International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.2, Issue No.12, pp : 1247-1251 1 Dec 2013

Detection of the 6.5- Base Periodicity in the *C.elegans* introns based on the Frequency Chaos Game Signal and the Complex Morlet Wavelet Analysis

Imen Messaoudi[1], Afef Elloumi Oueslati[1], Zied Lachiri[2]
[1]Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis,
LR-11-ES17 Signal, Images et Technologies de l'Information, BP 37, le Belvédère, 1002, Tunis, Tunisie.
[2]Département de Génie Physique et Instrumentation, INSAT, BP 676, Centre Urbain Cedex, 1080, Tunis, Tunisie.
Email : imen_messaoudi@ymail.com

Abstract— Here, we investigate introns' characterization in the C.elegans genes. Thus, we apply the complex Morlet wavelet analysis on a new coding technique: the Frequency Chaos Game Signal. Introns are showed to have a 6.5-base periodicity behavior. Direction of the periodic motifs indicates, moreover, the sequence orientation.

Keywords— 6.5-base periodicity, *C.elegans* Introns, Frequency Chaos Game Signal, Complex Morlet wavelet Transform.

I. Introduction

The gene analysis has been the subject of extensive studies in evolutionary and molecular biology. However, these studies remain insufficient and require much work to comprehend the genomes mechanisms secrets.

Actually, the available gene prediction tools are not accurate and introduce a high rate of errors [ix]: approximately only 40% of genes are correctly predicted [vii]. These errors arise from the fact that the non-coding regions (greatly vary in size and are known as introns) are recognised as intergenic DNA.

Typically, the gene prediction processes are focused on detecting the coding areas (exons) based on the 3-base periodicity property. Then, a number of exon prediction methods have been proposed. The most commonly used methods are based on the Discrete Fourier transform (DFT) [v,vii]. These methods lack good frequency resolution because they use an analysis window with fixed length. Other approaches, based on the digital filters, gave similar results to the DFT ones [iv]. With the advent of the Modified Gabor Wavelet Transform (MGWT), one can outperform the classical methods in terms of identification accuracy [iii]. Till now, approaches for identifying the non-coding portions in genes, have not yet emanated. Just as the 3 bp periodicity, it is known that introns involve a 6.5 bp periodicity [vi]. But this property remains to be proved.

In this work, we focus on the 6.5-base periodicity of intronic sequences as part of the genomic signal processing discipline. In this sense, we propose the complex Morlet wavelet analysis to characterize the non-coding regions in some genes of the *Caenorhabditis elegans* genome (*C.elegans*).

Application of the DSP technique on DNA characters requires their digitization. These characters can be bases (Adenine 'A', Cytosine 'C', Guanine 'G' and Thymine 'T') or strings of bases (such as 'AA', 'AC', 'AG', etc). The procedure is known as "the DNA coding". In line with this, we present a new coding technique: the Frequency Chaos Game Signal (FCGS). The technique is based on the Frequency Chaos Game Representation (FCGR) and allows us to follow the frequency evolution of words' occurrence along a given sequence.

The paper is organized as follows. Section 2 introduces the Frequency Chaos Game Signal and gives an overview on the Complex Morlet wavelet analysis. Section 3 deals with the characterization of the intronic sequences into the *C.elegans* genome by the 6.5- base periodicity. Finally, section 4 concludes the content of the paper.

II. Material and Methodology

1. The Frequency Chaos Game Signal

The Frequency Chaos Game Signal approach (FCGS) is based on the Chaos Game Representation theory (CGR), which permit in turn the representation of a DNA sequence into a unit square. For this aim, the nucleotides A, C, G and T are placed at the corners as described in figure 1.

The procedure consists on placing the first point X_0 at the center of the square. Then, whenever we read a DNA letter, we place a representative point in the square. Each letter U_{n+1} is represented by the point X_{n+1}; which is placed halfway between the previous plotted point X_n and the segment joining the vertex corresponding to the DNA character [x]. The coordinates of each plotted point are given by:

$$X_{n+1} = \frac{1}{2}\left(x_n + \ell_{U_{n+1}}\right) \quad (1)$$

Where $\ell_{U_{n+1}}$ can be:

$$\ell_A(0,0), \ell_C(0,1), \ell_G(1,1) \text{ or } \ell_T(1,0) \quad (2)$$

The figure 1 illustrates the procedure to represent the sequence "ACGGT".

To obtain a Frequency Chaos Game Representation, the CGR image must be divided into 4^k sub-squares. Each sub-square is, in turn, associated to a *k*-lengthen sub-pattern according to a specific organization. The frequency of a *k*-lengthen word occurrence is equal to the number of counted dots in the correspondent sub-square, divided by the complete length of the DNA sequence. Each frequency value in the FCGR matrix is then coded according to a color scale [x].

In figure 2, we furnish an example of the Frequency Chaos Game Representations with order 2 as well as the arrangement of all possible words into the FCGR's sub-squares.

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.2, Issue No.12, pp : 1247-1251 1 Dec 2013

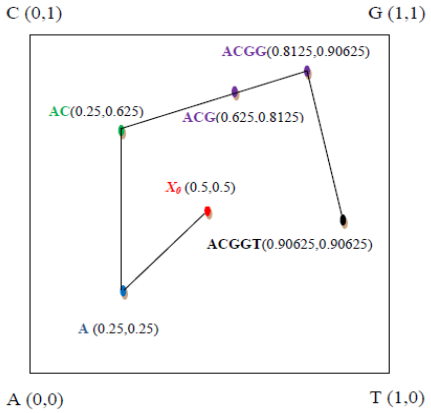


Fig. 1 Representation of the sequence "ACGGT" by the CGR technique: the word "A" is represented by the dot $X_1 = 1/2(X_0 + \ell_{U_1}) = 1/2((0.5,0.5)+(0,0))$.

Likewise, the word "AC" is represented by the point X_2, after that the word "ACG" is represented by X_3, then the word "ACGG" is represented by X_4 and finally the word "ACGGT" is represented by X_5. The set of points {X_1, X_2, X_3, X_4 and X_5} forms the final CGR plot.

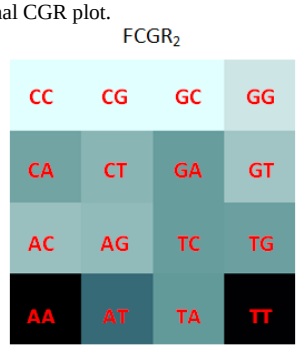


Fig. 2 $FCGR_2$ of the *C.elegans* chromosome V and the distribution of the possible dimers in the Frequency Chaos Game Representation' s space.

The principle of the FCGS technique is that we attribute the frequency value of each sub-pattern to the same group of nucleotides existing in the sequence [x]. To generate an FCGS signal, regarding its constitutive *n*-nucleotides, we have to:

- Calculate the n^{th}-order FCGR for the whole sequence ($FCGR_n$).
- Extract the first *n*-lengthen word in the sequence.
- Search for the frequency value of the considered word within the $FCGR_n$ matrix.
- Attribute, to the first position, the frequency value.
- Move to position 2 and redo the same procedure till reaching the position: Pos= *L*-*n*+1, where *L* is the DNA sequence length and *n* the word's length.

We consider, for example, the sequence:
S={ GAATTCCTAAGCCTAAGCCT}. The motifs contained in the sequence *S* with size two are as follows:
Dimers={ GA,AA,AT,TT,TC, CC, CT, TA, AA, AG, GC, CC, CT, TA, AA, AG, GC, CC, CT}.
While the sequence *S* belongs to the chromosome V of *C.elegans*, we must compute $FCGR_2$ for the whole chromosome.

$$FCGR_2 = \begin{bmatrix} 0.0333 & 0.0305 & 0.0333 & 0.033 \\ 0.0627 & 0.0509 & 0.0618 & 0.0487 \\ 0.0489 & 0.0506 & 0.0619 & 0.0628 \\ 0.1339 & 0.0893 & 0.0642 & 0.1341 \end{bmatrix}$$

Based on these frequencies, we assign the correspondent value to each of the dimers previously extracted. In figure 3, we present the $FCGS_2$ signal:

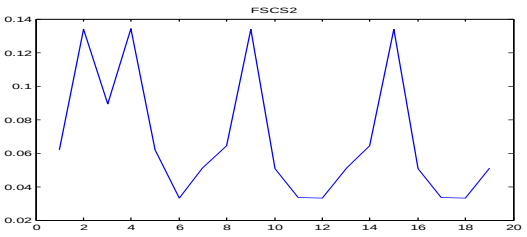


Fig. 3 $FCGS_2$ plot of the sequence S={GAATTCCTAAGCCTAAGCCT}which is taken from the *C.elegans* chromosome V.

2. The complex Morlet wavelet analysis

To localize biological information in both space (nucleotide position) and frequency, we propose the wavelet transform. The procedure consists in decomposing the genomic signal into a sum of basic functions obtained by translations and expansions of the so-called mother wavelet $\psi(t)$ [ii]. Sets of daughter wavelets are given by:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),\ a \succ 0,\ b \in \square \qquad (3)$$

Where *b* and *a* are respectively the time and scale parameters. Assuming that the basic wavelet is positioned around a central frequency f_0 (the maximum value of the mother wavelet's spectrum), the frequency set is proportional to scale one.

The continuous wavelet transform (CWT) of a function *X*(t) is defined as:

$$T_\psi(X)(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} X(t)\psi^*\left(\frac{t-b}{a}\right)dt \qquad (4)$$

Where * is the complex conjugate. The modulus of the coefficients: $|T_\psi(a,b)|$ is named scalogram.

As a mother wavelet, we choose the Complex Morlet function, which is a Gaussian-windowed complex sinusoid. The correspondent mathematical formulation is given by:

$$\psi(t) = \pi^{-\frac{1}{4}}\left(e^{i\omega_0 t} - e^{-\frac{1}{2}\omega_0^2}\right)e^{-\frac{1}{2}t^2} \qquad (5)$$

Where w_0 corresponds to the number of oscillations of the wavelet (w_0=2*pi* f_0). Note that w_0 must be greater than 5 to ensure the invertibility of the mother wavelet [i].

II. Results and Tables

In this study, we considered the *C.elegans* genome (NCBI ID=WBcel235) [viii] which we coded by the $FCGS_2$. Then, we applied the continuous wavelet analysis along 64 scales to each of the six chromosomic signals. For this aim, we took a complex Morlet wavelet with w_0 =5.4285 and having a temporal support of 600 points. Thus, to be able to discern significant information into the scalograms, we zoom using a window of 10^3 bp.

Among these results, we selected an example of gene on each chromosome (see description in Table I).
First of all, we start by studying the behavior of the H17B0 gene. The figure 4 presents the time-frequency representation of the considered sequence.

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.2, Issue No.12, pp : 1247-1251 1 Dec 2013

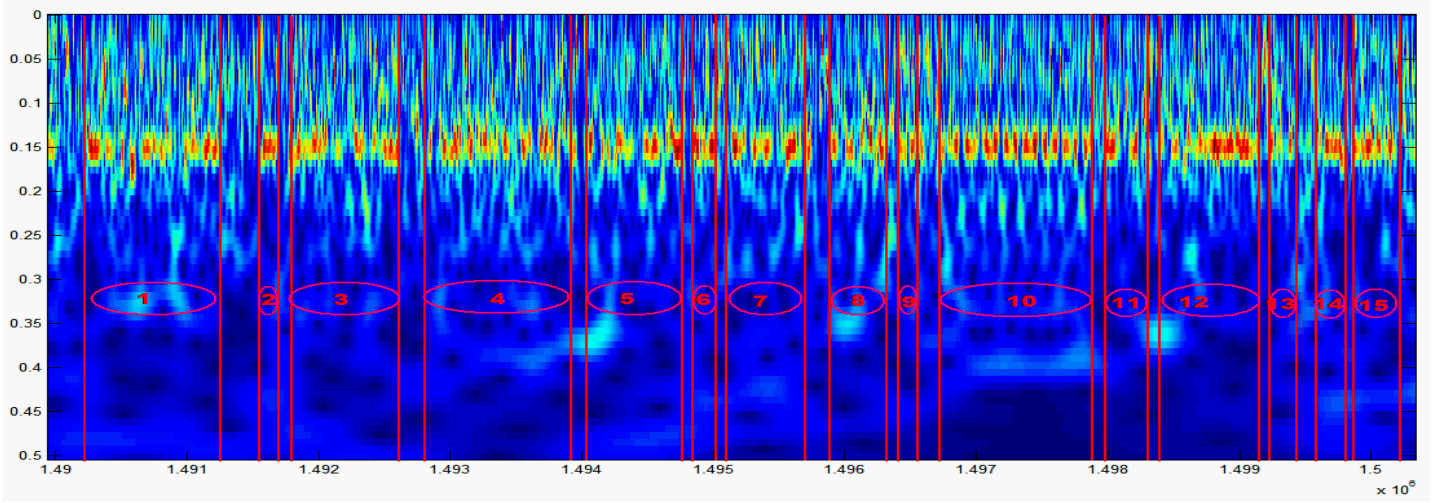


Fig. 4 Scalograms related to the fifteen introns present in the H17B0 gene when we code with $FCGS_2$. Exon-Intron boundaries are marked by red lines and intron zones are numbered.

As it can be noted in the $FCGS_2$ scalogram, it's the introns which are highlighted instead of exons. In fact, intronic zones are characterized by a considerable band with high energy which is concentrated around the frequency 0.15. This frequency value is equivalent to the 6.5 bp periodicity; whereas exons appear as blue regions with low energy. These findings correlate with the NCBI annotations as described in Table II.

Table I: Description of the six *C.elegans* genes.

Gene	Chromosome	Position	Length (bp)	Introns	Exons
Y34D9A.1	I	1019886-1029912	10027	4	7
H17B0	II	1489934-1500336	10403	15	18
Y50D7A.4	III	235841-253375	17535	11	13
F56B3.8	IV	773237-776628	3392	5	7
F33E11.6	V	305439-311536	6098	5	7
T08D2.6	X	180080- 181874	1795	1	2

Table II: Intron boundaries and length in the H17B0 gene.

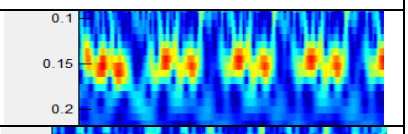
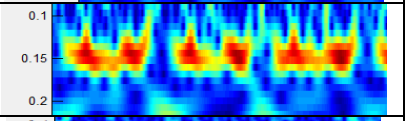
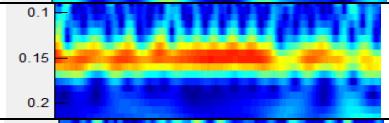
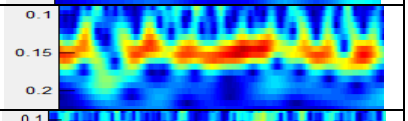
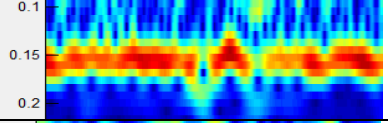
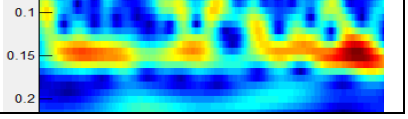
Intron	Start	End	Length (bp)
1	1490243	1491246	1004
2	1491561	1491698	138
3	1491802	1492616	815
4	1492919	1493847	929
5	1494047	1494763	717
6	1494834	1495002	169
7	1495089	1495646	558
8	1495900	1496313	414
9	1496480	1496600	121
10	1496694	1497880	1187
11	1498375	1499053	679
12	1499207	1499257	51
13	1499475	1499554	80
14	1499624	1499782	159
15	1499857	1500195	339

If we carefully analyze the behavior of each intron apart, we will note that they share the same characteristics. In fact, all the introns possess a high energy band around the frequency 1/6.5, independently of their lengths.

The specificity of this result is in proving the property of the 6.5 bp periodicity; which was stated in a few works (such as [vi]) and is not yet proved. Effectively, the majority of *C.elegans* introns are characterized by the same 6.5 periodicity. The figure 5 presents similar behavior in other *C.elegans* genes (see annotations in Table I). The six genes are characterized by specific textures that contain periodic motifs of 6.5 bp long.

When we take a closer look at these motifs, we can observe that they change direction depending on the sequence orientation (see Table III).

Table III: Orientation of the intronic motifs in the six genes.

Gene	Orientation	Motif
Y34D9A.1	-	
H17B0	-	
Y50D7A.4	+	
F56B3.8	+	
F33E11.6	-	
T08D2.6	-	

In fact, we can distinguish if the sequence belongs to a positive strand when the motifs are oriented to the left side. Whereas, motifs oriented to the right side indicate that the sequence belongs to a negative strand.

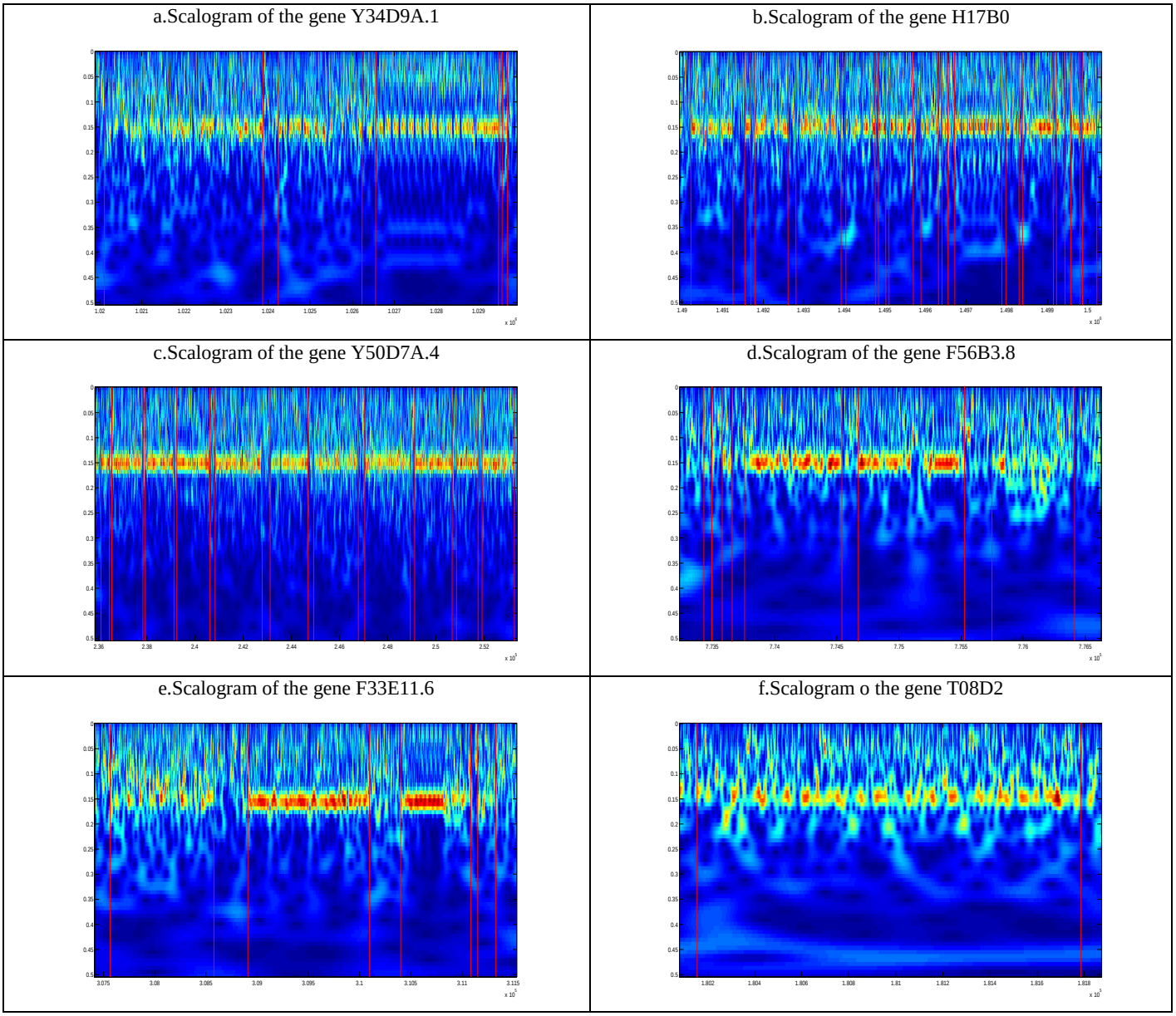


Fig. 5 Scalograms related to six gene-sequences which are coded with the $FCGS_2$ technique. The horizontal axes indicate the nucleotides positions on the *C.elegans* chromosomes and the vertical axes indicate the frequency values.The red lines delimit the Exon-Intron boundaries.

Overall, the colour multi-scale spectrum (scalogram) provides an efficient tool to characterize intron sequences as well as the related orientation. This will be a great help in the gene prediction enhancement as well as correcting the available annotations.

III. Conclusion

Aiming to provide an efficient tool that reduces errors on gene prediction, we focused on intron analysis. Thus, the complex Morlet wavelet transform applied to the "Frequency Chaos Game Signal" allowed the characterization of introns by the 6.5- base-periodicity. Moreover, the intronic periodic-patterns are directed depending on the sequence orientation.

References

i. H. Najmi and J. Sadowsky, —The Continuous Wavelet Transform and Variable Resolution Time–Frequency Analysis,‖ Johns Hopkins Apl Technical Digest, Vol. 18(1), 1997.

ii. 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.

iii. 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.

iv. M. Akhtar, J. Epps and E. Ambikairajah, —Signal Processing in Sequence

Analysis: Advances in Eukaryotic Gene Prediction,‖ IEEE Journal of Selected Topics in Signal Processing, vol. 2(3), pp. 310–321, 2008.

v. S. Datta and A. Asif, ―A Fast DFT-Based Gene Prediction Algorithm for Identification of Protein Coding Regions,‖ Proc. 30th IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, 2005, vol. 3, paper 113.116.

vi. L. Wang and L. D. Stein, ―Localizing triplet periodicity in DNA and cDNA sequences,‖ BMC Bioinformatics, 2010.

vii. E. V. Koonin and M. Y. Galperin, Sequence-Evolution-Function: Computational Approaches in Comparative Genomics, Kluwer Academic, 2003.

viii. (2013) The NCBI GenBank database. [Online]. Available: http://www.ncbi.nlm.nih.gov/Genbank/.

ix. J. Spieth and D. Lawson, ―Overview of gene structure,‖ Wormbook, 2006.

x. A.E. Oueslati, I. Messaoudi, Z. Lachiri, and N. Ellouze, ―Spectral Analysis of global behaviour of C. Elegans Chromosomes, Fourier Transform Applications, ed. by Salih Mohammed Salih, INTECH, 2012.