Digital Signal Processing 22 (2012) 869–877

Contents lists available at SciVerse ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Identification of protein coding regions using antinotch filters

Malaya Kumar Hota [a,*], Vinay Kumar Srivastava [b]

[a] Department of Electronics and Telecommunication Engineering, Synergy Institute of Engineering and Technology, Dhenkanal 759001, Odisha, India
[b] Department of Electronics and Communication Engineering, Motilal Nehru National Institute of Technology, Allahabad 211004, Uttar Pradesh, India

ARTICLE INFO

Article history:
Available online 19 June 2012

Keywords:
Genomic sequence analysis
Protein coding region
Period-3 property
Antinotch filter

ABSTRACT

A major area of research in genomic sequence analysis is the identification of protein coding regions using the period-3 property. Previously antinotch filter has been used for this purpose. In this paper, three antinotch filters, namely conjugate suppression antinotch filter, antinotch filter followed by moving average filter and harmonic suppression antinotch filter are proposed to improve the identification accuracy. Conjugate suppression antinotch filter suppresses the conjugate frequency component, antinotch filter followed by moving average filter reduces the background noise and harmonic suppression antinotch filter suppresses the harmonic frequency component. Several existing DNA to numerical mapping techniques are compared for GENSCAN test set and based on the result one mapping technique is recommended so that detailed analysis can be performed using various datasets. The computational complexity of the antinotch filters is evaluated in comparison with the ST-DFT method and it is found that the computational load is reduced to a greater extent in antinotch filter. The identification accuracy of the proposed antinotch filter methods is compared with the existing antinotch filter method at the nucleotide level for benchmark datasets. The results show that proposed methods outperform the existing method, giving improved identification of the protein coding regions.

1. Introduction

Deoxyribonucleic acid (DNA) contains the genetic information of living organisms. DNA consists of genes and intergenic regions. In eukaryotes (cells with a nucleus), a gene can be further subdivided into exons and introns. Cells without a nucleus are called prokaryotes and do not contain introns. Only the exons are involved in protein coding. Proteins are sequences made of amino acids. An important topic in genomic signal processing is the identification of protein coding regions using signal processing techniques. Anastassiou [1] and Vaidyanathan [2] have given an overview of some of the important aspects of genomic signal processing. The methods for the identification of protein coding regions are divided into model-dependent and model-independent methods [3]. Model-dependent methods are built upon some a priori information while model-independent methods do not assume such a priori information.

The protein coding regions in genes have a period-3 component, which is not found in other regions such as intergenic and introns in eukaryotes [1,2]. Tiwari et al. also observes that, some genes don't exhibit period-3 property at all in *S. Cerevisiae* [4]. W. Li observes that, such periodicity is also present in noncoding regions [5]. Therefore, identification of protein coding region is a complex problem. However, most of the digital signal processing (DSP) methods for the identification of protein coding regions are based upon the period-3 property, i.e., the period-3 component is found in the coding regions but is not found in the noncoding regions.

Short-time discrete Fourier transform (ST-DFT) is one of the earliest and traditional methods used for the identification of protein coding region based on period-3 property. It involves taking discrete Fourier transform (DFT) of a rectangular window of a defined size in a DNA sequence, which is then made to slide across the whole length of the sequence [4,6]. Many other methods such as optimal ST-DFT [1], spectral rotation [7], digital filters (antinotch, multistage) [2,6], single digital filter followed by quadratic window operation [8], time–frequency hybrid measure [9], signal boosting technique [10,11], multirate DSP model [12], modified Gabor-wavelet transform [13], and modified Gabor-wavelet transform with signal boosting technique [14] have also been used for the identification of protein coding region based on period-3 property.

The ST-DFT method can be regarded as digital filtering followed by a decimator which depends on the separation between adjacent positions of the window [15]. Digital filter methods are of interest because they are significantly faster than DFT based methods. In digital filter method a narrowband bandpass filter is designed whose passband is centered at $2\pi/3$. The narrowband bandpass filter can be regarded as an antinotch filter. If we pay more careful

* Corresponding author. Fax: +91 (6762) 225905/226708.
E-mail addresses: mkhota.mnnit@gmail.com, malaya_hota@rediffmail.com (M.K. Hota), vinay@mnnit.ac.in (V.K. Srivastava).

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2012.06.005

Table 1
Summary of benchmark datasets.

Dataset	Organism	Gene sequence	bp
GENSCAN test set	Human	65	602,030
HMR195	Mammalian	195	1,386,021
BG570	Vertebrate	570	2,892,149

attention to the design of the digital filter, we can isolate the period-3 behavior from the background noise [6]. In this paper, three efficient antinotch filters with improved noise suppression are proposed which improves the identification accuracy of protein coding regions.

2. Materials and methods

2.1. Data resources

For a detailed analysis, the commonly used gene F56F11.4 in the *C. elegans* Chromosome III of 8000 bp is used, which contains five coding exons in positions 928–1039, 2528–2857, 4114–4377, 5465–5644, and 7255–7605 (GenBank access number AF099922 and positions 7021–15020). To measure the computational complexity, two genes are also chosen randomly from NCBI GenBank database (http://www.ncbi.nlm.nih.gov). One is the Mus musculus homeobox containing nuclear transcriptional factor Hmx1 gene of 5195 bp, which contains two coding exons in positions 1267–1639 and 3888–4513 (GenBank access number AF009614). Other is the Mus musculus gene for PSMB5 of 5006 bp, which contains three coding exons in positions 1020–1217, 2207–2513 and 4543–4832 (GenBank access number AB003306). To measure the identification accuracy, three benchmark datasets as summarized in Table 1 are also considered. HMR195 is a data set of 195 single-gene human, mouse and rat sequences from [16]. BG570 is a genomic test data set of 570 single gene vertebrate sequences from [17]. GENSCAN test set comprises 65 available multiexon gene sequences listed in [18, Appendix B].

2.2. Mapping of DNA into numerical sequences

A DNA sequence is made from an alphabet of four elements, namely A, C, G and T (adenine, cytosine, guanine, and thymine respectively). The mapping of DNA alphabet into digital signals is central to DSP based DNA sequence analysis. Perhaps the most widely used mapping technique for this purpose is Voss representation (or binary indicator sequences) [19]. Each strand of DNA consists of four nucleotides (or bases) which can be mapped into four signals. The binary indicator sequence $u_A(n)$ takes the value of either 1 or 0 depending upon the presence and absence of base A at position n. Other binary indicator sequences $u_T(n)$, $u_C(n)$, and $u_G(n)$ can be obtained in a similar fashion. In tetrahedron [1,20] and z-curve [21,22] mapping techniques three indicator sequences are used instead of four indicator sequences. Both tetrahedron and z-curve indicator sequences can be calculated from Voss indicator sequences [1,22]. Many other single indicator sequence have also been introduced, such as complex [1,23,24], electron–ion interaction potential (EIIP) [25,26], real numbers [22,27,28] and paired numeric [29]. In this paper we have considered only those mapping techniques which do not assume a priori information. We have not considered frequency of nucleotide occurrence [29] and nucleotide bias [30] because these mapping techniques need a priori information.

Using the proposed antinotch filters and the existing antinotch filter several existing DNA to numerical mapping techniques are compared for the GENSCAN test set at the nucleotide level. Based on the results, one mapping technique is recommended so that detailed analysis can be performed using various data sets. A comparative evaluation of the proposed and existing antinotch filter methods for the GENSCAN test set is also made using these mapping techniques. Further, the performance of these antinotch filters is evaluated in terms of computational complexity in comparison with the ST-DFT method using recommended mapping technique for different data sets. Furthermore, the identification accuracy of the proposed methods is compared with the existing method at the nucleotide level for benchmark data sets.

2.3. Antinotch filters for the identification of protein coding regions

Consider a digital filter whose magnitude response has a sharp peak at $2\pi/3$. This digital filter can be regarded as an antinotch filter. If we give the binary indicator sequences $u_A(n)$, $u_T(n)$, $u_C(n)$ and $u_G(n)$ separately as inputs to the antinotch filter, the corresponding outputs $y_A(n)$, $y_T(n)$, $y_C(n)$ and $y_G(n)$ will be relatively large in the coding regions as the inputs to the filter are having the period-3 property and the filter has a passband around $2\pi/3$. Thus, using digital filters, the feature can be computed as

$$Y(n) = \sum_{i \in F} |y_i(n)|^2, \qquad F = \{A, T, C, G\}. \tag{1}$$

A plot of $Y(n)$ will have peaks in the coding regions where as such peaks are absent in noncoding regions. So, this feature can be utilized for the identification of protein coding regions in a DNA segment.

2.3.1. Antinotch filter (ANF)

The design and implementation of existing IIR antinotch filter for the identification of protein coding regions has been proposed by Vaidyanathan and Yoon in [6,31]. The antinotch filter is a narrowband bandpass filter which provides high gain in the passband region. The antinotch filter can be obtained by starting from a second order real coefficient allpass filter. The transfer function of the antinotch filter is

$$H(z) = \frac{1}{2} \frac{(1 - R^2)(1 - z^{-2})}{(1 - 2R\cos\theta z^{-1} + R^2 z^{-2})}. \tag{2}$$

This antinotch filter has two poles at $Re^{\pm j\theta}$ and two zeros at ± 1. The value of pole radius R should be less than one for stability. The magnitude response and pole–zero plot of the antinotch filter is shown in Fig. 1. It is observed that the filter has two passband with center frequencies at $2\pi/3$ and $4\pi/3$ because of the two poles at those frequencies. This second order IIR antinotch filter is stable and real.

2.3.2. Conjugate suppression antinotch filter (CSANF)

The antinotch filter proposed by Vaidyanathan and Yoon [6,31] passes the frequency component at $2\pi/3$ along with its conjugate at $-2\pi/3$ or $4\pi/3$. Conjugate frequency component is present due to the complex conjugate nature of poles and zeros. This conjugate frequency component may contribute towards the peak strength in exons and introns, giving inaccurate measure of the protein coding regions. Thus, the passband due to this conjugate frequency component should be suppressed.

In the proposed filter, ANF is used at the first stage and a first order complex FIR filter is used at the second stage whose transfer function is

$$H_1(z) = 1 - e^{j4\pi/3} z^{-1}. \tag{3}$$

The second stage complex FIR filter has one zero on the unit circle at $\omega = 4\pi/3$ and one pole at origin. The magnitude response and pole–zero plot of complex FIR filter is shown in Fig. 2. Proposed second stage filter is able to suppress the frequency component

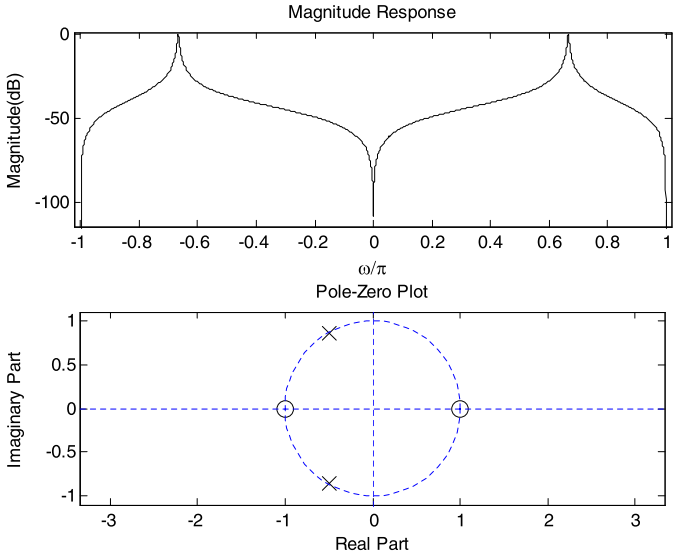

Fig. 1. Magnitude response and pole–zero plot of ANF for pole radius $R = 0.992$.

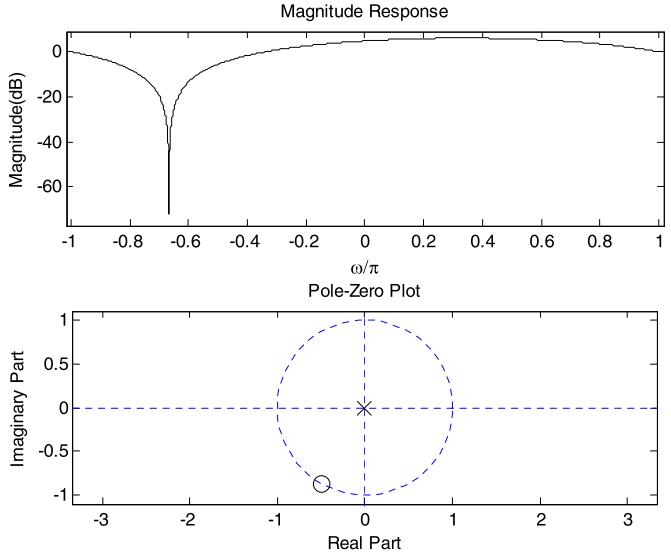

Fig. 2. Magnitude response and pole–zero plot of the first order complex FIR filter which is used at the second stage of CSANF.

present at $\omega = 4\pi/3$. Thus the resulting filter is a conjugate suppression antinotch filter with center frequency at $\omega = 2\pi/3$. The magnitude response and pole–zero plot of conjugate suppression antinotch filter is shown in Fig. 3.

Schematic data flow diagram of the proposed CSANF method for the identification of protein coding regions is shown in Fig. 4. The method has four steps:

(i) numerical mapping of DNA sequence into four binary indicator sequences,
(ii) filtering of each indicator sequence by proposed antinotch filter to get the filtered output sequence,
(iii) summation of magnitude square of each filter output sequence to get the feature, and
(iv) thresholding of the feature for location of the coding and the noncoding regions.

The proposed third order filter is stable but complex since the second stage filter is a first order complex FIR filter. There is one complex filter coefficient present at the second stage FIR filter but

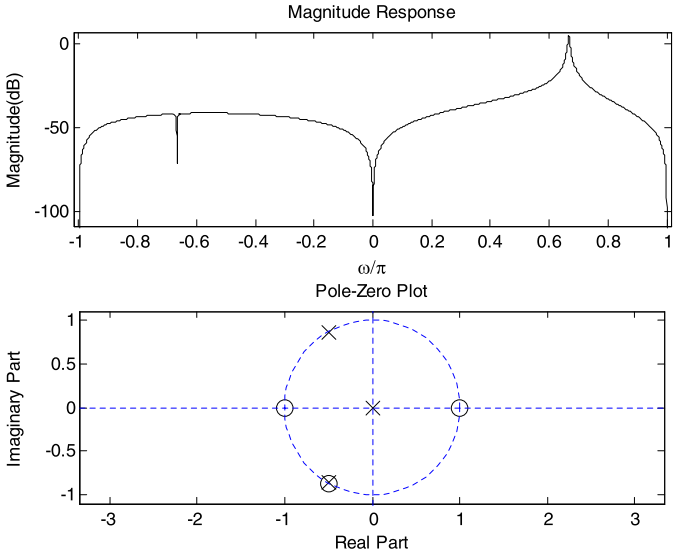

Fig. 3. Magnitude response and pole–zero plot of the CSANF for pole radius $R = 0.992$.

the inputs to the second stage filter may be real (in case of Voss mapping) or complex. The complex filter outputs are governed by the following equations where both inputs and filter coefficients are complex. Let

$$y_r(n) = \sum_{k=0}^{M-1} \left[b_{rk} x_r(n-k) - b_{ik} x_i(n-k) \right],$$
$$y_i(n) = \sum_{k=0}^{M-1} \left[b_{ik} x_r(n-k) + b_{rk} x_i(n-k) \right], \tag{4}$$

where $y_r(n)$ is real output, $y_i(n)$ is imaginary output, $x_r(n)$ is real input, $x_i(n)$ is imaginary input, b_r is real coefficient, b_i is imaginary coefficient, and M is the length of the filter. A generalized first order complex FIR filter structure is shown in Fig. 5 where both inputs and filter coefficients are complex. In this complex filter, the real and imaginary outputs are calculated separately by two FIR filters, namely FIR_{real} and $\text{FIR}_{\text{imaginary}}$, respectively. In our case $b_{r0} = 1$, $b_{i0} = 0$, $b_{r1} = -\cos(4\pi/3)$, and $b_{i1} = -\sin(4\pi/3)$. Therefore, there will be four real multiplications if the input is complex.

2.3.3. Antinotch filter followed by moving average filter (ANFMA)

There is an inherent tradeoff between filter design technique and background noise reduction at the output of the filter. The conjugate suppression antinotch filter provides better noise reduction but is harder to design. Therefore, in proposed antinotch filter, IIR antinotch filter is followed by a moving average filter. The moving average filter is a second order lowpass FIR filter whose transfer function is

$$H_2(z) = \left(1 + z^{-1} + z^{-2}\right)/3. \tag{5}$$

Moving average filter is used for smoothing purpose because it attenuates the background noise. In the proposed antinotch filter, there is slightly higher complexity because of the second stage, but its characteristics are significantly better than ANF. Schematic data flow diagram of the proposed ANFMA method for the identification of protein coding regions is shown in Fig. 6. This method has similar four steps as discussed with reference to CSANF.

2.3.4. Harmonic suppression antinotch filter (HSANF)

Instead of suppressing conjugate frequency component, we can also suppress the harmonic frequency components. In the pro-

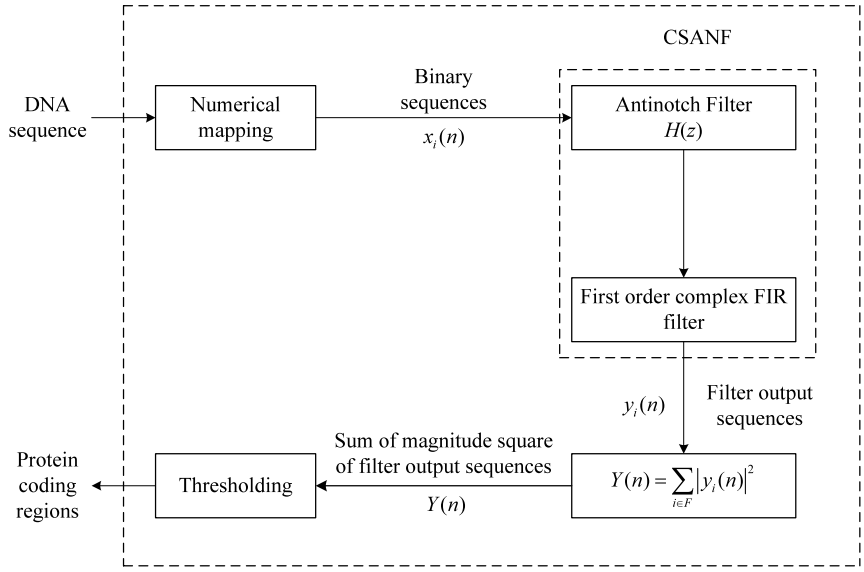

Fig. 4. Schematic data flow diagram of the proposed CSANF method for the identification of eukaryotic protein coding regions.

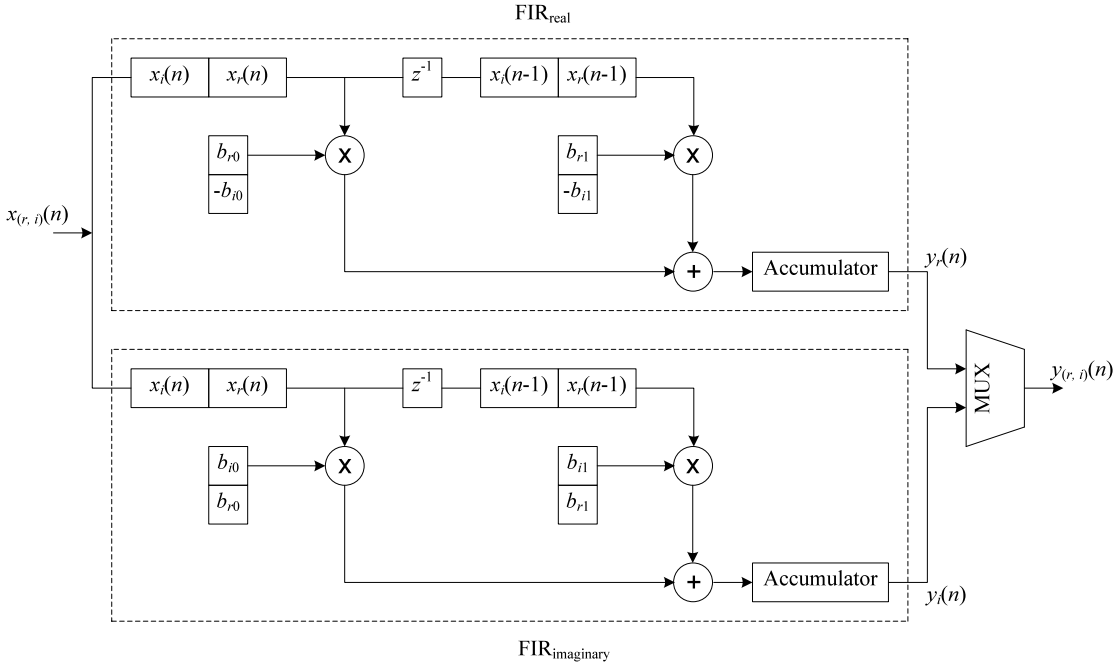

Fig. 5. First order complex FIR filter structure.

posed filter, ANF is used at the first stage and a third order complex FIR filter is used at the second stage whose transfer function is

$$H_3(z) = 1 + e^{j2\pi/3}z^{-1} - e^{j2\pi/3}z^{-2} - e^{j4\pi/3}z^{-3}. \quad (6)$$

The second stage complex FIR filter has three zeros on the unit circle at $\omega = 2\pi/6, 4\pi/3$ and $10\pi/6$ and three poles at origin. The magnitude response and pole–zero plot of complex FIR filter is shown in Fig. 7. Thus the resulting filter is a harmonic suppression antinotch filter having dominant zeros at the multiples of frequency $2\pi/6$ except at $2\pi/3$ and dominant pole at $2\pi/3$. The magnitude response and pole–zero plot of harmonic suppression antinotch filter is shown in Fig. 8. The harmonics of $2\pi/6$ is suppressed. Suppression of higher harmonic is not necessary because there contribution is negligible. Proposed fifth order HSANF is stable but complex since the second stage filter is a third order complex FIR filter. A generalized third order complex FIR filter structure is shown in Fig. 9 where both inputs and filter coefficients are complex. In our case $b_{r0} = 1$, $b_{i0} = 0$, $b_{r1} = \cos(2\pi/3)$, $b_{i1} = \sin(2\pi/3)$, $b_{r2} = -\cos(2\pi/3)$, $b_{i2} = -\sin(2\pi/3)$, $b_{r3} = -\cos(4\pi/3)$, and $b_{i3} = -\sin(4\pi/3)$. Therefore, there will be twelve real multiplications if the input is complex. Thus, computational complexity increases. Schematic data flow diagram of the proposed HSANF method for the identification of protein coding regions is shown in Fig. 10. This method has similar four steps as discussed with reference to CSANF.

2.4. Performance assessment

Identification of protein coding region results are compared at the nucleotide level [13] rather than at exon level [7]. Per-

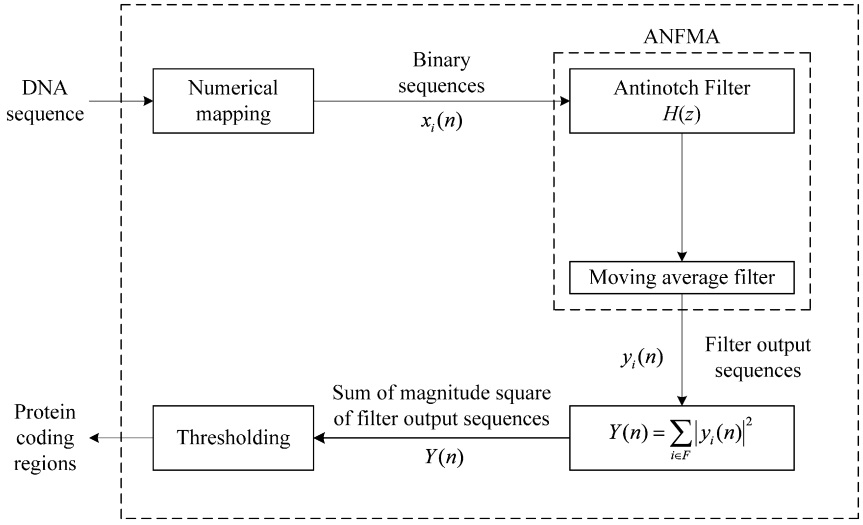

Fig. 6. Schematic data flow diagram of the proposed ANFMA method for the identification of eukaryotic protein coding regions.

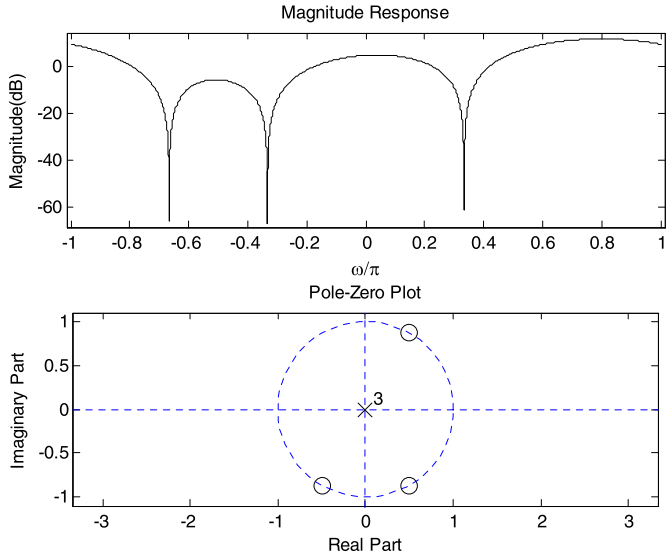

Fig. 7. Magnitude response and pole–zero plot of the third order complex FIR filter which is used at the second stage of HSANF.

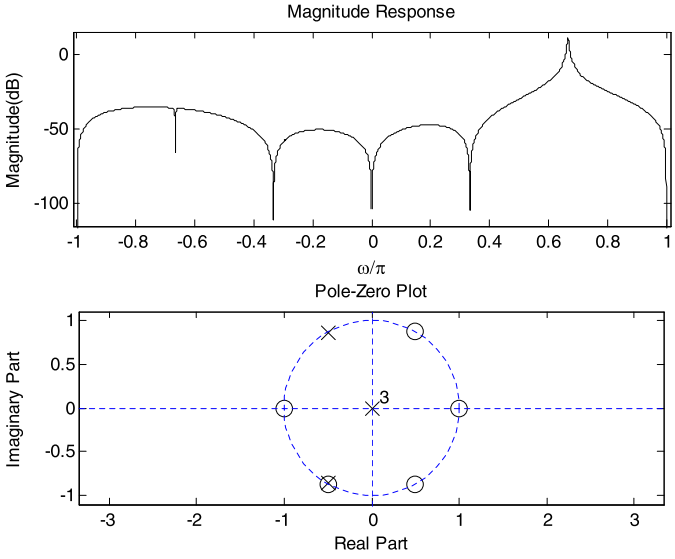

Fig. 8. Magnitude response and pole–zero plot of the HSANF for pole radius $R = 0.992$.

formance at the nucleotide level is measured in terms of the true-positive (*TP*), true-negative (*TN*), false-positive (*FP*), and false-negative (*FN*). For comparison, the performance is evaluated by the measures of sensitivity (S_n), specificity (S_p), approximate correlation (*AC*) and correlation coefficient (*CC*). These measurements are commonly used for the identification of protein coding regions [17].

Receiver operating characteristic (*ROC*) curve [32] is also used as a measure to evaluate the performance for the identification of protein coding regions. The *ROC* curve is commonly defined as the plot of the *true-positive rate* as a function of the *false-positive rate* for all possible thresholds. The *ROC* curve can be characterized as a single number using the area under the *ROC* curve (*AUC*), with large area indicating more accurate detection methods.

In order to set up the general form of comparison for the identification of protein coding regions, threshold percentiles within the interval $(1, 99)$ are used. Thus, the *true-positive rate*, *false-positive rate*, S_n, S_p, *AC*, *CC* and *AUC* calculation are obtained under the same conditions at different threshold values.

3. Results and discussion

The features obtained for the identification of protein coding regions using different methods applied to gene F56F11.4 are shown in Fig. 11. The peaks of the graph clearly show the exon regions of the gene. We normalized the peak by its maximum value. The resolutions of the peaks are improved depending upon the antinotch filter used for identification of the exon regions. It can be observed that, the background noise is reduced greatly by all the proposed antinotch filters.

While designing the antinotch filter, proper selection of pole radius R is required. The features obtained using CSANF method applied to gene F56F11.4 for pole radius $R = 0.986$ and 0.994 are shown in Fig. 12. In Table 2, *AUC* values for the same gene have been given using all antinotch filters for different values of R. It can be observed that maximum *AUC* value using CSANF is obtained when the value of R is 0.986. But from Fig. 12 it can be viewed that the background noise is also more with the same value of R. If we increase the value of R, the background noise decreases at the filter output at the same time the accuracy in location of exons decreases since *AUC* decreases. Therefore, we have to make tradeoff

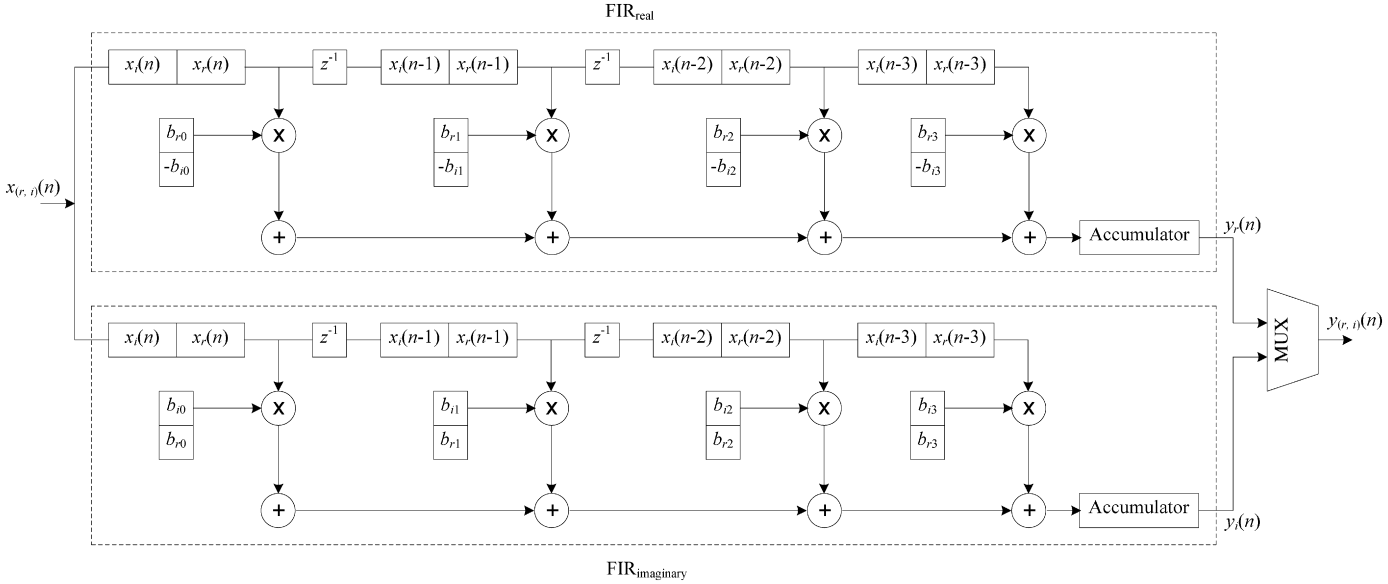

Fig. 9. Third order complex FIR filter structure.

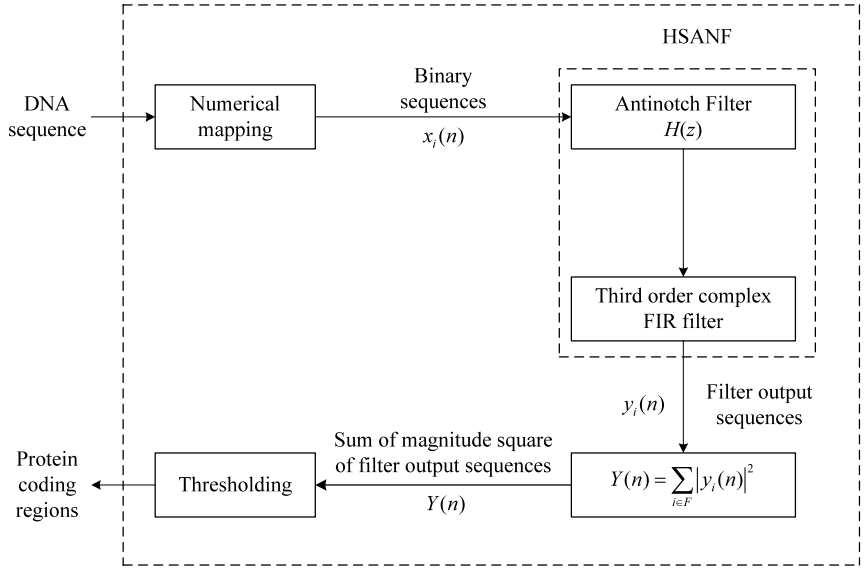

Fig. 10. Schematic data flow diagram of the proposed HSANF method for the identification of eukaryotic protein coding regions.

between these two things. Empirically it is found that, $R = 0.992$ is the most suitable value to enhance the identification accuracy. Also in [6,31], the value of R has been taken as 0.992. Therefore, in this work the value of R is considered as 0.992.

3.1. Performance analysis of DNA to numerical mapping techniques

Several existing DNA to numerical mapping techniques are compared for the identification of protein coding regions, using antinotch filters applied to GENSCAN test set. Fig. 13 shows *ROC* curve for comparison of DNA to numerical mapping techniques for the identification of protein coding regions using CSANF applied to GENSCAN test set. Table 3 shows *AUC* values for GENSCAN test set using antinotch filters for different DNA to numerical mapping techniques. Note that in Fig. 13 and Table 3 'complex 1' refers to $A = 1 + j$, $T = 1 - j$, $C = -1 + j$, $G = -1 - j$; 'complex 2' refers to $A = 1$, $T = j$, $C = -j$, $G = -1$; 'real 1' refers to $T = 0$, $C = 1$, $A = 2$, $G = 3$; 'real 2' refers to $A = 0$, $G = 1$, $C = 2$, $T = 3$; and 'real 3' refers to $A = 1.5$, $T = -1.5$, $C = 0.5$, $G = -0.5$. For identification of protein coding regions, Z-curve and tetrahedron mapping techniques gives almost same result as Voss mapping technique in terms of *ROC* and *AUC*. But other single sequence mapping techniques gives poor result than Voss mapping technique in terms of *ROC* and *AUC*. Voss mapping technique is the simplest, most popular and gives maximum *AUC* values. Voss mapping technique does not depend on any particular adapted labeling and no relevant harmonic structure of biological meaning is hidden or exposed [33]. Therefore, for detailed analysis of antinotch filters Voss mapping technique is preferred in this paper. A comparison of DNA to numerical mapping technique using ST-DFT method for GENSCAN test set can be found in [34].

3.2. Performance analysis of antinotch filters

In [35,36], computational complexity of different methods has been evaluated by the consumed CPU time for the identification of protein coding regions. In this work, the computational efficiency of antinotch filters with respect to the ST-DFT method is

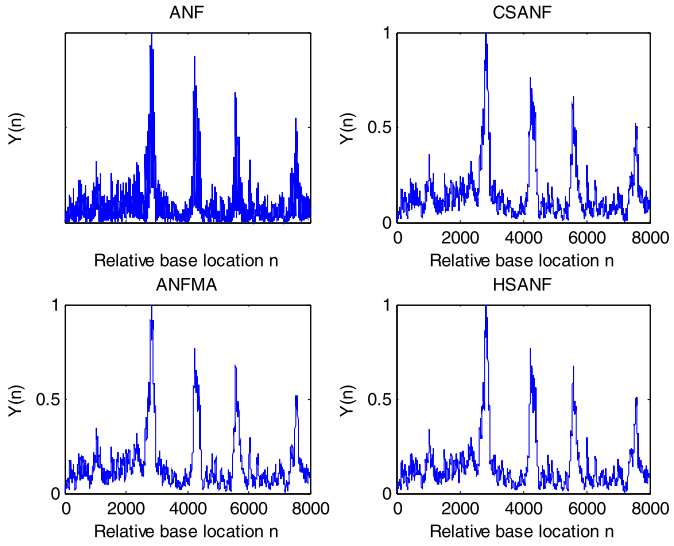

Fig. 11. The features obtained using ANF, CSANF, HSANF and ANFMA methods applied to gene F56F11.4 (clockwise from top left).

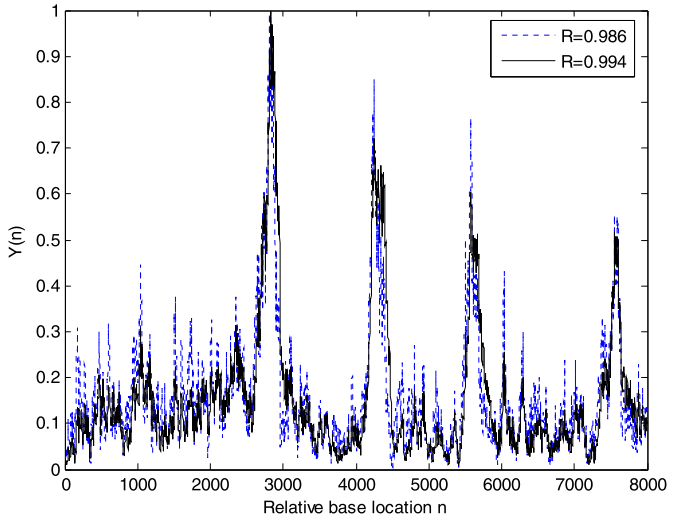

Fig. 12. The features obtained using CSANF method applied to gene F56F11.4 for pole radius $R = 0.986$ and 0.994.

Table 2
AUC values using antinotch filters applied to gene F56F11.4 for different values of pole radius *R*.

Different value of pole radius *R*	*AUC* value using antinotch filter			
	ANF	CSANF	ANFMA	HSANF
0.98	0.7640	0.8723	0.8729	0.8718
0.982	0.7678	0.8759	0.8761	0.8751
0.984	0.7708	0.8786	0.8785	0.8776
0.986	0.7732	0.8801	0.8800	0.8789
0.987	0.7740	0.8800	0.8798	0.8791
0.988	0.7745	0.8795	0.8792	0.8786
0.99	0.7739	0.8756	0.8753	0.8745
0.992	0.7693	0.8656	0.8654	0.8645
0.994	0.7570	0.8447	0.8442	0.8434
0.996	0.7245	0.7999	0.7990	0.7980
0.998	0.6348	0.6845	0.6836	0.6825

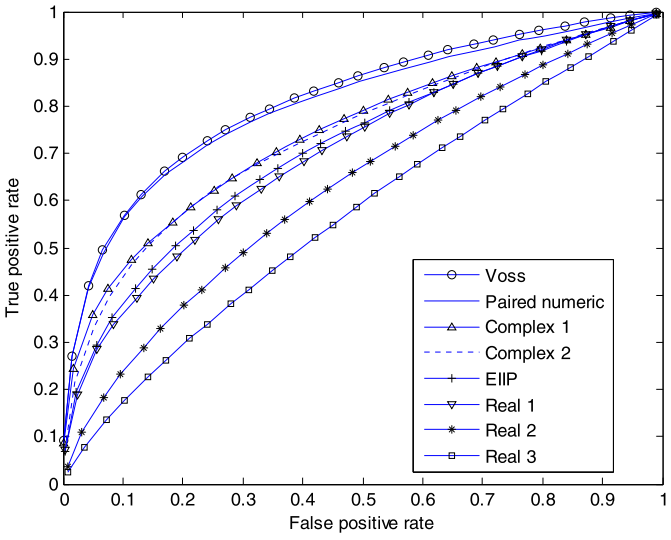

Fig. 13. *ROC* curve for comparison of DNA to numerical mapping techniques for the identification of protein coding regions using CSANF applied to GENSCAN test set.

Table 3
AUC values for the identification of protein coding regions using antinotch filters applied to GENSCAN test set.

DNA symbolic to numerical mapping method	*AUC* value using antinotch filter			
	ANF	CSANF	ANFMA	HSANF
Voss	0.7235	0.8037	0.8035	0.8032
Z-curve	0.7235	0.8037	0.8035	0.8032
Tetrahedron	0.7235	0.8037	0.8035	0.8032
Paired numeric	0.6947	0.7940	0.7938	0.7934
Complex 1	0.6979	0.7339	0.7864	0.7335
Complex 2	0.6979	0.7276	0.7864	0.7274
EIIP	0.6335	0.7030	0.7031	0.7030
Real 1	0.6253	0.6934	0.6935	0.6935
Real 2	0.5728	0.6190	0.6191	0.6190
Real 3	0.5349	0.5601	0.5603	0.5605

Table 4
Computational complexity comparison (average CPU times) for the identification of protein coding regions using different methods applied to different data sets.

Data set	Average CPU time (seconds) using different method				
	ST-DFT	ANF	ANFMA	ANFCS	ANFHS
F56F11.4 of *C. elegans*	1.7675	0.0015	0.0018	0.0064	0.0066
Mus musculus homeobox containing nuclear transcriptional factor Hmx1	1.1342	0.0010	0.0012	0.0041	0.0042
Mus musculus for PSMB5	1.0902	0.00093	0.0011	0.0040	0.0041

evaluated by computing the average CPU time of 100 runs. The computational complexity comparison using different methods and computational load required by the antinotch filters with respect to the ST-DFT method are shown in Table 4 and Table 5, respectively. The antinotch filters require only about 0.08% to 0.37% of the computational load required by the ST-DFT method. But this computational saving is achieved at the cost of identification accuracy compared to the ST-DFT method. Further, it is found from Table 5 that the computational load required by ANFMA is less, HSANF is more and CSANF is in-between.

AUC values shown in Table 3 indicate that, for all the DNA to numerical mapping techniques, the proposed CSANF, ANFMA and HSANF methods provide better identification accuracy than the ANF method for the GENSCAN test set. The performance of pro-

Table 5
Computational load required by the proposed method for the identification of protein coding regions with respect to the ST-DFT method applied to different data sets.

Data set	Computational load required by different antinotch filter with respect to ST-DFT method			
	ANF	ANFMA	ANFCS	ANFHS
F56F11.4 of *C. elegans*	0.0849%	0.1018%	0.3621%	0.3734%
Mus musculus homeobox containing nuclear transcriptional factor Hmx1	0.0882%	0.1058%	0.3615%	0.3703%
Mus musculus for PSMB5	0.0853%	0.1009%	0.3669%	0.3761%

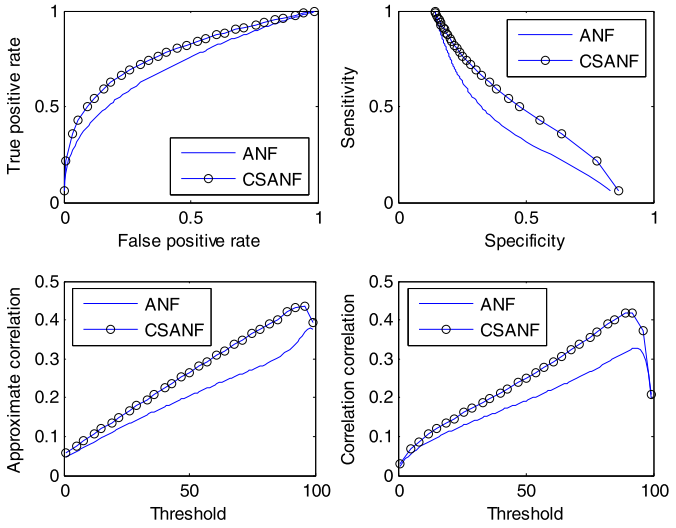

Fig. 14. Performance evaluation of ANF and CSANF methods in terms of *ROC* curve, S_n versus S_p, *CC* versus threshold and *AC* versus threshold applied to HMR195 dataset (clockwise form top left).

posed methods is analyzed in comparison with existing method at the nucleotide level using benchmark data sets. Figs. 14, 15 and 16 show *ROC* curve, S_n versus S_p, *CC* versus threshold and *AC* versus threshold using different antinotch filters for the HMR195 dataset. Further, *AUC* values using different antinotch filters for benchmark datasets are shown in Table 6. Simulation results reveal that the CSANF, ANFMA and HSANF methods outperform existing ANF method, giving improved identification of the protein coding regions. The *AUC* values indicate that the identification accuracy of the proposed antinotch filters is very close to each other. From Table 6, it is observed that, maximum *AUC* value is obtained in the CSANF method for benchmark datasets. Among proposed antinotch filters, identification accuracy is more in the CSANF method, less in the HSANF method whereas in-between in the ANFMA method.

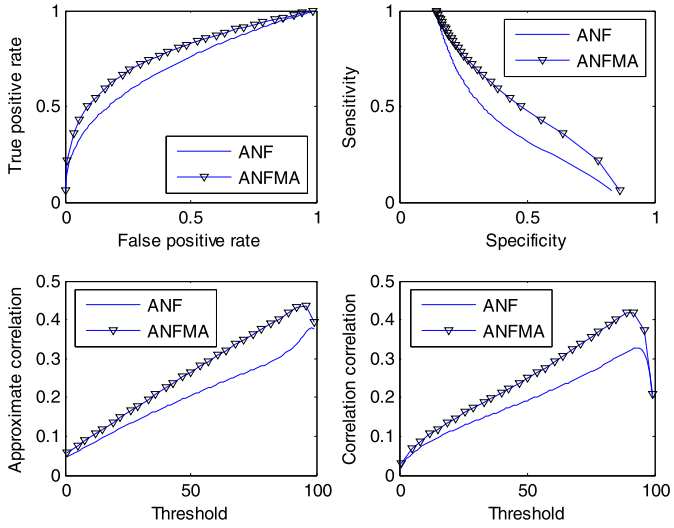

Fig. 15. Performance evaluation of ANF and ANFMA methods in terms of *ROC* curve, S_n versus S_p, *CC* versus threshold and *AC* versus threshold applied to HMR195 dataset (clockwise from top left).

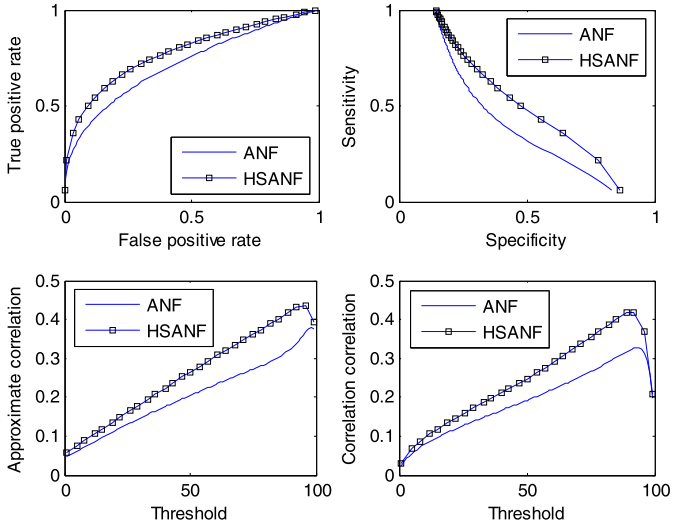

Fig. 16. Performance evaluation of ANF and HSANF methods in terms of *ROC* curve, S_n versus S_p, *CC* versus threshold and *AC* versus threshold applied to HMR195 dataset (clockwise from top left).

4. Conclusion

Three types of antinotch filters (i.e., CSANF, ANFMA and HSANF) have been proposed in this paper for the identification of protein coding regions. The performance of DNA to numerical mapping techniques using antinotch filters has been compared for the GENSCAN test set. It has been observed that Voss, *Z*-curve and tetrahedron mapping techniques give almost similar results and pro-

Table 6
AUC values for the identification of protein coding regions using antinotch filters applied to benchmark datasets.

Antinotch filter	Gene F56F11.4		HMR195		BG570		GENSCAN test set	
	AUC	% Improvement with respect to ANF	*AUC*	% Improvement with respect to ANF	*AUC*	% Improvement with respect to ANF	*AUC*	% Improvement with respect to ANF
ANF	0.7693	–	0.7066	–	0.6876	–	0.7235	–
CSANF	0.8656	12.52	0.7706	9.06	0.7470	8.64	0.8037	11.09
ANFMA	0.8654	12.49	0.7701	8.99	0.7463	8.54	0.8035	11.06
HSANF	0.8645	12.37	0.7694	8.89	0.7453	8.39	0.8032	11.02

vides maximum identification accuracy. For detailed analysis using various datasets, Voss mapping technique is preferred due to its simplicity. For the entire mapping techniques, proposed antinotch filters provide better identification accuracy than existing antinotch filter. Further, the performance of antinotch filters has been analyzed in terms of computational complexity in comparison with the ST-DFT method [4] for different datasets. Simulation results indicate that antinotch filters require only about 0.08% to 0.37% of the computational load required by the ST-DFT method. Furthermore, the identification accuracy of the proposed CSANF, ANFMA and HSANF methods has been analyzed in comparison with the existing ANF method [6,31]. Various evaluation measures reveal that the proposed methods outperform the existing method, giving improved identification of the protein coding regions. Among the proposed methods, CSANF method has slightly better identification accuracy than the ANFMA and HSANF methods.

Acknowledgments

The authors wish to thank the anonymous reviewers for their comments and valuable suggestions, which helped in improving this paper.

References

[1] D. Anastassiou, Genomic signal processing, IEEE Signal Process. Mag. 18 (2001) 8–20.
[2] P.P. Vaidyanathan, Genomics and proteomics: A signal processor's tour, IEEE Circuit Syst. Mag. (2004) 6–29.
[3] R. Guigo, DNA composition, codon usage and exon prediction, in: Genetic Databases, Academic Press, 1999, pp. 53–80.
[4] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, CABIOS 13 (1997) 263–270.
[5] W. Li, The study of correlation structures of DNA sequences: a critical review, Comput. Chem. 21 (1997) 257–272.
[6] P.P. Vaidyanathan, B.-J. Yoon, The role of signal-processing concepts in genomics and proteomics, J. Franklin Inst. 341 (2004) 111–135 (Special Issue on Genomics).
[7] D. Kotlar, Y. Levner, Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions, Genome Res. 13 (2003) 1930–1937.
[8] T.W. Fox, A. Carreira, A digital signal processing method for gene prediction with improved noise suppression, EURASIP J. Appl. Signal Process. (2004) 108–114.
[9] M. Akhtar, J. Epps, E. Ambikairajah, Time and frequency domain methods for gene and exon prediction in eukaryotes, in: Proc. IEEE ICASSP, 2007, pp. 573–576.
[10] T.S. Gunawan, E. Ambikairajah, J. Epps, A signal boosting technique for gene prediction, in: Proc. IEEE ICICS, 2007, pp. 1–4.
[11] T.S. Gunawan, J. Epps, E. Ambikairajah, Boosting approach to exon detection in DNA sequences, Electron. Lett. 44 (2008) 323–324.
[12] J. Tuqan, A. Rushdi, A DSP approaches for finding the codon bias in DNA sequences, IEEE J. Sel. Top. Signal Process. 2 (2008) 343–356.
[13] J. Mena-Chalco, H. Carrer, Y. Zana, R.M. Cesar Jr., Identification of protein coding regions using the modified Gabor-wavelet transform, IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (2008) 198–207.
[14] M.K. Hota, V.K. Srivastava, Identification of protein-coding regions using modified Gabor-wavelet transform with signal boosting technique, Int. J. Comput. Biol. Drug Des. 3 (4) (2010) 259–270.
[15] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[16] S. Rogic, A.K. Mackworth, B.F. Ouellette, Evaluation of genefinding programs on mammalian sequences, Genome Res. 11 (2001) 817–832.
[17] M. Burset, R. Guigo, Evaluation of gene structure prediction programs, Genomics 34 (1996) 353–367.
[18] C. Burge, Identification of genes in human genomic DNA, PhD dissertation, Stanford University, Stanford, CA, 1997.
[19] R.F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805–3808.
[20] B.D. Silverman, R. Linsker, A measure of DNA periodicity, J. Theoret. Biol. 118 (1986) 295–300.
[21] R. Zhang, C.T. Zhang, Z curves, an intuitive tool for visualizing and analyzing the DNA sequences, J. Biomol. Struct. Dyn. 11 (1994) 767–782.
[22] A. Rushdi, J. Tuqan, Gene identification using the z-curve representation, in: Proc. IEEE ICASSP 2006, 2006, pp. 1024–1027.
[23] P.D. Cristea, Conversion of nucleotides sequences into genomic signals, J. Cell. Mol. Med. 6 (2002) 279–303.
[24] M.K. Hota, V.K. Srivastava, DSP technique for gene and exon prediction taking complex indicator sequence, in: Proc. IEEE TENCON 2008, 2008, pp. 1–6.
[25] J. Ning, C.N. Moore, J.C. Nelson, Preliminary wavelet analysis of genomic sequences, in: Proc. IEEE Bioinformatics Conf., 2003, pp. 509–510.
[26] A.S. Nair, S.P. Sreenadhan, A coding measure scheme employing electron–ion interaction pseudopotential (EIIP), Bioinformation 1 (6) (2006) 197–202.
[27] G.L. Rosen, Signal processing for biologically-inspired gradient source localization and DNA sequence analysis, PhD thesis, Georgia Institute of Technology, 2006.
[28] N. Chakravarthy, A. Spanias, L.D. Lasemidis, K. Tsakalis, Autoregressive modeling and feature analysis of DNA sequences, EURASIP J. Appl. Signal Process. 1 (2004) 13–28.
[29] M. Akhtar, J. Epps, E. Ambikairajah, On DNA numerical representations for period-3 based exon prediction, in: IEEE Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland, 2007.
[30] A.S. Nair, S. Sreenadhan, An improved digital filtering technique using nucleotide frequency indicators for locating exons, J. Comput. Soc. India 36 (2006) 60–66.
[31] P.P. Vaidyanathan, B.-J. Yoon, Gene and exon prediction using allpass-based filters, in: Workshop on Genomic Signal Process. Stat., Raleigh, NC, 2002.
[32] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.
[33] V. Afreixo, P.J.S.G. Ferreira, D. Santos, Fourier analysis of symbolic data: A brief review, Digital Signal Process. 14 (2004) 523–530.
[34] M. Akhtar, J. Epps, E. Ambikairajah, Signal processing in sequence analysis: advances in eukaryotic gene prediction, IEEE J. Sel. Top. Signal Process. 2 (2008) 310–321.
[35] N. Rao, X. Lei, J. Guo, H. Huang, Z. Ren, An efficient sliding window strategy for accurate location of eukaryotic protein coding regions, Comput. Biol. Med. 39 (2009) 392–395.
[36] P. Ramachandran, W.-S. Lu, A. Antoniou, Location of exons in DNA sequences using digital filters, in: Proceedings of IEEE ISCAS, 2009, pp. 2337–2340.

Malaya Kumar Hota received his M.Tech. in Electronics Engineering from Visvesvaraya National Institute of Technology, Nagpur, India, in 2002 and Ph.D. in Electronics and Communication Engineering from Motilal Nehru National Institute of Technology, Allahabad, India, in 2011. Presently, he is a Professor in the Department of Electronics and Telecommunication Engineering, Synergy Institute of Engineering and Technology, Dhenkanal, Odisha, India. He has authored or co-authored about twelve publications. His biography has been included in Marquis Who's Who in Science and Engineering, 11th edition, USA. His main research interest is in genomic signal processing with special focus on DNA to numerical mapping techniques and DSP methods for the identification of protein coding regions.

Vinay Kumar Srivastava received his B.E. in ETC from GEC Rewa, MP, India, in 1989, M.Tech. in Communication from IT-BHU, Varanasi, India, in 1991 and Ph.D. in Electrical Engineering from I.I.T. Kanpur, India, in 2001. Presently, he is a Professor in the Department of ECE, MNNIT, Allahabad, India. He has about twenty years of teaching and research experience in the area of signal and image processing. He has chaired many sessions in conferences. He has authored or co-authored about thirty-five publications. His current research interest includes image compression, post-processing, digital watermarking, stability of 2D PSV system, DSP methods for the identification of protein coding regions, design and analysis of IDMA systems.