International Journal of Computing and Applications | **ISSN: 0973-5704**
Volume 13, Number 1, (January-June 2018), pp 125-130

EXON PREDICTION BY PEAKING FILTER IMPLEMENTED USING FDA TOOL BOX

Amit Kumar Singh[1] and Vinay Kumar Srivastava[2]

[1-2]Electronics and Communication Engineering Department, MNNIT, Allahabad, India.
Email: [1]mr.amitbel@gmail.com, [2]vinay@mnnit.ac.in

Abstract: Proteins are the essential building blocks of the life, it governs almost every physiological processes of a living being, and in eukaryotic (cells with a nucleus) the information to create these proteins are contained in the specific section of the genome called exon. In biological science identification of these protein coding sections (exons) are useful in many ways. Signal processing plays an important role in exon prediction in DNA sequences. Despite many signs of progresses being made in the exon prediction in recent years, the performances of these methods still need to be improved. This paper presents an easy way to design peaking filter (narrow band-pass filter) on FDA toolbox for exon prediction from a given DNA sequence. This filter is further followed by the moving average filter to suppress the background noise. The simulation results show that the proposed method gives the better result than conventional anti notch filter.

Keywords: DNA, Protein coding regions; FDA toolbox; Digital filters; period-3 property.

1. INTRODUCTION

DNA's (Deoxyribonucleic acids) are molecules to store genetic information of a living organism; these are made of just four simple nucleotide adenine (A), guanine (G), cytosine (C) and thymine (T), and arranged in a double helix structure where A pairs with T and C pairs with G [1]. Instructions that governs the synthesis of protein is also hidden somewhere in DNA. Since only a small proportion of whole DNA contributes to protein coding, it is always very hard to extract out these regions. As an example, for the human genome, only 3%-5% nucleotides of DNA sequence participate in coding; rest 97% to 95% is probably the junk DNA of little relevance [2]. Specific sections of DNA that contribute to coding are known as gene. For eukaryotic cells, the gene is further divided into exons and introns and only the exons are involved in protein coding. This relationship among different section of DNA is clearly mentioned in Figure 1.

The cell machinery inside the body distinguishes coding sections from non-coding sections with the help of start and stop codons, a start codon indicates the beginning while a stop codon indicates the end of a coding section. Codons are basically triplets of nucleotides that codes for amino acids and amino acids are smaller components that linked together to synthesize different proteins. There are 64 possible codons and 20 amino acids; detail description of the mapping of codons into amino acids is mentioned in [1].

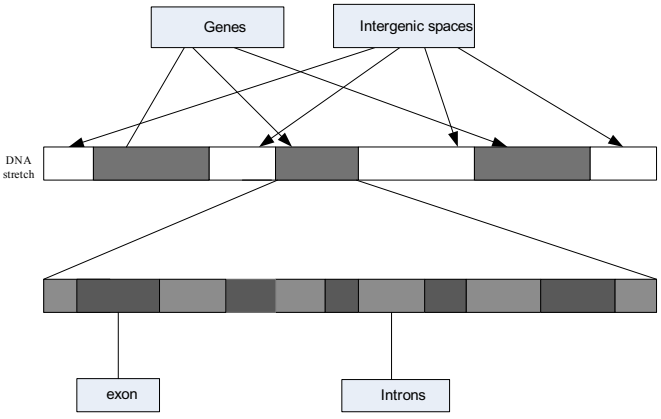


Figure 1: Illustration of different regions of a DNA stretch

It has been observed that the DNA sections related to protein coding regions have a propensity to display

strong spectral component at the frequency of $2\Pi/3$, this is called period-3 behavior. That means DFT power spectrum of a coding DNA section of length N, typically has a peak at frequency $k = N/3$. Figure 2 represents DFT power spectrum for a coding section of a length N = 2142 base pairs (bp) starting from location 309 of the gene TUP 1for a DNA stretch of S.cerevisiae chromosome III (accession number NC001135), demonstrating a peak at frequency $k = 714$. Such peaks are usually absent in non-coding regions.

Most of the signal processing techniques that are used to distinguish coding and noncoding region exploit this inherent period-3 behavior of coding region. The reason behind this periodicity is the uneven distribution of codons, also known as codon bias i.e. some codons are used more frequently in protein synthesis than others [3].

The remaining paper is organized as follows: section II includes discussion about fundamental steps involved in filtering approach along with conventional filter used in literature; in section III proposed method is described. Result and analysis are discussed in section IV. Finally, the paper is concluded in section V.

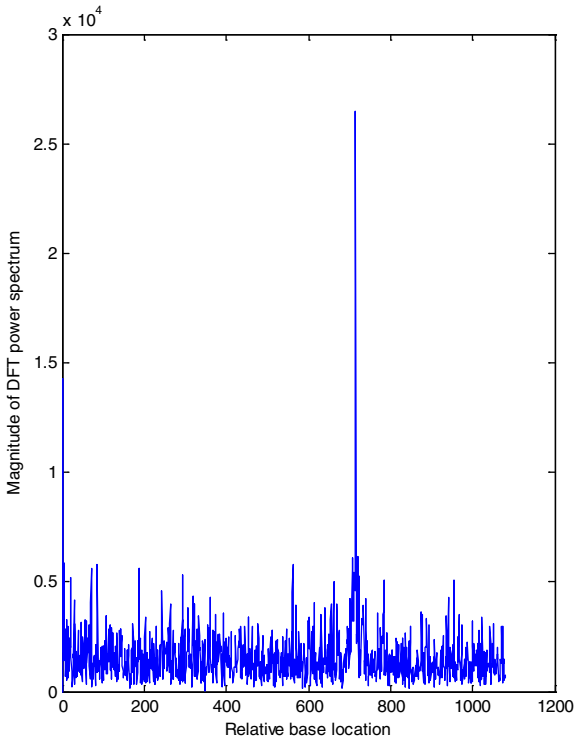


Figure 2: DFT power spectrum for a coding region gives peak at $k = N/3$

2. FILTERING APPROACH

Signal processing methods that has been widely used for protein coding region identification are basically of two types one that uses short-time discrete Fourier transform (ST-DFT) for the period-3 measure and other that uses digital filters. Digital filtering techniques have the advantage over ST-DFT methods in terms of computational speed, they are significantly faster than ST-DFT methods.

A. DNA to Numerical Mapping

Prior to applying a filtering technique for exon prediction, it is required to convert the character string of DNA sequence into numeric signals. The simplest and widely used DNA to numerical mapping method is Voss sequencing method (or binary indicator sequences) [4]. In this technique, the DNA sequence is mapped into four numeric signals corresponding to each nucleotide i.e. $x_i(n)$, $i \in B = \{A, T, G, C\}$. The numeric signal $x_i(n)$ of length M takes the value of 0 or 1 depending on the absence or presence of the base i at position n as shown in Table 1.

Table 1
Example of Voss mapping for a given DNA Sequence

DNA sequence	*G*	*T*	*A*	*C*	*G*	*T*	*A*	*C*	*G*	*T*
$x_G(n)$	1	0	0	0	1	0	0	0	1	0
$x_T(n)$	0	1	0	0	0	1	0	0	0	1
$x_A(n)$	0	0	1	0	0	0	1	0	0	0
$x_C(n)$	0	0	0	1	0	0	0	1	0	0

B. Extraction of Coding Section by Narrow Band Band-Pass Filter

The idea is to design a filter that passes the period-3 spectral component of numeric sequences and suppress the other spectral component. A narrow band band-pass filter (Anti notch filter) with center frequency $\omega_o = 2\pi/3$ is an obvious choice for this objective. The filter must provide high gain in passband region. If we give the binary sequences $x_i(n)$, $i \in B = \{A, T, G, C\}$ separately as input to this filter, and let assume their corresponding output as $y_i(n)$, $i \in B = \{A, T, G, C\}$ then power spectrum can be evaluated by equation (1)

$$Y(n) = \sum_{i \in B} \left| y_i(n) \right|^2, \; B = \{A, T, G, C\} \qquad (1)$$

A plot of $Y(n)$ gives a relatively larger peak in the coding section due to period-3 property.

On the basis of same principal, the narrowband-pass filter of various type and structure has been used. The most common filter is allpass-based IIR antinothch filter with lattice structure used in [2, 5]. A multistage filter with better stopband attenuation is also used in [2, 6]. A narrow band-pass inverse-Chebyshev kind of filter is used in [7]. Few modified antinotch filter structures are used in [8], and an improved comb filter based approach is followed in [9].

3. PROPOSED METHODOLOGY

The filter designing and analysis tool (FDA Tool) is a powerful user interface for implementing and analyzing digital filters quickly. In this paper, we have used this toolbox to implement a narrow band band-pass IIR filter (peaking filter) with direct form –II structure for exon prediction. To further improve the performance, moving average filter is used. The moving average filter is widely used for data smoothing purpose and it can also be used to eliminate background noise from the period-3 spectrums [8]. The complete data flow diagram of the proposed scheme is shown in Figure 3.

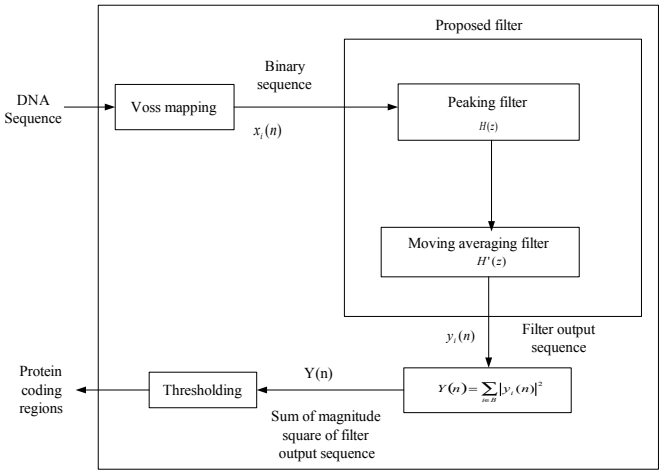


Figure 3: Data flow diagram of proposed method

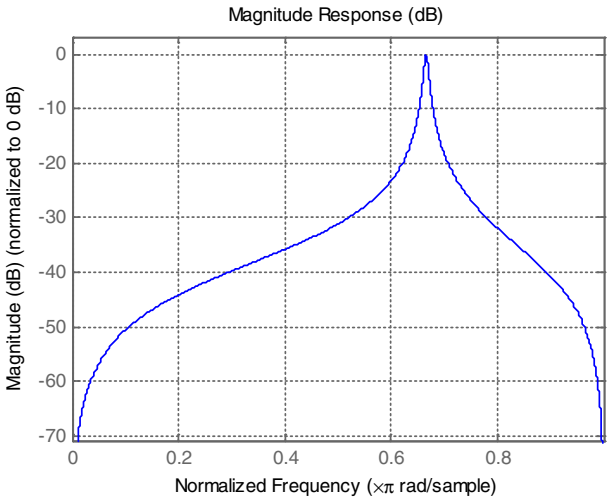
A. Peaking Filter in FDA Tool Box

A second order peaking filter with a single peak centered at $\omega_0/\pi = 2/3 = 0.666$ (normalized frequency), and with quality factor Q = 150is implemented. Magnitude Response of this filter is shown in Figure 4. Let equation (2) represents transfer function of the filter

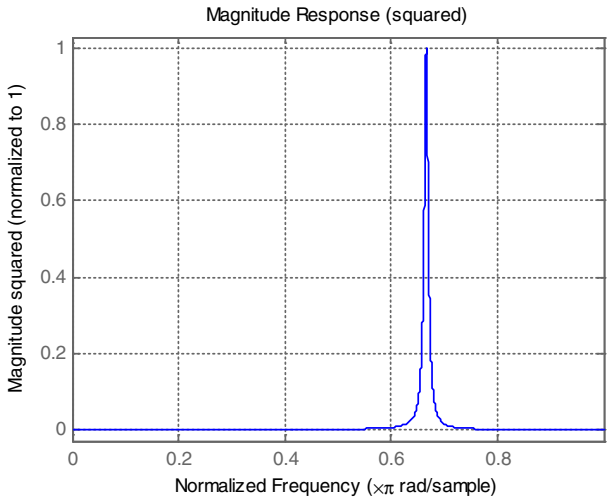
$$H(Z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \qquad (2)$$

Then obtained coefficient for this realization will be as follows:

$$b_6 = 0.013521048607837316$$
$$b_1 = 0$$
$$b_2 = -0.013521048607837316$$
$$a_1 = 0.98289824062898989$$
$$a_2 = 0.98289824062898989$$



(a)



(b)

Figure 4: Magnitude response of peaking filter: (a) decibel (dB) scale, (b) linear scale

B. Moving Average Filter

Moving average filter used in proposed algorithm is basically a second order lowpass FIR filter whose transfer function is given in (3)

$$H'(z) = \left(\frac{1 + z^{-1} + z^{-2}}{3} \right) \quad (3)$$

Figure 5 shows the pole–zero placements for moving average filter

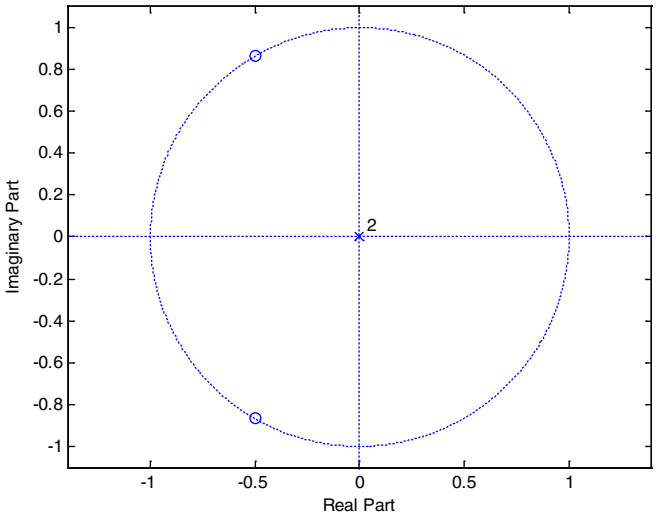


Figure 5: Pole-zero plot of second order moving average filter

C. Performance Parameter

The performance evaluation at nucleotide level is carried out with the help of receiver operating characteristic (ROC) curve. The ROC curve is usually defined as the plot of the true-positive rate (TPR) as a function of the false-positive rate (FPR) for all possible thresholds. Where TPR and FPR can be defined in terms of sensitivity (S_n) and specificity (S_p) follows:

$$TPR = S_n \quad (4)$$

$$FPR = 1 - S_p \quad (5)$$

Further, S_n and S_p are defined in terms of true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

$$S_n = \frac{TP}{TP + FN} \quad (6)$$

$$S_p = \frac{TP}{TP + FP} \quad (7)$$

where, TP denotes the number of nucleotides correctly predicted as exons, FP denotes the number of intron nucleotides predicted as exon, TN denotes the number of nucleotides correctly predicted as introns and FN denotes the number of exon nucleotides predicted as intron.

The ROC curve can be characterized as a single number using the area under the ROC curve (AUC). Closer the AUC value towards unity indicates better prediction.

4. RESULTS AND ANALYSIS

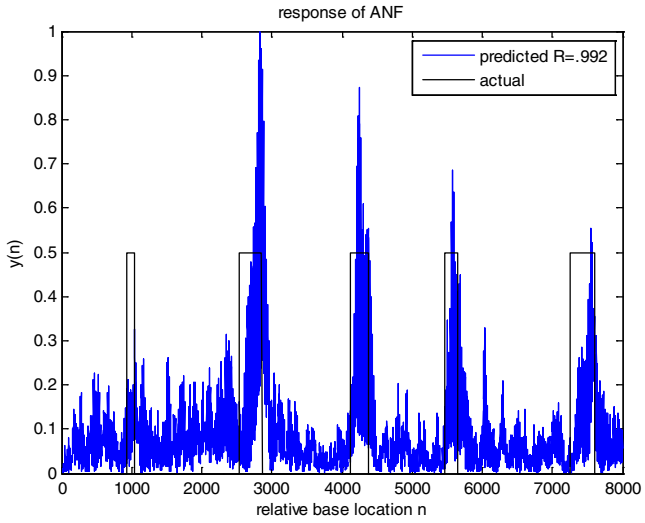
To evaluate the performance of the proposed technique we have chosen two genes. The first one is Gene F56F11.4 (Genbank accession number AF099922) of C.elegens chromosome III. This particular gene has been used by several researchers [1, 3, 6] and is, therefore, a good benchmark for comparing results of different exon prediction schemes. It has total five exons at location (928:1039, 2528:2857, 4114:4377, 5465:5644, and 7255:7605). The second gene is Gene PP32R1 (accession number AF008216) of Homo sapiens. This gene is part of well-known dataset HMR195 [10] and also used in [9].It has single exon located at position (4453:5157). The result is compared with allpass based antinotch filter with pole radius R = 0.992 as reported in several papers [3, 8].

A. Power Spectrum Analysis

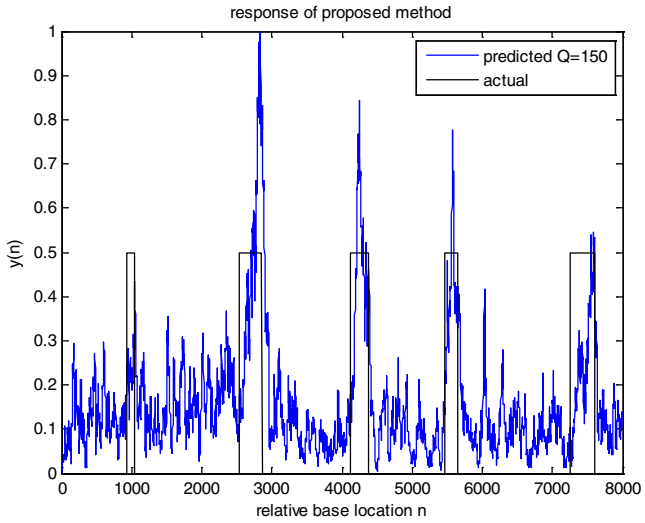
Figure 7 represents the power spectrum plot for gene F56F11.4 and Fig.8 represents the power spectrum plot for gene PP32R1. Detection of the smaller coding region is always a challenging task, we can see that the spectral peak of smaller exon i.e. the first exon (112 nucleotides) of gene F56F11.4 is more prominent in the proposed method. It is also clearly visible from both the Figure 6 (b) & 7(b), that the Background noise is effectively reduced with the proposed method.

B. Reciever Operating Characteristic Curve (ROC) Analysis

Figure 8 (a) & (b) represents the ROC curve for the gene F56F11.4 and gene PP32R1resspectively and

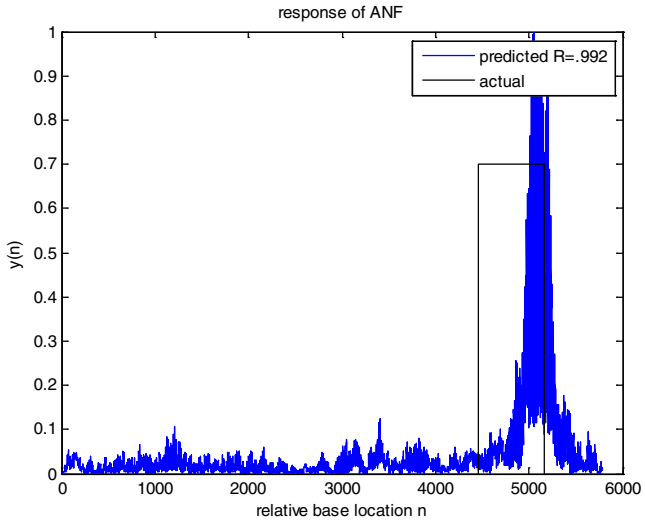


(a)

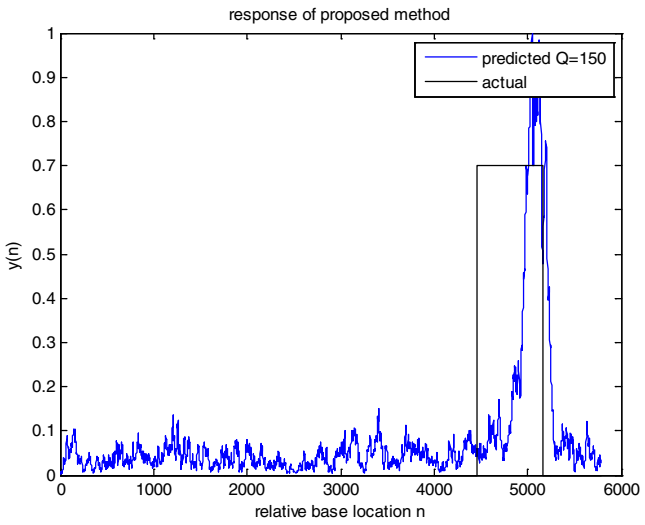


(b)

Figure 6: Output power spectrum for gene F56F11.4 (a) ANF (b) Proposed



(a)



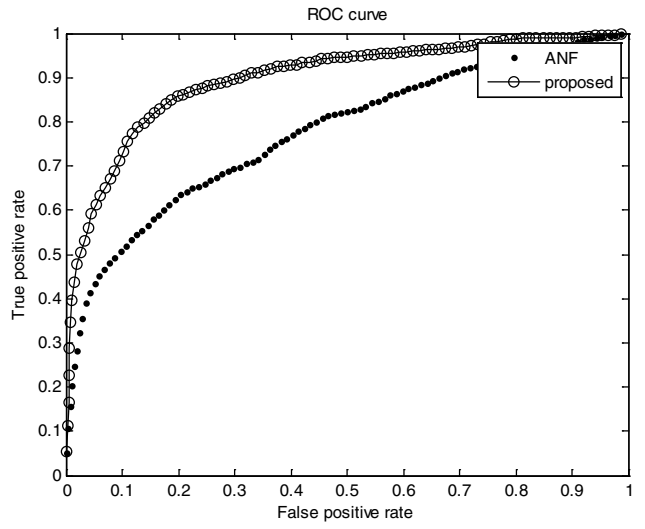
(b)

Figure 7: Output power spectrum for gene PP32R1 (a) ANF (b) Proposed

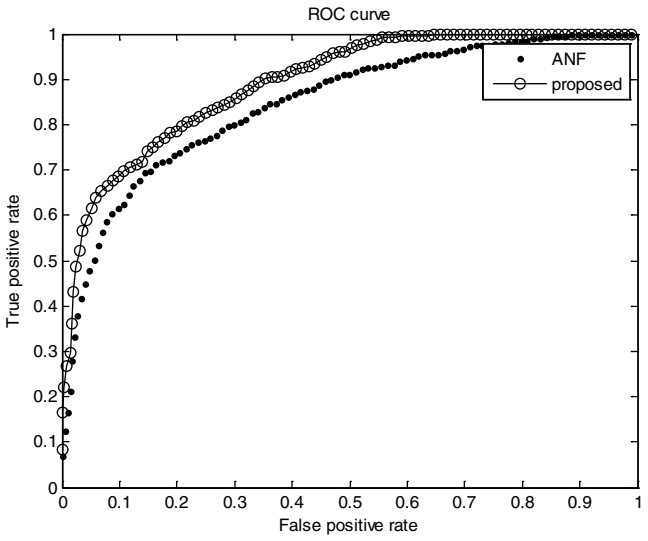
performance in terms of AUC is mentioned in Table 2. It is found that the area under curve obtained by proposed method (AUC = 0.8858 & AUC = 0.8814) is larger than the area under curve obtained for allpass based antinotch filter (AUC = 0.7659 & AUC = 0.8359).

Table
Comparison of auc value

Gene	*AUC Value*	
	ANF mentioned in [8] (R = 0.992)	*Proposed (Q = 150)*
F56F11.4	0.7659	0.8858
PP32R1	0.8359	0.8814



(a)



(b)

Figure 8: ROC curve (a) gene F56F11.4 (b) gene PP32R1

5. CONCLUSION

In this paper, a peaking filter has been designed using FDA toolbox and applied for exon prediction. Before analyzing the Gene sequence by the proposed filter, it is converted into numeric signals by using Voss mapping technique. Second order moving average filter is used to diminish the background noise. It is found that peaking filter is able to enhance peaks in power spectrum due to the coding region. It has been observed that the proposed method outperforms over well-known all pass based antinotch filter in predicting smaller protein coding section. In addition, the performance at the nucleotide level has also been improved.

Reference

[1] D.Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, Vol. 18, No. 4, pp. 8-20, 2001.

[2] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah, "Signal processing in sequence analysis: Advances in eukaryotic gene prediction," IEEE Journal of selected topics in signal processing, Vol. 2, No. 3, pp. 310-321, June 2008.

[3] P.P. Vaidyanathan, and BJ. Yoon, "The role of signal-processing concepts in genomics and proteomics," Journal of the Franklin Institute, Vol. 341, pp. 111-135, 2004.

[4] RF. Voss, "Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences," Phys. Rev. Lett., Vol. 68, pp. 3805-3808, 1992.

[5] P.P. Vaidyanathan, and B.J. Yoon, "Gene and exon prediction using all pass-based filters," in Workshop on Genomic Signal Processing and Statistics, Raleigb, NC, USA, October 2002.

[6] M.K. Hota, V.K. Srivastava, "DSP technique for gene and exon prediction taking complex indicator sequence," Proc. IEEE TENCON 2008, pp. 1-6, 2008.

[7] P. Ramachandran, Wu-Sheng Lu, and Andreas Antoniou, "Location of exons in DNA sequences using digital filters," in Proc. IEEE International Symposium on Circuits and Systems, Taipei, May 2009.

[8] M..K. Hota, V.K. Srivastava, "Identification of protien coding regions using antinotch filters," Digital Signal Processing, Vol. 22, No. 6, pp. 869-877, 2012.

[9] Jayakishan Meher, Pramod K. Meher, Gananath Das, "Improved comb filter based approach for effective prediction of protein coding regions in DNA sequences," Journal of Signal and Information Processing, Vol. 2, pp. 88-99, May 2011.

[10] S. Rogic, A.K. Mackworth, and F.B.F. Ouellette, "Evaluation of gene finding programs on mammalian sequences," Genome Research, Vol. 11, No. 5, pp. 817-832, May 2001.