

A Novel Fast Algorithm for Exon Prediction in Eukaryotic Genes using Linear Predictive Coding Model and Goertzel Algorithm based on the Z-Curve

Hamidreza Saberhari, Mousa Shamsi, Hamed Heravi, Mohammad Hossein Sedaaghi
Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran.

ABSTRACT

Punctual identification of protein-coding regions in Deoxyribonucleic Acid (DNA) sequences because of their 3-base periodicity has been a challenging issue in bioinformatics. Many DSP (Digital Signal Processing) techniques have been applied for identification task and concentrated on assigning numerical values to the symbolic DNA sequence and then applying spectral analysis tools such as the short-time discrete Fourier transform (ST-DFT) to locate periodicity components. In this paper, first, the symbolic DNA sequences are converted to digital signal using the Z-curve method, which is a unique 3-D plot to illustrate DNA sequence and presents the biological behavior of DNA sequence. Then a novel fast algorithm is proposed to investigate the location of exons in DNA strand based on the combination of Linear Predictive Coding Model (LPCM) and Goertzel algorithm. The proposed algorithm leads to increase the speed of process and therefore reduce the computational complexity. Detection of small size exons in DNA sequences, exactly, is another advantage of our algorithm. The proposed algorithm ability in exon prediction is compared with several existing methods at the nucleotide level using: (i) specificity - sensitivity values; (ii) Receiver Operating Curves (ROC); and (iii) area under ROC curve. Simulation results show that our algorithm increases the accuracy of exon detection relative to other methods for exon prediction. In this paper, we have also developed a useful user friendly package to analyze DNA sequences.

Keywords

DNA sequence; Protein coding regions; Signal processing; Exon; Linear predictive coding model; Goertzel algorithm.

1. INTRODUCTION

Deoxyribonucleic acid (DNA) is of the most important chemical compounds in living cells, bacteria and some viruses [1]. A sequence of DNA is a long molecule of biopolymer origin which carries the genetic information and has the ability of expression and replication. It consists of two strands of linear polymers, each made of monomer nucleotide units. Figure 1 shows the DNA molecule structure [2]. Each nucleotide is made of three chemical components: a sugar (deoxyribose), a phosphate group, and a nitrogenous base. There are four possible bases: Adenine and guanine which are *purines* and have bicyclic structures, and cytosine and thymine which are *pyrimidines*, and have monocyclic structures. The nucleotides, based on containing bases, are often abbreviated as A, G, C, and T, respectively.

In eukaryotic cells, the DNA is divided into genes and intergenic spaces. Only genes are responsible for protein synthesis. The information encoded in the genes, is copied to a molecule called messenger RNA (mRNA). This process is called *transcription*. The next step is called *translation* which

converts mRNA into chains of linked amino acids called *polypeptides*. Proteins consist of a single type of polypeptides or combinations of different types. Every three nucleotide combination is called a *codon*. Each codon specifies one amino acid, so sequences of codons in mRNA determine specific polypeptide chains. Each gene consists of *exon* and *intron* regions (Figure 2). In transcription of eukaryotic DNA into mRNA, introns are omitted (spliced away) and only exons are translated into proteins. Therefore, exons are called protein-coding regions because they carry the necessary information for protein coding [3-5]. The proteins are the machinery of the cell and determine its properties, so detection of protein-coding regions is very important in understanding of biological functions. Protein-coding regions exhibit a period-3 behavior that is not found in other parts of the DNA and this property can be used in digital processing methods for gene detection and exon prediction purposes [6].

The attendance of long-range correlation which is considered as the background noise results in more difficulty of exon finding in DNA sequences [7-8]. Also because of the complex nature of the gene identification problem, we usually need a more efficient model that can effectively represent the characteristics of protein-coding regions. Up to now, distinctive methods have been proposed to overcome these problems which in a comprehensive categorization they can be divided into two groups; Model-dependent or supervised methods and Model-independent or Filter-based methods. Model-dependent methods like Hidden Markov Model (HMM) [9], neural network [10] and pattern recognition [11], are based on some former information gathered from the available datasets, and have been successfully used to predict exons in genes. In [9] an HMM model is proposed for gene identification which resolves three basic problems; Evaluation, Decoding and Learning problems, which can be solved using Forward and Backward, Viterbi and Baum-Welch algorithms, respectively. In [10] by introducing two new methods, one-vs-others and all-vs-all methods, and using Support Vector Machine (SVM) and Neural Network (NN) methods as base classifiers, protein fold recognition has been done.

The major drawback of using model-dependent techniques is that coding regions may not be represented in the accessible datasets but exist in the sequenced organism. To overcome this problem, Model-independent or Filter-based techniques which are based on the Fourier spectral [12-13] have been mentioned in recent years as a successful method in gene finding. Several Model-independent algorithms have been proposed in literatures for determining the protein coding regions of genes based on period-3 property. In [14] Fourier transform is used for this purpose. In this way, a fixed-length window is selected and moved on the numerical sequence. Then, we determine the exonic regions by calculating the power spectrum. In our previous work [15], we used the

Notch filter with the central frequency of $2\pi/3$ in order to remove the background noise. In [15], first, the numerical DNA sequence is passed through a notch filter and then a sliding windowed Discrete Fourier Transform (DFT) is applied on the filtered sequence. In [16] a windowless technique based on the Z-curve has been proposed to identify gene islands in total DNA sequence which called cumulative GC-Profile method. The main characteristic of the proposed method in [16] is that the resolution of the algorithm output in displaying the genomic GC content is high since no sliding window is used, but the computational complexity of this method is also high. In [17] an appropriate method is proposed to predict the protein regions by combining the DFT and Continuous Wavelet Transform (CWT). CWT leads to eliminate the high frequency noise and therefore improves the accuracy of the prediction. In [18] a new algorithm is proposed based on Fourier Transform using Bartlett window to suppress the non-exonic regions. Authors in [19] used time domain algorithms to determine the coding regions in DNA sequences. Adaptive filters [20] are one of the best tools for prediction tasks. In [21-22] authors proposed two adaptive filtering approaches based on Kalman filter and Least Mean Squares (LMS). However, the major problem with LMS is that the convergence behavior of the algorithm is slow which leads to high computational complexity. In [23] a parametric method based on autoregressive (AR) model is proposed for spectral estimation. The AR model has the advantage over the DFT that they work with smaller window sizes and, thus, shorter sequences. Using the fixed-length window is the major restriction of discussed filter-based algorithms. In many cases, size of the selected window is not successful to predict the small size coding regions. So these methods have no sensitivity to determine the protein coding regions, especially small size of exon regions. Also the run process in the mentioned algorithms is high, so there is a demand for presenting a fast and efficient method to overcome these limitations.

In this paper, a fast model-independent method based on a Linear Predictive Coding (LPCM) and Goertzel algorithm is proposed to identify the exon regions based on period-3 property in DNA sequences. First, the symbolic DNA sequence is converted into the numerical values by the Z-curve method. Then an appropriate model which named Linear Predictive Coding Model (LPCM) has been provided to remove the redundancy information in numerical DNA signal (i.e., high frequency noise components). Finally the Goertzel algorithm is applied to the estimated sequence in order to extract the period-3 components. Compared with the other existed techniques, increasing the speed of process and reducing the computational complexity are the major advantages of our proposed algorithm. Also the amount of background noise is greatly reduced in our method and the protein coding regions with small sizes are well recognized. The rest of the paper is organized as follows: In section 2 the proposed algorithm is described in details. The evaluation criteria at nucleonic level are expressed in section 3. Simulation results using Genbank database are reported in section 4. Finally, conclusion is mentioned in section 5.

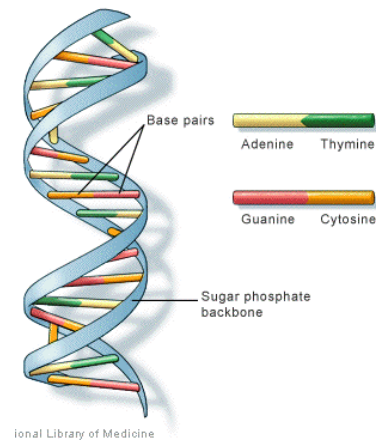


Fig 1: DNA molecule structure.

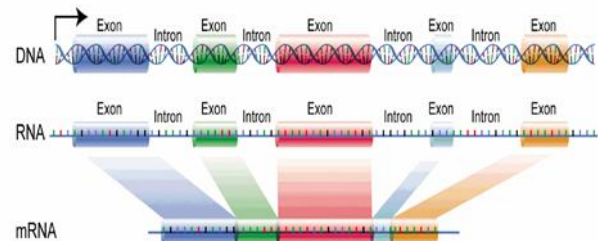


Fig 2: Exon/Intron regions for eukaryotic DNA.

2. PROPOSED ALGORITHM

Figure 3 shows the block diagram of our proposed algorithm to identify protein coding regions. The main steps of the algorithm are as follows that will be discussed in more details in this section.

- Mapping of DNA sequence to the numerical form using the Z-curve method,
- Using Linear Predictive Coding Model (LPCM) to remove the noise from the numerical sequence and achieve an appropriate model of numerical sequence,
- Using Goertzel algorithm to extract the period-3 components.

2.1 DNA numerical representation

Converting the DNA sequences into digital signals [24, 25] opens the possibility to apply signal processing methods for analyzing of genomic data and reveals features of chromosomes. The genomic signal approach has already proven its potential in revealing large scale features of DNA sequences maintained over distance of $10^6 - 10^8$ base pairs, including both coding and non-coding regions, at the scale of whole genomes or chromosomes [26-28].

There are many methods for converting the DNA sequences into numerical signals. In [5] and [12], Voss mapping technique is adapted to convert the DNA sequence into a numerical sequence. In [29] the EIIP method which is based on the electron-ion-interaction potential associated with each nucleotide is used for this purpose. In this paper, the Z-curve

representation is used to map DNA character string into numerical sequence. Based on this curve, each DNA sequence could be described separately by three independent distributions X_n , Y_n and Z_n , which each of these distributions is as follows [30]:

$$\begin{aligned} x_i &= (A_i + G_i) - (C_i + T_i) \\ y_i &= (A_i + C_i) - (G_i + T_i) \\ z_i &= (A_i + T_i) - (G_i + C_i) \end{aligned} \quad x_i, y_i, z_i \in [-1, 1], i = 1, 2, 3 \quad (1)$$

where A_i , C_i , G_i , and T_i are the cumulative numbers of the bases A, C, G and T, respectively, occurring in the sub-sequence from the first to the n^{th} base in the DNA sequence. We define $A_0 = C_0 = G_0 = T_0 = 0$. Each of these components has a biological interpretation. The first sequence X_n indicates the existence of either A or G which represents a differentiation between the purines/pyrimidine (R/Y) bases along the DNA strand. Similarly, the second

sequence Y_n represents the distribution of the amino/keto (M/K) types bases along the DNA sequence while the third sequence Z_n represents the distribution of the strong/weak hydrogen bonds (S/W). For the sub-sequence which made of the first base to the n^{th} base of DNA sequence, when purine bases (A or G) are more than pyrimidine ones (C or T) then $X_n > 0$ and if not $X_n < 0$. When the amounts of purines and pyrimidine bases are equal then $X_n = 0$. Similarly when Amino bases (A or G) are more than keto bases (C or T) then $Y_n > 0$ and if not $Y_n < 0$ and if these bases are equal then $Y_n = 0$ and finally when strong hydrogen linkage bases (A or T) are more than weak hydrogen linkage bases (G or C) $Z_n > 0$ and if not $Z_n < 0$ and in case amounts of these bases are equal then $Z_n = 0$. Therefore Z curve includes all these data that has a DNA sequence adapted to itself [31].

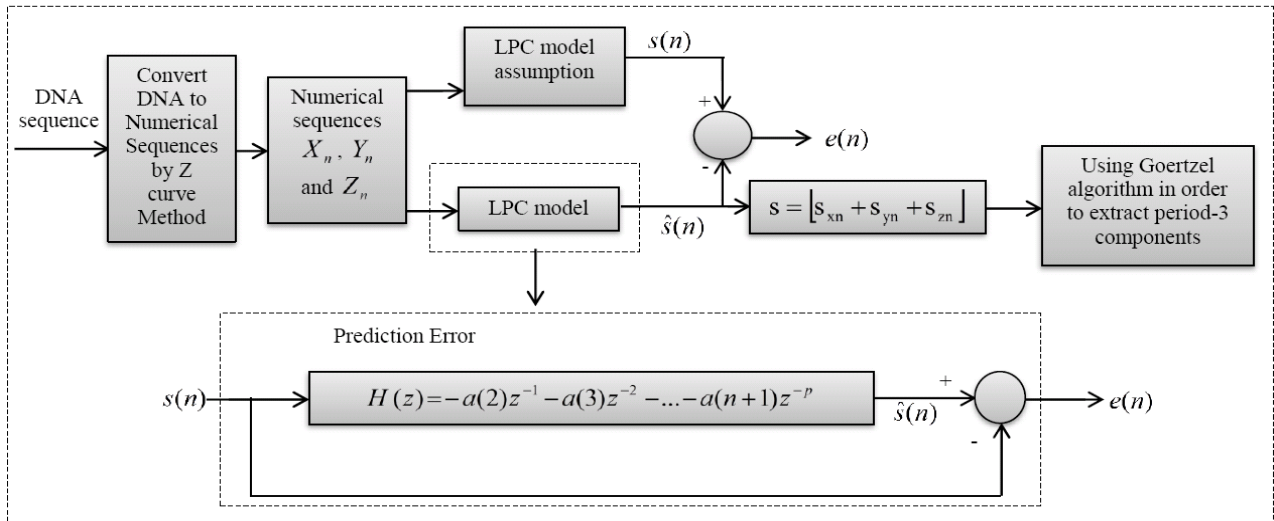


Fig 3: Block diagram of the proposed algorithm.

2.2 Linear Predictive Coding Model

Linear Predictive Coding Model [32] is one of the more efficient techniques for analyzing of non-stationary signals. The importance of this method lies to its ability to provide extremely accurate estimates of the signal parameters, and in its relative speed of computation. The main idea of this model is that a signal sample can be estimated as a linear combination of its previous samples. By minimizing the sum of the squared error between the actual signal samples and the estimated ones, a set of estimated coefficients can be calculated.

Let $s(n), n=1, 2, \dots, N$ be a protein sequence with length N whose elements are represented the Z-curve values for the corresponding amino acids. The estimated Z-curve value of each amino acid at position n is shown by $\hat{s}(n)$ and it can be calculated as a linear combination of p previous Z-curve values as follows:

$$\hat{s}(n) = \sum_{k=1}^p a_k s(n-k) \quad (2)$$

where $\{a_k\}$ are called the linear prediction coefficients. In this paper, the value of p is chosen 6.

The prediction error $e(n)$ between the observed Z-curve value $s(n)$ and the estimated Z-curve value $\hat{s}(n)$ is defined:

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

The estimated coefficients $\{a_k\}$ can be efficiently determined by minimizing the sum of squared error as follows:

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

To solve Eq. (4), we differentiate E with respect to each a_k and equate result to zero:

$$\frac{\partial E}{\partial a_k} = 0, k = 1, \dots, p. \quad (5)$$

By solving Eq. (5), a set of p linear equations is obtained as follows:

$$\sum_{k=1}^p a_k r(m-k) = r(m), m = 1, \dots, p \quad (6)$$

in which, $r(m)$ is the autocorrelation of $s(n)$, that is:

$$r(m) = \sum_{n=1}^N s(n)s(n+m) \quad (7)$$

The matrix form of Eq. (6) can be expressed as:

$$\mathbf{R}\mathbf{a} = \mathbf{r} \quad (8)$$

where \mathbf{R} is a $p \times p$ autocorrelation matrix, \mathbf{r} is a $p \times 1$ autocorrelation vector, and \mathbf{a} is a $p \times 1$ vector of prediction coefficients. So, the three parameters \mathbf{R} , \mathbf{r} and \mathbf{a} are defined as follows:

$$R = \begin{bmatrix} r(1) & 0 & \dots & 0 \\ r(2) & r(1) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 0 \\ r(p) & \dots & \dots & r(1) \\ 0 & r(p) & \dots & r(2) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & r(p) \end{bmatrix}, a = \begin{bmatrix} 1 \\ a(2) \\ \dots \\ a(p+1) \end{bmatrix}, r = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} \quad (9)$$

2.3 Goertzel algorithm

In order to apply the Goertzel algorithm, Blackman window is used to segment the estimated sequence from LPC model. The impulse response of the FIR windows has been discussed in [33]. According to [33], Blackman window has the highest amount of attenuation between the other windows. So the background noise is more suppressed by Blackman. The Goertzel algorithm is a digital signal processing technique which provides a means for efficient evaluation of individual terms of the Discrete Fourier Transform (DFT), thus results in useful certain practical application, such as dual-tone multi frequency (DTMF) signals [34], digital multi frequency (MF) receiver [35] and in a very small aperture terminal (VSAT) satellite communication system [36]. In order to express the functionality of this algorithm, it must be noted that the DFT can be formulated in terms of a convolution as follow:

$$\begin{aligned} X[k] &= \sum_{n=0}^{N-1} x[n] W_N^{nk}, \quad W_N = e^{-j\frac{2\pi}{N}nk} \\ &= \sum_{n=0}^{N-1} x[n] W_N^{-k(N-n)}, \quad W_N^{-kN} = 1 \\ &= \sum_{r=0}^{N-1} x[r] W_N^{-k(N-r)} \\ &= (x[n] * W_N^{-nk}) \\ &= [x[n] * (W_N^{-nk} u[n])] \end{aligned} \quad (10)$$

Processing of the signal $x[n]$ through an LTI filter with impulse response $h[n] = W_N^{-nk} u[n]$ and evaluating the result, $y_k[n]$ at $n=N$ will give the corresponding N-point DFT coefficient $X[k] = y_k[N]$. This LTI filtering process is illustrated as follows:

$$x[n] \Rightarrow \boxed{h[n] = W_N^{-nk} u[n]} \Rightarrow y_k[n] = x[n] * (W_N^{-nk} u[n])$$

Based on the Z-transform of the filter, we have:

$$x[n] \Rightarrow \boxed{\sum_{m=0}^{\infty} W_N^{-mk} z^{-m}} \Rightarrow y_k[n]$$

$$\begin{aligned} H(z) &= \sum_{m=0}^{\infty} W_N^{-mk} z^{-m} \\ &= \frac{1 - W_N^{-k} z^{-1}}{1 - W_N^{-k} z^{-1}} \sum_{m=0}^{\infty} W_N^{-mk} z^{-m} \\ &= \frac{\sum_{m=0}^{\infty} W_N^{-mk} z^{-m} - \sum_{m=0}^{\infty} W_N^{-(m+1)k} z^{-(m+1)}}{1 - W_N^{-k} z^{-1}} \\ &= \frac{W_N^0 z^0}{1 - W_N^{-k} z^{-1}} \\ &= \frac{1}{1 - W_N^{-k} z^{-1}} \end{aligned} \quad (11)$$

The filtering operation can be equivalently performed by the system:

$$x[n] \Rightarrow \boxed{\frac{1}{1 - W_N^{-k} z^{-1}}} \Rightarrow y_k[n]$$

Note that:

$$\begin{aligned} H(z) &= \frac{1}{1 - W_N^{-k} z^{-1}} \\ &= \frac{1 - W_N^k z^{-1}}{1 - W_N^k z^{-1}} \cdot \frac{1}{1 - W_N^{-k} z^{-1}} \\ &= \frac{1 - W_N^k z^{-1}}{1 - \left(2\cos\frac{2k\pi}{N}\right)z^{-1} + z^{-2}} \end{aligned} \quad (12)$$

So, the filtering operation can also be equivalently performed by:

$$x[n] \Rightarrow \boxed{\frac{1 - W_N^k z^{-1}}{1 - \left(2\cos\frac{2\pi k}{N}\right)z^{-1} + z^{-2}}} \Rightarrow y_k[n]$$

Hence, as can be seen from Figure 4, the Goertzel filter is composed of a recursive part and a non-recursive part. The DFT coefficients are obtained as the output of the system after N iterations. The recursive part is a second-order IIR filter (resonator) with a direct form structure. The resonant frequency of the first stage filter is set at equally spaced frequency points; that is, $\omega_k = \frac{2\pi k}{N}$ (This value is chosen $\frac{2\pi}{3}$ in our work to extract the period-3 components). The second stage filter can be observed to be an FIR filter, since its calculations do not use of the previous values of the output. In fact, we only compute the recursive part of the filter at each sample and the non-recursive part is computed only after the N th time instant when the Fourier coefficients are to be determined.

The major advantage of Goertzel algorithm is its ability to reduce the computational complexity relative to other existence methods such as DFT. This algorithm requires N real multiplications and a single complex multiplication to compute $X[k]$ for a given k . However, DFT and decimation in time FFT require N^2 and $N \log_2 N$ complex multiplications to compute $X[k]$, respectively [33].

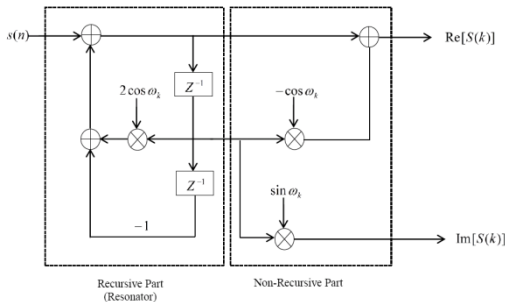


Fig 4: Filter realization of the Goertzel algorithm [33].

3. Evaluation criteria at nucleotide level

In order to compare accuracy of the different methods for protein coding regions detection the evaluation is done at nucleotide level. For this purpose, we introduce some parameters that are listed as follows:

Sensitivity and Specificity: Figure 5 shows these parameters definition, where true positive (TP) is the number of coding nucleotides correctly predicted as coding, false negative (FN) is the number of coding nucleotides predicted as non-coding. Similarly, true negative (TN) is the number of non-coding nucleotides correctly predicted as non-coding, and false positive (FP) is the number of non-coding nucleotides predicted as coding. By definition of these four quantities, the parameters sensitivity (S_n) and specificity (S_p) and precision (P) are defined as follows [37]:

$$\begin{aligned} S_n &= \frac{TP}{TP + FN} \\ S_p &= \frac{TP}{TP + FP} \\ P &= \frac{TP + TN}{TP + FP + TN + FN} \end{aligned} \quad (13)$$

Receiver Operating Characteristic (ROC) curves: The receiver operating characteristic (ROC) curves were developed in the 1950s as a tool for evaluating prediction

techniques based on their performance [38]. An ROC curve explores the effects on TP and FP as the position of an arbitrary decision threshold is varied. The ROC curve can be approximated using an exponential model as follow [39]:

$$y = \alpha \left(1 - e^{-[\beta_1 \sqrt{x} + \beta_2 x]} \right) \quad (14)$$

in which, parameters α , β_1 and β_2 can be determined by minimizing the error function:

$$E(p) = \sum_{i=1}^n \left[\alpha - \left(1 - e^{-[\beta_1 \sqrt{x_i} + \beta_2 x_i]} \right) - y_i \right]^2 \quad (15)$$

where $p = [\alpha \ \beta_1 \ \beta_2]^T$ and $\{x_i, y_i\}$ are points in the ROC plane.

Area under the ROC curve (AUC): This parameter is also a good indicator of the overall performance of an exon-location technique. The greater the AUC leads to the better performance of the tested algorithm [37].

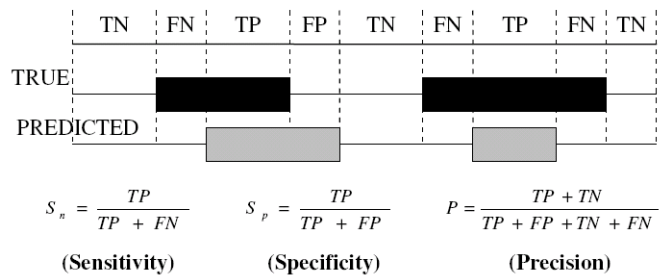


Fig 5: Parameters for evaluation of the accuracy of the algorithms.

4. Simulation Results

In order to demonstrate the performance of the discussed methods, the proposed algorithm is applied on four gene; F56F11.4, AF009962, AF019074 and AJ223321 from Genbank dataset [40]. The gene sequence F56F11.4 (GenBank No. AF099922) is on chromosome III of *Caenorhabditis elegans*. *C. elegans* is a free living nematode, about 1mm in length, which lives in temperate soil environment. It has five distinct exons, relative to nucleotide position 7021 according to the NCBI database. These regions are 3156-3267, 4756-5085, 6342-6605, 7693-7872 and 9483-9833. AF009962 is the accession number for single exon which has one coding region at position 3934-4581. The gene sequence AF019074 has the length of 6350 which has three distinct exons, 3101-3187, 3761-4574, and 5832-6007. AJ223321 is in the HMR195 dataset. This dataset consists of 195 mammalian sequences with exactly one complete either single-exon or multi-exon gene. All sequences contain exactly one gene which starts with the 'ATG' initial codon and end with a stop codon (TAA, TAG, or TGA). There is one coding region existed in AJ223321 gene sequence which its location is 1196-2764. All mentioned sequences are converted to numerical sequences using Z-curve method.

Figure 6 shows our user friendly package designed to analyze DNA sequences. This tool has been designed by our research group on genomic signal processing at Sahand University of Technology, Tabriz, Iran and consists of two main parts: the graphic display and the DSP tools for analyzing the DNA

sequences. The graphic display allows the user to view the structure record either as a graphic or as a text record in txt formats. Also it can be useful to search option for special patterns in the sequences (for example, start and stop codons in DNA sequences). The DSP tools are applying to DNA sequences in order to spectral analysis.

Briefly, there are some advantages for this tool as mentioned below:

- Loading of any DNA sequences,
- Genomic sequence representation,
- Conversion of the genomic sequence into digital values by EIIP, binary methods and z-curve,
- Search option for special patterns in the sequence,
- Applying of DSP and non-DSP methods such as DFT on the signal, and
- Prediction of the protein coding regions.



Fig 6: A view of the designed user friendly package for analyzing of the DNA sequence.

Figure 7 (a) shows the Z curve for a sample gene sequence (F56F11.4). As we can observe, it is hard to display such a long character-based sequence on computer screen. Even though the display is made possible, the extraction of any feature from the sequence is still difficult. By applying the Z curve for visualize this sequence; some features of global and local nucleotide composition of a genome can be displayed in a perceivable form. Although the screen resolution in insufficient to convey the details of the curve and help a user to display local features of the Z curve involved at single nucleotide level. To provide more detailed information than the original Z curve, three 1-D projection curves are determined. Figure 7 (b) shows three 1-D curves, which are obtained by projecting the Z curve onto x, y and z axes, respectively.

In this paper, to compare the performance of the proposed algorithm and other tested methods, we use the parameters S_n , S_p and P which were described in section 3.

Amounts of these parameters achieved from equations (13). The amounts of TP, FP, TN and FN are calculated by changing threshold level in range of 0 and 1 with small steps according Figure 8 (In this figure the value of threshold is 0.161). It can be observed in Figure 8 that if the decision threshold is very high, then there will be almost no false positives, but it won't be really identified many true positives either.

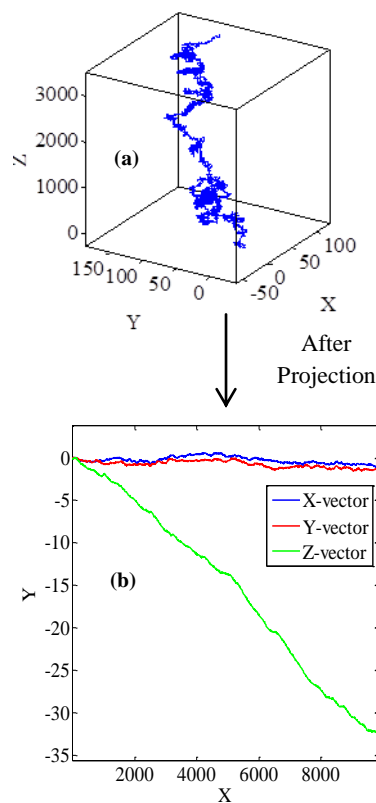


Fig 7: Z-curve representation for gene sequence F56F11.4. (a) 3-D Z curve (b) Projection curves of the original Z curve onto x, y and z axes, respectively.

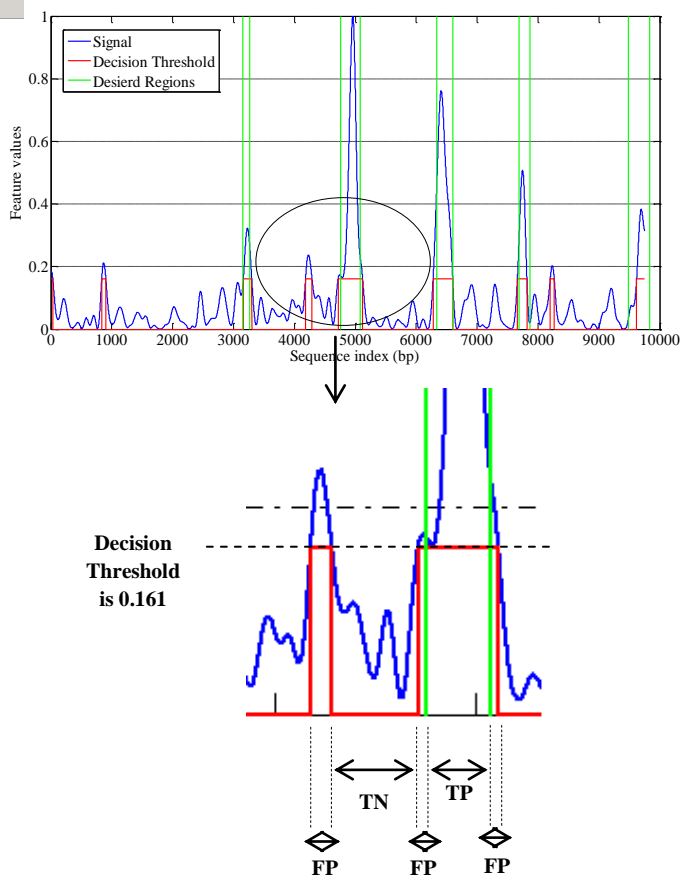


Fig 8. Parameters for exon-intron separation problem.

In this paper, to evaluate the performance of the proposed algorithm, the DFT [14] and Multi-Stage filter (MS) [41] methods are implemented. Figures 9-12 (a) and (b) show results of implementation of these methods and the proposed algorithm in identifying protein coding regions in four gene sequences explained above. As can be seen, accuracy of the DFT method for protein coding regions estimation is not high due to the noise associated with the original signal. However, the MS filter resulted a good spectral component compared to DFT and reduced the computational complexity. Also the non-coding regions are relatively suppressed in it, but this method cannot recognize the small size exonic regions. As shown in Figures 9-12 (c), the large amount of noise is removed in the proposed method due to applying the LPC model, and small size exons (For example, first exon in F56F11.4 gene sequence) can be identified because of using the Goertzel algorithm.

Figures 13 to 16 show the bar chart of sensitivity, specificity and precision in the proposed algorithm and other methods in different thresholds (from $Th=0.2$ to $Th=0.8$). As can be seen, the proposed algorithm yields the highest of these values in all threshold levels. By way of illustration, at $Th=0.2$, this algorithm exhibits relative improvements of 40% and 87% over DFT and MS filter in the Sensitivity in a typical gene sequence F56F11.4, respectively. Also the parameter S_p is improved by the factors of 3.1 and 5.7 relative to the DFT and MS filter in the sequence, respectively. Finally, the proposed algorithm shows relative improvements of 15% and 33% over the MS filter and DFT methods, respectively, in terms of Precision measure. Similar results of the proposed algorithm are apparent for the other gene sequences as shown in Figures 14-16.

In Table I, the number of false positive nucleotides, specificity and precision for specified sensitivities are

presented for the proposed and the other tested methods. According to this table, the proposed algorithm has the minimum nucleotides incorrectly identified as exons in all four gene sequences. For example in F56F11.4, at the sensitivity of 0.5, the number of false positives in the proposed method is 18bp, while this quantity for MS filter and DFT are 1052 and 1183, respectively. Also, the proposed algorithm shows relative improvements of 18.1% and 19.8% over the MS filter and DFT methods, respectively, in terms of the precision measure in the same gene sequence. Similar results of the proposed algorithm are apparent for the other three gene sequences which are shown in Table I.

To compare the computational efficiencies of the algorithms, the average CPU times over 1000 runs of the techniques were computed for the four gene sequences. Note that all of the implemented algorithms were run on a PC with a 1.6 Ghz processor (Intel (R) Pentium (R) M processor) and 2 GB of RAM. Table II summarizes results of the average CPU times. It is observed that the proposed algorithm has improved the average CPU times by the factor of 60.2, 25.75, 23.42 and 19.23 relative to the next-best performing method, DFT in F56F11.4, AF009962, AF019074 and AJ223321 gene sequences, respectively.

Finally, Figures 17-20 illustrate the ROC's of the algorithms. It is obvious that the proposed algorithm has the highest value of its parameter over the other methods. By way of illustration, the area under the ROC curve is improved by the factor of 1.35, 1.74, 1.39 and 1.76 over the DFT and 1.9, 1.73, 1.57 and 1.2 over the MS filter in F56F11.4, AF009962, AF019074 and AJ223321 gene sequences, respectively. This implies that the proposed algorithm is superior to the other methods for identifying exonic gene regions.

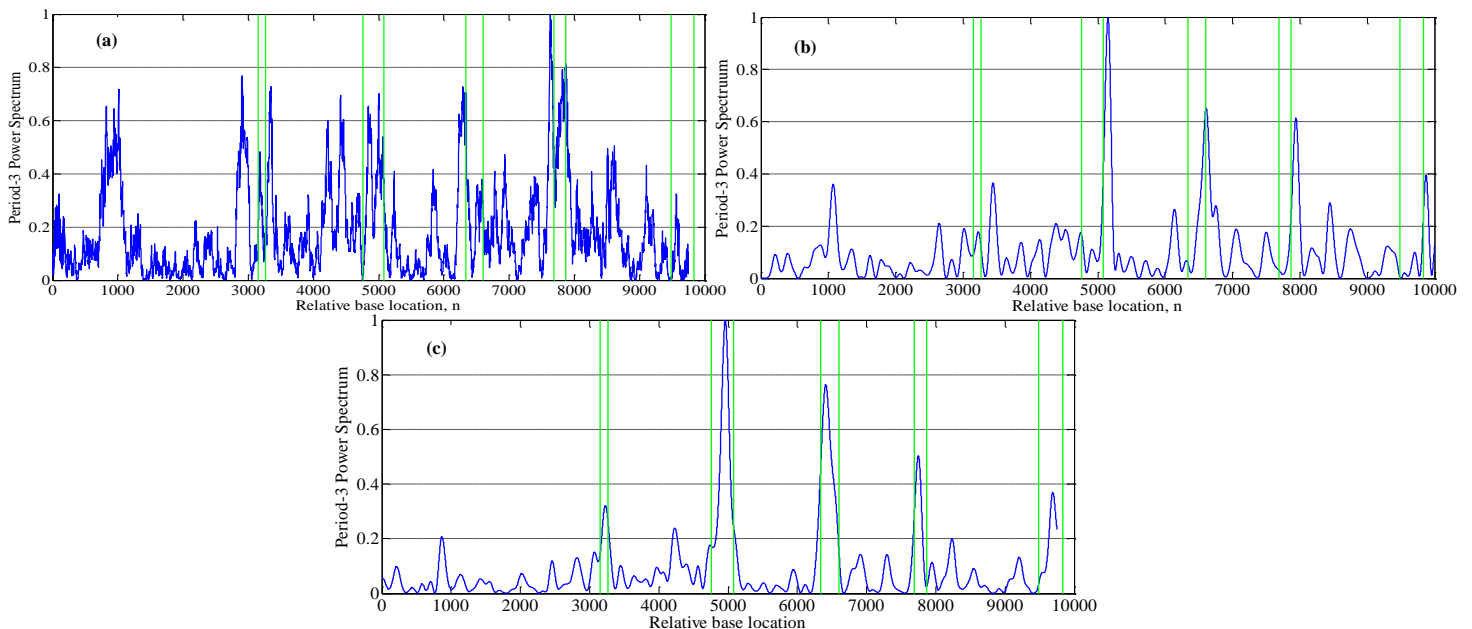


Fig 9: Results of the (a) DFT, (b) MS-filter and (c) Proposed algorithms for identification of the exonic regions on the gene sequence F56F11.4.

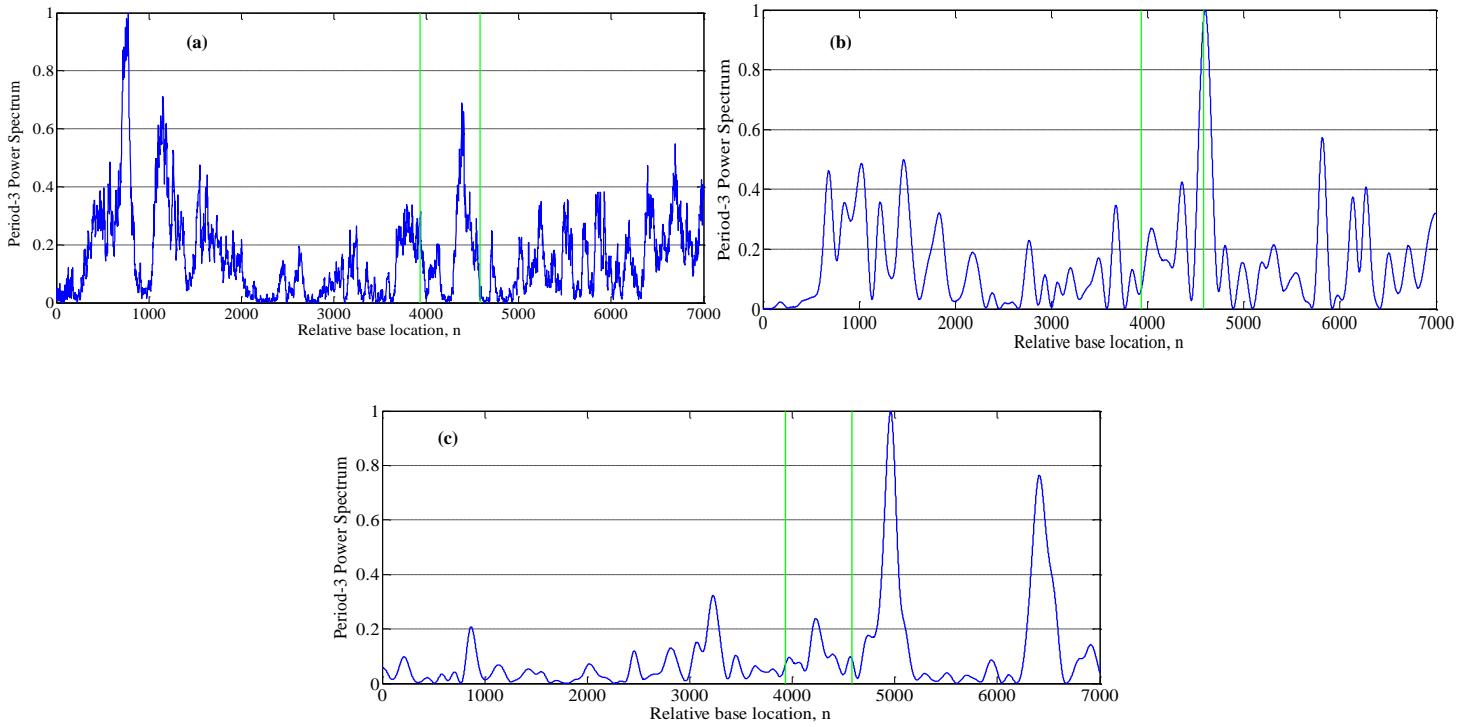


Fig 10: Results of the (a) DFT, (b) MS-filter and (c) Proposed algorithms for identification of the exonic regions on the gene sequence AF009962.

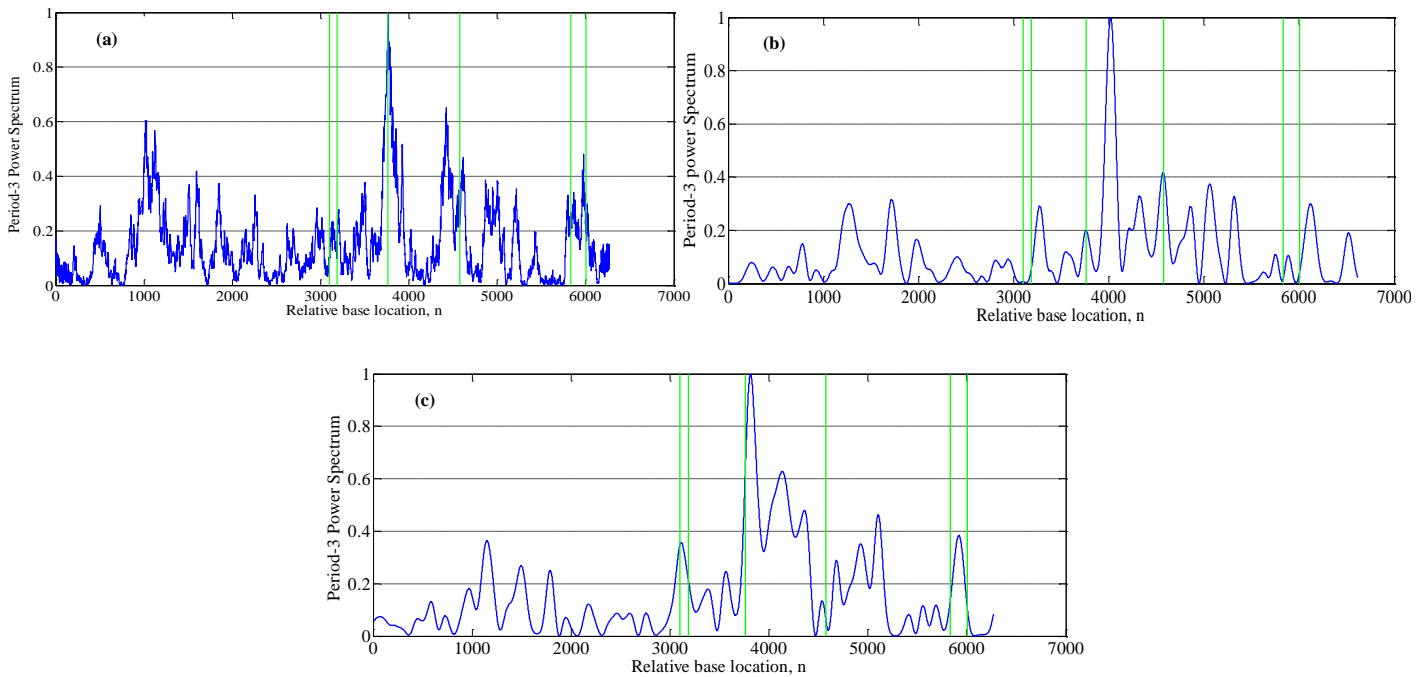


Fig 11: Results of the (a) DFT, (b) MS-filter and (c) Proposed algorithms for identification of the exonic regions on the gene sequence AF019074.

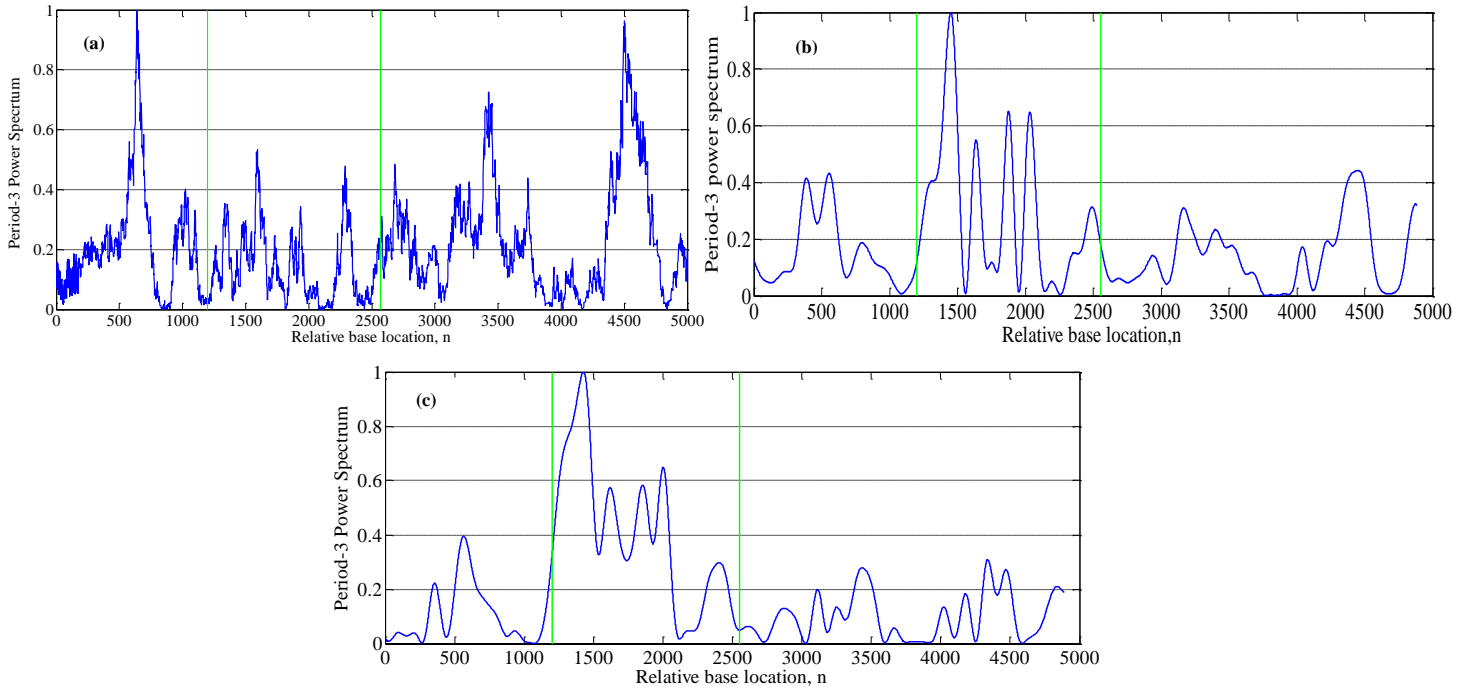


Fig 12: Results of the (a) DFT, (b) MS-filter and (c) Proposed algorithms for identification of the exonic regions on the gene sequence AJ223321.

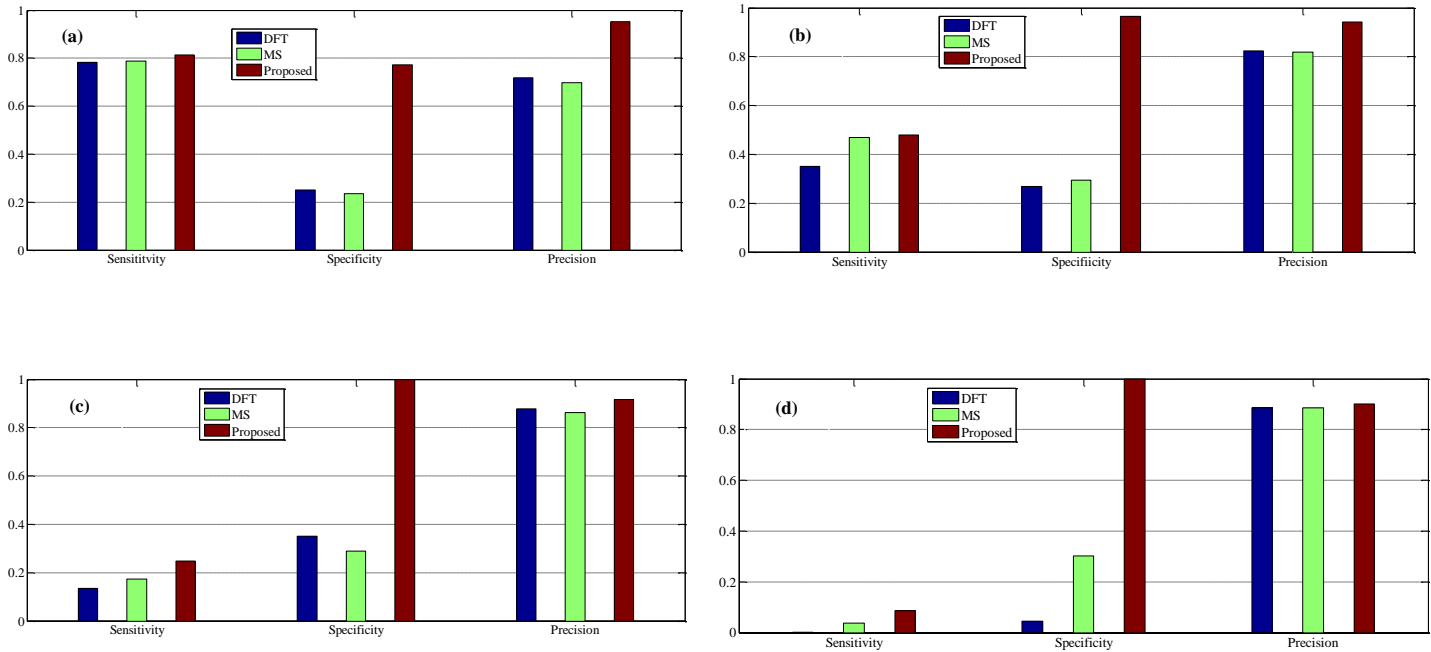


Fig 13. Comparison of sensitivity, specificity, and precision of the algorithms applied on F56F11.4 gene sequence by selecting different threshold levels. (a) Th=0.2, (b) Th=0.4, (c) Th=0.6, and (d) Th=0.8.

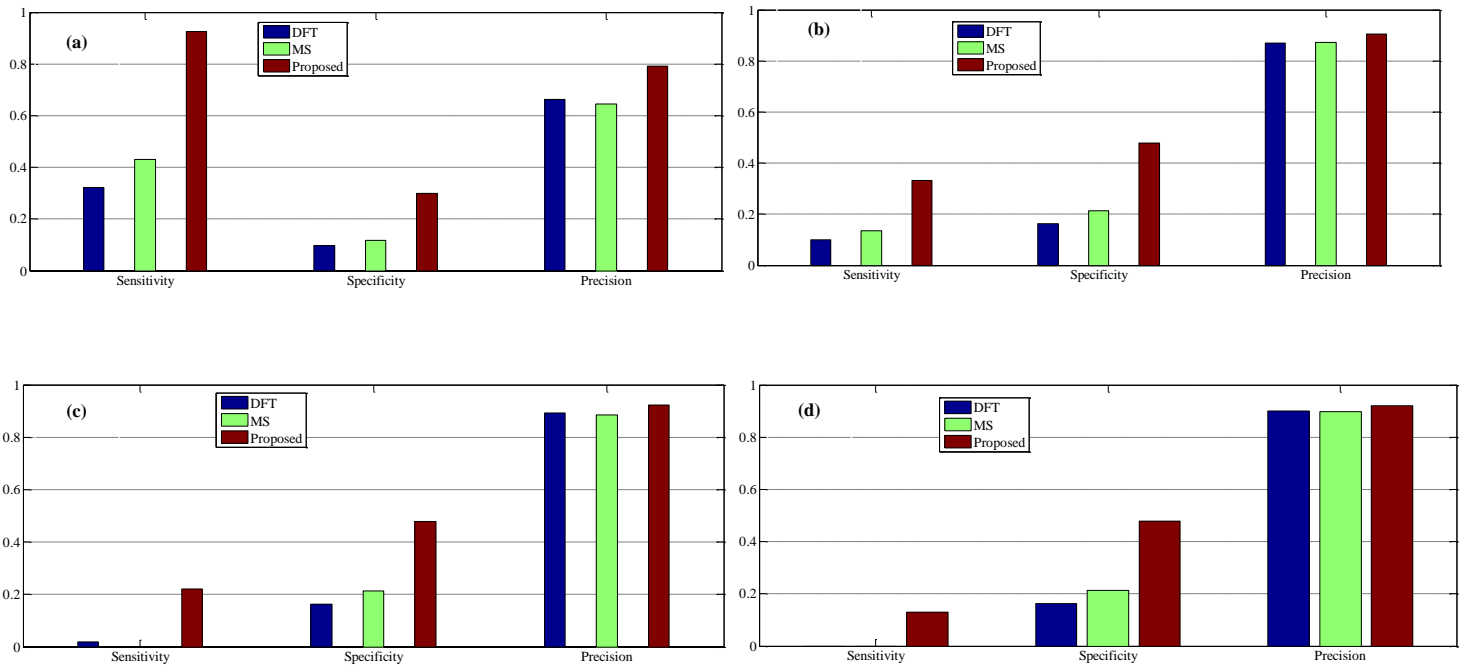


Fig 14. Comparison of sensitivity, specificity, and precision of the algorithms applied on AF009962 gene sequence by selecting different threshold levels. (a) Th=0.2, (b) Th=0.4, (c) Th=0.6, and (d) Th=0.8.

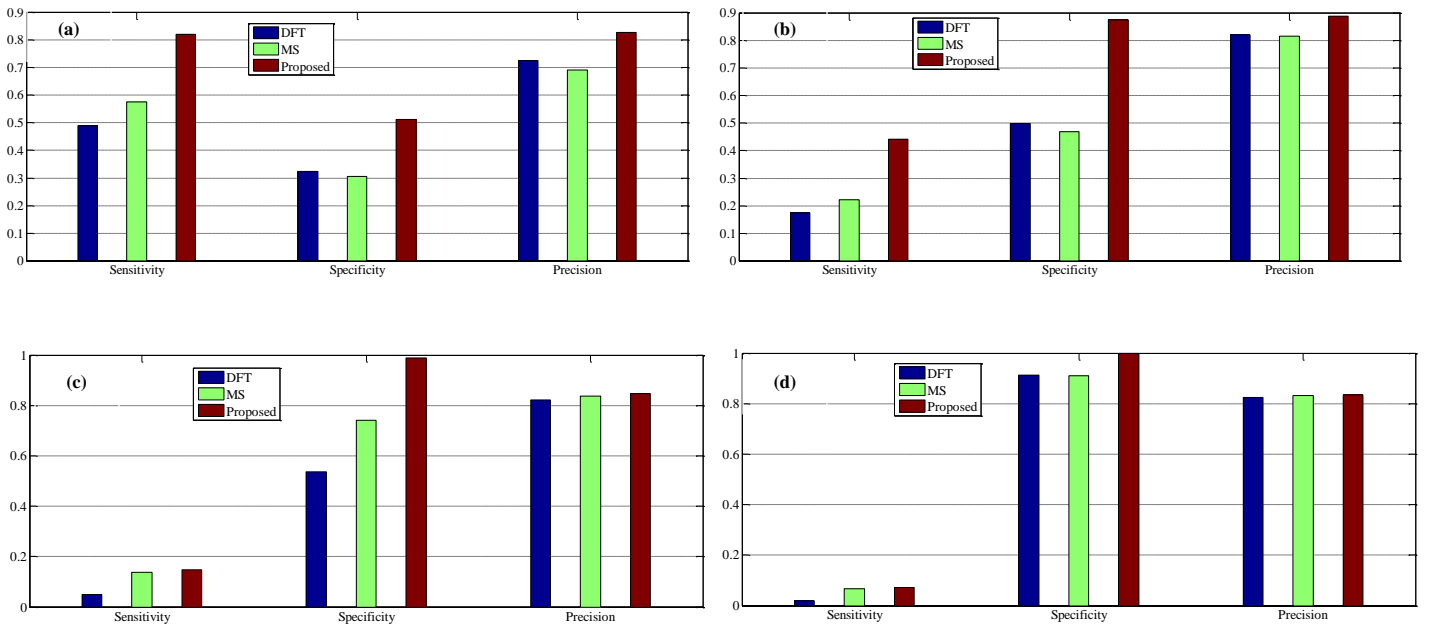


Fig 15. Comparison of sensitivity, specificity, and precision of the algorithms applied on AF019074 gene sequence by selecting different threshold levels. (a) Th=0.2, (b) Th=0.4, (c) Th=0.6, and (d) Th=0.8.

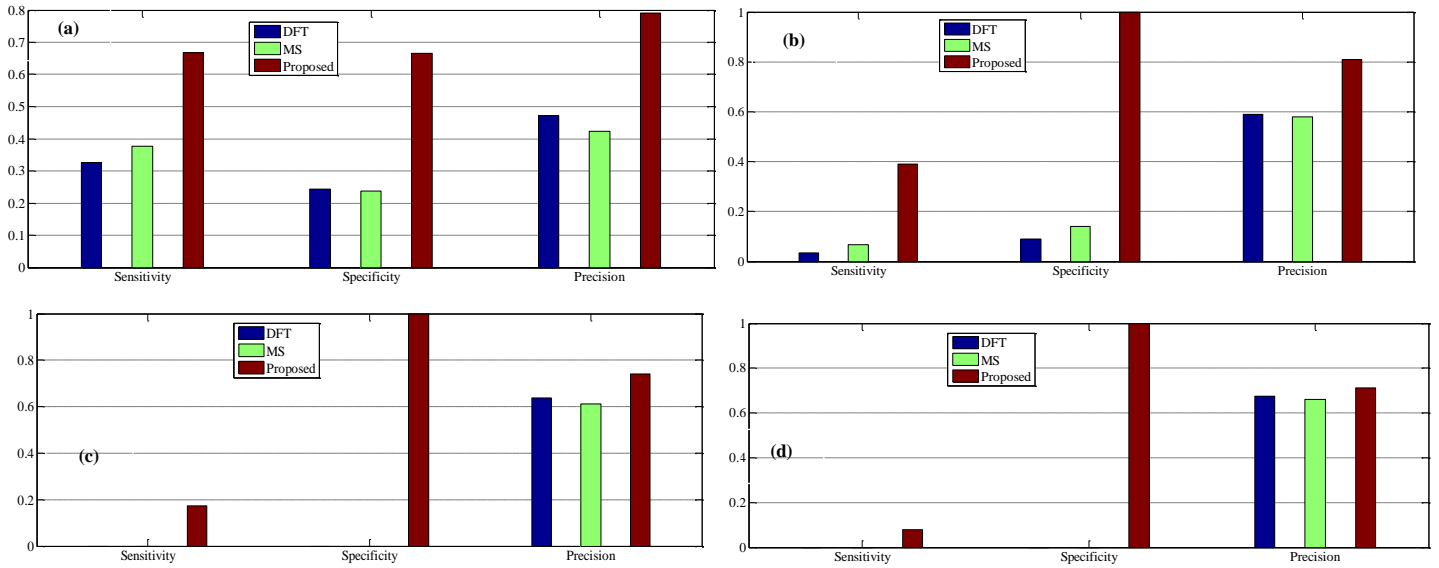


Fig 16. Comparison of sensitivity, specificity, and precision of the algorithms applied on AJ223321 gene sequence by selecting different threshold levels. (a) Th=0.2, (b) Th=0.4, (c) Th=0.6, and (d) Th=0.8.

TABLE I. Quantitative evaluation of the algorithms result using Genbank datasets.

Sequence	Method	Sn								
		0.1			0.3			0.5		
		FP	Sp	P	FP	Sp	P	FP	Sp	P
F56F11.4	Proposed	0	1	0.90	0	1	0.92	18	0.96	0.96
	MS-filter	222	0.29	0.87	620	0.29	0.842	1052	0.30	0.81
	DFT	180	0.33	0.88	711	0.27	0.83	1183	0.27	0.80
AF009962	Proposed	0	1	0.92	183	0.52	0.91	477	0.40	0.88
	MS-filter	239	0.21	0.88	1421	0.12	0.73	2467	0.12	0.60
	DFT	2791	0.11	0.55	1791	0.10	0.68	2791	0.10	0.55
AF019074	Proposed	0	1	0.84	14	0.96	0.87	79	0.87	0.90
	MS-filter	24	0.82	0.83	478	0.40	0.79	1036	0.34	0.74
	DFT	83	0.57	0.83	479	0.40	0.79	1177	0.31	0.71
AJ223321	Proposed	0	1	0.72	0	1	0.78	84	0.90	0.83
	MS-filter	2128	0.27	0.42	1660	0.22	0.45	2128	0.27	0.42
	DFT	757	0.17	0.57	1468	0.24	0.49	2173	0.27	0.41

TABLE II. Average computational time computed for the different algorithms.

Gene identifier	Sequence Length (bp)	Average Computational Time (Second)		
		Proposed algorithm	Multi-Stage filter	DFT
F56F11.4	9833	11.93	714.97	718.40
AF009962	7422	14.62	712.24	391.06
AF019074	6350	12.04	710.17	282.04
AJ223321	5321	10.05	710.51	193.29

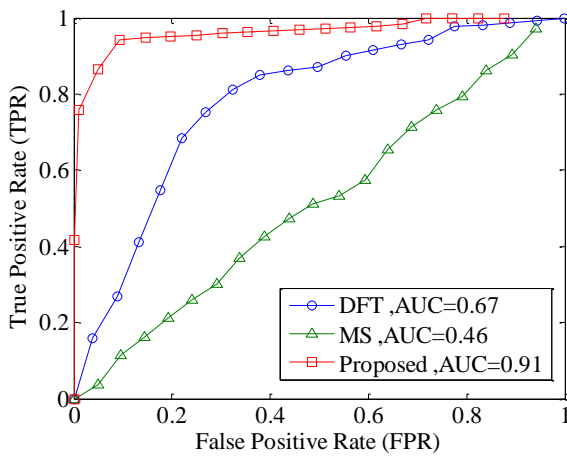


Fig 17: ROC curves of different methods for the gene sequence F56F11.4.

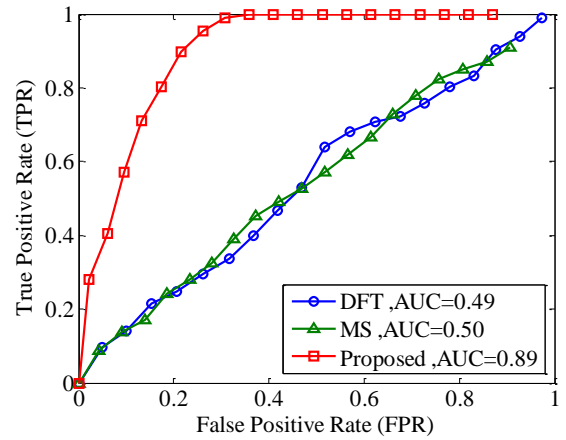


Fig 18: ROC curves of different methods for the gene sequence AF009962.

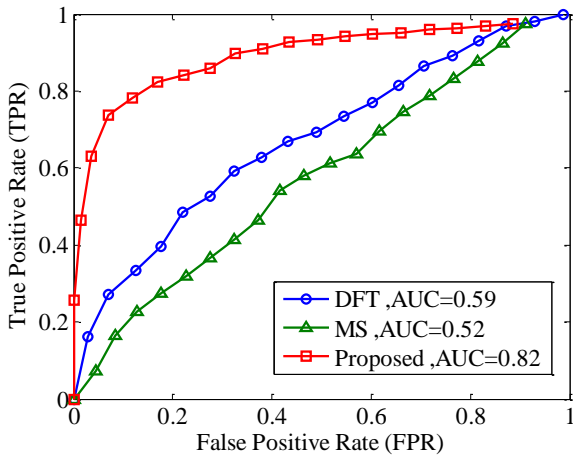


Fig 19: ROC curves of different methods for the gene sequence AF019074.

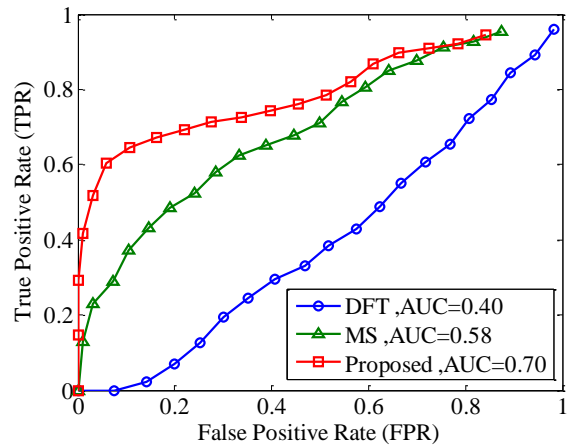


Fig 20: ROC curves of different methods for the gene sequence AJ223321.

5. Conclusion

Gene identification is a complicated problem, and the detection of the period-3 patterns is a first step towards gene and exon prediction. Many different DSP techniques have been successfully applied for the identification task but still improvement in this direction is needed. In this paper, a fast

model-independent algorithm is presented for exon detection in DNA sequences. First, Z-curve representation was used to convert the symbolic sequence into digital signal. The Z-curve method decreases the computational cost by removing the one redundant sequence from the four binary indicator sequences in the Voss representation explained in [5] and [12]. Then, the

Linear Predictive Coding Model was used to reduce the correlation between the numerical data and therefore reduce the high frequency noise. Finally, the Goertzel algorithm was applied to the estimated sequence for the period-3 detection. The proposed algorithm minimizes the number of nucleotides incorrectly predicted as coding regions which leads to increase the specificity. Also, area under the ROC curve is improved in the proposed algorithm over the other tested methods. High speed characteristic is the major advantage of our algorithm which leads to decrease the run process of the algorithm.

6. REFERENCES

- [1] Snustad D.P. and Simmons M.J., Principles of Genetics, John Wiley & Sons Inc., 2000.
- [2] Dougherty E. R, et al., Genomic signal processing and statistics, *EURASIP Book Series on Signal Processing and Communications*, 2005.
- [3] Fickett, J.W. and Tung CS, "Assessment of protein coding measures," *Nucleic Acids Res*, PP. 6441-6450, 1992.
- [4] Fickett, J.W, "The gene identification problem: an overview for developers," *Comput Chem*, vol. 20, PP. 103-118, 1996.
- [5] Vaidyanathan, P.P. and Yoon, B.J, "The role of signal-processing concepts in genomics and proteomics," *J. Franklin Inst*, PP. 111-135, 2004.
- [6] Tsonis, A. A., et al., "Periodicity in DNA coding sequences: implications in gene evolution," *J. Theor. Biol*, vol. 151, pp. 323-351, 1991.
- [7] Voss, R. F., "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Phy. Rev. Lett*, vol. 85, pp. 1342-1345, 1992.
- [8] Chatzidimitriou-Dreismann, C. A., and Larhammar, D., "Long-range correlations in DNA," *Nature*, vol. 361, pp. 212-213, 1993.
- [9] Henderson, J., et al., "Finding genes in DNA with a Hidden Markov Model," *J. Comput. Biol*, vol. 4, pp. 127-141, 1997.
- [10] Ding, C. H., and Dubchak, I., "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349-358, 2001.
- [11] Eftestel, T., et al., "Eukaryotic gene prediction by spectral analysis and pattern recognition techniques," *In Proceedings of the Seventh IEEE Nordic Signal Processing Symposium*, pp. 146-149, 2006.
- [12] Anastassiou, D., "Genomic signal processing," *IEEE Sign. Proc. Mag*, vol. 18, pp. 8-20, 2001.
- [13] Fox, T. W., and Carreira, A., "A digital signal processing method for gene prediction with improved noise suppression," *EURASIP J. Appl. Aign. Proc*, pp. 108-114, 2004.
- [14] Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., and Ramaswamy R., "Prediction of probable genes by Fourier analysis of genomic sequences," *Comput Appl Biosci*, vol. 13, pp. 263-270, 1997.
- [15] Saberkari H., Shamsi M., Sedaaghi M. H., and Golabi F., "Prediction of protein coding regions in DNA sequences using signal processing methods," *2012 IEEE Symposium on Industrial Electronics and Applications (ISIEA 2012)*, Bandung, Indonesia, pp. 354-359, September 2012.
- [16] Saberkari H., Shamsi M., and Sedaaghi M. H., "Identification of genomic islands in DNA sequences using a non-DSP technique based on the Z-Curve," *11th Iranian Conference on Intelligent Systems (ICIS 2013)*, Tehran, Iran, 27-28 February 2013.
- [17] Deng S., et al., "Prediction of Protein Coding Regions by Combining Fourier and Wavelet Transform", *International Conference on Image and Signal processing (ICISP)*, 2010.
- [18] Datta S., Asif A., "A Fast DFT-Based Gene Prediction Algorithm for Identification of Protein Coding Regions," *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing*, 2005.
- [19] Akhtar M., Epps J., Ambikairajah E., "Signal Processing in sequence Analysis: advanced in Eukaryotic gene Prediction," *IEEE journal of selected topics in signal processing*, 2008, vol. 2, pp. 310-321.
- [20] Haykin S., Adaptive Filter Theory, Fourth Edition, Prentice Hall, 2001.
- [21] Ma Baoshan., Zhu Yi-Sheng., "Kalman Filtering Approach for Human Gene Identification," *2nd International Conference on Signal Processing Systems (ICSPS 2010)*, 2010.
- [22] Ma Baoshan, "A novel adaptive filtering approach for genomic signal processing," *IEEE 10th International Conference on Signal Processing (ICSP)*, 2010, pp. 1805-1808.
- [23] Chakravarthy, N., et al., "Autoregressive modeling and feature analysis of DNA sequence," *EURASIP J. Appl. Sign. Proc*, pp. 13-28, 2004.
- [24] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279-303, 2002.
- [25] P. D. Cristea, "Genetic signal representation and analysis," *In SPIE Conference, International Biomedical Optics Symposium, Molecular Analysis and Informatics (BIOS '02)*, vol. 4623 of Proceedings of SPIE, pp. 77-84, San Jose, Calif, USA, January 2002.
- [26] J. M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences," *Hum. Mol. Genet*, vol. 6, no. 10, PP. 1735-1744, 1997.
- [27] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, no. 5423, pp.2124-2128, 1999.
- [28] P. D. Cristea, "Genomic signals of chromosomes and of concatenated reoriented coding regions," *In SPIE Conference, Biomedical Optics (BIOS '04)*, vol. 5322 of Proceedings of SPIE, pp. 29-41, SanJose, Calif, USA, January 2004, Progress in Biomedical Optics and Imaging, Vol. 5, No. 11.
- [29] K.D. Rao and M.N.S. Swamy "Analysis of genomics and proteomics using DSP techniques," *IEEE Transactions on Circuits and Systems-I*, vol. 55, no. 1, pp. 370-378, February 2008.
- [30] Zhang R. and Zhang C. T., "Z curves, an intuitive tool for visualizing and analysing the DNA sequences," *J. on Biom. Struc. Dyn.*, vol. 11, pp. 767-782, 1994.

- [31] Yan M., Lin Z. S., and Zhang C. T., “A new Fourier Transform approach for protein coding measure based on the format of the Z-curve,” *Bioinformatics*, vol. 14, no. 8, 1998.
- [32] Rabiner L. R., Schafer R. W., *Digital Processing of Speech Signals*, Prentice-Hall, Inc, 1987.
- [33] A.V. Oppenheim and R.W.Schafer, “Discrete Time Signal Processing,” *Prentice Hall, Inc*, NJ, 1999.
- [34] Braun F. Q., “Nonrecursive digital filters for detecting multifrequency code signals,” *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-23, no. 3, pp. 250-256, 1975.
- [35] Koval I., Gara G., “Digital MF receiver using discrete Fourier transform,” *IEEE transactions on Commiunication*, vol. COM-21, no. 12, pp. 1331-1335, 1973.
- [36] Simington R. A. Z., and Percival T. M. P., “New frequency domain technique for DSP based VSAT modems,” *Proc. IREE Conference*, pp. 428-431, 1991.
- [37] Burset M., and Guigo R., “Evaluation of gene structure prediction programs,” *genomics*, pp. 353-367, 1996.
- [38] Fawcett T., *ROC Graphs: Notes and Practical Considerations for Researchers* HP Laboratories, 2003.
- [39] Ramachandran P., Lu W. S., and 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.
- [40] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, <http://www.ncbi.nlm.nih.gov/Genbank/index.html>.
- [41] Vaidyanathan P. P., Yoon B. J., “Digital filters for gene prediction applications,” *Proceeding of the 36th Asilomar Conference on Signals, Systems, and Computers*, 2002.