INTERNATIONAL JOURNAL OF BIOLOGY AND BIOMEDICAL ENGINEERING
DOI: 10.46300/91011.2021.15.7 Volume 15, 2021

A Novel Approach For Identification Of Exon Locations In DNA Sequences Using GLC Window

P.Kamala Kumari[1], J.B. Seventline[2],
[1]Muffakham Jah College of Engineering and Technology,
Hyderabad,500049, India.
[2]GITAM University, Visakhapatnam, 530045, India.

Received: October 4, 2020. Revised: February 18, 2021. Accepted: March 15, 2021. Published: March 30, 2021.

Abstract— **The application of signal processing techniques for identification of exons in Deoxyribonucleic acid (DNA) sequence is a challenging task. The objective of this paper is to introduce a combinational window approach for locating exons in DNA sequence. In contrast to the traditional single window function for evaluation of short time Fourier transform (STFT), this work proposes a novel method for evaluating STFT coefficients using a combinational window function comprising of Gaussian, Lanczos and Chebyshev (GLC) windows. The chosen combinational window GLC has the highest relative side lobe attenuation values compared to other window functions introduced by various researchers. The proposed algorithm incorporates GLC window function for evaluating STFT coefficients and in the design of FIR bandpass filter. Simulation results revealed its effectiveness in improving the evaluation parameters like Sensitivity, Specificity, Accuracy, Area under curve (AUC), Discrimination Measure (DM). Furthermore, the proposed algorithm has been applied successfully to some universal benchmark datasets like C. elegans, Homosapiens, etc., The proposed method has shown to be an efficient approach for the prediction of protein coding regions compared to other existing methods. All the simulations are done using the MATLAB 2016a.**

Keywords—DNA sequences, Dolph-Chebyshev window .Exons, Gaussian window, Lanczos window.

I. INTRODUCTION

DNA is the genetic material responsible for the growth and genetic transfer of individuals. This genetic information is stored in the form of a particular order comprising of the four nucleotides namely Cytosine(C) Adenine(A), Thiamine(T) and Guanine(G) . The entire DNA sequences are segregated into genic and intergenic region. Genic region stores information for making proteins. The genes are further subdivided into two regions: exons (coding regions) and introns (noncoding regions) as depicted in Figure.1

Proteins are considered as the essential component of every cell in the body. Proteins are the most profuse kind of molecules present in the body, next to water. They are made up of hundreds of compact units called amino acids that are linked together by peptide bonds to form a long chain. There are 20 different amino acids present in the body. Each amino acid is encoded as a sequence of three successive nucleotides in protein coding regions. Human genome constitutes approximately 3 billion basepairs Out of which only 2% are associated with protein coding regions while the rest 98% are probably junk DNA which is associated with either intergenic or introns. Locating such a low-density coding sequences makes it an arduous task.

A wide variety of Digital signal processing (DSP) tools and algorithms have emerged to solve the problem of identification of protein coding regions[1,2,3]. "Researchers revealed that exonic regions have strong power spectrum density (PSD) peak at f=1/3, which is absent in introns"[4,5]. This three-base property (TBP) is used extensively by the researchers for exon prediction in DNA sequences in conjunction with signal processing techniques. Prior to applying DSP algorithms, the DNA sequence has to be converted in discrete sequence by using mapping techniques. "The two major objectives of this genomics is 1) To discover families of genes or gene products that can be used to classify disease, thereby leading to molecular-based diagnosis and prognosis. 2) To characterize genomic regulation thereby leading to a functional understanding of disease and the development of system based medical solutions"[6].

In recent decade, quite a good number of mapping schemes have been introduced to convert alphabetical DNA sequence into numerical values. These include Voss, the Electron Ion Interaction Potential (EIIP), the Pseudo EIIP, the tetrahedron, the Paired numeric, the complex, the trigonometric, the integer, the variable mapping using twiddle factor, the real and the quaternion.

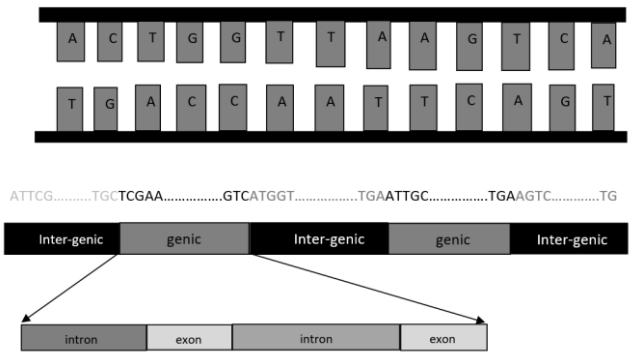


Fig.1 : Structure of DNA

The main DSP approaches comprises of the Discrete Fourier transform(DFT), spectral analysis using parametric models , entropy measures and digital filters. Digital filters and Fourier transform are widely used for genomic sequence analysis. Vaidyanathan and Yoon[7] "proposed a method which deploys an anti-notch digital filter to find the signal energy at the $2\pi/3$ frequency". Parametric methods such as autoregressive(AR) modelling of DNA sequences were addressed by Chakravarthy et al[8]. Choong and Yan[9] further proposed multi scale parametric spectral analysis for exon detection based on AR model. This method is proven to be better than the DFT and previous AR based models. Sajid and Stefan[10] introduced wide range wavelet window (WRWW) which is dominant in analyzing both short and long coding regions. Lopamudra et al[11] proposed an integrated approach using recursive Gauss Newton adaptive Kaiser window. The parameter β of the Kaiser window was made adaptive using error correlation by gauss newton tuning in conjunction with trigonometric mapping scheme. Saikat and Soma [12] proposed polyphase filtering with variable mapping. This DSP approach had multistage filtering technique which includes IIR anti notch filter, low pass FIR filter with Blackman window and polyphase filter structure to suppress the undesired noise. Hota and Srivastava[13] deployed Windowed DFT on two existing mapping schemes EIIP and Complex indicator and they stated that complex indicator sequence outperformed in exon prediction accuracy.

It is observed in frequency domain analysis, most researchers deployed anti notch filter to extract the period -3 peaks and digital filter framework with different window functions for denoising. Also, window functions are mostly used to evaluate Short Time Fourier transform (STFT) and then further to obtain PSD.

Table 1 shows the list of various window functions available in literature, stated by different authors for various applications apart from exon prediction. The Fourier transform is applied to a signal to obtain its frequency spectrum. For long sequences the signal is segmented into smaller signal by convolving with a noise suppression window. In frequency domain analysis, spectral leakage is defined as shifting of signal energy from main lobe to side lobes. This concept is witnessed in terms of the "width of the main lobe and its sharp peaks around the side lobes which results in energy leakage from the sharp energy levels to the lower levels"[14]. To counteract this issue, combination of windows with inclusion of an adjustable parameter can be framed to adjust the spectral characteristics. It is observed that prediction accuracy equally depends on both DNA mapping as well as the DSP approach. Mapping technique that works well for a DSP approach may not give same accuracy levels to another DSP approach. Therefore, DSP approaches and mapping techniques are trail and error methods. It is also observed from Table 1 that for exon prediction a combinational window was never been deployed. In all the previous literature only the traditional single window functions are deployed in their design.

This paper is organized as follows: the proposed algorithm is explained in Section 2, the parameters that has been used for evaluation are presented in Section 3, the results of simulation and discussions are presented in Section 4 and, finally section 5 presents the conclusion and future scope

TABLE I : WINDOW FUNCTION PROPOSED IN THE LITERATURE FOR VARIOUS APPLICATIONS

Authors	Window	Application
Kamala and Seventline[15]	Gaussian +Lanczos + Chebyshev	FIR band pass filter design
Lopamudra Das et al[11]	RGNAK	Exon Prediction
Sajid and Stefan[10]	WRWW	Exon Prediction
Tapash[16]	Blackman+Lanczos	Noise reduction in ECG signal
Vivek Kumar[17]	Gaussian + Hann	FIR low pass filter
Chavan et al[18]	Kaiser window	ECG Processing
Mena Chalco et al[19]	Gaussian	Exon prediction
Sahu and Panda[20]	Rectangular window	Exon prediction
Abbasi et al[21]	Hamming	Exon prediction
Nair and Sreenadhan[22]	Kaiser	Exon prediction
Shakya[23]	Bartlett	Exon Prediction
Mitun Shil[24]	Tan+Cosine	FIR low pass filter

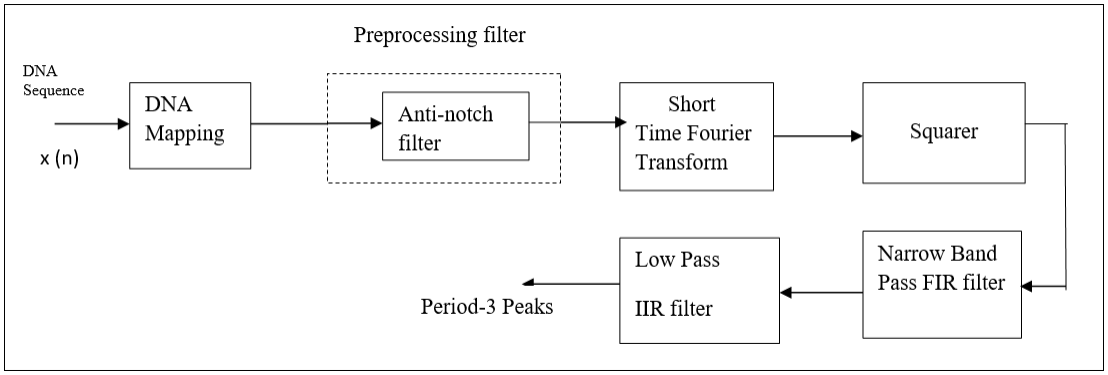


Fig.2 DSP Framework for identification of Exons

II. Materials and Methods

In order to improve the accuracy in the prediction of exons, the DSP approach shown in the Fig.2 have been framed. The alphabetic DNA sequence is first converted into a numerical sequence using mapping schemes. Furthermore, an Anti-Notch filter is followed by applying STFT, taking PSD, passing through Narrow bandpass FIR filter and Low pass IIR filter. Finally, the result is graphically depicted in the form of PSD plot having peaks. These peaks describe the unique feature of exon regions which is due to their three-base periodicity (TBP). Using the proposed DSP framework and the existing window functions, four DNA mapping schemes are chosen and tested. Based on the results, one of the mapping techniques is proposed so that detailed analysis can be performed at nucleotide level using various datasets form the Gene Bank.

A. DNA Mapping

It is a prerequisite to convert the alphabetical DNA sequence into numerical sequence using a mapping scheme before applying DSP algorithms [25]. One of the earliest and popularly used mapping technique is the Voss mapping [26]. It maps the four protein coding bases A,C,G, and T into four binary indicator sequences. In integer mapping [5] the four nitrogenous bases were mapped to the four integers values as follows A=2,T=0, G=3,C=1. This method incorporates DNA structural properties, such as purines (A,G)>pyrimidines(C,T) that introduces bias in the DNA sequence analysis. Electron Ion Interaction Potential(EIIP)[21] is defined as the average energy of delocalized electrons of the nucleotide.

The EIIP indicator sequence values are defined as T=0.1335, A=0.1260, C=0.1340 and G=0.0806. In Pseudo EIIP Mapping [27] the optimized characteristics values of nucleotides are obtained using the Quazi-Newton algorithm. The numerical values are normalized in such a way that their sum is always equal to the sum of EIIP values. The Pseudo-EIIP values are defined as A=0.1994, T=0.1033, G= 0.0123, C=0.0682. In Trigonometric mapping[11] the DNA nucleotides are given the values as follows

$$A = \cos(\Theta)+j\sin(\Theta) \quad (1)$$

$$C = -\cos(\Theta)-j\sin(\Theta) \quad (2)$$

$$G = -\cos(\Theta)+j\sin(\Theta) \quad (3)$$

$$T = \cos(\Theta)-j\sin(\Theta) \quad (4)$$

Purines A-G and Pyrimidines T-C have the same imaginary but opposite real parts. The value of Θ is chosen to be π/3 in this approach. These mapping schemes are tabulated in Table-II

B.Anti-Notch filter.

Anti-notch filters are widely used in Digital filter methods for identification exons. These methods are considered as faster methods than DFT based methods. In these methods the narrow bandpass filter has a center frequency of 2π/3. The narrow band pass filter is regarded as Anti-notch filter.[13] To detect the TBP of genomic sequence an IIR digital filter with a magnitude response showing a sharp peak at Θ =2π/3 is employed. IIR anti-notch filter has high gain in the pass band region. When input is passed through this filter, it gives an output with sharp gain at the frequency 2π/3. The transfer function of an IIR anti-notch filter can be obtained from a second order all pass filter whose transfer function specified as Eq.5

$$A(z) = \frac{R^2-2R\cos\Theta\, z^{-1}+z^{-2}}{1-2R\cos\Theta z^{-1}+R^2z^{-2}} \quad (5)$$

Numerator is mirror image of denominator with poles at $Re^{\pm j\theta}$ and zeros at $1/Re^{\pm j\theta}$. Anti-notch filter is obtained as

$$H(z) = \frac{1-A(z)}{2} \quad (6)$$

$$= \frac{1}{2}\left[\frac{1+R^2z^{-2}-z^{-2}-R^2}{1-2R\cos\Theta z^{-1}+R^2z^{-2}}\right] \quad (7)$$

$$= \frac{1-R^2}{2}\left[\frac{1-z^{-2}}{1-2R\cos\,\Theta z^{-1}+R^2z^{-2}}\right] \quad (8)$$

With R=0.993 and Θ= 2π/3, the above equation reduces to Eq. 9

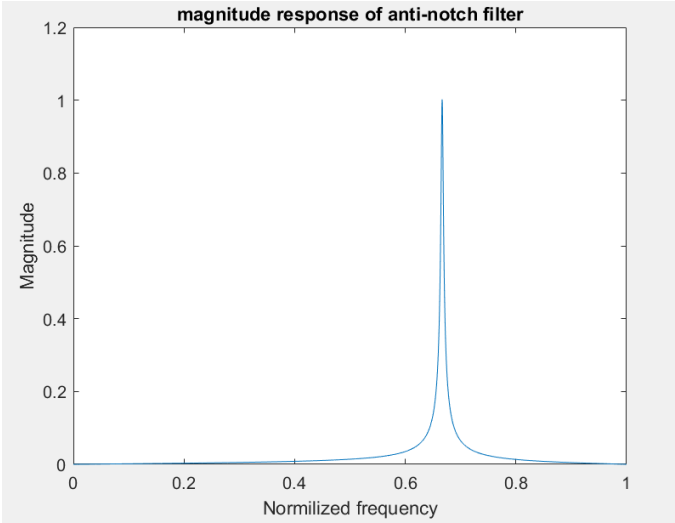
$$H(z) = \frac{0.007\,z^2-0.007}{z^2+0.992z+0.9841} \quad (9)$$

The magnitude response and pole/zero are plotted as shown in Fig.3

TABLE II :NUMERICAL MAPPING TECHNIQUES FOR THE FOUR NUCLEOTIDES

Method	Representation			
	A	***T***	***C***	***G***
Integer	2	0	1	3
Pseudo EIIP	0.1994,	0.1033	0.0682	0.0123,
EIIP	0.1260	0.1335	0.1340	0.0806
Trigonometric	0.5+j0.866	-0.5-j0.866	-0.5+j0.866	0.5-j0.866

In the pole zero plot of Fig.3 it reveals that there are two zeros at ±1. There are two poles located at an angle Θ= ±2π/3 and at a radius of R=0.993. The value of R chosen to be 0.993 which is less than 1 for stability. In the magnitude response, the response exhibits a peak value at the normalized frequency value of 0.666.



(a)

(b)

Fig.3 Magnitude response and Pole/zero plot of Anti-notch filter

C. Window Functions.

In the proposed algorithm, window functions are used in evaluating STFT and design of narrow Bandpass FIR filter. The performance window function extremely affects the prediction accuracy of exons regions. In evaluation of STDFT, gradual truncation of DNA sequence leads to unwanted peaks in the DFT spectrum. Therefore, instead of abrupt truncation, signals are segmented using a window function of specific length. The window is slid by one or more base pair positions along the sequence, and the processing is repeated on that window until the end of the sequence [28]. In the literature, many authors applied different windows and analyzed their performance in the prediction methods. The desired specifications of a window function are main lobe width and ripple ratio. But these specifications are contrary. A window with a thinner main lobe has a poor side lobe rejection and vice versa. The following window functions are incorporated in the design flow.

- *Blackman window [29]*:

$$w_b(n) = \begin{cases} 0.42 - 0.5\cos\left(\frac{2\pi n}{M-1}\right) + 0.08\cos\left(\frac{4\pi n}{M-1}\right), 0 \le n \le M-1 \\ 0, \qquad Otherwise \end{cases} \tag{10}$$

- *Hamming window [29]*

$$w_{hm}(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{M-1}), 0 \le n \le M-1 \\ 0 \qquad , Otherwise \end{cases} \tag{11}$$

- *Hanning window [28]*

$$w_{hn}(n) = \begin{cases} 0.5 * \left[1 - cos\left(\frac{2\pi n}{M-1}\right)\right], 0 \le n \le M-1 \\ 0, \quad Otherwise \end{cases} \tag{12}$$

- *Kaiser window*

The coefficient of Kaiser window [30] with length L=N+1 is computed as

$$w_k(n) = \frac{I_o\left(\beta\sqrt{1-\left(\frac{n-\frac{N}{2}}{\frac{N}{2}}\right)^2}\right)}{I_o(\beta)}, 0 \le n \le N \tag{13}$$

Where I_o is the zeroth order Bessel function.

- *Lanczos and Blackman window(LB):*

The adjustable window function introduced in [15] is a combination of Lanczos and Blackman window with the controlling parameter, p is given as

$$w_{lb}(n) = [w_l(n)w_b(n)(\frac{1}{n}w_l(n) - 1)]^p \quad (14)$$

where , $$w_l(n) = sinc^M(\frac{2n}{N} - 1) \quad (15)$$

The value of M is fixed at 2.

- *Gaussian and Hanning window(GH):*

The window function introduced in [16] is the product of Gaussian and Hanning window given as

$$w_{gh}(n) = [\, w_g(n) * w_{hn}(n)] \quad (16)$$

where, $w_g(n)$ is Gaussian window for α =2.3 is given by

$$w_g(n) = e^{\frac{1}{2}(\alpha\frac{n}{N/2})^2} \quad (17)$$

- *GLC window*

This a combination of Gaussian, Lanczos and Dolph Chebyshev window functions [14]

$$w_{glc}(n) = [\, w_g(n) * w_l(n) * w_c(n)]^r \quad (18)$$

Where $w_c(n)$ is zero phase Dolph-Chebyshev window which is obtained from the optimal Dolph Chebyshev window transform given by

$$W_0(k) = \frac{\cos(Ncos^{-1}[\beta\cos(\frac{\pi k}{N})])}{\cosh[Ncosh^{-1}(\beta)]} \quad (19)$$

where $$\beta = \cosh[\frac{1}{N}cosh^{-1}(10^{\alpha})] \quad (20)$$

$w_c(n) = IDFT(W_o(k))$. The parameter α controls the side lobe attenuation.

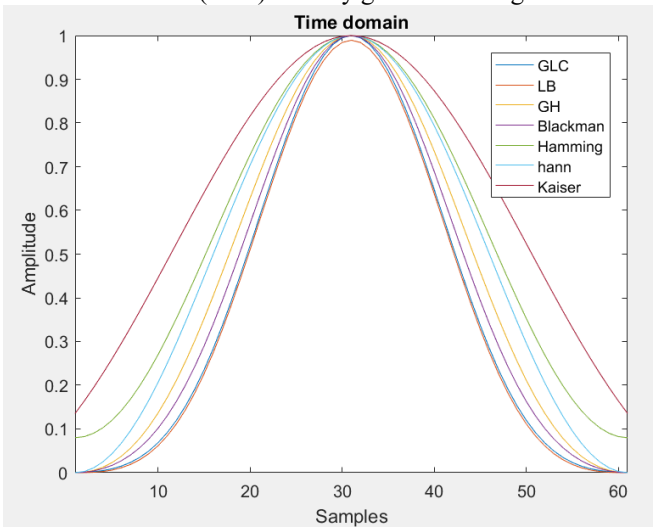
The comparison of these windows based on their spectral parameters are shown in the Table-III and Fig.4. From the Fig.4 it can observed that GLC window is having wider main lobe width as well as the lower sidelobes. These features of GLC window gives better suppression of background noise. From frequency domain characteristics illustrated in Fig.4 it is evident that the relative side lobe width is much better (-101.3dB) compared to other window functions The performance of different windows is dealt in many literatures [15,20,31]

TABLE III : COMPARISON OF SPECTRAL PARAMETERS OF WINDOWS FOR N=60

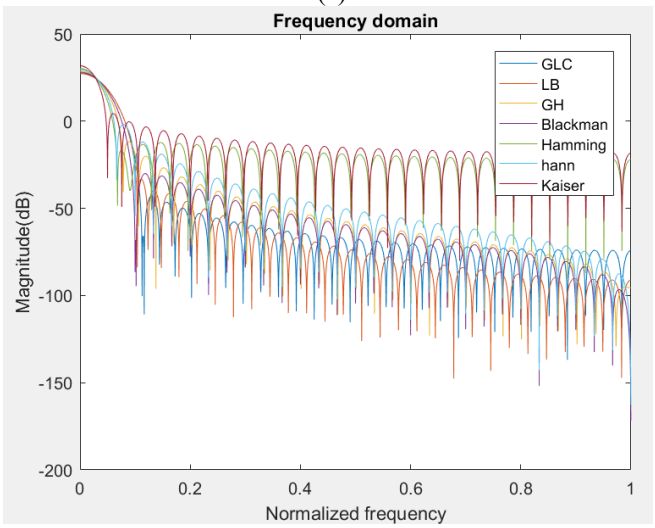
Window	Main lobe width	Relative side lobe attenuation(dB)
Kaiser,β=3.5	0.035	-27.4
Hanning	0.046	-31.5
Hamming	0.042	-42.4
Blackman	0.054	-58.2
GH	0.05	-40.4
LB	0.058	-60.5
GLC,r=0.6	0.066	-101.3

D.Short Time Fourier Transform(STFT).

Fourier Transform is a tool that transforms a time domain signal into frequency domain and is used to visualize the spectral components in it. Fourier transform based methods are widely used for identification of exon regions. DNA strand being very long sequences, evaluation of Discrete Fourier transform (DFT) directly give the average



(a)



(b)

Fig. 4 Time domain and Frequency domain illustrations of different window functions.

frequency content over the entire signal time interval. Therefore, Short time Fourier Transform (STFT) is deployed instead of DFT. STFT is one of the earliest and traditional methods used for identification of protein coding regions based on period-3 property. It necessitates taking DFT of a window of defined size in a DNA sequence, which is then made to slide across the whole length of the sequence.[32,33] "The STFT is regarded as digital filter followed by decimator which depends on the separation between adjacent positions of the window"[34].

STFT is chain of Fourier transforms of a windowed sequence. It is a method that yields localized spectrum in time domain by applying Fourier transform in a localized time window. DFT of a sequence x(n) of the length L is given by

$$X(k) = \sum_{n=0}^{L-1} x(n)e^{-j\frac{2\pi}{L}kn} \qquad k = 0,1,2, \dots \dots L-1 \quad (21)$$

For better localized spectrum , the entire sequence is partitioned into N samples by sliding a window by one entry in the sequence

$$X_s(k) = \sum_{n=0}^{N-1} w(n)x(n)\, e^{-j\frac{2\pi}{N}kn} \quad k = 0,1,2 \dots \dots \dots N-1 \quad (22)$$

w(n) is a window function of length N=351.

Window length is one important factor that effects the performance of STFT in suppressing the background noise present in the Fourier spectrum. Larger window size may miss small coding regions and smaller size window may introduce statistical fluctuations. Therefore, empirically the window size is chosen to be 351 as most of the researchers suggested[35,36,37]

The power spectral density of the discrete sequence is

$$P_s(k) = \sum_{n=0}^{N-1} |X_s(k)|^2 \quad (23)$$

When N is a multiple of 3, the value of DFT at k=N/3 attains the maximum peak value witnessing the period -3 property of a DNA sequence. The period -3 property is associated with the different statistical distribution of codons between exons and introns in genomic sequence. This property is used to locate exons regions in a DNA sequence.

If the DNA section is an exon region , the STFT coefficient $X_s(N/3)$ is remarkably larger than the adjacent STFT coefficients consequently P(N/3) is large in a exon region.

E.Filtering for exon prediction

The power spectral density plot of DNA sequence is much noisy and is hard to identity the exact location of exons. Therefore, to locate the exact location of protein coding regions in this presence of noise it is required to filter the spectrum. A digital filter is a specific type of discrete system able to realize some transformations to discrete numerical sequence input. Digital filters mostly appear in two shapes: FIR and IIR. Both of them can proceed a filter that passes or rejects frequencies bands, but the mathematical implementations differ totally. Two stages of filtering are deployed. The PSD of the protein coding sequence $P_s(k)$ is given as input to Narrow Bandpass FIR filter. It enhances the peaks of the PSD plot. Narrow Bandpass FIR filter is designed with various window functions stated earlier having specifications: order=100, passband frequency = [0.6666 0.6667].

The output of the Narrow Bandpass FIR filter is given as the input to IIR low pass filter. The low pass filter reduces 1/f noise and smoothens the PSD plots. In frequency analysis of a signal, the spectral leakage is due to diffusion of coding and noncoding regions which appears as 1/f noise in DNA sequence spectrum. Butterworth approximation is incorporated in the design of IIR filter with the following specifications: order=10 and normalized cutoff frequency (ω_c) =0.38 rad/sec are chosen by trail and error method. The filtered PSD plotted against the nucleotide locations will reveal peaks only in exon regions and no such peaks will be visible in intron regions.

TABLE IV : DATASETS DEPLOYED FOR ANALYSIS

Organism name	Data length(base-pairs)	Accession No.	Coding sequence(CDS)
Homo sapiens tubulin	1-1770	BC111374	28..1371
Caenorhabditis.elegans	7020-15,021	F56F11.4	7949..8059, 9548... 9877, 11134...11397, 12485...12664,14275..14625
HMR195 dataset	1-3036	AF058762	115..482, 1867..2662
Pseudanabaena Bacteria	1-731	A00447.1	146..364

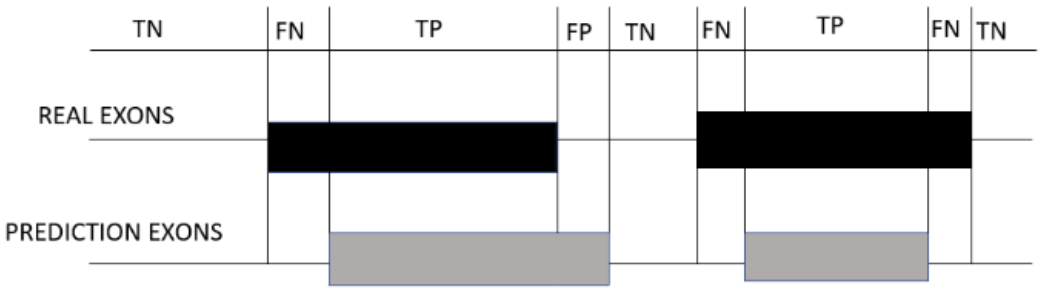


Fig. 5 Illustration of Nucleotide level measures for prediction accuracy

III. Evaluation Parameters

To evaluate the prognostic accuracy of the proposed algorithm in identifying coding regions, the DNA sequences are fetched from GeneBank of National Center of Biotechnology Information (NCBI) https://www.ncbi.nlm.gov/ in FASTA format. The datasets used for testing purpose are Tabulated in Table-IV

The actual locations are named as CDS for coding regions in the Genebank documentation. From the simulated results the exons locations are identified and are compared with that of the CDS for exons. Identification of exons results are compared at exon level. Performance at nucleotide level is measured in terms of the True positive(TP) , True Negative(TN), False positive(FP) and False Negative (FN) as shown in Fig. 5. From these terms the following parameters are evaluated. These measurements are commonly used for identification of protein coding regions.[38-40]

- *Specificity (S_p):*

It measures the proportion of intron regions correctly recognized as intron regions. Conventionally S_p is specified by Eq.24

$$S_p = \frac{TN}{TN + FP} \quad (24)$$

- *Sensitivity (S_n)*

It measures the proportion of exon regions correctly recognized as exon regions. Conventionally S_n is specified by Eq.25

$$S_n = \frac{TP}{TP + FP} \quad (25)$$

- *Accuracy (AC)* It is given by Eq.26

$$Accuracy = \frac{S_n + S_p}{2} \quad (26)$$

When AC value is nearer to 1 implies that the identification methodology is more accurate.

- *Geometric Mean (gm)*

It is specified by Eq.27

$$gm = \sqrt{S_n \times S_p} \quad (27)$$

Geometric mean attains high value when Sensitivity and Specificity have close and higher value.

- *Discrimination measure(DM)*

$$DM = \frac{Lowest\ peak\ amplitude\ in\ Exon\ region}{Highest\ peak\ amplitude\ in\ Intron\ region} \quad (28)$$

When DM> 1 implies that all the coding regions are well distinct and for DM<1 at least one coding region is not having an ample amount of signal power to be discriminate from noncoding region.

- *Signal to Noise Ratio (SNR)*

It generally depicts the ratio of signal strength to noise strength. Coding regions have high value and noncoding regions have low value of SNR. More the SNR better is the performance [11].

- *Receiver Operating Characteristic (ROC) curve*

ROC [21]is also used as a measure to evaluate the performance accuracy. ROC is characterized by the single number using area under the ROC curve (AUC). AUC is calculated from ROC plot. ROC curves are plotted TPR against TFR for different values of threshold. Larger area indicates more accurate prediction. "The closer the ROC curve to 1 more is the area under the curve and better is the technique and less is the AUC and poorer is the method" [11].

IV. Results And Discussion

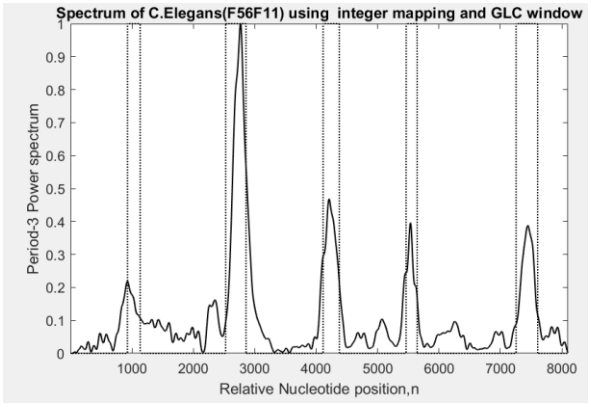
The proposed algorithm is analyzed into the following three cases.

A. Case-I (Comparative analysis of different mapping techniques)

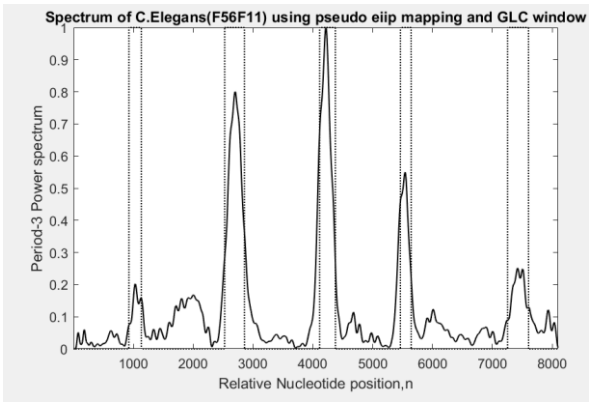
In the literature there are more than 15 mapping techniques available. Out of them considering most frequently and latest mapping techniques , the four techniques mentioned in Table II are chosen. In the proposed flow, GLC window function is incorporated in evaluating STFT and design of narrow bandpass filter. The window length is chosen to be 351. All the four mapping techniques are tested on three gene sets with the Accession numbers F56F11.4, BC111374, A00447.1 from NCBI homepage. The PSD plots for all the mapping are shown in Fig. 6. Table V summarizes the performance of the said mappings on the proposed algorithm. It can be observed from Fig.6 and Table V that protein coding regions are quite noisy in Trigonometric mapping. Fig. 9,10,11 shows the ROC curves depicting FPR Vs TPR for comparison of mapping schemes. In the analysis of Homosapiens BC111374 the value of AUC is more for Pseudo EIIP mapping. In the analysis of C,elegans F56F11.4 the value of accuracy (AC) and geometric mean (gm) are larger for EIIP mapping but AUC is less. The accuracy (AC) , geometric mean(gm) and Area under curve(AUC) values are calculated to be highest in the integer mapping compared to rest. The evaluation parameters are calculated with a threshold values of 0.15, 0.16 and 0.035 for C.elegans F56F11.4, Pseudanabaema A00447.1 and Homosapiens BC111374.1 respectively. This reveals that integer mapping with GLC window incorporated design flow yields better results and best suits for the proposed algorithm.

B. Case II(Comparative analysis of different window techniques).

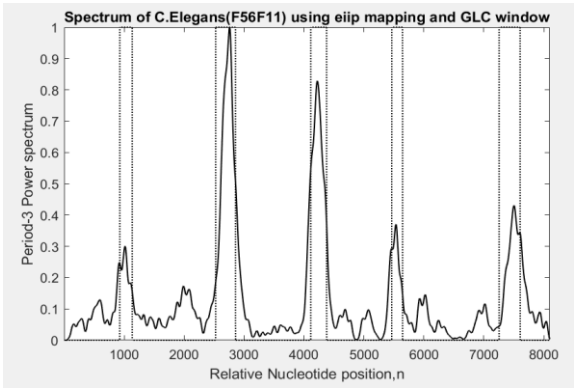
In the proposed algorithm window functions are used for evaluation of STFT and in the design of Narrow band FIR filter. Window functions greatly influence the prediction of exons in DNA sequences. The window functions used for comparation are of two types. Blackman window, Hamming window, Hanning window have one fixed parameter which controls width of the main lobe. Kaiser window, Lanczos Blackman window, Gaussian Hann window and GLC window have two independent parameters which control main lobe width and relative side lobe attenuation. The proposed algorithm with Integer mapping and the stated window functions are tested on two DNA sequences. The two DNA sequences are universally used gene sequences C.elegans and a sequence from HMR195 data set. The PSD plots are shown in Fig.7 and Fig.8. The dotted lines in the PSD plots indicate the actual exon locations. The results are summarized in the Table VI. It can be observed from the PSD plots that the peaks related to exons are clearly


Spectrum of C.Elegans(F56F11) using integer mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

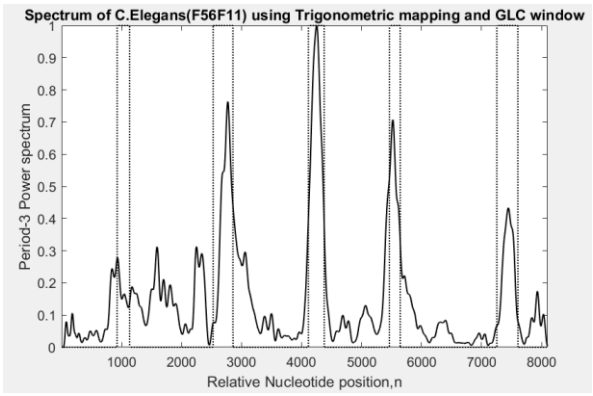
(a)


Spectrum of C.Elegans(F56F11) using pseudo eiip mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

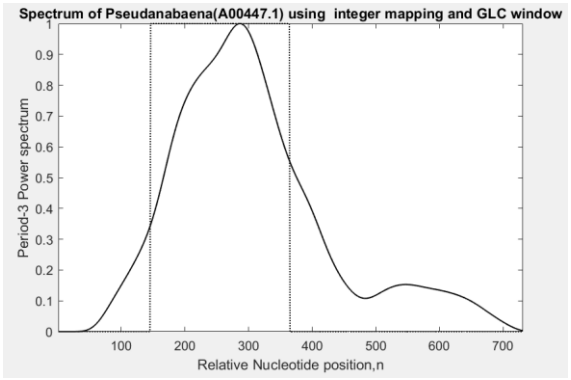
(b)


Spectrum of C.Elegans(F56F11) using eiip mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

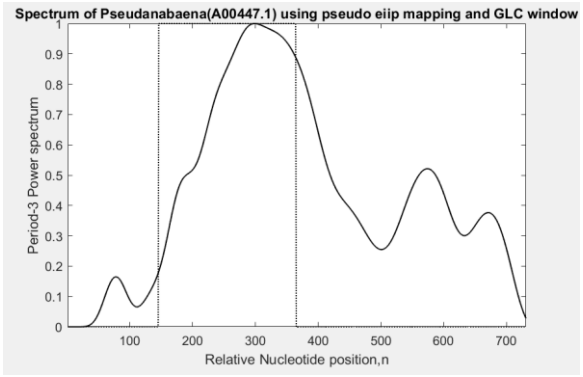
(c)


Spectrum of C.Elegans(F56F11) using Trigonometric mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

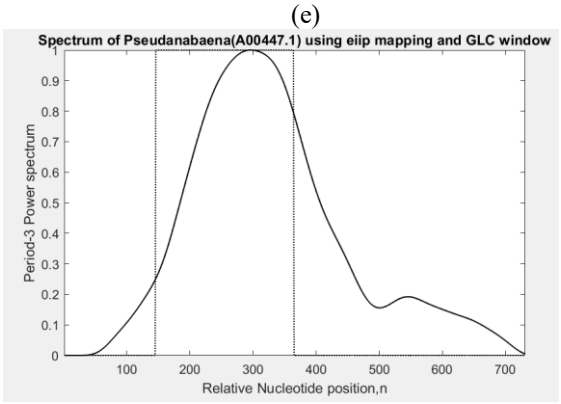
(d)


Spectrum of Pseudanabaena(A00447.1) using integer mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

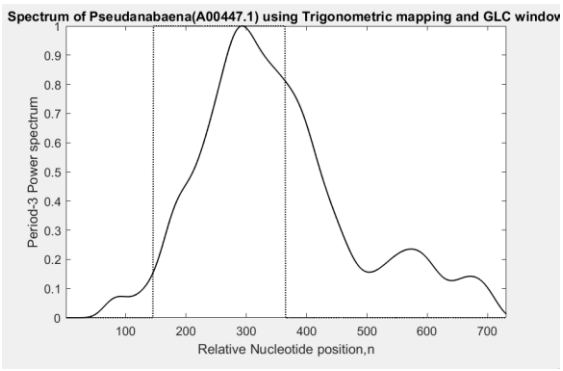
(e)


Spectrum of Pseudanabaena(A00447.1) using pseudo eiip mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

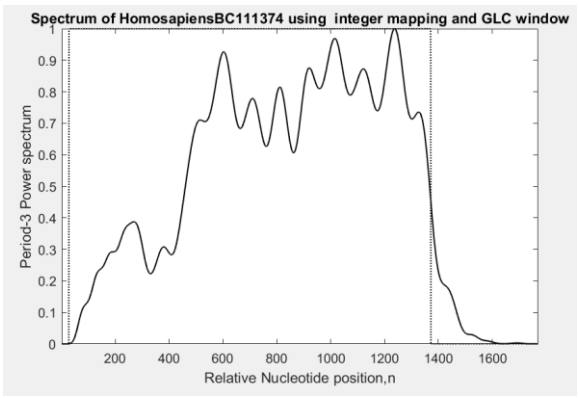
(f)


Spectrum of Pseudanabaena(A00447.1) using eiip mapping and GLC window
Period-3 Power spectrum
Relative Nucleotide position,n

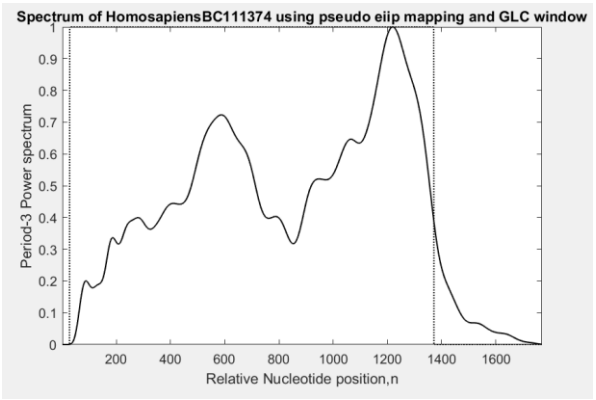
(g)


Spectrum of Pseudanabaena(A00447.1) using Trigonometric mapping and GLC windov
Period-3 Power spectrum
Relative Nucleotide position,n

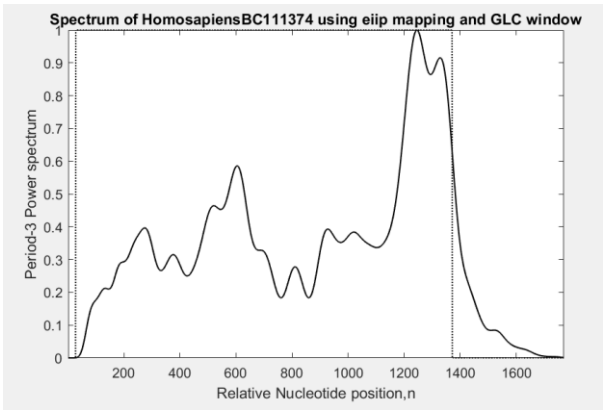
(h)



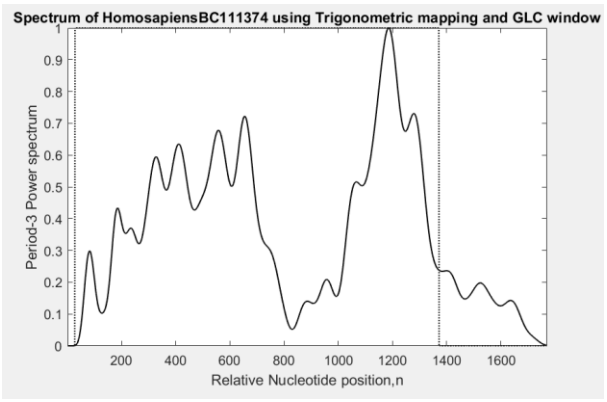
(i)



(j)



(k)



(l)

Fig. 6 PSD plots showing the effect of mapping techniques integer, Pseudo EIIP, EIIP and Trigonometric on genes C.elegans (F56F11.4) (a),(b),(c),(d) ,Pseudanabaema(A00447.1) (e),(f),(g),(h) and Homosapiens (BC111374) (i),(j),(k),(l) respectively.

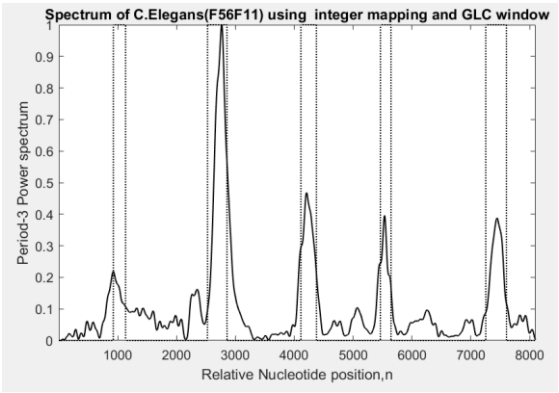
TABLE V: COMPARATIVE ANALYSIS OF DIFFERENT MAPPING SCHEMES(CASE I)

	gene set														
	C.elegansF56F11 Th=0.15					PseudanabaemaA00447.1 Th=0.16					HomosapiesBC111374.1 Th=0.035				
Mapping	Sp	Sn	AC	gm	AUC	Sp	Sn	AC	gm	AUC	Sp	Sn	AC	gm	AUC
Trigonometric	0.7811	0.8177	0.7994	0.7992	0.9102	0.541	0.9954	0.7682	0.7338	0.9145	0.2042	0.9866	0.5954	0.4488	0.8569
EIIP	0.9096	0.9415	0.9255	0.9254	0.9588	0.5449	1	0.7725	0.7382	0.956	0.4765	0.9807	0.7286	0.6836	0.8967
Pseudo EIIP	0.9158	0.8372	0.8765	0.8753	0.9609	0.2988	1	0.6494	0.5466	0.8872	0.3991	0.9836	0.6914	0.6265	0.958
Integer	0.9394	0.8522	0.8958	0.8947	0.9732	0.748	1	0.874	0.8649	0.9825	0.6995	0.9725	0.836	0.8248	0.9525

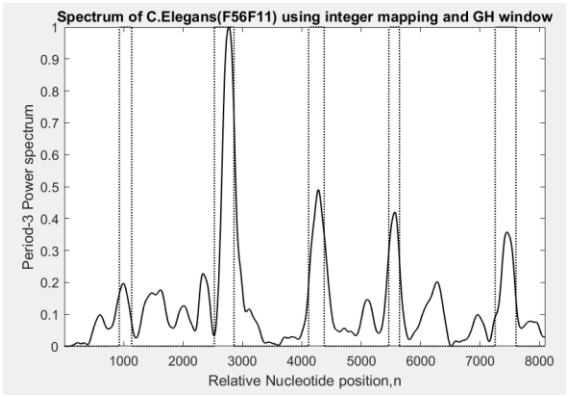
depicted in the Fig 7 and 8 which is the combination of integer mapping and GLC window. And also, from the Table VI it reveals that GLC window outperforms all the other windows in terms of the evaluation parameters AUC, DM and SNR. In case of C.elegans, with a threshold value of 0.15 and HMR195 dataset with a threshold value of 0.35 , the AUC value of 0.9732 and 0.8204 is very high compared to all , which indicates most of exons locations are predicted very closely to the actual exons as mentioned in the NCBI homepage. It also reveals that the SNR value is also significantly high i.e., 1.112 and 1.38 which indicates its noise sensitivity. Fig. 10, Fig 11 Fig12 and Fig. 13 shows the ROC curves depicting FPR Vs TPR for comparing various window functions. From the ROC curves it is evident that GLC window with integer mapping outperforms as the curves are close to 1.

C. Case III (Comparative analysis of proposed method with other prediction methods).

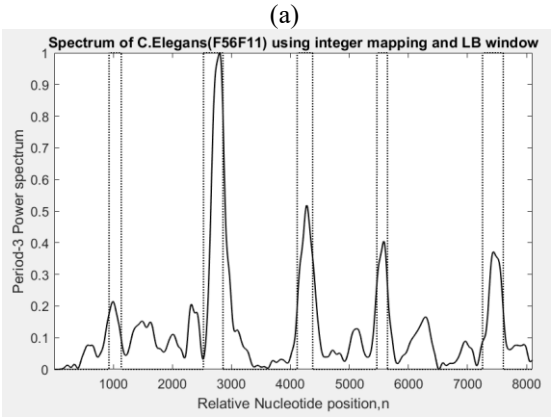
The proposed method is compared with various other exon prediction techniques like RGNAK[11], Geortzel algorithm[41], STFT[37],



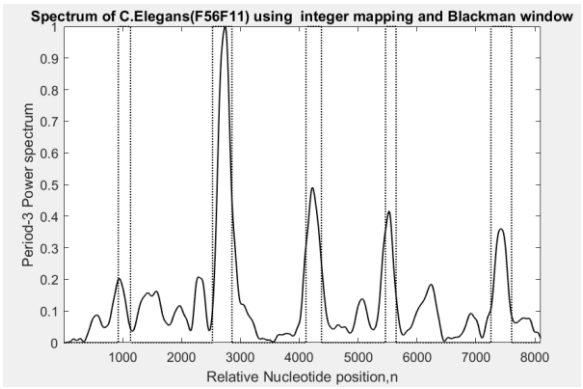
(a)



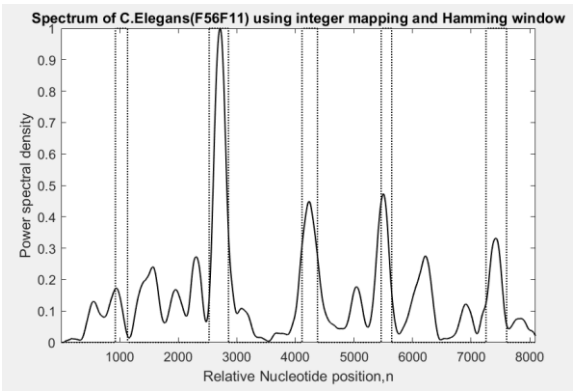
(b)



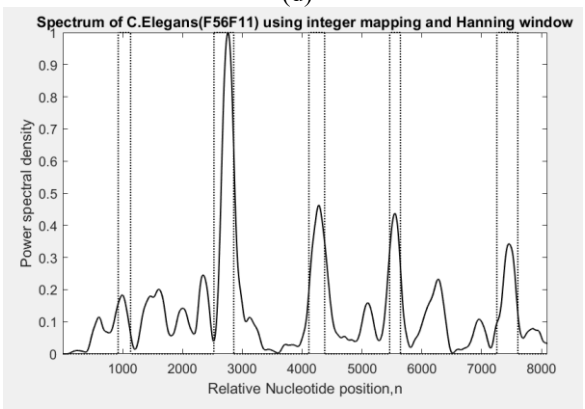
(c)



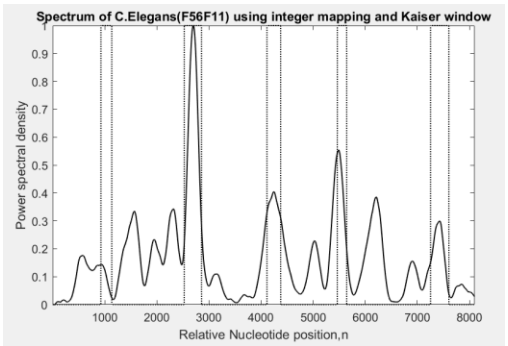
(d)



(e)

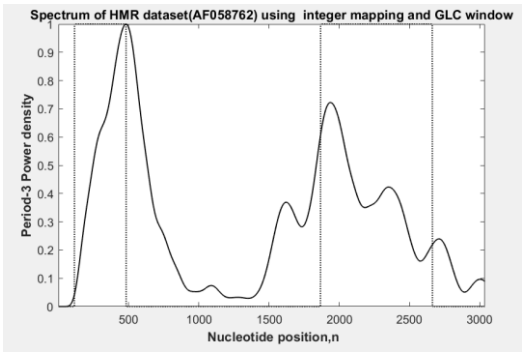


(f)

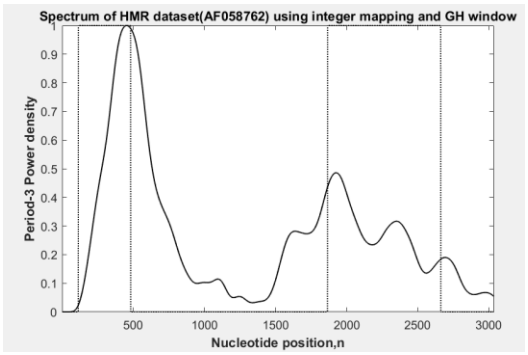


(g)

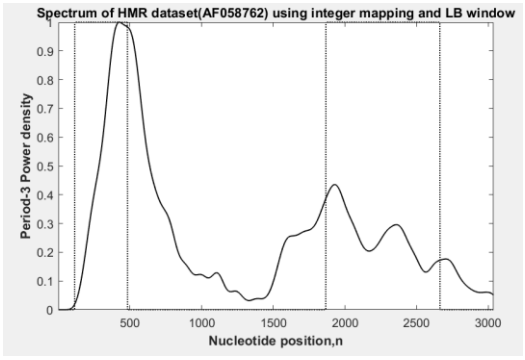
Fig . 7 PSD plots showing effect of window functions GLC, Gaussian Hann, Lanczos Blackman, Blackman,Hamming, Hannig and Kaiser on the proposed algorithm for the DNA sequence C.elegans(F56F11.4).

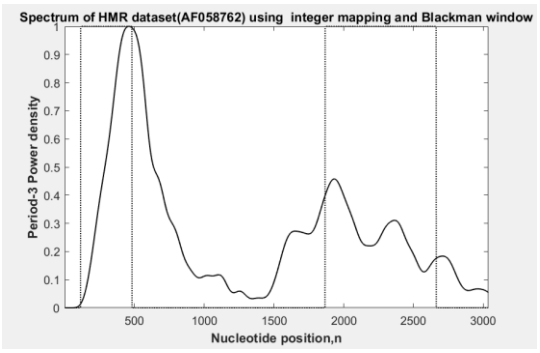


(a)

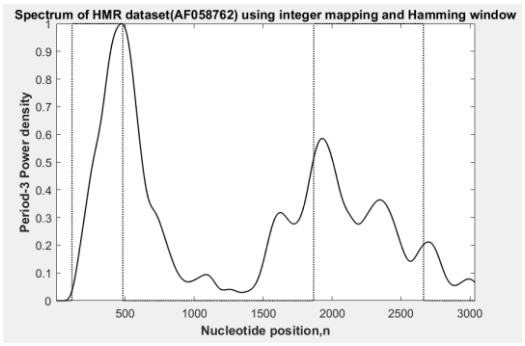


(b)

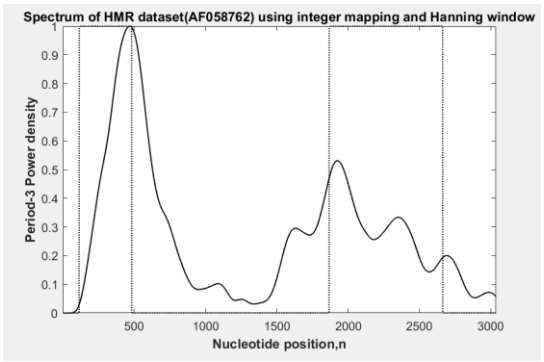




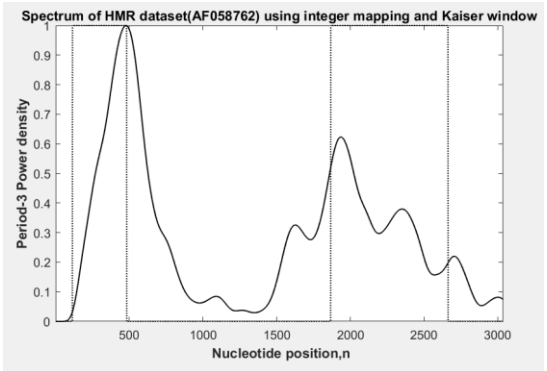
(d)



(e)



(f)



(g)

Fig . 8 PSD plots showing effect of window functions GLC, Gaussian Hann, Lanczos Blackman, Blackman,Hamming, Hannig and Kaiser on the proposed algorithm for the DNA sequence in HMR195 dataset.

TABLE VI: COMPARATIVE ANALYSIS OF DIFFERENT WINDOW FUNCTIONS ON PROPOSED ALGORITHM(CASE II)

window	Geneset F56F11.4 Th=0.15						HMR195 data set Th=0.35					
	Sp	Sn	AC	AUC	DM	SNR	Sp	Sn	AC	AUC	DM	SNR
Kaiser	0.6677	0.7914	0.7295	0.8395	0.4176	0.5718	0.8579	0.5541	0.706	0.8078	1.14	1.29
Hanning	0.8205	0.8357	0.8281	0.9165	0.7403	0.7874	0.8494	0.3849	0.6172	0.7736	1.13	1.21
Hamming	0.7758	0.8327	0.8042	0.8944	0.623	0.7331	0.8584	0.4863	0.6724	0.7995	1.14	1.27
Blackman	0.8708	0.8552	0.863	0.9421	0.9351	0.922	0.8526	0.3548	0.6037	0.7419	1.13	1.08
Gaussian+Hanning	0.8493	0.8492	0.8492	0.9208	0.8594	0.8395	0.8488	0.3686	0.6087	0.762	1.12	1.14
Lanczos+Blackman	0.8954	0.8954	0.8775	0.9339	1.04	0.8938	0.8483	0.3393	0.5938	0.7376	1.12	1.08
GLC(Proposed)	0.9394	0.8537	0.8965	0.9732	1.38	1.112	0.8168	0.7388	0.7778	0.8204	1.14	1.38

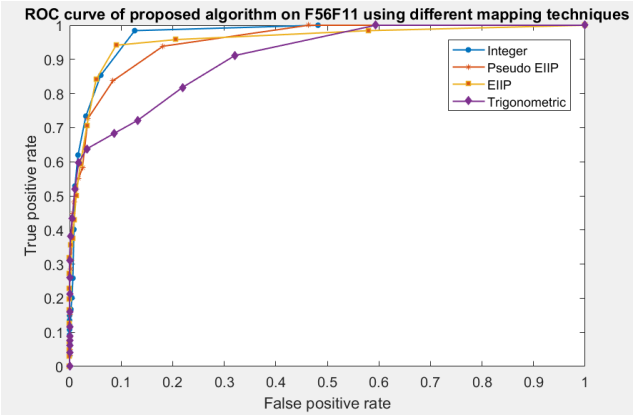


Fig. 9 The ROC curve showing the performance comparison of Mapping schemes on DNA sequence F56F11.4

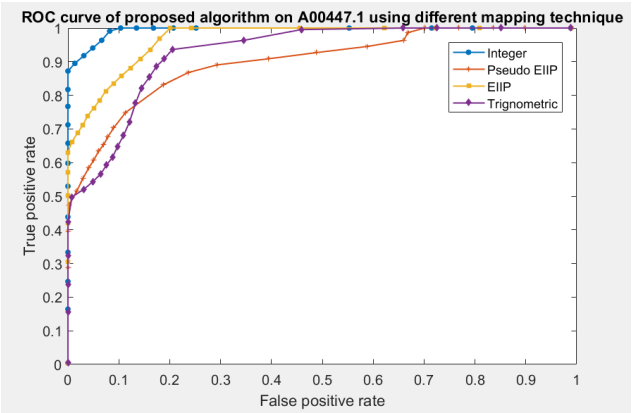


Fig.10 The ROC curve showing the performance comparison of Mapping schemes on DNA sequence A00447.1

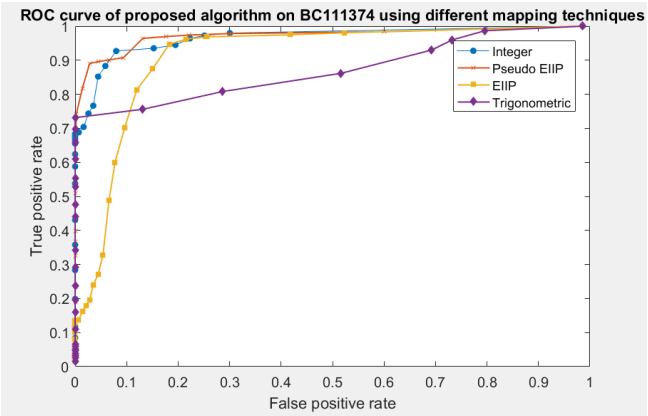


Fig.11 The ROC curve showing the performance comparison of Mapping schemes on DNA sequence BC111374.

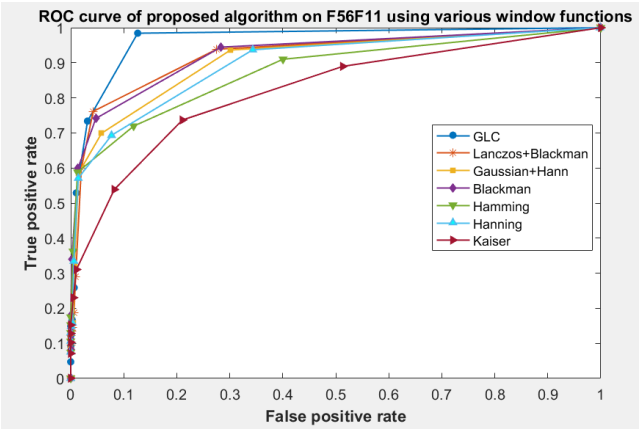


Fig. 12 The ROC curve showing the performance of window functions on DNA sequence F56F11.4

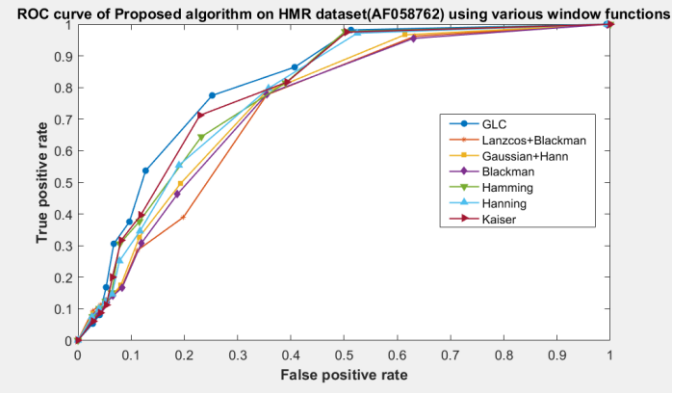


Fig. 13 The ROC curve showing the performance of window functions on DNA sequence from HMR195 dataset

Anit-notch filter method[13] which used the universally tested gene sequence C.elegans. Table VII reveals that the proposed methods outperforms the other existing methods in terms of the evaluation parameter Area Under Curve(AUC).

TABLE VII: COMPARATIVE ANALYSIS OF PROPOSED ALGORITHM WITH OTHER EXCISTING METHODS(CASE III)

Method	**Mapping**	**AUC**
Anti-notch filters [13]	Voss	0.8656
Goertzel algorithm [41]	EIIP	0.912
RGNAK [11]	Trigonometric	0.92
STFT Hamming Window [37]	Paired Numeric	0.9435
Proposed	**Integer**	**0.9732**

V. CONCLUSION AND FUTURE SCOPE

Prediction of exons is imperative for the proper diagnosis of any disease and hence needs to be done timely and accurately. In existing DSP techniques, specially the frequency domain methods, STFT as the key transform. All these methods stated that spectral leakage is the limitation of STFT. First integer mapping is applied to symbolic DNA sequence and applied to the DSP framework. This paper introduces a novel approach of using a combination of window function, i.e Gaussian, Lanczos and Chebyshev(GLC). In our proposed approach, GLC window is deployed for evaluating STFT with window length 351 and in the design of Narrow bandpass FIR filter. This window exhibits lowest sidelobes and wider main lobe width overcoming the limitations. This feature solely improved the evaluation parameter AUC ranging from 3% to 11% for the DNA sequences C.Elegans F56F11.4 and HMR195 dataset. Furthermore, the proposed approach also maximizes the number of nucleotides correctly predicted as coding regions. Table VI reveals that the proposed algorithm shows significant improvement in the evaluation parameters AUC, DM and SNR, that indicates the accuracy in identification of exons. Further the work may be extended in implementing GLC window incorporating adaptive methodology.

References

[1] Lan Zhan , Application of spectral anlayis to DNA sequences,CSD TR #06-003, January 2006

[2] D. Anastassiou, Frequency -domain Analysis of Biomolecular sequences, Bioinformatics 16,pp 1073-1081.

[3] D.Anastassiou, DSP in genomics: processing and frequency domain analysis of character strings, IEEE,0-7803-7041,2001

[4] P.P. Vaidyanathan, B.-J. Yoon, The role of signal-processing concepts in genomics and proteomics, J. Franklin Inst. 341 (2004) 111–135 (Special Issue on Genomics).

[5] P.D.Cristea, Genetic signal representation and analysis [c]. in Proc.SPIE Inter. Conf. on Biomedical Optics , 2002,4623:77-84.

[6] S. Chakraborty and V. Gupta, "DWT Based Cancer Identification Using EIIP," *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*, Ghaziabad, 2016, pp. 718-723, doi: 10.1109/CICT.2016.148.

[7] P.P. Vaidyanathan, B.-J. Yoon, Gene and exon prediction using allpass-based filters, in: Workshop on Genomic Signal Process. Stat., Raleigh, NC, 2002.

[8] Niranjan Chakravarthy, A. Spanias, L. D. Iasemidis, K. Tsakalis, Autoregressive Modeling and Feature Analysis of DNA Sequences, EURASIP Journal on Applied Signal Processing 2004:1, 13–28

[9] M.K.Choon, Hong 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 curve, Bioinformation,2008; 2(7): 273–278

[10] Sajid A. Marhon, Stefan.C.Kremer, Prediction of protein coding regions using a wide range wavelet window method, IEEE/ACM Trans. Comput. Biol. Bioinform. Vol. 13, No. 4, 2016

[11] L.Das,S.Nanda, J.K.Das, An integrated approach for identification of exon locations using recursive Gauss Newtion tuned adaptive Kaiser window, Genomics,https://doi.org/10.1016/j.ygeno.2018.10.008

[12] S.S.Roy, S.Burman , Polyphase filtering with variable mapping rule in protein coding region prediction, Microsyst Technol 2016 ,doi 10.1007/s00542-016-2884-5

[13] M.K Hota,V.K.Srivastava, Identification of protein coding regions using anti-notch filters,J.Digital signal processing 22(2012) 869-877

[14] M.Cerna, A.F.Harvey, The fundamentals of FFT-based signal analysis and measurements, Natiional instruments,Junho,2000.

[15] P.Kamala Kumari, J.B.Seventline, Improved spectral characteristics of bandpass filter using a novel adjustable window function, International journal of circuits,systems and signal processing, vol 13,2019.

[16] Tapash karmaker et al , A new adjustable window function to design FIR filter and its application to noise reduction from contaminated ECG signal, 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)

[17] Vivek Kumar et al , Design of Effective Window Function for FIR Filters, IEEE International Conference on Advances in Engineering &Technology Research 2014

[18] M.S.Chavan,R.A.Agarwala,M.D.Uplane,Use of Kaiser window for ECG processing, in: Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Robotics and Automation , 2006 February, Madrid, Spain 2006.

[19] J.Mena Chalco, H.Carrer,Y.Zana,R.M.Cesar Jr, Identification of protein coding regionsmusing using modified Gabor wavelet transform, IEEE/ACM Trans. Comput. Biol. Bioinform.,5(2)(2008) 198-207

[20] S.S.Sahu, G.Panda, Identification of protein coding regions in DNA sequences using time-frequency filtering approach , Genom,Proteom,Bioinform, 9(1)(2011) 45-55

[21] O.Abbasi,A.Rostami, G.Karimian, Identification of exonic regions in DNA sequencing using cross-correlation and noise suppression by discrete wavelet transform,BMC Bioinfom. 12(1)(2011) 1.

[22] A.S Nair,S.P Sreenadhan , A coding measure scheme employing electron ion interaction pseudopotential(EIIP), Bioinformation 1 (6) (2006)197-202

[23] D.K. Shakya, R.Saxena,S.N.Sharma , An adaptive window length strategy for eukaryotic CDS prediction,IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 10(5)(2013) 1241-1252.

[24] Mitun Shil, Hrishi Rakshit, Hadaate Ullah, An adjustable window function to design an FIR filter, 2017 IEEE (icIVPR)

[25] P.Kamala Kumari, J.B.Seventline, A survey on numerical representations of DNA sequences, Asian journal of convergence in technology 4, Issue 1, 2018

[26] R.F Voss,Evolution of long range fractal correlations and 1/f noise in DNA base sequences .Physical review letters,1992,68(25):3805-3808.

[27] P. Ramachandran, W. Lu, and A. Antoniou, Filter-based methodology for the location of hot spots in proteins and exons in DNA,IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1598-1609, June 2012.

[28] Sajid A. Marhon, Stefen C. Kremer, Gene Prediction Based on DNA Spectral Analysis: A Literature Review, Journal of Computational Biology volume 18, number 4, 2011

[29] Proakis JG, Manolakis D, Digital signal processing , Prentice- Hall of India Pvt Ltd, Fourth 2007.

[30] A. Oppenheim, R. Schafer and J. Buck, “Discrete –Time Signal Processing”, Prentice –Hall, second edition ,1999.

[31] M.Ahmad, et al , From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? A review, J.Biomedical signal processing and control 34(2017)44-63

[32] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, CABIOS 13 (1997) 263–270.

[33] W. Li, The study of correlation structures of DNA sequences: a critical review, Comput. Chem. 21 (1997) 257–272.

[34] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice–Hall, Englewood Cliffs, NJ, 1993.

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

[36] Mohhammed Abo-Zahhad, Sabah M. Ahmed, Shimma A. Abd-Elrahman, Genomic Anlysis and Classification of exons and introns sequences using DNA numerical mapping techniques, I.J.Information Technology and Computer Science, 2012,8,22-36

[37] A. K. Singh and V. K. Srivastava, "Performance Evaluation of Different Window Functions for STDFT Based Exon Prediction Technique Taking Paired Numeric Mapping Scheme." *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2019, pp. 739-743, doi: 10.1109/SPIN.2019.8711741.

[38] M. Akhtar, J. Epps and E. Ambikairajah, "Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 310-321, June 2008, doi: 10.1109/JSTSP.2008.923854.

[39] Heba Mohammed Wassfy, et al Advanced DNA Mapping schemes for exon prediction using Digital filters, American Journal of Biomedical Engineering 2016,6(1):25-31

[40] I. M. El-Badawy, S. Gasser, M. E. Khedr and A. M. Aziz, "Improved time-domain approaches for locating exons in DNA using zero-phase filtering," *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, 2014, pp. 1334-1337, doi: 10.1109/GlobalSIP.2014.7032340.

[41] Hamidreza Saberkari, Mousa Shamsi,Hamed Heravi, Mohammad Houssein Sedaghi, A Fast algorithm for exonic regions prediction in DNA sequences, J. of Medical Signals &Sensors, Vol 3. Issue 3, 2013.