Exon Mapping in Long Noncoding RNAs Using Digital Filters

Genomics Insights
Volume 10: 1–12
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1178631017732029

Tina P George and Tessamma Thomas

Department of Electronics, Cochin University of Science and Technology (CUSAT), Kochi, India.

ABSTRACT: Long noncoding RNAs (lncRNAs) which were initially dismissed as "transcriptional noise" have become a vital area of study after their roles in biological regulation were discovered. Long noncoding RNAs have been implicated in various developmental processes and diseases. Here, we perform exon mapping of human lncRNA sequences (taken from National Center for Biotechnology Information GenBank) using digital filters. Antinotch digital filters are used to map out the exons of the lncRNA sequences analyzed. The period 3 property which is an established indicator for locating exons in genes is used here. Discrete wavelet transform filter bank is used to fine-tune the exon plots by selectively removing the spectral noise. The exon locations conform to the ranges specified in GenBank. In addition to exon prediction, G-C concentrations of lncRNA sequences are found, and the sequences are searched for START and STOP codons as these are indicators of coding potential.

KEYWORDS: Long noncoding RNA, exon prediction, antinotch filter, digital filter bank, DWT denoising, DSP methods

RECEIVED: May 5, 2017. **ACCEPTED:** August 18, 2017.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 520 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Tina P George, Research Scholar, Department of Electronics, Cochin University of Science And Technology (CUSAT), Kerala 682022, India. Email: ptinageorge@gmail.com

Introduction

The number of protein-coding genes has come down to around 20 000 with the completion of the human genome project in 2003, from the estimated 60 000 in the mid 1990s. Evolutionary studies proved that there is a large amount of apparently functional, yet noncoding, DNA contained in the human genome, the volume of which was estimated to be 4 times the amount of protein-coding sequences.[1] Long noncoding RNAs (lncRNAs) are a part of these "functional yet noncoding" sequences. Long noncoding RNAs started finding prime importance in microbiology research from the beginning of the current decade. Although initially ignored as "transcriptional noise," "dark matter," and so on, lncRNAs are now recognized as crucial elements in biological regulation. They have been found to be a diverse class of RNAs that engage in numerous biological processes across every branch of life.[2,3] It could be said that a new lncRNA is found to be upregulated or downregulated in a particular disease almost on a weekly basis.[4]

In the study of vertebrate genome, thousands of genes that code for lncRNAs have been identified. Eukaryotic genomes transcribe[5] a wide spectrum of RNA molecules which include long protein-coding messenger RNAs (mRNAs) to short noncoding transcripts. One of the striking observations made from transcriptome studies is that a much larger fraction of the genome is represented as exons in mature RNAs than what would be predicted from the amount of DNA covered by exons of protein-coding genes. Long noncoding RNAs are the major component of this all-encompassing transcription.[5] Early studies revealed that only around 5% to 10% of the genome is accounted for, by mRNA sequences and spliced noncoding RNAs that are transcribed in cell lines. It means that only around 1% of the human genome encodes proteins, leaving around 4% to 9% that is transcribed but whose functions are largely unknown.[5] Recent studies suggest that out of the human genome transcribed, only 2% accounts for protein-coding exons.[6] The exonic portion of human lncRNAs accounts for 1% of the genome which is about the same amount of DNA as protein-coding exons.[7]

Long noncoding RNAs constitute a heterogenic class of RNAs that include intergenic lncRNAs, antisense transcripts, and enhancer RNAs. Long noncoding RNAs generally refer to those sequences that are more than 200 nucleotides in length.[5,8] Defining lncRNAs by the virtue of what they are not, namely, neither short nor protein coding, is rather inapt. Nevertheless, the current imperfect level of understanding of their functions makes such a categorization practical. Long noncoding RNAs were called so primarily to distinguish them from small noncoding RNAs. They could be categorized based on their diverse empirical features, namely, genomic context,[4] origin of transcription, tissue specificity, molecular function, or mechanism of action. For example, based on the origin of transcription, we could have the following different categories: lncRNAs transcribed from intergenic regions are called long intervening noncoding RNAs, those transcribed from within introns of protein-coding genes are called intronic lncRNAs, those transcribed from the antisense strand of a given gene are called natural antisense transcripts, and so on.[9] However, their classification is not standardized. For example, defining a transcript, or its locus, as being coding or noncoding is unsatisfactory simply because of the inherent contrariness. Very often human genes possess both coding and noncoding transcripts which are difficult to distinguish without detailed experimental studies. It is equally difficult to label a transcript as being "intergenic."[1]

In this context, it also needs to be mentioned that many methods attempting to classify RNAs into protein coding and noncoding have come up. Some noncoding RNA sequences could actually code for peptides and some which are thought to be coding RNAs might not be so. Besides, protein-coding and noncoding transcripts often overlap as already mentioned. Such factors make it practically impossible to classify RNAs under this feature. RNAs cannot be unequivocally classified as being protein coding or nonprotein coding.[10] The functionality of any transcript at the RNA level should not be discounted. Hence, the very name "lncRNA" is not always truly descriptive of the function of a sequence.

Long noncoding RNAs of all kinds have been implicated in a range of developmental processes and diseases, but knowledge of the mechanisms by which they act is still surprisingly limited. At the same time, there are a small number of lncRNAs which have been well-studied from which we have been able to deduce important clues about the biology of these molecules. For example, metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) and myocardial infarction–associated transcript (MIAT) were shown to affect endothelial cell functions, whereas lincRNA-p21 controls neointima formation.[6] The Xist lncRNA has been found to be essential in X-chromosome inactivation during female eutherian mammalian development.[11] However, it is to be noted that the functions/involvement of lncRNAs is not limited to the ones mentioned here. The same lncRNA can be implicated more than one disease/function.

In the vertebrate genomes studied so far, thousands of genes encoding lncRNAs have been identified.[7] The human genome consists of many thousands of lncRNA. Analyses show that human lncRNAs are generated through pathways similar to that of protein-coding genes, with similar histone-modification profiles, splicing signals, and intron/exon lengths.[12] It has also been seen that lncRNAs exhibit a striking bias toward 2-exon transcripts unlike protein-coding genes.[12] Expression analyses have shown that lncRNAs are generally lower expressed than protein-coding genes. They show positive correlation with the expression of antisense strand of coding genes.[12,13]

Recent studies suggest the need for in-depth study of sequence, structural features, and genomic architecture of lncRNA.[14] In this work, we focus on lncRNA, typically said to be those with more than 200 nucleotides,[3,4] a heterogeneous group of sequences which are implied in diseases and cell development. Digital signal processing methods inherently have simplicity of implementation and ease of use. The aim of this work is to study lncRNA sequences using digital signal processing techniques and search for similarity/differences they have with coding genes as regards the signal/spectral properties of their sequences. Here, we apply digital filtering technique to map out the exons in lncRNA sequences taken from public database (National Center for Biotechnology Information [NCBI] GenBank). The property used here is the period 3 property which is an established feature[15–17] in locating exons in coding regions. Long noncoding RNAs have been found to have low values of GC concentration[14] which is considered to be one of the reasons for their lack of protein-coding capability. In this work, G-C content of the sequences in percentage relative to the net nucleotide content is computed. In this study, the sequences are also searched for START codon (both AUG and the alternate START codons) and the STOP codon patterns.

Materials and methods

Materials

In this work, we make use of human lncRNA sequences which are available in the NCBI GenBank. Only sequences of length more than 200 nucleotides are considered. The list of lncRNAs analyzed in this work is a random assorted list. It includes stand-alone lncRNAs (eg, MALAT1), natural antisense transcripts (BACE-AS1, FOXC2-AS1), lncRNAs implicated in diseases (CAHM, CCEPR), and so on. The list of sequences used in this work is shown in Table 1. Column 2 of Table 1 gives the name of the lncRNA, column 3 has the NCBI GenBank accession number, column 4 gives the location of the sequence in the Genome Browser,[18] and column 5 gives a brief description of the sequence as found in the corresponding NCBI record. More details about the sequences can be found at the NCBI Web site.[15]

Methods

The period 3 property which is an established digital signal processing (DSP) method to detect protein-coding regions in genes[16,17] and in gene detection[19–21] is used here. The base sequences in the coding regions (exons) of genes exhibit a strong period 3 component. This was observed by Trifonov and Sussman[22] as early as 1980. They maintained that this is due to the nonuniform codon usage in the formation of amino acids. Even though there are several codons that could possibly code a given amino acid, they are not used with uniform probability and this creates a codon bias. There is an excess Guanine in position 1, which leads to a strong period 3 oscillation.[23] There are other authors[20] who think this explanation is rather incomplete. But all authors do agree to the fact that the spectrum of protein-coding DNA has a peak at every third component (ie, at frequency component $k = N/3$, in a sequence of length N), and this property still remains widely accepted in predicting exons in eukaryotic coding regions. Nevertheless, it is not to be forgotten that such periodicity was observed 2 decades ago in noncoding regions for prokaryotes and some viral and mitochondrial base sequences.[24]

In this work, we make use of a DSP-based algorithm to map out the exons in lncRNA sequences using the period 3 property. Algorithms which exploit the period 3 property proceed

Table 1. lncRNA sequence list.

S. NO.	NAME OF THE LNCRNA	GENBANK ACCESSION NO.	LOCATION IN GENOME BROWSER	GENBANK INFORMATION OF THE SEQUENCE
1	CCEPR	NR_131782.1	chr6:163413065-163413950	*Homo sapiens* cervical carcinoma expressed PCNA regulatory lncRNA (CCEPR)
2	BACE1-AS	NR_037803.2	chr11:117,291,346-117,292,170	*Homo sapiens* BACE1 antisense RNA (BACE1-AS), antisense RNA
3	CAHM	NR_037593.1	chr6:163,413,065-163,413,950	*Homo sapiens* colon adenocarcinoma hypermethylated (nonprotein coding) (CAHM)
4	BGLT3	NR_121648.1	chr11:5,244,554-5,245,546	*Homo sapiens* β-globin locus transcript 3 (nonprotein coding) (BGLT3), lncRNA
5	ABALON	NR_131907.1	chr20:31,721,507-31,723,409	*Homo sapiens* apoptotic BCL2L1-antisense lncRNA (ABALON)
6	DISC2	NR_002227.2	chr1:231,814,626-231,818,517	*Homo sapiens* disrupted in schizophrenia 2 (nonprotein coding) (DISC2), lncRNA
7	GHET1	NR_130107.1	chr7:148,987,527-148,989,429	*Homo sapiens* gastric carcinoma proliferation enhancing transcript 1 (GHET1), lncRNA
8	HEIH	NR_045680.1	chr5:180,829,954-180,831,618	*Homo sapiens* hepatocellular carcinoma upregulated EZH2-associated lncRNA (HEIH)
9	NEAT1	NR_028272	chr11:65,422,798-65,426,532	*Homo sapiens* nuclear paraspeckle assembly transcript 1 (NEAT1), transcript variant MENepsilon, lncRNA
10	MALAT1 (TV1[a])	NR_002819.4	chr11:65,497,679-65,504,494	*Homo sapiens* metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), transcript variant 1, lncRNA
11	NKILA	NR_131157.1	chr20:57,710,183-57,712,780	*Homo sapiens* NF-κB interacting lncRNA (NKILA), lncRNA
12	FOXC2-AS1	NR_125795.1	chr16:86,565,145-86,567,761	*Homo sapiens* FOXC2 antisense RNA1 lncRNA
13	DLEU1 (TV2[b])	NR_002605.2	chr13:50082169-50107218	*Homo sapiens* deleted in lymphocytic leukemia 1 (DLEU1), transcript variant 2, lncRNA
14	HULC	NR_004855.2	chr6:8,652,209-8,653,846	*Homo sapiens* hepatocellular carcinoma upregulated lncRNA (HULC)
15	KIAA0087	NR_022006.1	chr7:26,533,121-26,538,825	*Homo sapiens* KIAA0087 lncRNA (KIAA0087), lncRNA
16	MHENCR	NR_132417.1	chr20:63,627,235-63,628,824	*Homo sapiens* melanoma highly expressed competing endogenous lncRNA for miR-425 and miR-489 (MHENCR), transcript variant 1, lncRNA
17	FALEC	NR_051960.1	chr1:150,515,757-150,518,032	*Homo sapiens* antisense of IGF2R nonprotein coding RNA (AIRN), transcript variant 1
18	PRNT	NR_024267.1	chr1:150,515,757-150,518,032	*Homo sapiens* prion protein (testis specific) (PRNT), transcript variant 1, lncRNA
19	HOTAIRM1	NR_038366.1	chr7:27,096,094-27,100,258	*Homo sapiens* HOXA transcript antisense RNA, myeloid-specific 1 (HOTAIRM1), transcript variant 1, lncRNA
20	CISTR (TV1[a])	NR_104332.1	chr12:53,750,447-53,757,034	*Homo sapiens* chondrogenesis-associated transcript (CISTR), transcript variant 1, lncRNA

Abbreviations: lncRNA, long noncoding RNA; PCNA, proliferating cell nuclear antigen.
[a]TV1—transcript variant 1.
[b]TV2—transcript variant 2.

by computing the discrete Fourier transform (DFT)[16,17] which is expected to exhibit a peak at frequency $2\pi/3$ in the spectrum. From the spectrum, the component at frequency $\omega = 2\pi/3$ can be located using a sharp single-frequency filter.

The mathematical mapping of the sequence string $x[n]$ is done making use of binary indicator sequences.[17] $[n]$, $u_u[n]$, $u_c[n]$, and $u_g[n]$ are the binary indicator sequences corresponding to A, U, C, and G which take on a value of 0 or 1 at

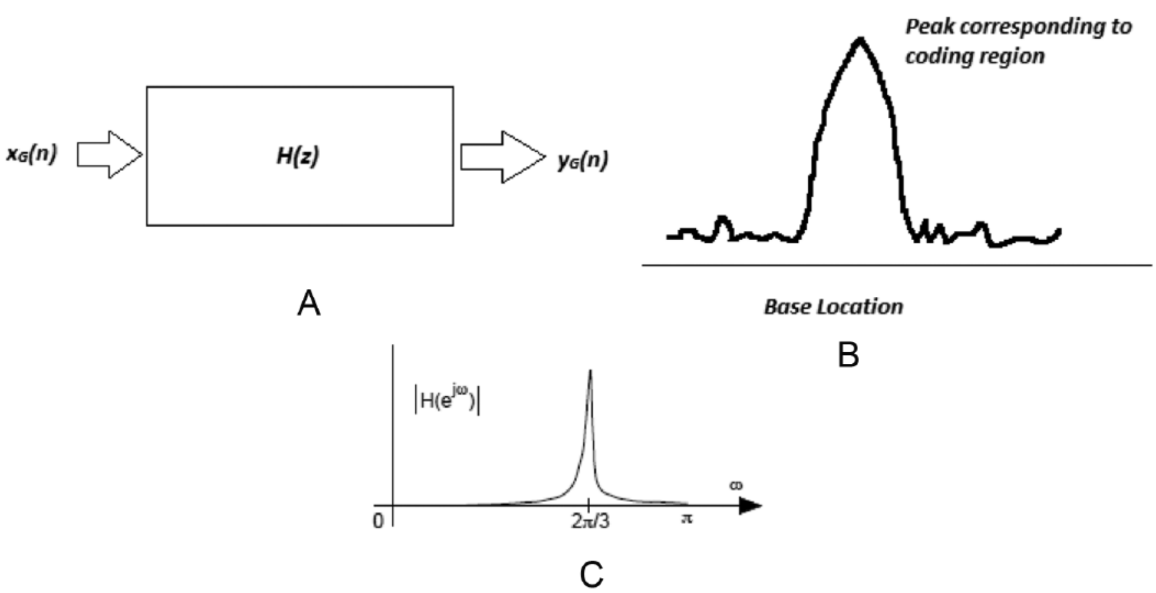

Figure 1. Expected O/P of the single peaking/antinotch filter. (A) $x_G(n)$ —Indicator sequence, $y_G(n)$ —output of the filter. (B) Expected output of antinotch filter. (C) $H(z)$ —antinotch filter with pass band centered at $2\pi/3$.

location n, depending on whether the corresponding character exists or not at n such that

$$u_a[n]+u_u[n]+u_c[n]+u_g[n]=1 \quad (1)$$

Discrete Fourier transform of a sequence[25,26] $y[n]$, of length N, is itself another sequence $Y[k]$, of the same length N, expressed mathematically as follows:

$$Y(k)=\sum_{n=0}^{N-1} y(n)e^{-(jk2n\pi)/N}, \text{ for } k=0,1,2,3,\ldots,(N-1) \quad (2)$$

Discrete Fourier transforms of individual indicator sequences $u_a[n]$, $u_u[n]$, $u_c[n]$, $u_g[n]$ ($U_a(k)$, $U_u(k)$, $U_c(k)$, $U_g(k)$, respectively) are computed as per equation (2) and the power spectrum is obtained as follows:

$$X(k)=\left|U_a(k)^2\right|+\left|U_u(k)^2\right|+\left|U_c(k)^2\right|+\left|U_g(k)^2\right| \quad (3)$$

We make use of sliding overlapping windows for better time resolution and compute the short-time Fourier transform (STFT). The length of the window has to be a multiple of 3 as the spectral content at the one-third of every window is to be picked out. In a former work,[17] we have found that the window length should be selected based on the length of the sequence used for optimum results. No rule of thumb is followed in the selection of the window length. The optimum window length is arrived at based on trial and error as to which gave the more accurate value of the position and length of exons.

Due to the period 3 property, we expect a peak in the spectrum at frequency $2\pi/3$, as shown in Figure 1, and this can be detected using a digital infinite impulse response (IIR)[25,26] antinotch filter.[16] This IIR antinotch filter is a lattice implementation. However, a direct form II implementation of the IIR single peaking is used here. The single peaking IIR filter is designed using the in-built filter design utility of the platform MATLAB 2016.[27]

The filter coefficients of the IIR single peaking design are as follows:

Numerator: $2.05798 \times 10^{-10}, 0, 2.05798 \times 10^{-10}$.

Denominator: 1, 0.99999, 0.999999.

The magnitude and phase responses of the filter are given in Figure 2, and the pole-zero plots are given in Figure 3. It is a single peaking, direct form II, transposed, stable filter of order 2 with very high Q factor.[25,26] The general form of the direct form II filter is given in Figure 4, and the general transfer function for Direct form II implementation is as follows:

$$H(z)=\frac{\sum_{k=0}^{M} b_k z^{-k}}{1+\sum_{k=1}^{N} a_k z^{-k}} \quad (4)$$

For order 2, $M=N=2$. An antinotch filter[16] with a sharp gain at the frequency $2\pi/3$ has an impulse response given as follows:

$$w(n)=\begin{cases} e^{j\omega_0 n} & 0 \le n \le N-1 \\ 0 & otherwise \end{cases} \quad (5)$$

The design starts by considering the second-order all-pass filter which has poles at $\mathrm{Re}^{\pm j\theta}$ and transfer function $A(z)$:

$$A(z)=\frac{R^2-2R\cos\theta z^{-1}+z^{-2}}{1-2R\cos\theta\, z^{-1}+R^2\, z^{-2}} \quad (6)$$

Let there be a filter $G(z)$ such that

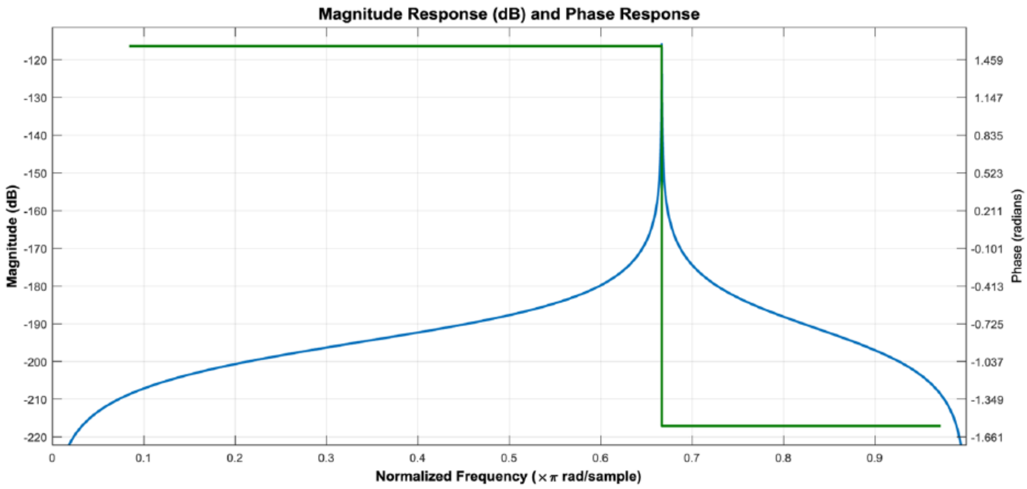

Figure 2. Magnitude and phase responses. Magnitude and phase responses of the infinite impulse response single peaking filter. Single peak at 2π/3^c. The magnitude response is shown in blue and phase response in green.

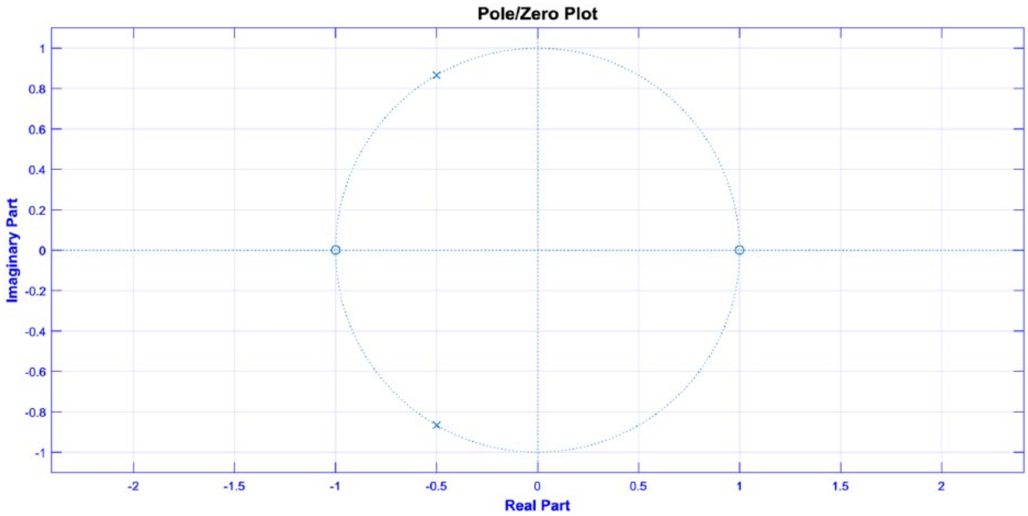

Figure 3. Pole-zero plot. Pole-zero plot of the infinite impulse response single peaking filter.

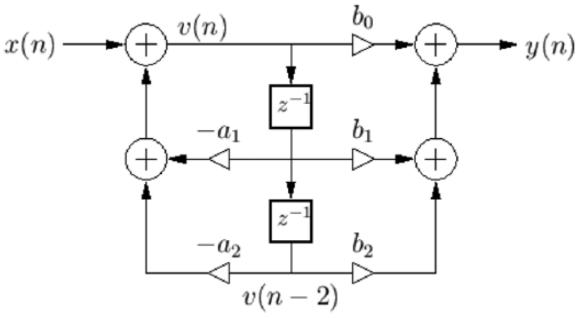

Figure 4. Direct form II. General form of direct form II infinite impulse response filter.

$$G(z)=\frac{1+A(z)}{2} \tag{7}$$

This can be simplified to the following form:

$$G(z)=K\left(\frac{1-2\cos\varnothing z^{-1}+z^{-2}}{1-2R\cos\theta z^{-1}+R^2z^{-2}}\right) \tag{8}$$

$G(z)$ represents a notch filter[16] with notching frequency at $\varnothing$.

If another filter $H(z)$ is designed as the difference such that

$$H(z)=\frac{1-A(z)}{2} \tag{9}$$

Then, $H(z)$ will have antinotching property; ie, $H(z)$ would be a single peaking filter.

Let $G(z)$ and $H(z)$ can be expressed together as follows:

$$\begin{bmatrix} G(z) \\ H(z) \end{bmatrix}=\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ A(z) \end{bmatrix} \tag{10}$$

$|Ae^{j\omega}|=1$ also, the 2×2 matrix is unitary, hence it follows that

$$\left|G\left(e^{j\omega}\right)\right|^2+\left|H\left(e^{j\omega}\right)\right|^2=1 \tag{11}$$

$G(z)$ and $H(z)$ are power complimentary and because $G(z)$ is a notch filter, $H(z)$ would be antinotching/single peaking filter. The IIR single peaking filter[17] gives a far better result than the

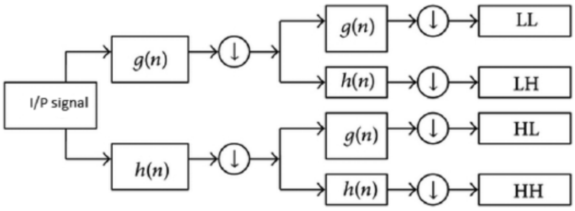

Figure 5. The discrete wavelet transform (DWT) filter bank. Representation of the DWT filter bank that performs frequency decimation.

IIR antinotch filter.[16] The subsequent filter bank[16] is not needed for removal of noise. This can be attributed to the high attenuation in the stop band of the peaking filter.

The exon map is improved by denoising the exon plot with the discrete wavelet transform (DWT).[28] The DWT is a digital filter bank which performs subband coding. A simple schematic representation is shown in Figure 5. The signal is passed through a filter bank consisting of low-pass and high-pass filters followed by scaling. The scale is altered by upsampling and downsampling or subsampling operations. Subsampling reduces sampling rate, whereas upsampling increases it. The incoming signal is split into 2 frequency-specific halves: the low-frequency (LF) half $g(n)$ and the high-frequency (HF) half $h(n)$.

This can be expressed by the following mathematical equations:

$$y_{high}\left[k\right]=\sum_{n}x\left[n\right]\cdot g\left[2k-1\right] \quad (12)$$

$$y_{low}\left[k\right]=\sum_{n}x\left[n\right]\cdot h\left[2k-1\right] \quad (13)$$

$g(n)$ and $h(n)$ are the impulse responses of the high-pass and low-pass filters, and $y_{high}[k]$ and $y_{low}[k]$ are their outputs.

The first half-band low-pass filter removes the higher half of frequencies, and the first half-band high-pass filter removes the lower half of frequencies. This is the first level of decimation which gives the level 1 (first level) DWT coefficients. Each of the LF and the HF halves are further decimated. If the signal contains a maximum frequency of 1000 Hz, then half-band low-pass filtering removes all the frequencies above 500 Hz. When an analogue signal with maximum frequency F_m is converted into digital, the perfect sampling frequency is to be maintained at $2F_m$, and the maximum digital frequency ω is 2π radians over a span from $-\pi$ to $+\pi$. Decimation with the half-band filters outputs of the first half-band filters to occupy frequency ranges from 0 to $\pi/2$ and $\pi/2$ to π, and so on. In this work, 2 levels of decomposition and successive reconstruction with the Haar wavelet is found to give a noise-free exon plot. Noise is found to occupy the higher frequencies, and hence, higher frequencies are not used in reconstruction. The advantage of using DWT denoising is that good time resolution is obtained at high frequencies and good frequency resolution at low frequencies with effective removal of noise. There is no need to compromise one for the other.

Long noncoding RNA is reported to have lower G-C content when compared with coding regions.[14] The G-C content of the sequences is found, relative to the total number of nucleotides. Sequence matching is done to locate START and STOP codon patterns. The sequences were checked for ATG and the alternative START codons too, namely, ATG, CTG, and GTC, and also for the STOP codons, TAA, TAG, and TGA. The results of the study are detailed in the next section.

Results

The exon maps of lncRNA sequences obtained with the period 3 property making use of digital filters. Short-time Fourier transform is used to obtain the spectrum of the sequences. While computing the spectrum using STFT, optimum window size is mandatory for locating the exons. Window sizes depend on the length of the sequence analyzed.[17] Denoising of exon plots is done with the help of the DWT filter bank which filters out HF noise and the LF components of decimation alone are used in reconstruction. It is found that the reduced computation technique[16] which applied a quadratic window and reduced noise in the case of exon prediction of coding DNA sequences is not found to be of use here. Applying the quadratic window is seen to introduce additional spectral noise and is not used in the algorithm here.

Figures 6 and 7 give sample exon maps. Figure 6 shows the exon plot of lncRNA CCEPR (*Homo sapiens* cervical carcinoma–expressed proliferating cell nuclear antigen regulatory lncRNA) and its GenBank accession number is NR_131782.1. It is 2502 bases long and contains a single exon as per the NCBI record (https://www.ncbi.nlm.nih.gov/nuccore/NR_131782.1), from 1 to 2502, ie, spanning the entire length of the sequence taken. The exon plots has nucleotide location along the *X* axis and the power spectral density (PSD) along the *Y* axis. As per the period 3 property, the energy peaks (peaks in the PSD) should correspond to exons. The peak power in this plot is between 6×10^{-17} and 7×10^{-17}, the half-power value is between 3×10^{-17} and 3.5×10^{-17}. Although there are dips in the plot, on an average, the plot retains the half power throughout and does not touch the 0 PSD value at any point. Hence, we count only 1 peak in this plot. Thus, there is a single exon extending from 1 to around 2450 to 2500. This range conforms to the value given in the NCBI database.

The next sample plot given in Figure 7 is that of lncRNA focally amplified lncRNA in epithelial cancer (FALEC) with NCBI accession number NR_051960.1. As per the NCBI record (https://www.ncbi.nlm.nih.gov/nuccore/NR_051960.1), FALEC has 2 exons: 1-306 and 307-566. The exon plot given in Figure 7 shows 2 energy peaks corresponding to 2 exons. Peak power value is between 0.6×10^{-17} and 0.8×10^{-17} and half-power values between 0.3×10^{-17} and 0.4×10^{-17}. Based on the very definition of half power, PSD values less than the half power are not considered as peaks. The first exon in the plot spans from 1 to around 190 and the second from around 220 to 560. The net length of the sequences in terms of nucleotides is 566.

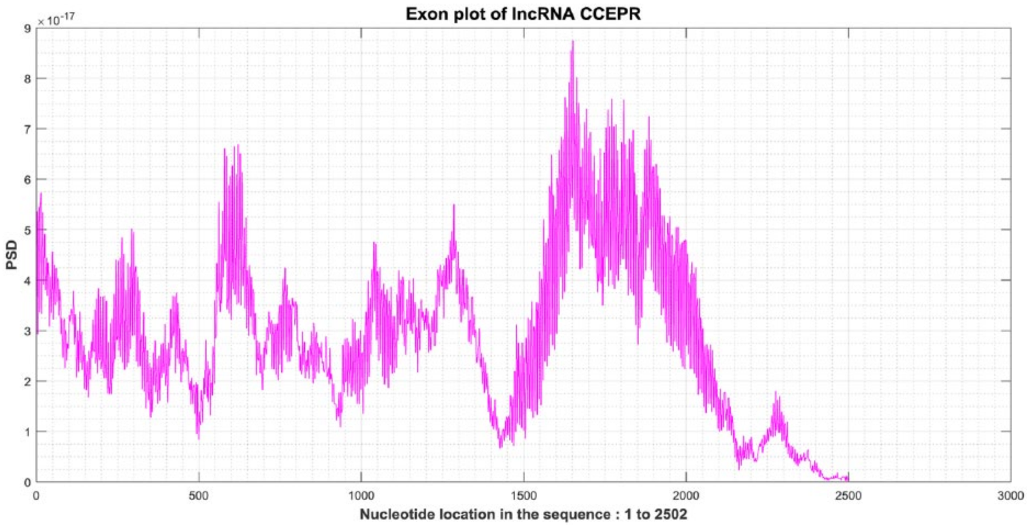

Figure 6. Exon plot of lncRNA CCEPR. Exon plot of lncRNA CCEPR with GenBank accession no. NR_131782.1. CCEPR indicates *Homo sapiens* cervical carcinoma expressed PCNA regulatory lncRNA; lncRNA, long noncoding RNA.

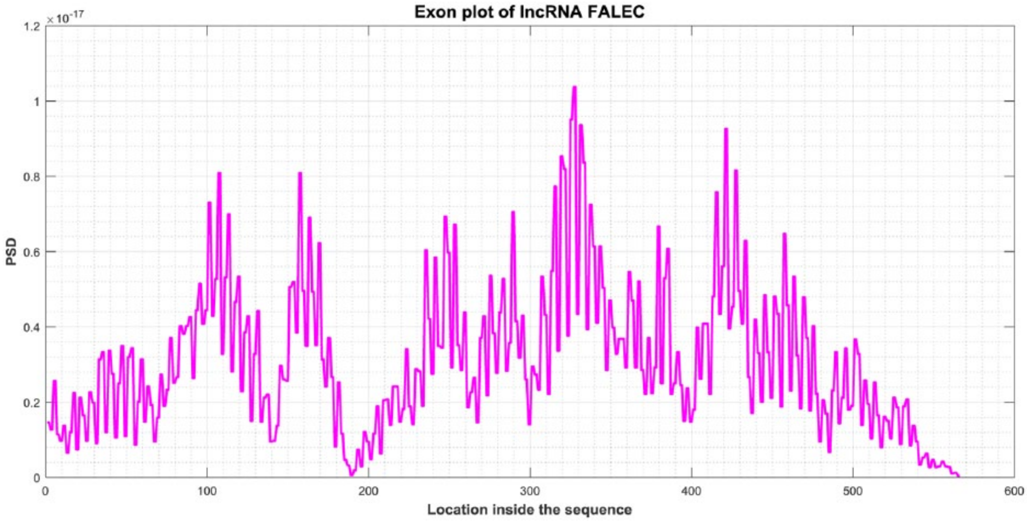

Figure 7. Exon plot of lncRNA FALEC. Exon plot of lncRNA FALEC with GenBank accession no. NR_051960.1. FALEC indicates focally amplified long noncoding RNA in epithelial cancer; lncRNA, long noncoding RNA.

A summary of the exon locations obtained in this work has been compared with wet lab results which are found in the NCBI records is given in Table 2. Column 2 of the table shows the name of the lncRNA sequence, column 3 gives the GenBank accession number, and column 4 gives the length of the sequence. Columns 5 and 6 show the start and end positions of exons as per the NCBI records which are results of wet lab methods, whereas columns 7 and 8 show the same obtained in this work. Columns 9 and 10 display the deviation in exon locations observed at the start and the end with reference to the NCBI records. Each of the NCBI records for these sequences site literature which ascertains that wet lab techniques have been used in the analysis of the lncRNAs. Sample references[29–31] are included in this article for the sequence CCEPR (https://www.ncbi.nlm.nih.gov/nuccore/NR_131782.1). Records corresponding to the NCBI accession numbers can be found for each of the sequences presented here.

Among sequences analyzed, CCEPR, BACE-AS1, CAHM, BGLT3, ABALON, DISC2, GHET1, NEAT1, HEIH, NEAT1, MALAT1, and NKILA have 1 exon each. Long noncoding RNAs FOXC2-AS1, DLEU1, HULC, KIAA0087, MHENCR, FALEC, and PRNT have 2 exons each. CISTR and HOTAIRM1 have 3 exons. The range of deviation of exon locations obtained in this work is around 36 to 100 nucleotides with respect to the exon ranges given in their NCBI records except for lncRNAs PRNT and HOTAIRM1. These 2 sequences show deviations of 180 and 175, respectively.

The G-C content of the sequences is computed relative to the total number of nucleotides and is shown in Table 3. It is found that of 20 sequences, 9 sequences have relative GC concentration more than 50%. The sequence with the highest GC content is lncRNA CAHM (GenBank accession number NR_037593.1) found to be 59.4684%. The average value of GC concentration is found to be 47.9865%.

Table 2. Exon location—comparison.

S. NO.	NAME OF THE LNCRNA	GENBANK ACCESSION NO.	LENGTH	EXON LOCATION				DEVIATION IN LOCATION	
				REFERENCE (GENBANK)		OBSERVED			
				START	END	START	END	START	END
1	CCEPR	NR_131782.1	2502	1	2502	1	2500	0	2
2	BACE-AS1	NR_037803.2	840	1	825	1	800	0	25
3	CAHM	NR_037593.1	903	1	886	1	850	0	36
4	BGLT3	NR_121648.1	1019	1	993	1	980	0	13
5	ABALON	KC505631.1	1903	1	1903	1	1900	0	3
6	DISC2	NR_002227.2	3892	1	3892	1	3880	0	12
7	GHET1	NR_130107.1	1913	1	1903	1	1890	0	13
8	HEIH	NR_045680.1	1681	1	1681	1	1680	0	1
9	NEAT1	NR_028272	3756	1	3735	1	3700	0	35
10	MALAT1 (TV1[a])	NR_002819.4.	8779	1	8779	1	8750	0	29
11	NKILA	NR_131157.1	2615	1	2598	1	2500	0	98
12	FOXC2-AS1	NR_125795.1	319	1	145	1	140	0	5
				146	319	140	319	6	0
13	DLEU1 (TV2[b])	NR_002605.2	2904	1	389	1	450	0	–61
				390	2904	500	2900	–110	4
14	HULC	NR_004855.2	500	1	182	1	230	0	–48
				183	484	250	480	–67	4
15	KIAA0087	NR_022006.1	4320	1	420	1	500	0	–80
				421	4320	500	4300	–79	20
16	MHENCR	NR_132417.1	793	1	158	1	200	0	–42
				159	793	200	793	–41	0
17	FALEC	NR_051960.1	566	1	306	1	200	0	106
				307	566	220	560	87	6
18	PRNT	NR_024267.1	2353	1	529	1	350	0	179
				530	2333	350	2300	180	33
19	HOTAIRM1	NR_038366.1	1052	1	295	1	300	0	–5
				296	564	300	740	–4	–176
				565	1044	740	1000	–175	44
20	CISTR (TV1[a])	NR_104332.1	856	1	221	1	200	0	21
				222	337	200	420	22	–83
				338	856	450	800	–112	56

Abbreviations: lncRNA, long noncoding RNA.
[a]TV1—transcript variant 1.
[b]TV2—transcript variant 2.

Table 3. G-C concentration.

S. NO.	LNCRNA	NCBI ACCESSION NO.	LENGTH	% GC CONCENTRATION
1	CCEPR	NR_131782.1	2502	41.9265
2	BACE1-AS	NR_037803.2	840	47.1429
3	CAHM	NR_037593.1	903	59.4684
4	BGLT3	NR_121648.1	1019	38.5672
5	ABALON	NR_131907.1	1903	56.5423
6	DISC2	NR_002227.2.	3892	37.7698
7	GHET1	NR_130107.1	1913	44.5896
8	HEIH	NR_045680.1	1681	58.5366
9	NEAT1	NR_028272.1	3756	47.9499
10	MALAT1 (TV1[a])	NR_002819.4	8779	40.3463
11	NKILA	NR_131157.1	2615	53.3461
12	FOXC2-AS1	NR_125795.1	319	54.8589
13	DLEU1 (TV2[b])	NR_002605.2	2904	38.6708
14	HULC	NR_004855.2	500	36
15	KIAA0087	NR_022006.1	4320	41.4583
16	MHENCR	NR_132417.1	793	55.9899
17	FALEC	NR_051960.1	566	56.8905
18	PRNT	NR_024267.1	2353	47.0463
19	HOTAIRM1	NR_038366.1	1052	50.7605
20	CISTR (TV1[a])	NR_104332.1	856	51.8692

Abbreviations: lncRNA, long noncoding RNA.
[a]TV1—transcript variant 1.
[b]TV2—transcript variant 2.

Sequence matching is done to locate START and STOP codon patterns. The sequences are checked for ATG and the alternative START codons too, namely, ATG, CTG, and GTC, and also for the STOP codons, TAA, TAG, and TGA. It is found that all the sequences have START codon patterns, but none of them have STOP codons.

Discussion

Period 3 property is a feature which has been combined with a proven signal processing–based automated method of detecting exons in coding DNA sequences.[16,17,19] It was observed in noncoding regions for prokaryotes and some viral and mitochondrial base sequences 2 decades ago,[24] and this concept is explored here. There are other signal processing–based automated methods[32,33] to detect exons in coding regions. One such method[32] detects short exons in DNA sequences by analyzing their structural properties, namely, DNA bending stiffness, disrupt energy, free energy, and propeller twist making use of the autoregressive model to arrive at linear prediction matrices for these features. The linear prediction matrices for the 4 features are combined to find the linear prediction coefficients from which the spectrum of the DNA sequence is estimated and exons detected based on the 1/3 frequency component. Short exons have also been detected by evaluating the complex wavelet transform of the structural features of DNA sequences.[33] The authors opted for period 3 property because of its proven robustness and relative simplicity.[16,17,19]

In a former work,[17] the authors have mapped out the exons for the sequence AF099922 (former GenBank accession number). The nucleotide sequence was taken from the gene SL1 trans-splice acceptor F56F11.4, which is a part of the F56F11 DNA sequence. Exons are located making use of the period 3 property, and the best approach was found to be the one using IIR peaking filters followed by DWT denoising using the Haar wavelet. This GenBank record for AF099922 is obsolete now, but it is mentioned here as the plot[17] is easy to relate to in the context of exon locations.

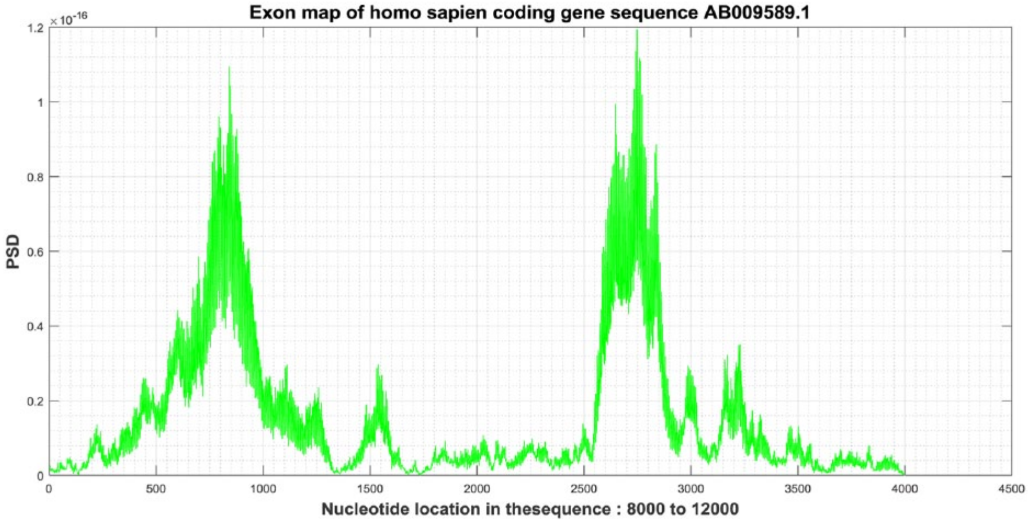

Figure 8. Sample exon plot of a coding sequence. The exon plot of 8000 to 11 000 locations of *Homo sapiens* gene for osteomodulin. GenBank accession number AB009589.1.

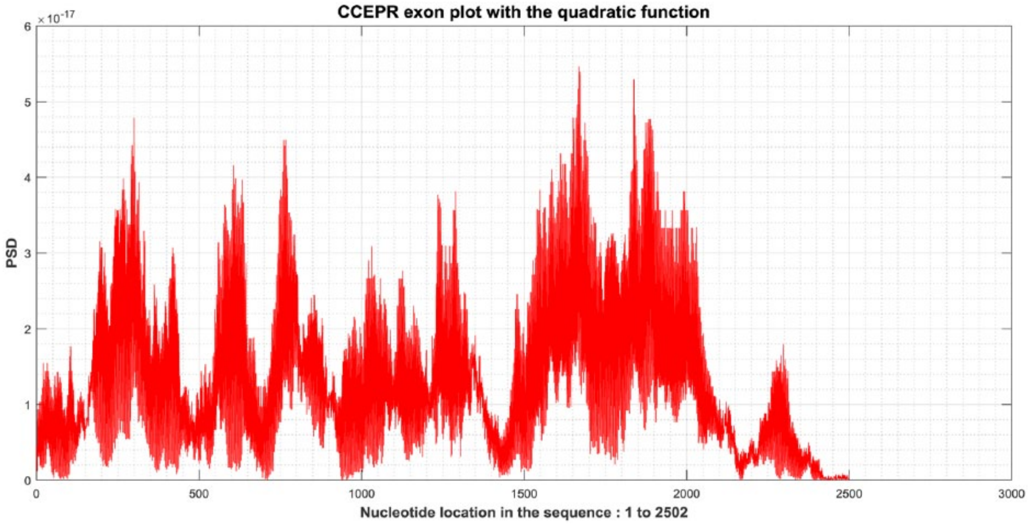

Figure 9. Sample exon plot of CCEPR with reduced computation technique. Exon plot of lncRNA CCEPR with GenBank accession no. NR_131782.1 obtained when the reduced computation technique is used. CCEPR indicates *Homo sapiens* cervical carcinoma expressed PCNA regulatory lncRNA; lncRNA, long noncoding RNA.

Figure 8 shows a sample exon plot of a coding region which is obtained by making use of period 3 property along with digital filtering. The sequence is that of *Homo sapiens* gene for osteomodulin with GenBank accession number AB009589.1. The region of the sequence considered is 8000 to 11 000. As per the GenBank record (https://www.ncbi.nlm.nih.gov/nuccore/AB009589.1), this region has 2 exons: 8524 to 9479 and 10 624 to 11 846. In the PSD plot (Figure 8), we find 2 energy peaks in the regions around 500 to 1350 and 2600 to 3500. As the segment considered here is 8000 to 12 000, the energy peaks are from around 8500 to 9350 and from around 10 600 to 11 500. These 2 energy peaks correspond to the exons in this particular segment (8000-11 000) of AB009589.1. The peak power is seen to be between 0.8×10^{-16} and 1×10^{-16}, and the half-power values are between 0.4×10^{-16} and 0.5×10^{-16}. The authors have adopted the same technique with minor variations in plotting the exon locations of lncRNAs.

While interpreting the exon plots, the following 2 points are to be considered:

- Even the most accurate of exon prediction algorithms do not pinpoint the exon locations to the precision of a nucleotide.
- While interpreting graphs of power spectrum, generally, the values below half power are not considered as signals.

As seen from the sample exon plots in Figures 6 and 7, period 3 property in conjunction with digital filtering techniques can be used to locate the exons in lncRNA sequences.

While computing the STFT for obtaining the spectrum, optimum window lengths are mandatory[17] in locating the exons from the sequences as they are in the case of coding DNA sequences. A reduced computation technique which makes use of a quadratic function (using only T and G sequences) was used to compute the spectrum in our former work.[17] This effectively reduces spectral noise with coding DNA sequences. When the same approach is used here with lncRNA sequences, it is found to insert spectral noise. The exon plot of CCEPR making use of this reduced computation technique is shown in Figure 9. The noise in the spectral plot is evident and the exon (1-2502; Figure 6) is not discernible from the noise. This means that T and G sequences are insufficient to represent the signal spectrum unlike the case of coding DNA sequences. This indicates the difference in spectral properties of coding and noncoding sequences.

Although certain lncRNA sequences contain exons, their coding ability is still not been confirmed yet due to a variety of reasons. Most of the lncRNAs were found to have low GC concentration when compared with coding sequences.[12] This also suggests poor coding capacity. G-C concentration of 20 lncRNA sequences analyzed is given in Table 3. Long noncoding RNAs, namely, CAHM, ABALON, HEIH, NKILA, FOXC2-AS1, MHENCR, FALEC, HOTAIRM1, and CISTR, have G-C concentrations above 50%. This could imply protein-coding capacity. But the lack of introns and the lack of STOP codons suggest otherwise. Computational analysis of functional lncRNA has been reported to reveal lack of protein-coding capacity and also was found to have similarities with 3′ untranslated regions. Long noncoding RNA sequences have been found to possess low G-C content and scantiness of introns. In previous studies, open reading frames (ORFs) were detected in some lncRNA sequences, but they have a poor start codon and ORF contexts which would make it unlikely for these lncRNAs to be protein coding.[14] The lncRNAs analyzed in this work have very short or practically nonexistent introns. These sequences have START codon corresponding to the exon locations mentioned in the reference database, and they do not have any of the STOP codon patterns within them (UAA, UAG, or UGA). Although such a stretch after the START codon might appear to be an ORF, there are no STOP codons, which would make it unlikely for the lncRNA to code for peptides. Most of the lncRNAs have been found to be spliced (98%), and they exhibit a striking bias toward 2 exon transcripts. About 42% of lncRNAs have only 2 exons as against 6% of the protein-coding genes.[12] Long noncoding RNAs with more than 2 exons are also included in the study to establish the robustness of the algorithm in locating exons.

Conclusions

Exons of human lncRNA sequences are predicted making use of the period 3 property, a widely accepted approach for predicting exons in coding DNA sequences. The IIR antinotch filter picks the spectral component at $2\pi / 3$, and denoising of the exon map with DWT filter bank refines it as the noise is seen to occupy the HF part of the spectrum. For obtaining the spectrum, the choice of the window used for computing the STFT is found to be crucial just as the case with coding DNA sequences. Window is to be selected based on the length of the sequence used. The reduced computation technique which makes use of the T and G binary sequences alone to compute the spectrum was found to suppress spectral noise with coding DNA sequences. But this is not so with lncRNA sequences. In this case, the quadratic function introduces spectral noise. Thus, it is clear that T and G binary sequences alone cannot represent the spectrum amply as in the case of coding DNA sequences. This indicates that the spectral properties of lncRNA sequences are different from those of coding DNA sequences. Long noncoding RNA sequences may contain information in their spectrum which could be made use of in further study. Comparing the exon plots for lncRNA sequences (Figures 6 and 7) with that of the exon plot of a coding DNA sequence (Figure 8), it is clear that the algorithm based on period 3 property followed by digital filtering techniques can effectively be extended to locate exons in lncRNA sequences.

Period 3 property which picks exons from coding DNA sequences has been used successfully in identifying genes[20] from DNA sequences. On parallel logic, it is to be investigated whether the technique used in locating exons within lncRNAs can be adapted to identify lncRNAs themselves. This could be yet another area in which this work could be taken forward. However, this has multiple constraints as the functional implications of the exons present in lncRNAs have not been unveiled yet.

There are many computational methods to predict functional features of lncRNA that are listed in literature.[34,35] The former[34] details a method to predict lncRNA functions based on a coding-noncoding gene co-expression network. Several in silico methods for the prediction of function and characterization of lncRNAs are outlined by Signal et al.[35] Computational prediction of lncRNA function using tissue-specific co-expression and from the expression of genes in various genes in different species is detailed by Perron et al.[36]

The works mentioned above are just a couple of examples of methods to predict functions of lncRNAs, it is not an all inclusive reference list.

The study presented here is not a method to identify lncRNA nor is it sufficient to predict regulatory properties/functions of lncRNA. The signal processing technique that is widely used to locate exons in coding genes is used here to detect exons in lncRNA. The similarity/differences of lncRNA sequences with sequences of coding genes in terms of their spectral properties have been highlighted. Such a study which predicts exons in lncRNAs using signal processing principles was not found in literature. The novelty of the study is this very fact, and hence, a comparative study of this work with existing techniques is not presented. Signal processing methods are inherently easy to implement and robust. The authors expect that this novel approach to analyze lncRNA would be helpful in bringing to light many of their sequence and spectral properties. The future direction of this work would be to explore the possibility of predicting the regulatory functions of lncRNA from their sequence properties or by frequency domain analysis of sequences.

Acknowledgements

I, Tina P George, would like to thank the authorities of the Department of Electronics, Cochin University of Science And Technology (CUSAT) for providing access to the resources to carry out this work under the guidance of the second author, Dr Tessamma Thomas.

Author Contributions

TPG and TT conceived and designed the experiments, agree with manuscript results and conclusions, jointly developed the structure and arguments for the paper, made critical revisions and approved final version, reviewed, and approved the final manuscript. TPG collected and analyzed the data, wrote the first draft of the manuscript, and contributed to the writing of the manuscript.

Disclosures and Ethics

As a requirement of publication, authors, Tina P George and Tessamma Thomas, will provide to the publisher at the time of publication, signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. The external blind peer reviewers report no conflicts of interest.

REFERENCES

1. Ponting CP, Belgard TG. Transcribed dark matter: meaning or myth? *Hum Mol Genet*. 2010;19:R162–R168. doi:10.1093/hmg/ddq362.
2. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet*. 2016;17:47–62.
3. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol*. 2013;20:300–307.
4. Johny T, Kung Y, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. *Genetics*. 2013;193:651–669.
5. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell*. 2009;136:629–641.
6. Boon RA, Jae N, Holdt L, Dimmeler S. Long noncoding RNAs: from clinical genetics to therapeutic targets? *J Am Coll Cardiol*. 2016;67:1214–1226.
7. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet*. 2014;30: 439–452.
8. Mercer TR, Dinger ME, Mattick JS. Long noncoding RNAs: insights into functions. *Nat Rev Genet*. 2009;10:155–159.
9. Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. *RNA Biol*. 2013;10:925–933.
10. Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol*. 2008;4:e1000176. doi:10.1371/journal.pcbi.1000176.
11. Calabrese JM, Magnuon T. Roles of long non-coding RNAs in X-chromosome inactivation. In: *Molecular Biology of Long Non-Coding RNAs*. 2013. doi:10.1007/978-1-4614-8621-3_3.
12. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22:1775–1789.
13. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res*. 2012;22: 1760–1774.
14. Niazi F, Vlaladkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. *RNA*. 2012;8:825–843.
15. https://www.ncbi.nlm.nih.gov/nucleotide.
16. Vaidyanathan PP, Yoon BJ. Digital filters for gene prediction applications. Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers. Monterey, CA, 2002;1:306–310.
17. George TP, Thomas T. Discrete wavelet transform de-noising in eukaryotic gene splicing. *BMC Bioinformatics*. 2010;11:S50. doi:10.1186/1471-2105-11-S1-S50.
18. https://genome.ucsc.edu/.
19. Anastassiou D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*. 2005;16:1073–1081.
20. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput Appl Biosci CABIOS*. 1997;13:263–270.
21. Kakumani R. Prediction of protein-coding regions in DNA sequences using a model-based approach. Paper presented at: International Symposium on Circuits and Systems IEEE; May 18-21, 2008. doi:10.1109/ISCAS.2008.4541818.
22. Trifonov EN, Sussman JL. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci USA*. 1980;77:3816–3820.
23. Herzela H, Trifonovb EN, Weissa O, Groβec I. Interpreting correlations in biosequences. *Physica A*. 1998;249:449–459.
24. Li W. The study of correlation structures of DNA sequences: a critical review. *Comput Chem*. 1997;21:257–271.
25. Proakis JG, Manolakis DK. *Digital Signal Processing*. 4th ed. New Delhi: Pearson; 2006.
26. Oppenheim AV, Schafer RW. *Signal Processing Series: Discrete-Time Signal Processing*. 3rd ed. Prentice Hall USA; 2009.
27. MATLAB Release 2016b. The MathWorks Inc, Natick, Massachusetts, United States.
28. Soman KP, Ramachandran KI. *Insights Into Wavelets: From Theory to Practice*. 2nd ed. Prentice-Hall Of India Pvt. Ltd. 2005.
29. Peng W, Fang H. Long noncoding RNA CCHE1 indicates a poor prognosis of hepatocellular carcinoma and promotes carcinogenesis via activation of the ERK/MAPK pathway. *Biomed Pharmacother*. 2016;83:450–455.
30. Yang M, Zhai X, Xia B, Wang Y, Lou G. Long noncoding RNA CCHE1 promotes cervical cancer cell proliferation via upregulating PCNA. *Tumour Biol*. 2015;36:7615–7622.
31. Choy KW, Wang CC, Ogura A, et al. Genomic annotation of 15,809 ESTs identified from pooled early gestation human eyes. *Physiol Genomics*. 2006;25:9–15.
32. Song NY, Yan H. Short exon detection in DNA sequences based on multifeature spectral analysis. *Eurasip J Adv Sig Pr*. 2011;2011:780794.
33. Provazník I, Kubicová V, Škutková H, et al. Detection of short exons in DNA sequences using complex wavelet transform of structural features. Proceedings IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS); December 2-4, 2012; Washington, DC. doi:10.1109/GENSIPS.2012.6507740.
34. Zhao Y, Luo H, Chen X, Xiao Y, Chen R. Computational methods to predict long noncoding RNA functions based on co-expression network. *Meth Mol Biol*. 2014;1182:209–218.
35. Signal B, Gloss BS, Dinger ME. Computational approaches for functional prediction and characterisation of long noncoding RNAs. *Trends Genet*. 2016; 32:620–637.
36. Perron U, Provero P, Molineris I. In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression. *BMC Bioinformatics*. 2017;18:144. doi:10.1186/s12859-017-15.