IDENTIFICATION OF GENOMIC SEQUENCES USING CORRELATION TECHNIQUE

K Satya Sravya, G Raghava, K Gireesh Varma, G Prudhivi Raj, G Santhi Kumari

Abstract— In this paper genomic sequences are analyzed by developing the pattern filtering approach. By using this approach the distance of certain pattern is first translated into a "**gap sequences**" consisting of integer numbers. These patterns result in different gap sequences, comparing of two genomic sequences can be made based upon the processing of gap sequences which are generated by a set of pre-selected patterns. For the enhancement of two sequences "**correlation technique**" is applied used for checking the similarity between two genomic sequences. Conventional techniques of non-numeric data analysis consist of assigning numeric values to non-numeric symbols and using numeric techniques for processing the resultant sequences.

Index Terms— Genomic sequences, Gap Sequences, Correlation technique, Pre –Selected Patterns, Non-numerical sequences, and Integer numbers.

I INTRODUCTION

Genomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA and proteins, have been mathematically represented by character strings, in which each character is a letter of an alphabet. In the case of DNA, the alphabet is size 4 and consists of the letters A, T, C and G; in the case of proteins, the size of the corresponding alphabet is 20.

'Gene Prediction' refers to detecting locations of the protein-coding regions (exons) of genes in a long DNA sequence. The problem constitutes one of the first steps in understanding life processes. For most prokaryotic DNA sequences, the problem is to determine which ORFs, in the given sequence, are really coding sequences coding for proteins. For eukaryotic DNA sequences, the problem is to determine how many exons and introns (non-coding regions) are there in the given sequence and what are the exact boundaries between the exons and introns.

The output from the matched filter needs enhancement for key information extraction. Several post processing techniques are introduced and applied to the filtered result for signal enhancement. For example, the **modified Butterworth window (MBW)** is used to remove the edge effect of the matched filter output, and the uncertain region is beleaguered by the **advanced similarity test (AST) algorithm**. The match between gap sequences is called a "frame match" or a "structural match". The actual match of two genomic sequences demands both frame match and stuffing match. The proposed approach is useful for sequence analysis based on the frame match with desirable patterns.

Extensive experimental results will be presented to demonstrate the performance of the proposed method. The obtained results justify the use of the gap sequence as an effective tool for genomic DNA sequence analysis. Beyond that, the gap structure can go further to the core issues of the DNA encoding such as morphological DNA structure, sequence decomposition, and advanced pattern filtering. This paper is organized in the following way: Pattern Filtering and Gap Sequences, Matched Filter, Correlation Enhancement, Modified Butterworth Window (MBW), Advanced Similarity Test (AST).

• *Author names is currently pursuing degree in electronic & communications enginerring,Jntu, Kakinada, India, PH-8106102971,9700853271, 9491767818,7893151050E-mail: sravya.satya@gmail.com,sidhuraghava@gmail.com,gireesh.k1992@gmail.com,santhikumari53@gmail.com*

II. GENOMIC SIGNALS

The amount of data stored in the field of Bioinformatics is increasing every day including genomic data such as DNA, RNA and Protein sequences. DNA sequences are in the nucleons of cells. These sequences are life codes for an organism. They specify the various tasks of life such as inheritance. DNA sequences consist of four nucleotides (symbols): Adenine (A), Cytosine (C), Guanine (G) and, Thymine (T). These nucleotides construct DNA double helix structures.

DNA sequences are transformed to protein sequences in a two-step process to complete their mission. First, the ***transcription step*** in which the DNA sequences are converted to RNA sequences. Next, the ***translation process*** occurs, in which the RNA sequences are transformed into protein sequences. Specific sites in a DNA sequence are known as ***promoters***, which determine the start position of the transcription process. After these sites every three nucleotides specify a ***codon***. In fact, each codon encodes amino acids that construct protein sequences. Protein sequences consist of twenty amino acids. Each amino acid is represented by a letter as listed in figure 8. We hereby refer to these sequences as ***genomic signals.***

1	*A*	Ala	Alanine	GCA, GCC, GCG, GCT
2	*C*	Cys	Cysteine (has *S*)	TGC, TGT
3	*D*	Asp	Aspartic acid	GAC, GAT
4	*E*	Glu	Glutamic acid	GAA, GAG
5	*F*	Phe	Phenylalanine[1]	TTC, TTT
6	*G*	Gly	Glycine	GGA, GGC, GGG, GGT
7	*H*	His	Histidine[2]	CAC, CAT
8	*I*	Ile	Isoleucine[3]	ATA, ATC, ATT
9	*K*	Lys	Lysine[4]	AAA, AAG
10	*L*	Leu	Leucine[5]	TTA, TTG, CTA, CTC, CTG, T
11	*M*	Met	Methionine[6] (has *S*)	ATG
12	*N*	Asn	Asparagine	AAC, AAT
13	*P*	Pro	Proline	CCA, CCC, CCG, CCT
14	*Q*	Gln	Glutamine	CAA, CAG
15	*R*	Arg	Arginine[7]	AGA, AGG, CGA, CGC, GG, GT
16	*S*	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
17	*T*	Thr	Threonine[8]	ACA, ACC, ACG, ACT
18	*V*	Val	Valine[9]	GTA, GTC, GTG, GTT
19	*W*	Trp	Tryptophan[10]	TGG
20	*Y*	Tyr	Tyrosine[11]	TAC, TAT

Figure 8. A list of the twenty amino acids, and codons which generate them (from Fig. 7). For example the amino acid alanine (*A*) can be generated by any one of four possible codons *GCA, GCC, GCG,* or *GCT*. The superscipts 1 to 11 indicate the eleven *essential amino acids* (some references say there are fewer than eleven). These by definition are the amino acids animals cannot manufacture–they need to eat them. Milk provides all essential amino acids, and so does a combination of grains and beans.

III. SIGNAL PROCESSING TECHNIQUES IN DNA SEQUENCE ANALYSIS

The DNA sequence contains the instructions that control nearly everything about how an organism lives, such as its development, metabolism, and sensitivity to infection. Its analysis is an important research project in genomic signal processing. With the exponential generation of complete DNA sequences, it is particularly urgent for us to decode these inherent sequence features. Many studies have been carried out to extract the characteristic segments, to reveal some hidden structures, to distinguish coding from non coding regions in DNA sequences, and to explore structural similarity among DNA sequences. Signal processing will play an important role in reaching this goal, and indeed many computational techniques have already been applied, including the artificial neural network (ANN), nonlinear model, spectrogram, and statistical techniques .In this section, the applications of WT in DNA sequence analysis will be reviewed below separately according to their different analysis tasks.

IV. GAP SEQUENCES

The distance of a certain pattern is first translated into a "gap sequence" consisting of integer numbers. Different patterns result in different gap sequences, and the similarity measure of two genomic sequences can be made based upon the processing of gap sequences generated by a set of pre-selected patterns. Furthermore, several post-processing techniques are applied to the filtered result for signal enhancement.

A. Pattern Filtering and Gap Sequences

A structure mapping technique, pattern filtering, is introduced here which keeps only the structural information in the translated sequence. We start out with an easy case of pattern filtering. Let S be a DNA sequence of length *n. S[i]* is a **DNA** character at location ***i*** of the sequence *S,* that is, *S[i]* E *{a,* ***c , g , t*** *}* , ***i*** = 1,2,. . . , *n.* To point out the locations of some specific character, say 'a', in the sequence *S,* we need an indicator .sequence, which is defined as,

$$\widetilde{I_a[i]} = \begin{cases} 0, \text{if } S[i] \neq \text{'a'} \\ 1, \text{if } S[i] = \text{'a'} \end{cases}, \text{ where } i = 1, \ldots, n - j_a + 1. \quad (1)$$

The pattern length, $j_{,,}$ is unity in this case. Now we have a binary indicator sequence which is the intermediate step on the translation to sequence of numbers. The pattern filtering is to read the gap between two successive occurrences of some specific character or pattern. **To** precisely describe the location of the specific pattern, we added two virtual values **'1'** to the head and the tail of ***I,[i].*** The modified indicator sequence is

$$I_a[i] = \begin{cases} \widetilde{I_a[i]}, & \text{where } i = 1, 2, \ldots, n - j_a + 1 \\ 1, & \text{where } i = 0, n - j_a + 2 . \end{cases} \quad (2)$$

Finally the 'a'-filtered sequence ***F,[i]*** is defined as the number of steps from the ***it''*** **'1'** to the ***(i +l)t''*** **'1'** in $I_{,,}$ where $i = 0,1, \ldots,$ ***nu.*** All elements in ***F,*** are positive integers, and ***n,*** **is** the number **of** occurrences of the selected pattern 'a'. The relationships between the sequences above are demonstrated in Figure,

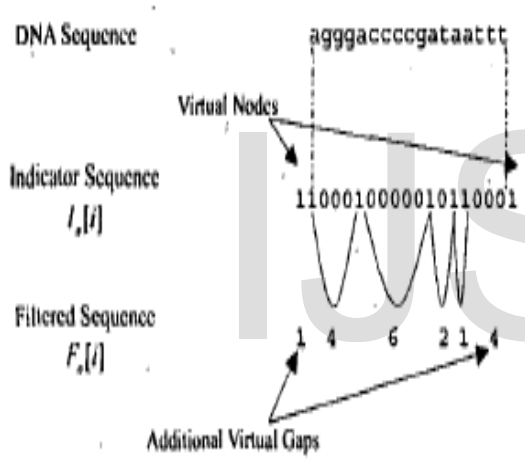


Figure 2 of the relation between DNA sequence, the indicator sequence, and the filtered gap sequence

When we measure the similarity between two sequences, the correlation operation is a good measurement. We are not able to pick up the similarity if we correlate the sequences in the head-to-head manner; i.e., correlating from the beginning of both sequences. To allow checking in every possible starting location, the correlation must be collected in every possible shift between the two sequences. This is equivalent to convoluting these two sequences in the opposite direction.

We can expect to obtain some spike(s), denoted by ***C[i],*** in the matched filter output by the convolution operation for the query and the target sequences. To check the significant similarity, a detection threshold will be set to sweep out the insignificant similarities. Regarding to finding the locations of similarities, the location of the spikes will indicate the amount of corresponding shifts between the query and the target sequences. At the matched filter output is a collection **of** different shifting-amount correlations. Let F_t ***[i]*** and F_q ***[i]*** be the target and the **query** gap sequences, respectively. The matched filter output is,

$$C[i] = F_t[i] \otimes F_q[-i] = \sum_{k=-\infty}^{\infty} F_t[i] F_q[i + k]. \quad (3)$$

Let n_q and n_t denote the length of corresponding gap sequences F_y***[i]*** and F_t***1z],*** the length of ***C[i]*** could be as long as n_t+ n_q- ***1.*** The time complexity to compute *C[i]* is *O* (n_t+ n_q)***,*** including the FFT, multiplication in Fourier domain, and IFFT. The result is shown in Fig. 2(a), from which we can find at least two problems. The first one **is** that the raw output signal sequence of matched filter is not normalized, which makes the decision of spike locations more complicated. Another problem **is** the edge effect at both beginning and end portion of the output signal due to the insufficient correlating points. The edge effect introduces high variation to the *C[i]* values close to the head and tail of the whole signal. We have several basic solutions for both problems.

IV.POST PROCESSING TECHNIQUE

Two problems have been pointed out by judging the result signal sequence at the output of matched filter in Figure. One is that the output signal is not normalized, and the other is the **edge** effect appearing at the head and the tail portion of the output signal. Two processes are designated to solve these problems. Before fed info the matched filter, we first apply the normalization process to both query gap sequence F_p and target gap sequence F_t, Then use the proposed edge effect reduction process with the output signal of matched filter to reduce the edge effect.

A. Correlation Enhancement

The bias component of a sequence is the sequence mean. We have the matched filter collecting the cross-covariance rather than the cross-correlation, as in Figure by setting

$$C[il = (\boldsymbol{Ft}[-\boldsymbol{i}\,]\, mt) \otimes (\boldsymbol{Fg}[-\boldsymbol{i}]\, \boldsymbol{m}\, \boldsymbol{q})$$

To cope with the edge effect, it is desirable to multiply the output signal with another signal sequence, which serves as a weighting coefficient sequence emphasizing the middle portion of the target sequence. The multiplication is done in the value by-value manner, and is called modulation. For simplicity, we apply a half-period sin-e wave (HPS) to modulate *C[i]. T*hus, we can obtain,

$$C'[i] = C[i].\sin\left(\frac{i\pi}{nq + nt - 1}\right)$$

The resulting *C'[i]* of the correlation enhanced matched filter, which applies the combination of both normalization and edge effect reduction processes. The edge effect reduction process suppresses the uncertain regions in the output sequence **of** matched filter. The correlations close to the edge shall look like and thus reduce their intensities by modulating with a sequence of weighting coefficients. HPS is a fixed-shape sequence of weighting coefficients. Although it works quite well judging from the result in our specific case where the lengths of F_q and F_t are not far away from each other, the modulation result will turn worse when $|n_q|$, - $|n_t|$ becomes large.

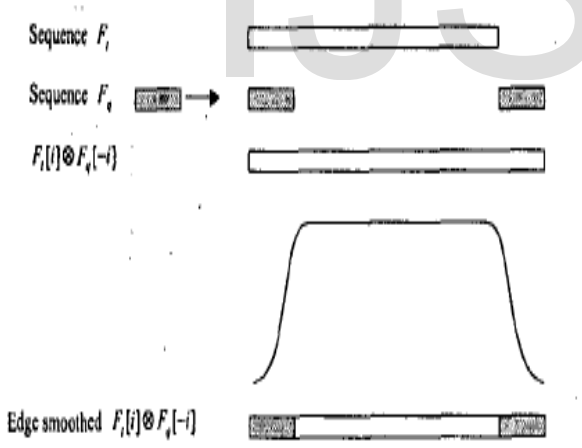


Figure 3: The regions with and without confidence: The uncertain regions are the shaded segments in the lowest sequence.

VI. EXPERIMENTAL RESULTS

In this section the proposed method is applied on real DNA and protein sequences and the results are compared with some previously proposed methods. The two DNA sequences have been taken from the Gen- Bank at the National Center for Biotechnology Information (NCBI). The accession numbers of the two sequences are AF320294 and AF324494.Demonstrates the correlation between the two DNA sequences using the proposed method.

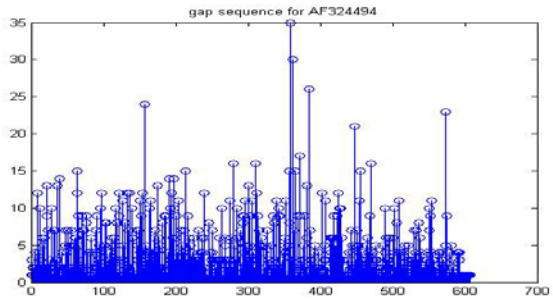


Figure (a)

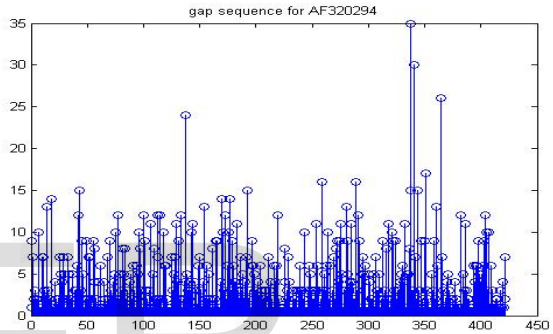
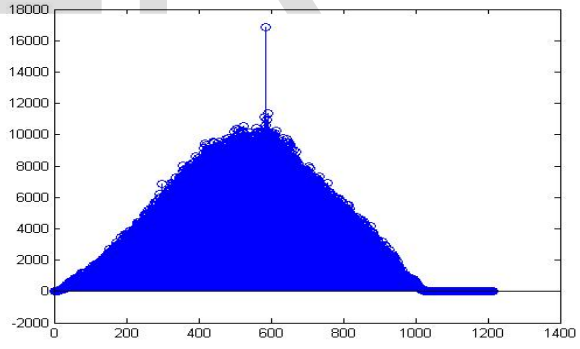


Figure (b)



Figure(c)

Figure 4: (a) Gap Sequence for AF324494 (b) gap sequence for AF320394 (C) Correlation between above two gap sequences.

This approach can be used in measuring the similarity of two protein sequences. Only one indicator sequence has been used and the effects of the other nucleotides in DNA sequences have been ignored. Apparently, this method does not use other nucleotides to compute the similarity between two sequences.

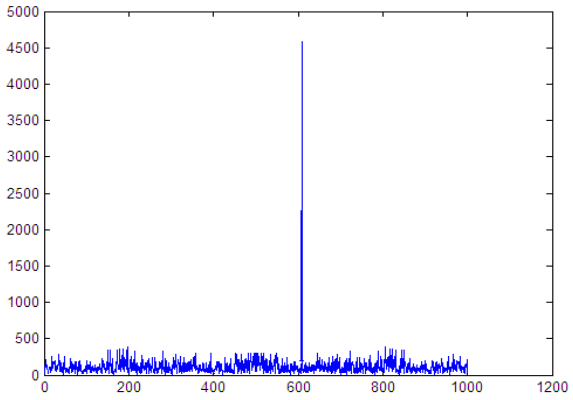


Figure (a)

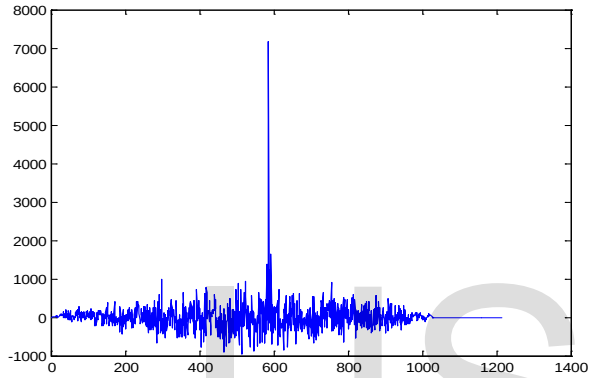
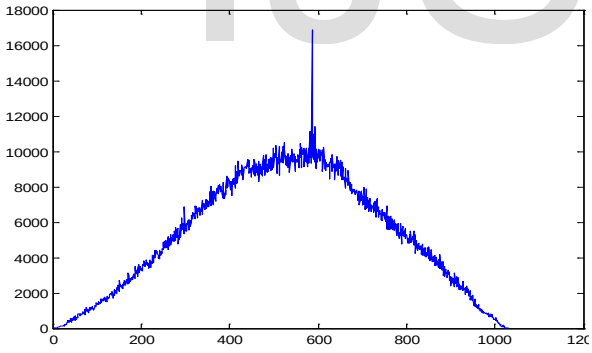


Figure (b)



Figure(c)

Figure 5 of (a) the original output has been enhanced by reducing edge effect, while (b) is the normalized version of original output. (c) Has been through both of these processes. Both sequences car1 be downloaded from Gen Bank at National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov).

The two DNA sequences have been taken from the Gen- Bank at the National Center for Biotechnology Information (NCBI). The accession numbers of the two sequences are AF065986 and AF015224. Demonstrates the correlation between the two DNA sequences using the proposed method. Here there is no peak occurs in the correlation graph. So, we can say that these two sequences are dissimilar sequences.

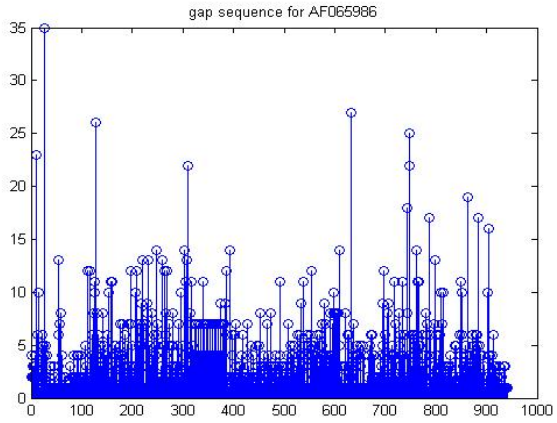


Figure (a)

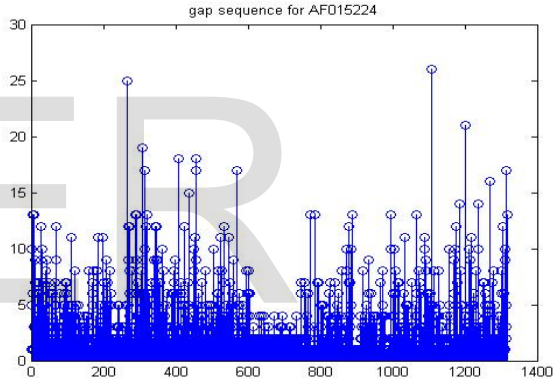


Figure (b)

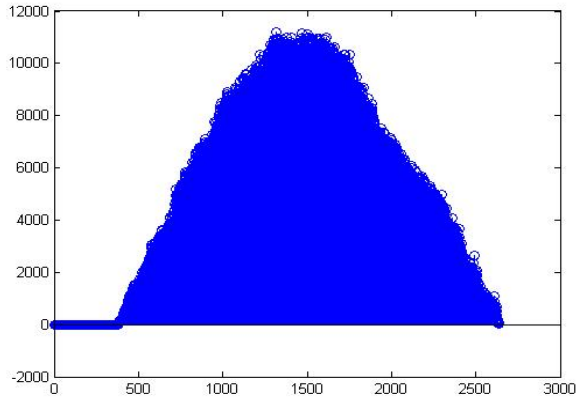


Figure (c)

***Figure: 6* of *(a)* Gap sequence for AF065986 (b) Gap Sequence for AF015224 (c) Correlation between above two gap sequences.**

VII. CONCLUSION

A technique for matching genomic sequences using the gap sequences was proposed in this work. We studied the behavior **of** gap sequences and proposed the matched filtering approach to find similar segments between gap sequences. By detecting spikes in the filtered output, we are able to locate and align similar segments between two sequences. Our major achievement in this research was to demonstrate that, given the partial knowledge of a genomic sequence segment described by a gap sequence, we can predict remaining portions **of** this segment with accurate knowledge. Simulation results demonstrated the good performance of the proposed scheme, including accurate results and a fast processing speed. In the near future, we would like: to extend the technique to find some desirable patterns in genomic sequences.

REFERENCES

[1] *Shih-Chief Su, Chia H. Yeh and C.-C Jay Kuo* **,***Structural Analysis of Genomic Sequences with Matched Filterin*,,2003:17-21

[2] Oppenheim AV, Schafer R. *Discrete-Time Signal Processing (3rd Edition)(*Prentice-Hall, NY 2009.)

[3] Proakis J G, Manolakis DK, *Digital Signal Processing (4th Edition)(* Prentice Hall, NY 2006).

[4] Stoica P, Moses RL, *Spectral Analysis of Signals*, Prentice-Hall, NY 2005.

[5] Akhtar M, Epps J, Ambikairajah E *Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction, IEEE J Select Topics Sign Proc* 2008; 3: 310-21.

[6] Hayes M.H., *Statistical digital signal processing and modeling*, John Wiley & Sons, Inc., New York, USA, 1996.

[7] Juan V. Lorenzo-Ginori , *Digital Signal Processing in the Analysis of Genomic Sequences* 2009:28-40

[8] Anastassiou, D,*Genomic Signal Processing*, IEEE Signal Processing Magazine 18,no. 4 2001):8-20.

[9] E. A. Cheever, D. B. Searls, W. Karunaratne, *and G. C. Overton,Using signal processing techniques for DNA sequence comparison, in Bioengineering Conference, 1989, pp. 173–174.*

[10] Vaidyanathan, P. P. and Byung-Jun Yoon., *Gene and Exon Prediction Using Allpass-Based Filters.*, In IEEE International Workshop on Genomic Signal Processing and Statistics, CP2-02. Piscataway, NJ: IEEE Press, 2002

[11] Tiwari, S., S. Ramachandran, A. Bhattacharya., S. thattacharya, and R. Ramaswamy. *Prediction of Probable Genes by Fourier analysis of Genomic Sequences*, Computer Applications in the Biosciences 113, no. 3 (1997):263-270.

[12] Chakrabarty Niranjan, Spanias A., Lesmidis L.D. and Tsakalis K., *Autoregressive Modeling and Feature Analysis of DNA Sequences*, EURASIP Journal on Applied Signal Processing 2004:I, 13-28.

[13] J. Gao, Y. Cao, Y. Qi, and J. Hu, "Building Innovative Representations of DNA Sequences to Facilitate Gene Finding," IEEE INTELLIGENT SYSTEMS, pp. 34-39, 2005.

[14] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "New approaches to genome sequence analysis based on digital signal processing," University of California, 2002.

[15] J. Pevsner, *Bioinformatics and functional genomics*. Wiley-Blackwell, 2009.

[16] J. Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. Scribner, 2011.

[17] B. Tropp and D. Freifelder, *Molecular biology: genes to proteins*. Jones and Bartlett Publishers, 2008.

[18] J. U. Pontius, L. Wagner, and G. D. Schuler, *The NCBI Handbook, Bethesda (MD)*. The NCBI Handbook, 2003, ch. 21. Uni Gene: a unified view of the transcriptome.

[19] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & systems (2nd ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[20] A. Papoulis, Ed., *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Companies, 1991.

[21] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Gen Bank." *Nucleic acids research*, vol. 37, no. Database issue, pp. D26–31, Jan. 2009.

[22] Vaidyanathan PP, Yoon BJ. The role of signal-processing concepts in genomics and proteomics. *J Franklin Inst* **2004**; 341: 111-35.

[23] Tuqan J, Rushdi A. A DSP perspective to the period-3 detection problem**.** *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS '06* **2006**: 53-54.

[24] Fox TW, Carreira A. A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression. *EURASIP J ApplSign Proc* **2004**; 1: 108-11.

[25] Afreixo V, Ferreira PJSG, Santos D. Fourier analysis of symbolic data: A brief review. *Digit Sign Proc* **2004**; 14: 523-30.

[26] Tiwari S, Rama chandran S, Bhattacharya A, Bhattacharya S,
Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS* **1997**; 13: 263-70.

[27] Yan M, Lin ZS, Zhang CT. A new Fourier Transform approach for protein coding measure based on the format of the Z curve.
Bioinformatics **1998**; 14: 685-90.

[28] Datta S, Asif A, Wang H. Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics. *Proceedings of the IEEE Sixth International Symposium on Multimedia Software***2004**: 160-63.

[29] Datta S, Asif A. A fast DFT based gene prediction algorithm for identification of protein coding regions. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing,(ICASSP '05)* **2005**; 5: 653-56.

[30] Datta S, Asif A. DFT based DNA splicing algorithms for prediction of protein coding regions. *Proceedings of the IEEE Thirty-EighthAsilomar Conference on Signals, Systems and Computers* **2004**; 1: 45-49.

[31]Isaac B, Singh H, Kaur H, Raghava GPS. Locating probable genes using Fourier Transform approach. *Bioinform Appl Note* **2002**; 18:196-97.

[32] Stoffer D, Ombao HC, Tyler DE. Local spectral envelope: an approach using dyadic tree-based adaptive segmentation. *Ann Inst*
Statist Math **2002.** 54: 201-23.

[33] Epps J, Ambikairajah E, Akhtar M. An integer period DFT for biological sequence processing. *Proceedings of the IEEE InternationalWorkshop on Genomic Signal Processing and StatisticsGENSIPS* **2008**: 1-4.

[34] Rushdi A, Tuqan J. Trigonometric transforms for finding repeats in DNA sequences. *Proceedings of the IEEE International Workshopon Genomic Signal Processing and Statistics, GENSIPS '08* **2008**: 1-4.

[35] Berger JA, Mitra SK, Astola J. Power spectrum analysis for DNA sequences. *Proc IEEE Seventh Int Symp Sign Proc Appl* **2003**; 2: 29-32.

[36] Rodríguez-Fuentes A, Lorenzo-Ginori JV, Grau-Ábalo R. Detection of coding regions in large DNA sequences using the short time Fourier Transform. *Lect Notes Comput Sci* **2006**. 4225: 902 909.
[37] Kotlar D, Lavner Y. Gene Prediction by Spectral Rotation Measure: A New Method for identifying Protein-Coding Regions. *GenomeRes* **2003**; 13: 1930-1937.

[38] Rushdi A, Tuqan J. The Filtered Spectral Rotation Measure. *Proc the IEEE Fortieth Asilomar Conf Sign Sys Comput ACSSC '06*
2006: 1875-79.

[39] Yin C, Yau SS-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J Theor Biol* **2007**;
247: 687-94.

[40] Akhtar M, Ambikairajah E, Epps J. Optimizing period-3 methods for eukaryotic gene prediction. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP***2008**: 621-24.