A Fast Algorithm for Exonic Regions Prediction in DNA Sequences

Hamidreza Saberkari, Mousa Shamsi, Hamed Heravi, Mohammad Hossein Sedaaghi

Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran

Submission: 15-03-2013 Accepted: 04-05-2013

ABSTRACT

The main purpose of this paper is to introduce a fast method for gene prediction in DNA sequences based on the period-3 property in exons. First, the symbolic DNA sequences were converted to digital signal using the electron ion interaction potential method. Then, to reduce the effect of background noise in the period-3 spectrum, we used the discrete wavelet transform at three levels and applied it on the input digital signal. Finally, the Goertzel algorithm was used to extract period-3 components in the filtered DNA sequence. The proposed algorithm leads to decrease the computational complexity and hence, increases the speed of the process. Detection of small size exons in DNA sequences, exactly, is another advantage of the algorithm. The proposed algorithm ability in exon prediction was 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 confirmed that the proposed method can be used as a promising tool for exon prediction in DNA sequences.

Key words: *Algorithm, DNA sequence, discrete wavelet transform, Exon, Goertzel, protein coding region, signal processing*

INTRODUCTION

Deoxyribonucleic Acid (DNA) is of the most important chemical compounds in living cells, bacteria, and some viruses.[1] It is composed of four types of different nucleotides, namely adenine (A), cytosine (C), guanine (G), and thymine (T).[2] However, only some specific areas of the DNA molecule, which called as genes, carry the coding information for protein synthesis. In eukaryotic cells, the DNA is divided into genes and inter-genic spaces. Genes are further divided into exon and intron, which is shown in Figure 1. Genes are responsible for protein synthesis; therefore, they are called protein-coding regions because they carry the necessary information for protein coding.[3-5] Protein-coding regions exhibit a period-3 behavior due to the codon bias involved in the translation process. This phenomenon caused background noise, which leads to more difficult of exon finding in DNA sequences.[6,7]

Nowadays, there are many digital signal processing (DSP) methods presented in literatures to identify the protein coding regions and also reduce the background noise in DNA sequences, which are based on Fourier spectral. In Tiwari *et al*,[8] Fourier transform is used for this purpose. In this way, a fixed-length window is selected and moved on the numerical sequence. Then, the exonic regions are determined by calculating the power spectrum. In our previous work,[9] the notch filter with the central frequency of 2π/3 was used in order to remove the background noise. First, the DNA sequence is passed through a notch filter and then a sliding windowed discrete Fourier transform (DFT) is applied on the filtered sequence. In Saberkari *et al*,[10] a windowless technique based on the Z-curve was implemented to identify gene islands in total DNA sequence which called cumulative GC-Profile method. The main characteristic of this method 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 Deng *et al*,[11] an appropriate method is proposed to predict the protein regions by combining the DFT and continues wavelet transform (CWT). CWT leads to eliminate the high frequency noise and, therefore, improves the accuracy of the prediction. In Datta *et al*,[12] a new algorithm is proposed based on Fourier transform using Bartlett window to suppress the non-exonic regions. In Akhtar *et al*,[13] the time domain algorithms have been used to determine the coding regions in DNA sequences. Adaptive filters[14] are one of the best tools for prediction tasks. In Baoshan *et al*,[15,16] two adaptive filtering approaches based on Kalman filter and least mean squares (LMS) are proposed for human gene identification. However, the major problem with LMS is that

Address for correspondence:
Dr. Mousa Shamsi, Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran. E-mail: shamsi@sut.ac.ir

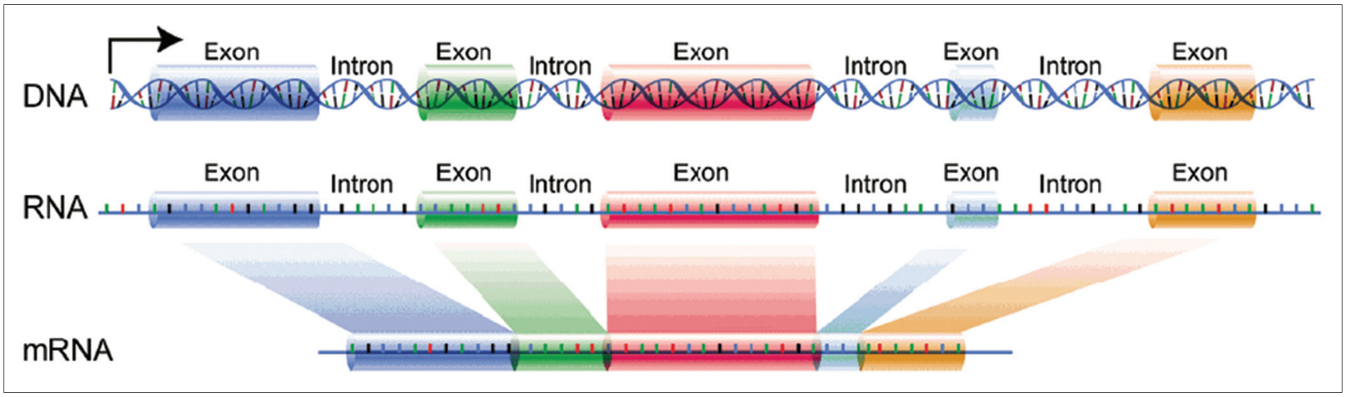


Figure 1: Exon/Intron regions for eukaryotic DNA[2]

the convergence behavior of the algorithm is slow, which leads to high computational complexity. A parametric method based on autoregressive (AR) model proposed in Chakravarthy *et al*.,[17] for spectral estimation. The AR model has the advantage over the DFT that it works with smaller window sizes and, thus, shorter sequences.

In this paper, a fast method based on DWT and Goertzel algorithm is proposed to determine the location of exons in DNA sequences. The proposed algorithm improves the accuracy of the prediction, especially in detection of the small size exons. The rest of the paper is organized as follows: Section II describes the proposed algorithm in details. The evaluation criteria at nucleonic level are expressed in Section III. Section IV shows simulation results using Genbank database. Finally, Section V concludes the experiments and algorithms.

THE PROPOSED ALGORITHM

Figure 2 shows block diagram of the 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.

- Numerical mapping of DNA sequence using EIIP method,
- Using DWT to remove the noise from the numerical sequence,
- Choosing Blackman window with the length 351 and sliding it on the filtered sequence, and
- Using Goertzel algorithm to extract the period-3 components.

DNA Numerical Representation

Converting the DNA sequences into digital signals[18,19] opens the possibility to apply signal processing methods for analyzing 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.[20-22]

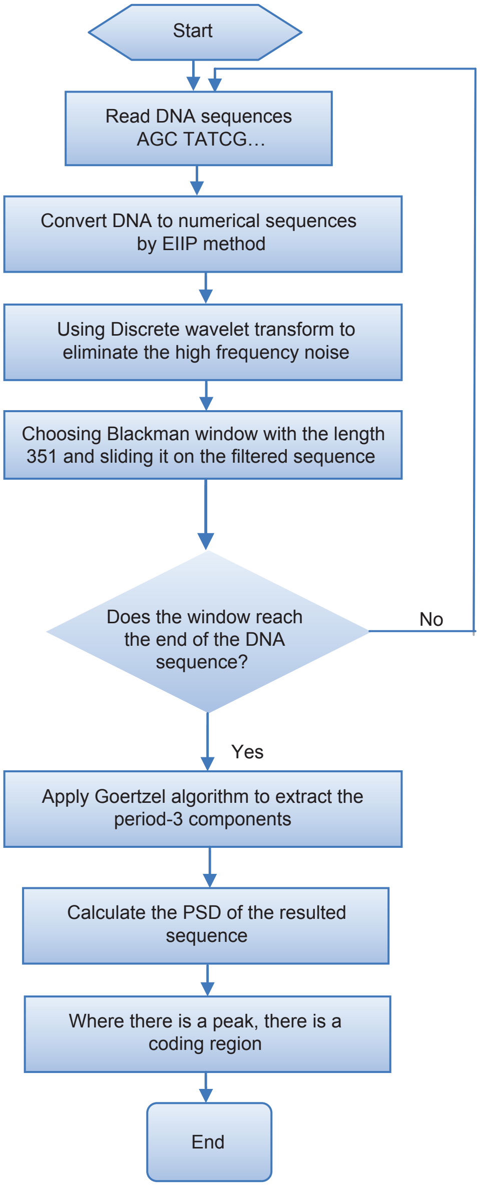

Start
Read DNA sequences AGC TATCG…
Convert DNA to numerical sequences by EIIP method
Using Discrete wavelet transform to eliminate the high frequency noise
Choosing Blackman window with the length 351 and sliding it on the filtered sequence
Does the window reach the end of the DNA sequence?
No
Yes
Apply Goertzel algorithm to extract the period-3 components
Calculate the PSD of the resulted sequence
Where there is a peak, there is a coding region
End

Figure 2: Block diagram of the proposed algorithm

There are many methods for converting the DNA sequences into numerical signals like VOSS mapping,[6] Z-curve,[23] and

EIIP.[24] In VOSS technique, the background noise is more dominant because the magnitude for each base is the same (i.e. 1 represents the presence of the nucleotide and 0 for its absence). The Z-curve technique is a 3-D curve for representing the DNA sequence. In this method, the dimension is reduced by projecting each 3-D curve into x-y axes, which leads to more computational complexity. In this paper, we have used EIIP method to convert DNA sequence into numerical signal. This approach allows DNA representations with either one or four sequence (s). It can be noticed that the EIIP representation has a different magnitude for each base and the distances among them are unequal. In this method, the electron-ion-interaction potential associated with each nucleotide is used for mapping of the DNA sequence. The EII*P* values for the nucleotides are: A = 0.1260, G = 0.0806, T = 0.1335, C = 0.1340.[24]

Using Discrete Wavelet Transform to Reduce the High Frequency Noise

In this paper, DWT is applied on the input numerical sequence to remove the high frequency noise and hence, improve the accuracy of the algorithm for exonic region identification. In DWT, the signal is passed first through the high and low pass filters, then by down-sampling the filtered signal, samples are divided into two signals; high frequency samples (detail signals) and low frequency ones (approximation signals). The DNA numerical signal, *x[n]*, is passed first through the high pass filter, $g[n]$, then through the low pass filter, $h[n]$. So, we have:

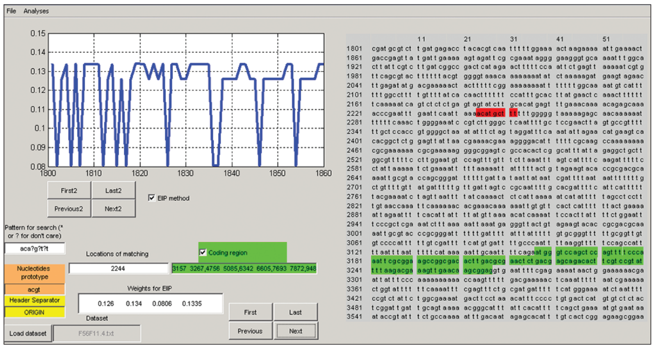
$$s_{high}[k] = \sum_n x[n].\,g[2k-n]$$

$$s_{low}[k] = \sum_n x[n].\,h[2k-n] \qquad (1)$$

Figure 3 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 or binary methods
- Search option for special patterns in the sequence



aca?g?t?t

2244

3157 3267,4756 5085,6342 6605,7693 7872,948

acgt

0.126 0.134 0.0806 0.1335

F56F11.4.txt

```
              11          21          31          41          51
1801  cgat gcgt ct  t gat gagacc  t acacgt caa  t t t t t ggaaa  act aagaaaa  at t gaaaact
1861  gaccgagt t a  t gat t gaaaa  agt agat t cg  cgaaat aggg  gaagggt gca  aaat t t ggca
1921  ct t at t cgt c  t t gat cggcc  gact cat aga  act t t t t cca  at t ct gagt t  aaaaat cgt g
1981  t t cagcgt ac  t t t t t t acgt  ggggt aaaca  aaaaaaat at  ct aaaaagat  gaagt agaac
2041  t t gagat at g  acgaaaaact  act t t t t cgg  aaaaaaaat t  t t t t t ggcaa  aat gt cat t t
2101  t t t ggcct t t  t gt t t t at ca  caact t t t t t  ccat t t gcac  t t at gaact c  aaact t t t t t
2161  t caaaaat ca  gt ct ct ct ga  gt agt at ct t  gcacat gagt  t t gaaacaaa  acagagcaaa
2221  acccgaat t t  gaat t caat t  aaaacat gct  t t t t t ggggg  t aaaaagagc  aacaaaaaat
2281  t t t t t caaac  t ggggaaat c  cgt ct t gggc  t caat t t t gc  t ccgaact t a  gt gccgt t t t
2341  t t gct ccacc  gt ggggct aa  at at t t ct ag  t aggat t t ca  aat at t agaa  cat gaagt ca
2401  cacggct ct g  gagt t at t aa  cgaaaacgaa  aggggacat t  t t t t cgcaag  ccaaaaaaaa
2461  cgcgaaaaaa  cgcgaaaaag  gggcggagt c  gccacact cg  gcat t t at t a  gaggct gct t
2521  ggcgt t t t t c  ct t ggaat gt  ccagt t t gt t  t ct t aaat t t  agt cat t t t c  aagat t t t t c
2581  ct at t aaaaa  t ct gaaaat t  t t t aaaaat t  at t t t t act g  t agaagt aca  ccgcgacgca
2641  aaat t gcgt a  ccagcgggat  t t t t t gat t a  t aat t at aat  cgat t t t aat  aat t t t t t t g
2701  ct gt t t t gt t  at gat t t t t g  t t gat t agt c  cgcaat t t t g  cacgat t t t c  at t cat t t t t
2761  t acgaaaat c  t agt t aat t t  t at caaaact  ct cat t aaaa  at cat at t t t  t agcct cct t
2821  t gt aaccaaa  t t caaaaaac  acgaaacaaa  aaat t gt gt t  cact cat t t t  t t act gaaaa
2881  at t agaat t t  t cacat t at t  t t at gt t aaa  acat caaaat  t ccact t at t  t t ct ggaat t
2941  t cccgct cga  aat ct t t aaa  aat t aaat ga  gggt t act gt  agaagt acac  cgcgacgcaa
3001  aat t gcgt ac  ccgcgggat t  t t t t gat t t t  at t t at t t t t  gt gcgggt t t  t t gcggt t gt
3061  gt ccccat t t  t t gt cgat t t  t cat gt gat t  t t gcccaat t  t t aaggt t t t  t ccagccat t
3121  t t aat t t aat  t t t t cat aaa  aat t gcaat t  t t cagaat gg  gt ccagct cc  agt t t t ccca
3181  aat t cgcgga  agccggcgac  act t gacgcg  aact ct gacg  agcagacact  t cgt ccgt at
3241  t t t aagacga  aagt t gaaca  agcggaggt g  aat t t t cgag  aaaat aact g  gaaaaaacga
3301  at t at t t ccc  aaaaaaaaaa  ccaagt t t t t  gacgaaaaac  t caaat t t t t  aat cgaaaaa
3361  ct ggt at t t t  t t caaaat t t  cgagt t t ct g  cgat gat t t t  t cat gggaaa  at ct aat t t t
3421  ccgt ct at t c  t ggcgaaaat  gact t cct aa  acat t t cccc  t gt gact cat  gt gt ct ct ac
3481  t cggat t gat  t gcagt aaaa  acgggcat t t  at t cacat t t  t cagct gaaa  at gt gaat aa
3541  at accgt at t  t ct gccaaaa  at t t gacaat  agagcacat t  t gt cact cgg  agaagcggaa
```

Figure 3: A view of the designed user-friendly package for analyzing DNA sequences

- Applying of DSP methods such as DFT on the signal
- Prediction of the protein coding regions.

Figure 4a and b show the result of applying DWT algorithm on the sequence F56F11.4. The power spectrum of the signal is smoothed by removing the high frequency components. Hence, the noise effect is decreased, which leads to improve the accuracy of identification task.

Choosing Blackman Window and Sliding it on the Estimated Sequence

In DNA sequence analysis, it is important to make the window size sufficiently large. In this paper, like many other researches such as Tiwari *et al*.,[8] and Akhtar *et al*.,[13] we have taken the window length equal to 351. Since there are very few intron-containing genes in these sequences, open reading frames (ORFs) of length less than 300 bp are not frequently encountered. A window length in the range of 250-400 gives the similar results. The windows of length less than 250 increase the noise level resulting in unacceptable statistics, while those greater than 400 tend to miss the ORFs due to numerous overlaps.[8]

In the proposed algorithm, we have employed a Blackman window to segment the filtered sequence. The Blackman window gives high weight to the codon positions residing in center of window and much less weight to the codons near the window boundaries. Hence, the noise cancelling level in Blackman window is higher than the other windows. The impulse response of the FIR windows is depicted in Figure 5. As can be seen, Blackman window has the highest amount of attenuation between the other windows. So, the background noise is more suppressed by Blackman window.

Goertzel Algorithm

The Goertzel algorithm is a digital signal processing technique that provides a means for efficient evaluation of individual terms of DFT, thus making it useful in

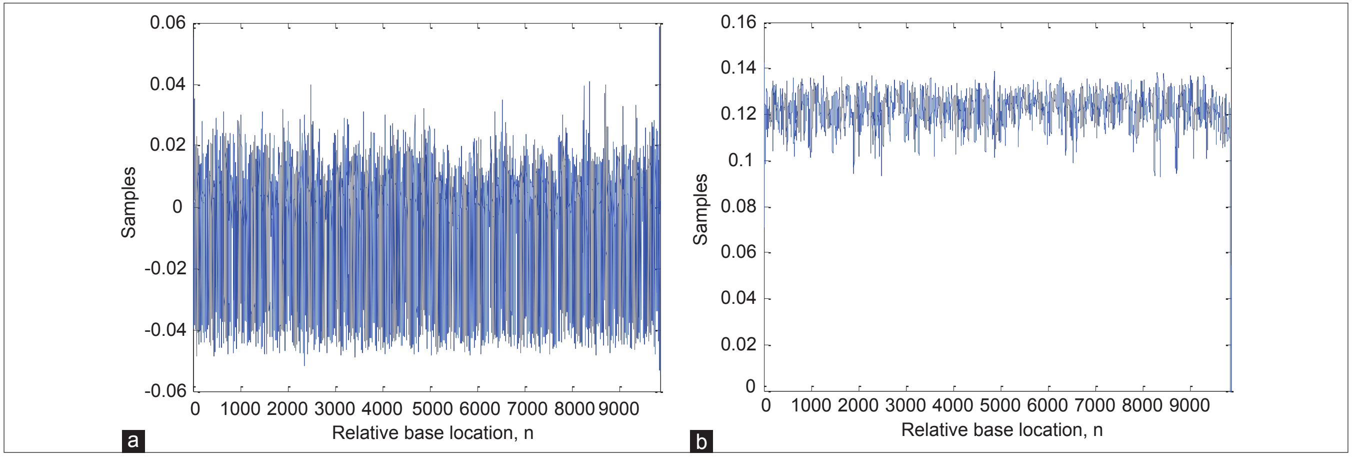


Figure 4: Applying DWT to the numerical sequence. (a) High frequency components of level 3 DWT decomposition (detail signal). (b) Low frequency components of level 3 DWT decomposition (approximation signal)

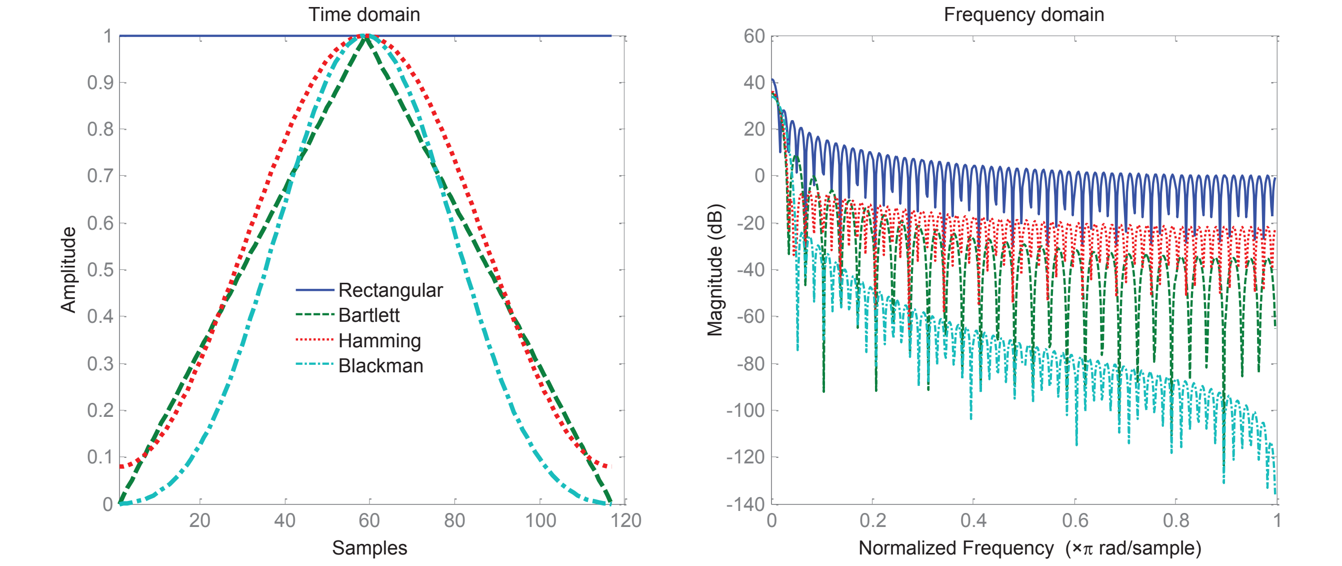


Figure 5: Comparison of the different FIR windows and their frequency impulse responses

certain practical application, such as dual-tone multi frequency (DTMF) signals,[25] digital multi frequency (MF) receiver,[26] and in a very small aperture terminal (VSAT) satellite communication system.[27]

The Goertzel filter is composed of a recursive part and a non-recursive part [Figure 6]. The DFT coefficients are obtained as the output of the system after N iterations which N is the input signal length. 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 this work to extract the period-3 components, exactly). 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 every sample update and the non-recursive part is computed only after the N^{th} time instant when the Fourier coefficients are to be determined.[28]

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 a sample. However, DFT and decimation in time FFT require N^2 and $N \log_2 N$ complex multiplications, respectively.[28]

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:

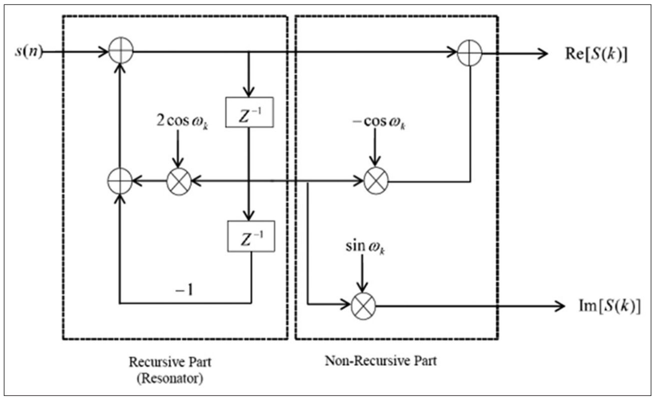


Figure 6: Filter realization of the Goertzel algorithm[28]

Sensitivity, Specificity, and Precision

These parameters are defined as follow according to[13] and:[29]

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

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

$$P = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

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.

Receiver Operating Characteristic Curves

The receiver operating characteristic (ROC) curves were developed in the 1950s as a tool for evaluating prediction techniques based on their performance.[30] 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:[31]

$$y = \alpha\left(1 - e^{\left[-\beta_1 \sqrt{x} + \beta_2 x\right]}\right) \tag{3}$$

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

$$E(p) = \sum_{i=1}^{n}\left[\alpha - \left(1 - e^{-\left[\beta_1 \sqrt{x_i} + \beta_2 x_i\right]}\right) - y_i\right]^2 \tag{4}$$

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

Area Under the ROC Curve

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.[29]

SIMULATION RESULTS

In order to demonstrate the performance of the methods, we apply them on four gene sequences; F56F11.4, AF009962, AF019074.1, and AJ223321 from GenBank database.[32] The gene sequence F56F11.4 (GenBank No. AF099922) is on chromosome III of *Caenorhabditiselegans*. *C elegans* is a free living nematode, about 1 mm 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.[32] AF009962 is the accession number for single exon, which has one coding region at position 3934-4581. The gene sequence AF019074.1 has the length of 6350, which has three distinct exons, 3101-3187, 3761-4574, and 5832-6007. AJ223321.1 is in the HMR195 dataset. This database 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 ends with a stop codon (TAA, TAG, or TGA). There is one coding region existed in AJ223321.1 gene sequence, which its location is 1196-2764. All mentioned sequences are converted to numerical sequences using EIIP method.

In this paper, to compare the performance of the proposed algorithm and other tested methods, we used the parameters S_n, $S_p$and *P*, which were described in section III. Amounts of these parameters achieved from equation (2). 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 7. In this Figure, the value of threshold is 0.161. It can be observed in Figure 7 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.

In this paper, to evaluate the performance of the proposed algorithm, DFT[8] and Multi-Stage filter (MS)[33] methods are implemented. Figures 8-11a 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, the 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 8-11c, the large amount of noise is removed in the proposed method due to applying

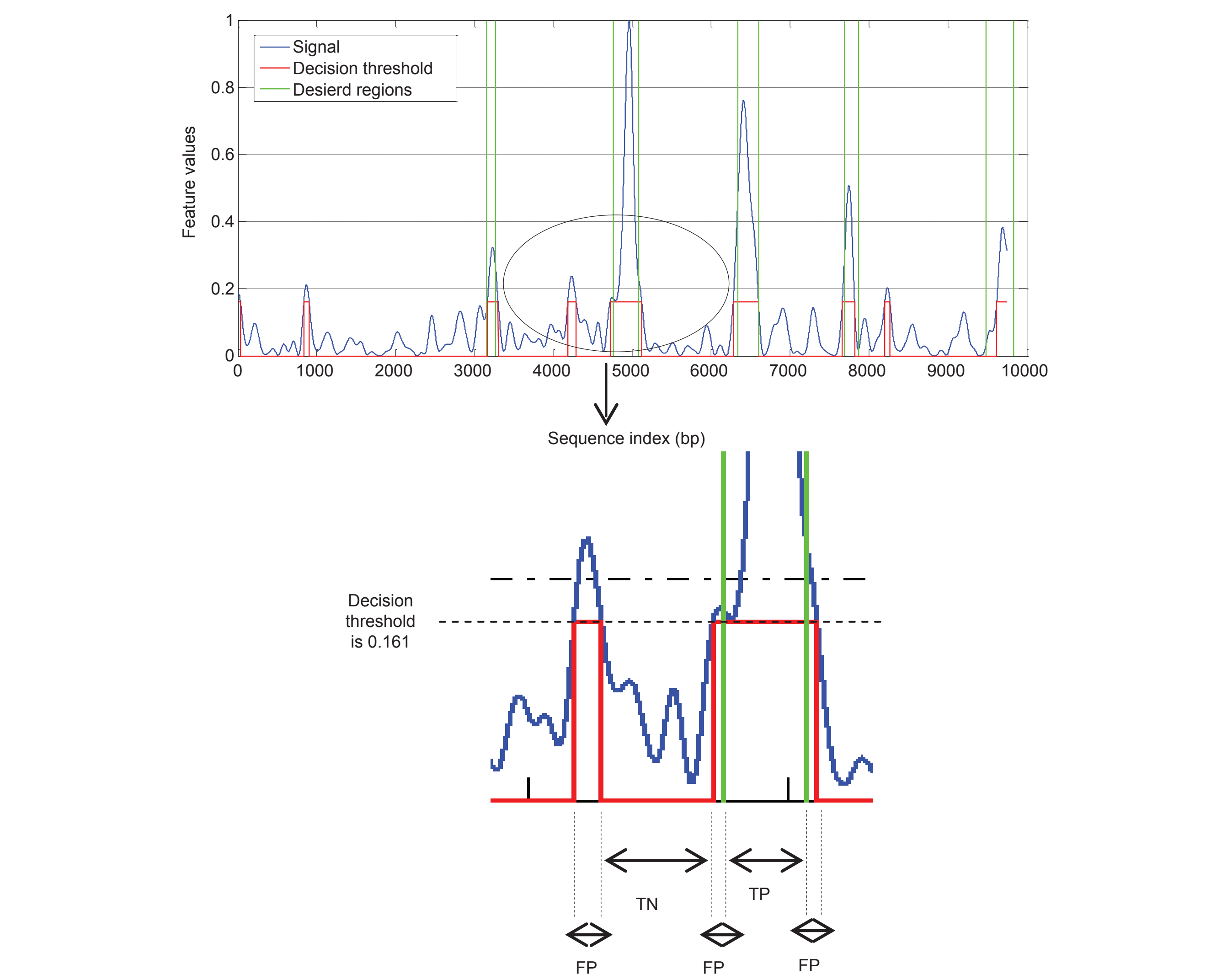


Figure 7: Parameters for exon-intron separation problem

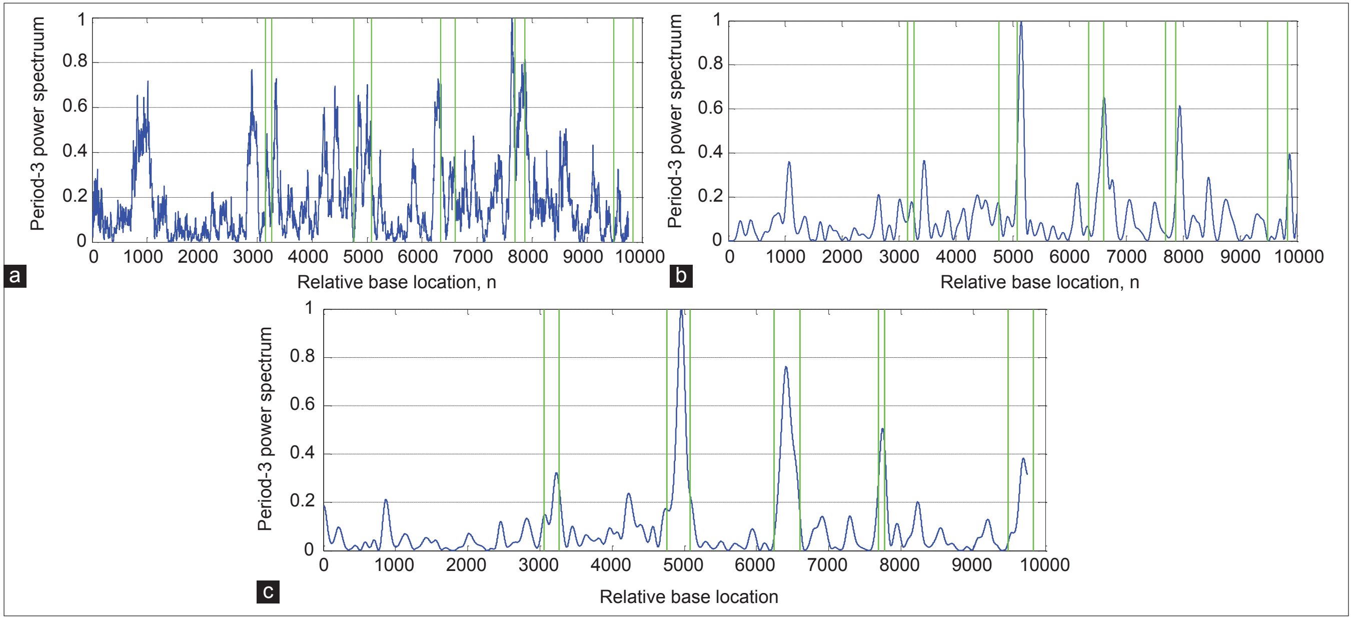


Figure 8: Results of the algorithms for identification of the exonic regions on the gene sequence F56F11.4: (a) DFT, (b) MS-filter, and (c) Proposed algorithm

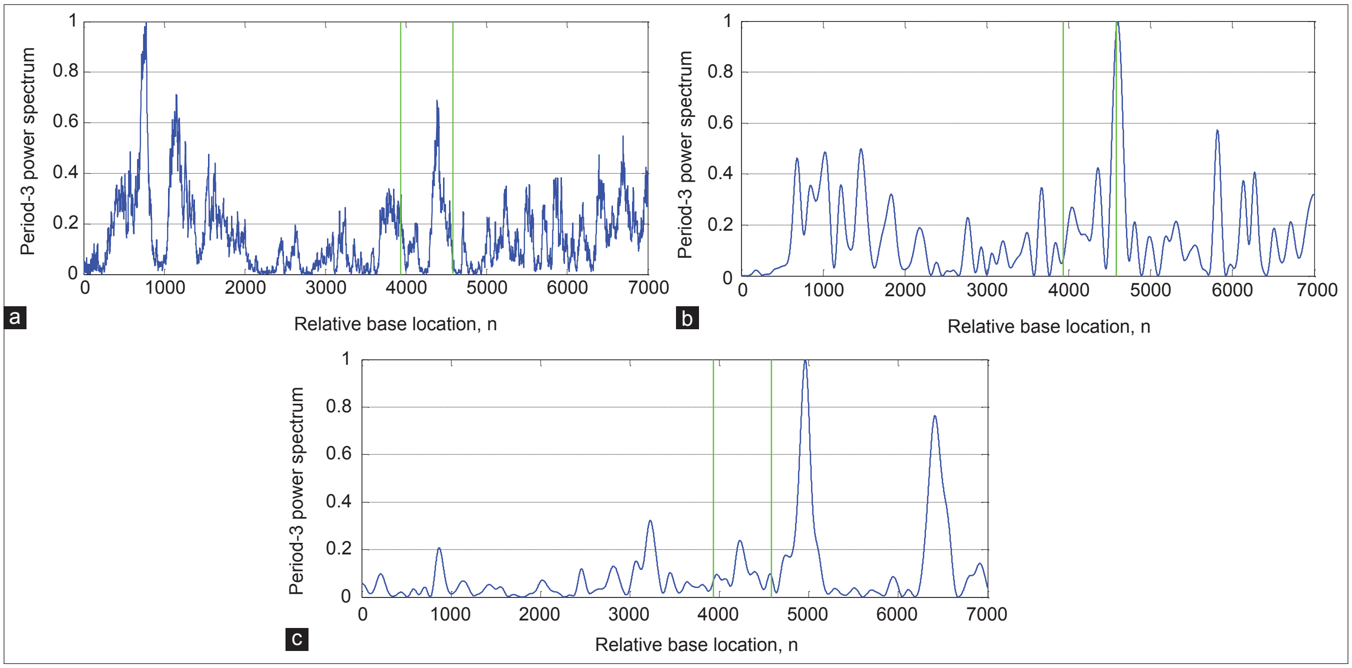


Figure 9: Results of the algorithms for identification of the exonic regions on the gene sequence AF009962: (a) DFT, (b) MS-filter, and (c) Proposed algorithm

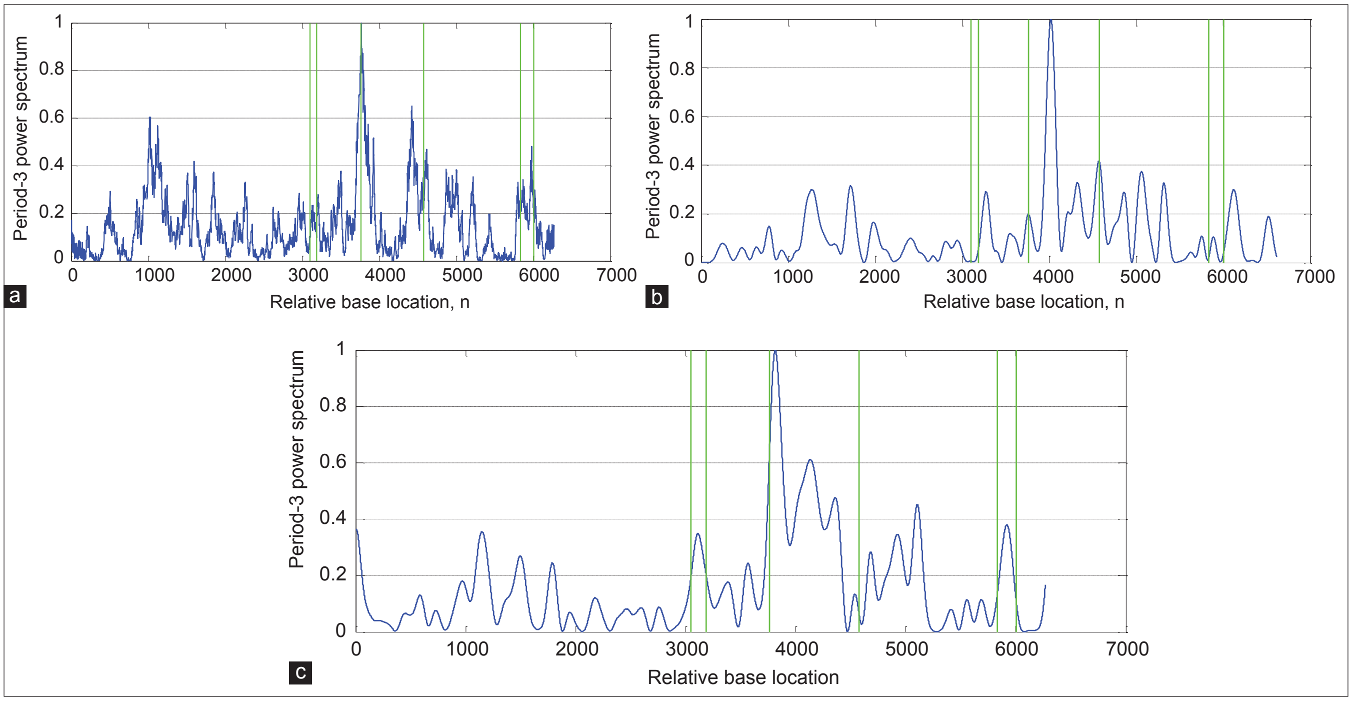


Figure 10: Results of the algorithms for identification of the exonic regions on the gene sequence AF019074.1: (a) DFT, (b) MS-filter, and (c) Proposed algorithm

the DWT, and small size of exons (For example, first exon in F56F11.4 gene sequence) can be identified because of using the Goertzel algorithm.

Table 1 shows the estimated exons by methods DFT, MS filter, and the proposed algorithm compared with the locations of exons in a sample gene sequence F56F11.4 from NCBI database. As can be seen, the proposed algorithm result is better than the other methods because of using the Goertzel algorithm. In Table 2, 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

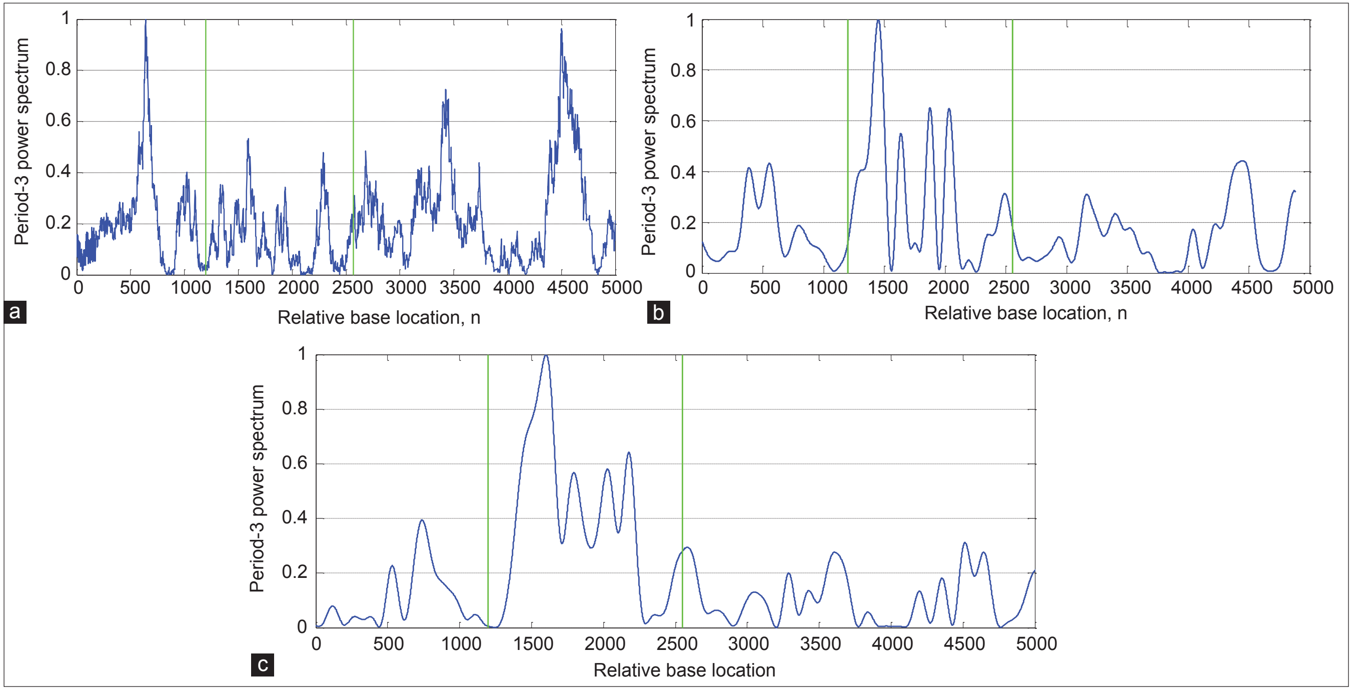


Figure 11: Results of the algorithms for identification of the exonic regions on the gene sequence AJ223321.1: (a) DFT, (b) MS-filter, and (c) Proposed algorithm

Table 1: Comparison of the proposed algorithm and the other methods in determining protein coding regions using F56F11.4 gene sequence

Proposed algorithm	MS-filter	DFT	Exon locations in NCBI	# Exons
3167-3262- (95)	3177-3386 (209)	3167-3410 (243)	3157-3267 (110)	1
4759-5139 (380)	4749-5208 (459)	4771-5226 (455)	4756-5085 (329)	2
6326-6620 (294)	6310-6685 (375)	6278-6667 (389)	6342-6605 (263)	3
7672-7878 (206)	7708-8010 (302)	7700-7897 (197)	7693-7872 (179)	4
9502-9827 (325)	9630-9958 (328)	9608-10028 (420)	9483-9833 (350)	5

DFT – Discrete fourier transform; MS – Multistage Filter; NCBI – National Center for Biotechnology Information

Table 2: Quantitative evaluation of the algorithms using Genbank datasets

Sequence	Methods	Sn								
		10 (%)			30 (%)			50 (%)		
		FP (#)	Sp (%)	*P* (%)	FP (#)	Sp (%)	P (%)	FP (#)	Sp (%)	P (%)
F56F11.4	Proposed	0	100	90	0	100	91	18	94	94
	MS-filter	222	28	87	620	28	84	1052	29	81
	DFT	180	33	88	711	27	83	1183	27	80
AF009962	Proposed	0	100	90	183	53	90	477	40	86
	MS-filter	239	21	88	1421	12	73	2467	11	60
	DFT	2791	11	55	1791	10	68	2791	10	55
AF019074.1	Proposed	0	100	82	14	95	86	79	90	88
	MS-filter	24	81	83	478	40	79	1036	34	73
	DFT	83	57	82	479	40	79	1177	31	71
AJ223321.1	Proposed	0	100	71	0	100	75	84	90	81
	MS-filter	2128	27	41	1660	22	44	2128	26	41
	DFT	757	17	56	1468	24	48	2173	26	40

DFT – Discrete fourier transform; MS – Multistage Filter; FP – False positive; Sp – Specificity

number of false positives in the proposed method is 18 bp, while this quantity for MS filter and DFT are 1052 and 1183, respectively. Also, the proposed algorithm shows relative improvement of 11.1% and 12.5% 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 2.

To compare the computational efficiencies of the proposed algorithm and other tested methods, the average CPU time is computed over 1000 runs of the techniques 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 3 summarizes results of the average CPU times. It is observed that the proposed algorithm has improved the average CPU time by the factors of 65.9, 28.7, 23.5, and 16.2 relative to the next-best performing method, DFT in F56F11.4, AF009962, AF019074.1, and AJ223321.1 gene sequences, respectively.

Table 3: Average computational time of the algorithms

Gene identifier	Sequence length (bp)	Average computational time (second)		
		Proposed algorithm	Multi-stage filter	DFT
F56F11.4	9833	10.9	714.9	718.4
AF009962	7422	13.6	712.2	391.0
AF019074.1	6350	12.0	710.1	282.0
AJ223321.1	5321	11.9	710.5	193.3

DFT – Discrete fourier transform

Finally, Figures 12-15 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 factors of 1.36, 1.84, 1.38, and 1.83 over the DFT and 2, 1.82, 1.56, and 1.25 over the MS filter methods in F56F11.4, AF009962, AF019074.1, and AJ223321.1 gene sequences, respectively. This implies that the proposed algorithm is superior to the other methods for identifying exonic gene regions.

CONCLUSION

Gene identification is a complicated problem, and the detection of the period-3 patterns is a first step towards

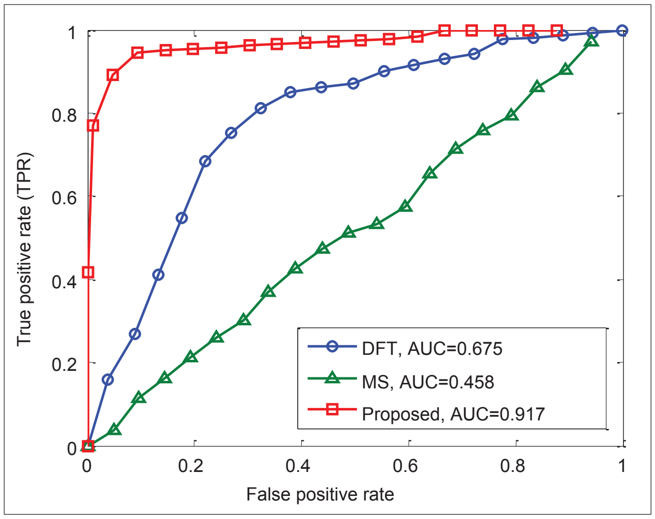


Figure 12: ROC curves of the methods for gene sequence F56F11.4

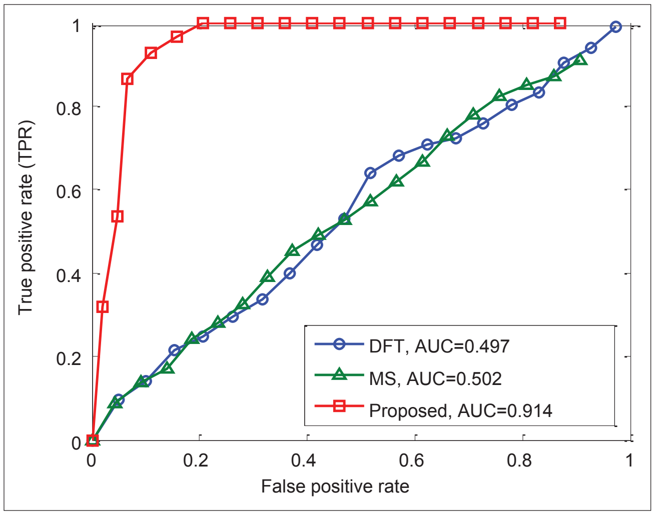


Figure 13: ROC curves of the methods for gene sequence AF009962.

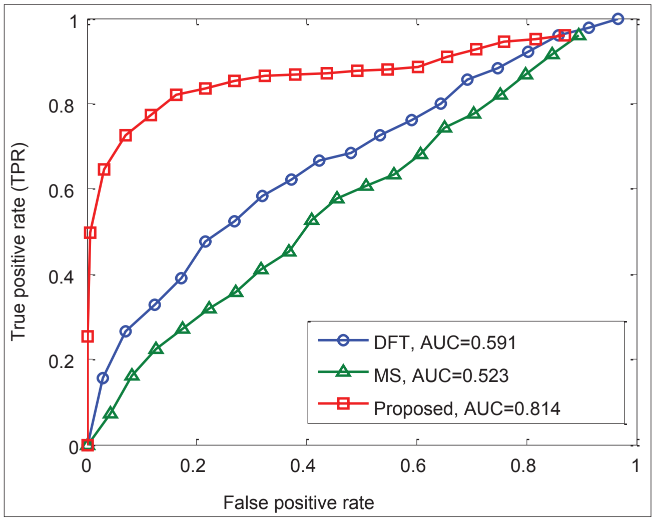


Figure 14: ROC curves of the methods for gene sequence AF019074.1.

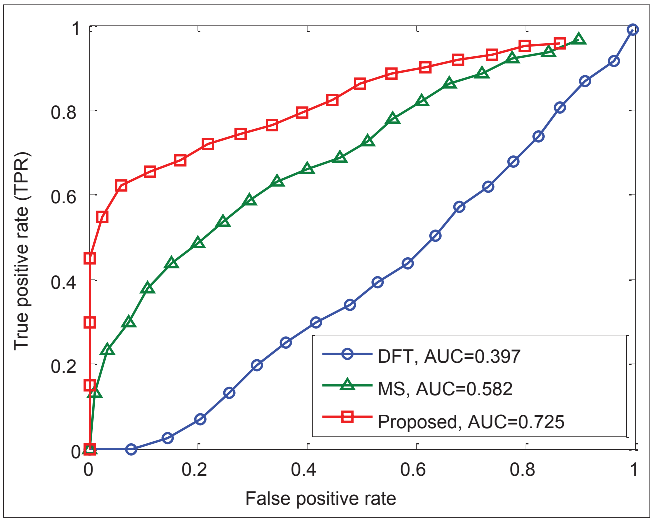


Figure 15: ROC curves of the methods for gene sequence AJ223321.1

gene and exon prediction. Due to the complex nature of the gene identification problem, we usually need a powerful model that can effectively represent the characteristics of protein-coding regions. 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, EIIP method is used to convert the symbolic sequence into digital signal. Then, we applied discrete wavelet transform to reduce the correlation between the numerical data and, therefore, reduce the high frequency noise. Finally, the Goertzel algorithm was applied to the filtered 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 methods. The main advantage of the proposed algorithm is its high speed characteristic, which leads to less run process.

REFERENCES

1. Snustad DP, Simmons MJ. Principles of Genetics. United States: John Wiley and Sons Inc; 2000.
2. Dougherty ER, Shnulevich L, Chen J, Wang ZJ. Genomic signal processing and statistics. EURASIP Book Series on Signal Processing and Communications. 2005; 2.
3. Fickett JW, Tung CS. Assessment of protein coding measures. Nucleic Acids Res 1992;20:6441-50.
4. Fickett JW. The gene identification problem: An overview for developers. Comput Chem 1996;20:103-18.
5. Vaidyanathan PP, Yoon BJ. The role of signal-processing concepts in genomics and proteomics. J Franklin Inst 2004;341:111-35.
6. Voss RF. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 1992;68:3805-8.
7. Chatzidimitriou-Dreismann CA, Larhammar D. Long-range correlations in DNA. Nature 1993;361:213-3.
8. 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.
9. Saberkari H, Shamsi M, Sedaaghi MH, 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: 2012. p. 354-9.
10. Saberkari H, Shamsi M, Sedaaghi MH. 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: 2013. p. 27-8.
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). Yantai, Vol 9. 2010; p. 4113-7.
12. 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.
13. Akhtar M, Epps J, Ambikairajah E. Signal processing in sequence analysis: Advanced in eukaryotic gene prediction. IEEE J Sel Top Signal Process 2008;2:310-21.
14. Haykin S. Adaptive Filter Theory. 4th ed. United States: Prentice Hall Inc; 2001.
15. Baoshan Ma, Zhu Yi-Sheng. Kalman filtering approach for human gene identification. 2nd ed. International Conference on Signal Processing Systems (ICSPS 2010). Vol 1. Dalian, 2010; P. V1-525-V1-528.
16. Baoshan Ma, Dongdong Qu. A novel adaptive filtering approach for genomic signal processing, IEEE 10th International Conference on Signal Processing (ICSP). Beijing, 2010. p. 1805-8.
17. Chakravarthy N, Spanias A, Lasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequence. EURASIP J Appl Signal Processing 2004;1:13-28.
18. Cristea PD. Conversion of nucleotides sequences into genomic signals. J Cell Mol Med 2002;6:279-303.
19. Cristea PD. Genetic signal representation and analysis, In SPIE Conference, International Biomedical Optics Symposium, Molecular Analysis and Informatics (BIOS '02), Vol. 4623 of Proceedings of SPIE. San Jose, Calif, USA. 2002. p. 77-84.
20. Claverie JM. Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet 1997;6:1735-44.
21. Doolittle WF. Phylogenetic classification and the universal tree. Science 1999;284:2124-8.
22. Cristea PD. Genomic signals of chromosomes and of concatenated reoriented coding regions, In SPIE Conference, Biomedical Optics (BIOS '04), Vol. 5. 5322 of Proceedings of SPIE. San Jose, Calif, USA: Progress in Biomedical Optics and Imaging; 2004. p. 29-41.
23. Zhang R, Zhang CT. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. J Bioml Struc Dyn 1994;11:767-82.
24. Rao KD, Swamy MN. Analysis of genomics and proteomics using DSP techniques. IEEE Trans Circuits Syst 2008;55:370-8.
25. Rabiner LR, Schafer RW. Digital Processing of Speech Signals. United States: Prentice-Hall Inc; 1987.
26. Braun FQ. Nonrecursive digital filters for detecting multi frequency code signals. IEEE Trans Acoust 1975;23:250-6.
27. Koval I, Gara G. Digital MF receiver using discrete Fourier transform. IEEE Tran Commun 1973;12:1331-5.
28. Oppenheim AV, Schafer RW. Discrete Time Signal Processing. United States: Prentice Hall Inc; 1999.
29. Burset M, Guigo R. Evaluation of gene structure prediction programs. Genomics 1996;34:353-67.
30. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers HP Laboratories. 2003; 4.
31. 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). Paris, 2010; p. 2231-4.
32. 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 2013 Apr 19].
33. Vaidayanathan PP, Yoon BJ. Digital filters for gene prediction applications, Proceeding of the 36th Asilomar Conference on Signals, Systems, and Computers, 2002.

How to cite this article: Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A fast algorithm for exonic regions prediction in DNA sequences. J Med Sign Sens 2012;3:139-49.

Source of Support: Nil, **Conflict of Interest:** None declared

BIOGRAPHIES

Hamidreza Saberkari was born in Rasht, Iran. He received the B.Sc. degree in Electrical Engineering from Guilan University, Rasht, IRAN, in 2011. In 2013, he received his M.Sc. degree in Communication Engineering from Sahand University of Technology, Tabriz, IRAN. Now, he is Ph.D. student in Electrical Engineering at Sahand University of Technology, Tabriz, Iran. His research interests include Bio-MEMS, RF MEMS, RF IC design, genomic signal processing, Bioinformatics, signal processing, pattern recognition.

E-mail: h_saberkari@sut.ac.ir

Mousa Shamsi received his B.Sc. degree in Electrical Engineering (major: electronics) from Tabriz University, Tabriz, IRAN, in 1995. In 1996, he joined the University of Tehran, Tehran, IRAN. He received his M.Sc. degree in Electrical Engineering (major: Biomedical Engineering) from this university in 1999. From 1999 to 2002, he taught as a lecturer at Sahand University of Technology, Tabriz, Iran. From 2002 to 2008, he was a PhD student at the University of Tehran in Bioelectrical Engineering. In 2006, he was granted with the Iranian government scholarship as a visiting researcher at the Ryukyus University, Okinawa, Japan. From December 2006 to May 2008, he was a visiting researcher at this University. He received his PhD degree in Electrical Engineering (major: Biomedical Engineering) from University of Tehran in December 2008. From December 2008 to April 2013, he was an assistant professor at Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran. From April 2013, he is an associate professor at Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran. His research interests include medical image and signal processing, genomic signal processing, pattern recognition, adaptive networks, and facial surgical planning.

E-mail: shamsi@sut.ac.ir

Hamed Heravi was born in Mashad, Iran in 1989. He received the B.Sc. degree in Biomedical Engineering from Sahand University of Technology in 2012. Now, he is M.Sc. student in Communication Engineering at this University. His research interests are Biomedical Signal Processing, Biomedical Image Processing, Machine Vision And Biometrics.

E-mail: h_heravi@sut.ac.ir

Mohammad Hossein Sedaaghi was born in Tehran, Iran. He received the B.Sc. and M.Sc. degrees from the Sharif University of Technology, Tehran, IRAN, in 1986 and 1987, respectively. In 1998, he received the Ph.D. degree from Liverpool University. He is now a professor at Sahand University of Technology, Tabriz. His research interests include signal/image processing, pattern recognition, machine learning and biometrics.

E-mail: sedaaghi@sut.ac.ir