Prediction of Protein Coding Regions in the DNA Based On Period-3 Property of Nucleotides

Vikrant Singh Tomar, Dipesh Gandhi, Vijaykumar Chakka

Abstract—An important problem in genomic signal processing is the prediction of gene locations in a genomic sequence. This is a big issue as the length of a DNA sequence can be of the order of billions of nucleotides. A period-3 behavior of nucleotides has been observed in the protein coding regions (exons) of DNA sequences. This $f = 1/3$ $(or\ \omega = 2\pi/3)$ property in the exons of eukaryotic gene sequences enables signal processing methods to identify these regions, which in-turn helps in predicting the *presence* of genes in the DNA strand. Existing techniques are not effective in detecting small exons. This paper presents a number of new techniques in order to minimize the power in introns (non-coding regions), and predicting the genes more robustly. Through a series of experiments, it is demonstrated that these techniques are able to detect smaller exons, and hence gene locations to a high degree of accuracy.

Keywords—DNA, gene prediction, 3-periodicity, anti-notch filter, minimum variance spectrum estimation, teager energy operator.

I. Introduction

GEnomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. For this reason digital signal processing (DSP) techniques offer a great promise in analysing genomic data. Such sequences, like DNA and proteins, are typically represented by character strings, in which each character is a letter of an alphabet. In case of DNA, the alphabet is of size 4 (for proteins, it's 20) and consists of the letters A, T, C and G. For example,

$$...ATCGCTGATAGGATGGTTAACC...$$

The specific problem targeted in this paper is 'gene prediction' which refers to locate the protein-coding regions (exons) of genes in a long DNA sequence. The problem constitutes of two major steps, first is to map the alphabetical DNA sequence onto a numerical sequence and then apply DSP algorithms to predict the gene locations. For most prokaryotic[1] DNA sequences, the problem is to determine which open read frames (ORF), in the given sequence are really coding for proteins. For eukaryotic[2] DNA sequences, the problem is to determine how many exons and introns are there in the given sequence, and what are the exact boundaries between the exons and introns.

It has been observed by many researchers that the protein-coding regions of DNA sequences exhibit a period-3 behaviour due to codon structure [1]–[4]. Thus Fourier analysis can be utilized to detect the probable coding regions in DNA sequences. In [3], Tiwari *et al.* demonstrated, by computing the amplitude profile of its spectral component, that a sharp peak at angular frequency $\omega = 2\pi/3$ is prominent in the power spectrum of a DNA single strand, as shown in Fig. 1. Since Tiwari [3], there has been a great deal of work done in applying signal processing and statistical methods to DNA. The antinotch filter proposed in [5] has been so far one of the most promising algorithms in gene filtering. However, none of the existing work has focused on attenuating the harmonics of frequency $2\pi/3$ which contribute to spurious peaks in the estimate of the power spectral density (PSD) of the DNA sequence leading to false positives for gene detection. For these reasons, such techniques methods fail to detect or differentiate the short or closely spaced exons.

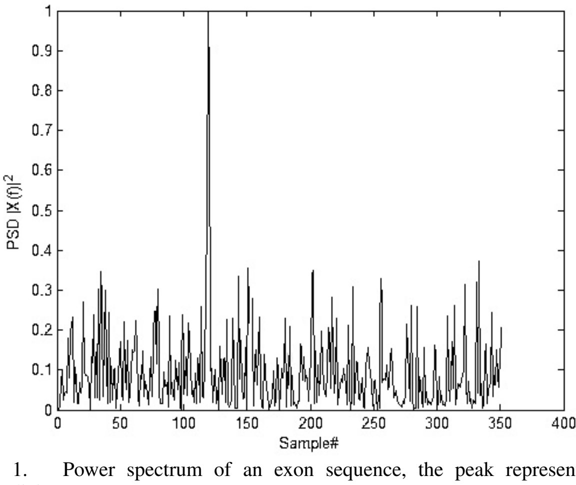

Fig. 1. Power spectrum of an exon sequence, the peak represents 3-periodicity.

In this paper, several techniques are applied in order to address the limitations of signal processing techniques as applied to gene prediction. First we try to improve the aforementioned antinotch filter using a Blackmanharris windowing function, instead of traditional rectangular window. Such a window helps reduce the spectral leakage, increases the stop band attenuation, and thus smoothens the filter output. However,

V. S. Tomar is with the Department of Electrical and Computer Engineering, McGill University, Monteral, QC, Canada e-mail: vikrant.tomar@mail.mcgill.ca.

D. Gandhi is a researcher at ST Microelectronics ltd., India.

Dr. V. Chakka is with Dhirubhai Ambani Institute of Information and Communication Technology.

[1] Single-celled organisms like bacteria do not have a nucleus, and are called Prokaryotes. Their DNA just resides in the cell and genes do not have intron regions.

[2] Higher organisms (worms, insects, plants, mammals, etc.) have cells with nucleus, and are called Eukaryotes. These have the DNA residing in the nucleus and their genes have both intron and exon regions. An exception is the red blood cell which has no nucleus. Cells also have a small quantity of DNA in the mitochondria.

it does not improve on the fact that the antinotch filter allows the harmonics of frequency $2\pi/3$ to pass, which contributes to false peaks strength. To this end, a harmonic suppression comb filtering technique is proposed which removes such harmonics by introducing zeros and poles at specific positions. The comb filtering technique perfectly attenuates the higher order harmonics of $2\pi/3$. However, the conjugate frequency components, i.e., at $-2\pi/3$, are still not suppressed. To improve on this, an adaptive Minimum Variance Spectrum Estimator is proposed which minimizes the out-of-band power in the side lobes and thus suppresses the conjugate harmonics as well. Furthermore, the variable length Teager energy operator (VTEO) is utilized in lieu of basic PSD calculation, to optimize the energy calculation in the filtered output [6].

Rest of the paper is organized as follows. Section II and III briefly discuss the required basic background of Genomic Signal Processing. Section IV focuses on the numerical mapping of the alphabetical genomic sequences. Section V elaborates on the work that has been done in the area. In Section VI new proposed techniques for improvement in gene prediction are discussed. Section VII contains the discussions. Section VIII concludes the work.

II. A Brief Background

A very short introduction of the required biological background, such as DNA sequences, genes, exons, *etc.*, which feature repeatedly in this document, is provided in this Section. Interested reader should refer to [4], [5] for detailed information.

A. DNA

DNA – deoxyribo nucleic acid – is the part of a cell that contains and controls all the genetic information, and the thousands of genes necessary to reproduce it. DNA is a double-stranded, helix shaped, long and linear biopolymer formed from many linked, basic structural units called nucleotides. The four nucleotides in DNA contain the bases: Adenine, Thymine, Cytosine, and Guanine, which are designated by the letters A, T, C, and G, respectively. A DNA double strand contains two complementary strands connected in *Watson-Crick base pairing* through weak hydrogen bonds [7], [8]. The two strands run in opposite directions, as illustrated in Fig. 2. Typically in any given region of the DNA molecule, at most one of the two strands is active in gene expression.

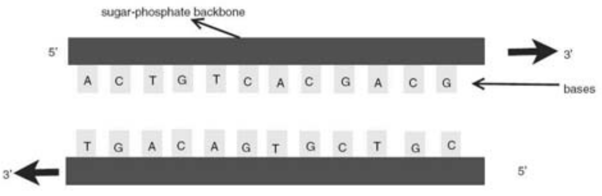

Fig. 2. The two complementary DNA strands

B. Genes, and Exons and Introns

A DNA sequence can be discriminated into *genes* and *inter-genic* spaces. Within genes there are two types of sub regions called the *exons* and *introns*. Exons are the regions which contributes toward the protein synthesis, and introns are those which do not.

C. Proteins and Protein Synthesis

A protein is composed of amino acids linked together in a particular order specified by a gene sequence. Protein functions are ultimately determined by the DNA character string because it is the digital information in the DNA nucleotide sequence that determine the amino acid sequence. Each protein character string is generated based on information in the exons in genes. The protein synthesis process is expressed schematically in Fig.3.

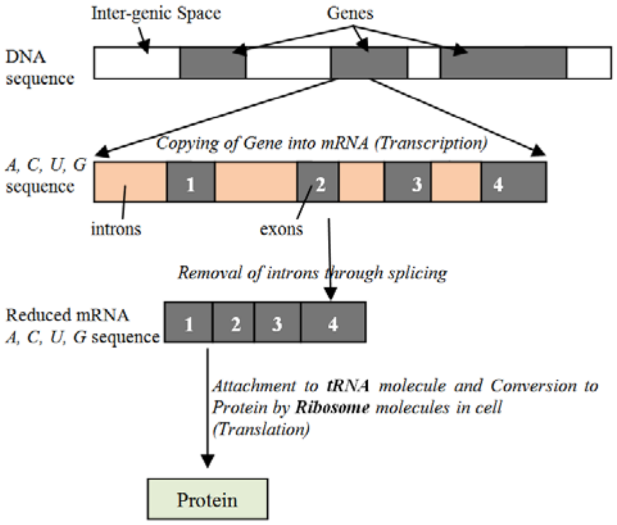
Inter-genic Space
Genes
DNA sequence
Copying of Gene into mRNA (Transcription)
A, C, U, G sequence
1
2
3
4
introns
exons
Removal of introns through splicing
Reduced mRNA A, C, U, G sequence
1
2
3
4
Attachment to tRNA molecule and Conversion to Protein by Ribosome molecules in cell (Translation)
Protein

Fig. 3. Synthesis of protein from DNA. The gene is first copied into a single stranded chain called the messenger RNA or mRNA molecule. The introns are then removed from the mRNA by a process called *splicing*. The spliced mRNA can be divided into groups of three adjacent bases. Each triplet is called a *codon*. Evidently there are 64 possible codons. Each codon instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein.

The next section discusses the two properties of DNA sequences that make the signal processing techniques very effective in exploiting the gene positions.

III. Properties and DNA sequence

This section discussed two important properties of DNA sequences that make them amenable to signal processing techniques. This section also discusses the DNA sequence that is used in various experiments in this work.

A. 3-Periodicity

For most of DNA sequences, one of the principal features is the periodic 3-nucleotide pattern in exons, whereas no such

International Science Index 74, 2013 waset.org/publications/17163

periodicity is detected in introns. This has been a known phenomenon for eukaryotic genes [9], [10]. DNA periodicity in exons is determined by codon usage frequencies. At this point, it is essential to differentiate between DNA periodicity itself and the length of the period equal to 3. Periodicity itself is a result of certain combinations of codons with different frequencies typical for a species. The length of period equal to 3, instead, is caused by the triplet nature of genetic code [10]. This periodicity reflects correlations between nucleotide positions along coding sequences which is caused by the asymmetry in base composition at the three coding positions [11]. This property makes the exon regions tractable by Fourier analysis, as shown in Fig. 1.

B. $1/f$ Noise

$1/f$ noise or the long range correlation is a signal or process with a frequency spectrum such that the power spectral density (PSD) follows:

$$S(f) \propto \frac{1}{f^{\beta}}, \tag{1}$$

where $0 < \beta < 2$. It is also called pink noise as it is white noise $(1/f^0)$ and red noise $(1/f^2)$. It has been found that in DNA sequence correlation exists between the base sequences at large distance as well. H. Voss has established that the PSD follows a $1/f^{\beta}$ law, where β depends on the gene [12], [13]. Hence, it is important to choose the window size N, while taking a Fourier Transform, such that the peak at frequency $\omega = 2\pi/3$ is visible over the long-range correlation.

C. DNA Sequence Data

For comparison purpose, throughout this paper, the commonly used sequence 'F56F11.4' (indices 2858700-2865376) from 'Chromosome III' of the organism 'C.elegans' is used which has five exons, as shown in Table I. Use accession number AF 099922 in [14], [15] to download the gene.

TABLE I
EXON LOCATIONS IN CHROMOSOME III OF C.ELEGANS

exon #	Start location	End location
1	1000	1111
2	2600	2929
3	4186	4449
4	5537	5716
5	7327	7677

IV. NUMERICAL REPRESENTATION OF A DNA CHARACTER SEQUENCE

The simplest form of representation of DNA sequences is the FASTA representation that uses the alphabet A, C, T, G showing a DNA sequence as a string of characters. But to apply signal processing algorithms one needs to assign numerical values to these characters. An ideal mapping should preserve the period-3 property of the DNA sequence, which is possible only through symmetric mapping [9]. Many rules have been proposed for this purpose, some of the important ones are briefly discussed here.

- *Voss Mapping [13]:* For each symbol $\alpha \in \{A, C, T, G\}$, a binary indicator sequence $x_\alpha[n]$ is defined. In the indicator sequence $x_\alpha[n]$, $'1'$ indicates the presence of base α and $'0'$ indicates its absence. Voss mapping is a four dimensional mapping, because each base in the DNA sequence is represented by a four dimensional vector composed of $'0'$s and $'1'$s. The number of $'1'$s in any vector is exactly one. For example, given a DNA sequence $x[n] = ATTGTCACTCGG....$ the indicator sequence would be as given below:

  $$\begin{aligned} &x_A[n] : \{1,0,0,0,0,0,1,0,0,0,0,0\},\\ &x_C[n] : \{0,0,0,0,0,1,0,1,0,1,0,0\},\\ &x_G[n] : \{0,0,0,1,0,0,0,0,0,0,1,1\},\\ &x_T[n] : \{0,1,1,0,1,0,0,0,1,0,0,0\},\end{aligned}$$

  where $x_A[n] + x_C[n] + x_T[n] + x_G[n] = 1$ and 'n' represents the base index.
- *Complex Mappings:* It has been shown that instead of a binary digital signal, a complex signal of the form $x[n] = ax_A[n] + cx_C[n] + tx_T[n] + gx_G[n]$, where $\{a, t, c, g\}$ are complex numbers, would be more advantageous from signal-processing perspective. Depending upon the chosen values of variables $\{a, t, c, g\}$ several variants of complex mapping are available [4], [9], [16].
- *Optimized Multipliers:* In [4] and [17], Anastassiou has shown, by means of defining an optimization problem, that the performance of the optimized spectral content measure $|aA + tT + cC + gG|^2$ is significantly superior to that of the traditional one from, $|A|^2 + |T|^2 + |C|^2 + |G|^2$, in terms of their capabilities to distinguish between coding and non-coding regions in DNA sequences. One such example of a, t, c and g is:
  $a = 0.10 + 0.12j; \quad t = -0.20 - 0.30j$
  $c = 0; \quad g = 0.45 - 0.19j$

V. EXISTING APPROACHES TO GENE PREDICTION

In this Section, various existing signal processing techniques for gene prediction are briefly discussed.

A. Sliding Window Discrete Fourier Transform

The periodicity of 3 suggests a peak at $k = N/3$ in exons in the Discrete Fourier Transform (DFT) spectrum of a DNA sequence, where N is window size, so that calculation of DFT at that point should be sufficient (that is 117th sample in frequency spectrum for window size 351, see Fig. 1. The window can then be slided by one or more points. If the sequence is from a gene, the DFT spectrum $X_A[k], X_C[k], X_T[k]$ and $X_G[k]$ of these signals each indicate a peak at $\omega = 2\pi/3$. Hence the power spectral density (PSD) for the DNA sequence, $S[k]$ would indicate a peak at $\omega = 2\pi/3$.

$$S[k] = \sum_m |X_m[k]|^2, \tag{2}$$

where $m \in \{A, T, C, G\}$. For the complex representation, $|X[k]|^2$ can be used in lieu of $S[k]$. Fig. 4 depicts the output of

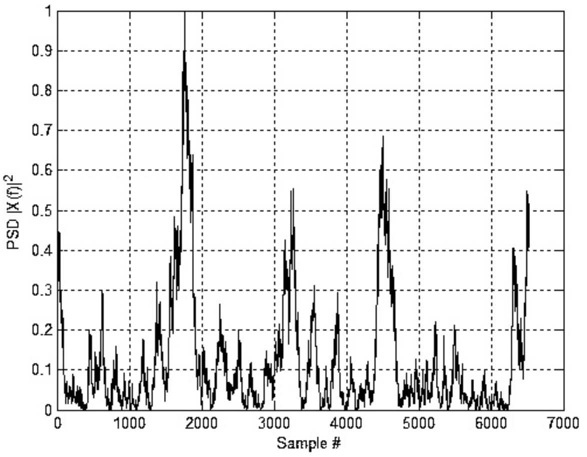

Fig. 4. Output of sliding window DFT technique

sliding window DFT for window length of 351 samples with complex mapping. The peaks in the spectrum implies a strong correlation at a period of 3 among the nucleotides at these positions, and hence the positions of coding regions. In order to differentiate the spectrum of exons and introns, only peaks having power greater than the threshold P are considered to be representing exon regions [3], where,

$$P = S(2\pi/3)/S_{avg} \tag{3}$$

S_{avg} is the frequency averaged PSD for the given window, and $S(2\pi/3)$ represents the PSD at $\omega = 2\pi/3$.

B. Antinotch Filter

An infinite impulse response (IIR) digital filter with a magnitude response showing a sharp peak $\omega = 2\pi/3$, known as *antinotch filter* can be used to filter the periodicity of 3 in a DNA sequence [18]. The magnitude of frequency response of such a filter is depicted in Fig. 5. The binary indicator sequences $x_\alpha[n]$ where $\alpha \in \{A, C, T, G\}$ are applied to the input of such an antinotch filter producing output $y_\alpha[n]$. The expression

$$Y[n] = \sum_m |y_m[n]|^2, \tag{4}$$

where $m \in \{A, T, C, G\}$, can be used as a feature to predict the coding regions in a DNA sequence. The output of such a filter is shown in Fig. 6.

C. Average Magnitude Difference Function

The AMDF for a bandpass-filtered binary indicator DNA sequence $x[n]$ as a function of the period $k = 3$ is defined as:

$$AMDF[k] = \frac{1}{N}\sum_{n=1}^{N} |x[n] - x[n-k]|, \tag{5}$$

where N is the window size. The deep null produced by the AMDF at $k = 3$ can be used for exon prediction. A linear combination of the AMDF outputs for each of the four indicator sequences gives the final feature values for exon prediction [19]. The output of such a filter for window length = 117 samples is shown in Fig. 17.

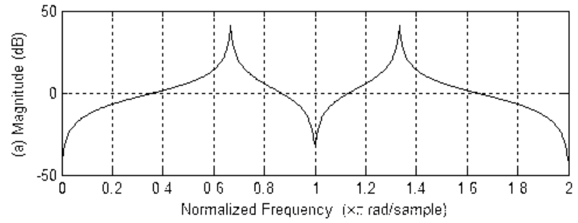

Fig. 5. Magnitude of frequency responses of IIR antinotch filter

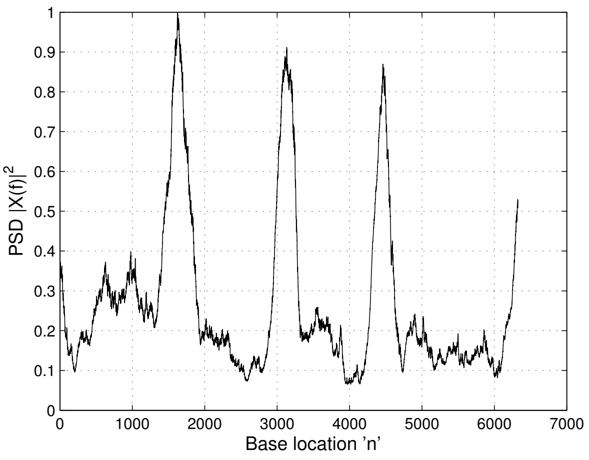

Fig. 6. Filter output of IIR Antinotch Filter

Having discussed the existing approaches for gene prediction, in the next section we discuss the proposed approaches and how those outperform the existing ones.

VI. Proposed Algorithms

This section describes the proposed algorithms for gene filtering. Four different techniques are described, each successively building upon the shortcomings of the previous one.

A. Improved Antinotch Filter

We improved the antinotch filter, discussed in Section V-B, by introducing more attenuation to the unwanted frequencies. In order to boost the $2\pi/3$ component, we introduced poles at an angle $\omega = 2\pi/3$ with magnitude close to 1, *i.e.*, 0.998. Furthermore, in order to smoothen the output, we used Blackmanharris windows instead of the traditional rectangular window. The window length was taken same as that of the rectangular windows, *i.e.*, 351 samples. Fig. 7 (a) shows the frequency responses of the proposed antinotch filter, and (b) shows the designed windowing function for smoothing. Fig. 8 and Fig. 9 show the output of the designed antinotch filter without and with smoothing window, respectively. Clearly, the filter has boosted the peaks in exon regions and suppressed the $1/f$ noise as compared to the existing antinotch filter.

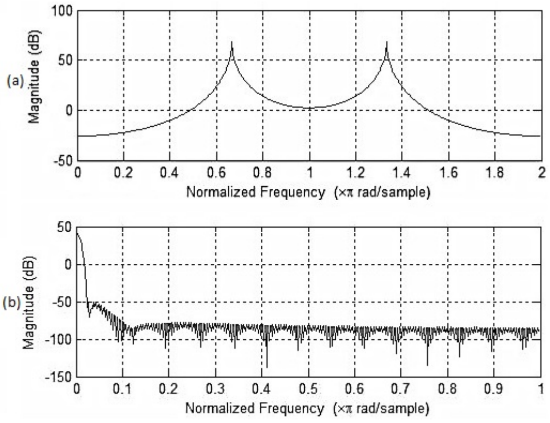

Fig. 7. (a) Magnitude of frequency response of the proposed antinotch filter. (b) Magnitude of the designed windowing functions for smoothing.

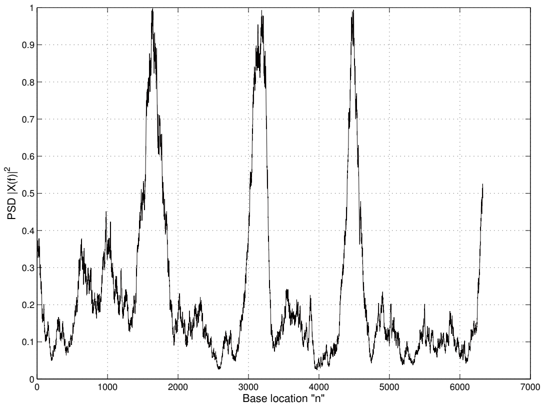

Fig. 8. Output of improved Antinotch filter.

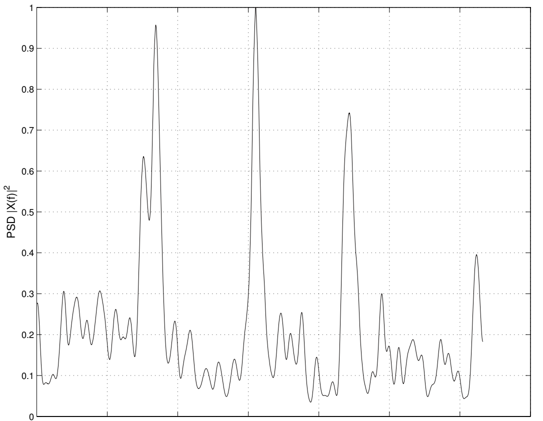

Fig. 9. Output of improved Antinotch filter with smoothing window.

Smoothing has further reduced the $1/f$ noise and thus made the peaks more visible.

Both the antinotch filters, discussed here and in Section V-B, allow passing the harmonics of the frequency $2\pi/3$ which also contribute to the peak strength. Thus outputs of these filters do not give a true measure of the 3-periodicity. To suppress the harmonics, notches at the harmonic frequencies can be created, as discussed next.

B. Harmonic Suppressing Comb Filtering Technique

To suppress the harmonic frequencies, a harmonics suppressing comb (HaSCo) filter was designed having dominant zeros at the multiples of frequency $2\pi/6$, except $2\pi/3$, and dominant pole at $2\pi/3$, as depicted in Fig. 10 (b). Such a filter perceptibly suppresses the samples of harmonic frequencies of $2\pi/3$ while allowing the samples of frequency $2\pi/3$ to pass.

Fig. 10(a) shows frequency response of the comb filter on a digital frequency scale of 24 KHz. The peak at 16 KHz is evidence of the pole at angular frequency $2\pi/3$ radians. A pole-zero plot of designed filter is illustrated in Fig. 10 (b) having poles of magnitude 0.898, 0.898, 0.998, 0.898 at angular frequencies ω=0, $2\pi/6$, $2\pi/3$, π radians, respectively, and zeros of magnitude 0.998, 0 .998, 0.898, 0.998 at angular frequencies ω=0, $2\pi/6$, $2\pi/3$, π radians, respectively. Thus the transfer function of the above filter can be written as,

$$\begin{aligned} H(z) =& \frac{1 - 2R_2 \cos\theta_2 z^{-1} + R_2^2 z^{-2}}{1 - 2R_1 \cos\theta_2 z^{-1} + R_1^2 z^{-2}} \cdot \\ & \prod_{i=1}^{3} \frac{1 - 2R_1 \cos\theta_i z^{-1} + R_1^2 z^{-2}}{1 - 2R_2 \cos\theta_i z^{-1} + R_2^2 z^{-2}}, \end{aligned} \quad (6)$$

where $\theta_i \in \{0, 2\pi/6, \pi\}$ for $i = 1, 2, 3$ and $R_1 = 0.998$ and $R_2 = 0.898$. Here it is important to note that it is enough to suppress the harmonics of $2\pi/6$, and suppression of the higher period harmonics is not necessary because their contribution in a window size of only 351 samples is negligible.

Fig. 10 (c) shows output of designed comb filter. It is apparent from the figure that the proposed comb filter suppresses the harmonics which are allowed to pass in antinotch filter, as shown in Fig. 6.

C. Minimum Variance Spectrum Estimator

Up to this point, we have been considering non-parametric techniques for estimating the power spectrum of a process. Relying on the DTFT of an estimated autocorrelation sequence, the performance of these methods is limited by the length of the data record. Furthermore, all of these filters are data independent. As a result, *when a random process contains a significant amount of power in frequency bands within the side lobes of the bandpass filter, leakage through the side lobes will lead to significant distortion in the power estimates*. Therefore, a better approach would be to allow the filter to be data adaptive so that the filter may be designed to be *optimum* in the sense of rejecting as much out-of-band signal power as possible, while giving a distortion-less output at the central frequency. The motivation for this approach can also be seen by looking to the problem of harmonic suppression in the aforementioned filters. Though, in the previous section, we have suppressed the higher order harmonics frequencies using the comb filter but still the conjugate frequency component, *i.e.*, $-2\pi/3$ or $4\pi/3$ could not be suppressed due to the complex conjugate pairing nature of zeros and poles, as illustrated

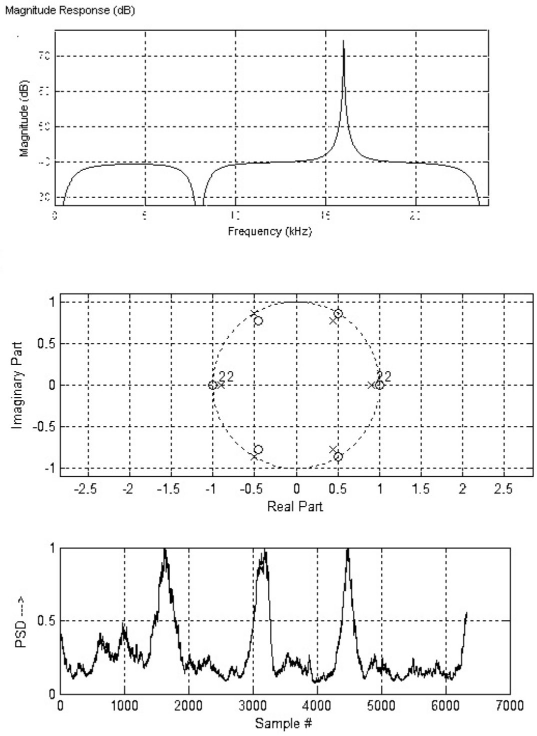

Fig. 10. The harmonics suppressing filter. (a) Frequency Response, (b) pole-zero plot, (c) Output.

in Fig. 5(a). In this Section, we apply the minimum variance (MV) method for spectrum estimation of the DNA sequence. Minimum Variance method gives us ability to minimize the power in the side lobe frequencies thus maximizing the power in main lobe [20].

In the MV method, the power spectrum is estimated by filtering a process with a bank of narrow band bandpass filters. As in the problem at hand we are concerned only with one frequency component, i.e. $2\pi/3$, hence now onwards we will limit our discussion to one frequency only. The MV spectrum estimation technique described in this Section involves the following steps:

1) Design a bandpass filter $g(n)$ with center frequency $\omega = 2\pi/3$ so that the filter rejects the maximum amount of out-of-band power while passing the component at frequency ω with no distortion.
2) Filter the DNA sequence $x(n)$ with the filter and estimate the power in each output process $y(n)$.

$$Hence, \qquad G(e^{j\omega}) = 1 \tag{7}$$

$$or, \quad \sum_{n=0}^{p-1} g(n)e^{-j\omega n} = 1 \tag{8}$$

where p is the filter tap length or the window size, *i.e.*, the number of samples, and $g(n)$ is the impulse response of the MV filter with band-pass frequency ω. Also,

$$E[|Y(n)|^2] = \mathbf{g}^H \mathbf{R_x} \mathbf{g} \tag{9}$$

where, $\mathbf{g}^H$ represents the Hermitian (complex conjugate) of vector $\mathbf{g}$ and $\mathbf{R_x}$ is the $p \times p$ autocorrelation matrix of the samples (**x**) in the current window.

Our goal is to minimize the power in (9) for frequencies other than ω, *i.e.*, we want to minimize (9) taking the constraint (8) in consideration. Using matrix representation, (8) can be written as:

$$\mathbf{g}^H \mathbf{e} = \mathbf{e}^H \mathbf{g} = 1 \tag{10}$$

Using Lagrange multiplier, we effectively need to minimize

$$\mathbf{Q} = \mathbf{g}^H \mathbf{R_x} \mathbf{g} + \lambda(1 - \mathbf{g}^H \mathbf{e}) \tag{11}$$

Hence, on differentiating **Q** with respect to $\mathbf{g}^H$ and equating to zero, we get

$$\lambda = \frac{1}{\mathbf{e}^H \mathbf{R_x}^{-1} \mathbf{e}} \tag{12}$$

and,

$$\mathbf{g} = \frac{\mathbf{R_x}^{-1} \mathbf{e}}{\mathbf{e}^H \mathbf{R_x}^{-1} \mathbf{e}}. \tag{13}$$

MV spectrum estimator can be realized as a multi-stage filter. It is better to realize the MV filters using lattice structure instead of direct forms, because direct form structures are extremely sensitive to parameter quantization. The lattice structure implementation is as illustrated in Fig. 11 the MV filter. The reflection coefficients K_m's depend on the input sequence where M is the window length.

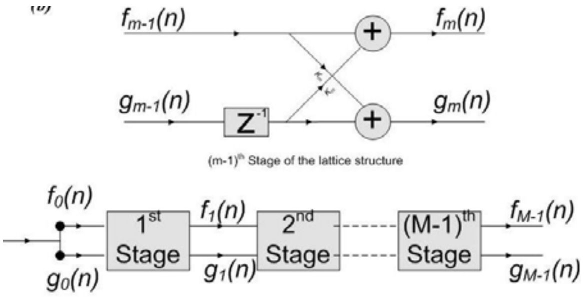

Fig. 11. Lattice structure realization for MV filter.

Fig. 12 shows the magnitude of frequency response of MV spectrum estimator filter for a particular window of 351 samples. It should be noted that MV spectrum estimator is *an adaptive filter* and the frequency response changes with the input samples. It is evident from the figure that there is no first harmonic or conjugate frequency component at $4\pi/3$ in the frequency response which was the problem with the exiting filtering techniques, as in Fig. 5 (a), thus the harmonics are completely suppressed. MV filter is also effective in the case of small length genes, as in Fig. 13 the exon from base location 1 – 100 has come out strongly which, most of the time, was *not-visible* and sometimes below threshold in the existing filtering techniques' output. This is also an outcome of complete suppression of the harmonics of $2\pi/3$ frequency because now the peak strength reflects the true 3-periodic components. A further analysis of MV spectrum analysis in comparison the the antinotch filter is presented in Section VII-C.

As a conclusion, the presented gene filtering techniques are optimal in the sense that they remove almost all harmonics and

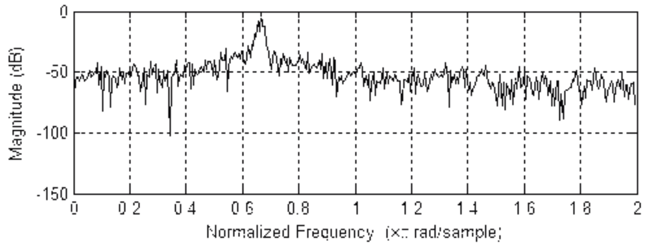

Fig. 12. The Minimum Variance spectrum estimator: magnitude of frequency response.

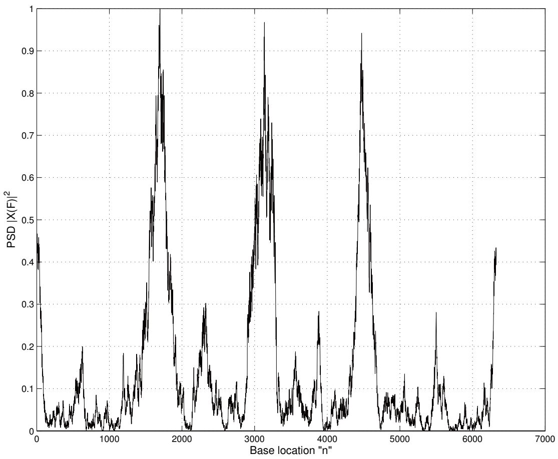

Fig. 13. The Minimum Variance spectrum estimator: output filtered DNA sequence.

false peaks, and minimize the power leakage in side lobes. To further optimize the process, we look at the methods of energy tracking. Traditional methods calculate the energy of a signal as sum of square of samples in the sliding window. A more localised way of energy tracking is required so that the fluctuations in the energy content of the DNA sequence be more readily accessible. The next section talks about this issue in detail, and proposes to use an alternative L^2 energy tracking technique.

D. Variable length Teager Energy Operator

Variable length Teager energy operator (VTEO) is an extension of the powerful Teager energy operator proposed by Kaiser and Teager [6], [21]. VTEO gives the running estimate of energy as to be directly proportional to the square of amplitude and instantaneous frequency of the signal samples unlike the traditional methods, *i.e.*,

$$E \propto A^2\omega^2. \tag{14}$$

The expression for VTEO for discrete time signals can be written as,

$$\psi[x(n)] = x^2(n) - x(n+i)x(n-i) \tag{15}$$

where $\psi[.]$ represents the VTEO operator which gives the running estimate of energy of the discrete time signal $x(n)$. 'i' is the dependency or context index. In this work, the optimal value of i has been empirically derived to be 4. One should refer to [6] for further details on this topic.

VTEO response for band-pass filtered (using MV technique) DNA sequences, where the $2\pi/3$ frequency should be dominant, is illustrated in Fig. 14. It is important to note that the VTEO is to be used to calculate the energy estimate in lieu of PSD after pre-processing the data with any-one of the discussed filtering techniques. One could also use the aforementioned comb filtering technique as a preprocessing step, for example see Section VII-B. It is evident from the Fig. 14 that peaks in exon regions are more dominant and the $1/f$ noise (non-coding regions) is almost entirely suppressed. This is because the output of the filters has more energy in the exon regions. The figures bring out the beauty of VTEO that it boosts the desired $2\pi/3$ components while compressing the noise (non coding regions) irrespective of the filtering techniques used. Furthermore, VTEO also minimizes the computational complexity as it takes only 3 samples into calculation.

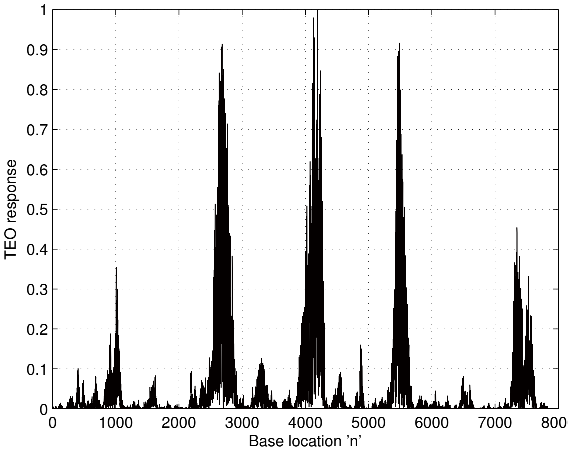

Fig. 14. VTEO response of bandpass filtered DNA sequence, where the MV filtering technique is used.

VII. Discussion

In this Section, we compare the different mapping functions, discussed in Section IV, and discuss their efficiency with different gene prediction techniques. We also discuss the choice of mapping function depending upon the available resources and requirements. The comparisons are done against the following two constraints.

A. Peaks difference

A criterion to decide on a mapping function is the difference between the peaks in exon regions and the intron regions, i.e., the visibility of the peaks in exon regions. So far, in all the mapping functions tested, Voss mapping [13] has emerged as the mapping with maximum peak difference.

An important factor to decide on a filtering technique is the frequency resolution and the window size, i.e., number of samples taken into computation once. *Ideally to calculate the periodicity of a signal one needs infinite number of samples and infinite time to process [22]*, which for obvious reasons is not practical. It should be understood that the smaller the window the lesser the computational overhead. Thus the task at hand is to reduce the number of samples while successfully capturing the 3-periodicity in the samples of DNA sequence. For the perfect case we want to capture the 3-periodicity with only 3 samples. But as we reduce the window size in time domain, the trade-off is reflected in the frequency domain, *i.e.*, width of main lobe of the sinc function increases, covering more frequency components and giving lesser frequency resolution, for example, in case of rectangular window the width of the main lobe is $4\pi/N$ where N is window size in time domain. On convolving the sinc function with the DTFT of the DNA sequence, all the undesired frequency components in the sinc main lobe add up in the DFT spectrum including the concerned $2\pi/3$ component hence giving *spurious* peaks in the PSD plot.

B. 3-dB Power Differentiator

As discussed in Section V-A and in Eq. (3), the criteria for deciding on whether a peak represents a gene lies within the threshold power. This threshold power depends significantly on the gene sequence [3]. Alternatively, the 3-dB power of the maximum power can be chosen as the threshold power, which shows significant improvement over the other thresholds for all the tested genes. Thus, another important method for differentiating the exon regions from intron regions is to *supercharge* the peaks having power more than or equal to the threshold power. This significantly increases the strengths of peaks representing gene while leaving the spurious peaks as it is. An illustration for this method, when applied to output of the HaSCo filter, is depicted in Fig. 15, and VTEO response of the same in Fig. 16. The effectiveness is evident in the figures where all the five exons are captured clearly with complete suppression of the introns. However, this method can only be used in confidence when the power in introns is expected to be below 3-dB threshold, which is generally the case after applying VTEO, as illustrated in Fig. 14.

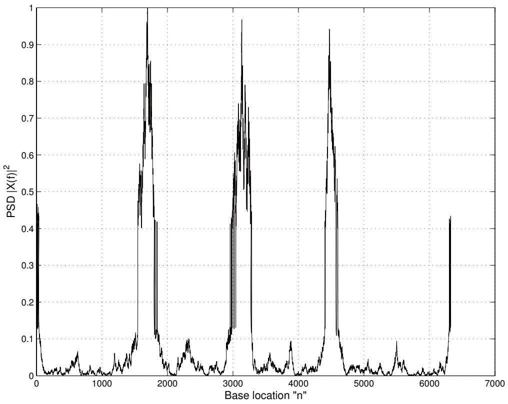

Fig. 15. The 3-dB power differentiator output for the comb filter combined with VTEO.

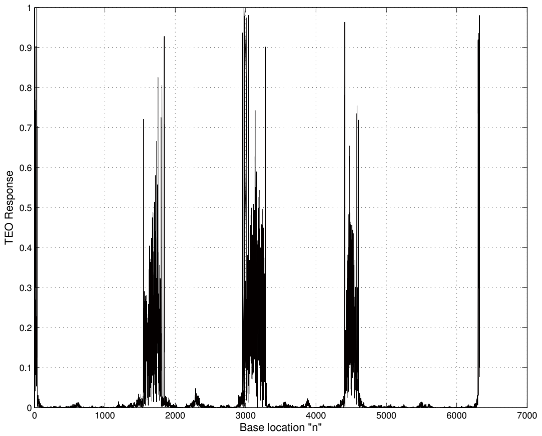

Fig. 16. The 3-dB power differentiator output for MV-VTEO response

C. Window Size

For all the existing techniques discussed in Section V, the optimal minimum window size is 351 samples, below which spurious peaks are introduced making it difficult to differentiate between the exon and intron regions. Though, in [16] and [19] authors have claimed to achieve comparable results with lower window sizes, the improvements are not very encouraging as demonstrated in Fig. 17. This particular problem can be tackled by the VTEO energy estimator, which yields good results at a window size of 117 samples, as depicted in Fig. 18.

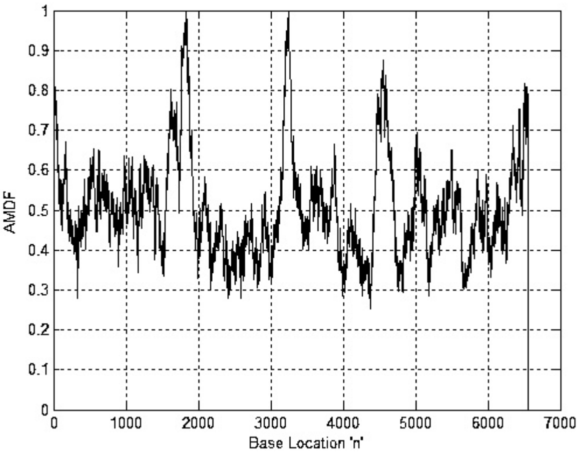

Fig. 17. AMDF response for a band pass filtered DNA sequence at windows size= 117

The proposed MV spectrum estimation technique also yields perceptibly good results at smaller window sizes. Figure 19 shows a qualitative comparison between the antinotch filter and the proposed Minimum Variance spectrum estimator for a range of window sizes. In Fig. 19, the horizontal dashed lines represent the threshold P as defined in Eq. 3. The circles over some regions of the figures highlight the spurious peaks in the PSD that contribute to false gene detection.

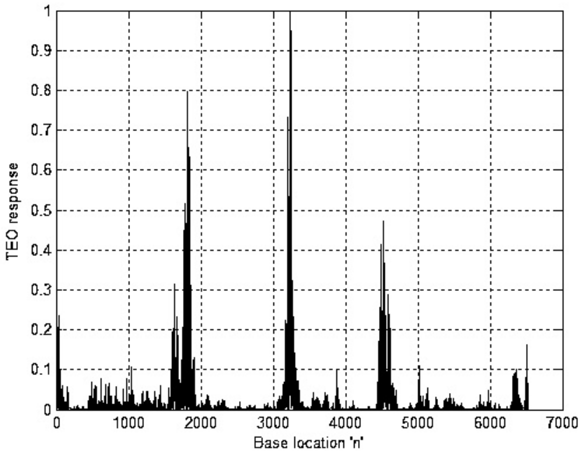

Fig. 18. VTEO response for a HaSCo band pass filtered DNA sequence at window size =117

It is clear from Fig. 19 that, for MV spectrum estimator, the noise in introns regions is suppressed giving more peaks difference among the exons and introns regions than in the case of antinotch filter. This is highly useful in case of smaller length exons, as in the range 1-100 in the figures, which is clearly visible in case of MV whereas it is confused among the introns region peaks, e.g., 100-1500, in the antinotch output. Even at a window size of mere 90 samples, the MV filter output can be perfectly utilized after enhancing using VTEO and 3-dB power differentiator which will lead to false results in case of antinotch filter because of the equivalent power in introns regions. It is important to remember that the threshold power depends on the gene as well as on the window size.

VIII. Conclusion

The classical DSP tools like Fourier transforms and time-frequency analysis have been used for a long time in studying DNA sequences and genes. But these tools and techniques have *not* been very effective due to their inability to eliminate the harmonic frequency components and thus in detecting smaller length exons. In this paper, these approaches were enhanced, and use of parametric spectral analysis to overcome this inability by completely eliminating the harmonic frequencies and thus capturing even smaller exons was introduced. The analyses were of two basic types. First, the mechanisms to suppress the harmonic frequencies, in non-parametric methods, were suggested by introducing smoothing window functions and HaSCo filtering technique. Second, the parametric MV spectral estimation tools to analyse DNA sequences was proposed. Furthermore, use of the traditional methods of calculating L^2 energy has some pitfalls which lack in tracking the small changes in frequency. To overcome this, VTEO was used to calculate the running estimate of energy in lieu of PSD, and it was demonstrated that VTEO was able to eliminate the $1/f$ noise and hence the peaks in intron regions to a greater extent.

Genomic Signal Processing holds many promises for a bright future of bio-informatics. Once gene positions are predicted, gene identification is also possible by utilizing simple techniques like linear prediction [23]. The gene prediction algorithms can then further be extended to identify the particular protein synthesized by the process of splicing from this gene, however this procedure is made tough by the process of alternative splicing. Also, newly developed methodologies and tools in signal processing for modelling signals and processes appear to be most promising for genomic research and DNA sequence compression, and opens the possibilities of further work [24].

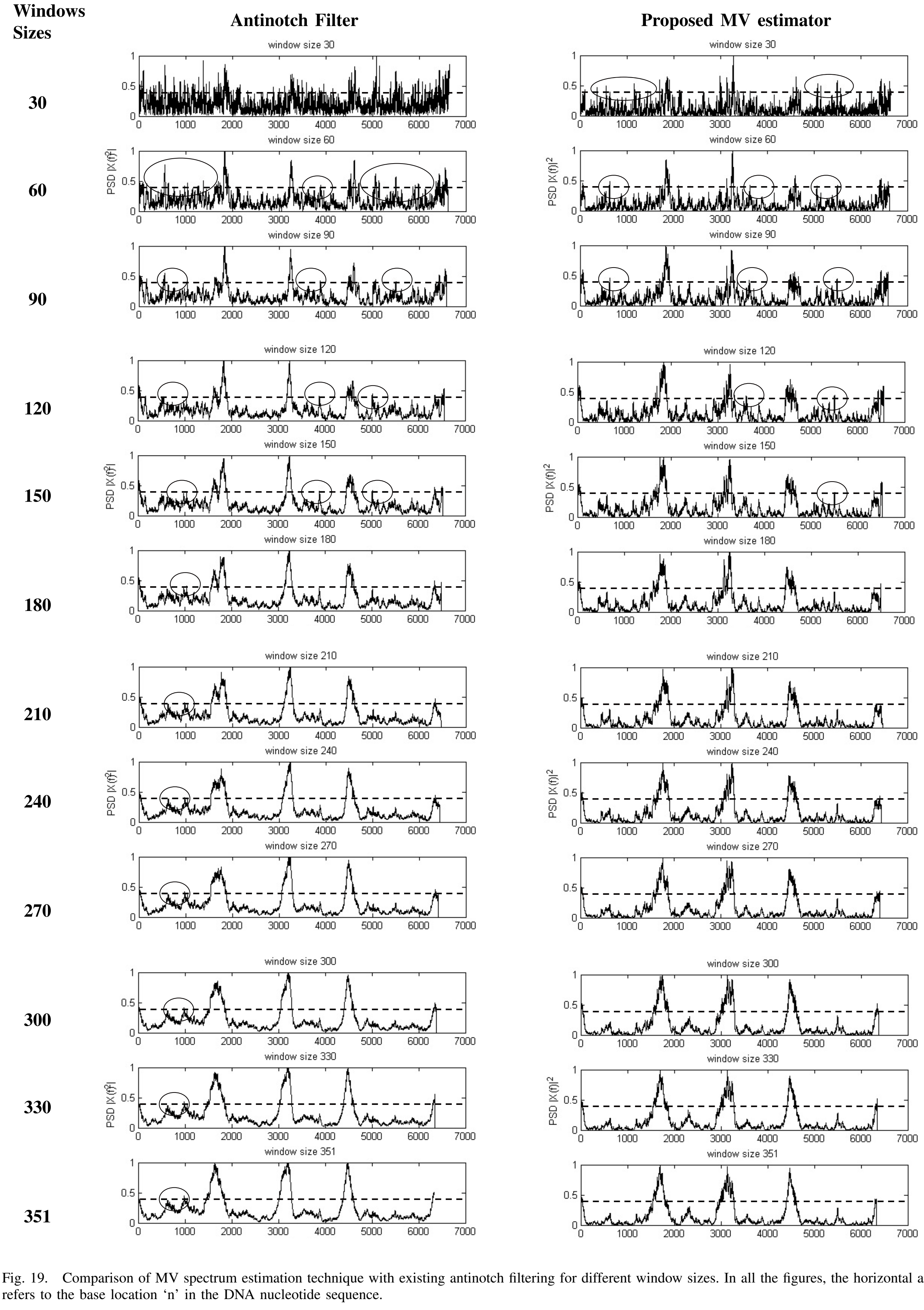

Fig. 19. Comparison of MV spectrum estimation technique with existing antinotch filtering for different window sizes. In all the figures, the horizontal axis refers to the base location ‘n’ in the DNA nucleotide sequence.

REFERENCES

[1] E. R. Dougherty, A. Datta, and S. C., "Research Issues in Genomic Signal Processing," *IEEE Signal Processing Magazine*, pp. 46–68, November 2005.

[2] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Res.*, vol. 10, pp. 5303–5318, 1982.

[3] S. Tiwari, S. Ramachandran, and A. Bhattachalya, *et al*, "Prediction of probable gene by Fourier analysis of genomic sequences," *CABIOS*, vol. 13, no. 3, pp. 263–270, 1997.

[4] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine*, pp. 8 – 20, July 2001.

[5] P. P. Vaidyanathan, "Genomics and Proteomics: A Signal Processor's Tour," *IEEE Circuits and Systems Magazine*, November 2004.

[6] V. Tomar and H. A. Patil, "On the Development of Variable length Teager Energy Operator (VTEO)," *Interspeech*, Australia 2008.

[7] Human Genome Project. [Online]. Available: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[8] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[9] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *Journal of Theoretical Biology*, vol. 118, no. 3, pp. 295–300, Feb 1986.

[10] S. T. Eskesen, F. N. Eskesen, K. Brian, and A. Ruvinsky, "Periodicity of DNA in exons," *BMC Molecular Biology*, vol. 5, pp. 12–24, 2004.

[11] R. Guigo, "DNA composition, codon usage and exon prediction." [Online]. Available: http://www.pdg.cnb.uam.es/cursos/FVi2001/GenomAna/GeneIdentification/SearchContent/

[12] C. K. Peng, S. V. Buldyrev, A. Goldberger, and S. Havlin, *et al*, "Long-range Correlations in nucleotide sequences," *Nature*, vol. 356, pp. 168–170, March 1992.

[13] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, June 1992.

[14] "National Center for Biotechnology Information, US National Library of Medicine, National Institute of Health." [Online]. Available: http://www.ncbi.nlm.nih.gov/

[15] "Wormbase." [Online]. Available: http://www.wormbase.org/

[16] N. Rao and S. J. Shepherd, "Detection of 3-periodicity for small genomic sequences based on AR technique," *in Proc. International Conference on Communications, Circuits and Systems, ICCCAS*, vol. 2, pp. 1032–1036, June 2004.

[17] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Oxford University Press, Bioinformatics*, vol. 16, pp. 1073–1081, 2000.

[18] P. P. Vaidyanathan and B. J. Yoon, "The Role of Signal-Processing Concepts in Genomics and Proteomics," *Invited Paper, Journal of the Franklin Institute, Special Issue on Genomics*, 2004.

[19] E. Ambikairajah, E. J., and A. M., "Gene and exon prediction using time-domain algorithms," *IEEE 8th Int. Symp.on Sig. Proc. and its Appl.*, pp. 199–202, 2005.

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

[21] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Pro.*, vol. 28, pp. 599–601, 1980.

[22] J. G. Proakis and D. G. Manolakis, "Digital Signal Processing: Principles, Algorithms and Applications," *3rd edition, Printice Hall*, India.

[23] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive Modeling and Feature Analysis of DNA Sequences," *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 13–28, 2004.

[24] G. Korodi, I. Tabus, J. Rissanen, and J. Astola, "DNA Sequence Compression," *IEEE Signal Processing Magazine*, pp. 47–53, July 2007.