

A Coding Technique Based on the Frequency Evolution Creates with a Time Frequency Analysis a New Genome's Landscape

Imen MESSAOUDI, Afef ELLOUMI

Université de Tunis El Manar, Ecole Nationale
d'Ingénieurs de Tunis, LR Signal, Images et Technologies
de l'Information, BP 37, le Belvédère, 1002,
Tunis, Tunisie.

Zied LACHIRI

Département de Génie Physique et Instrumentation,
INSAT, BP 676, Centre Urbain Cedex, 1080,
Tunis, Tunisie.

Abstract—In recent years, considerable effort has been devoted to study the biological data sets within the framework of the genomic signal processing field. However, the enormous amount of data deposited into public databases makes the search for useful information a difficult task. Effectively, the choice for a convenient analysis approach is not at all obvious at all. In this work, we provide a new way to map the genomes within the form of images. The mapping uses the Complex Morlet wavelet as analysis technique and the Frequency Chaos Game Signal (FCGS) as digital dataset. Before processing the wavelet analysis, we build the FCGS in such a way that we can follow the frequency evolution of nucleotides' occurrence along the genome. The time-frequency analysis of the FCGS signals constitutes a pertinent tool for exploring the DNA structures in the *C.elegans* genome-wide landscape.

Keywords—*C.elegans*; Complex Morlet wavelet; Frequency Chaos Game Signal; Genome exploration

I. INTRODUCTION

With the advances in the field of genomics, the sequencing techniques keep improving; which speeds up the collection of the biological data. Consequently, the amounts as well as the types of the data are continually increasing. Hence the need for new tool that permits an easy navigation within the genomes. Nowadays, researchers in genomics rely on a standard graphical representation of chromosomes called ideogram. Ideograms allow genomic data visualization using points, lines and other shapes to indicate the location of particular sites along the chromosomes [1]. However, the ideograms' annotations must be updated once one discovers new DNA hotspots within the chromosomal sequences. Thus, it is better to find other tool that permanently describes all the chromosome structures independently from their complexity. The idea consists in finding an adequate representation tool that directly maps the DNA produced by sequencing (i.e. in its character form) [2] [3] without a need for biological experiments or alignment algorithms [4] [5] [6] or automatic pipelines [7] to annotate all the inherent components.

II. RELATED WORK

Within the context of the genomic signal processing, the joint time-frequency analysis based on the Fourier transform has played a key role in the data characterization and visualization [8-11]. In fact, the color spectrograms were

shown to provide significant information about periodicities and recurrent motifs along the bio-molecular sequences [8] [9]. Furthermore, the tri-color spectrograms, which are obtained by the reduction of dimensionality, give a unique visual signature of specific regions of the DNA [12]. Nevertheless, the problem with STFT goes back to its limited time-frequency resolution capability. Indeed, according to the Heisenberg uncertainty principle, one cannot obtain a good resolution in both the time and the frequency domains due to the fixed STFT window's length. Thus, the wavelet transform appears to be a good solution to overcome the STFT resolution limitation [13-17]. Given the localization wavelet properties, recent works were oriented towards investigating the DNA correlations [18-20] as well as the identification of the coding regions in genes [21] and some of the repeating protein motifs [22]. In this framework, we propose the continuous wavelet analysis to map the genomic DNA as scalogram images based on the complex Morlet wavelet. The main objective of this work consists in unraveling the localized spectral behaviors of different DNA structures. Since the DNA sequences are stored in the biological databases within the form of strings, it is necessary to convert them into numerical data; which will enable in turn the signal processing based applications. This operation defines the so-called "DNA coding". Here, we are concerned with a new numerical assignment scheme, which is the Frequency Chaos Game Signal (FCGS). The latter, allows following the evolution of the oligomers frequency of occurrence. Furthermore, by combining the FCGS with the wavelet analysis we create a new perspective to represent the information within any given genomic sequence. The method is not only effective and original; it also enhances specific information about the sequence at each FCGS level. To prove the efficacy of the work in terms of the genomes landscapes visualizing, we consider the *C.elegans* genome.

III. MAPPING THE DNA SEQUENCES: FROM CHARACTER STRINGS TOWARDS THE FREQUENCY CHAOS GAME SIGNAL

The Frequency Chaos Game Signal is a new DNA coding technique consisting in a linear form of the Frequency Chaos Game Representation (FCGR). The latter method illustrates a DNA sequence in the form of a 2D image [23-25], based on the Chaos Game theory [26] [27]. The method (we mean the FCGR) encodes the oligomers frequency of apparition according to given color scale; where each frequency occupies

a precise emplacement in the representation area. Consider the following sequence:

Seq_{comp}=ACGATACAGATCAGATTTAGACAGACCGATA
GTAGACGATCAGATCACCAGTGAC.

The monomers, dimmers and timers frequencies of apparition as well as the related words organization are given by TABLE I and TABLE II.

In our coding approach, we take the frequency matrices as assignment base [28] [29]. For this purpose, we fix the representation order k and we generate the FCGR _{k} for the totality of the entry sequence. For this coding we generally take

the chromosomic sequence as entry data set, because we want that the FCGS reflects the statistical properties of the genome itself. Then, we read the bases' succession by a group of k -nucleotides and we assign to each position the correspondent frequency value. For example **seq**=ACGATACA, which is a portion of the **seq_{comp}**. So, based on the FCGRs matrices previously extracted, we attribute to the monomers, dimmers and timers the frequency values as illustrated in TABLE III.

If we want to encode the totality of the sequence **Seq_{comp}**, we do the same thing. In this case, the resulting FCGS₁, FCGS₂ and FCGS₃ plots are given in Figure 1.

TABLE I. FCGR MATRICES ORDER 1 AND 2 WITH THE ASSOCIATED MATRICES OF WORDS: EACH WORD HAS THE FREQUENCY OF APPARITION WHICH APPEARS AT THE SAME EMPLACEMENT

FCGR2				FCGR ₁	
Frequency matrix				Frequency matrix	
0.0556	0.0741	0.0185	0.0185	0.2182	0.2182
0.1296	0.0185	0.2037	0.0556	0.3818	0.1818
0.1481	0.1667	0.0741	0.0370		
0.0185	0.1296	0.0926	0.0556		
Dimers organization				Monomers organization	
CC	CG	GC	GG	C	G
CA	CT	GA	GT	A	T
AC	AG	TC	TG		
AA	AT	TA	TT		

TABLE II. FCGR₃ MATRIX WITH THE ASSOCIATED MATRIX OF WORDS: EACH WORD HAS THE FREQUENCY OF APPARITION WHICH APPEARS AT THE SAME EMPLACEMENT

FCGR ₃							
Frequency matrix							
0.0189	0.0377	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189
0.0377	0.0189	0.0755	0.0189	0.0189	0.0189	0.0189	0.0189
0.0377	0.1132	0.0189	0.0189	0.0943	0.0189	0.0189	0.0377
0.0189	0.0189	0.0189	0.0189	0.0189	0.1321	0.0377	0.0189
0.0566	0.0566	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189
0.0566	0.0189	0.1321	0.0566	0.0755	0.0189	0.0377	0.0189
0.0189	0.0189	0.0755	0.0189	0.0377	0.0755	0.0189	0.0189
0.0189	0.0189	0.0566	0.0377	0.0189	0.0189	0.0377	0.0377
Trimers organization							
CCC	CCG	CGC	CGG	GCC	GCG	GGC	GGG
CCA	CCT	CGA	CGT	GCA	GCT	GGA	GGT
CAC	CAG	CTC	CTG	GAC	GAG	GTC	GTG
CAA	CAT	CTA	CTT	GAA	GAT	GTA	GTT
ACC	ACG	AGC	AGG	TCC	TCG	TGC	TGG
ACA	ACT	AGA	AGT	TCA	TCT	TGA	TGT
AAC	AAG	ATC	ATG	TAC	TAG	TTC	TTG
AAA	AAT	ATA	ATT	TAA	TAT	TTA	TTT

TABLE III. LINEARIZATION OF THE OLIGOMERS FREQUENCIES TO OBTAIN FCGSK WITH K={1, 2 AND 3}

Monomers (seq)=	A, C, G, A, T, A, C, A.
FCGS ₁	0.3818, 0.2182, 0.2182, 0.3818, 0.1818, 0.3818, 0.2182, 0.3818.
Dimers (seq)=	AC, CG, GA, AT, TA, AC, CA.
FCGS ₂	0.1481, 0.0741, 0.2037, 0.1296, 0.0926, 0.1481, 0.1296.
Trimers (seq)=	ACG, CGA, GAT, ATA, TAC, ACA.
FCGS ₃ =	0.0566, 0.0755, 0.1321, 0.0566, 0.0377, 0.0566.

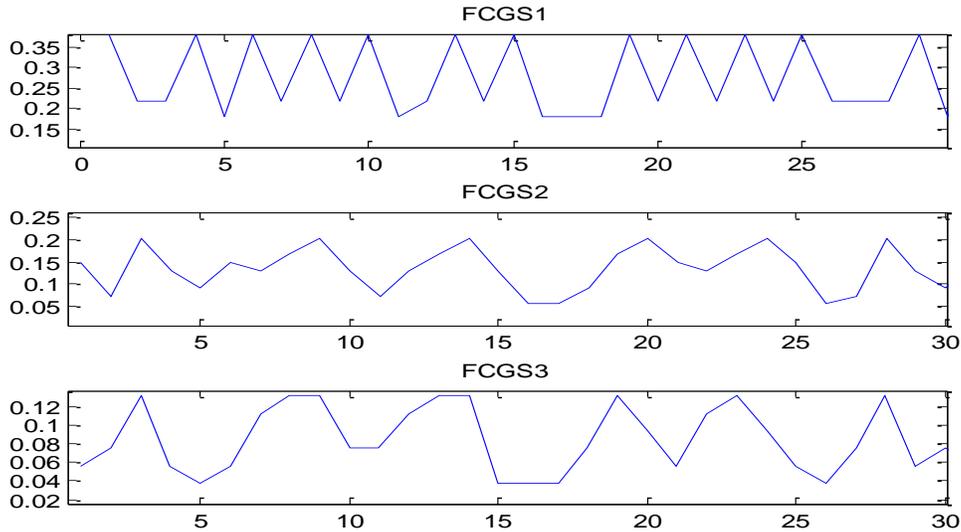


Fig. 1. Representation of the sequence “ACGATACAGATCAGATTTAGACAGACCGATAGTAGACGATCAGATCACCAGTGAC” by FCGS₁, FCGS₂ and FCGS₃.

IV. COMPLEX MORLET WAVELET ANALYSIS

Unlike the Fourier transform, which is based on the average of signals contents within a fixed window, the wavelet transform offers very good time-frequency localization as it satisfies the uncertainty principle [30]. Indeed, the wavelet transform adapts its window (called mother wavelet and denoted ψ) in such a way it shortens at high frequencies and expands at low frequencies depending on a scale parameter a . At each scale, the daughter wavelet shifts in time using a shift parameter b to permit the convolution of the input signal with the analysis window. This principle is the basis of the Continuous Wavelet Transform (CWT). Thus, the daughter wavelets are generated by equation (1); and the continuous wavelet transform is defined by equation (2).

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right), \quad a > 0, b \in \square \quad (1)$$

$$T_\psi(X)(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} X(t) \psi^*\left(\frac{t-b}{a}\right) dt \quad (2)$$

Here * represents the operation of complex conjugate and $X(t)$ is the input signal function. The most usual way to display wavelet transforms is to look at the absolute wavelet coefficients: $|T_\psi(a,b)|$, giving a time-scale representation. This representation is named scalogram. The scale set is proportional to the frequency one, it follows that one can project the wavelet coefficients in the time-frequency plan [31-

33]. In general, it is preferable to choose a continuously differentiable mother wavelet with compact support [34]. In our analysis, we choose the Complex Morlet wavelet which is a Gaussian-windowed complex sinusoid :

$$\psi(t) = \pi^{-\frac{1}{4}} \left(e^{i\omega_0 t} - e^{-\frac{1}{2}\omega_0^2} \right) e^{-\frac{1}{2}t^2} \quad (3)$$

In practice, one often takes $\omega_0 \gg 5$ to satisfy the admissibility condition which is required by CWT [34].

V. GENOME-WIDE VISUALIZATION OF C.ELEGANS

This paper concerns the DNA representation into time-frequency images based on the complex Morlet wavelet analysis. For this purpose, we turn our attention to the *C.elegans* genome, whose sequences and annotations are extracted from the NCBI database [35]. Concerning the coding, we consider the FCGS technique with order {1, 2 and 3}. As for the wavelet analysis, we take a mother wavelet with $\omega_0 > 5.4285$. We fix, then, the number of scales to 64. Afterward, we perform the CWT on the FCGS signals for the six chromosomes. Since the scalograms yield almost the same global behavior, we selected results relating to the chromosome1 for illustration. Observing the important length of the chromosome, we proceed by zooming into the scalogram. A myriad of structures are shown to possess specific time frequency behavior which exhibits a number of

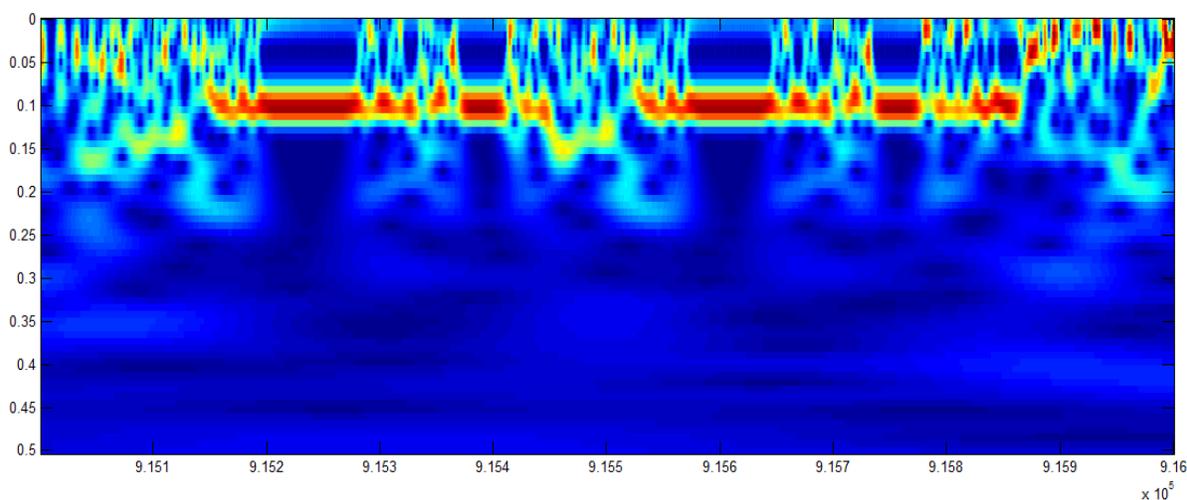
repetitive motifs. Here we are interested with (GCCTAA)_n, CEMUDR1, IR3_CE, MSAT1_CE, HelitronY4_CE, HelitronY1A_CE and Helitron2_CE (see description in TABLE IV).

- The first sequence consists of two (TTAGGC)_n motifs which are spread over 10³ bp. The Figure 2 illustrates the correspondent representation while coding with FCGS₁, FCGS₂ and FCGS₃.

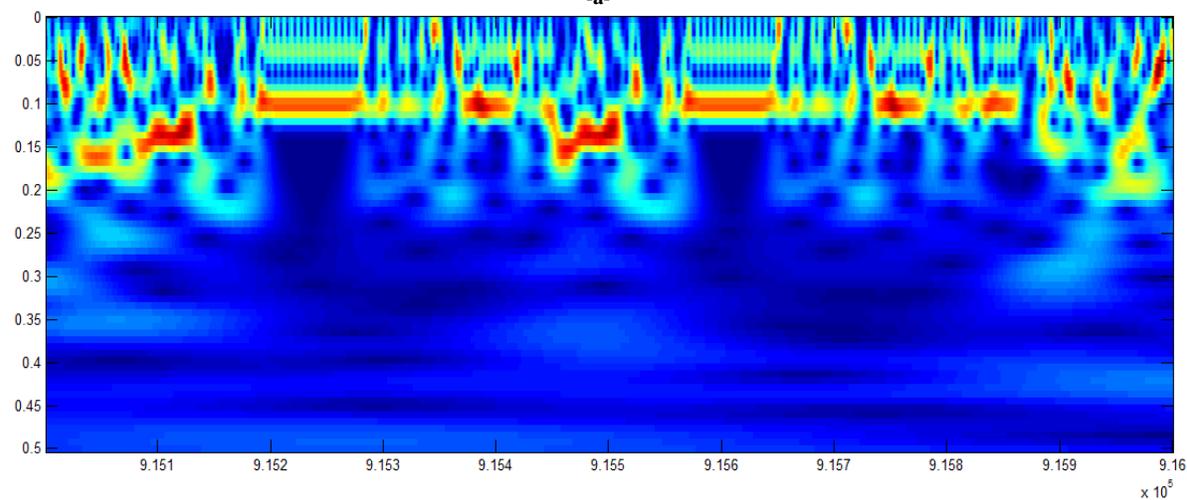
From subfigures (a) and (b), we mainly note the characterization of the structure by concentrated energy around the frequency 0.1 (which is equivalent to the 10bp-periodicity). The latter frequency is generally related to the nucleosome formation [10] [11] [36-39]. However, in the subfigure (c), energy level decreases. By comparing all the subfigures, we conclude that energy indicating the 10 periodicity decreases instead of the enhancement of other shapes within other frequency bands when we raise the FCGS order.

TABLE IV. BOUNDARIES AND LENGTHS OF THE SEQUENCES HAVING SPECIFIC TIME-FREQUENCY SIGNATURES

	Start	End	Illustrated Motifs	Sequence Length (bp)
Sequence n°1	915001	916000	(GCCTAA) _n	1000
Sequence n°2	13935001	13937000	CEMUDR1	2000
Sequence n°3	1922001	1926000	IR3_CE	4000
Sequence n°4	9816001	9824000	MSAT1_CE HelitronY4_CE HelitronY1A_CE Helitron2_CE	8000



-a-



-b-

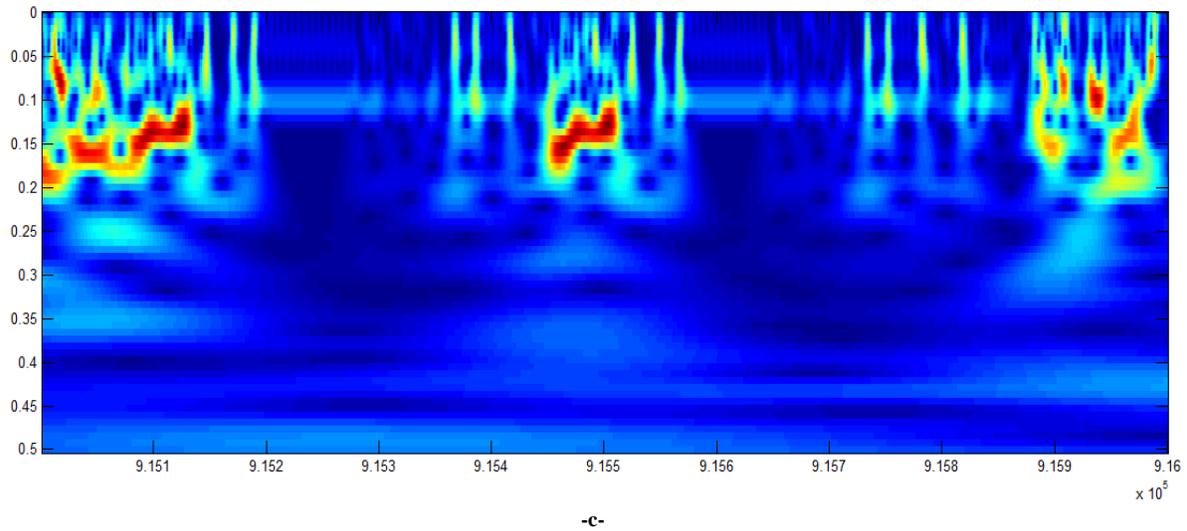
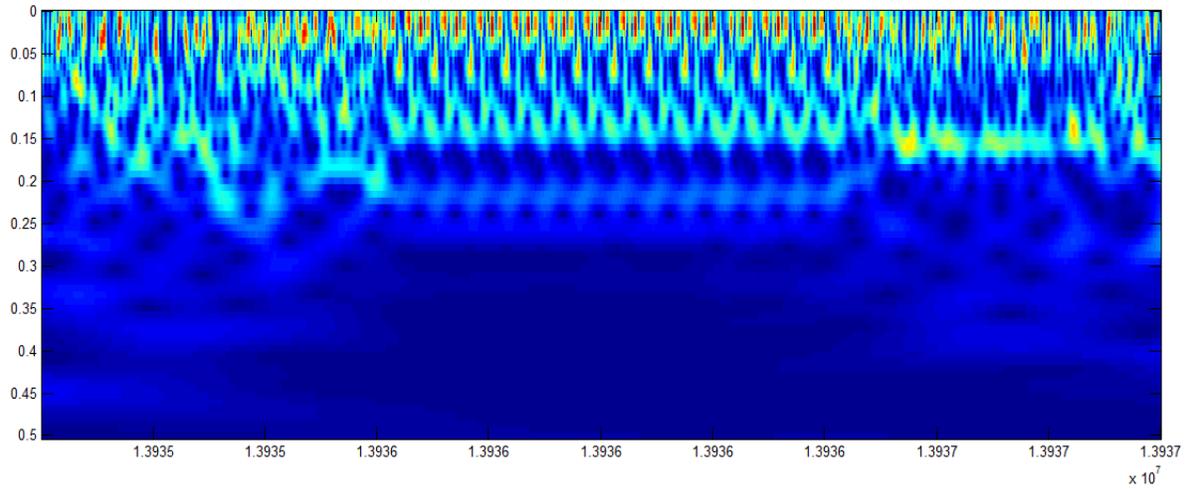


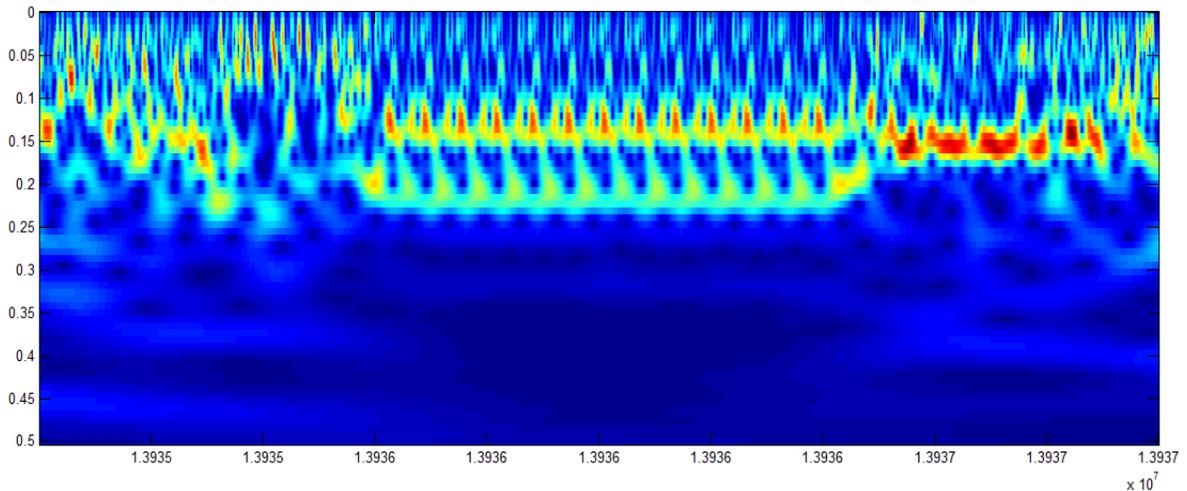
Fig. 2. Scalogram representation of two successive (TTAGGC)_n sequences when coded with FCGS₁ (a), FCGS₂ (b) and FCGS₃ (c)

- The Figure 3 reveals the time-frequency behavior of a CEMUDR1 example. At first glance, we see periodic motifs around the frequency 0.15 (corresponding to the 6.5 bp periodicity which is generally found in non-coding DNA of genes [21]); which offers an easy way

to detect the presence of the structure. A careful inspection of subfigures (from a to c) reveals that each level of the FCGS coding highlights specific repetitive shapes within a given frequency band while keeping the overall aspect.



-a-



-b-

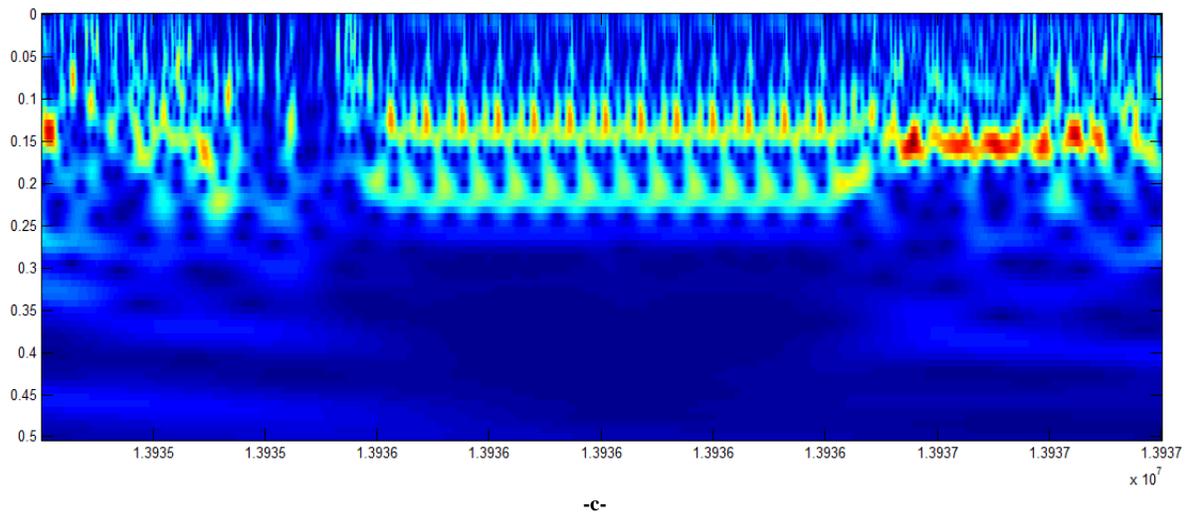
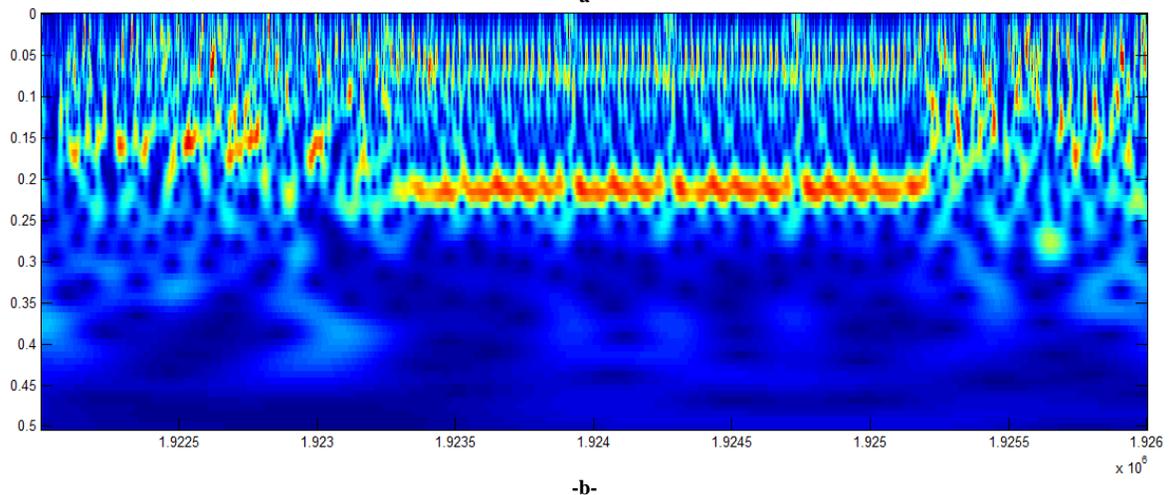
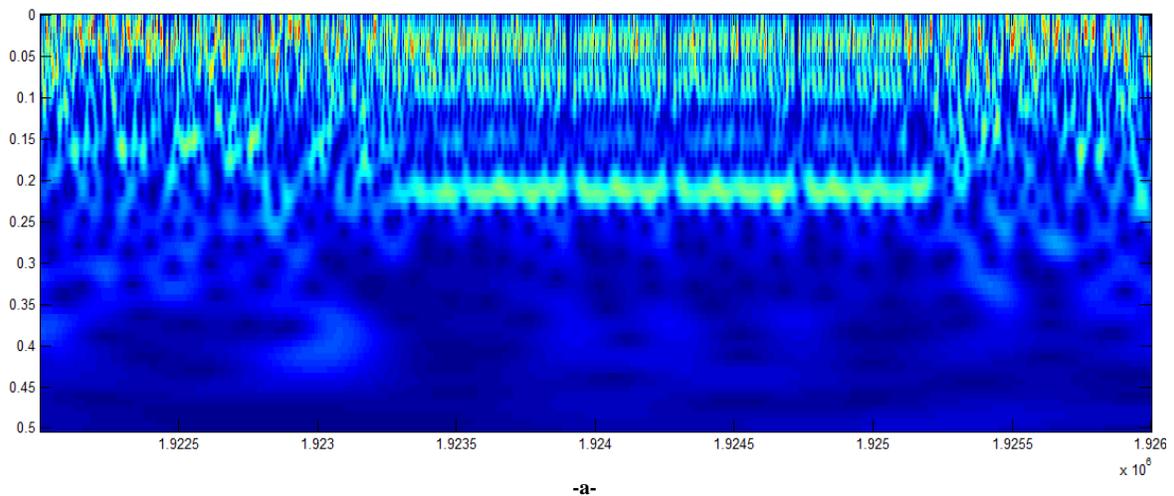


Fig. 3. Scalogram representation of the CEMUDR1 structure when coded with FCGS₁ (a), FCGS₂ (b) and FCGS₃ (c)

- Concerning the IR3_CE structure (Figure 4), the FCGS scalograms show characteristic signature having as main attribute the high energy around the frequency 0.2 (equivalent to the 5bp-periodicity). Moreover, changing the

FCGS scale is useful for visualizing the different frequency bands that characterize the sequence for the considered order.



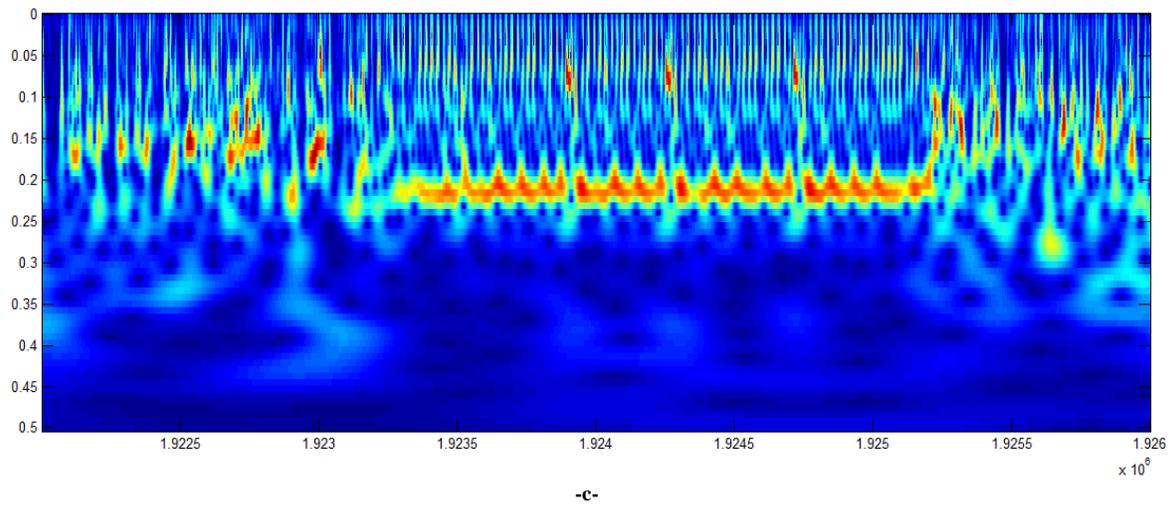
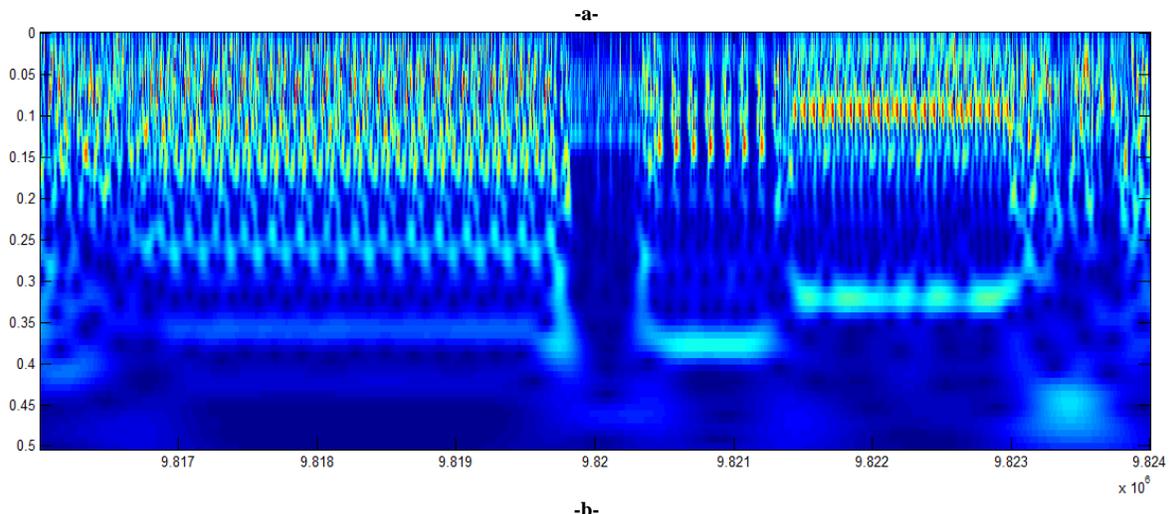
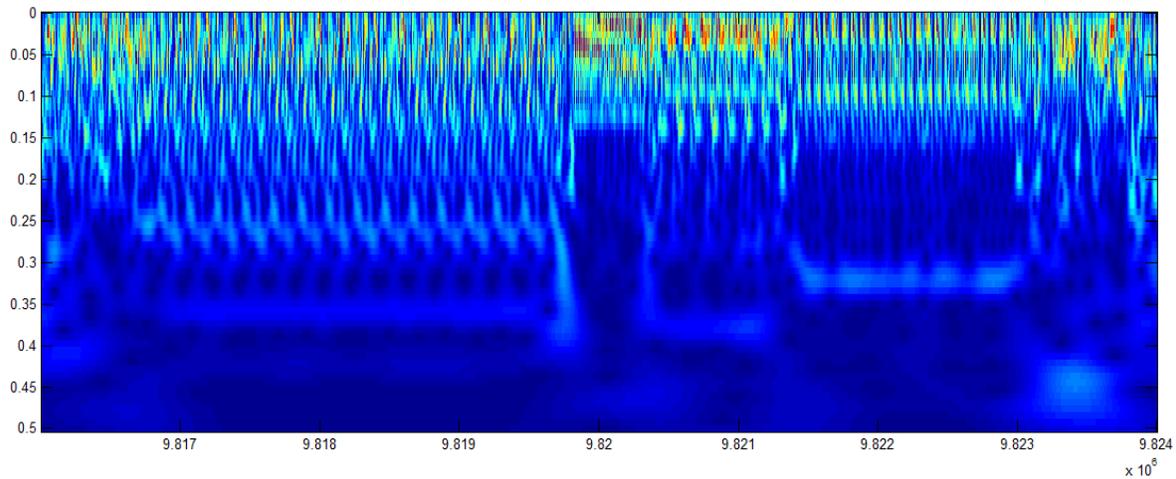


Fig. 4. Scalogram representation of the IR3_CE structure when coded with FCGS₁ (a), FCGS₂ (b) and FCGS₃ (c)

- The final example consists on a succession of a minisatellite (MSAT1_CE) and three helitrons from different classes (HelitronY4_CE, HelitronY1A_CE and Helitron2_CE). Each of these structures is shown to possess unique signature where the global behavior does not change too

much when the FCGS order changes (Figure 5). However at each scale, the difference lays in the energy band repartition. Further, thanks to our representation we can easily detect the boundaries of each sequence.



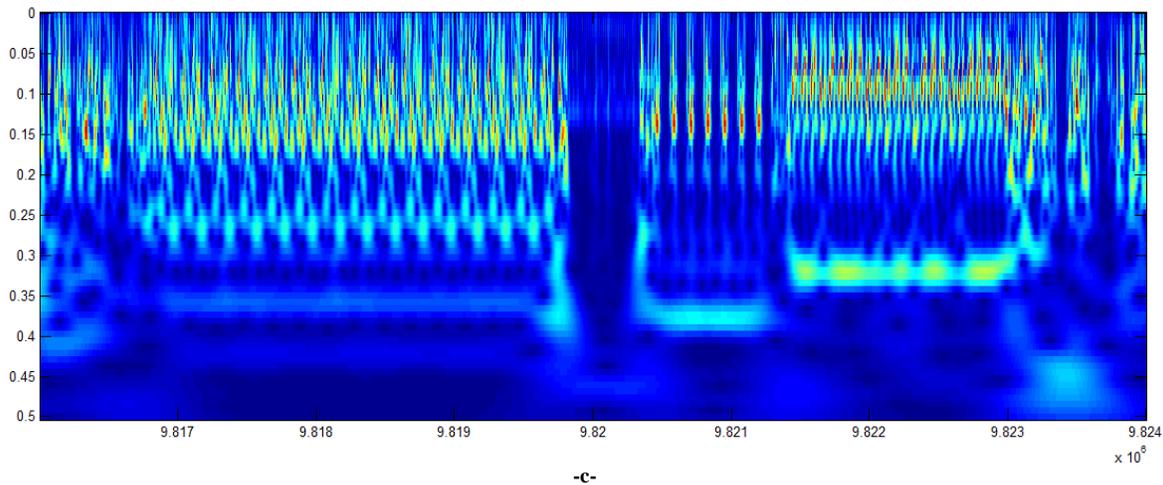


Fig. 5. Scalogram representation of the succession of MSAT1_CE, HelitronY4_CE, HelitronY1A_CE and Helitron2_CE when coded with FCGS₁ (a), FCGS₂ (b) and FCGS₃ (c)

Apart from these examples, we succeeded in identifying other structures due to their unique time-frequency features; such as: Introns, CeRep, Satellites, Minisatellites, Microsatellites, Cele14b, Tir9ta, Longpal, Npalta, etc.

VI. CONCLUSION

The work presented here, described a new way to represent the genomic DNA in the form of images. Our goal was to offer visual navigation of the genome where different structures can be easily distinguished through a specific signature in the representation's plan. This has the advantage of avoiding the inaccurate and unavailable annotations as well as the long and expensive experiments. This is why; we based our study on the Frequency Chaos Game Signal (FCGS) and the Complex Morlet Wavelet analysis. The FCGS approach is a new coding tool to convert the DNA strings into signals that reflect the statistical properties of genomes. These signals are converted in turn into scalogram images that reflect the periodic features of DNA using the complex Morlet wavelet. The analysis reveals the characterization of different structures by particular periodic motifs localized within particular frequency bands with different energy levels. Furthermore, the variation of the FCGS order offers variability in terms of information while keeping the global aspect of the sequence's behavior.

VII. FUTURE WORK

Overall, our analysis is a promising and fruitful direction in the sense that it forms a strong base for classifying the DNA structures as well as recognizing unknown sequences and rectifying the available annotations. Then, future work will focus on extracting pertinent parameters for DNA structures' classification based on the generated scalograms.

REFERENCES

- [1] C. O'Connor, "Chromosome mapping: Idiograms", *Nature Education*, vol. 1, no. 1: 107, 2008.
- [2] H.T chang , C.J Kuo, N.W. Lo and W.Z Lv, "DNA sequence Representation and comparison Based on Quaternion Number System", *International Journal of Advanced Computer Science & Applications (IJACSA)*, vol. 3, no 11, 2012.
- [3] K.S. Sathish and N. Duraipandian, "An effectice identification of Species from DNA Sequence: A Classification Technique by integrating

- DM and ANN", *International Journal of Advanced Computer Science & Applications (IJACSA)*, vol. 3, no 8, 2012.
- [4] P. Stothard and D. Wishart, "Automated bacterial genome analysis and annotation", *Elsevier, Current Opinion in Microbiology*, vol. 9, pp. 505-510, 2006.
- [5] X. Huang, M. D. Adams, H. Zhou and A. R. Kerlavage, "A tool for analyzing and annotating genomic sequences", *Genomics*, vol. 46, pp. 37-45, 1997.
- [6] R. Gupta, P. Agarwal and A. K. Soni, "Genetic Algorithm Based Approach for Obtaining Alignment of Multiple Sequences", *International Journal of Advanced Computer Science & Applications (IJACSA)* vol. 3, no 12, 2012.
- [7] E. J. Richardson and M. Watson, "The automatic annotation of bacterial genomes", *Briefings in Bioinformatics*, pp. 1-12, 2012.
- [8] D. Anastassiou, "Genomic signal processing", *IEEE Signal Processing Magazine*, 2001.
- [9] L. Wang and L.D. Stein, "Localizing triplet periodicity in DNA and cDNA sequences", *BMC Bioinformatics*, 2010.
- [10] A. E. Oueslati, Z. Lachiri, and N. Ellouze, "3D Spectrum Analysis of DNA Sequence: Application to Caenorhabditis elegans Genome", *Bioinformatics and Bioengineering BIBE, Proceedings of the 7th IEEE International Conference on*, pp. 864- 871, 2007.
- [11] A.E. Oueslati, I. Messaoudi, Z. Lachiri, and N. Ellouze, "Spectral Analysis of global behaviour of C. Elegans Chromosomes", *Fourier Transform Applications*, Intech, 2011.
- [12] D. Sussillo, "Spectrogram Analysis of Genomes", *EURASIP Journal on Applied Signal Processing* 1, pp. 29-42, 2004.
- [13] K. B. Alsberg, , A. M. Woodward and D. B. Kell, "An introduction to wavelet transforms for chemometricians: A time-frequency approach", *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no.2, pp.215-239, 1997.
- [14] P.J. Mena-Chalco, H. Carrer, Y. Zana, and R.M. Cesar, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5(2), 2008.
- [15] J.A. Tenreiro Machado, A.C. Costa and M. D. Quelhas, "Wavelet analysis of human DNA", *Genomics*, vol. 98, pp. 155-163, 2011.
- [16] B. Bhosale, B. S. Ahmed and A. Biswas, "Wavelet Based Analysis in Bio-informatics", *Life Science Journal*, vol. 10, no. 2, 2013.
- [17] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, Z. Lacroix, and Ch. Legendre, "Waveform Mapping and Time-Frequency Processing of DNA and Protein Sequences", *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 59, no. 9, pp. 4210-4224, 2011.
- [18] A. Arneodo, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy and C. Thermes, "What can we learn with wavelets about DNA sequences?", *Physica A* 249, pp. 439-448, 1998.
- [19] G. Dodin, P. Vanderghyest, P. Levoir, C. Cordier and L. Marcourt, "Fourier and Wavelet Transform Analysis, a Tool for Visualizing

- Regular Patterns in DNA Sequences”, *J. Theor. Biol.* 206, pp.323-326, 2000.
- [20] A. A. Tsonis, P. Kumar, J. B. . Elsner, P. A. Tsonis, “Wavelet Analysis of DNA sequences”. *Physical Review E* 53, pp. 1828-1834, 1996.
- [21] P. J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar, “Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5(2), 2008.
- [22] B. K. Murray, D. Gorse and J. M. Thornton, “Wavelet Transforms for the Characterization and Detection of Repeating Motifs”, *J. Mol. Biol.* 316, pp. 341-363, 2002.
- [23] J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, M. Fletcher M, “Analysis of genomic sequences by Chaos Game Representation”, *Bioinformatics*, vol. 17(5), pp.429-37, 2001.
- [24] A. Fiser, G. E. Tusnady and I. Simon, “Chaos game representation of protein structures”, *J.Mol Graphics*, vol. 12, pp. 295-304, 1994.
- [25] Y. W. Wang, K. Hill, S. Singh, L. Kari, “The spectrum of genomic signatures: from dinucleotides to chaos game representation”, *Gene*, vol. 346, pp. 73-185, 2005.
- [26] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, “Genomic signature: characterization and classification of species assessed by chaos game representation of sequences”, *Mol.Biol.E*, vol. 16(10), pp.1391-1399, 1999.
- [27] P. Deschavanne, A. Giron, J. Vilain, CH. Dufraigneand and B. Fertil “Genomic Signature Is Preserved in Short DNA Fragment”, *International Symposium on Bio-Informatics and Biomedical Engineering, IEEE*, pp. 161-167, 2000.
- [28] I. Messaoudi, A. E. Oueslati and Z. Lachiri, “Complex Morlet Wavelet Analysis of the DNA Frequency Chaos Game Signal and Revealing Specific Motifs of Introns in *C.elegans*”, *International Conference on Control, Engineering & Information Technology (CEIT'13)*, 2013.
- [29] I. Messaoudi, A. E. Oueslati and Z. Lachiri, “Revealing Helitron signatures in *Cænorhabditis elegans* by the Complex Morlet Analysis based on the Frequency Chaos Game Signals”, 2nd *International Work on Bioinformatics and Biomedical Engineering (IWBIO)*, 2014.
- [30] A. Grossmann and J. Morlet, “Decomposition of Hardy functions into square integrable wavelets of constant shape.” *SIAM: Journal on Mathematical Analysis*”, vol. 15, pp.723-736, 1984.
- [31] M. Steinbuch and M.J.G. van de Molengraft, Eindhoven University of Technology, Control Systems Technology Group Eindhoven, “Wavelet Theory and Applications”, a literature study, R.J.E. Merry, DCT, 2005.
- [32] N. C. F. Tse and L. L. Lai, “Wavelet-Based Algorithm for Signal Analysis”, *EURASIP Journal on Advances in Signal Processing*, 2007.
- [33] A.H. Najmi and J. Sadowsky, “The Continuous Wavelet Transform and Variable Resolution Time–Frequency Analysis”, *Johns Hopkins Apl Technical Digest*, vol. 18, No.1, 1997.
- [34] X. P. Zhang, M. D. Desai and Y. N. Peng, “Orthogonal Complex Filter Banks and Wavelets: Some Properties and Design”, *IEEE Transactions on signal processing*, vol. 47, n°. 4, 1999.
- [35] <http://www.ncbi.nlm.nih.gov/Genbank/>.
- [36] E. Segal, Y. Fondufe-Mittendorf, L. Chen, L. , A. Thamstrom, Y. Field et al., “a genomic code for nucleosome positioning”, *Nature* 442, pp.772-778, 2006.
- [37] J. Widom and P. T. Lowary, “Nucleosome packaging and nucleosome positioning of genomic dna”, *In Proceedings of the National Academy of sciences of the United States of America* 94, pp. 1183-1188, 1997.
- [38] R. D. Kornberg, “Chromatin structure: a repeating unit of histones and DNA”, *Science* 184, pp. 868-871, 1974.
- [39] E. N. Trifonov and 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*, 77, 7, part 2: Biological Sciences, pp. 3816-3820, 1980.