA, GTT, GTC, GTG<br>GCA, GCT, GCC, GCG, GGA, GGT, GGC, GGG |

Table 1. Codon associated to each base

### 3.1.2 Pnuc: the structural coding techniques

The second coding technique is the Pnuc which is based on local bending and flexibility properties of the double helix; it is deduced experimentally from nucleosome positioning (Pnuc). By considering the matching of both stalks (A-T and C-G) along the helix, one base's pair defines a plane and a direction in this plane. A description of the double helix shows the overlapping of the plans (Fig. 4). When considering that the planes are parallel, passing between planes needs translation and rotation of 34,3° of the orientation of the connection of the plan.

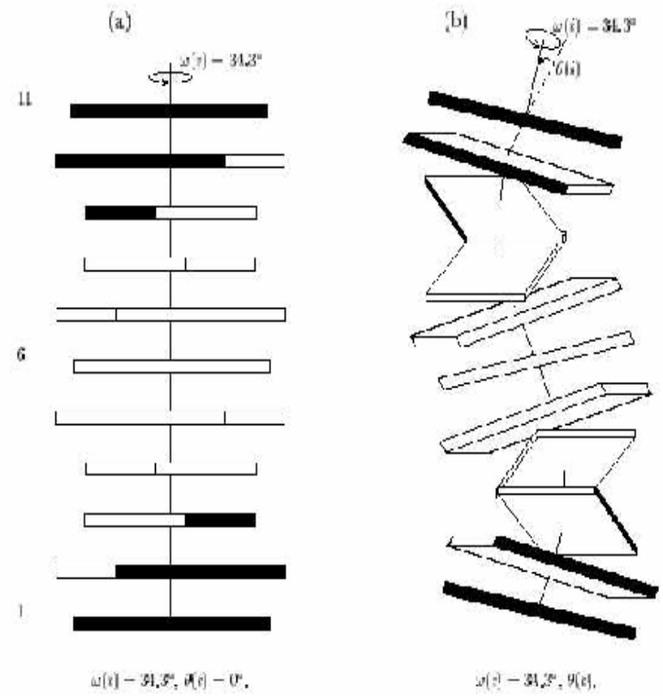



Fig. 4. A description of the double helix shows the overlapping of the plans

Now the plans are not parallel and the axis of the double helix presents curvature. By considering the interaction between a protein, a histone and a DNA's sequence, this interaction is stronger when the contact area between both objects is the biggest. To increase this surface, it is necessary to roll up as much as possible the segment of DNA around the protein, in this way, we have two properties:

If the segment of DNA is not rolled up around the protein, it is in position of equilibrium, the curvature is static

The stalk must be flexible to allow the additional curvature around the protein. These two properties generate the nucleosome which generates an excessive curvature of the stalk.

Each trinucleotide is replaced by its numerical value given by the Pnuc table. The $S_{DNA}$ is then replaced by the numerical sequence $C_{PNUC}$.

$S_{DNA}$ = 'AAT**CGCG**ACACTCATT**CG**G'
$C_{PNUC}$ = 0.7 5.3 8.3 7.5 7.5 6.0 5.4 5.2 6.5 5.8 5.4 5.4 6.7 0.7 3.0 8.3 4.7

| Trinucleotide | PNUC | Trinucleotide | PNUC |
|---|---|---|---|
| AAA/TTT | 0.0 | CAG/CTG | 0.042 |
| AAC/GTT | 0.037 | CCA/TGG | 0.054 |
| AAG/CTT | 0.052 | CCC/GGG | 0.060 |
| AAT/ATT | 0.07 | CCG/CGG | 0.047 |
| ACA/TGT | 0.052 | CGA/TCG | 0.083 |
| ACC/GGT | 0.054 | CGC/GCG | 0.075 |
| ACG/CGT | 0.054 | CTA/TAG | 0.022 |
| ACT/AGT | 0.058 | CTC/GAG | 0.054 |
| AGA/TCT | 0.033 | GAA/TTC | 0.030 |
| AGC/GCT | 0.075 | GAC/CTG | 0.054 |
| AGG/CCT | 0.054 | GCA/TGC | 0.060 |
| ATA/TAT | 0.028 | GCC/GGC | 0.0100 |
| ATC/GAT | 0.053 | GGA/TCC | 0.038 |
| ATG/CAT | 0.067 | GTA/TAC | 0.037 |
| CAA/TTG | 0.033 | TAA/TTA | 0.020 |
| CAC/GTG | 0.065 | TCA/TGA | 0.054 |

Table 2. The PNuc table

The signal generated from this coding for a part of chromosome is given by Fig. 5. For clarity purpose the signal is multiplied by 10. Fig. 6 illustrate the stft method applied on this resulting signal. First, subfigure a shows a mean spectrum for distinct window of length $5*10^5$.. The spectrum obtained needs smoothing so for the second figure (subfigure b) a blackman smoothing window is applied on each signal part before calculating the mean spectrum of equation 7. In the third and last figures (subfigure c and d) the equation 8 is used and the parameters chosen are: Blackman window, $M=5*10^5$ , $N=5*10^4$ and overlap 50% for subfigure c and $N=5*10^3$ with overlap 10% for subfigure d. The figure shows that meaning and smoothing are very efficient to have the best signal (subfigure d). In this signal, the periodicity 10 is enhanced to prove that this is a characteristic of helix flexibility.

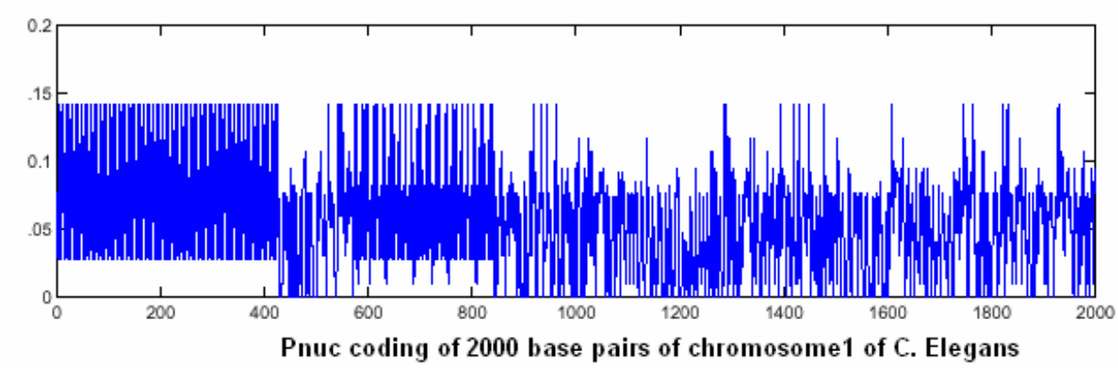



Fig. 5. Pnuc signal of 2000 base pairs of chromosome 1 of C. Elegans genome

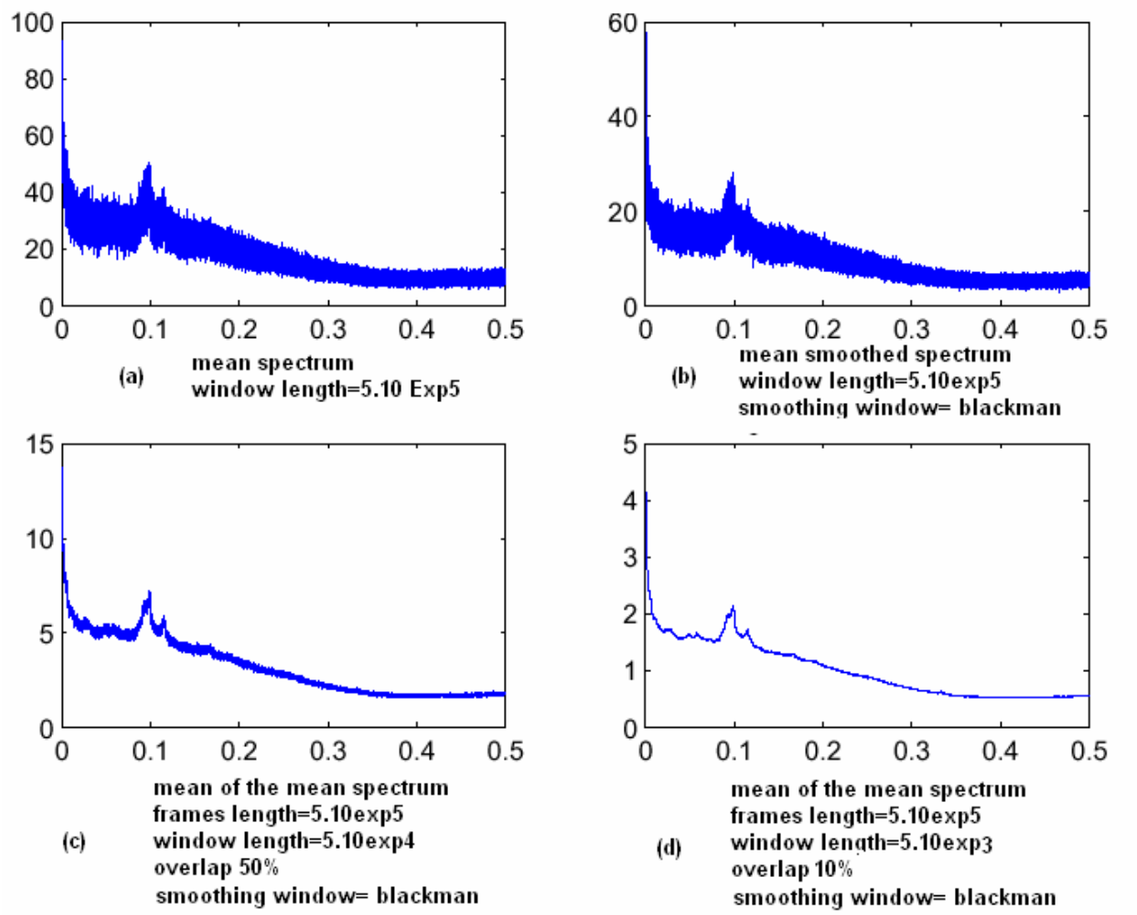



Fig. 6. Illustration of the smoothed mean spectrum applied on Pnuc signal of 2000 base pairs of chromosome 1 of C. Elegans genome

### 3.1.3 Fcgr: the two dimensionnal coding techniques

The third technique is submerged from the Chaos Game Representation (CGR) images which can forms a global signature of bio-sequences (Almeida et al, 2001; Cenac et al, 2004; Deshavanne et al, 1999, 2000; Joseph & Sasikumar, 2006; Oliver et al 1993; Fiser et al, 1994). The CGR paradigm is a holistic way of DNA representation. It provides a unique scatter pictures. In 1999, H. Joel Jeffrey uses for the first time this representation for studying the "non-randomness" of genomic sequences (Jeffrey, 1990). The CGR is an iterative algorithm for drawing fractal images to any desired scale. It maps nucleotide sequences in the [0,1]x[0,1] square. The four letters A, C, G and T are placed at the corners. The binary CGR vertices are assigned to the four nucleotides as:

$$l_A = (0,0), l_C = (0,1), l_G = (1,1), l_T = (1,0) \tag{15}$$

Deriving scatter pictures, the CGR's construction algorithm consists of three steps. First, the four letters A, C, G and T are placed at the corners of a rectangular unit square. Second, the first point is plotted halfway between the center of the square, and the corner corresponding

to the first nucleotide of the sequence. Third, the new point are marked successively half way between the previous point and the corner corresponding to the base of each nucleotide read from the sequence (Almeida et al, 2001; Joseph 2006). A generated CGR image can be viewed as an image of distributed dots. Subdividing the unit square into a set of square entries of equal size n, the number of square entries obtained is equal to 2n ×2n. The number of points counted in each sub-square represents the number of occurrence of a particular n-lengthen pattern.

For illustration, let's consider a DNA sequence $S = \{S_1, S_2, ...., S_N\}$ of N nucleotides, the CGR value along this sequence is defined by equation 16. The result will be a square uniformly and randomly filled with dots.

$$X_{n+1} = \frac{1}{2}\left(X_n + l_{s_{n+1}}\right) \tag{16}$$

The first point $X_0$ is usually placed at the center of the square having thus the coordinates (0.5, 0.5). Then, the next point $X_{n+1}$ is repeatedly placed halfway between the previous plotted point $X_n$ and the segment joining the vertex corresponding to the letter $s_{n+1}$ of the sequence. Fig. 7 illustrates the construction process of CGR trajectory for sequence "ATCGG".

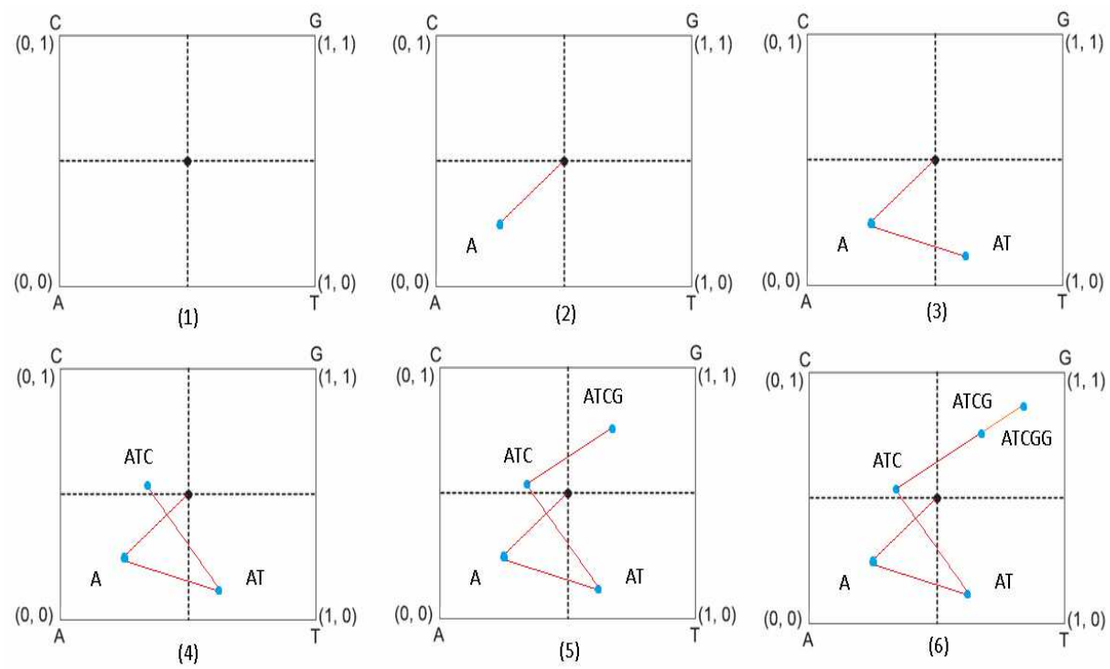



Fig. 7. An illustration of CGR trajectory for sequence "ATCGG"

To derive the CGR plot, the following steps are taken: First place $X_0$ at the square's center and the four letters at the corners as described before (subfigure 1). From center to vertex A, mark midpoint 1 ( address A) (subfigure 2). From 1 to T, mark midpoint 2 (address AT) (subfigure 3). From 2 to C, mark midpoint 3 (address ATC) (subfigure 4). From 3 to G, mark midpoint 4 (address ATCG) (subfigure 5). From 4 to G, mark midpoint 5 (address ATCGG) (subfigure 6).

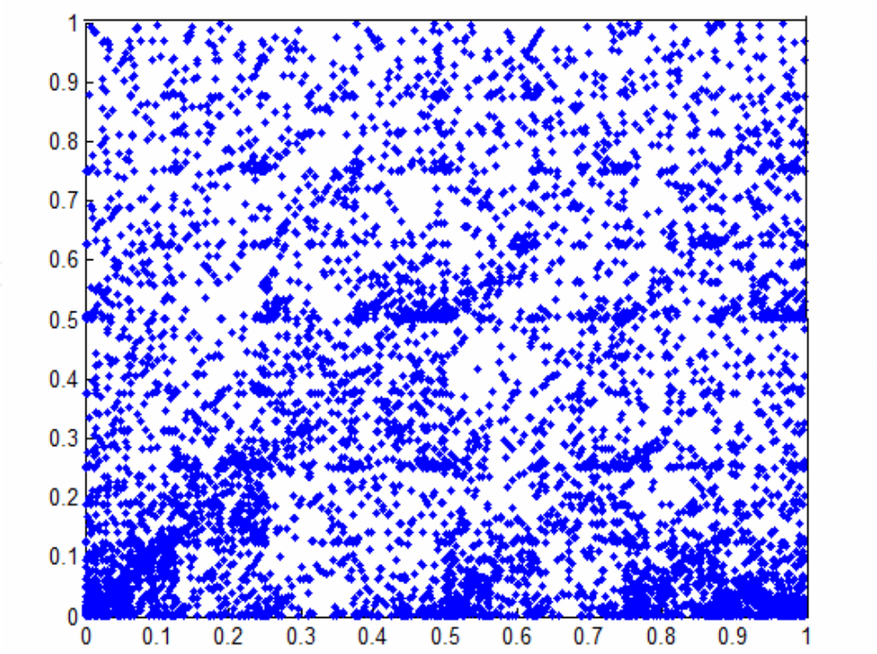



Fig. 8. Chaos Game Representation of the C. Elegans's gene F56F11.4

By identifying local patterns displayed in the CGR square, it is possible to identify correspondent features of DNA sequences (Yu et al, 2008). The fractal nature of this kind of DNA representation can be observed Fig. 8. The clustering dots in the lower corners indicate a slightly high concentration in A and T. It is known that CGR patterns depict base composition. In fact, we divide the CGR space with a grid of size k (i.e $(2^k \times 2^k)$ pixels) and we count occurrence in each quadrant, the frequency of k-lengthen words occurrence can be estimated and the frequency matrix then extracted is called FCGR (Frequency Chaos Game Representation) (Almeida et al, 2001; Deshavanne et al, 2000; Jeffrey, 1990).

The FCGR was first investigated by Deschavanne in (Deshavanne et al, 1999) and later by Almeida in (Almeida et al, 2001). To show the frequencies of the K-tuples, a color scheme normalized to the distribution of frequency of occurrence of associated patterns is used (Joseph & Sasikumar, 2006; Oliver et al, 1993; Tavassoly, 2007a; Tavassoly, 2007b; Makula, 2009; Goldman, 1993; Cénac, 2006; Tino, 1999; reference 44). A grayscale color mapping may also be used. In Fig. 9, the dinucleotide and trinucleotide frequency matrices (k ={2,3}) are obtained for the gene F56F11.4 of C.elegans. Thus, $2^2x2^2=16$ cells are needed for motifs of length two and $2^3x2^3=64$ regions to count motifs of length 3. The darker pixels represent the most frequently used words; when the clearest ones represent the fewer used words. CGRs were used for displaying the behavior of sub-patterns within the same input sequence and depicting oligo_mer composition. It forms the basis for similarity and self-similarity algorithms in a different way from traditional alignment of nucleotides.

This FCGR cannot follow the evolution of frequencies from the beginning to the end of a given sequence. So, we propose to generate signals from FCGR. We Generate the nth-order FCGR for the hole sequence, and we replace the reading the first n-lengthen word in the sequence, by the correspondent frequency of the same sub-pattern in the FCGRn matrix.

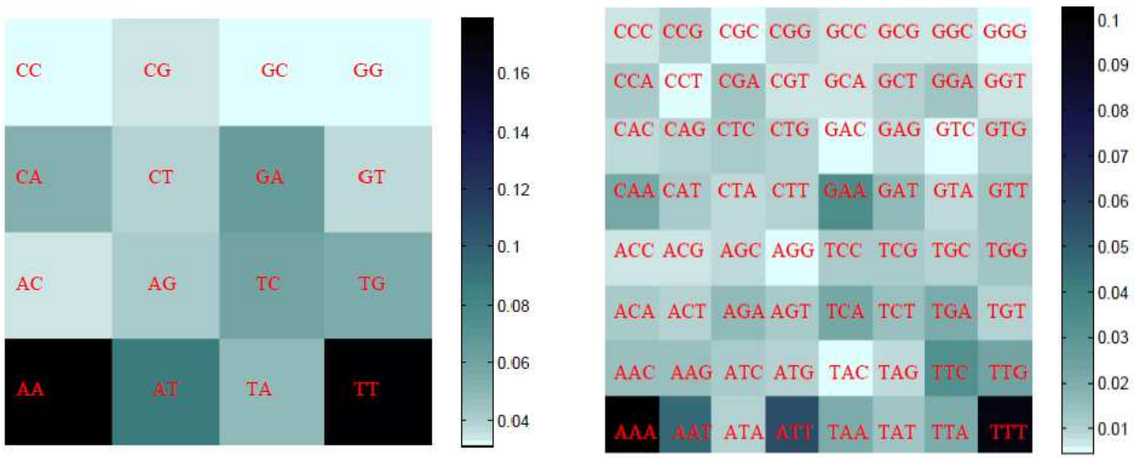



Fig. 9. The FCGR2 (k=2) and the FCGR3 (k=3) for the gene F56F11.4 of C. Elegans

Generating signals from FCGRs was a good way to capture such variability. For this fact, a new 1D graphical representation of DNA sequences is introduced, which provide useful insights into local and global characteristics of genomic sequences. This novel algorithm of DNA coding consists of computing the $k^{th}$-order FCGR for the whole sequence and assigning then the value of the correspondent frequency to each k-lengthen word in the sequence. Thus allows us to follow the frequencies' evolution along a given sequence. The the obtained plot set is called $FCGR_k$-signal.

Let's We consider the given sequence $S_{DNA}$
$S_{DNA}$ = ' TTTAAAAGCTCGCGCTAAAA'

The given sequence is divided with a k-length sliding window. A set of K-frames are obtained which are denoted by K-mers. For example when k= {2, 3, 6}, we have 2-mers ($S_{DNA}$), 3-mers ($S_{DNA}$) and 6-mers ($S_{DNA}$).

$F_K(s)$ is defined to be the frequencies' set of the k-substrings that appear in the sequence S. Obviously; these frequencies derive from the appropriate $FCGR_k$ matrices. It follows that:

$F_2(S_{DNA})$= {0.1579, 0.1579, 0.1579, 0.3684, 0.3684, 0.3684, 0.1053, 0.2105, 0.1579, 0.1053, 0.1579, 0.2105, 0.1579, 0.2105, 0.1579, 0.1579, 0.3684, 0.3684, 0.3684}

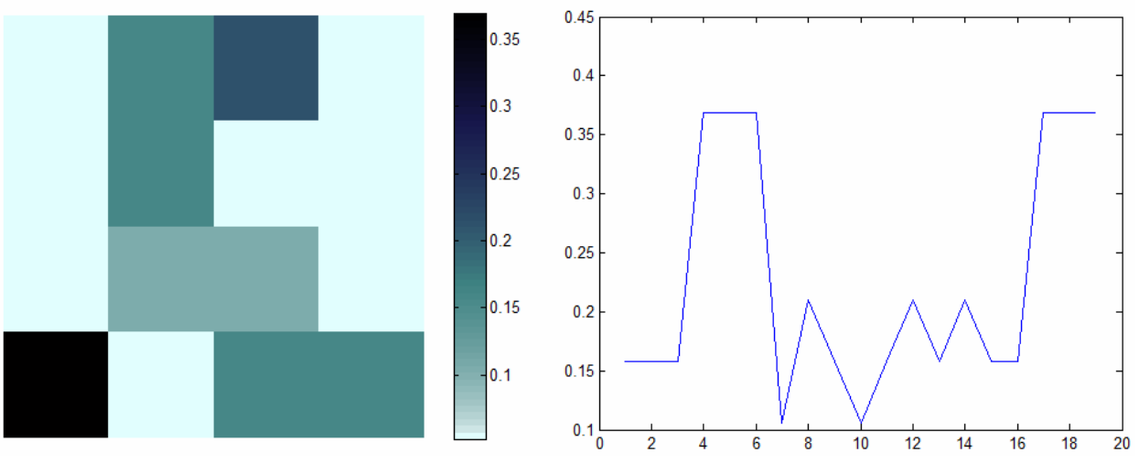
$F_3(S_{DNA})$= {0.1111, 0.1111, 0.1667, 0.2778, 0.2778, 0.1111, 0.1111, 0.1667, 0.1111, 0.1111, 0.1667, 0.1111, 0.1667, 0.1667, 0.1111, 0.1667, 0.2778, 0.2778}

$F_6(S_{DNA})$= {0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333, 0.1333}

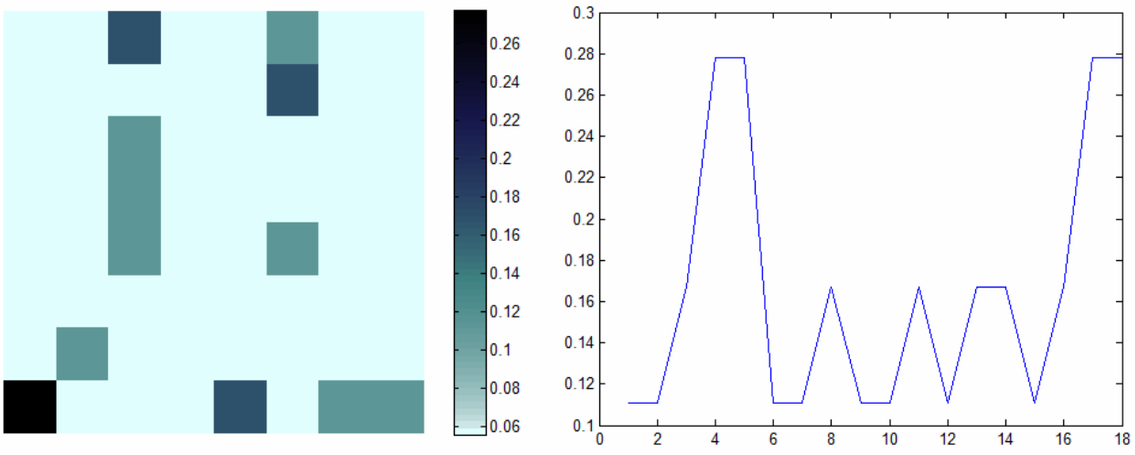
Fig. 10. illustrates the FCGR drawings for the case of k = {2,3and 6} and the corresponding plot sets for the considered sequence.

Fig. 10. Also shows the slightly high concentration in AA and AAA motifs in FCR2 and FCGR3 which are expressed by the high-rise blocks in the correspondent signals.

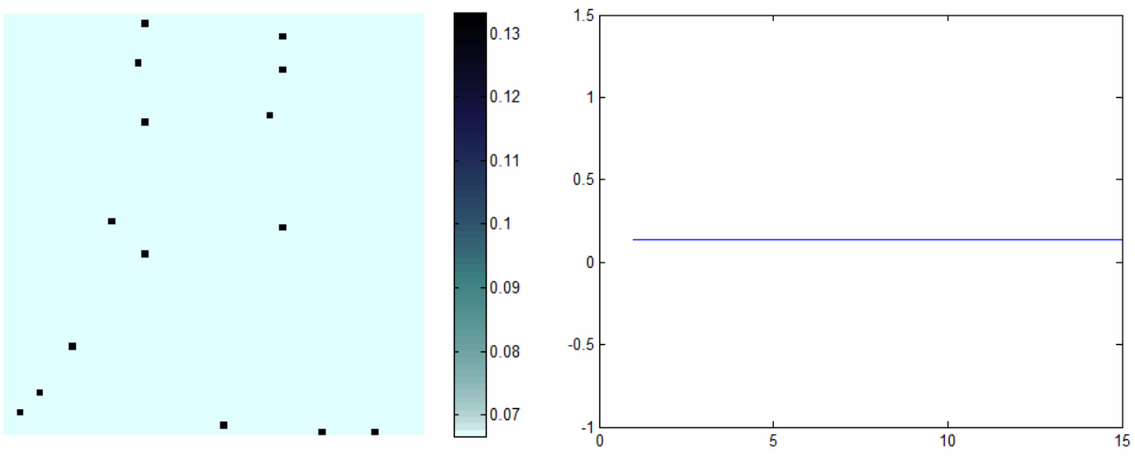
On the signals obtained, a spectral analysis is applied to detect the frequency global behaviour in the spectrum for each C. Elegans chromosome.




(a) FCGR2 (b) FCGR2-signal




(c) FCGR3 (d) FCGR3-signal




(e) FCGR6 (f) FCGR6-signal

Fig. 10. FCG representation for k=2, 3 and 6 and the FCGR_signals associated

### 3.2 The Fourier analysis method steps

The short time analysis is the technique used in order to locate specific regions in a DNA sequence. In this purpose, a mean values of Smoothed Discrete Fourier Transform is applied on sliding window along the DNA sequence to follow the peak's evolution for specific frequencies points. The Fourier analysis algorithm steps are:

The converted DNA sequence x[n] is divided into frames of M length with an overlap Δm. Each of these frames is also divided into N frames by multiplication with a sliding analysis window w[n]:

$$x_w[n,i] = x[n]w[n - i\Delta n] \tag{17}$$

Where i is the window index, and Δn the overlap. The weighting w[n] is assumed to be non zero in the interval [0, N-1]. The frame length value N is chosen in such a way that, on the one hand, the parameters to be measured remain constant and, on the other hand, that there are enough samples of x[n] within the frame to guarantee reliable frequency parameter determination. The choice of the windowing function influences the values of the short term parameters, the shorter the window the greater his influence (Mallat, 1999). We select N and M frame length as power of two to apply the Fast Fourier Transform algorithm.

Each weighted block $x_w[n]$, of the frame is transformed in the spectral domain using Discrete Fourier Transform (DFT), in order to extract the spectral parameters $X_w[k]$, where k represents the index of the frequency ([0, N-1]). The DFT of each frame (in one of M sequence parts) is expressed as follows:

$$X_w^i[k] = \sum_{n=0}^{N-1} x_w[n,i] e^{-j\frac{2\pi}{N}nk} \tag{18}$$

Using the mean values, we calculate a DFT mean value for each frame (1: M). The expression of mean DFT is expressed as:

$$Xm_w^j[k] = \frac{1}{N} \sum_{i=0}^{N-1} X_w^i[k] \tag{19}$$

Where *i* correspond to the index frame of N frames ([1...N]), *k* is the index of the frequency and *j* correspond to the index frame of M frames ([1: M]).
We constitute the matrix

$$MAT(j,k) = Xm_w^j[k] \tag{20}$$

With these obtained values, we can constitute the matrix to represent restricted join time frequency information, known as 2D or 3D DNA spectrograms. This 2D or 3D representation consists of the spectrogram amplitude for a specific index periodicity in a specific nucleotide position in the chromosome.

## 4. Results

The method has been applied on C. Elegans genome. The chromosomes have been divided on 1- million's parts. The M frames have a length of 1024 bp and an overlap Δm=256, the

N frames of each M frames have length of 256 with Δn =128. The fig. 11 presents some examples for the spectrum related to each of the three coding technique used. In this figure, we show particularly the periodicities 3 and 10 which are closely depending on coding.

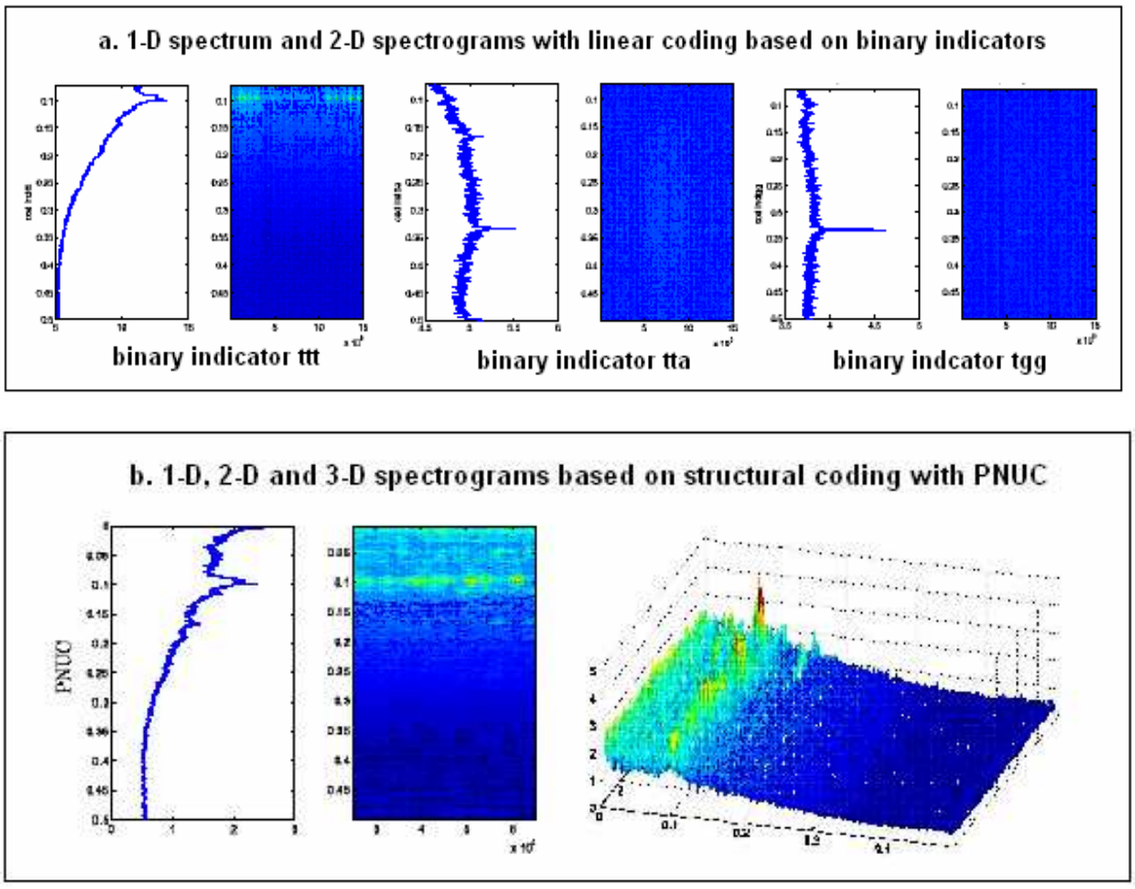



Fig. 11. Examples of spectrums and spectrograms generated with a mean valued technique based on smoothed Discrete Fourier Transform applied on sliding window along the DNA sequence parts of C. elegans genome. Two coding methods are used: a- linear coding technique (binary indicator) (subfigure a), b- structural coding technique (PNUC) (subfigure b)

In order to highlight the various frequencies characteristic of an organism, the tests were carried out with various coding over various sizes of segments and various widths. The example presented in the Table 3 presents the percentage of contribution of the trinucleotides in the highlighting of the various characteristic frequencies at the frequencies 1/3, 1/6.5, 1/9 and 1/10. The table shows that the organism C. Elegans is rich in periodicities and that these periodicities are raised by more than the 3/4 of these coding technique. We notice clearly that for periodicity 3, the rate has raised more 97 %, followed by periodicity 10 which has 90 % and periodicity 9 with 85 %. Periodicity 6.5 is a periodicity which is very marked for this organism 70 % of code contributes to its raising. It translates the existence with a very high rate of 6 bases groups at the periodicity 6. The majority of these groups represent polyA, generally associated for gene purposes.

| period | chromosomes | | | | |
|---|---|---|---|---|---|
| | Ch1 | Ch2 | Ch3 | Ch4 | Ch10 |
| P=3 | 96.8% | 96.8% | 96.8% | 96.8% | 96.8% |
| P=6.5 | 64% | 60.9% | 81.25% | 73.43% | 71.9% |
| P=9 | 85.9% | 89% | 89.1% | 82.81% | 79.7% |
| P=10 | 70.3% | 90.6% | 90.5% | 90.6% | 84.3% |

Table 3. The proportion of contribution of the trinucleotides in the highlighting of the various characteristic frequencies

The Fig. 11 presents some spectrum with linear coding based on binary indicator. Each indicator contributes on a specific periodicity enhancement. The ttt binary indicator enhances the periodicity 10 when for the indicators tta et tgg the periodicity 3 is picked up. The 3D spectrograms give more precision on the power's spread around these periodicities. In fact, the peaks in these frequency locations have different power values (Fig. 12).

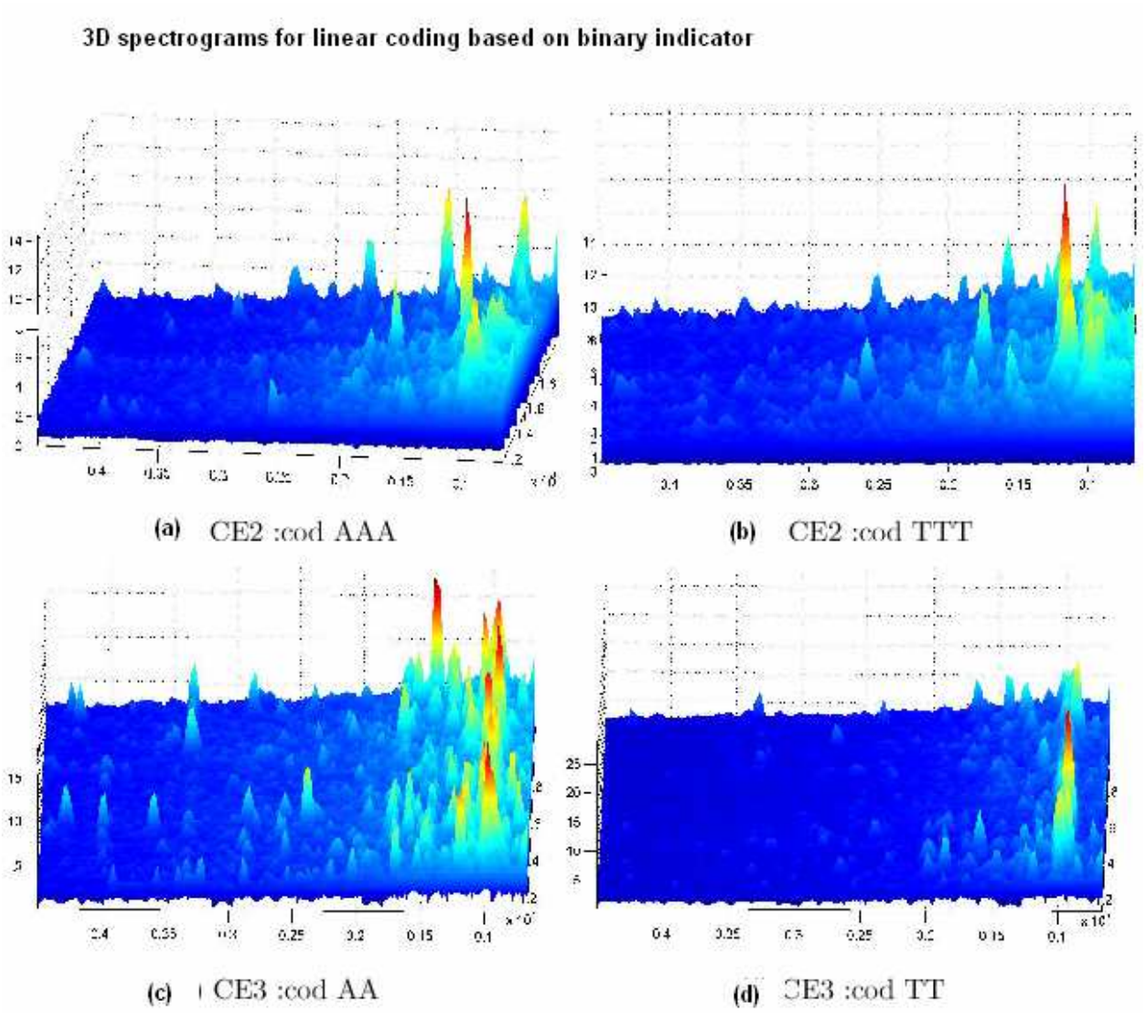



Fig. 12. 3-D spectrograms for binary indicators coding

The spectrogram 3D adds a third element to the representation 2D. In addition to the localization of the periodicities in the segment, we visualize power associated with each peak. We can distinguish between the peaks which we can find in all the segments for a given periodicity: 10 and 3 and the peaks which are present in certain segments and which were eliminated by carrying out the average

The Fig. 12 is divided on 4 subfigures. Each one add to the 2D spectrograms the power values and locations of the periodicities Enhanced: it represents the 3-D spectrograms. Subfigures (a) and (b) are related to chromosomes2 of C. Elegans when the subfigures (c) and (d) concern chromosome 3.

This figure shows that for the binary indicator 'AA', 'TT', 'AAA' and 'TTT', the peaks around the frequency 1/10.5 are very pronounced. The variation of the degree view angle demonstrates that the peaks are locally spread in the chromosome part. In the literrature, it has been demonstrated both with the biochemical and signal processing studies, that the periodicity 10.5 related to the nucleosomes is varying. That's why, these figures shows in one hand that there is peaks around this periodicity and in the other hand the peaks are spread in specific regions in the chromosome.

The Fig. 13 represents the spectrograms recovered after PNUC coding. The analysis breaks up the chromosome made up of 15,2Mbp into 15 parts of 1Mbp. We find the localization of the periodicity in the ends. In reality, the periodicity peaks are missed or have of very weak power in the sequence going of 6 Mbp with 12 Mbp, it is not localized on the centromer but it is around ends of the helix. We find it towards the position 13 Mbp until the end. In the parts where it exists it is not continuous, it is localized in specific time's lapses.

In Fig. 14 mean valued technique based on smoothed Discrete Fourier Transform was applied along the parts 6, 9 and 13 of the chromosome 1 of C.elegans. From the 1D, 2D and 3D plots, it is observed that coding with $FCGR_2$ reveal the presence of both 10.5 and 3 periodicities. The peaks are spread with different values according to parts around each of these periodicities. Each part has each own specificity. In fact, in part 9 (subfigure a) , periodicities 3 and 10 just submerge from the frequency behavior with peaks of modest values. For the part 6 these periodicities have the same behavior, the specificity is the presence of horizontal peaks around the location 750 in this part. When the part 13 is rich in periodicities 10 and 12 and poor in periodicity 3.

For coding with FCGR-3 (Fig.15), the very pronounced peaks correspond to the 10.5 periodicity; just in the left side other peaks appear around the frequency 0.11 which corresponds to the 9 periodicity; in the right side a few peaks occur around the frequency 1/12. The 3 periodicity disappears in the majority of the parts and when it appears, it is present only on a few areas with very low amplitudes. In Fig 15, we can distinguish between frequency behavior in the three parts represented. The periodicity 10 is more pronounced for the part 16 (subfigure b) when comparing with part 9 (subfigure a) and 10 (subfigure c).

As for the hexamers coding ($FCGR_6$), we find that it enhances the frequency 1/10.5; upon rare zones the frequency 1/12 is observed (Fig.16). We clearly notice that this coding technique enhances the periodicity 10 and his neighbor in opposition to periodicity 3. The three parts shows different aspect of the repartition of the periodicities. In part 9 (subfigure a), the peaks are spread in a "large" frequency band around periodicity. The band is reduced for part 16 (subfigure b) to be located in two frequencies then the power is grouped in one frequency for part 12 (subfigure c).

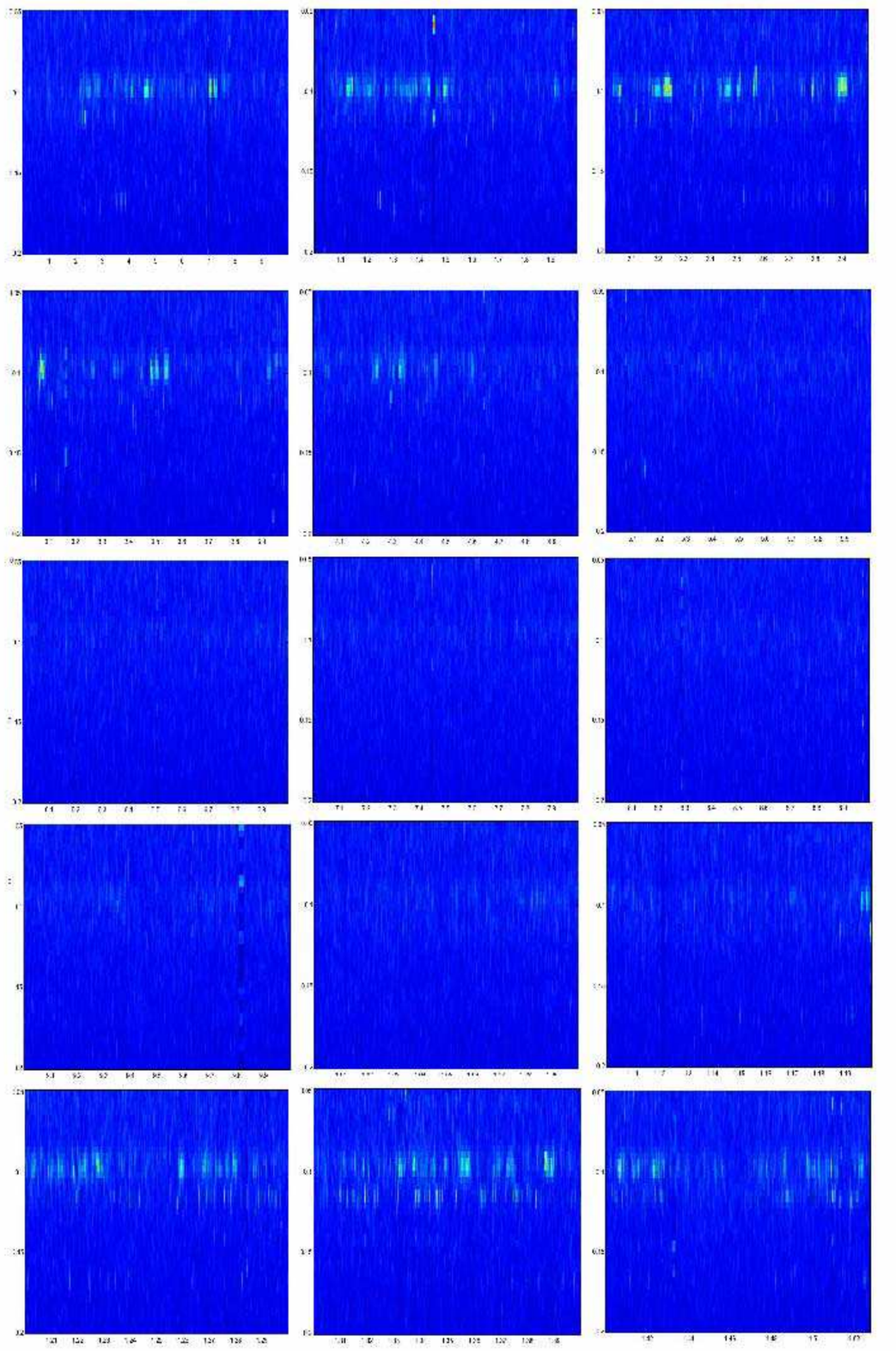
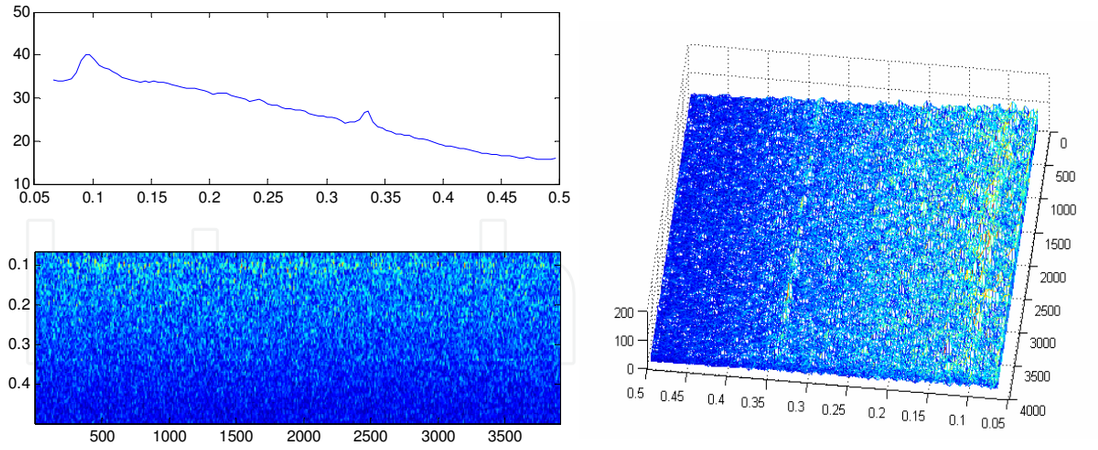
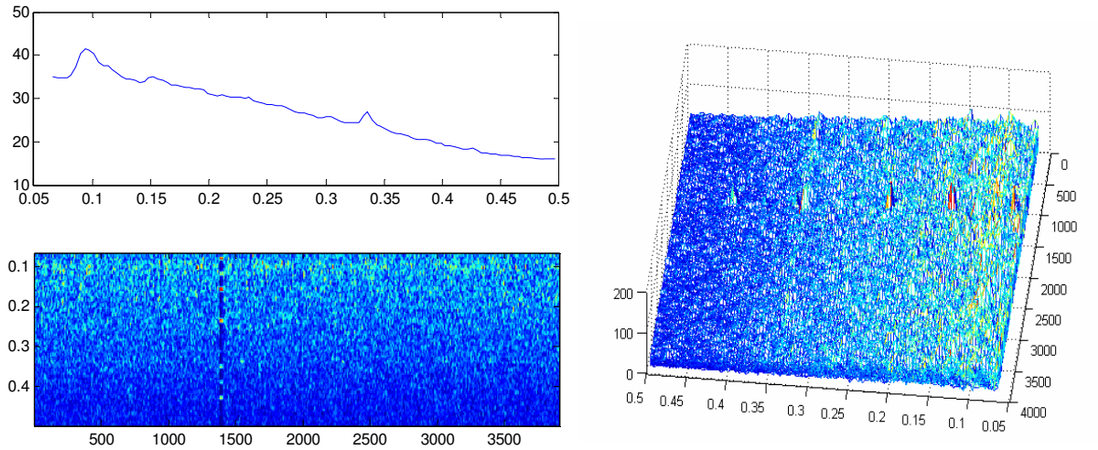


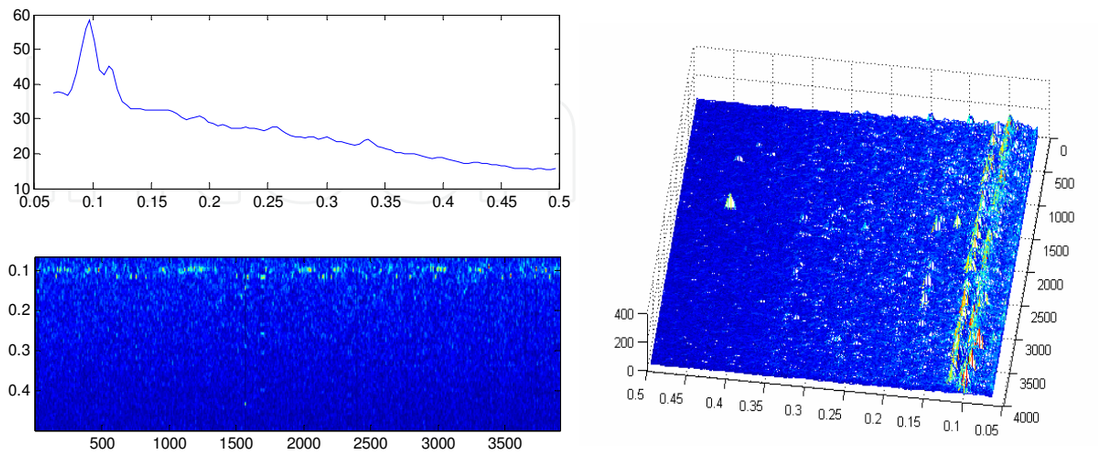
Fig. 13. Distribution of periodicity 10 for coding pnuc along chromosome 2



a- Fourier analysis of part 9 of chromosome 1

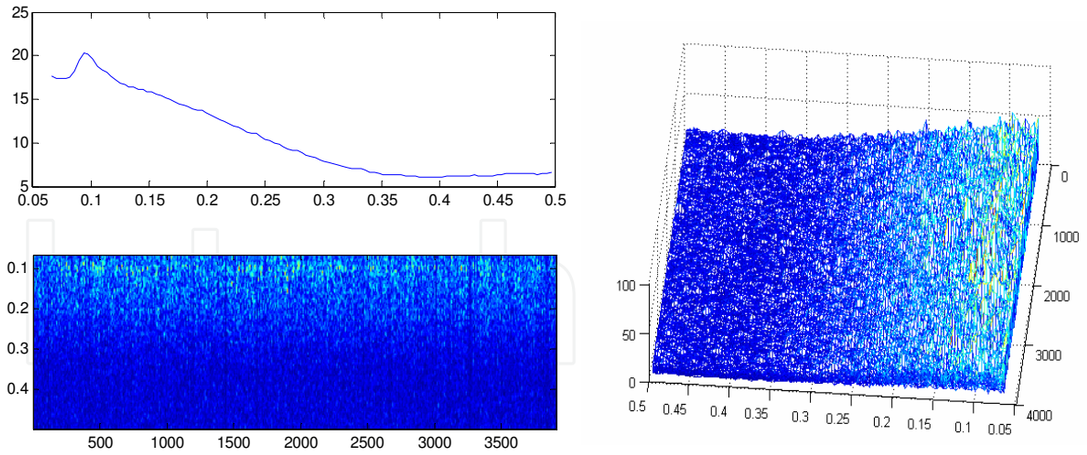


b- Fourier analysis of part 6 of chromosome 1

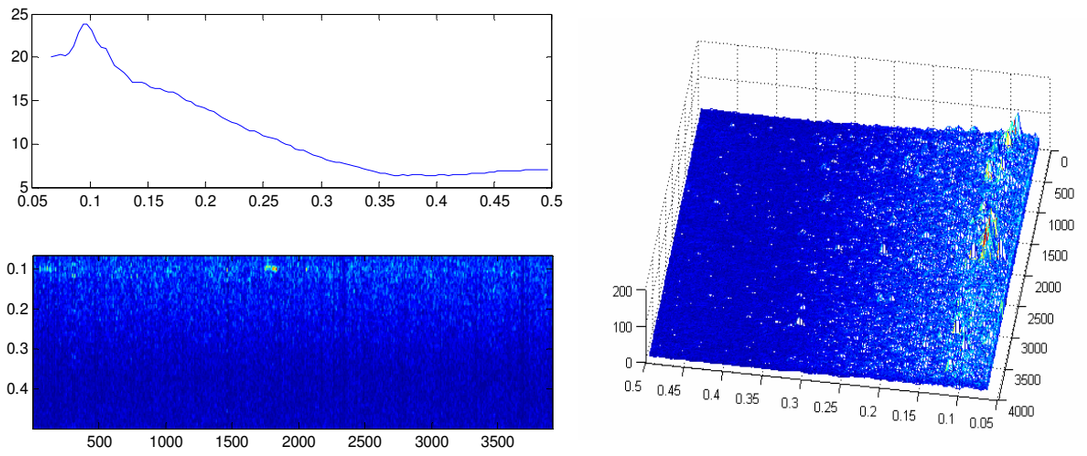


c- Fourier analysis of part 13 of chromosome 1

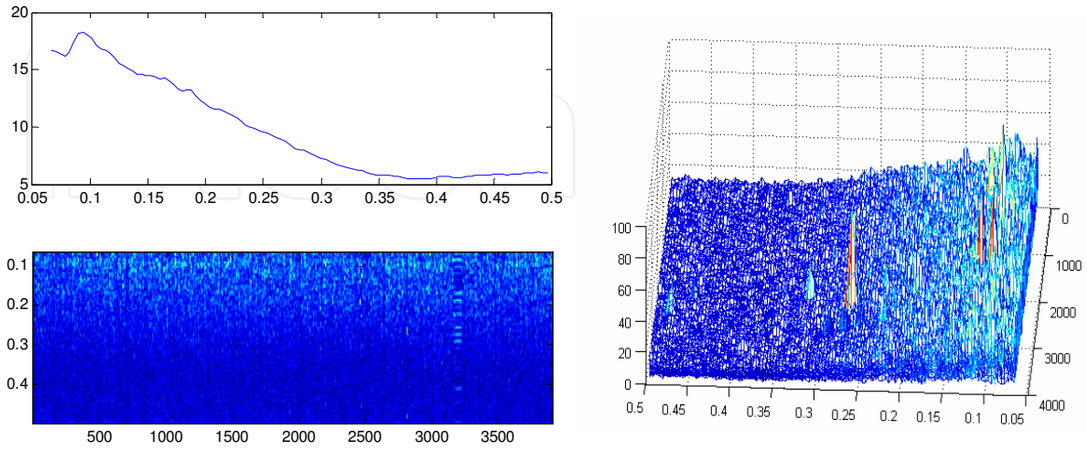
Fig. 14. Examples of spectrums and spectrograms of chromosome's parts with FCGR-2 signal coding




a- Fourier analysis of part 9 of chromosome 1

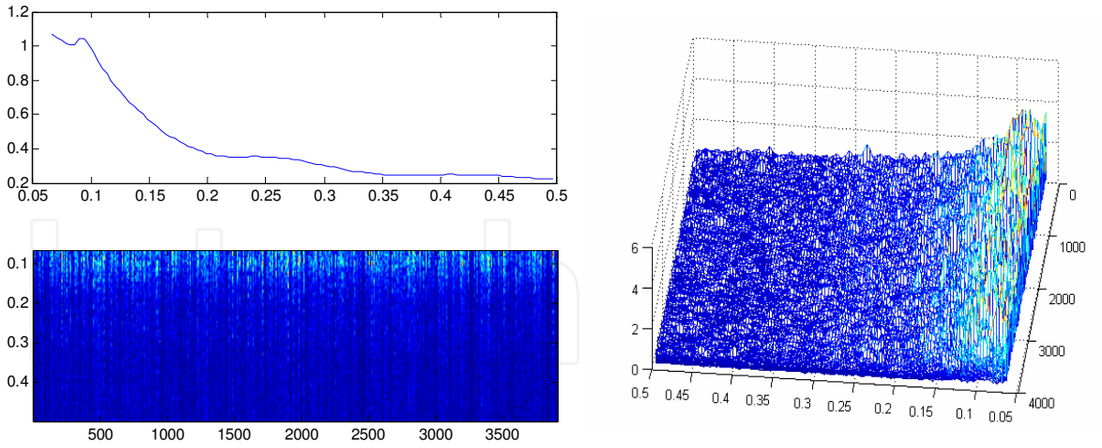



b- Fourier analysis of part 16 of chromosome 1

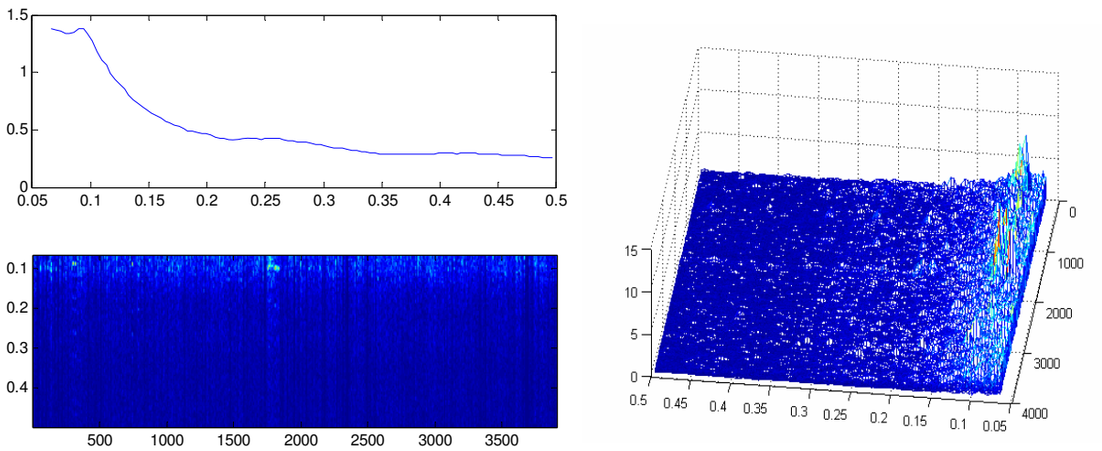



c- Fourier analysis of part 10 of chromosome 1

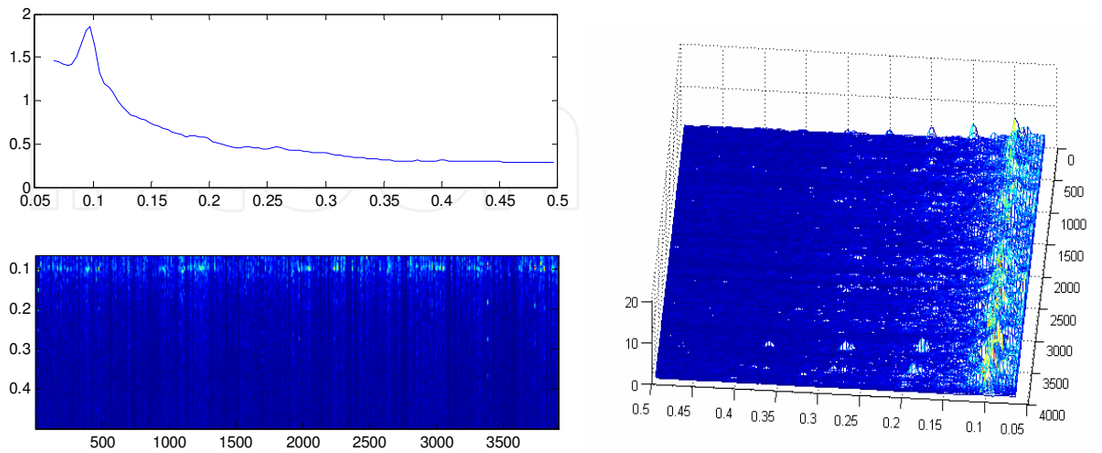
Fig. 15. Examples of spectrums and spectrograms of chromosome's parts with FCGR-3 signal coding




a- Fourier analysis of part 9 of chromosome 1




b- Fourier analysis of part 16 of chromosome 1




c- Fourier analysis of part 12 of chromosome 1

Fig. 16. Examples of spectrums and spectrograms of chromosome's parts with FCGR-6 signal coding

A peak around the frequency 1 / 4 nearby at position 2500 corresponds to a satellite (Fig. 16 subfigure a). This frequency derives from repetitions of certain dinucleotides in the area. The spectrogram reveals the presence of a satellite with multiple frequencies; this is manifested clearly in the 3D graph in the form of horizontally aligned peaks colored in red, the higher frequencies.

## 5. Conclusion

In This chapter, we investigate the contribution of each coding technique: the linear, the two-dimensional and the structural one in the enhancement of the peaks related to the C. elegans genome periodicities. For this purpose, we use a mean values of smoothed Discrete Fourier Transform applied on sliding window along the DNA sequence to follow the peak evolution for specific frequency points around the frequencies. We detect periodicities around 3, 6, 9 and 10 and found periodicities 3 and 10 related respectively to genes and the positions of the nucleosomes. First we evaluate the frequencies spread through the chromosomes with a 1-D spectrum. Second, we consider the 2-D and 3-D DNA spectrograms to visually detect the specific parts of chromosomes related with protein coding regions, nucleosomes positioning regions, and other particular regions.

The time frequency analysis made it possible to follow the periodicities' evolution. We studied the contribution of a range of binary indicators for the raising of exons' peak frequency. We also studied the localization of the areas being able to form nucleosomes. Thanks to the spectrogram with two dimensions, we visualized the localization of the areas corresponding to periodicity 10 in the limits and not in the center of the helix. The three-dimensional spectrogram showed that the raised peaks do not correspond to the periodicity 10 but we see clearly in certain sequences and for some indicators two lines of peaks of variable powers around this periodicity. This result can explain the variation between 10 and 10.7 of the periodicities associated with the nucleosomes presented in the literature. It is also observable that these peaks are alternated around two periodicities; this result could be associated with the phenomena of chromatin compaction.

## 6. References

Almeida, J.S., Carrico, J.A., Maretzek, A., Noble P.A. & Fletcher M. (2001) "*Analysis of genomic sequences by Chaos GameRepresentation*", Bioinformatics Vol. 17, n°5, pp 429–437.

Anastassiou D. (2001), "*Genomic Signal processing*", IEEE Signal Processing Magazine, 18 (4), pp: 8-20.

Berger J. A., Mitra S. K.& Astola J. (2003), "*Power spectrum analysis for DNA sequences*", Proc. of ISSPA 2003, pp 29-32, France, 1-4 July.

Cénac, P. (2006) "Étude statistique de séquences biologiques et convergence de martingales", PhD thesis on Applied Mathematics, Paul Sabatier University, Toulouse III, pp 17−25.

Cénac P., Fayolle G., Lasgouttes J.M., (2004) " *Dynamical Systems in the Analysis of Biological Sequences*", research report n° 5351,pp 3–50.

Cohanim, A.B., Kashi Y. & Trifinov E.N. (2005), "*Yeast Nucleosome DNA Pattern: Deconvolution from Genome Squences of S. cerevisiae*, Journal of Biomolecular Structure & Dynamics ISSN 0739-1102, volume 22, Issue Number 6, Adenine Press, pp: 687-693.

Deschavanne, P., Giron, A., Vilain, J. Dufraigne, CH., & Fertil, B. (2000), "*Genomic Signature Is Preserved in Short DNA Fragment*", International Symposium on Bio-Informatics and Biomedical Engineering, IEEE, pp 161–167.

Deschavanne, P., Giron, A., Vilain, J., Fagot, G. & Fertil, B. (1999) "*Genomic signature: characterization and classification of species assessed by chaos game representation of sequences*", Mol Biol E, Vol 16, n°10, pp 1391–1399.

Fiser, A., Tusnady, G.E.& Simon, I.(1994) "*Chaos game representation of protein structures*", J.Mol Graphics, Vol 12, pp 295,302–304.

Fukushima, A., Ikemurab, T., Kinouchie, M., Oshima, T., Kudod, Y., Morig, H. & Kanaya, S. (2002) "*Periodicity in prokaryotic and eukaryotic genomes identified by powerspectrum analysis*", Elsevier, Gene 300, pp 203–211.

Goldman, N. (1993) "*Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences*", Nucleic Acids Research Vol.21, n°10, pp 2487–2491.

Godsell, D.S. & Dickerson, R.E. (1994) "*Bending and curvature calculations in b-dna*", Nucl. Acids Res, vol 22, pp 5497-5503.

Hayes, J.J., Tullius, T.D. & wolffe, A. P. (1990) "*The structure of DNA in a nucleosome*", Proceedings of the National Academy of sciences of the United States of America, vol 87 No 19, pp 7405-7409, October.

Jeffrey, H.J., (1990) "*Chaos game visualization of sequences*", Computers & Graphics, Elsevier, Vol.16, n°1, pp 25–33.

Joseph, J. & Sasikumar, R. (2006) "*Chaos game representation for comparison of whole genomes*", BMC Bioinformatics, Vol 7, n°1, pp 1-10..

Kornberg, R.D. (1974), "*Chromatin structure: a repeating unit of histones and DNA*." Science 184, pp:868-871,

Kornberg, R.D. (1977), "*Structure of Chromatin*", Annu Rev Biochem. 46 , pp 931-954.

Makula, M., (2009) "*Interactive visualization of oligomer frequency in DNA*" Computing and Informatics, Vol. 28, pp 1001–1016.

Nicorici, D., Berger, J. A. , Astola, J. & Mitra, S. K. ,(2003), "*Finding borders between coding and non coding DNA regions using recursive segmentation and statistics of stop codons", Finnish Signal Processing Symposium (FINSIG'03)*, Tampere, Finland, pp. 231-235.

Oliver, J. L., Bernaola-Galvan, P., Guerrero, G. & Foman-Roldan, R. (1993), "*Entropic profiles of DNA sequences through chaos-game-derived images*," J. Theor. Biol.,Vol 160, n°4, pp 457–470.

Oppenheim, A. V., Schafer ,R. W. & Buck, J. R., (1999) "*Discrete Time Signal Processing*", 2nd Edition, Prentice Hall.

Oudet, P., Germond, J.E., Bellard, M., Spadafora, C. & Chambon, P. (1978) "*Structure of Eucaryotic Chromosomes and Chromatin",* Phylosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol 283, No 997, pp: 241-258,

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thamstrom, A., Field, Y., Moore, I. K., Wang, J.P.Z. & widom, J. (2006) " *a genomic code for nucleosome positioning, nature* vol 442, pp: 772-778, August

Sussillo, D., Kundaje, A. & Anastassiou, D., 2003 "*Spectrogram analysis of genomes*," *Eurasip* Journal of Applied Signal Processing, vol. 2003, no. 4, .

Tavassoly, I., Tavassoly, O., Rad, M.S.R & Dastjerdi, N.M. (2007a) "*Multifractal Analysis of Chaos Game Representation Images of Mitochondrial DNA*", Frontiers in the Convergence of Bioscience and Information Technologies FBIT, IEEE, pp 224–229.

Tavassoly, I., Tavassoly, O., Rad, M.S.R & Dastjerdi, N.M. (2007b) "*Three dimensional Chaos Game Representation of genomic sequences*", Frontiers in the Convergence of Bioscience and Information Technologies FBIT, IEEE, pp 219–223.

Tino, P. (1999) "*Spatial representation of symbolic sequences through Iterative Function System*", IEEE, Vol 29, n°4, pp 386–393.

Trifonov, E. N. & Sussman, J.L. (1980) " *The Pitch of chromatin DNA is Reflected in its Nucleotide Sequence*", Proceedings of the National Academy of Sciences of the United States of America, Vol 77, No 7, part 2: Biological Sciences, pp:3816-3820.

Trifonov, E. N. (1998), "3-*, 10.5-, 200- and 400-base periodicities in genome sequences*", Elsevier Physica A 249, pp :511-516,.

Vaidyanathan, P. P. & Yoon, B. J. (2004) "*The role of signal processing concepts in genomics and proteomics*", Journal of the *Franklin Institute (Special Issue on Genomics*), vol. 341, pp. 111-135.

Widom, J. (1996) "*Short-range Order in Two Eucaryotic Genomes: Relation to chromosome Structure*" J. Mol. Biol. 259 pp 579-588

Worcel, A., Strogatz, S. & Riley, D. "*Structure of chromatin and the linking number of DNA*", Proceedings of the National Academy of Sciences of the United States of America, Vol 78, No 3, part 2: Biological Sciences, pp:1461-1465, 1981

Yu, Z.G., Shi, L., Xiao, Q.J. & Anh, V., (2008) "*Chaos game representation of genomes and their simulation by recurrent iterated function systems*", Bioinformatics and Biomedical Engineering ICBBE, IEEE , pp 41–46.


**Fourier Transform Applications**

Edited by Dr Salih Salih

ISBN 978-953-51-0518-3
Hard cover, 300 pages
**Publisher** InTech
**Published online** 25, April, 2012
**Published in print edition** April, 2012

The book focuses on Fourier transform applications in electromagnetic field and microwave, medical applications, error control coding, methods for option pricing, and Helbert transform application. It is hoped that this book will provide the background, reference and incentive to encourage further research and results in these fields as well as provide tools for practical applications. It provides an applications-oriented analysis written primarily for electrical engineers, control engineers, signal processing engineers, medical researchers, and the academic researchers. In addition the graduate students will also find it useful as a reference for their research activities.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Afef Elloumi Oueslati, Imen Messaoudi, Zied Lachiri and Noureddine Ellouze (2012). Spectral Analysis of Global Behaviour of C. Elegans Chromosomes, Fourier Transform Applications, Dr Salih Salih (Ed.), ISBN: 978-953-51-0518-3, InTech, Available from: http://www.intechopen.com/books/fourier-transform-applications/spectral-analysis-of-global-behaviour-of-c-elegans-chromosomes

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen