An efficient way of identification of protein coding regions of eukaryotic genes using digital FIR filter governed by Ramanujan's Sum

Subhajit Kar* and Madhabi Ganguly

West Bengal State University,
Barasat, Kolkata, West Bengal-700126, India
Email: subhajitkar.wbsu@gmail.com
Email: ray_madhabi@yahoo.co.in
*Corresponding author

Abstract: Finding protein coding regions, i.e., exons in a gene is a complex problem due to its diverse nature. In this paper, a novel FIR filtering governed by Ramanujan's Sum is proposed for identification of protein coding regions in gene. The efficacy of the designed algorithms is tested on Caenorhabditis Elegans cosmid F56F11.4a, various benchmark datasets like GENSCAN, HMR195, ASP67, and, BG570, and compared to well-established algorithms based on Antinotch, Butterworth, and Comb filters. The numerical conversion of the biological sequence here is an integer sequence and Ramanujan's Sum always generates a periodic sequence of integer numbers. This results in reduced quantisation error and simple hardware implementation. The evaluation of the designed Ramanujan's Sum governed filtering is done at the exonic level, nucleotide level, and through ROC plots. The results obtained on gene F56F11.4 attain specificity of 82%, sensitivity 97%, and precision of 85% while the AUC value of ROC curve was calculated as 0.96 square units. These evaluation parameters reveal that the proposed method gives enhanced results while comparing it to other existing exon-finding techniques.

Keywords: FIR filter; Ramanujan's Sum; wavelet transform; exons.

Reference to this paper should be made as follows: Kar, S. and Ganguly, M. (2023) 'An efficient way of identification of protein coding regions of eukaryotic genes using digital FIR filter governed by Ramanujan's Sum', *Int. J. Biomedical Engineering and Technology*, Vol. 43, No. 2, pp.152–184.

Biographical notes: Subhajit Kar is currently pursuing his PhD in Signal Processing from the department of Electronics, West Bengal State University, India. He worked as Assistant Professors at Brainware College of Professional Studies and Brainware University where he taught Electronic Science and took analogue and digital electronics lab classes. He has obtained his BSc in Electronics from University of Calcutta in 2009 and completed his MSc in Electronics from West Bengal State University in 2011. He has published many research articles in the field of genomic signal processing published in various international journals and conference proceedings. His research interests include genomic signal processing, image processing, machine and deep learning based biological applications.

Madhabi Ganguly is an Assistant Professor of the Department of Electronics, West Bengal State University, Kolkata, India. She received her BE in Electrical Engineering from IIEST, Shibpur and she obtained her ME in Control System specialisation from Department of Electronics and Telecommunication

Engineering, Jadavpur University, Kolkata, India. She received her PhD from Department of Electronics and Telecommunication Engineering, Jadavpur University. She was a CSIR Fellow. She has published many papers in various international journals and conference proceedings. Her research area includes digital and genomic signal processing, application of nanotechnology, control system, etc.

1 Introduction

Genomic signal processing (GSP) is defined as the analysis, processing, and use of genomic data to gain biological knowledge, and the translation of that knowledge into systems-based applications that can be used to diagnose and treat genetic diseases. The scientific tool working behind the GSP algorithm is digital signal processing and its vast evolved techniques. Various DSP techniques like discrete Fourier transform, Wavelet transform, Digital filters, Parametric models, and Hilbert transform are widely used to capture various biological information such as identification of protein-coding regions, detection of motif and tandem repeats in gene and protein, protein structure prediction, pattern recognition, DNA or RNA sequence comparison, reading frame identification and many more (Rao and Swamy, 2008; Garg and Sharma, 2020; Singh and Dehuri, 2022; Cunha et al., 2022).

The genes of the eukaryotic genome are not continuous and consist of coding (exon) and non-coding regions (intron). Furthermore, the lengths of coding regions are not consistent in various genes suggesting very small as well as long coding regions could exist. The presence of an exitron, dual exon in the genetic sequence makes it further difficult to identify all regions in a precise way. The study of coding regions is very important as it is responsible for coding various proteins. Any mutation in coding regions will change the structure of the coded protein and therefore normal biological processes will be hampered.

The splicing algorithm of exons in the gene is based on the period-3 property which states that the spectrum of protein coding DNA has a peak at every third component at the frequency of $2\pi/3$ radian/sec (Anastassiou, 2001). Previous experimental results showed that DSP tools-based algorithms to identify protein coding regions in a DNA sequence are not accurate to find optimal boundaries of exons and introns. Hence, various statistical parameters are also considered in designing the filters such as the Kalman filter or statistically optimised notched filter to improve the efficiency (Zhang et al., 2012). The basic filter-based approach to the problem was first addressed by Baidyanathan and Yoon. They designed a band-pass filter called antinotch filter with a passband centred at $\omega = 2\pi / 3$ and minimum stopband attenuation of about 13 dB which can effectively eliminate $1 / f$ noise present in the power spectra (Vaidyanathan and Yoon, 2002a). Barman et al. (2012) proposed an IIR antinotch filter with higher stop-band attenuation and an increasing number of multipliers to remove harmonic distortion. Sharma et al. (2013) showed FIR filters using windows and optimal least square techniques are also effective in the prediction of coding regions. The advantage of the FIR filter over IIR is its inherent stability (Kar and Ganguly, 2022). Guan and Tuqan (2004) introduced a multitrate DSP model that used a real IIR filter and an ideal low pass filter which further improve the detection process by effectively suppressing the background noise. Hota and

Srivastava (2012a) proposed three models based on an antinotch filter known as harmonic suppressor antinotch filter, conjugate suppressor antinotch filter, and moving average antinotch filter to remove unwanted noise present in output power spectra due to the leaking of conjugate and harmonic frequencies. Furthermore, various adaptive comb filters are employed successfully for the precise prediction of introns and exons (Meher et al., 2011). A single peaking IIR Butterworth filter is utilised by George and Thomas (2010) which showed good accuracy in exon prediction. Kumari and Seventline (2021) designed a FIR filter with a combinational window function comprising Gaussian, Lanczos, and Chebyshev windows to fine-tune period-3 property in exons. Adaptive and optimal filtering techniques are also employed to address exon finding problems (Kar et al., 2022; Rahman et al., 2022). Adaptive exon predictors based algorithms update the filter coefficients according to the feedback received from the output and compared them with the desired signal for further modification. In this way, precise filter design can be obtained. Very recently, Tenneti and Vaidyanathan (2016a, 2016b, 2018) designed the Ramanujan filter bank which found wide application in epileptic seizure detection, tandem repeats, and protein repeats detection in DNA sequences. Although previous methods for predicting exons found promising results, there are still remains some issues to address. Most of the techniques analyse gene sequences based on a sliding window technique where a rectangular or any other shaped window is walked through the sequence. Determination of the correct window length is tedious work which can affect the outcome of the method. In the case of wavelet transform-based algorithms to determine the boundaries of exonic regions one have to choose the proper mother wavelet from various available options making the method extensive. Some of the previous well-established methods are much more complex in nature and hence computational complexity is very high. The proposed method provides a window-independent, only integer-based computation which provides a fast and efficient prediction of intron-exon interfaces.

Ramanujan filter bank theory is based on Ramanujan's Summation provided by Ramanujan in 1918. The Sum proved to be very useful to represent several well-defined arithmetic functions like Eulers Totient function, Von Sterneck's arithmetic function, and Brauer-Rademacher identity (Ramanujan, 1918). The Sum has many properties which are very essential in terms of DSP (Vaidyanathan, 2014). Even though the concept of Ramanujan's Sum was purely designed for the mathematical purpose, very recently it has been applied to various engineering branches like information theory, digital communication, signal and image processing, system theory and machine learning, VLSI signal processing, quantum information theory to produce great results (Vaidyanathan and Tenneti, 2020; Planat et al., 2009).

In this study, we have designed an FIR filter whose impulse response is determined by Ramanujan's Sum. Ramanujan's Sum is known to generate periodic integer terms which are used to design a band-pass filter to detect $\omega = 2 / 3 \pi$ frequency components. The designed filter provides a non-adaptive and data-independent way to compute the period-3 components from a given gene, thus making the design computationally efficient. The conventional antinotch filters used for detecting exons are not accurate as they pass harmonic and conjugate frequencies of the original detected frequency such as period $P = 3$ in our case. Thus, additional filters are required in the second stage along with this filter for efficient prediction of period-3 components, resulting increase in complexity. Similarly, generalised comb filters that are extensively used for exon identification either require adaptive structure or complex cascaded algorithms to

accurately find out period-3 frequencies and consequently exon sequences. The simple comb filter cannot distinguish a period $P = 3$ and $P = 3m$ ($m > 1$) signal allowing additional frequencies to be passed through the band-pass filter along with period-3 frequency. The IIR filters although required less hardware but unstable in nature and should be designed to provide stability. To eliminate such drawbacks, Ramanujan sum based FIR filter is designed that provide less computational requirement to detect exact period-3 frequencies to increase the accuracy of so called exon finding problem.

2 Methodology

This paper proposed an FIR filter governed by Ramanujan's sum to compute the power spectra of a given nucleotide sequence and predict the boundaries of exons and introns. To apply the biological sequence in a digital system the sequence must be converted into appropriate mapping. The mapping we adopted is a simple integer mapping to reduce the computational complexity of the algorithm. The filter output has noise buried within the spectra which could affect the threshold selection procedure. To eliminate the noise discrete wavelet transform together with the Gaussian window is applied. Previously DWT based denoising techniques are adopted successfully by using various mother wavelets such as Haar, Coiflet, Daubechies, and Symlets (Liu and Luan, 2014; Abbasi et al., 2011). In this paper, we utilised discrete Meyer wavelet (Dmey) and obtained enhanced noise removal. To evaluate the model ROC plot is computed on various datasets that determine the accuracy of the model.

Figure 1 Block diagram of the proposed FIR filter based protein coding detection approach (see online version for colours)

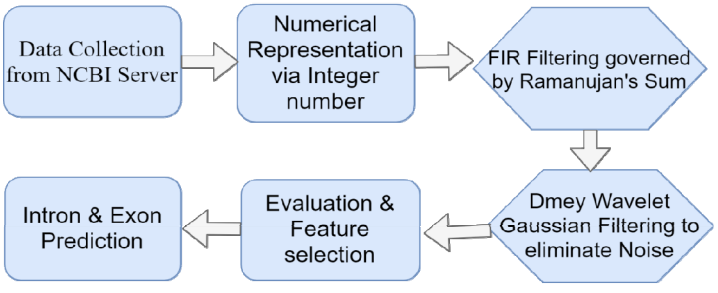


Figure 1 shows the proposed algorithm to identify introns and exons using digital filters. The process mainly comprises six steps. In the first step, nucleotide sequences in FASTA format are downloaded from the NCBI website. In the next step, it is converted into a digital sequence by integer mapping. The third step comprises bandpass FIR filtering governed by Ramanujan's Sum. The power spectral density plot is formulated from the output of the band-pass filter. The spectrum obtained in the previous step is then passed through a wavelet-based denoising filter. In the final step, proper thresholding and feature selection is done to recognise coding regions.

2.1 Data acquisition

Data collection for the experiment is a very important aspect of research. However, due to rapid progress in genome research various genomic and proteonomic sequences are easily available on various online databases such as NCBI, UNIPROT, UCSC, and ENSEMBL. Such databases contain a huge amount of data resources and provide them in various formats like FASTA, XML, ASN, and many more. The proposed method has experimented with DNA sequence having accession number AF099922 (precisely F56F11.4a) as it is the benchmark gene for the testing of exon finding algorithms. The gene belongs to the Caenorhabditis Elegans which is a type of worm. Other relevant genes which are taken into account for evaluation, are described in Table 1. The gene sequences are collected from National Centre of Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov) in the FASTA format.

Table 1 Description of DNA sequences used to evaluate the accuracy of the proposed algorithm

GenBank accession number	*Species*	*Number of exons*	*Information*
M62420	Homo sapiens	3	Length: 3,326 bp
			Locus: HUMCBRG
			Definition: Homo sapiens carbonyl reductase gene
			Coded protein: Carbonyl reductase [NADPH] 1
NC004843	Buchnera aphidicola	2	Length: 2,308 bp
			Locus: NC_004843
			Definition: Buchnera aphidicola Ps plasmid pBPS1
M13145	Bos taurus	3	Length: 1,769 bp
			Locus: BOVANPA
			Definition: Bovine atrial natriuretic peptide (ANP) gene
			Coded protein: Natriuretic peptides A
AF042784	Mus musculus	2	Length: 2234 bp
			Locus: AF042784
			Definition: Mus musculus galanin receptor type 2 (GalR2) gene
			Coded protein: Galanin receptor type 2
AF099922	Caenorhabditis elegans	5	Length: 42,799 but only 7,021 to 15,020 bp are considered
			Locus: AF099922
			Definition: Caenorhabditis elegans cosmid F56F11

These sequences originally belong to dataset GENSCAN containing 64 multi-exon Human genome sequences developed by Burge, dataset BG570 consists of 570 single-gene vertebrate sequences developed by Burset and Guigo, HMR195 consists of

195 single-gene human, mouse, and rat sequences assembled by Sanga-Rogic, ASP67 dataset consists of 67 multiple-gene sequences of Aspergillus fumigatus, which are part of the TIGR dataset, and KEGG dataset prepared by Kanehisa and Goto (Meher et al., 2012b; Mena-Chalco et al., 2008; Akhtar et al., 2007). The proposed algorithm was evaluated using these datasets by plotting the ROC curve and then computing AUC for each dataset. Relevant details about these datasets are given in Table 2.

Table 2 Description of datasets used in this study

Dataset	*No. of sequences*	*Avg. coding %*	*C-G content*	*No. of exons*
GENSCAN test	64	10.2	56 %	381
BG570	570	15.37	55 %	2,649
HMR195	195	14	56 %	948
ASP67	67	45	52 %	778

These datasets contain genes taken from human, mouse, mammals, vertebrates, fungi, plants and bacteria. Also, these genes can have any number of exons with varying lengths. Therefore, a large pool of versatile gene sequences is considered to evaluate the exon finding problem. Unlike machine learning-based methods where a huge number of sequences must be collected in order to prepare test and training sets, the proposed method can be evaluated on any number of sequences since it is a data-independent method. Datasets HMR195, Asp67, and BG570 are available from http://www.vision.ime.usp.br/~jmena/mgwt/datasets whereas GENSCAN test dataset is available from http://www.imtech.res.in/raghava/genebench/datasets/Kulp-Reese/Human/. The web address for accessing KEGG dataset is https://www.genome.jp/kegg/genome/. All these websites are last accessed on August 2022.

2.2 Numerical mapping of biological sequence

A DNA sequence is a combination of four nucleotide bases which are adenine (A), guanine (G), cytocine (C), and thymine (T). These nucleotides are complementary in nature means 'A' always paired up with 'T' and 'C' is always paired up with 'G' making a stable double-helical DNA structure. As genes are part of DNA structure, they eventually have the same structure. Genomic information is digital in nature as it is constituted by a finite number of entities. The DNA character sequence is not applicable to a signal processing tool. Therefore, they must be converted into an equivalent numerical sequence so that the original information content of DNA remains plenary.

In this paper, a fixed real number-based mapping technique is adopted where each nucleotide base is replaced by an integer number. The assignment was proposed by Cristea (2002) and can be obtained by attaching four digits to the nucleotides as T = 0, C = 1, A = 2, and G = 3. The mapping is chosen for three reasons.

1 It reflects the physicochemical property of DNA as the purine bases (A or G) consist of two carbon-nitrogen rings they are bigger in size compared to pyrimidine bases (C or T) which contain one carbon-nitrogen ring i.e (A or G) > (C or T).

2 The integer mapping is compatible with Ramanujan's sum as Ramanujan's sum always generates a periodic integer series. The computation time will be less as integer calculation takes less clock cycle.

3 The output representation will be a one-dimensional mapping which will reduce the computational burden by 75% compared to the four-dimensional Voss representation.

For the DNA sequence x(n) = [A T A C C A G……A A G] the corresponding numerical sequence will be X(n) = [2 0 2 1 1 2 3…….2 2 3] according to the above rule.

Various numerical representations can be found in works of literature which include classical representations like Voss, Atomic number, and Molecular mass, and modern representations like Z-curve, EIIP, hydration enthalpy, minimum entropy, dinucleotide representation, I-ching representation, and GCC representation (Kwan and Arniker, 2009; Yin and Yau, 2008; Das and Turkoglu, 2018). Although all the representations are applied successfully to address GSP-related problems, EIIP and Voss representations gained much attention in the field of exon identification problems (Yu et al., 2018). Numerical mapping is an important aspect of the algorithm since the power spectral output of the filter is vastly correlated with the adopted mapping technique.

2.3 Ramanujan filter to detect period – 3 components in gene

A digital band-pass filter is designed which can effectively pluck the period-3 components from a gene so that exact boundaries of coding regions can be predicted. The impulse response of the filter is computed using Ramanujan's Sum which has the following important properties with respect to signal processing applications:

1 Periodicity – The sequence generated by the sum repeats itself, i.e.,

$$C_q(n) = C_q(n+q) \quad \forall n \tag{1}$$

2 Symmetric – If the generated sum is $C_q(n)$ then $C_q(-n) = C_q(n)$.

$$C_q(n) = \sum_{a=1}^{q} e^{2\pi i \frac{a}{q} n} \tag{2}$$

If $C_q(n)$ is expanded using Euler's rule then it can be written as:

$$C_q(n) = \sum_{(a,q)=1} \cos\left(2\pi a \frac{n}{q}\right) \tag{3}$$

Which implies, $C_q(n) = C_q(-n)$.

3 Real integer terms – Despite the presence of the complex j in the summation, the numbers generated are all integers. If $(a, q) = 1$ then $(a - k, q) = 1$, so $e^{j2\pi a \frac{n}{q}}$ and $e^{-j2\pi a \frac{n}{q}}$ will appear in the sum.

The convolution operation is mathematically equivalent to FIR filtering. Using convolution one can easily find the impulse response of the LTI system (Hansen, 2014). If the input of the system is a unit impulse, then the output of the system is known as impulse response which is given by:

$$h(n) = T[\partial(n)] \tag{4}$$

For an LTI system, if the input sequence $x(n)$ and the impulse response $h(n)$ are given, the output sequence $y(n)$ in the time domain can be given by:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) \tag{5}$$

Since the input numerical sequence can be assumed t start from index zero, the output can be modified for the right-handed input sequence as:

$$y(n) = \sum_{k=0}^{D-1} h_k x(n-k) = x(n) * h(n) \tag{6}$$

where h_k = impulse response coefficient, D = length of the impulse response, and k = index to sum over. Again,'*' denotes the convolution operator.

Here the input sequence represents a numerically converted DNA nucleotide sequence encoded by an integer number, thereby making it suitable to be applied in a digital filter. In order to extract a predefined period from an input signal using the convolution sum, the impulse response of the filter must be modified according to the design specification. We have replaced the impulse response of the filter with Ramanujan's Sum in order to locate protein coding regions in the gene characterised by the period three property.

Ramanujan's sum denoted by $C_q(n)$ is a function of two positive integer variables q and n provided by Ramanujan and defined by the formula:

$$C_q(n) = \sum_{a=1}^{q} e^{2\pi i \frac{a}{q} n} = \sum_{a=1}^{q} W_q^{-an} \tag{7}$$

where a and q are prime to each other, and $W_q = e^{-2\pi i/q}$. W_q is a rotating vector quantity defined as the Twiddle factor which is used here to reduce the computational complexity. Hua et al. (2014) showed that the computation of Ramanujan's sum only needs arithmetic operation instead of exponential calculation therefore making it less complex. It can be mathematically proved that $C_q(n)$ has the periodicity of q. Very important properties of Ramanujan's sum include symmetricity and orthogonality making it very efficient in signal processing by offering excellent energy conservation. The DTFT of $C_q(n)$ could be found as:

$$C_q\left(e^{iw}\right) = \sum_{n=-\infty}^{\infty} \sum_{a=1}^{q} e^{2\pi i \frac{a}{q} n} e^{-iwn} \tag{8}$$

$$= \sum_{n=-\infty}^{\infty} \sum_{a=1}^{q} 1.e^{-i\left(w-2\frac{a}{q}\right)n} \tag{9}$$

Applying time shifting property of DTFT,

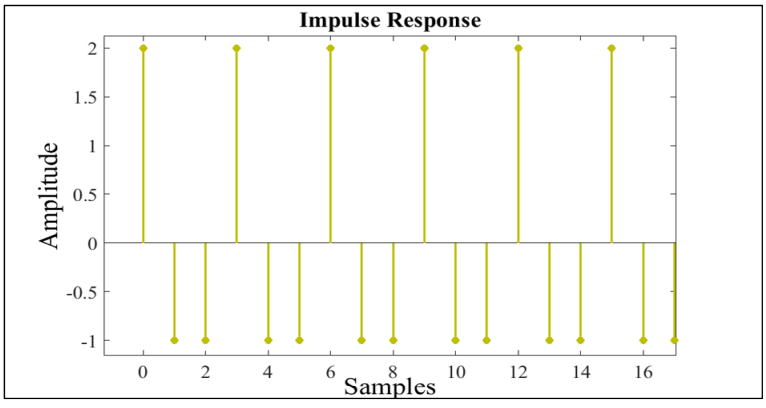
$$= 2\pi \sum_{a=1}^{q} \delta\left(w - \frac{2\pi a}{q}\right) \tag{10}$$

where $0 \le w \le 2\pi$.

Clearly, $C_q(e^{jw})$ is an impulse train whose values are zero everywhere except at the frequencies $2\pi a/q$ where 'a' is coprime to 'q'. Now if a signal $x(n)$ having period P is applied to the filter then the output will be non-zero if P is multiple of q as the Fourier transform of a periodic signal is a line spectrum with period $2\pi k/P$ where k lies between 0 to $P-1$.

Let the impulse response of the proposed filter be $C_p(n)$ $0 \le n \le KP-1$. Where P is the period of the input signal and K is the filter length. K must be adjusted so that better time-frequency resolution is obtained. A very small value of K may result in poor time resolution. In the proposed method, the value of K is chosen as 93 as it provides the highest accuracy for gene sequence F56F11.4. The method is discussed in the result section.

Figure 2 Ramanajun's summation for $q = 3$ and $K = 6$ (see online version for colours)



It is well known that exons in genes contain period-3 components. Hence, a Ramanujan filter can be efficiently employed to predict the exons by capturing base three periodicity. Therefore, in the case of filtering period-3 frequencies of coding regions, $P = 3$ and $q = 3$ are chosen and the corresponding Ramanujan's Sum can be written using equation (7) as:

$$C_3(n) = \sum_{a=1}^{3} e^{2\pi i \frac{a}{3} n} \tag{11}$$

$$= e^{2\pi i \frac{1}{3} n} + e^{2\pi i \frac{2}{3} n} \tag{12}$$

Using Euler's formula, the above equation can be expanded as:

$$\begin{aligned} &\cos 2\pi \frac{1}{3} n + i \sin 2\pi \frac{1}{3} n + \cos 2\pi \frac{2}{3} n + i \sin 2\pi \frac{2}{3} n \\ &= 2 \cos \frac{2}{3} n\pi \\ &= \{2, -1, -1\} \end{aligned} \tag{13}$$

Thus, $C_3(n)$ is a set of three integers.

Figure 3 Magnitude-phase response and pole-zero plot of the designed filter for $K = 93$ (see online version for colours)

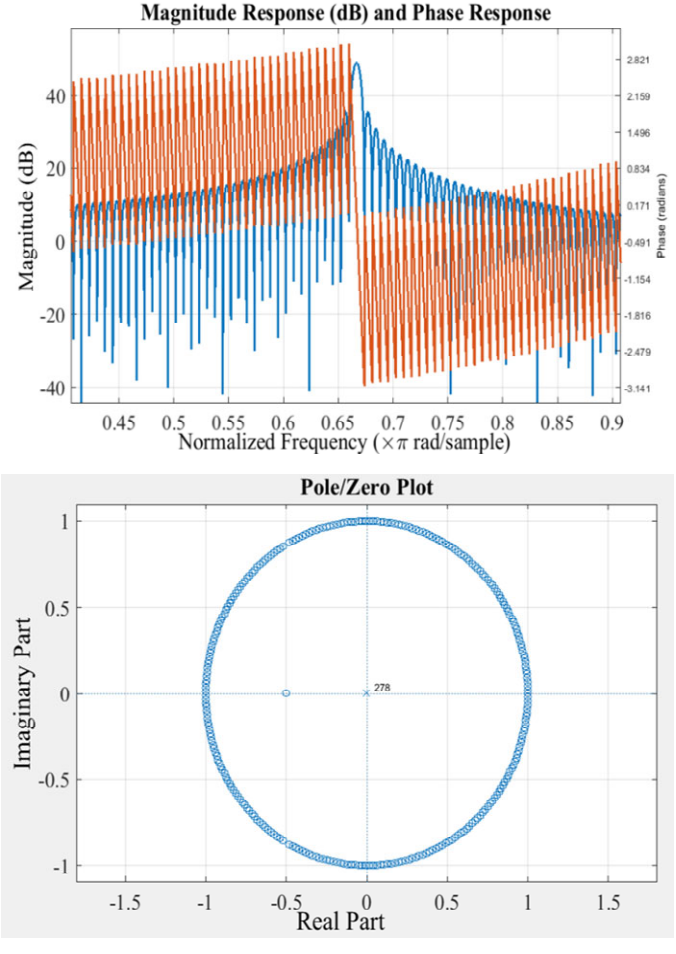
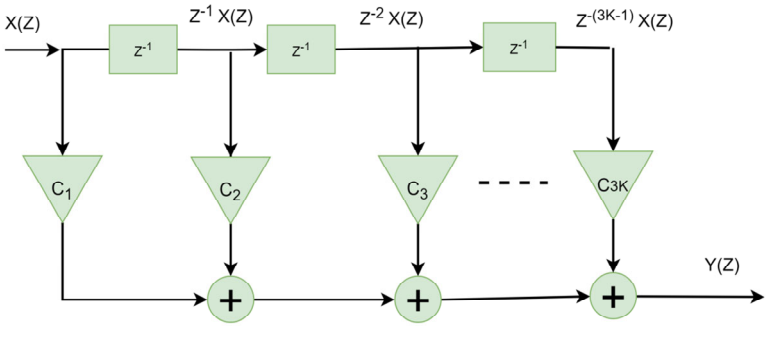


Figure 4 Schematic diagram of the designed FIR filter with $K = 93$ (see online version for colours)



Clearly, $C_3(n)$ is periodic in nature and it is depicted in Figure 2 for $K = 6$. $C_3(n)$ would be the one period of the impulse response of the designed Ramanujan's filter.

The designed filter is very much stable since an FIR system is inherently stable. In order to show poles and zeros in a complex plane, we must find the Z-transform of the $C_3(n)$ which is also known as a system function. The pole-zero and phase-magnitude plots of the designed filter are shown in Figure 3. Figure 4 shows the schematic diagram of the proposed filter.

The z transform of a right-handed discrete-time sequence $x(n)$ is defined as:

$$X(z) = \sum_{n=0}^{\infty} x(n) z^{-n} \tag{14}$$

Therefore, system function $H(z)$ of one period of the designed filter can be calculated as:

$$\begin{aligned} H(z) = C_3(z) &= \sum_{n=0}^{\infty} C_3(n) z^{-n} \\ &= 2 - z^{-1} - z^{-2} \end{aligned} \tag{15}$$

where z is a complex variable given as $z = e^{j\omega}$. From the transfer function of the filter, it is clear that the filter has $3*K$ zeros and a single pole situated at the origin. In this regard, it is essential to mention that the Z transform of Ramanujan's Sum can be derived using the method proposed by Samadi et al. (2005) which resulted in an IIR system function. Relationship between the system function $H(z)$ and the frequency response function is given by (Oppenheim et al., 1999):

$$H(\omega) = \sum_{-\infty}^{\infty} h(n) e^{-j\omega n} \tag{16}$$

Consequently, the frequency response by Fourier transforms for one period of the filter will be:

$$\begin{aligned} C_3(\omega) = F\{C_3(n)\} &= \sum_{n=-\infty}^{\infty} C_3(n) e^{-jwn} \\ = 2 - e^{-jw} - e^{-2jw} \end{aligned} \tag{17}$$

Hence, the final output of the filter in the time domain and frequency domain can be calculated as:

$$y(n) = x(n) * C_3(n) \tag{18}$$

$$Y(\omega) = X(\omega).C_3(\omega) \tag{19}$$

where $X(\omega)$ is the discrete-time Fourier transform of input numerical sequence $X(n)$ obtained by encoding selected gene sequences. Since the impulse response consists of fewer numbers of samples compared to the input sequences proper zero padding is required to eliminate aliasing.

IIR realisation

The Ramanujan's Sum can be expressed as:

$$C_q(n) = \sum_{a=1}^{q} e^{2\pi i \frac{a}{q} n} = \sum_{a=1}^{q} W_q^{-an} \tag{20}$$

where a and q are prime to each other and $W_q = e^{-2\pi i/q}$. If $C_q(Z)$ is the one-sided z-transform of the Ramanujan's Sum $C_q(n)$ then it is expressed as:

$$C_q(Z) = \sum_{n=0}^{\infty} C_q(n) z^{-n} \tag{21}$$

This can be expanded in the form of Twidle factor as:

$$C_q(Z) = \sum_{n=0}^{\infty} \sum_{a=1}^{q} W_q^{-an} z^{-n} \tag{22}$$

$$= \sum_{a=1}^{q} \sum_{n=0}^{\infty} \left(W_q^{-a} z^{-1} \right)^n \tag{23}$$

$$= \sum_{a=1}^{q} \frac{1}{1 - Q_q^{-a} z^{-1}} \tag{24}$$

$$= \frac{z \frac{d}{dz} F_q(Z)}{F_q(Z)} \tag{25}$$

For $q = 3$, the 2nd order cyclotomic polynomial can be found as (Dickson et al., 1923):

$$F_3(Z) = z^2 + z + 1 \tag{26}$$

Replacing it in equation (25) we get,

$$C_3(Z) = \frac{z \frac{d}{dz} (z^2 + z + 1)}{z^2 + z + 1} \tag{27}$$

$$= \frac{z(2z+1)}{z^2 + z + 1} \tag{28}$$

$$= \frac{2 + z^{-1}}{1 + z^{-1} + z^{-2}} \tag{29}$$

Stability must be designed in the IIR filter in order to practically utilise it for this particular problem. Also, the filter length must be adjusted so that the period three components can be enhanced by the filter.

2.4 Power spectral density of filtered data

The power spectral density plot is useful to assess the power content of a signal at different frequency levels. In the exon finding problem, it will help to find boundaries of exon and intron regions if the PSD plot is computed against the nucleotide bases. If $X[n]$

is the encoded numerical sequence then the designed bandpass filter will produce the output as $Y(\omega)$. Then the period-3 power spectrum of the input sequence can be expressed as:

$$S_k = |Y(\omega)|^2 \tag{30}$$

The output power spectrum is normalised to visualise the energy content of the signal in unit scale. It is advantageous to compute the normalised power spectrum for the approximation of threshold levels to distinguish intron-exon boundaries. The equation of normalised power spectrum is given as:

$$S_{nn} = \frac{S_k}{Max[S_k]} \tag{31}$$

where $Max[S_k]$ is the maximum value of the power spectrum.

2.5 Decreasing the noise using of discrete Meyer wavelet Gaussian filter

The output power spectrum of the designed bandpass filter needed to be processed in order to eliminate noise and thereafter proper selection of a threshold level is necessary. The threshold level determines the value below which the nucleotide regions will be considered as introns. The discrete Meyer wavelet transform at level 5 is assigned here to eliminate background noise present in the output power spectrum and predict exons and introns in a given nucleotide sequence.

A waveform of a small duration and zero average value is called a wavelet. Wavelet transform is calculated using a mother wavelet function φ_t, by convolving the original signal with the scaled and shifted version of the mother wavelet (Saini and Dewan, 2016). The discrete wavelet transform of a signal $f(t)$ is given by:

$$C(m,n) = < f, \varphi_{m,n} > = a_0^{-m/2} \int_{-\infty}^{\infty} f(t)\varphi^*\left(a_0^{-m}t - nb_0\right)dt \tag{32}$$

where $\varphi_{m,n}$ is the mother wavelet which in our case is 'Dmey' wavelet. Again, a_0 and b_0 are the scaling and translation coefficients respectively. Meyer wavelet belongs to a family of orthogonal continuous wavelet functions defined in the frequency domain and represented by $\varphi(\omega)$. 'Dmey' wavelet is the discrete variant of Meyer wavelet. The denoising of a signal is essentially estimating its true value from its noisy version. The model of the noisy signal is given by:

$$y(k) = x(k) + s(k) \tag{33}$$

where $x(k)$ is the original signal and $s(k)$ is the noise associated with it. In the case of the power spectral density plot obtained using bandpass filtering at the first stage, the mixed noise is pink noise which is inversely proportional to the frequency of the input signal. The noise components are smaller in magnitude but consist of wider bandwidth. To eliminate such noise, we used soft thresholding which denoises the output without degradation of the signal. The denoising algorithm can be explained in four steps.

1 The level-5 DWT $Y(k)$ of the noisy signal $y(k)$ is computed, resulting in the approximation component $Y_a(k)$ and detailed component $Y_d(k)$.

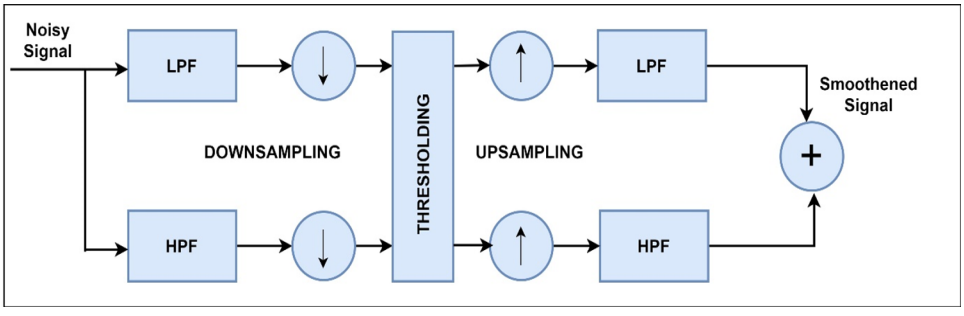
2 The detail component $Y_d(k)$ is subjected to a threshold operation to yield $Y_{dt}(k)$.

3 The modified DWT $Y_m(k)$ is constructed by concatenating $Y_a(k)$ and $Y_{dt}(k)$.

4 The level-5 inverse discrete wavelet transform of $Y_m(k)$ is computed resulting in the denoised version $\widetilde{x(k)} \approx x(k)$ of $y(k)$.

Denoising process using wavelet is described in Figure 5.

Figure 5 Signal decomposition and reconstruction at level 1 using wavelet transform (see online version for colours)



In computing DWT, a modified form of convolution is used as only one-half of the convolution is required to compute in DWT. This is known as double shifted convolution and is given as: $Y_{hp} = \sum_n x(n)g(2k-n)$ and $Y_{lp} = \sum_n x(n)h(2k-n)$ for the high pass and low pass filters respectively. $g(n)$ and $h(n)$ are the impulse responses of the filters.

The Gaussian window function efficiently applied in the image processing domain for noise removal, is connected with the wavelet filter to increase the overall efficiency. The effectiveness of the designed filter can be controlled by adjusting the window length of the Gaussian function. The coefficients of a Gaussian window can be computed from the following equation:

$$\omega(n) = e^{-\left(\frac{\alpha n}{L-1}\right)^2} = e^{\frac{-n^2}{2\sigma^2}} \tag{34}$$

where $\{-(L-1)/2 \le n \le (L-1)/2\}$ and α are inversely proportional to the standard deviation σ. The designed parameters and characteristics of the designed filter are provided in Figures 6 and 7, respectively.

2.6 Performance assessment tools

The performance of the proposed algorithm is performed in three ways:

1 evaluation at the exon level

2 evaluation at the nucleotide level

3 evaluation through ROC curve.

Various evaluation parameters used in this experiment for exon level and nucleotide level are described in Table 3. Evaluation parameters at the exon level determine the numbers

of correctly and wrongly identified exons and introns whereas nucleotide level evaluations are performed to measure the proportion of exonic nucleotides and intronic nucleotides that are correctly identified.

Figure 6 Residuals and various coefficients of the designed wavelet filter (see online version for colours)

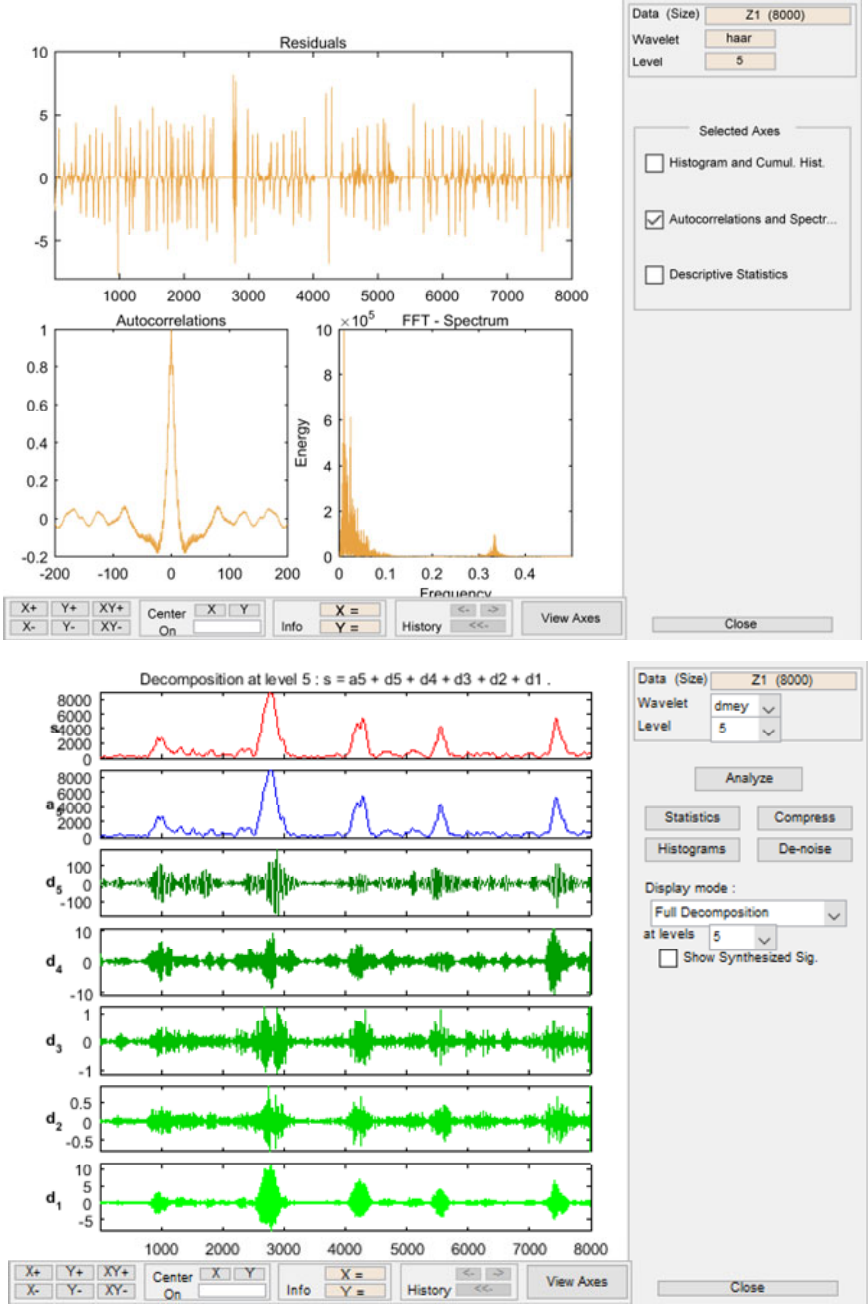
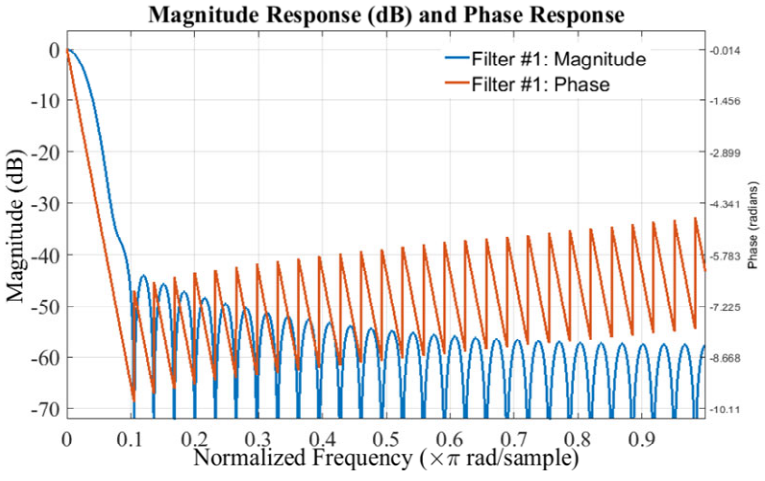


Figure 7 Magnitude and phase characteristic of designed Gaussian window for noise removal (see online version for colours)



Definitions of these evaluation parameters can be found in the literature (Rogic et al., 2001; El-Badawy et al., 2013; Meher et al., 2012a). We have segregated the signal-to-noise ratio so that it can be measured at nucleotide and exon levels. They are defined as:

$$SNR1 = \frac{Energy\ in\ the\ coding\ region}{Energy\ in\ the\ non\text{-}coding\ region} \tag{35}$$

$$SNR2 = \frac{Maximum\ value\ in\ spectrum}{Mean\ value\ of\ the\ spectrum} \tag{36}$$

The ROC curve is an important assessment tool which is used to visualise the performance of an algorithm model at all classification thresholds. A ROC curve plots true positive rate versus false-positive rates at different threshold levels. The area occupied by the ROC curve is measured as the area under the curve or AUC. AUC measures the degree of separability. Therefore, it suggests how good the designed model is to differentiate between coding and non-coding regions.

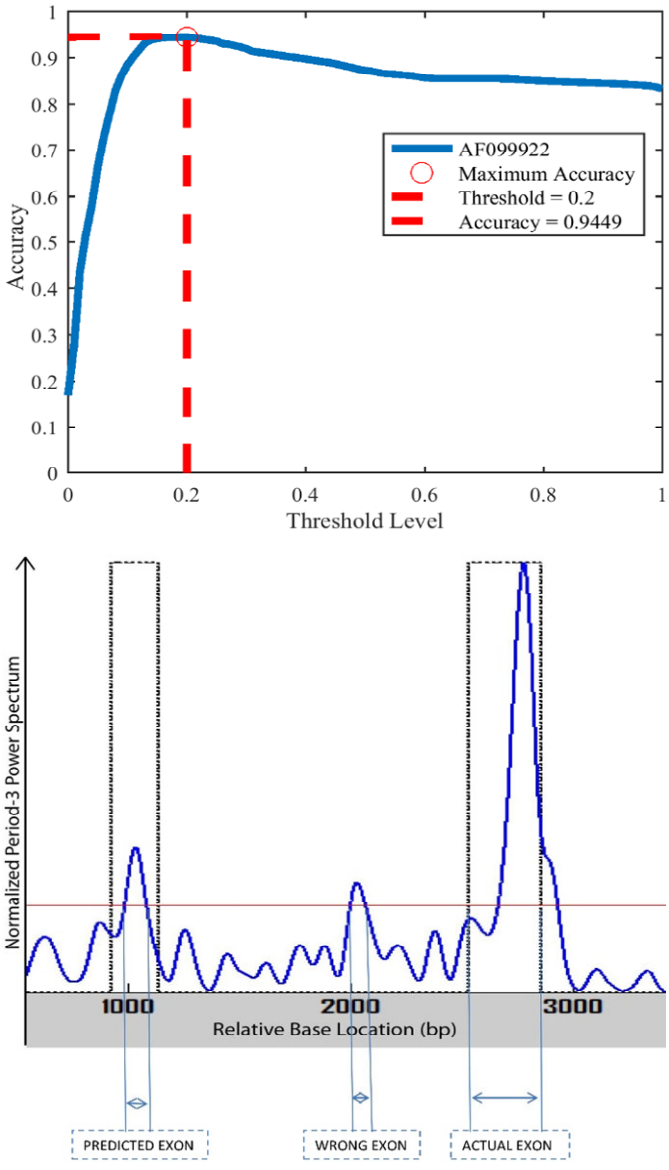
Table 3 Various evaluation parameters used in this study

Evaluation at the nucleotide level	*Evaluation at the Exon level*
Sensitivity, specificity, accuracy, F1 measure, precision, SNR1	Discrimination measure, wrong rate, missrate, INCLASS, EXCLASS, precision, SNR2

2.7 Selection of proper threshold value

A threshold level is superimposed on the output power spectrum to segregate exonic and intronic regions on a given nucleotide. Threshold selection plays an important role in the calculation of various evaluation parameters and hence could modify the accuracy of the implemented algorithm.

Figure 8 Selection of appropriate threshold value (see online version for colours)



Note: The red line in the PSD plot refers to the threshold line.

Various methods of selecting a threshold level are presented in previous literature. They are mainly based on sequence drive classification, Machine learning-based classification, and dynamic threshold classification (Marhon and Kremer, 2011). Kwan et al suggested an optimal threshold selection criterion based on the mean and standard deviation of the power spectrum. They have defined three threshold values namely the mid threshold value (T_m), the proportional threshold value (T_p), and the cumulative threshold value (T_c) for clustering between coding and non-coding regions (Kwan et al., 2012).

$$T_m = \frac{(meanP_{3i} + meanP_{3e}) + (stdP_{3i} - stdP_{3e})}{2} \tag{37}$$

$$T_p = \frac{sdP_{3e} \times meanP_{3i} + sdP_{3i} \times meanP_{3e}}{sdP_{3e} + sdP_{3i}} \tag{38}$$

$$T_c = Period - 3 \; value \; at \; minimum \left| F(P_{3e}) - F_c(P_{3i}) \right| \tag{39}$$

where $meanP_{3i}$ and sdP_{3i} are the mean and standard deviation of period-3 values obtained from introns. $F(P_{3e})$ is the cumulative distribution of all the exon period-3 values and $F_c(P_{3i})$ is the complementary cumulative distribution of all the intron values.

In addition, we have selected an optimum threshold (T_o) by looking at the threshold having maximum accuracy. In Figure 8, accuracy of the proposed method is plotted against various threshold values to determine the optimum threshold (T_o) for the gene AF099922.

3 Results

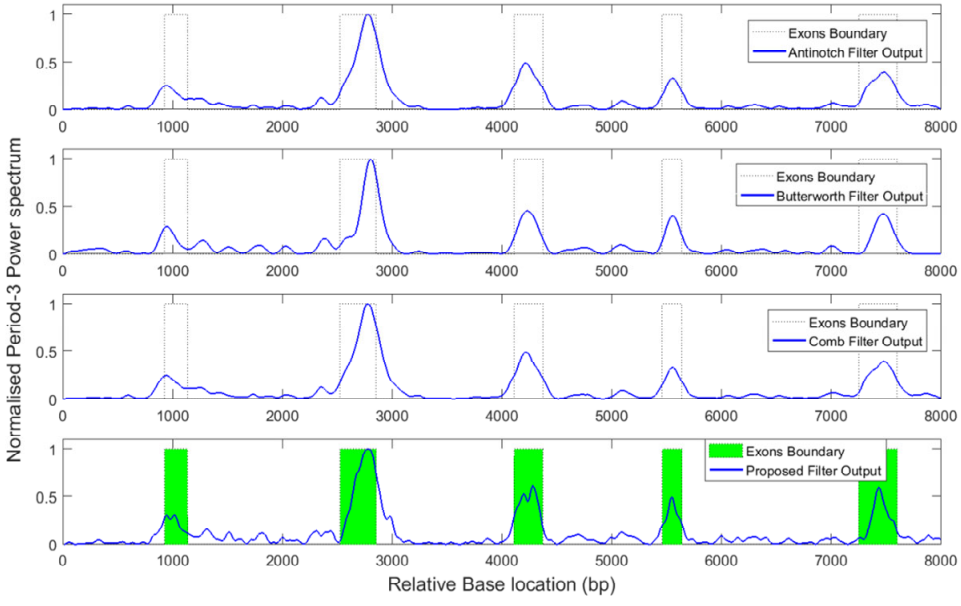
The proposed method is first tested on benchmark DNA sequence F56F11.4 in the C.Elegans chromosome III. The sequence has been experimented with in almost every literature involving exon finding problems. The sequence consists of five well-separated exons. The sequence is first passed through the proposed bandpass filter to strengthen three base periodicity then smoothened by the Gaussian 'Dmey' wavelet filter. To evaluate the performance of the designed filter it is compared with previously adopted three well-studied filters namely antinotch filter, generalised comb filter, and Butterworth filter (Vaidyanathan and Yoon, 2002b; Meher et al., 2011; Kar and Ganguly, 2022). The filters are chosen according to their superior performance and compatibility with the Ramanujan filter. Spectrums obtained using various filtering techniques on F56F11.4 is shown in Figure 9.

The peaks in the diagram correspond to regions where three base periodicity is present and hence it is identified as exons. The dotted lines and green shades are applied to locate exon boundaries. The exon boundaries are computed using the proposed algorithm and compared with the Genebank database to prove its effectiveness in exon identification. The result is given in Table 4.

Table 4 Actual and predicted exon locations of DNA sequence F56F11.4 using the proposed method

Specification	*Actual exon location (NCBI) in bp*		*Predicted exon location in bp*		*Measured deviation in bp*	
	Start	*End*	*Start*	*End*	*Start*	*End*
Exon 1	928	1,039	905	1,059	23	20
Exon 2	2,527	2,856	2,569	3,012	42	156
Exon 3	4,113	4,376	4,096	4,371	17	5
Exon 4	5,464	5,643	5,462	5,643	2	0
Exon 5	7,254	7,605	7,351	7,586	97	19

Figure 9 Normalised power spectrum of Genebank Accession number F56F11.4 using various filters (see online version for colours)



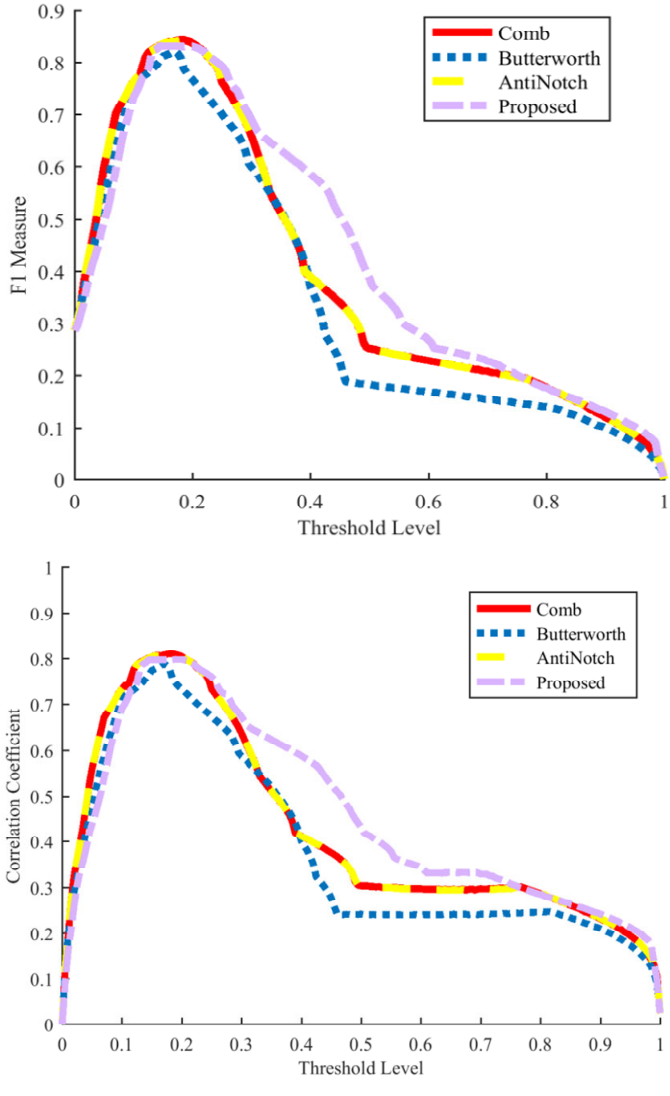
From the table, it is evident that all the exons are correctly identified by the proposed method and also very close to the original data provided by the NCBI website. The small deviation between the actual exons start and end location suggests another unique feature of our proposed approach compared with existing methods. Threshold versus F1 scores and threshold versus Matthews correlation coefficients (MCC) are plotted in Figure 10 below to establish the accuracy and applicability of this model where the performances of the four filters are specified in various colours in the plot.

The plots in Figure 10 show a peak at the threshold level close to '0.2' pointing to the optimal threshold value for gene F56F11.4. Another very important metric is the ROC curve which suggests how well the classifier performed to distinguish between two binary classes. The ROC plot for F56F11.4 is provided in Figure 11(b). The AUC is measured as 0.96 using the Ramanujan filter method.

In the exon identification problem, the choice of an optimal filter length is another parameter that has to be done properly to improve classification efficiency. In order to determine the filter length with the highest accuracy, we had to plot accuracy with respect to window lengths. The diagram is given in Figure 11(a).

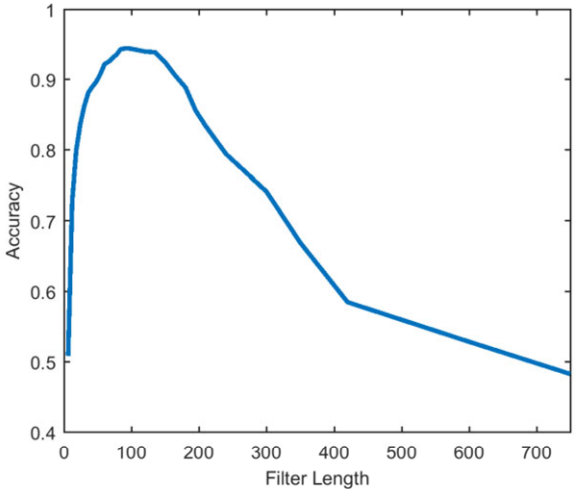
From Figure 11(a), we have measured the period of window length which performs at the highest accuracy in the case of Ramanujan filter is given by $K = 93$. Since each period consists of three components so the filter length can be computed as FL = 3 × 93 = 279. Next, we investigated the accuracy of the proposed algorithm using another four gene sequences. For further assessment of the sequences, we have incorporated threshold levels for each sequence using T_m, T_p, T_c, and T_o shown in Table 5.

Figure 10 Threshold versus F1 scores and threshold versus MCC measurements for gene sequence F56F11.4 (see online version for colours)

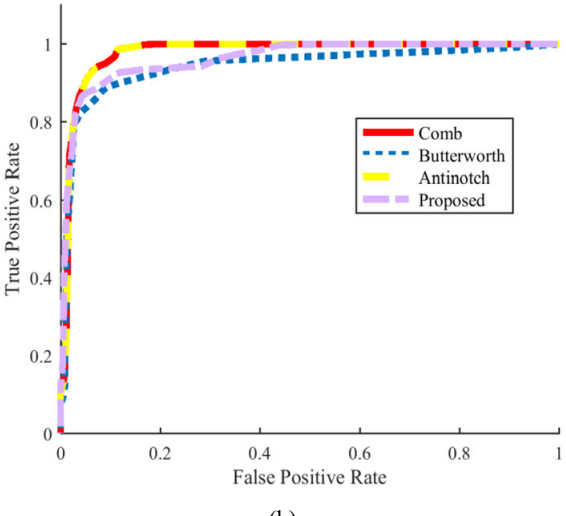


For evaluation purposes, only the optimal threshold level is considered. After the proper selection of threshold level various exon level and nucleotide level evaluation parameters are computed for five DNA sequences which are listed in Tables 6 and 7. Other threshold values are listed here to show that computed threshold values are similar to the other statistically defined threshold levels available in the literature. The power spectral density plots for gene M62420, NC004843, M13145, and AF042784 are provided in Figures 12(a), 12(b), 12(c) and 12(d), respectively.

Figure 11 (a) Selection of optimal filter length (b) ROC plot of F56F11.4 using proposed and other efficient filters (see online version for colours)



(a)

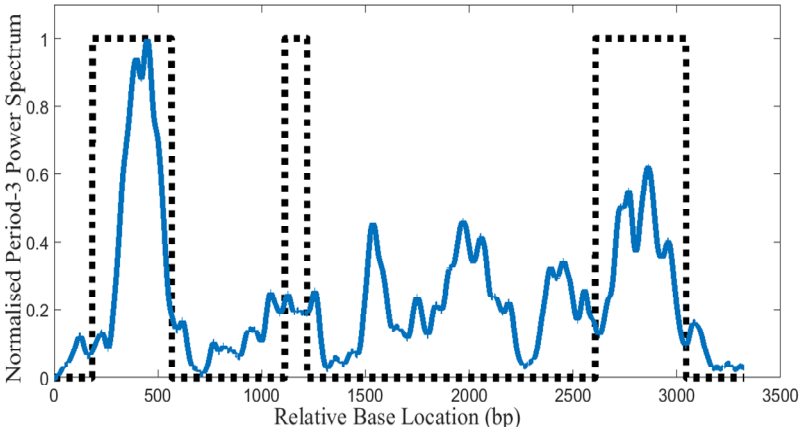


(b)

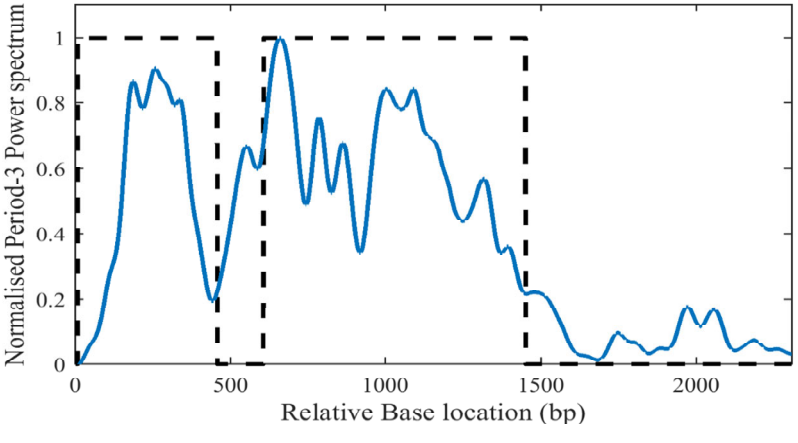
Table 5 Threshold selection parameters and their values for various genes given as NCBI accession numbers

Sequence ID	*Mid threshold (T_m)*	*Proportional threshold (T_p)*	*Cumulative threshold (T_c)*	*Optimal threshold (T_o)*
M62420	0.5	0.24	0.42	0.39
NC004843	0.69	0.33	0.26	0.23
M13145	0.6	0.28	0.5	0.46
AF042784	0.27	0.1	0.12	0.1
AF099922	0.37	0.13	0.22	0.2

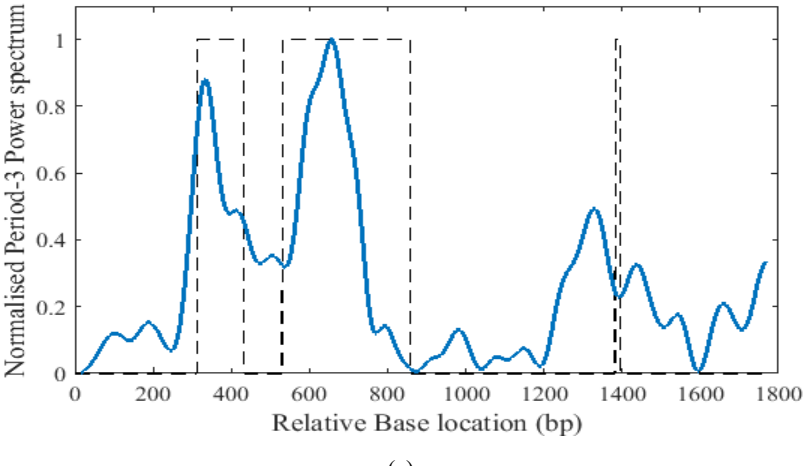
Figure 12 Power spectral density plot computed using the proposed filter for gene (a) M62420, (b) NC004843, (c) M13145 and (d) AF042784 (see online version for colours)



(a)

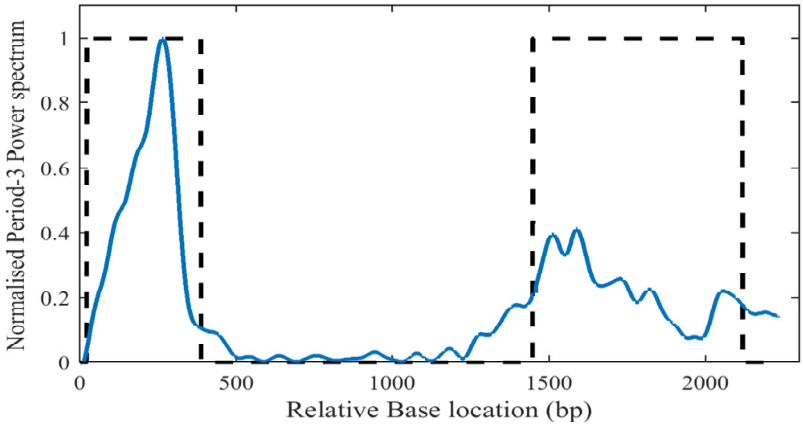


(b)



(c)

Figure 12 Power spectral density plot computed using the proposed filter for gene (a) M62420, (b) NC004843, (c) M13145 and (d) AF042784 (continued) (see online version for colours)



(d)

Table 6 Measurement of evaluation parameters at the exon level for selected gene sequences

Gene ID	*Discrimination factor*	*Miss rate*	*Wrong rate*	*INCLASS*	*EXCLASS*	*Precision*	*SNR2*
M62420	1.4	33%	0	100%	67%	86%	4.34
NC004843	3.35	0	33%	100%	100%	100%	2.55
M13145	1.8	33%	0	75%	67%	71%	3.82
AF042784	1.3	0	0	100%	100%	100%	5.63
AF099922	2	0	0	100%	100%	100%	8.8

The peaks in the PSD plots in Figure 12 are the locations of exons as determined by the proposed algorithm, whereas the valleys are the introns. The black dotted lines are denoting the actual boundaries of exons as obtained from NCBI website. Close observation suggests that the predicted exon boundaries are similar to the original values provided on the NCBI website. Various evaluation parameters at exon level classification are provided in Table 6.

Table 7 Measurement of evaluation parameters at the nucleotide level for selected genes

Gene ID	*Optimum threshold*	*Sensitivity*	*Specificity*	*F1 measure*	*Precision*	*Accuracy*	*SNR1*
M62420	0.39	0.49	0.96	0.62	0.84	0.83	1.5
NC004843	0.23	0.89	0.86	0.89	0.89	0.88	4.85
M13145	0.46	0.62	0.96	0.72	0.85	0.87	1.1
AF042784	0.1	0.91	0.85	0.88	0.86	0.88	6.75
AF099922	0.2	0.82	0.97	0.83	0.85	0.95	1.5

The obtained result is comparable with the recent state of the art as most of the parameters provided good accuracy. Miss rate and wrong rate values are zero for

sequences AF042784 and AF099922 suggesting there none of the exons are missed and no spurious peak in the spectrum. Intron and exon classification rates are also of the highest class. The exon-level signal-to-noise ratio is above one means clear discrimination between introns and exons. All the five sequences considered here have a precision rate of over 80% except for the sequence M13145. Table 7 depicts the evaluation parameters measured at the nucleotide level.

Higher sensitivity and specificity imply a high prediction rate for exons and introns respectively. F1 score is defined as the weighted mean of the test's precision and recall. If the F1 is high both the precision and recall of the classification indicate good results. F1 score is a very crucial parameter as it signifies the preciseness as well as the robustness of the designed classifier. The overall accuracy of the five gene sequences listed in Table 7 is 88.2 %. The signal to noise ratios are greater than 1 for every genes suggesting very good discrimination. In, addition to the tabled parameters we have also measured the false discovery rate, false negative rate, and negative predictive values. While FDR and FNR measured 0.15 and 0.02, NPV measured as 0.96 in the case of gene sequence F56F11.4.

4 Discussion

To show the effectiveness of the proposed filter in the current state of the art we compared it with others exon-finding methods found in recent literature. Most of these exon predicting techniques are steered by robust signal processing tools like discrete Fourier transform, digital filtering, and wavelet Transform. The performance of the proposed method is compared with three conventional filters which are IIR antinotch, comb, and Butterworth as well as other modern hybrid techniques.

4.1 Performance evaluation on benchmark sequence AF099922 (F56F11.4)

NCBI sequence AF099922 is the complete sequence of 'Caenorhabditis Elegans. It is 42799 base-pair long containing six protein-coding genes and two non-coding genes. The gene F56F11.4 is 8,000 bp long and stretched from base location 7021 to 15020 of AF099922. This gene contains five distinct exons having boundaries 928-1138, 2527-2856, 4113-4376, 5464-5643, and 7254-7605. We have taken the gene for performance evaluation as most of the exon predicting methods used it as a test sequence. Various evaluation parameters obtained by employing IIR antinotch, comb, Butterworth, and proposed Ramanujan filter when applied on F56F11.4 is provided in Table 8.

Table 8 Evaluation parameters for gene F56F11.4 at Th = 0.2

Filtering technique	*Sensitivity*	*Specificity*	*F1 score*	*Accuracy*	*Execution time (in sec)*
IIR Antinotch	0.83	0.97	0.837	0.94	0.065
Comb	0.83	0.97	0.837	0.94	0.070
Butterworth	0.69	0.98	0.678	0.93	0.165
Ramanujan filter	0.82	0.97	0.832	0.95	0.073

The results clearly indicate that Ramanujan filter acquired the best accuracy when comparing to the other filters. Although the execution time of the proposed method is slightly greater than antinotch and comb filter it is better than Butterworth and other hybrid methods. The FAGWT method developed by Das et al computed the PSD of the specified gene in 0.591 seconds (Das et al., 2020). The execution time of Saberkari et al is approximately 12 seconds for running the sequence F56F11.4 which is very high compared to the proposed algorithm (Saberkari et al., 2013). In order to verify the efficacy of our method we have compared two important parameters, i.e., AUC and approximate correlation (AC) with other renowned exon prediction methods. The corresponding details are available in Table 9.

Table 9 AUC and AC values as obtained by various state of art methods

Parameter	*Sequence F56F11.4*								
	MEMD +MGWT (Zheng et al., 2021)	*AST + PCA (Sharma et al., 2019)*	*Multistage filter (Hota and Srivastava, 2012)*	*AR + YW classifier (Roy and Barman, 2016)*	*ANF + MA (Hota and Srivastava, 2012)*	*Cross-correlation +WT (Abbasi et al., 2011)*	*WC + MBHW (Vaegae, 2020)*	*MA + WPT (Liu and Luan, 2014)*	*Proposed*
AUC	-	-	0.79	0.88	0.88	0.69	0.87	-	*0.96*
AC	0.84	-	-	-	0.77	0.78	-	0.42	*0.80*

The comparison model shows the proposed model attains the highest AUC and AC values compared to the recently developed and well-efficient exon finding algorithms. Similarly, other vital evaluation parameters which are specificity, sensitivity, and accuracy/precision are obtained from previous methods and compared with our method in the subsequent section. The proposed Ramanujan's Sum-based filtering generates a precision of 85% which measures the rate of correctly identified exons. Furthermore, the computed specificity and sensitivity values are found to be 97% and 82% respectively suggesting some exonic nucleotides are misclassified. Statistically optimised null filter (SONF) based method developed by Zhang et al. (2012) produced 91.4% sensitivity, 96% specificity, and 93.7% precision respectively. A multi-resolution-based wavelet transform method proposed by Marhon and Kremer (2015) measured 0.90% specificity and sensitivity when slid through F56F11.4 gene. Sahu and Panda (2011) suggested an S-transform based method which possesses superior time-frequency resolution compared with continuous wavelet transform and short time Fourier transform. The method produced a specificity value of 98%, a sensitivity value of 88 % and precision value of 96% (Sahu and Panda, 2011). To find the exon boundaries of genes Saberkari et al. (2016) used minimum variance spectrum estimation along with wavelet transformation. The method is effective to predict small exons but the accuracy of this method is only 68% while working with F56F11.4 sequence (Saberkari et al., 2016). Kar et al. (2019) proposed a fast Fourier transform method for identifying exons in a specific gene. The method proved to be computationally efficient in terms of time complexity. However, the sensitivity, specificity, and accuracy parameters derived as 80%, 90%, and 88% respectively which are well below the proposed model (Kar et al., 2019). Putluri and

Rahman (2018) experimented with various adaptive filtering techniques supposed to provide higher accuracy. For that, they employed various error functions to modify the filter coefficients. The evaluated results are moderate with respect to the exon identification problem as they calculated a sensitivity value of 76%, specificity value of 78%, and precision of 73% (Putluri and Rahman, 2018). The discrimination value of 2 resulted in our method by integer mapping is very good when compared to the previously designed methods based on EIIP and complex indicator mapping (Hota and Srivastava, 2008; Nair and Sreenadhan, 2006). Roy and Barman (2017) designed an IIR filter with the polyphase structure to find the protein coding regions in gene F56F11.4 resulting in a signal-to-noise ratio 1.90 (Roy and Barman, 2017). We achieved a similar SNR value at the optimal threshold suggesting the algorithm can accurately identify actual exons and introns locations in a given gene sequence.

4.2 Performance evaluation on benchmark datasets

In this section, the model is compared on the basis of results obtained by applying the proposed algorithm on four benchmark datasets which are GENSCAN, BG570, ASP67, and HMR195. The description of these four datasets are provided in the data acquisition section. At first, ROC plots are drawn for each dataset employing antinotch, comb, Butterworth, and Ramanujan filter. ROC plots are efficient as they compute sensitivity and specificity at different threshold values and thus provide an efficient way to estimate the performance of the designed algorithm. These plots are shown in Figure 13.

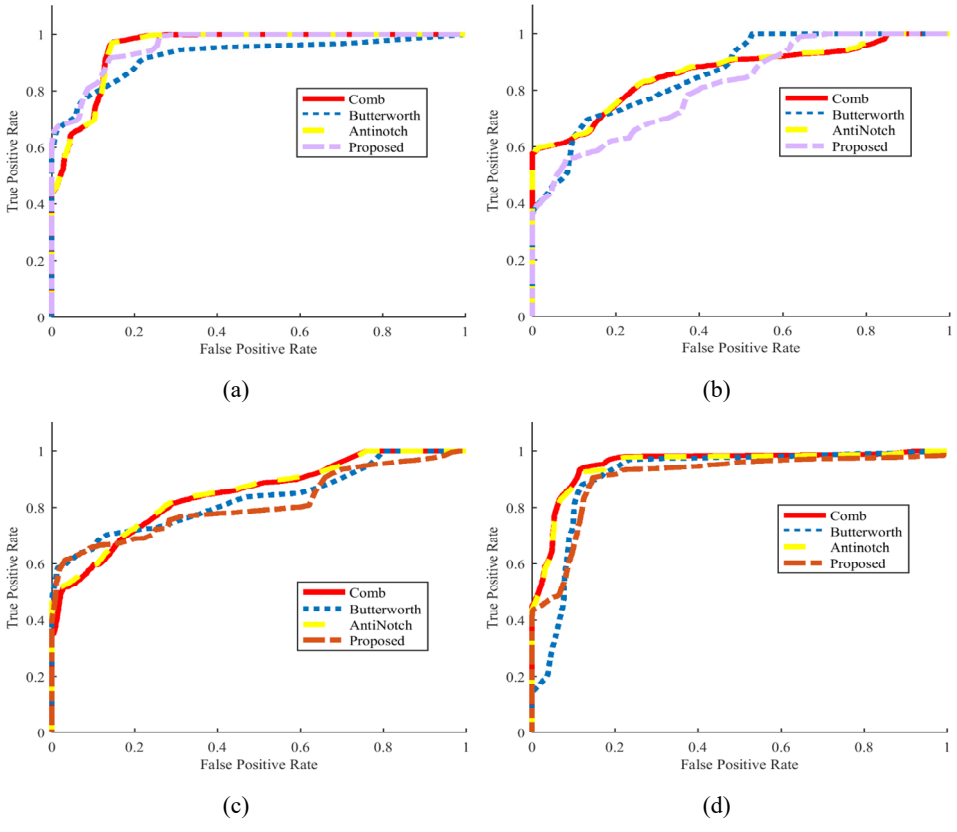
Table 10 Comparison of AUC values obtained from ROC plots using other state-of-art algorithms

Dataset	*AUC values*								
	MEMD +MGWT (Zheng et al., 2021)	*AST + PCA (Sharma et al., 2019)*	*Multistage filter (Hota and Srivastava, 2012)*	*AR + YW classifier (Roy and Barman, 2016)*	*ANF + MA (Hota and Srivastava, 2012)*	*Cross-correlation +WT (Abbasi et al., 2011)*	*WC + MBHW (Vaegae, 2020)*	*MA + WPT (Liu and Luan, 2014)*	*Proposed*
HMR195	0.74	0.83	0.74	0.71	0.77	0.83	0.70	0.72	*0.89*
BG570	0.67	0.83	0.72	0.70	0.75	0.81	0.66	0.65	*0.80*
ASP67	0.88	-	-	-	-	-	0.80	-	*0.77*
GENSCAN	-	0.85	0.76	-	0.80	-	-	0.68	*0.85*

The AUC value ranges from zero to one where values greater than 0.9 signify excellent classification accuracy between exons and introns. The proposed filter has performed exceptionally well for all datasets experimented. The AUC values are measured as 0.89, 0.85, 0.80, and 0.77 for HMR195, GENSCAN, BG670, and ASP67 respectively. While plotting the ROC plots only the average values of sensitivity and specificity are considered. The AUC values obtained from each dataset are very similar for antinotch and comb filters as seen in Figure 13. Both the filters belong to IIR category having filter order three. Although IIR filters produce the same results with lower filter stability must

be designed in IIR filter in order to apply it in practical applications. A comparative study of AUC values for various other algorithms available in the literatures are provided in Table 10.

Figure 13 ROC plots of various filters when applied on datasets (a) HMR195 (b) BG670 (c) ASP67 (d) GENSCAN64 (see online version for colours)



From the demonstrated Table 10, it can be visualised that the proposed method is able to provide the highest results for most of the datasets when compared to other algorithms designed for the prediction of exons and introns. HMR195 and BG570 datasets have been widely experimented in exon finding problems. The best AUC value obtained for HMR195 dataset is 0.89 obtained through the proposed filtering technique. While comparing the values obtained from other equivalent studies, we have taken the actual values resulted from the original studies. AC is a very good parameter which can be used solely to measure the accuracy of exon predicting problems as it combines both sensitivity and specificity into the calculation. Owing to this, we have calculated the AC values employing antinotch, comb, Butterworth, and Ramanujan's filter and compared them with MEMD + MGWT, Multistage filter, and MA + WPT techniques in Table 11 for evaluation. The AC values are calculated using the following two equations:

$$AC = (ACP - 0.5) * 2 \tag{40}$$

where ACP is the average conditional probability given as:

$$ACP = \frac{1}{4}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FN} + \frac{TN}{TN+FP}\right) \tag{41}$$

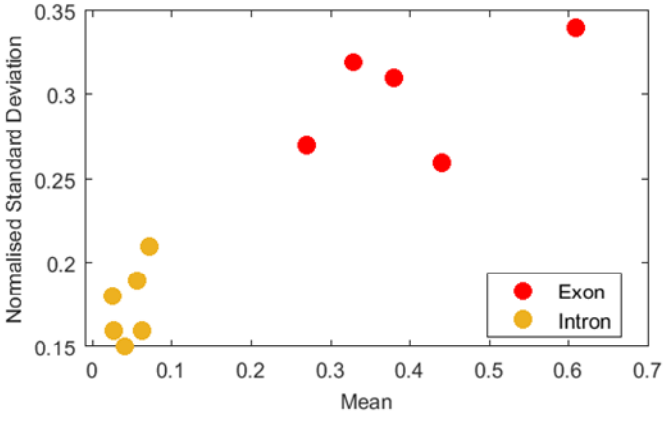
TP, *TN*, *FP*, *FN* are the true positive, true negative, false positive, and false negative values that can be obtained from the final PSD.

The above results suggest our method is very much proficient to predict exons and introns when large numbers of gene sequences are considered for testing. The highest obtained AC value is 0.755 found by evaluating HMR195 dataset. This value is better than the values obtained using methods like multistage filter, MEMD + MGWT, and MA + WPT. Notably, our research is mainly aimed to show digital FIR and IIR filters governed by Ramanujans Sum can produce very good exon prediction accuracy while providing simple integer-based computation. Additionally, the method is free from spectral leakage found in DFT. The major drawback of the method is to rely solely on the period three-based feature for exon detection whereas many exon sequences do not exhibit this feature thereby producing poor results for some datasets. The mechanism could be improved by introducing machine learning based method and including multiple properties of introns and exons for identification. Another disadvantage of the proposed method is the assignment of arbitrary integers to four nucleotides which results in an imaginary notion of distance between them. It can be improved by adopting reversible mapping techniques.

Table 11 Comparison of AC values obtained from various methods

Dataset	*AC values*						
	Antinotch filter	*Comb filter*	*Butterworth filter*	*Multistage filter*	*MEMD + MGWT*	*MA + WPT*	*Proposed*
HMR195	0.795	0.80	0.695	0.40	0.62	0.25	*0.755*
BG570	0.55	0.55	0.535	-	0.59	0.21	*0.560*
ASP67	0.51	0.49	0.565	-	0.69	-	*0.655*
GENSCAN	0.795	0.805	0.75	-	-	-	*0.750*

Figure 14 Machine learning approach to design a model in exon identification (see online version for colours)



We have demonstrated a scatter plot of normalised standard deviation versus mean of output PSD curve in Figure 14 as a prototype to discriminate between exons and intron regions by machine learning algorithm.

In the current scenario where data science and machine learning models are gaining much attention, it is inevitable to apply machine learning classifiers to solve the exon identification problem which could be extended to a large volume of genomes. The accuracy of such a machine learning model can be increased by enhancing the number of features extracted. The deep learning networks have the ability to extract detail characteristics of images at various filtering levels (Daş et al., 2020). Therefore, exon and intron sequences must be converted into images in order to apply deep learning algorithms to them.

5 Conclusions

The paper presented an FIR filter governed by Ramanujan's Sum which efficiently predicted the correct exons and introns in given nucleotide sequences. The main advantage of the method is its integer-based representation and processing. The filtered bandpass sequence can be achieved by the convolution of two integer sequences making it computationally more efficient. The proposed method is tested on various benchmark datasets to validate the results from a broader perspective. In the recent years, Ramanujans's Sum is gaining popularity in the field of engineering and physics due to its lucid properties. In digital signal processing fields, the Ramanujan's Sum introduces a breakthrough in solving various rigorous problems in speech, image and biomedical applications. In addition, we have employed a robust wavelet-based filtering based on the Gaussian window which was found very efficient in image processing applications. Although the computational efficiency of the algorithm is O(MN) where M and N are the lengths of two convolving sequences, it turns out to be very fast as it only deals with integers numbers that occupy less memory cells and takes less clock cycle to compute. The method did not utilise any prior knowledge of the analysed sequences which were used on the various statically optimised filters and hence it can be applied to all unknown sequences to predict intron-exon boundaries. One of the drawbacks of the proposed algorithm is that the integer number based numerical conversion introduces bias in the output spectra and could increase background noise. We have also designed the IIR model of the Ramanujan filter which can be studied in near future. Lastly, it is worth mentioning that most GSP algorithms to find exon location is based on period three property which may be unsuccessful for certain gene sequences due to inconsistent distribution of period-3 properties in exons and introns. The problem should be addressed in the coming studies so that period-3 based exon prediction methods find more practical applications.

References

Abbasi, O., Rostami, A. and Karimian, G. (2011) 'Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform', *BMC Bioinformatics*, Vol. 12, No. 1, pp.1–10.

Akhtar, M., Epps, J. and Ambikairajah, E. (2007) 'On DNA numerical representations for period-3 based exon prediction', in *2007 IEEE International Workshop on Genomic Signal Processing and Statistics*, IEEE, June, pp.1–4.

Anastassiou, D. (2001) 'Genomic signal processing', *IEEE Signal Processing Magazine*, Vol. 18, No. 4, pp.8–20.

Barman, S., Biswas, S., Das, S. and Roy, M. (2012) 'Performance analysis and simulation of IIR anti-notch filter with various structures for gene prediction application', in *2012 5th International Conference on Computers and Devices for Communication (CODEC)*, IEEE, December, pp.1–4.

Cristea, P.D. (2002) 'Genetic signal representation and analysis', in *Functional Monitoring and Drug-Tissue Interaction*, SPIE, June, Vol. 4623, pp.77–84.

Cunha, R.M., Silva, G., Alves, M., Bispo, B.C., Alves, D., Garrett, C. and Rodrigues, P.M. (2022) 'EEG wavelet packet power spectrum tool for checking Alzheimer's disease progression', *International Journal of Biomedical Engineering and Technology*, Vol. 40, No. 3, pp.289–302.

Das, B. and Turkoglu, I. (2018) 'A novel numerical mapping method based on entropy for digitizing DNA sequences', *Neural Computing and Applications*, Vol. 29, No. 8, pp.207–215.

Daş, B., Toraman, S. and Türkoğlu, İ. (2020) 'A novel genome analysis method with the entropy-based numerical technique using pretrained convolutional neural networks', *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol. 28, No. 4, pp.1932–1948.

Das, L., Das, J.K. and Nanda, S. (2020) 'Detection of exon location in eukaryotic DNA using a fuzzy adaptive Gabor wavelet transform', *Genomics*, Vol. 112, No. 6, pp.4406–4416.

Dickson, L.E., Mitchell, H.H. and Vandiver, H.S. (1923) *Algebraic Numbers: Report of the Committee on Algebraic Numbers*, National Research Council (No. 28), National Research Council of the National Academy of Sciences.

El-Badawy, I.M., Aziz, A.M., Gasser, S. and Khedr, M.E. (2013) 'A new multiple classifiers soft decisions fusion approach for exons prediction in DNA sequences', in *2013 IEEE International Conference on Signal and Image Processing Applications*, IEEE, October, pp.281–286.

Garg, P. and Sharma, S. (2020) 'Identification of CpG islands in DNA sequences using short-time Fourier transform', *Interdisciplinary Sciences: Computational Life Sciences*, Vol. 12, No. 3, pp.355–367.

George, T.P. and Thomas, T. (2010) 'Discrete wavelet transform de-noising in eukaryotic gene splicing', *BMC Bioinformatics*, Vol. 11, No. 1, pp.1–8.

Guan, R. and Tuqan, J. (2004) 'Multirate DSP models for gene detection', in *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, IEEE, November, Vol. 2, pp.1641–1645.

Hansen, E.W. (2014) *Fourier Transforms: Principles and Applications*, 1st ed., John Wiley & Sons, New Jersey, USA.

Hota, M.K. and Srivastava, V.K. (2008) 'DSP technique for gene and exon prediction taking complex indicator sequence', in *TENCON 2008-2008 IEEE Region 10 Conference*, IEEE, November, pp.1–6.

Hota, M.K. and Srivastava, V.K. (2012a) 'Identification of protein coding regions using antinotch filters', *Digital Signal Processing*, Vol. 22, No. 6, pp.869–877.

Hota, M.K. and Srivastava, V.K. (2012b) 'Multistage filters for identification of eukaryotic protein coding regions', *International Journal of Biomathematics*, Vol. 5, No. 2, p.1250018.

Hua, W., Wang, J. and Zhao, J. (2014) 'Discrete Ramanujan transform for distinguishing the protein coding regions from other regions', *Molecular and Cellular Probes*, Vol. 28, Nos. 5–6, pp.228–236.

Kar, S. and Ganguly, M. (2022) 'Study of effectiveness of FIR and IIR filters in Exon identification: a comparative approach', *Materials Today: Proceedings*.

Kar, S., Ganguly, M. and Das, S. (2019) 'Using DIT-FFT algorithm for identification of protein coding region in eukaryotic gene', *Biomedical Engineering: Applications, Basis and Communications*, Vol. 31, No. 1, p.1950002.

Kar, S., Ganguly, M. and Ganguly, A. (2022) 'Spectral analysis of DNA on 1-D hydration enthalpy-based numerical mapping using optimal filtering', in *Emerging Technologies for Computing, Communication and Smart Cities*, pp.137–149, Springer, Singapore.

Kumari, P. and Seventline, J. (2021) 'A novel approach for identification of exon locations in DNA sequences using GLC window', *International Journal of Biology and Biomedical Engineering*, Vol. 15, No. 1, pp.47–60.

Kwan, H.K. and Arniker, S.B. (2009) 'Numerical representation of DNA sequences', in *2009 IEEE International Conference on Electro/Information Technology*, IEEE, June, pp.307–310.

Kwan, H.K., Kwan, B.Y. and Kwan, J.Y. (2012) 'Novel methodologies for spectral classification of exon and intron sequences', *EURASIP Journal on Advances in Signal Processing*, Vol. 2012, No. 1, pp.1–14.

Liu, G. and Luan, Y. (2014) 'Identification of protein coding regions in the eukaryotic DNA sequences based on Marple algorithm and wavelet packets transform', in *Abstract and Applied Analysis*, July, Vol. 2014, Hindawi.

Marhon, S.A. and Kremer, S.C. (2011) 'Gene prediction based on DNA spectral analysis: a literature review', *Journal of Computational Biology*, Vol. 18, No. 4, pp.639–676.

Marhon, S.A. and Kremer, S.C. (2015) 'Prediction of protein coding regions using a wide-range wavelet window method', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 13, No. 4, pp.742–753.

Meher, J., Meher, P.K. and Dash, G. (2011) 'Improved comb filter based approach for effective prediction of protein coding regions in DNA sequences', *Journal of Signal and Information Processing*, Vol. 2, No. 2, p.88.

Meher, J.K., Meher, P.K., Dash, G.N. and Raval, M.K. (2012a) 'New encoded single-indicator sequences based on physico-chemical parameters for efficient exon identification', *International Journal of Bioinformatics Research and Applications*, Vol. 8, Nos. 1–2, pp.126–140.

Meher, J.K., Panigrahi, M.R., Dash, G.N. and Meher, P.K. (2012b) 'Wavelet based lossless DNA sequence compression for faster detection of eukaryotic protein coding regions', *International Journal of Image, Graphics and Signal Processing*, Vol. 4, No. 7, p.47.

Mena-Chalco, J., Carrer, H., Zana, Y. and Cesar Jr., R.M. (2008) 'Identification of protein coding regions using the modified Gabor-wavelet transform', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 5, No. 2, pp.198–207.

Nair, A.S. and Sreenadhan, S.P. (2006) 'A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)', *Bioinformation*, Vol. 1, No. 6, p.197.

Oppenheim, A.V., Ronald W.S, and John R.B. (1999) *Discrete-Time Signal Processing*, Prentice Hall, Upper Saddle River, NJ.

Planat, M., Minarovjech, M. and Saniga, M. (2009) 'Ramanujan sums analysis of long-period sequences and 1/f noise', *EPL (Europhysics Letters)*, Vol. 85, No. 4, p.40005.

Putluri, S.R. and Rahman, M.Z.U. (2018) 'Identification of protein coding region in DNA sequence using novel adaptive exon predictor', *Journal of Scientific and Industrial Research*, Vol. 77, No. 1, pp.87–91.

Rahman, Z.U., Vardhan, B.V., Jenith, L., Rakesh Reddy, V., Surekha, S. and Srinivasareddy, P. (2022) 'adaptive exon prediction using maximum error normalized algorithms', in *Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications*, Springer, Singapore, pp.511–523.

Ramanujan, S. (1918) 'On certain trigonometrical sums and their applications in the theory of numbers', *Trans. Cambridge Philos. Soc.*, Vol. 22, No. 13, pp.259–276.

Rao, K.D. and Swamy, M.N.S. (2008) 'Analysis of genomics and proteomics using DSP techniques', *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 55, No. 1, pp.370–378.

Rogic, S., Mackworth, A.K. and Ouellette, F.B. (2001) 'Evaluation of gene-finding programs on mammalian sequences', *Genome Research*, Vol. 11, No. 5, pp.817–832.

Roy, M. and Barman, S. (2016) 'Improved gene prediction by principal component analysis based autoregressive Yule-Walker method', *Gene*, Vol. 575, No. 2, pp.488–497.

Roy, S.S. and Barman, S. (2017) 'Polyphase filtering with variable mapping rule in protein coding region prediction', *Microsystem Technologies*, Vol. 23, No. 9, pp.4111–4121.

Saberkari, H., Farsani, M.S., Aminkar, S. and Shamsi, M. (2016) 'An efficient algorithm for small gene prediction in DNA sequences', *International Journal of Computer Vision and Signal Processing*, Vol. 6, No. 1, pp.1–10.

Saberkari, H., Shamsi, M., Heravi, H. and Sedaaghi, M.H. (2013) 'A novel fast algorithm for exon prediction in eukaryotic genes using linear predictive coding model and goertzel algorithm based on the Z-curve', *International Journal of Computer Applications*, Vol. 67, No. 17, pp.25–38.

Sahu, S.S. and Panda, G. (2011) 'Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach', *Genomics, Proteomics & Bioinformatics*, Vol. 9, Nos. 1–2, pp.45–55.

Saini, S. and Dewan, L. (2016) 'Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis', *SpringerPlus*, Vol. 5, No. 1, pp.1–15.

Samadi, S., Ahmad, M.O. and Swamy, M.S. (2005) 'Ramanujan sums and discrete Fourier transforms', *IEEE Signal Processing Letters*, Vol. 12, No. 4, pp.293–296.

Sharma, R., Kaushik, A. and Sharma, N. (2013) 'Gene prediction using FIR filter', *International Journal of Biomedical Engineering and Technology*, Vol. 11, No. 1, pp.33–45.

Sharma, S., Sharma, S.N. and Saxena, R. (2019) 'Identification of short exons disunited by a short intron in eukaryotic DNA regions', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 17, No. 5, pp.1660–1670.

Singh, N. and Dehuri, S. (2022) 'Epilepsy detection from electroencephalogram signal using singular value decomposition and extreme learning machine classifier', *International Journal of Biomedical Engineering and Technology*, Vol. 39, No. 1, pp.22–39.

Tenneti, S.V. and Vaidyanathan, P.P. (2016a) 'Detecting tandem repeats in DNA using Ramanujan filter bank', in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May, pp.21–24.

Tenneti, S.V. and Vaidyanathan, P.P. (2016b) 'Detection of protein repeats using the Ramanujan filter bank', in *2016 50th Asilomar Conference on Signals, Systems and Computers*, IEEE, November, pp.343–348.

Tenneti, S.V. and Vaidyanathan, P.P. (2018) 'Absence seizure detection using Ramanujan filter banks', in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, IEEE, October, pp.1913–1917.

Vaegae, N.K. (2020) 'Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes', *Biomedical Signal Processing and Control*, Vol. 58, p.101859.

Vaidyanathan, P.P. (2014) 'Ramanujan-sum expansions for finite duration (FIR) sequences', in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May, pp.4933–4937.

Vaidyanathan, P.P. and Tenneti, S. (2020) 'Srinivasa Ramanujan and signal-processing problems', *Philosophical Transactions of the Royal Society A*, Vol. 378, No. 2163, p.20180446.

Vaidyanathan, P.P. and Yoon, B.J. (2002a) 'Digital filters for gene prediction applications', in *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*, IEEE, November, Vol. 1, pp.306–310.

Vaidyanathan, P.P. and Yoon, B.J. (2002b) 'Gene and exon prediction using all pass-based filters', in *Workshop on Genomic Sig. Proc. and Stat*., Raleigh, NC, October.

Yin, C. and Yau, S.S.T. (2008) 'Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes', in *2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE, September, pp.223–227.

Yu, N., Li, Z. and Yu, Z. (2018) 'Survey on encoding schemes for genomic data representation and feature learning – from signal processing to machine learning', *Big Data Mining and Analytics*, Vol. 1, No. 3, pp.191–210.

Zhang, L., Tian, F. and Wang, S. (2012) 'A modified statistically optimal null filter method for recognizing protein-coding regions', *Genomics, Proteomics & Bioinformatics*, Vol. 10, No. 3, pp.166–173.

Zheng, Q., Chen, T., Zhou, W., Xie, L. and Su, H. (2021) 'Gene prediction by the noise-assisted MEMD and wavelet transform for identifying the protein coding regions', *Biocybernetics and Biomedical Engineering*, Vol. 41, No. 1, pp.196–210.