Original Article

www.jmss.mui.ac.ir

Performance Improvement of the Goertzel Algorithm in Estimating of Protein Coding Regions Using Modified Anti-notch Filter and Linear Predictive Coding Model

Mahsa Saffari Farsani, Masoud Reza Aghabozorgi Sahhaf, Vahid Abootalebi

Department of Electrical and Computer Engineering, Yazd University, Yazd, Iran

Submission: 09-04-2016 Accepted: 28-05-2016

ABSTRACT

The aim of this paper is to improve the performance of the conventional Goertzel algorithm in determining the protein coding regions in deoxyribonucleic acid (DNA) sequences. First, the symbolic DNA sequences are converted into numerical signals using electron ion interaction potential method. Then by combining the modified anti-notch filter and linear predictive coding model, we proposed an efficient algorithm to achieve the performance improvement in the Goertzel algorithm for estimating genetic regions. Finally, a thresholding method is applied to precisely identify the exon and intron regions. The proposed algorithm is applied to several genes, including genes available in databases BG570 and HMR195 and the results are compared to other methods based on the nucleotide level evaluation criteria. Results demonstrate that our proposed method reduces the number of incorrect nucleotides which are estimated to be in the noncoding region. In addition, the area under the receiver operating characteristic curve has improved by the factor of 1.35 and 1.12 in HMR195 and BG570 datasets respectively, in comparison with the conventional Goertzel algorithm.

Key words: *Anti-notch filter, deoxyribonucleic acid, Goertzel, linear predictive coding, thresholding*

INTRODUCTION

The factor that controls the transfer of certain characteristics and specificities of a species from one generation to the next is the genetic material.[1] The genetic material carries the instructions, which determine the specificities of any living organism.[2] The genetic material is made up of nucleic acids, which is found in two types: Deoxyribonucleic acid (DNA) and ribonucleic acid.[2] DNA molecules are composed of two polymer strands.[3] Each polymer strand's formula is composed of DNA monomer units or nucleotides.[3] Each nucleotide within the polymer consists of three components; a sugar (furanose-derivative deoxyribose), a heterocyclic (5-carbonic) nitrogenous base, and a phosphate group. As part of the nucleotides, bases are categorized into four different types: Adenine (*A*) and guanine (*G*) of the purine category and thymine (*T*) and cytosine (*C*) of the pyrimidine category.[4] The sugar is attached to one of the four bases through a β-glycosidic bond and makes up one of the four nucleotides: Adenosine, guanosine, cytidine, and thymidine. A nucleotide is derived from the phosphorylation of a sugar with the hydroxyl group.[4]

Address for correspondence:
Mahsa Saffari Farsani, Faculty of Electrical and Computer Engineering, Yazd University, Yazd, Iran.
E-mail: m.saffari@stu.yazd.ac.ir

Amino acids are the building blocks of proteins. The basic concern of molecular biology in the twentieth century is to create a set of genetic codes by which a strand of protein is encoded in a DNA.[5] A sequence of three nucleotides in a DNA molecule is called a codon. As the primary unit for the encoding of amino acids, each codon specifies a particular amino acid. Since there are 64 different types of codon and 20 different types of amino acids, the mapping from codons to amino acids forms a multiple-to-one relation. This means that amino acids may be specified by more than one codon. The AUG codon, which is used for coding methionine amino acids, indicates the beginning of protein synthesis in the DNA sequence.[1,6] In addition, three TAA, TAG and TGA codons, known as a stop codon or termination codon, can mark the end of protein synthesis.

How to cite this article: Farsani MS, Sahhaf MRA, Abootalebi V. Performance Improvement of the Goertzel Algorithm in Estimating of Protein Coding Regions Using Modified Anti-notch Filter and Linear Predictive Coding Model. J Med Sign Sence 2016;6:130-140.

In eukaryotes, DNA is divided into genic and intergenic regions. Only the genic region does carry data for the synthesis of proteins. Each gene consists of exon and intron regions. Exons carry the codes for the production of proteins. That is why they are called protein coding regions. Coding regions account for only about 2–5% of the entire human DNA sequence.[7]

Unlike intron regions, exon regions feature oscillating patterns. There are different periods for exon regions in eukaryotic genomes including 10.5, 200, 400, and 3 bases. Among them, the period-3 property is known as the main feature of protein coding regions in eukaryotic genomes. This feature can be due to the nonhomogeneous use of codons (i.e., codon bias). In other words, even though several codons may codify a particular amino acid, not all of them appear with equal probability in living organisms. For example, the *G* nucleotide finds its place in the codons of exon regions in certain situations.[8,9]

Several algorithms have been proposed for determining period-3 regions using signal processing. The basic idea behind signal processing techniques, as proposed by Vaidyanathan and Yoon, rests on the use of the discrete Fourier transform (DFT) and the calculation of its power spectrum.[10] The main problem with DFT-based methods is that their performance is dependent on the length of the window. The length of the window, therefore, must be so high so that the peaks caused by the period-3 patterns overcome the background noise. In the same vein, the length of the window must not be so high as to cause computational complexity and reduce the resolution for determining the initial and final exon positions. As a result, coding regions with long or short lengths, which reduce the precision of estimation, are measured by window length in DFT-based methods. To resolve this problem, the continuous wavelet transform method was proposed in.[11] In addition, the modified wavelet transform method was proposed by Singh *et al*.[12] However, the theory of using wavelet transform as an efficient tool in bioinformatics had been discussed some time earlier by Lio.[13] Filters are widely used in determining genetic regions for their higher speed compared with DFT-based methods. Using time-frequency filters as proposed by Sahu and Panda,[14] null filters as recommended by Zhang *et al*.[15] and anti-notch filters as suggested by Hota and Srivastava[16] lead to an acceptable reduction in the volume of calculations and an increase in the precision of estimation. The amplitude response of such filters has a sharp peak at $\theta = 2\pi/3$. Multistage filters are recommended for gene estimation.[17] Multistage filters are very important in signal processing because they make possible sampling at different speeds and subsequently, facilitate the use of new equipment in accordance with already existing hardware.

This article proposes a new algorithm by combining a modified anti-notch filter with linear predictive coding (LPC) model to achieve performance improvement in the Goertzel Algorithm proposed in[18] for estimating genetic regions. Furthermore, a new thresholding method has been presented to precisely identify the exon/intron regions. Using the proposed algorithm leads to reduce the correlation of signal samples. Furthermore, the execution speed of the algorithm also rises due to the use of the Goertzel algorithm. The rest of the paper is organized as follows; Section 2 introduces the database(s) used in this paper. The main stages of the proposed algorithm are presented in Section 3. Evaluation criteria for the nucleotide level are also discussed in Section 4 for comparing the proposed algorithm with other methods. Implementation results of the proposed algorithm are described in Section 5. Finally, Section 6 contains a summary of the article.

DATABASES

The proposed algorithm was applied to the gene *F56F11.4* in the *Caenorhabditis elegans* chromosome III. *C. elegans* is an intestinal parasite containing five exon regions at positions 928–1039, 2528–2857, 4114–4377, 5465–5644, and 7265–7605. This gene was extracted from the GenScan test dataset of human genes (accession no. AF099922 from the GenBank database).[19] The proposed algorithm was also applied to other genes available in two other databases – HMR195 and BG570. HMR195 is a database containing 195 sets of human, mouse, and rat genes.[20] The assessment was carried out on the *AJ223321.1* gene from this database containing one exon region at position 1196-2764. BG570 contains 570 gene sequences related to vertebrate and was created in 1996 by Burset and Guigo; the *BABAPOE* gene was selected from this database for assessment which has three coding regions in 854–896, 2654–2846, and 3467–4184.[21] Table 1 summarizes the specificities of BG570 and HMR195 databases.

PROPOSED ALGORITHM

Figure 1 shows a block diagram of the proposed algorithm for locating protein coding regions in DNA sequences. The main stages of the proposed algorithm are as follows:

- Specifying symbolic DNA sequences
- Converting symbolic DNA sequences to numeric signals using electron-ion interaction potential (EIIP) method
- Reducing background noise using a modified anti-notch filter

Table 1: A summary of HMR195 and BG570 databases

Database	Organism	Number of base-pairs	Number of genes	Number of exons	Density of protein domains (%)
BG570[21]	Vertebrates	2,892,149	570	2649	15.37
HMR195[20]	Mammals	1,383,720	195	948	14

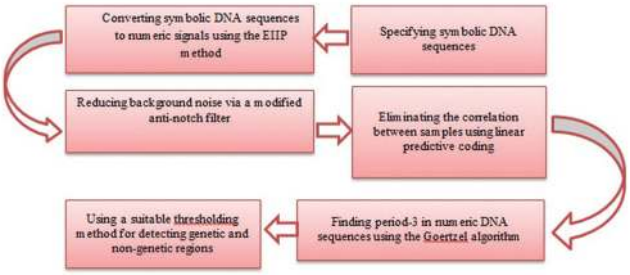

Figure 1: Block diagram of the proposed algorithm

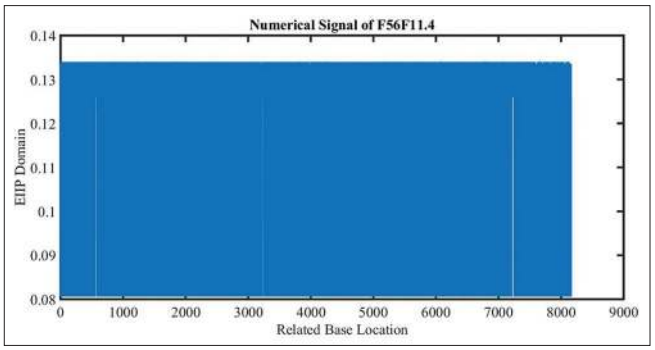

Figure 2: Primary signal of the *F56F11.4* gene sequence after converting to numerical signal by electron ion interaction potential method

- Eliminating the correlation between samples using LPC model
- Extracting period-3 patterns in numeric DNA sequences using the Goertzel algorithm, and
- Using a suitable thresholding method for detecting genetic and nongenetic regions.

Conversion of Symbolic Deoxyribonucleic Acid Sequences to Numeric Signals Using the Electron Ion Interaction Potential Method

In recent years, several methods have been proposed for mapping symbolic DNA sequences onto numerical values. Despite some differences, all these methods convert symbolic DNA sequences into numerical sequences – from at least one sequence up to four sequences. Some specificities of a desirable numeric representation of a DNA sequence are as follows:

- Each nucleotide has an equal weight
- The distance between each pair of nucleotides must be the same
- The numeric representation of a DNA sequence must be compressed; especially, the redundancy must be minimized, and
- The numeric representation of a DNA sequence must provide access to a range of mathematical tools for analysis.

In this paper, we used the EIIP mapping method for converting symbolic DNA sequences to numerical signals. This method is defined based on the electron-ion interaction in each nucleotide. EIIP values for each nucleotide are as follows; $A = 0.1260$, $G = 0.0806$, $T = 0.1335$, and $C = 0.1340$.[22]

Figure 2 shows the primary signal after converting to numerical signal for *F56F11.4* gene sequence.

Reduction of Background Noise Using a Modified Anti-notch Filter

Filters are the main tool for isolating particular frequencies in signal processing. To reduce the flux in the estimation of genetic regions, there is the need for a window with high dimensions. This leads to computational complexity and resolution reduction. To overcome this problem, we can use filters with an unlimited amplitude response known as anti-notch filters (ANF).[16] The amplitude response of such filters has a sharp peak at $\theta = 2\pi/3$. However, filters exhibit distortion at passband edges. In other words, they may detect higher or lower frequencies or attenuate some frequencies at borders. In this sense, combined filters are an efficient way to solve this problem since they overlap and ensure that no frequency will be attenuated within the desirable frequency spectrum.

ANFs are narrow band-pass filters whose its central frequency θ is $2\pi/3$. In other words, the amplitude response of ANFs has a sharp peak at $\theta = 2\pi/3$. An ANF can be calculated by reference to the real coefficients of second-order all-pass filters. It is a second order, stable and real infinite impulse response filter whose transfer function is defined as:

$$H(z)=\frac{1}{2}\times\frac{(1-R^2)(1-Z^2)}{(1-2R\cos\theta Z^{-1}+R^2Z^{-2})} \quad (1)$$

In Eq. 1, radius R at the poles is on the Z-plane. For a stable condition, we need an R^2 of lower than one ($R^2 < 1$). The frequency response of the ANF as defined in Eq. 1, can be specified by drawing radius R closer to the unit value for adjusting the sharpness of the filter. However, increasing radius R excessively to near 1 leads to visible round-off noise, and subsequently, to reduced resolution in locating exon regions. The ANF passes the frequency component at $2\pi/3$ along with its conjugate at - $2\pi/3$ and $4\pi/3$ [Figure 3a]. Conjugate frequency components are defined in relation to the complex conjugate nature of zeros and poles. These complex components contribute to the strength of the peaks in exon and intron regions. This could yield wrong measures of coding and noncoding regions. Therefore, the band-pass filter is suppressed because of the presence of conjugate frequency components. To resolve this problem, an ANF is applied in the first stage followed by a first-order complex finite impulse response (FIR) filter in the second stage. In the second phase, the first-order complex FIR filter has a zero in the unit circle at a theta rhythm of $4\pi/3$ and a pole

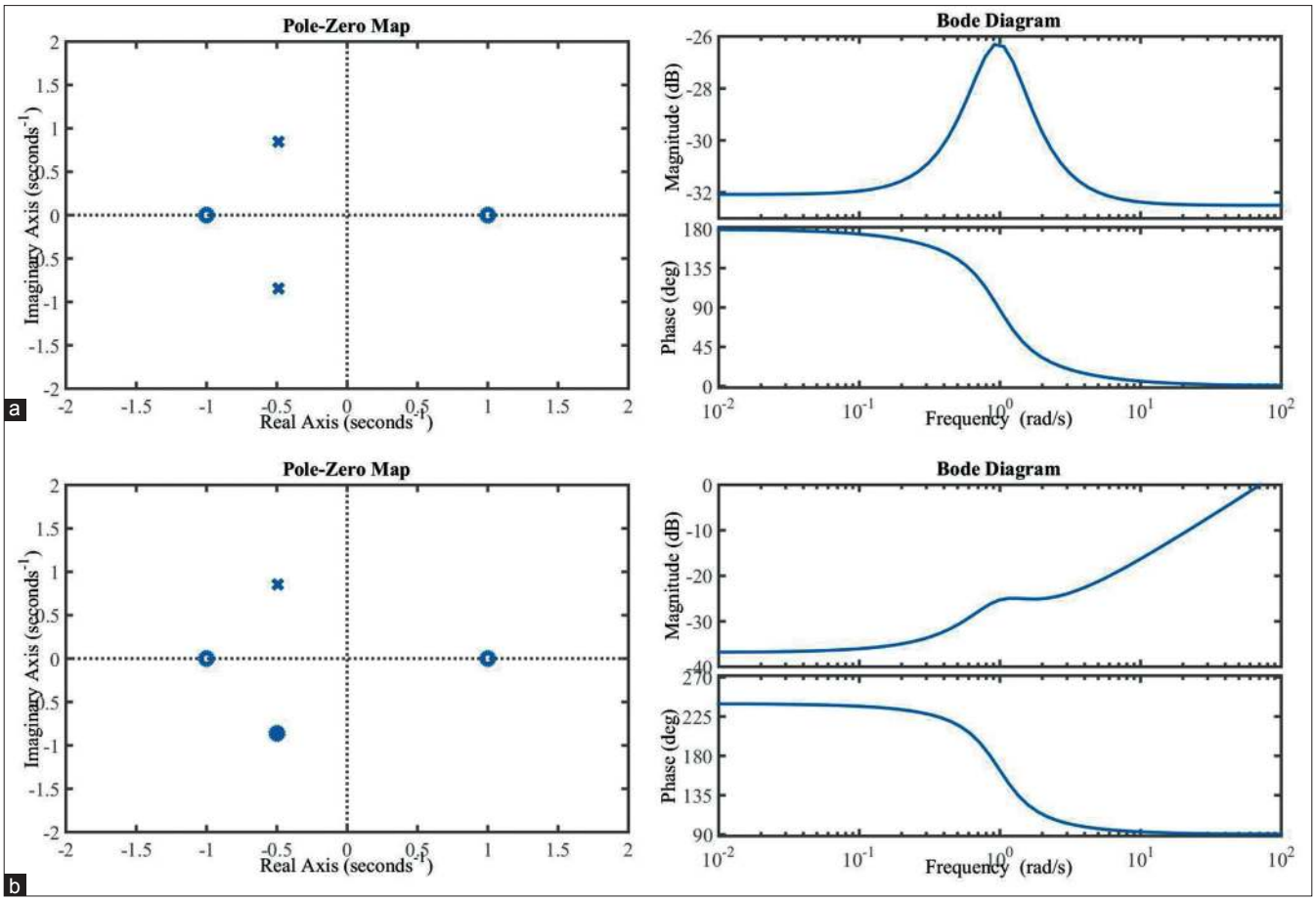

Figure 3: Zero-pole diagram and the frequency response of the (a) conventional anti-notch filters and (b) modified anti-notch filters filters

in its origin. The proposed second-stage filter is capable of suppressing frequency components at a theta rhythm of $4\pi/3$. Thus, this new filter is named conjugate suppression anti-notch filter (CSANF) with the central frequency of $2\pi/3$ [Figure 3b]. The transfer function of this filter is as follows:

$$H(z)=1-e^{j\frac{4\pi}{3}}z^{-1} \quad (2)$$

In Figure 4, the improvement impact of the proposed modified ANF filter has been shown as we discussed in theory.

Elimination of the Correlation between Samples Using Linear Predictive Coding Model

The theory of LPC in speech signal processing, and its summarization by a linear predictive coder, as effective specificities of the human speech signal, have numerous applications.[23,24] The proposed algorithm uses this technique to eliminate noise and reduce correlation between the samples. In coding region prediction methods based on spectrum estimation, sample correlation reduction is used as a technique for noise elimination and a more accurate detection of original frequencies.

Suppose that $s(n)$, $n \in 1, 2,\ldots, N$ is a DNA sequence, of N length, whose elements represent the values yielded by an EIIP mapping of the DNA strands. The objective is to estimate the volume of the $s(n)$ sample using a linear combination of N previous samples. The estimation value is represented by $\hat{s}(n)$ and is calculated with the following equation:

$$\hat{s}(n)=\sum_{k=1}^{p} a_k\, s(n-k) \quad (3)$$

where p is the level of linear prediction. a_k estimation coefficients can be calculated by minimizing the mean square error defined as follows:

$$E=\sum_{n=1}^{N} e^2(n)=\sum_{n=1}^{N}\left[s(n)-\sum_{k=1}^{p} a_k\, s(n-k)\right]^2 \quad (4)$$

by calculating the derivative of the above function in relation to a_k and equalizing it with zero, we have:

$$R_a = r \quad (5)$$

where

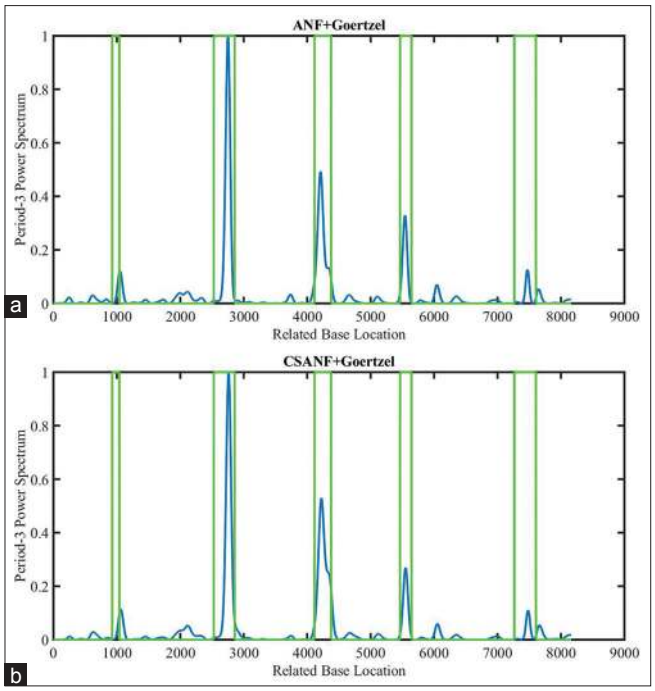

Figure 4: Results of the different algorithms for locating protein coding regions in the gene sequence *F56F11.4* (a) conventional anti-notch filters and (b) modified anti-notch filters filters

$$R=\begin{bmatrix} r(1) & 0 & . & . & . & 0 \\ r(2) & r(1) & . & & & . \\ . & r(2) & . & . & & . \\ . & . & . & . & . & . \\ . & . & & . & . & 0 \\ r(p) & . & & & . & r(1) \\ 0 & r(p) & . & & & r(2) \\ . & . & . & . & & . \\ . & & . & . & . & . \\ . & & & . & . & . \\ 0 & . & . & . & 0 & r(p) \end{bmatrix} \quad (6)$$

$$a=\begin{bmatrix} 1 \\ a(2) \\ . \\ . \\ a(p+1) \end{bmatrix} \quad (7)$$

$$r=\begin{bmatrix} 1 \\ 0 \\ . \\ . \\ 0 \end{bmatrix} \quad (8)$$

$$r(m)=\sum_{n=1}^{N} s(n)s(n+m) \quad (9)$$

where R is the autocorrelation matrix of $p \times p$, r is the autocorrelation matrix of $p \times 1$, and a is the estimated coefficient vector of $p \times 1$. LPC predicts signal samples by considering the correlation between the samples. The correlation between the samples is higher in exon regions than in intron regions. That is why the power spectrum has a peak in exon regions. In contrast, the correlation between the samples is lower in intron regions because of their biological nature. LPC causes new estimated samples to have lower/higher values in intron/exon regions, respectively. Therefore, it can be concluded that LPC yields effective specificities of gene sequences.

Calculating Period-3 Components in Numeric Deoxyribonucleic Acid Sequences Using the Goertzel Algorithm

The Goertzel algorithm is an optimal method for finding monotone components whereby the DFT obtains input data for the frequency index. This algorithm is used in the analysis of DNA sequences to extract period-3 components at $\omega = 2\pi/3$. The *Z*-transform of the Goertzel-algorithm-based FIR's is as follows:[18]

$$H_K\left(z^{-1}\right)=\frac{1-e^{j\omega_k}\ z^{-1}}{1-2\cos\omega_k\ z^{-1}+z^{-2}} \quad (10)$$

The Goertzel algorithm-based filter has two parts: Recursive and nonrecursive. DFT coefficients are obtained as the output of the system after N repetitions. The recursive section is a second order digital oscillator whose oscillation frequency is set at equal frequency intervals. In the proposed algorithm, the frequency is set at $\omega = 2\pi/3$. In practice, only the recursive section of the filter is calculated in each new sample whereas the nonrecursive section is calculated only after the N^{th} repetition, which reduces computational complexity.

EVALUATION CRITERIA AT NUCLEOTIDE LEVEL

To compare the performance of the proposed algorithm with other gene-finding methods in the literature, we use nucleotide level evaluation criteria whose parameters are defined by changing the output threshold level. The following parameters are determined to assess an algorithm:

- Number of exon nucleotides that have been identified correctly (*TP*),
- Number of exon nucleotides that have been identified as introns (*FN*),
- Number of intron nucleotides that have been identified correctly (*TN*), and

- Number of intron nucleotides that have been identified as exons (*FP*).

Based on the above parameters, the following criteria are defined.

Sensitivity, Specificity, Precision, Approximate Correlation and Mean Correlation Coefficient

The sensitivity (S_n) parameter is a measure of the proportion of encoding nucleotides that have been identified correctly. The specificity (S_p) parameter is a measure of the ratio of predicted coding nucleotides that belong to coding regions. Finally, the precision (*P*) parameter is a measure of the system's correct identification. These parameters are defined as follows:[25]

$$S_n = \frac{TP}{TP+FN} \tag{11}$$

$$S_p = \frac{TP}{TP+FP} \tag{12}$$

$$P = \frac{TP+TN}{TP+FP+TN+FN} \tag{13}$$

S_n and S_p parameters are not adequate for measuring the performance of proposed algorithms since S_p is low in high levels of S_n, and vice versa. Instead, the approximate correlation (*AC*) criterion, which is a combination of S_n and S_p, is defined as follows:[25]

$$ACP = \frac{1}{4}\left(\frac{TP}{TP+FP}+\frac{TP}{TP+FN}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right) \tag{14}$$

$$AC = 2\times(ACP-0.5)$$

Mean correlation coefficient (*Mcc*) is also another criterion, which is defined as follows:[25]

$$Mcc = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}} \tag{15}$$

System Performance Characteristic Curve

The receiver operating characteristic (ROC) curve evaluates *TP* and *FP* effects at different threshold levels and is defined as a diagram in which the true *TP* is plotted in function of *FP* via an exon-intron region separation technique at different threshold levels. The area under the ROC curve (AUC) in an algorithm is equivalent to the probability that the differentiation technique evaluates a positive, rather than a negative, random value. A higher AUC value, thus, represents a better algorithm performance.[26,27]

Sensitivity versus Specificity

Calculating S_p, *FP*, and *AC* with constant S_n, provides us with informative data for facilitating algorithm behavior analysis. In this sense, system performance improvement corresponds with lower levels of *FP* and higher levels of S_p.

IMPLEMENTATION RESULTS

Experiment 1: Gene F56F11.4 from the GenScan Database

Figure 5 shows the results of applying the proposed algorithm on gene *F56F11.4*. For comparing the performance of the proposed method, the simple Goertzel algorithm and CSANF + Goertzel methods were also implemented. As can be seen, background noise was eliminated to a large extent in the proposed method due to the use of both LPC and the modified anti-notch filter such that the correlation between samples is reduced. The simultaneous use of the Goertzel algorithm and CSANF + Goertzel methods facilitates the detection of the period-3 component in the proposed method such that a short-length exon (the first exon of the gene sequence *F56F11.4*) is identified with good precision.

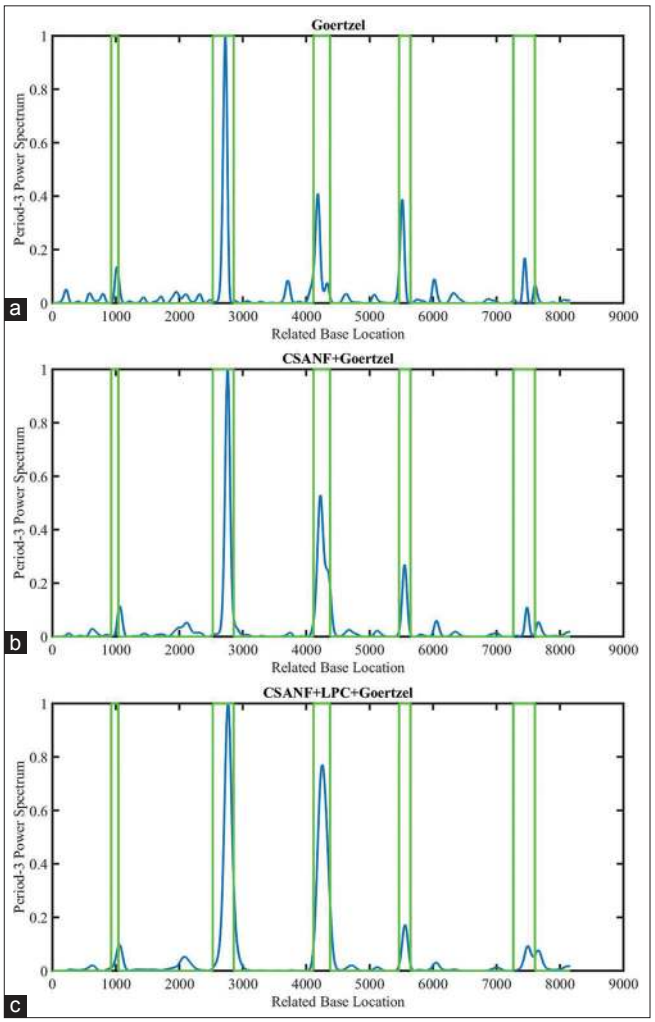

Figure 5: Results of the different algorithms for locating protein coding regions in the gene sequence *F56F11.4* (a) Goertzel (b) conjugate suppression anti-notch filter + Goertzel, and (c) proposed algorithms

Table 2 shows the quantitative values of *FP*, *AC*, S_p and AUC parameters for the gene *F56F11.4* in the proposed algorithm and the other two methods for different S_n values. As can be seen, the proposed algorithm features the highest AUC. AUC improvement in the proposed algorithm is 17.58% and 8.58% compared with the simple Goertzel algorithm and CSANF + Goertzel, respectively. In addition, the proposed algorithm features the lowest number of intron nucleotides identified as exons for all S_n values. For example, at S_n= 60%, the *FP* value is equal to 244 in the proposed algorithm, whereas it is equal to 651 and 410 in the conventional Goertzel algorithm and CSANF + Goertzel, respectively. The same condition applies to the S_p and *AC* parameters in the proposed algorithm. At S_n= 60%, the proposed algorithm yielded an improvement in the S_p parameter with coefficients of 1.41 and 1.17 compared with the conventional Goertzel algorithm and CSANF + Goertzel, respectively. The amount of *AC* also has the same superiority at S_n = 60% in the proposed algorithm. This improvement is 29.94% and 12.42% in comparison by Goertzel and CSANF + Goertzel methods, respectively. From Table 2, we can see that only at S_n = 40%, *AC* and S_p values in the proposed algorithm are lower than those in the Goertzel and CSANF + Goertzel methods.

Figure 6a plots, the ROC curve in the proposed algorithm and other methods for the gene sequence *F56F11.4*. Figure 6b and c show the S_n curve in the function of S_p and *AC* parameters based on the threshold. It should be noted that in locating genetic regions, the goal is to find the position of nucleotides in exon regions. To this end, we must search the regions near the peaks of the spectrum obtained in Figure 4 and select a suitable threshold level. A thresholding method is presented in the following equation:

$$\bar{X}=\begin{cases}1, & |X|\geq T\\ 0, & |X|<T\end{cases} \tag{16}$$

Table 2: Comparison of quantitative parameters in the proposed algorithm and other methods in the gene sequence *F56F11.4*

F56F11.4	AUC	S_n								
		40%			60%			80%		
		FP	S_p	*AC*	*FP*	S_p	*AC*	*FP*	S_p	*AC*
Method										
Goertzel	0.7832	40	0.9247	0.6113	651	0.5306	0.482	3590	0.2148	0.2145
CSANF + Goertzel	0.846	17	0.9667	0.635	410	0.6425	0.5571	2511	0.2818	0.3352
LPC + CSANF + Goertzel	0.9184	56	0.8976	0.5965	244	0.751	0.6263	693	0.5863	0.6244

CSANF – Conjugate suppression anti-notch filter, LPC – Linear predictive coding, AUC – Area under the receiver operating characteristic curve, *AC* – Approximate correlation, S_p – Specificity

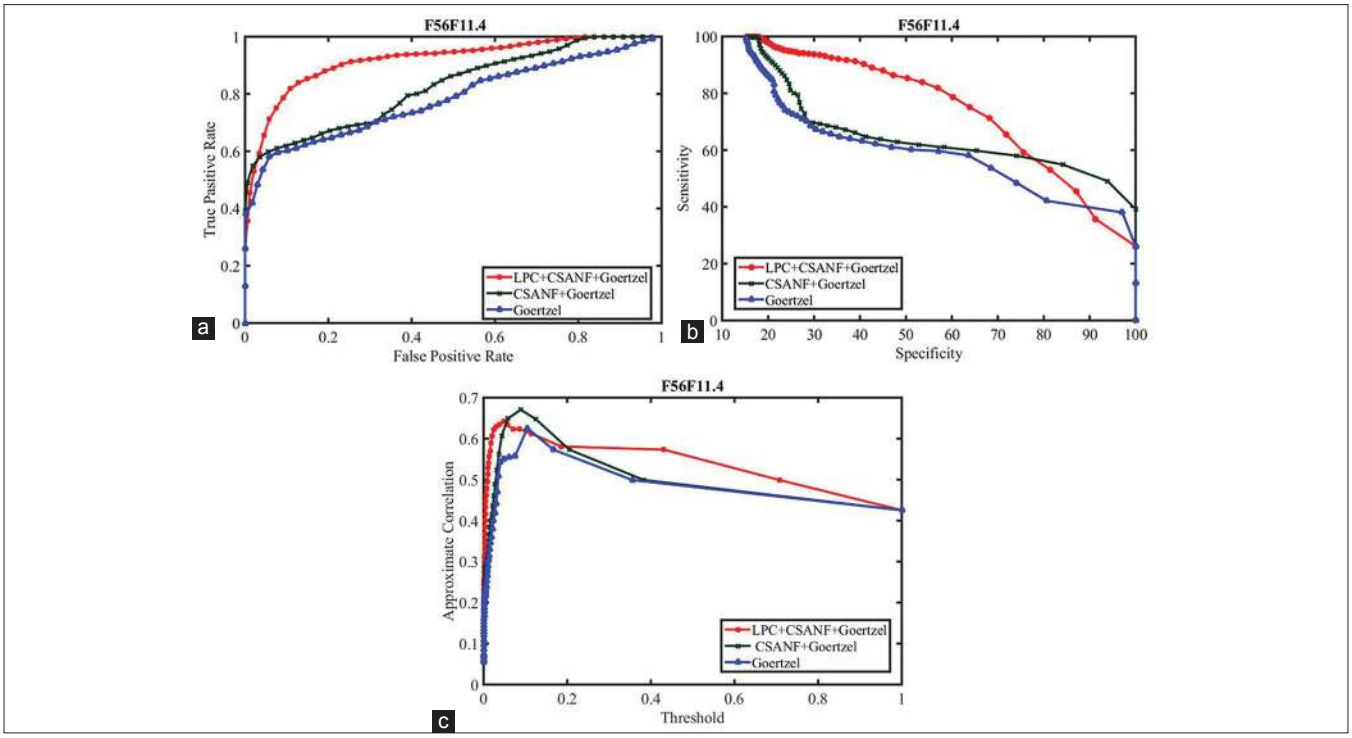

Figure 6: Comparison of different curves in the gene sequence *F56F11.4* (a) receiver operating characteristic curve, (b) sensitivity curve in terms of specificity, and (c) approximate correlation curve in terms of threshold

in Eq. 16, the borders of exon regions are detected by recourse to the start and end points of $\bar{X}$ signal's pulse 1. Selecting a suitable threshold level improves the precision of coding region detection. The important issue, thus, is how to choose a threshold level to increase the precision of detection. This paper uses the relation Eq. 17 to select the threshold level. So we have:

$$T = \frac{\text{sdP3e} \times \text{meanP3i} + \text{sdP3i} \times \text{meanP3e}}{\text{sdP3e} + \text{sdP3i}} \quad (17)$$

where sdP3e represents the standard deviation of exon regions and sdP3i represents the standard deviation of intron regions. Similarly, meanP3e represents the mean value of exon regions, and meanP3i represents the mean value of intron regions.

As shown in Figure 6c, the highest *AC* value for the proposed algorithm and CSANF + Goertzel methods occurs at the same threshold. However, in the simple Goertzel algorithm, the highest correlation occurs at a different threshold level. This indicates that the best performance for each algorithm occurs at a specific threshold. In more powerful algorithms, correlation with the threshold level decreases and an optimal yield is achieved at a fixed threshold.

Table 3 shows the position of exon regions in the gene sequence *F56F11.4* from the NCBI database as well as estimated positions using the proposed algorithm, the simple Goertzel method, and CSANF + Goertzel methods.

Table 3: The position of exon regions as yielded by the proposed algorithm, the simple Goertzel method, and conjugate suppression anti-notch filter + Goertzel methods

LPC + CSANF + Goertzel	CSANF + Goertzel	Goertzel	Exon positions in NCBI
1007-1104 (98)	1020-1105 (86)	978-1061 (84)	928-1039 (112)
2611-2940 (340)	2661-2869 (209)	2639-2793 (155)	2528-2857 (330)
4103-4431 (329)	4116-4419 (304)	3681-3741 (61)	4114-4377 (264)
5492-5623 (132)	5486-5616 (131)	4087-4352 (266)	5465-5644 (180)
7445-7699 (255)	7448-7516 (69)	5446-5574 (129)	7265-7605 (340)
		5991-6050 (60)	
		7408-7482 (75)	
		7596-7622 (27)	

CSANF – Conjugate suppression anti-notch filter, LPC – Linear predictive coding

As can be seen, exon positions obtained in the proposed algorithm are more in line with those in the NCBI database.

Experiment 2: Gene AJ223321.1 from the HMR195 Database

The second experiment was conducted on gene *AJ223321.1* from the HMR195 database. Figure 7 shows the results of the proposed algorithm applied on this gene and also the simple Goertzel, and CSANF + Goertzel methods. Furthermore, Table 4 shows the values of AUC, *FP*, *AC*, and S_p parameters for different S_n values. The superiority of the proposed algorithm is clearly visible in all of these

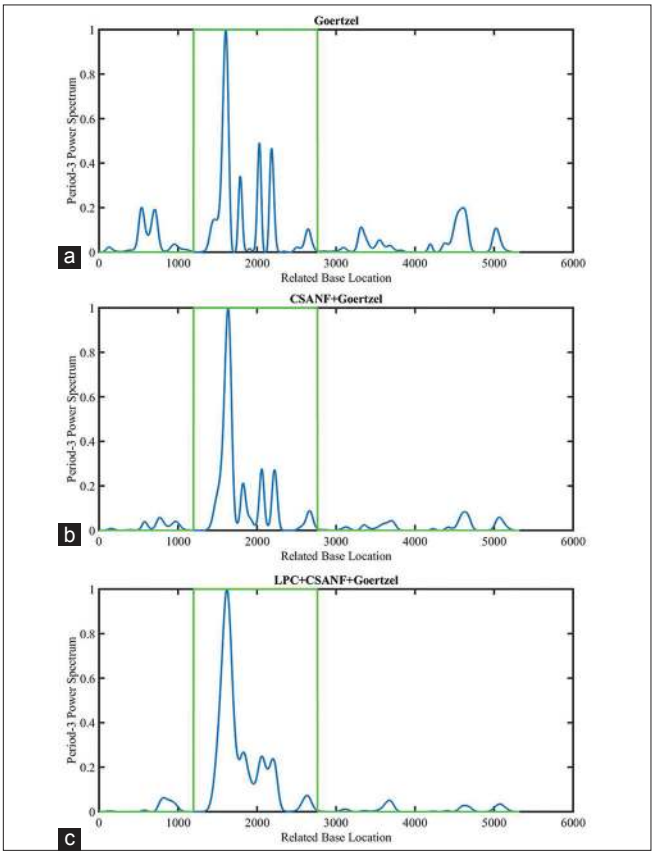

Figure 7: Results of the different algorithms for locating protein coding regions in the gene sequence *AJ223321.1* (a) Goertzel, (b) conjugate suppression anti-notch filter + Goertzel and (c) proposed algorithms

Table 4: Comparison of quantitative values of area under the receiver operating characteristic curve, *FP*, approximate correlation and specificity parameters in the proposed algorithm and other methods in the gene sequence *AJ223321.1*

AJ223321	AUC	S_n								
		40%			60%			80%		
		FP	S_p	*AC*	*FP*	S_p	*AC*	*FP*	S_p	*AC*
Method										
Goertzel	0.6224	799	0.4908	0.1728	1095	0.6333	0.3021	1236	0.2122	0.3693
CSANF + Goertzel	0.7528	383	0.7559	0.2596	596	0.6927	0.6201	776	0.5054	0.5368
LPC + CSANF Goertzel	0.8421	215	0.863	0.3053	426	0.9128	0.7285	555	0.6463	0.6211

CSANF – Conjugate suppression anti-notch filter, LPC – Linear predictive coding, AUC – Area under the receiver operating characteristic curve, *AC* – Approximate correlation, S_p – Specificity

parameters for all S_n values. At S_n = 40%, the quantity of S_P in our algorithm is improved by the factor of 2.05 and 1.1 compared with Goertzel and CSANF + Goertzel methods. Furthermore, the reduction ratio of *FP* in the proposed algorithm is more than 50% relative to two other methods. In Table 5, the values of *AC*, *Mcc*, *P*, S_n and S_P parameters by selecting the threshold as defined in Eq. 17 are shown.

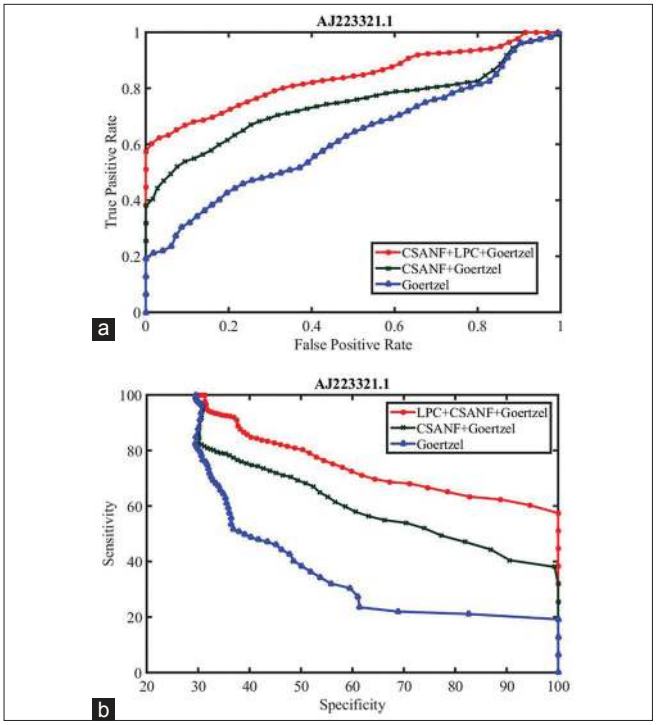

Figure 8: Comparison of different curves in the gene sequence *AJ223321.1* (a) receiver operating characteristic curve, and (b) sensitivity curve in terms of specificity

The proposed algorithm yielded a *Mcc* value of 0.6813 with improvement coefficients of 2.84 and 1.22 in relation to the simple Goertzel method and CSANF + Goertzel. This superiority can also be seen in Figure 8a and b.

Experiment 3: Gene BABAPOE from the BG570 Database

Finally, the proposed algorithm was applied to the gene sequence *BABAPOE* from the BG570 database and was compared with other methods. Results of the proposed algorithm and also the Goertzel algorithm and CSANF + Goertzel are shown in Figure 9. Table 6 presents the comparison of quantitative values of AUC, *FP*, *AC*, and S_P parameters for different S_n values. The superiority of the proposed algorithm for this gene is clearly visible. At S_n = 80%, the *FP* value in the proposed algorithm is equal to 108 whereas it is equal to 1269 in the best next method (i.e., CSANF + Goertzel). In the proposed algorithm, *AC* and S_P parameters exhibit an improvement of 13.36% and 28.30%, in comparison with CSANF + Goertzel, respectively. A similar advantage is obtained for this gene in *AC*, *Mcc*, *P*, S_n, and S_P parameters by selecting the threshold as defined in Eq. 17, [Table 7]. This advantage is also visible in Figure 10a and b, which represent the ROC curve and S_n curve in the function of S_P, respectively, based on the threshold as defined in Eq. 17.

CONCLUSION

In this paper, by combining the modified anti-notch filter and LPC model, an efficient algorithm has been presented to improve the performance of Goertzel

Table 5: Comparison of quantitative values of approximate correlation, mean correlation coefficient, accuracy, specificity, and sensitivity parameters in the proposed algorithm and other methods in the gene sequence *AJ223321.1* by selecting the threshold as defined in Eq. 17

AJ223321	Goertzel	CSANF + Goertzel	LPC + CSANF + Goertzel
AC	0.2394	0.5726	0.6892
Mcc	0.2392	0.5594	0.6813
Accuracy	0.6942	0.8265	0.8709
Specificity	0.479	0.8782	0.9375
Sensitivity	0.4219	0.478	0.6023

CSANF – Conjugate suppression anti-notch filter, LPC – Linear predictive coding, *AC* – Approximate correlation, *Mcc* – Mean correlation coefficient

Table 6: Comparison of quantitative values of area under the receiver operating characteristic curve, *FP*, approximate correlation and specificity parameters in the proposed algorithm and other methods in the gene sequence BABAPOE

BABAPOE	AUC	Sn								
		40%			60%			80%		
		FP	S_P	*AC*	*FP*	S_P	*AC*	*FP*	S_P	*AC*
Method										
Goertzel	0.8386	25	0.9386	0.6005	233	0.7109	0.577	1393	0.3542	0.3581
CSANF + Goertzel	0.8982	25	0.9387	0.6012	51	0.9184	0.7074	1269	0.3761	0.3877
LPC + CSANF + Goertzel	0.9405	0	1	0.6349	42	0.9317	0.7147	108	0.8761	0.7999

CSANF – Conjugate suppression anti-notch filter, LPC – Linear predictive coding, AUC – Area under the receiver operating characteristic curve, *AC* – Approximate correlation, S_P – Specificity

Table 7: Comparison of quantitative values of approximate correlation, mean correlation coefficient, accuracy, specificity, and sensitivity parameters in the proposed algorithm and other methods in the gene sequence BABAPOE by selecting the threshold as defined in Eq. 17

BABAPOE	Goertzel	CSANF + Goertzel	LPC + CSANF + Goertzel
AC	0.6047	0.7197	0.8202
Mcc	0.5923	0.7135	0.8201
Accuracy	0.8818	0.9139	0.9429
Specificity	0.8388	0.8965	0.8667
Sensitivity	0.5073	0.6447	0.8449

CSANF – Conjugate suppression anti-notch filter, LPC – Linear predictive coding, *AC* – Approximate correlation, *Mcc* – Mean correlation coefficient

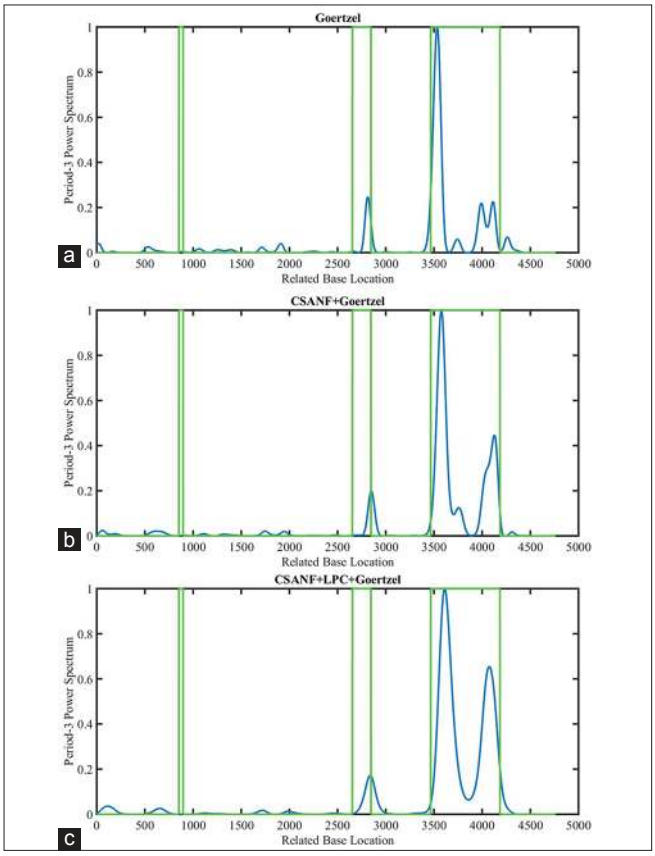

Figure 9: Results of the different algorithms for locating protein coding regions in the gene sequence *BABAPOE* (a) Goertzel, (b) conjugate suppression anti-notch filter + Goertzel and (c) proposed algorithms

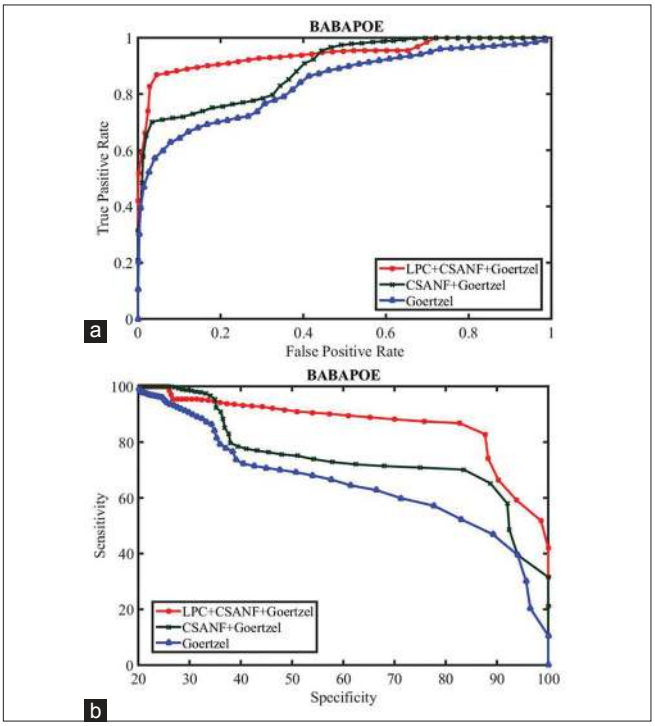

Figure 10: Comparison of different curves in the gene sequence *BABAPOE* (a) receiver operating characteristic curve, and (b) sensitivity curve in terms of specificity

algorithm in exon prediction in DNA sequences. An important advantage of the proposed algorithm is that the amount of noise reduction in it is high because of using LPC model. By comparing the performance of the proposed algorithm with other existing methods, it is seen that this algorithm, for datasets HMR195 and BG570, improves the AUC from 4.23% to 21.97%. Our proposed method also reduces the number of incorrect nucleotides which are estimated to be in the noncoding region. This reduction results in an increase of the S_p. For example, for $S_n = 0.80$, S_p recovery rate of the proposed algorithm relative to other methods is from 17.35% to 52.19% in HMR195 and BG570 database.

Many signal processing-based methods such as filter-based methods have been developed to improve the performance in gene prediction. In near future, we will consider of integrating the modified versions of LPC model and comparative methods for a hybrid signal-processing-based method in gene prediction.

Financial Support and Sponsorship

Nil.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Bolshoy A, Volkovich Z, Kirzhner V, Barzily Z. Genomic Clustering: From Lingustic Models to Classification of Genetic Texts, Studies

in Computational Intelligence. Vol. 286. Springer-Verlag Berlin Heidelberg: Springer; 2010.
2. Snustad DP, Simmons MJ. Principles of Genetics. USA: John Wiley & Sons Inc.; 2000.
3. Dougherty ER, Shmulevich L, Chen J, Wang ZJ. Genomic signal processing and statistics, EURASIP book series on signal processing and communications. USA: Hindawi Publishing Corporation; 2005.
4. Reece RJ. Analysis of Genes and Genomes. England: John Wiley & Sons Ltd.; 2004.
5. Anastassiou D. Genomic signal processing. IEEE Signal Proc Mag 2001;18:8-20.
6. Setubal J, Meidanis J. Intriduction to Computational Molecular Biology. California: PWS Publishing Company; 1999.
7. Vaidyanathan PP, Yoon BJ. The role of signal processing concepts in genomics and proteomics. J Franklin Inst Spec Issue Genomics 2004;341:111-35.
8. Wan XF, Xu D, Kleinhofs A, Zhou J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol 2004;4:19.
9. Herzel H, Trifonov EN, Weiss O, Groβe I. Interpreting correlations in biosequences. Physica A 1998;249:449-59.
10. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 1997;13:263-70.
11. Deng S, Chen Z, Ding G, Li Y. Prediction of Protein Coding Regions by Combining Fourier and Wavelet Transform. International Conference on Image and Signal processing (ICISP); 2010.
12. Singh G, Singh R, Kaur DP. Improved identification of protein coding region using wavelet transform. Int J Comput Appl 2014;92:32-7.
13. Lio P. Wavelets in bioinformatics and computational biology: State of art and perspectives. Bioinformatics 2003;19:2-9.
14. Sahu SS, Panda G. Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach. Genomics Proteomics Bioinformatics 2011;9:45-55.
15. Zhang L, Tian F, Wang S. A modified statistically optimal null filter method for recognizing protein-coding regions. Genomics Proteomics Bioinformatics 2012;10:166-73.
16. Hota MK, Srivastava VK. Identification of protein coding regions using anti-notch filters. Digit Signal Process 2012;22:869-77.
17. Guan R, Tuqan J. IIR Filter Design for Gene Identification. In Proceeding of IEEE Workshop on Genomic Signal Processing and Statistics; 2004.
18. Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A fast algorithm for exonic regions prediction in DNA sequences. J Med Signals Sens 2013;3:139-49.
19. National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine. Available from: http://www.ncbi.nlm.nih.gov/Genebank/index.html. [Last accessed on 2016 Jun].
20. Burset M, Guigo R. Evaluation of gene structure prediction programs. Genomics 1996;34:353-67. Available from: http://www.imim.es/GeneIdentification/Evaluation/Index.html.[Last accessed on 2016 Jun].
21. Rogic S, Mackworth AK, Ouellette BF. Evaluation of gene-finding programs on mammalian sequences. Genome Res 2001;11:817-32. Available from: http://www.cs.ubc.ca/~rogic/evaluation/dataset.html. [Last accessed on 2016 Jun].
22. Ning J, Moore CN, Nelson JC. Preliminary Wavelet Analysis of Genomic Sequence. In Proceeding of IEEE Bioinformatics Conference; 2003. p. 509-10.
23. Madane AR, Shah Z, Shah R, Thakur S. Speech Compression Using Linear Predictive Coding. In Proceeding of the International Workshop on Machine Intelligence Research; 2009.
24. Ganapathiraju A, Hamaker J, Picone J. Applications of support vector machines to speech recognition. IEEE Trans Signal Process 2004;52:2348-55.
25. Akhtar M, Ambikairajah E, Epps J. Detection of Period-3 Behavior in Genomic Sequences Using Singular Value Decomposition. International Conference on Emerging Technologies; 2005.
26. Fawcett T. ROC graphs: Notes and practical considerations for researchers HP laboratories. USA: Hewlett-Packard Company; 2003.
27. Ramachandran P, Lu WS, Antoniou A. Optimized Numerical Mapping Scheme for Filter-Based Exon Location in DNA Sing a Quasi-Newton Algorithm. IEEE International Symposium on Circuits and Systems (ISCAS 2010); 2010.