# Exon_Intron Separation Using Amino Acids Groups Frenquency Repartition as Coding Technique

Afef Elloumi Oueslati
Unite Signal, Image et Reconnaissance de Formes,
Département de Génie Electrique, ENIT, BP 37, Campus
Universitaire, Le Belvédère, 1002, Tunis

Noureddine Ellouze
Unite Signal, Image et Reconnaissance de Formes,
Département de Génie Electrique, ENIT, BP 37, Campus
Universitaire, Le Belvédère, 1002, Tunis

***Abstract*—this paper presents a new coding technique based on amino acids repartition in chromosome. The signal generated with this coding technique constitutes, after treatment, a new way to separate between exons and introns in a gene. The algorithm proposed is composed of six steps. We convert from ATCG to amino acids. We specify the amino acid order group. We constitute the signal based on group's repartition. We apply a smoothing technique on resulting signal. We inverse the exons peaks from minima to maxima. We show the separation between exons and introns regions. We present here the results obtained on the gene reference G 56F11.**

***Keywords*—exons; introns; amino acid coding technique; amino acid repartition; exons - introns separation**

## I. Introduction

Genomic signal processing consists in applying signal processing method to code and analyze the genome. These analyses can be made on DNA sequences, RNA sequences or proteins. All of them are represented by characters. DNA and RNA are represented by four characters and proteins are made of twenty letters. The succession of the DNA's bases: A, G, C, and T, constitutes the hereditary message. Each DNA fragment involves a specific protein synthesis process. A set of 20 different amino acids synthesize proteins, following subsequent order of three bases called codon. A total of 64 different combinations specify 20 amino acids and three stop codons, namely TAA, TAG, and TGA. Signal processing methods have focused on genomic signals analysis and many methods of coding and analyzing are proposed [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. The application here concerns the protein coding regions (exons) and non-coding regions (introns). These regions are sometimes located by interpreting coding techniques results directly. But generally, the revealing periodicities are necessitating some transforms to make decision. In fact, the weighted real and complex value proposed by Anastassiou in [18, 19] needs an analyzing technique to decide on the sequences' category. This coding technique needs the spectral analysis to enhance the specific features of exonic regions [6, 7, 9, 10, 11, 20, 21, 22, 23, 24, 25, 26, 27]. The DNA walks consist in summing the values of each base or nucleotides' class [28, 29, 30]. A grammar of alphabets defines the location of each base in a codon or three bases [31]. The tetraedrical representation reveals the exon's characteristics and contributes to the identification of these regions [32, 33]. The aim of each coding method is to improve the hidden information for further analysis. In this context, statistic analysis has been elaborated to define the similitude degree and the region's dependence. The purpose is to find differences between the coding and non coding regions in the DNA sequence. Li demonstrates the existence of a correlation between all the regions [34, 35]. Peng et al. show on the other hand that the long range correlation is related to the intronic regions [36]. We focus here on periods based methods to analyze data. A method, called periodicity transforms proposed by Sethares and Staley detects periodicities in data with projections onto periodic subspaces [37]. This technique isn't based on a predefined set basis such as Wavelet or Fourier Transform. It associates different algorithms to calculate a specific set of non orthonormal basis elements which are directly depending on the analyzed data. This technique reveals periodicities relevant to signals by decomposition into the basic periodic components. A method based on amino acid coding for coding region detection followed by principal component analysis and wavelet transform is proposed in [38]. In our previous work presented in the reference [39], we focus on a particular and a specific periodicity, so we propose to apply the pitch synchronous analysis on genomic DNA sequences. This technique is based on the wavelet transform. The proposed method in this paper is dealing with a coding technique based on amino acids and its frequency repartition in the chromosome. In fact, the protein sequence is traduced into numerical signal by replacing each amino acid group by its frequency repartition. We demonstrate that such signal is an efficient tool to separate between exon and intron in a gene.

The paper is divided into 5 sections. Section 2 describes the coding technique, detailing the frequency order groups of amino acids used. Section 3 presents the methodology used to convert the signal obtained by the coding technique to a segmenting tool allowing the separation of exon and introns. Section four illustrates the accuracy of the proposed method by presenting the results obtained. The last section concludes the paper.

## II. The Coding Technique Based On Amino Acids Groups Repartition

The coding technique proposed is based on the probability (frequency) of apparition of each amino acid group in the entire chromosome. The repartition order corresponds to the number of amino acid in a group. The order one specify the probability of repartition of each amino acid (AA) as: A, D, M, N…. The order 2 is related to a combination of two consecutives AA as MN, MA, NY,…. The third order is for three successive AA such MNA, ADA,…..

These probabilities are calculated by the following equation

$$Paag = \mathrm{Naag}/\ \mathrm{Naach} \qquad (1)$$

*With Paag represents the Probability of one amino acid group,* Naag *represents the number of one amino acid group in the entire chromosome and* Naach *represents the number of all amino acids in the chromosome*

For example:

Order 1:

*For amino acid M* $Pm=\mathrm{Nm}/\ \mathrm{Naach};$

*For amino acid Y* $Py=\mathrm{Ny}/\ \mathrm{Naach}$

Order 2:

*For amino acid group AM* $Pam=\mathrm{Nam}/\ \mathrm{Naach};$

*For amino acid group YZ* $Pyz=\mathrm{Nyz}/\ \mathrm{Naach}$

Order 3:

*For amino acid group DTS* $Pdts=\mathrm{Ndts}/\ \mathrm{Naach}$

*For amino acid group YYY* $Pyyy=\mathrm{Nyyy}/\ \mathrm{Naach}$

To code the DNA sequence, we use the probabilities calculated for each order. In fact, each value in the numeric signal is obtained by replacing each position k of one amino acid group by its probability of repartition values as expressed in equation 2

$$Saa\ (k) = \Sigma_i\ Paag\ (i,k) \qquad (2)$$

*The indice i determines the aag group and k represent the position in the sequence to code*

The entire numeric signal is then obtained by calculating the sum of these probabilities on the entire sequence as expressed in equation 3

$$Sa = \Sigma_\kappa\ Saa\ (k) \qquad (3)$$

The probabilities are calculated on the entire chromosome but the numeric signal can be applied only on the considered sequence to code, in our case the sequence is the gene. When the sequence is coded, the next step consists in dealing with the numeric signal to show the separation between exons and introns in a gene.

## III. The Exon Introns Separation Methodologie

The aim of the proposed coding technique is to distinguish between exons and introns in a gene. Using the probability of repartition into the coding technique is a way to avoid the analysis method. In fact, the numerical signal obtained after the coding is able to separate between the considered regions. We just add some preprocessing techniques. The methodology proposed begins from chromosome with its ATCG form and finishes with highlighting the exons-introns limits in a gene. The approach proposed is divided into six parts as follows:

- Converting chromosome from the ATCG form to amino acid form via the genetic code ATATCGATCTG→ISI*

  With *represents stop codons

- Specifying the repartition order which corresponds to the number of amino acid used in a group. Order 1 for one amino acid, order 2 for two amino acids…etc.

- Calculating the probabilities Paag on the entire chromosome. We present in the Table1 the Paag for each amino acid for the order 1. Table 2 presents some examples of the Paag for the order two.

TABLE I. Amino Acid Probabilty For Order 1

| ***Amino acid group (order 1)*** | ***Paag*** | ***Amino acid group(order 1)*** | ***Paag*** |
|---|---|---|---|
| L | 0.0940 | Q | 0.0349 |
| F | 0.0929 | P | 0.0349 |
| S | 0.0866 | G | 0.0348 |
| K | 0.0830 | A | 0.0342 |
| I | 0.0724 | Y | 0.0292 |
| R | 0.0601 | C | 0.0272 |
| N | 0.0553 | H | 0.0252 |
| T | 0.0461 | D | 0.0238 |
| V | 0.0460 | M | 0.0149 |
| E | 0.0398 | W | 0.0126 |
| * | 0.0520 | - | - |

TABLE II. Examples of Amino Acid Probabilties for Order 2

| ***Amino acid group (order2)*** | ***Paag*** |
|---|---|
| FF | 0,0115 |
| FL | 0,0107 |
| FS | 0,0103 |
| KK | 0,0102 |
| IF | 0,0099 |
| KI | 0,0095 |
| LL | 0,0094 |
| LK | 0,0089 |
| SS | 0,0085 |
| KL | 0,0079 |
| NF | 0,0078 |

- Coding the gene with the probabilities of the specified order. For example the numeric sequence of seqaa=ISI* for the first order is

  seqn= 0.0724 0.0866 0.0724 0.0520

- The resulting signal needs smoothing to reveal clearly the different region in a gene. We choose simply to apply the mean value on a number of consecutive neighbors. In our case we fix the number to 18.

  Each value k is replaced by the mean value of its 18 neighbors as expressed in equation 4

$$Smean(\kappa)= \Sigma i\ Saa\ (i) \qquad (4)$$

  With index i is varying from k to k+18. The smoothed signal is the sum for all the values k, the expression is given by equation 5

$$Ssmooth= \Sigma\kappa\, Smean\ (k) \qquad (5)$$

  With the index k is varying from 1 to the length of the gene.

  Finally, the obtained signal shows that exons are characterized with minima so we inverse signal to have the exons as maxima with the equation 6:

$$Sfinal= 1 - Ssmooth \qquad (6)$$

- The Sfinal of equation 6 is the signal on which we distinguish the exons and introns. In fact, the peaks represent the exons regions and the separation is clear.

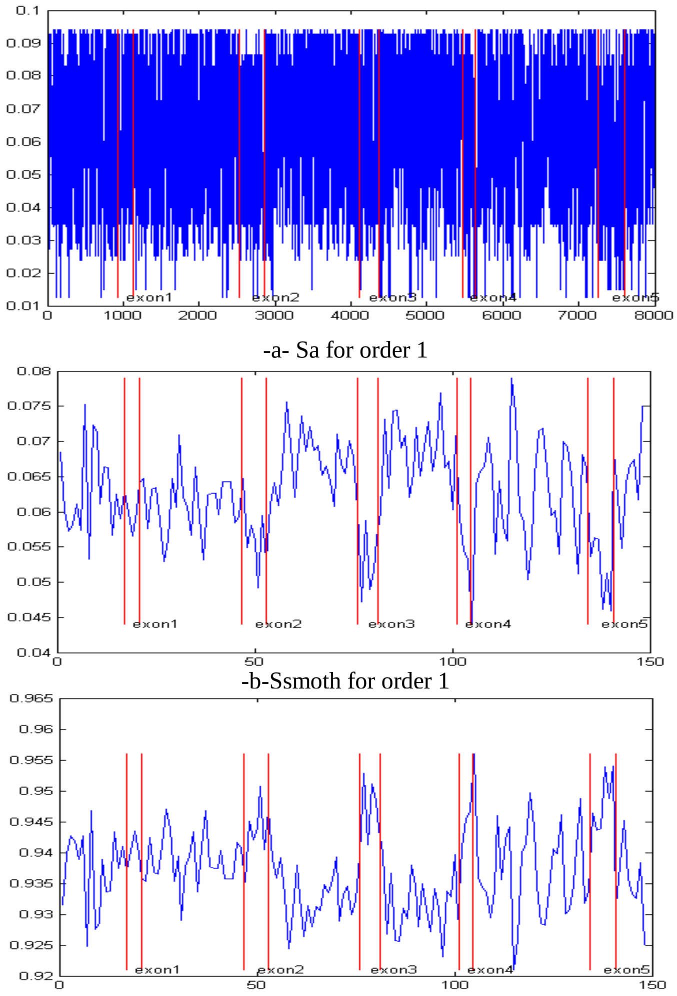
We test in our analysis different orders. We illustrate here the results for orders from order 1 to order 4. We remind that the order 4 is a combination of four consecutive amino acids. We present here the results obtained for these 4 orders for the reference gene G56F11. In the table III, we present the exon's positions for the five exons initially and after applying the mean value as smoothing technique with the mean value obtained for 18 neighbors.

TABLE III. EXONS' POSITION FOR GENE G56F11

| Exon's number | Exon's position | |
|---|---|---|
| | *Initial* | *After mean value* |
| 1 | 929-1135 | 17-21 |
| 2 | 2528-2857 | 46-52 |
| 3 | 4114-4377 | 76-81 |
| 4 | 5465-5644 | 101-104 |
| 5 | 7255-7605 | 134-140 |

The methodology's steps are illustrated by the figure 1for the reference gene G56F11.

For each order from order 1 to order 4, we consider 3 figures. The methodology's steps are illustrated by figure 1.



-a- Sa for order 1

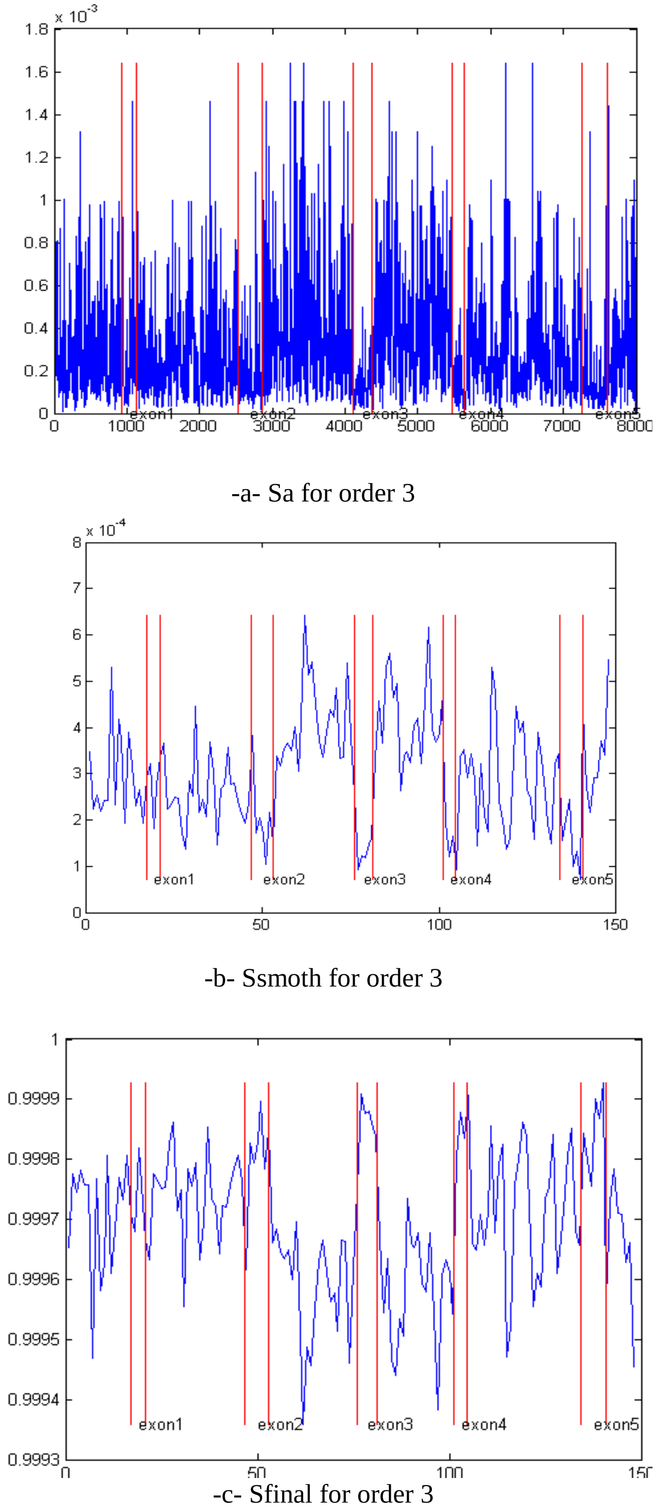
-b-Ssmoth for order 1

-c-Sfinal for order 1

Fig. 1. The methodology's steps for order 1. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal

We present in each figure the results obtained for each order. Figure 2 is related to order 2. Figure 3 exposes order 3 results and figure 4 the exon intron separation with the probabilities repartition of order 4.

These figures are subdivided as follows. The first one is the signal corresponding to the probabilities repartition Saa. It is calculated for the whole chromosome and applied to the gene G56F11. The initial exon's positions are delimited with red lines. The second signal is the smoothed one Ssmooth. The modified exon's positions after the mean value, given in table 3, are represented with red lines to highlight the regions separation. The third subfigure is the final signal enhancing the exon's peaks as maxima to clearly present the exons _introns separation.

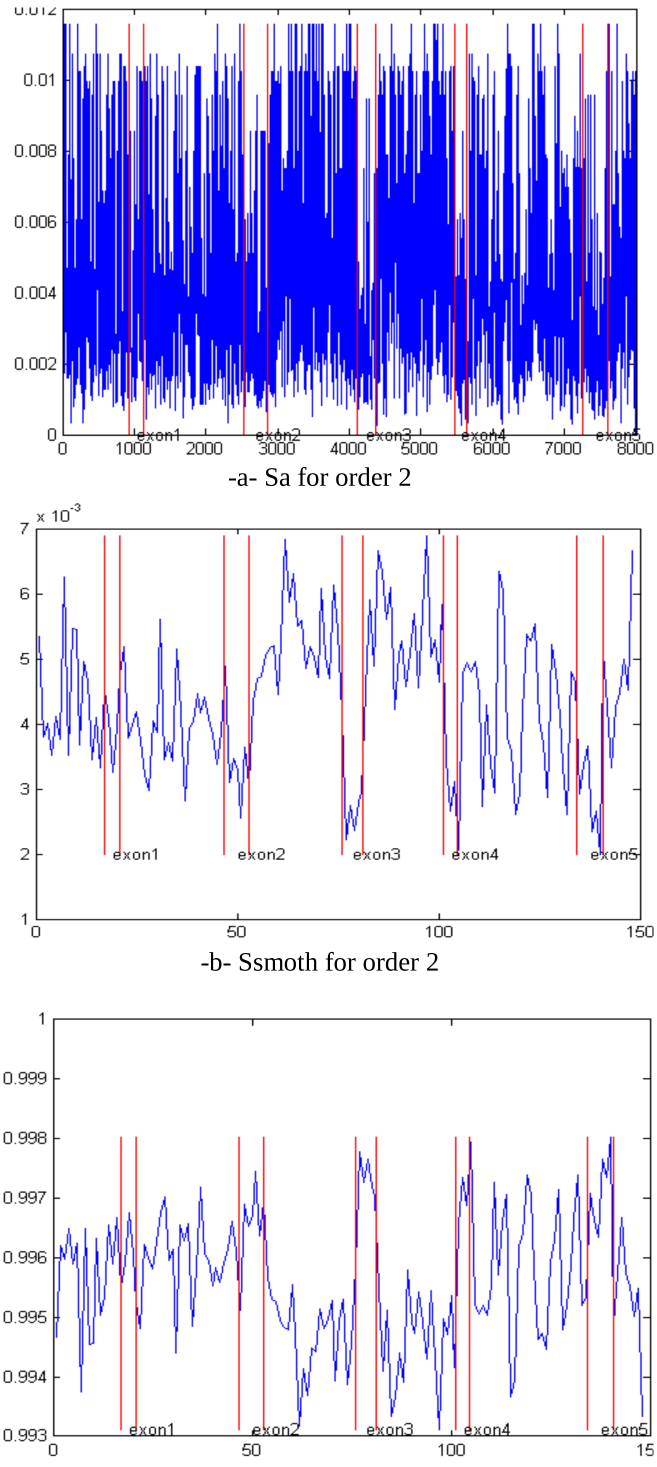


-a- Sa for order 3

-b- Ssmoth for order 3

-c- Sfinal for order 3

Fig. 2. The methodology's results for order 2. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal

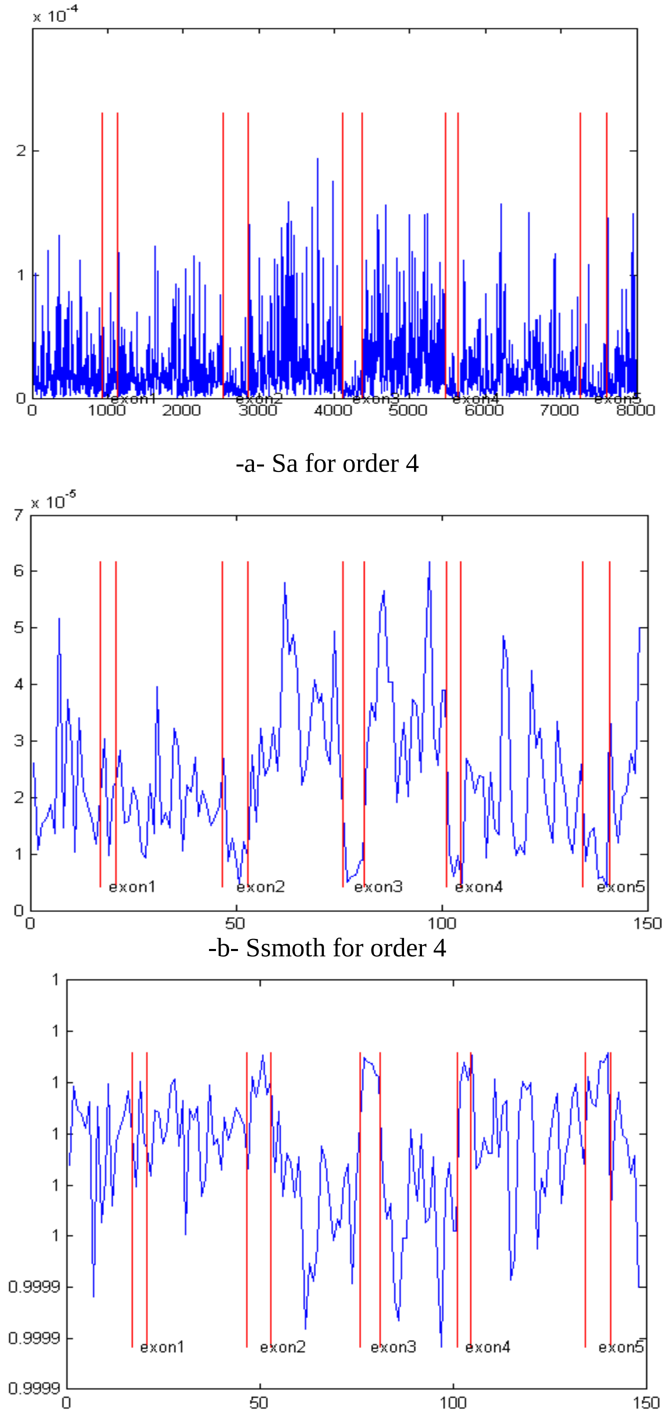


-a- Sa for order 2

-b- Ssmoth for order 2

-c- Sfinal for order 2

Fig. 3. The methodology's results for order 3. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal



-a- Sa for order 4

-b- Ssmoth for order 4

-c- Sfinal for order 4

Fig. 4. The methodology's results for order 4. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal

## References

[1] C. Mathé, M.F. Sagot, T. Schiex, and P. Rouzé, "Current methods of gene prediction, their strengths and weaknesses," Nucleic Acids Research, 2002, vol. 30, no. 19, pp. 4103–4117,

[2] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," Nucleic Acids Research, 1982, vol. 10, no. 17, pp. 5303–5318.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, UK, 1998.

[4] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," EURASIP Journal on Applied Signal Processing ,2004, vol. 2004, no. 1, pp. 13–28,.

[5] J. V. Lorenzo-Ginori, A. Rodrıguez-Fuentes, R. G.Abalo, and R. S. Rodrıguez, "Digital signal processing in the analysis of genomic sequences, "Current Bioinformatics, 2009, vol.4,no.1,pp.28–40.

[6] S. Nancy Yu and Y. Hong. Short exon detection in DNA sequences based on multifeature spectral analysis. EURASIP Journal on Advances in Signal Processing, 2011.vol no pp

[7] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, "Computer Applications in the Biosciences , 1997, vol. 13, no. 3, pp. 263–270.

[8] M. Akhtar, E. Ambikairajah, and J. Epps, "Optimizing period- 3 methods for eukaryotic gene prediction," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08), 2008, pp. 621–624.

[9] H. Yan and T. D. Pham, "Spectral estimation techniques for DNA sequence and microarray data analysis, 2007, "Current Bioinformatics, vol. 2, no. 2, pp. 145–156,.

[10] M. K. Choong, and H. Yan, Multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves. Bioinformation, 2008. 2(7), 273-278.

[11] R. Jiang and H. Yan, "Studies of spectral properties of short genes using the wavelet subspace Hilbert-Huang transform (WSHHT)," Physica A , 2008, vol. 387, no. 16-17, pp. 4223–4247.

[12] T. P. George and T. Thomas, "Discrete wavelet transform denoising in eukaryotic gene splicing," BMC Bioinformatics, 2010, vol.11, supplement 1, article S50.

[13] Y.Wu, A.W.-C.Liew,H.Yan,andM.Yang,"DB-Curve: a novel 2D method of DNA sequence visualization and representation," Chemical Physics Letters , vol. 367, no. 1-2, pp. 170–176, 2003

[14] M. Akhtar, J. Epps, J. and E. Ambikairajah, E. On DNA numerical representations for period-3 based exon prediction. In Genomic Signal Processing and Statistics, 2007. GENSIPS 2007. IEEE International Workshop on (pp. 1-4). IEEE.

[15] 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. 2012 vol 3, no 11

[16] K.S. Sathish and N. Duraipandian. An effectice identifcation of Species from DNA Sequence: A Classification Technique by integrating DM and ANN. International Journal of Advanced Computer Science & Applications. 2012 vol 3, no 8

[17] S.N. Devi and S.P. Rajagopalan. A study on Feature Selection Techniques in Bio-Informatics. International Journal of Advanced Computer Science & Applications. 2011 vol 2, no 1,pp 138-144

[18] D. Anastassiou. "Genomic signal processing". Signal Processing Magazine, IEEE, 2001, vol. 18, no 4, p. 8-20.

[19] D. Anastassiou. "DSP in genomics: processing and frequency-domain analysis of character strings". In : Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. IEEE, 2001. p. 1053-1056.

[20] R. P Costa," Gene prediction algorithms". Computational Biology, 2003, p. 1-7.

[21] A. Elloumi, Z. Lachiri and N. Ellouze. « DNA sequence analysis: From DNA sequencing to gene prediction". In : 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation. 2005.

[22] A. Elloumi, Z. Lachiri and N. Ellouze. Spectral Analysis of DNA Sequence: The Exon's Location Method. In : Digital Signal Processing, 2007 15th International Conference on. IEEE, 2007. p. 115-118.

[23] A. Elloumi, Z. Lachiri and N. Ellouze. 3D Spectrum Analysis of DNA Sequence: Application to Caenorhabditis elegans Genome. In :

Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on. IEEE, 2007. p. 864-871.

[24] Y. Changchuan and Y. S.-T Stephen. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of theoretical biology, 2007, vol. 247, no 4, p. 687-694.

[25] J. Xianyang, D. Lavenier and Y. S.-T Stephen. Coding region prediction based on a universal DNA sequence representation method. Journal of Computational Biology, 2008, vol. 15, no 10, p. 1237-1256.

[26] A. S Marhon, and S. C. kremer. Gene prediction based on DNA spectral analysis: a literature review. Journal of Computational Biology, 2011, vol. 18, no 4, p. 639-676.

[27] Y. Changchuan and Y. S.-T Stephen. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. Journal of Computational Biology, 2005, vol. 12, no 9, p. 1153-1165.

[28] J. A. Berger, S. K. Mitra, M. Carli, Marco, and al. New approaches to genome sequence analysis based on digital signal processing. University of California, 2002. Workshop on Genomic Signal Processing and Statistics (GENSIPS), IEEE, Raleigh, North Carolina, October,pp. 1–4, .

[29] J. A. Berger, S. K Mitra, M. Carli, Marco, and al. Visualization and analysis of DNA sequences using DNA walks. Journal of the Franklin Institute, 2004, vol. 341, no 1, p. 37-53.

[30] J. A. Berger, S. K. Mitra and M, J. Astola. Power spectrum analysis for DNA sequences. In : Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on. IEEE, 2003. p. 29-32.

[31] D. Nicorici and J. Astola. Segmentation of DNA into coding and noncoding regions based on recursive entropic segmentation and stop-codon statistics. EURASIP Journal on Applied Signal Processing, 2004, vol. 2004, p. 81-91.

[32] P.D. Cristea. Large scale features in DNA genomic signals. Signal Processing, 2003, vol. 83, no 4, p. 871-888.

[33] P.D. Cristea. Multiresolution phase analysis of genomic signals. In : Control, Communications and Signal Processing, 2004. First International Symposium on. IEEE, 2004. p. 743-746.

[34] W. Li and K. Kaneko. Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. EPL (Europhysics Letters), 1992, vol. 17, no 7, p. 655.

[35] W. Li. The study of correlation structures of DNA sequences: a critical review. Computers & chemistry, 1997, vol. 21, no 4, p. 257-271.

[36] C.K. Peng, S.V. Buldyrev,A. L. Goldberger and al. Long-range correlations in nucleotide sequences. Nature, 1992, vol. 356, no 6365, p. 168-170.

[37] W.A. Sethares and W. Thomas. Periodicity transforms. Signal Processing, IEEE Transactions on, 1999, vol. 47, no 11, p. 2953-2964.

[38] C-Y. Tsai and C-C Chiu. An efficient conserved region detection method for multiple protein sequences using principal component analysis and wavelet transform. Pattern Recognition Letters, 2008, vol. 29, no 5, p. 616-628.

[39] A. Elloumi, Z. Lachiri and N. Ellouze. Detecting particular features in C. elegans genomes using Synchronous Analysis based on Wavelet Transform. International Journal of Bioinformatics Research and Applications, 2011, vol. 7, no 2, p. 183-201.