3D Spectrum Analysis of DNA Sequence: *Application to Caenorhabditis elegans Genome*

Afef ELLOUMI OUESLATI, Noureddine ELLOUZE

Unité Signal, Image et Reconnaissance de Formes,
Département de Génie Electrique,
ENIT, BP 37, Campus Universitaire, Le Belvédère, 1002,
Tunis, Tunisie
afefelloumi@gmail.com, n.ellouze@enit.rnu.tn

Zied LACHIRI
Département de Génie Physique et Instrumentation
INSAT, BP 676, Centre Urbain Cedex, 1080, Tunis, Tunisie
zied.lachiri@enit.rnu.tn

***Abstract*— a Fourier Transform Technique has been used to enhance the genome periodicities when analyzing the distributions of independent nucleotides and dinucleotides. These periodicities are varying from 2 to 500bp. In this paper we focus on the 3 and 10.5 periodicities. The 3-base periodicity is characteristic for the protein-coding sequences only. The source of the approximately 10.5-base sequence period is related to the deformability of DNA. In fact, DNA folding in chromatin is facilitated by periodical positioning of some dinucleotides along the sequences, with the period close to 10.5 bases. When the DNA sequence is encoded for the signal 'AA' or 'TT' or 'TA' the peak at 10.5 is locally strengthened. For the Caenorhabditis elegans (C. Elegans) genome, this peak becomes the dominant feature in the transform. Studying one organism's genome requires three steps. First, the DNA coding method: the DNA's string data are transformed into numerical signal. Second, periodicities are detected by spectral analysis. Third, a 3D graphical representation allows following the evolution of this periodicity along the genome and facilitating the specific regions location.**

Keywords- Fourier Transform Technique; detecting specific periodicities; 3D graphical representation; specific region's location

I. INTRODUCTION

Genomic DNA sequence contains many signals that have not yet been detected or not well interpreted. Many techniques have been developed to analyze the genomes in both their similarity and their specific characters. DNA sequence is a string of nucleotides or bases that we represent usually by four letters, namely A (adenine), T (thymine), C (cytosine), G (guanine). These foregoing bases define the composition of all protein coding region. Consequently, they form the key of the biochemical function of all living organisms. Applying the signal processing techniques on the character string of the DNA sequence needs the conversion of the symbols into numerical sequences. Hence, different techniques come to light to achieve this particular conversion. As examples of these techniques, we mention the tetrahedral representation [10], the weighted real and complex value [2] and [3], the DNA walks [4, 5, 6], and using a grammar of alphabets [14]. The aim of each coding method is to enhance the hidden information for further analysis.

This paper is aimed to study the contribution of the signal processing methods on the analysis of some specific periodicities in DNA sequences. The work of E.N. Trifonov [21], related with the periodicities revealed by signal processing methods, has resumed their origin in the genomic sequence. The main specification rests on genomic signal processing in [2, 3]. Genomic signal processing consists on the processing of DNA sequences, RNA sequences and proteins. In this work, we proceed by following these periodicities evolution along a gene. We also study the role of some base-groups' indicators in these periodicities enhancement.

DNA spectral analysis based on Fourier Transform and correlation study are techniques proposed in the literature [2, 3, 4, 5, 8, 14, 19, 20, 21, 22, and 23]. It contributes in the systematic search of special DNA patterns which may correspond to biological important markers. By depicting the frequencies by a 3D spectrogram representation, specific regions appear distinctly. In this paper, we are concerned with the periodicity 3 and 10.5. The periodicity 3 is related with protein coding regions (called exons) in the gene. This periodicity is given by the combination of three bases, which are involved in the constitution of amino acids in the protein synthesis process. We compare the contribution of these groups and the other ones in this periodicity. And the second periodicity in which we will focus is the 10.5 periodicity. This periodicity is related with nucleosome's positions in the DNA sequence and the degree of deformability of the sequence in the DNA helix [11, 18, 21, 23, and 24]

This paper is divided into 6 sections. Section 2 introduces an overview on the DNA, detailing the different parts in a sequence, the protein synthesis process and the structure of DNA in a nucleosome. Section 3 focuses on the DNA coding methods: the conversion of the symbolic chain of character to numerical sequences using binary indicator sequence for each group adopted. Section 4 exposes the adopted spectral analysis steps. In section 5, we present some 3D spectrograms for the C. elegans genome and we make a comparison with the different results obtained for different chromosomes and different binary indicator. We conclude by section 6, which summarize the main results of this paper.

1-4244-1509-8/07/$25.00 ©2007 IEEE

II. THE DNA : WHAT IS SPECIFIC TO THE PROTEIN CODING REGIONS AND THE NUCLEOSOME POSITIONING REGIONS?

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 deoxyribose 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.

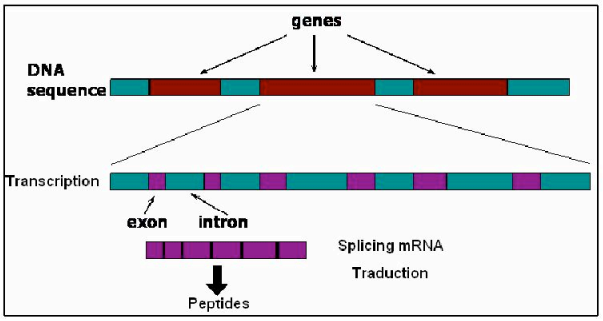

Figure 1. The protein's synthesis steps

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 (Fig.2). A model of this DNA structure in such regions is proposed by Kornberg [12, 13]. The chromatin is a dynamic structure, oscillating between the nucleosome and open structures depending on the environmental conditions [12, 13, and 16]. 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 (Fig.3). This number is varying according to the chromatin's origin [11, 12, 16, and 24]. In contrast a particle named 'nucleosome core' is invariant in its DNA content about 146 base pairs. Interesting electron microscopic evidence elaborated in [16] 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 [1]

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

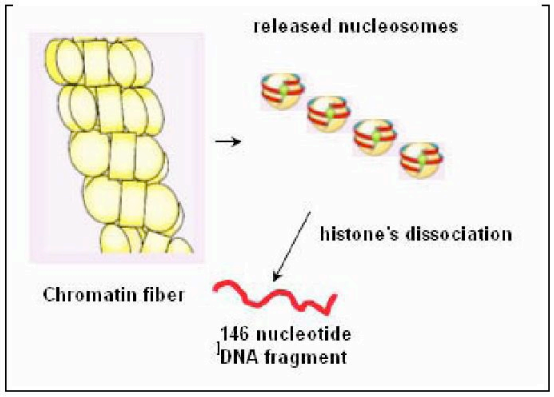

Figure 2. Chromatine's structure

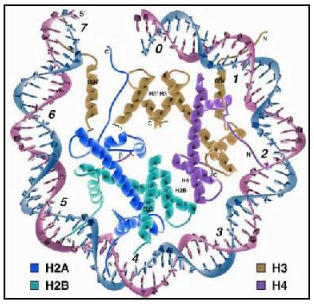

Figure 3. nucleosome's structure

III. DNA'S SYMBOLIC DATA CODING METHODS

In our analysis, we want to enhance the signals generated by the protein coding regions and the nucleosome regions. For that purpose we will consider each codon (three base group coding for amino acids) in studying the periodicity 3. The periodicity 10.5 specific to nucleosome is enhanced for specific dinucleotide as demonstrated in the studies with Fourier analysis and correlation studies [2, 4, 5, 19, and 22]. In our paper we will consider these dinucleotides, 'AA', 'TT', 'TA' and we will also consider the 'GC' signals. Thus, we extend this study to 68 base groups composed by the whole codon base groups (64 indicators) and the 4 dinucleotides named before. In order to evaluate the contribution of each one for the two periodicities 3 and 10.5, we attribute a binary value for each unit of 68 indicators. Note that the 68 indicators are included in {'AA', 'TT', 'TA', 'GC', 'AAA'… 'GGG'}

$$x[n] = u_{ind}[n] \tag{1}$$

Where *ind* in (1) represents "binary indicator sequences". This marker 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.

For the DNA sequence S, we generate 68 binary indicator sequences as follows:

$$S = 'aaaattacgc \quad ttgcatcgtg \quad tgc'$$

$$u_{aa} = 11100000000000000000$$

$$u_{aaa} = 11000000000000000000$$

On the signals obtained, we proceed to a spectral analysis to detect the periodicity in the spectrum of each indicator to enhance the exon's location.

IV. The Fourier analysis method steps

We consider the short time analysis for the specific frequencies 1/3 and 1/10.5 in order to locate specific regions in a DNA sequence. In this purpose, we have used a mean values of Smoothed Discrete Fourier Transform 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]:

$$x_w[n,i] = x[n]w[n - i\Delta n] \tag{2}$$

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 [15]. 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:

$$X_w^i[k] = \sum_{n=0}^{N-1} x_w[n,i] e^{-j\frac{2\pi}{N}nk} \tag{3}$$

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

$$Xm_w^j[k] = \frac{1}{N}\sum_{i=0}^{N-1} X_w^i[k] \tag{4}$$

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

$$MAT(j,k) = Xm_w^j[k] \tag{5}$$

We represent with this matrix the restricted join time frequency information , known as 3D DNA spectrograms. This 3D representation consists of the spectrogram amplitude for a specific index periodicity in a specific nucleotide position in the chromosome.

V. Experiments and Results

In Order to evaluate the method, computer simulation is carried out. The application of this method and the 3D spectrogram concern the C. elegans genome (Fig.4).

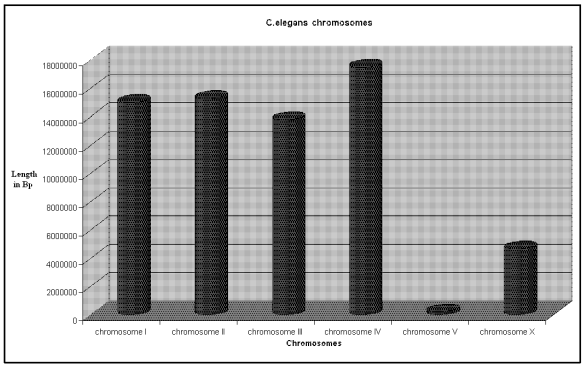

Figure 4. The C. Elegans chromosomes constitution in Bp

First, we consider the mean DFT in (4) for the three bases binary indicator and we evaluate the contribution of these groups in some periodicities enhancement in each chromosome (Fig.5). the main peridicities detected in the global frequency behavior are around 3, 6, 9 and 10.

Second, we present the results obtained for the 1 million parts of the chromosomes 1, 2, 3, 4 and 10 for the base pairs index indbp varying from 1000000 to 2000000. Each of 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. So in this case M= 3906 frames and N=8 .

As presented before, the experiments are made for DNA signal's obtained by binary indicator sequences of the dinucleotides : 'AA', 'TT', 'TA' which are the dinucleotides facilitating the DNA deformability for the nucleosome's constitution (Fig.6). And we want to study the 'GC' dinucleotide and its frequency behavior.

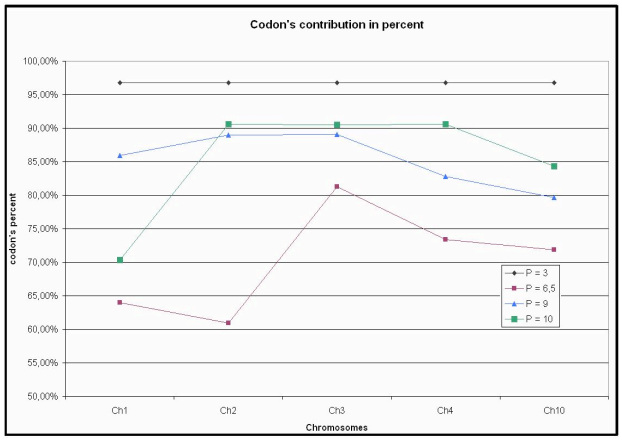

Figure 5. The codon's contribution in the periodicities enhancement for each chromosomes

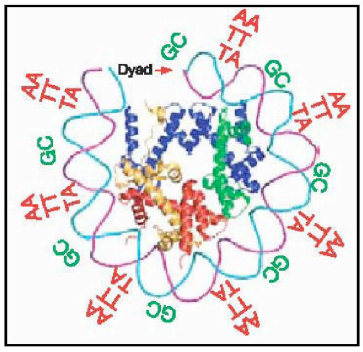

Figure 6. Dinucleotide position in the DNA structure of nucleosomes

For the periodicity 3, we will consider the 64 combinations of codons to evaluate the contribution of each one in the enhancement of the peaks in the protein coding regions.

The 3D spectrograms represent for a coding signal the spectrum amplitude for a nucleotide position at normalized frequency.

The Fig.7 shows that for the chromosome 1 of C. elegans and the binary indicator 'AA' and 'TT' (Fig 7 (a),(b)) the peaks around the frequency 1/10.5 are very pronounced. The variation of the degree view angle (Fig 7 (c)) 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.

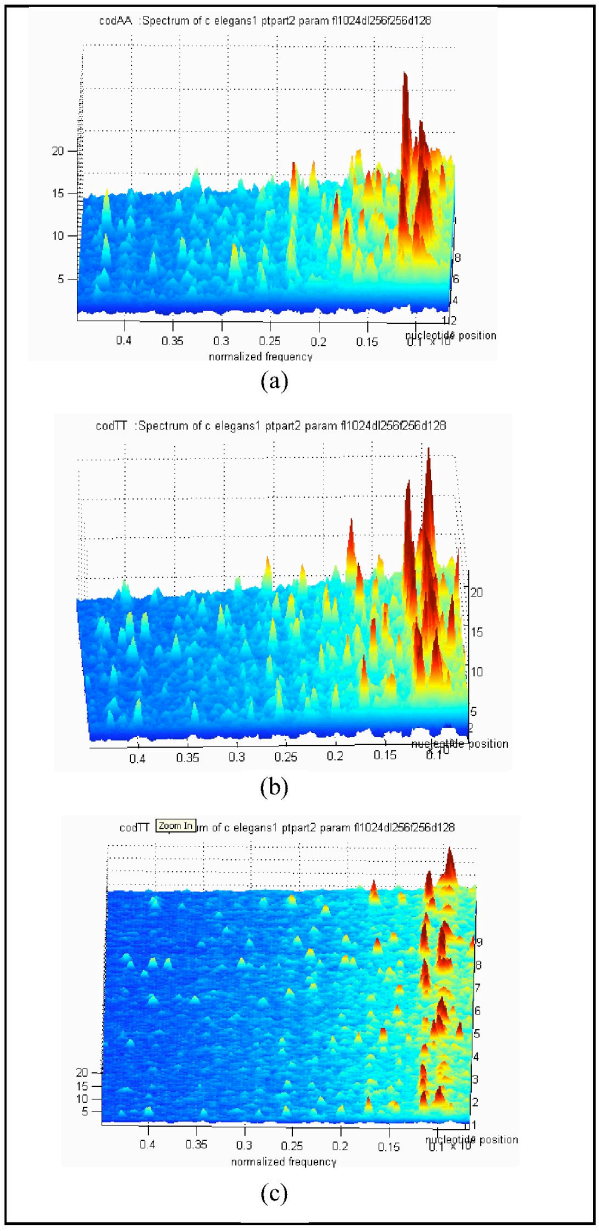

Figure 7. 3D spectrograms for the Chromosome 1 C. elegans with binary indicator dinucleotide 'AA' (a),'TT'(b,c)

These results are the same for all the chromosome tested for these specific indicator. The Fig 8 shows some results similar to the previous one but according to the chromosomes 3 and 4.

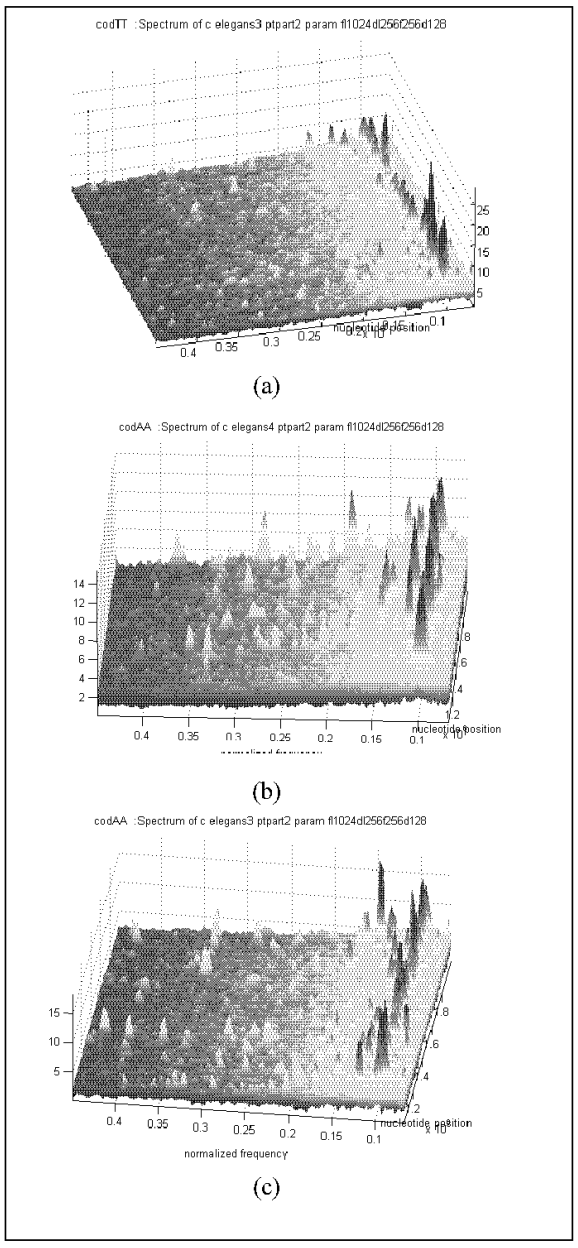

Figure 8. 3D spectrograms for the Chromosome 3 (a,c) and 4(b) C. elegans with binary indicator dinucleotides 'AA','TT'. The peaks of 1/10.5 are enhanced for all the chromosomes.

We extend our studies for this periodicity for the binary indicator trinucleotides 'AAA' and 'TTT' (Fig.9). The results proove that the periodicity 10.5 still enhanced for these indicators.

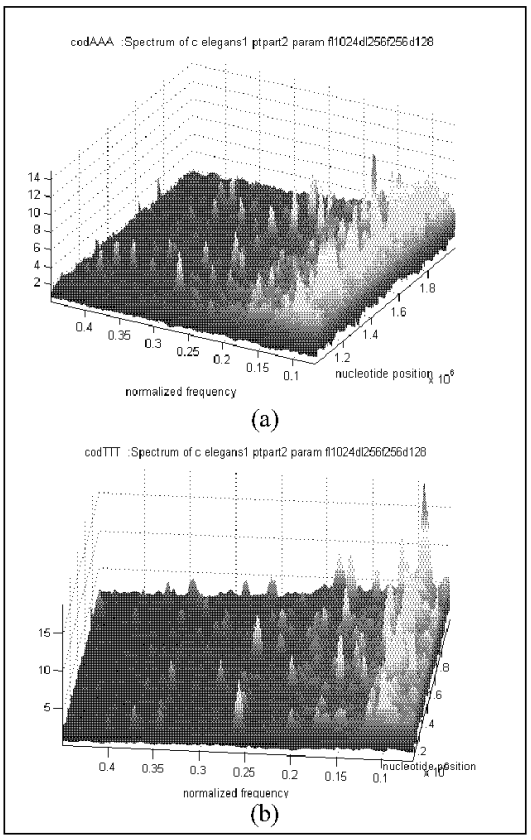

Figure 9. 3D spectrograms for the Chromosome 1 with binary indicator trinucleotides 'AAA' (a), 'TTT'(b). The peaks of 1/10.5 are also enhanced for all the chromosomes.

When we study the spectrogram's behavior for the binary indicator 'GC' and 'TA' (Fig.10) we find that they enhance the frequency 1/3. the peaks are also locally spread.

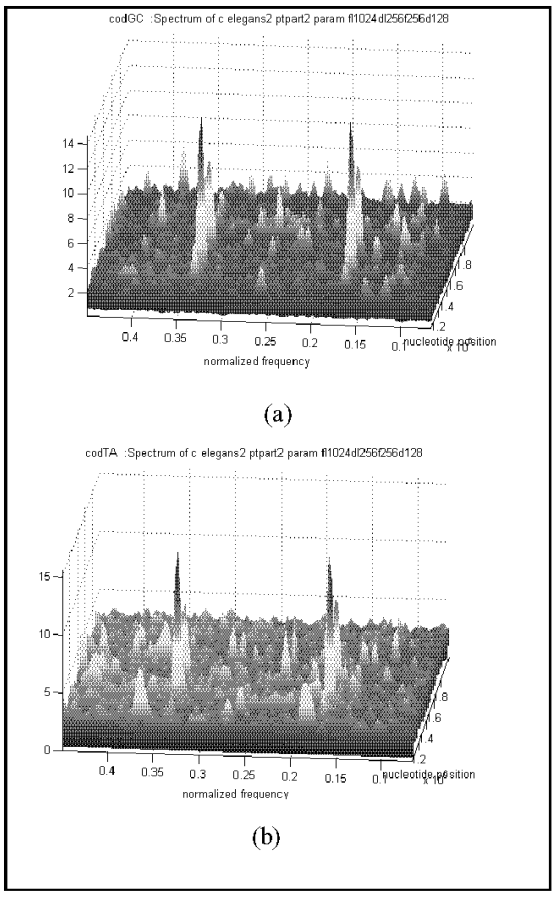

Figure 10. 3D spectrograms for the Chromosome 2 with binary indicator dinucleotides 'GC' (a),'TA'(b). The peaks of 1/3 are enhanced

Concerning the codon groups binary indicators, many periodicities are observed. For some codons the periodicity 3 is enhanced. The maximum values for the peaks related to this periodicity are somtimes less important than the amplitude of other periodicities.

In the Fig. 11, the peaks related to periodicity 3 are visible for many binary indicator codon. For the codons 'CCT' and 'GCT' the peaks are very pronounced. When we focus on the 'GCC', we find that there is a particular behavior involving various periodicities. However we can detect the periodicities 3 and 10.5. The same observation concern the 'CCA'. For the 'CTC', all the peaks have very reduced amplitude but we can see some peaks related to the periodicty 3.

In our study, we have investigated with all the codon indicators and we find various results. In fact, we observe the periodicities 3 and 10.5 object of our research but we find that some codon enhance other periodicities.

For some codon the periodicities are not present for specific frequencies but are present in specific part of the chromosome (Fig.12). This figure present particular frequency behavior according to specific codon.

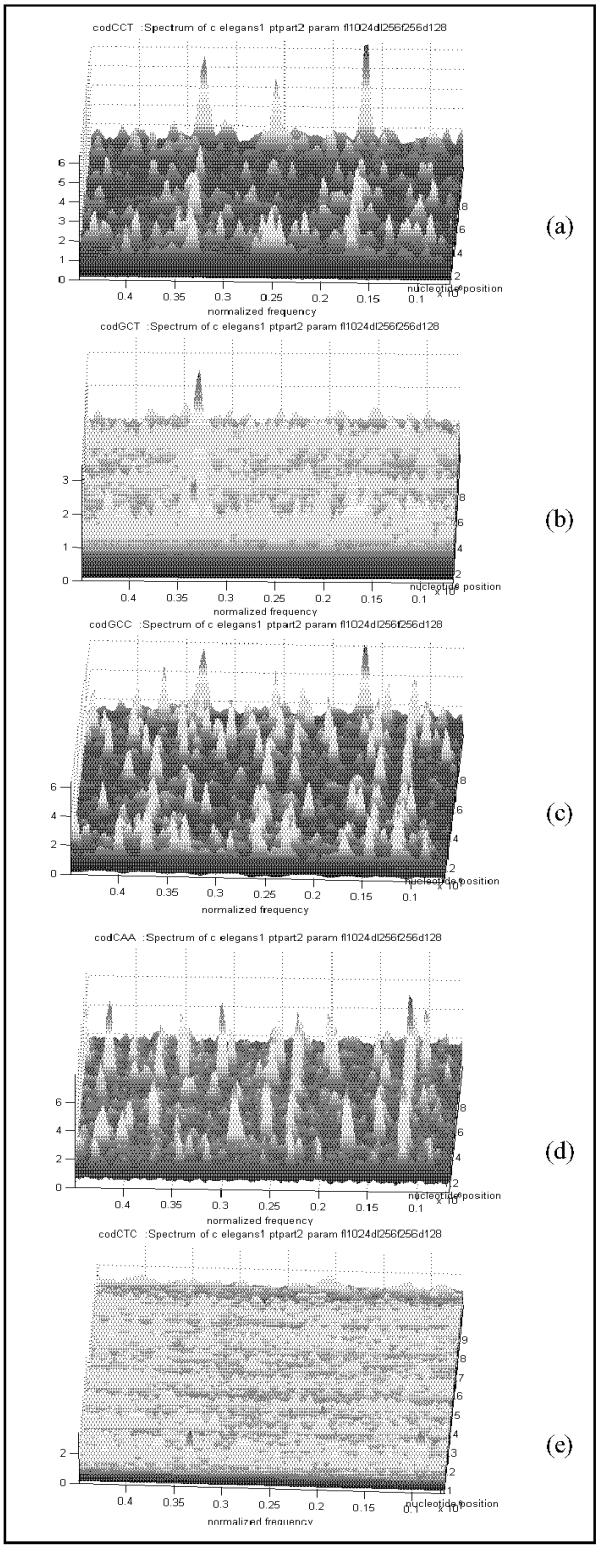

Figure 11. 3D spectrograms for the Chromosome 1 with binary indicator trinucleotides 'CCT' (a),'GCT'(b), 'GCC'(c),'CAA'(d), 'CTC'(e).The amplitude of The peaks related to the frequency 1/3 are varying according to codons

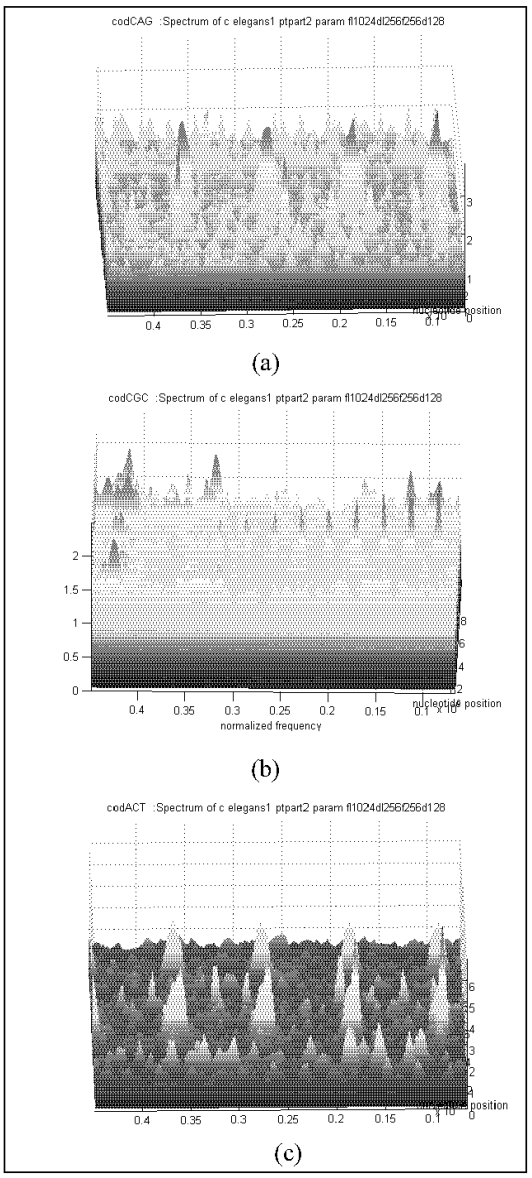

Figure 12. 3D spectrograms for the Chromosome 1 with binary indicator trinucleotides 'CAG' (a),'CGC'(b), 'ACT'(c),.These codons shows a very particular spectrograms frequency behavior

VI. CONCLUSIONS:

In This paper, we investigate the contribution of each base group indicator in the enhancement of the peaks related to periodicities 3 and 10.5 for the C. elegans genome. 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. First we evaluate the codon's contribution in the frequencies enhancement through the chromosomes. Second, we consider the 3D DNA spectrograms to visually detect the specific parts of chromosomes related with protein coding regions and nucleosomes positioning regions. In this paper, we propose a 3D representation of the specific regions. The future direction of this work is to exploit this method to analyze various types of genomes in order to study the percent of the base group indicators able to enhance specific periodicities.

VII. REFERENCES

[1] W. Altenburger, W. Horz, H. Zachau, "Nuclease cleavage of chromatin at 100-nucleotide pair intervals" Nature 264 pp 517-522 1976

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

[3] D. Anastassiou, "DSP In genomics: processing and frequency domain analysis of character strings" *in Proc. IEEE International conference acoustics speech and signal processing ICASSP 2001*, vol 2, pp: 1053-1056.

[4] J. A. Berger, S. K. Mitra, M. Carli and A. Neri, "New approaches to genome sequence analysis based on digital signal processing", *In Proc. GENSIPS2002*, pp: 1-4, Raleigh, North Carolina, USA, 11-13 October 2002,

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

[6] J. A. Berger, S. K. Mitra, M. Carli and A. Neri, "Visualization and analysis of DNA sequences using DNA walks", *Journal of the Franklin Institute*, 341, pp: 37-53,2004.

[7] C. B. Burge and S. Karlin, "Finding the genes in genomic DNA". *Current Opinion in Structural Biology*, 8, pp: 346-354, 1998.

[8] A.B. Cohanim, Y. Kashi and E.N. Trifinov, "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, 2005

[9] R. P. Costa, "Gene prediction algorithms", *Computational Biology*, pp: 1-7, May 2003

[10] P. Cristea "Large scale features in DNA genomic signals", *Signal Processing (Special issue on Genomic signal Processing)*, 83 (4), pp: 871-888, 2003.

[11] J.J. Hayes, T.D. Tullius, A. P. wolffe, " 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 1990

[12] R.D. Kornberg, "Chromatin structure: a repeating unit of histones and DNA." Science 184, pp:868-871,1974

[13] R.D. Kornberg, "Structure of Chromatin", Annu Rev Biochem. 46 , pp 931-954, 1977

[14] D. Nicorici, J. A. Berger, J. Astola and S. K. Mitra "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, May 2003.

[15] A. V. Oppenheim, R. W. Schafer and J. R. Buck, "*Discrete Time Signal Processing*", 2nd Edition, Prentice Hall, 1999.

[16] P.Oudet, J.E.Germond, M.Bellard, C. Spadafora,P. Chambon, " 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, 1978

[17] F. Rechenmann and C. Gautier, "Interpreting the Genome", *La Recherche*, N° 332, pp 39-45, June 2000.

[18] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thamstrom, Y.Field, I. K. Moore, J.P.Z. Wang and J. widom, " a genomic code for nucleosome positioning, nature 04979 vol 442, pp: 772-778, August 2006

[19] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes," *Eurasip Journal of Applied Signal Processing*, vol. 2003, no. 4, Dec. 2003.

[20] E. N. Trifonov, J.L. Sussman, " 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, 1980

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

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

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

[24] A. Worcel, S. Strogatz, D. Riley, " 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