International Journal of Computer Vision and Signal Processing, 6(1), 1-10(2016) ORIGINAL ARTICLE

An Efficient Algorithm for Small Gene Prediction in DNA Sequences

Hamidreza Saberkari*

Department of Electrical Enginnering , Rasht Branch
Islamic Azad University, Rasht, Iran

Mahsa Saffari Farsani

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

Sahar Aminkar

Department of Agricultural Biotechnology and Plant Breeding
Tarbiat Modares University, Tehran, Iran

Mousa Shamsi

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

Abstract

The main purpose of this paper is to introduce a new method for gene prediction in DNA sequences based on the period-3 property in exons. First, the symbolic DNA sequences converted to digital signal by using maximum homogeny estimation modeling method. Then, to reduce the effect of background noise in the period-3 spectrum, we have used the discrete wavelet transform (DWT) at four levels and apply it on the input numerical strand. Finally, to extract the period-3 components in smoothed sequence, we have used the minimum variance spectrum estimating technique. Using the proposed algorithm leads to increase the speed of process and therefore to reduce the computational complexity. The ability of detect small size exons in DNA sequences is another advantage of our algorithm. Performance of the proposed algorithm in exon prediction is compared with several existing methods at the nucleotide level using: (i) specificity vs. sensitivity; (ii) receiver operating curves (ROC) curve; (iii) area under ROC curve. Simulation results show that our algorithm increase the accuracy of exon detection relative to the most common digital signal processing (DSP) tested methods for gene prediction.

Keywords: DNA sequence, coding region, exon, signal processing, intron, wavelet

IJCVSP

ISSN: 2186-1390
http://cennser.org/IJCVSP

Article History:
Received: 11 August 2015
Revised: 11 March 2016
Accepted: 11 April 2016
Published Online: 12 April 2016

1. 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 Figure1. Genes are responsible for protein synthesis; therefore, they are called protein-coding regions because they carry the necessary information for protein coding

*Corresponding author
Email addresses: h_saberkari@sut.ac.ir (Hamidreza Saberkari), m.saffari@stu.yazd.ac.ir (Mahsa Saffari Farsani), Sahar.Aminkar@modares.ac.ir (Sahar Aminkar), shamsi@sut.ac.ir (Mousa Shamsi)

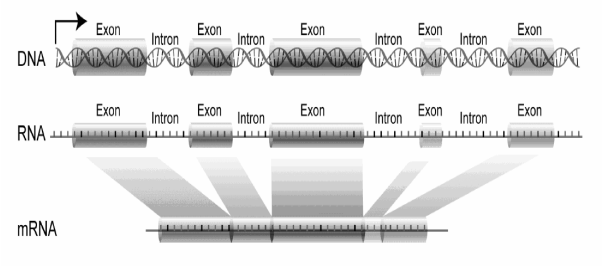


Figure 1: Exon/Intron regions for eukaryotic DNA.

[3, 4, 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)-based methods presented in literatures to identify the protein coding regions in DNA sequences which are based on Fourier spectral. In [8] Fourier transform is used for this purpose. In this way, by choosing a fixed-length window and sliding it on the numerical DNA sequences and then applying a discrete Fourier transform (DFT) and calculate the power of the resulted spectrum, we can determine the exonic regions. In our previous work [9] we used an anti-notch filter (AN filter) with the central frequency of 2/3 in order to capture the background noise. In this work, the DNA sequence is first passed through a notch filter and then a sliding windowed DFT is applied on the filtered sequence. In [10] we proposed a windowless technique based on the Z-curve to identify gene islands in total DNA sequence which called cumulative GC-Profile method. The main characteristic of our proposed method is that the resolution of it in displaying the genomic GC content is high since no sliding window is used, but the computational complexity of this method is high. In [11] an appropriate method is proposed to predict the protein regions by combining the DFT and continues wavelet transform (CWT). CWT leads to reduce the high frequency noise and therefore improve the accuracy of prediction. In [12] authors proposed a new algorithm based on Fourier Transform and using Bartlett window to suppress the non-exonic regions. Authors in [13] used Time Domain algorithms to determine the coding regions in DNA sequences. Adaptive filters [14] are one the best tools for predictions task. In [15] and [16] authors proposed two adaptive filtering approaches based on Kalman filter and least mean squares (LMS) algorithm. However, the major problem with LMS is that the convergence behavior of it is slow which leads to high computational complexity in it. In [17] a parametric method estimation of spectrum based on autoregressive model (ARM) is proposed. The ARM has the advantage over the DFT that they work with smaller window sizes and, thus, shorter sequences. In this paper, a new method based on discrete wavelet transform (DWT) and the minumum variance spectrum estimating technique is proposed to determine the location of exons based on their period-3 properties. Using the proposed algorithm leads to improve the prediction accuracy of coding regions, especially small size of exons in DNA sequences. The rest of the paper is organized as follows: In Section 2, DNA numerical representation and maximum homogeny estimation modeling method for mapping genomic sequence into digital values is discussed. In Section 3, DWT is introduced to reduce the noise in indicator sequence. The minumum variance spectrum estimating technique which is used to extract period-3 patterns is discussed in Section 4. The proposed algorithm is described in section 5 as a flow graph. Results and discussion is explained in Section 6 using Genbank database. Finally, conclusion is mentioned in Section 7.

2. NUMERICAL REPRESENTATION OF DNA SEQUENCE

Converting the DNA sequences into digital signals [18, 19] provides the possibility to apply signal processing tools in order to analysis 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 base pairs, including both exon and intron zones, at the scale of whole genomes or chromosomes [20, 21, 19]. In this paper, we have used the maximum homogeny estimation modeling method to convert the symbolic sequence of DNA to numerical signals. In this method, it is assumed that each symbol of DNA sequence is produced from a data source with a given probability density function, and the series $D = [D_1, D_2, ...D_N]$ is produced through drawing symbols from these data sources in a cycle. The number of sources is equal to the number of hidden alternation in the sequence. Accordingly, maximum homogeny estimation in P alternation is computed as:

$$MLE = \arg \max_{P \in B} \log P(W \mid M^P, P) \qquad (1)$$

where $B = [1, ...N_0]$ is searching area for the $P(N_0 < N)$ parameter and $W = [W_1, D_2, ...W_N]$ is a series of vectors to represent D. Also, M is a matrix which its elements indicate the probability density functions of data sources. So that M_{ij} represents the probability that the ith source causes the jth symbol of S which $S\epsilon(A, C, G, T)$ [22].

3. USING DWT TO REDUCE THE HIGH FREQUENCY NOISE

In this paper, we use discrete wavelet transform and applied it on the input numerical sequence to eliminate the high frequency noise and hence improve the accuracy of exonic region identification. In DWT, the signal is passed first through the high and low pass filters, then by downsampling the filtered signal, samples are divided into two

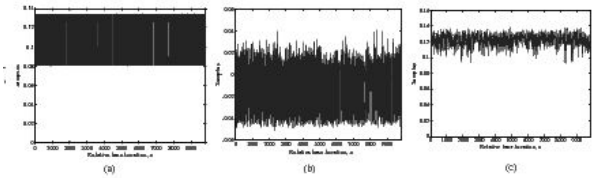


Figure 2: Applying DWT to the input numerical sequence. (a). Original signal (b). High frequency components of level 4 DWT decomposition (detail signal). (c). Low frequency components of level 3 DWT decomposition (approximation signal).

signals; high frequency samples (detail signals) and low frequency one (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:

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

$$s_{low}[k] = \sum_{n} x[n].h[2k-n] \tag{3}$$

Figures 2 (a) to (c) show the approximation and detail signal for the output spectrum of the gene sequence F56F11.4. By removing the detail signals and considering only the approximation signal, the extra frequencies are eliminated and the output power spectrum is smoothed. Hence, the noise effect is decreased which leads to improve the accuracy of identification task.

4. THE MINIMUM VARIANCE SPECTRUM ESTIMATION TECHNIQUE

All the algorithms proposed for protein coding regions detection are non-parametric techniques for power spectrum estimation of a random process. So their performance based on short-term Fourier transform is limited by the length of the input series. As a result, when the input series is containing remarkable amounts of energy within the frequency spectrum of lateral lobes, the leakage found in lateral lobes leads to distortion in the estimated spectrum. The leakage in the spectrum makes signals with lower energy in the input series fade away. Thus, designing an optimal filtering method is necessary, where it provides facility of adaptation with the input data. In this paper, minimum variance technique is used to estimate signal spectrum. This technique is the adapted sample of maximum homogeny method, which was first used by Capon to analyze power spectrum densities of multidimensional arrays [23]. Minimum variance method is created by minimizing output variance of a band-limited filter so that it is set on the spectral content of input process in the desired frequency (f_0). In order to get the impulse response of this filter, an inverse filter with $1+p$ coefficients of $a[0], a[1], ..., a[p]$ is considered. According to the convolution relationship, filter output ($y[n]$) towards input series ($x[n]$) is computed as Eq. (4):

$$y[n] = \sum_{k=0}^{P} a[k]x[n-k] = X^T[n]a \tag{4}$$

where ($x[n]$) and a are two vectors with $1+p$ dimensions and defined as follows:

$$X[n] = \begin{bmatrix} x[n] \\ x[n-1] \\ . \\ . \\ x[n-p] \end{bmatrix} a = \begin{bmatrix} a[0] \\ a[1] \\ . \\ . \\ a[p] \end{bmatrix} \tag{5}$$

Output variance of the filter equals to:

$$\rho = E[\mid y[n] \mid^2] = E[a^H X^*[n] X^T[n] a] \tag{6}$$
$$= a^H E[X^*[n] X^T[n]] a = a^H R_P a$$

where R_p is estimated autocorrelation matrix with dimensions of $(p+1) \times (p+1)$ and defined as follows:

$$R_p = \begin{bmatrix} r_{xx}[0] & . & . & . & r^*_{xx}[p] \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ r_{xx}[p] & . & . & . & r_{xx}[0] \end{bmatrix} \tag{7}$$

Output variance achieved in Eq. (5) is similar to the variance of linear predictor filter, with the difference that the coefficient[0] is selected arbitrarily to unitilize gain. The coefficients of minimum variance filter must be selected in such a way that in the desired frequency (f_0), the frequency response of the filter becomes unit, which means:

$$\sum_{k=0}^{p} a[k] exp(-j2\pi f_0 kT) = e^H(f_0)a = 1 \tag{8}$$

which $e(f_0)$ is a vector with dimensions of $1+p$ and is defined as follows:

$$e(f) = \begin{bmatrix} 1 \\ exp(j2\pi f) \\ . \\ . \\ exp(j2\pi f_0 T) \end{bmatrix} \tag{9}$$

Therefore, impulse response of the minimum variance filter for an input series is achieved as Eq. (9):

$$a_{MV} = \frac{R_p^{-1} e(f_0)}{e^H(f_0) R_p^{-1} e(f_0)} \tag{10}$$

Figure 3 shows the impulse response of minimum variance spectrum estimation filter, selecting window with fixed length of 351. As mentioned before, this filter is an adaptive filter such that its impulse response varies with the input samples. As it is seen, frequency conjugate components in $4\pi/3$ are eliminated.

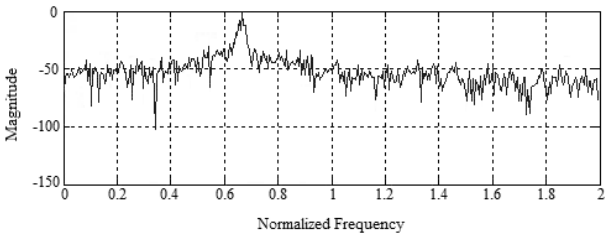


Figure 3: The impulse response of minimum variance spectrum estimation filter.

5. PROPOSED ALGORITHM

Details of the proposed method are shown in Figure 4 as a flow graph.

- Numerical mapping of DNA sequence using maximum homogeny estimation modeling method,
- Using DWT and applied it on the numerical sequence to remove the high frequency noise.
- Choosing Bartlett window and sliding it on the filtered sequence, and
- Using minimum variance spectrum estimating technique.

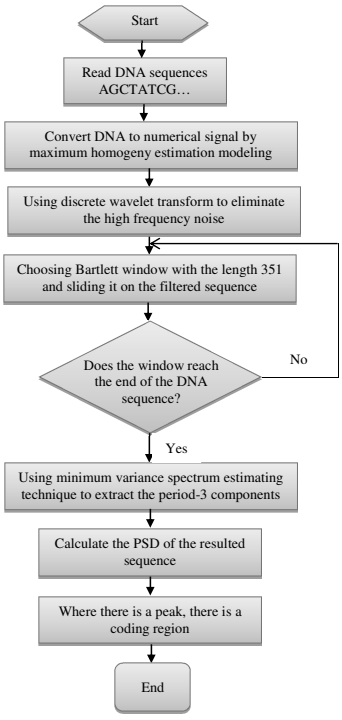


Figure 4: Flow graph of the proposed algorithm.

6. RESULTS AND DISCUSSION

To accurate comparison the different methods in identify the protein coding regions; the evaluation is done at

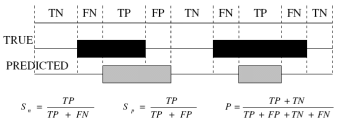


Figure 5: Nucleotide level measures of prediction accuracy.

nucleotide level. To determine the genomic regions by signal processing methods, some parameters are defined by changing the threshold level at system output. In this section, we introduce these parameters that listed as below: Sensitivity and Specificity: These parameters can be defined with the help of Figure 5, 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), specificity (S_p) and precision (P) are defines as [24]:

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

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

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

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 [25]. An ROC curve explores the effects on TP and FP as the position of an arbitrary decision threshold is varied. The ROC curve can approximated using an exponential model as below [26]:

$$y = a(1 - e^{[-\beta_1\sqrt{x} + \beta_2 x]}) \tag{14}$$

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

$$E(P) = \sum_{i=1}^{n} [a - (1 - e^{-[\beta_1\sqrt{x_i} + \beta_2 x_i]}) - y_i)]^2 \tag{15}$$

where $p = [\alpha \ \beta_1 \ \beta_2]^T$ and x_i, y_j are points in the ROC plane. The area under the ROC curve (AUC): This parameter is also a good indicator of the overall performance of an exon-location technique. The greater of AUC leads to the better performance of the tested algorithm [24].

In order to demonstrate the performance of the discussed methods, we have used the DNA sequence of gene F56F11.4a (GenBank No. AF099922) 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 [1]. We have used three other gene sequences named AF009962, AF019074.1 and AJ223321 for further assessments. 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 has only one coding region which its location is 1196-2764. We also have utilized two other datasets which describe below: Dataset HMR195: This dataset contains 195 genes of the human, mouse and rat, was established by Rogic and his colleagues in 2001 with purpose of evaluating different gene finding programs in DNA sequences. It includes 43 single-exon and 152 multi-exon genes and its coding area density is 14%. The maximum length of the sequences in the database is 1,383,720 bp and the number of sequences related to humans, mice and rats are 103, 82 and 10, respectively [24]. Dataset BG570: This dataset contains 570 multi-exon vertebrate gene sequences and is established by Burset and Guigo in 1996 to evaluate different programs designed for prediction of protein coding regions in genomic sequences. Each sequence in this database includes at least one intron and two exons. The total number of base pairs in this database is 2,892,149 bp containing 2,649 exons with total lengths of 444,498 bp. In addition its coding area density is 15.37% [26]. In this paper, to evaluate performance of the proposed algorithm, DFT method [8] and also AN-filter [9] are implemented. In addition to these methods, results of two Asif [12] and AMDF [13] methods are given in tables to compare them with the proposed algorithm. Figures 6 (a), (b) and (c) show the results of DFT, AN-filter and the proposed algorithms in determination of protein encoding regions in the gene F56F11.4, respectively. In DFT method, because of the accompaniment of noise with the main signal, estimation of protein coding regions does not have high accuracy. AN-filter has higher accuracy when compared to DFT, and non-protein regions are almost removed in them. Also, this method decreases computational complexity in comparison to DFT. In the proposed algorithm shown in Fig. 6 (c) the noise is highly removed due to the use of DWT and gene regions with small sizes (for example the first Exon in the gene F56F11.4a) could be identified because of the use of minimum variance spectrum estimation filter. In Figures (7) to (12) results of applying DFT, AN-filter and our proposed algorithm in identifying the coding regions on the gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 are shown by choosing the worst and best window types (Rectangular and Bartlett windows). Note that the length of the windows is chosen 351 (bp) for all of these simulations. Superiority of the proposed algorithm in determining the coding regions is clearly visible for both situations. In Table 1, S_p and AC amounts are given for a fixed amount of S_n in the proposed algorithm and other algorithms in sequence of the gene F56F11.4a. As seen, the proposed algorithm has the maximum amount of these two parameters; So that S_p and AC amounts are 0.95 and 0.68, respectively. Results of applying of the proposed algorithm and other methods on a set of genes from BG570 database is shown in Table 2. In order to apply the proposed algorithm to the genes in this database, exons and introns with length of 100bp or longer are extracted which includes 1768 exons and 1844 introns. As can be seen the proposed algorithm has the least amount of FP. In case of S_n equal to 0.40, the number of incorrect nucleotides at the proposed algorithm improves by the factor of 14.26 in comparison to the best method, AN- filter. A similar superiority of the proposed algorithm is shown in Table 3 which relates to the HMR195 database. The value of AC of the proposed algorithm for = 0.40 equals to 0.752 while its value for the AN-filter is 0.324. To compare the computational efficiencies of our proposed algorithm and other methods, the average CPU times over 1000 runs of the techniques for the four gene sequences, F56F11.4, AF009962, AF019074.1 and AJ223321.1, was computed. 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 4 summarizes results of the average CPU times. We observe that our algorithm has improved the average CPU times by the factor of 30.23, 37.77, 46.32 and 55.56 relative to the next-best performing method, AN-filter in F56F11.4, AF009962, AF019074.1 and AJ223321.1 gene sequences, respectively.

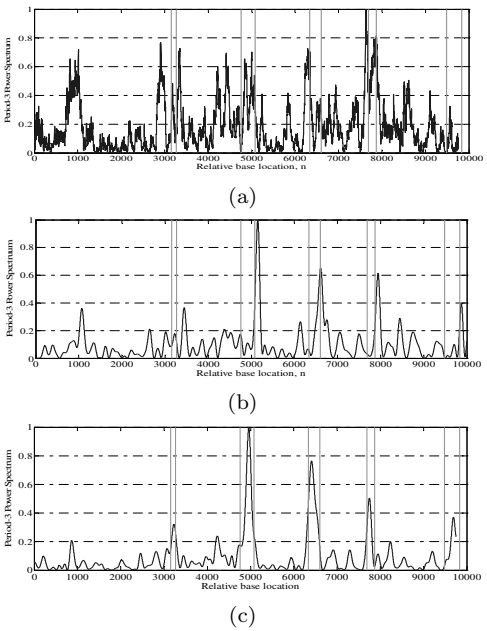
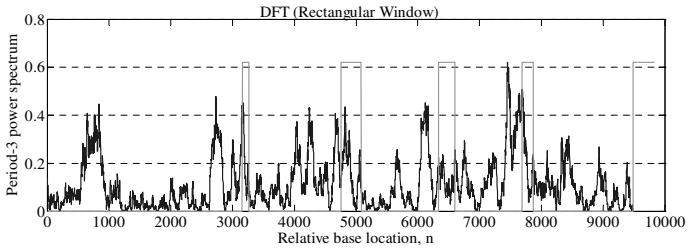
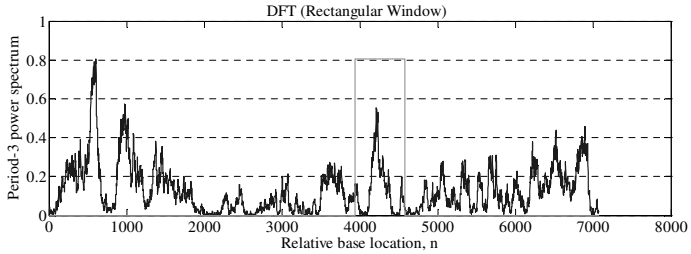


Figure 6: Exonic regions predicted on the gene sequence F56F11.4a by applying: (a). DFT, (b). AN-Filter and (c). Proposed Algorithm.

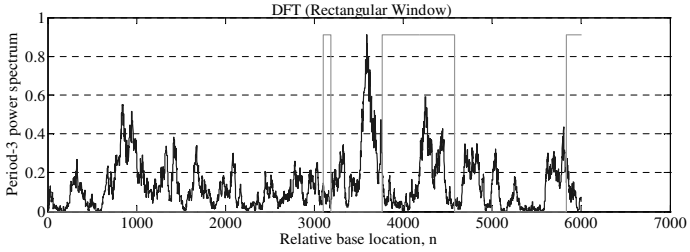
[1] http://www.ncbi.nlm.nih.gov/Genebank/index.html



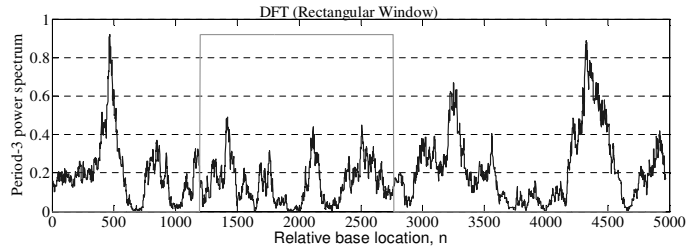
(a) F56F11.4a



(b) AF009962

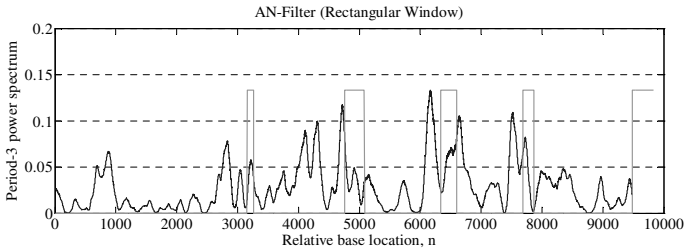


(c) AF019074.1

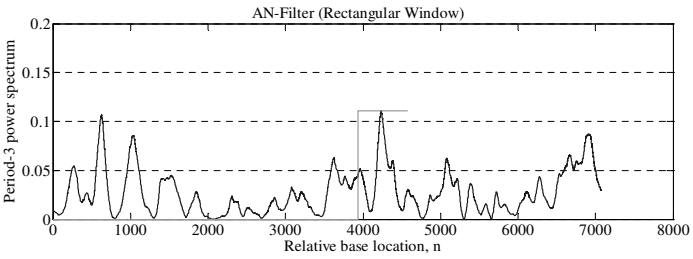


(d) AJ223321

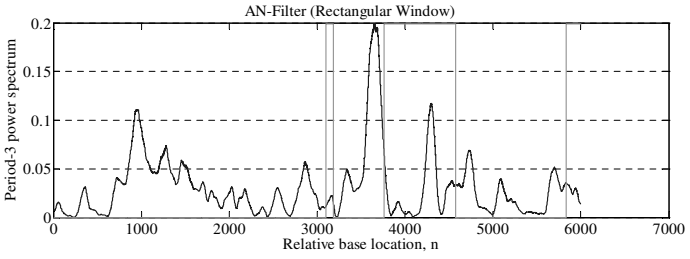
Figure 7: Identifying the coding regions in gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 by applying DFT method with the rectangular window size of 351 (bp).



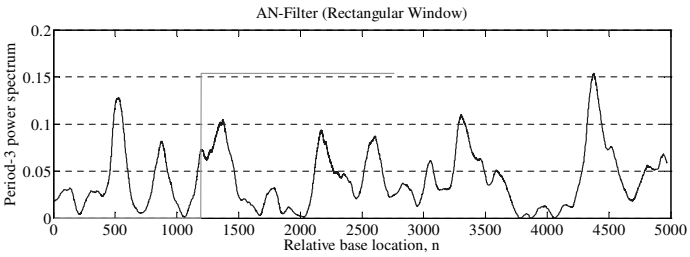
(a) F56F11.4a



(b) AF009962

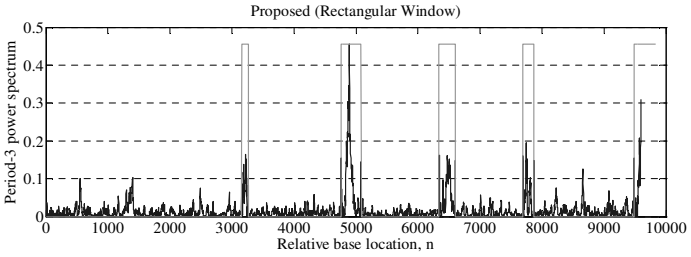


(c) AF019074.1

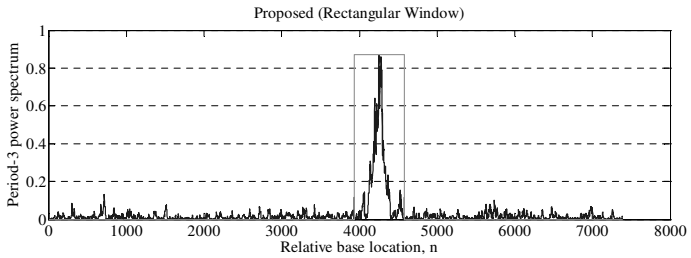


(d) AJ223321

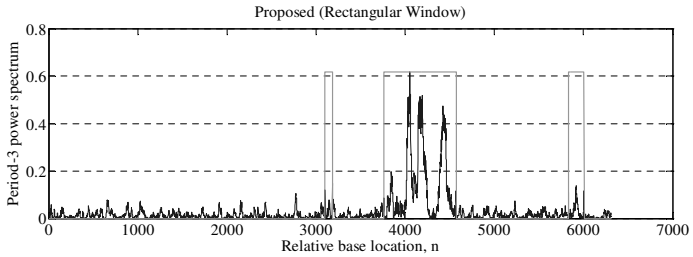
Figure 8: Identifying the coding regions in gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 by applying AN-Filter method with the rectangular window size of 351 (bp).



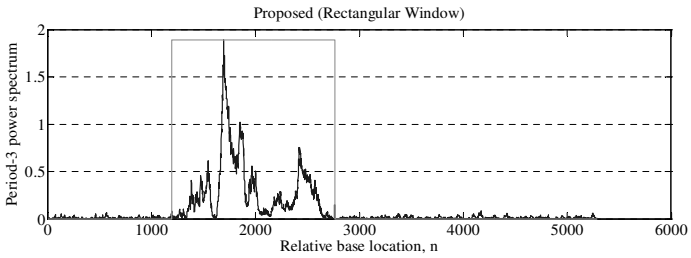
(a) F56F11.4a



(b) AF009962

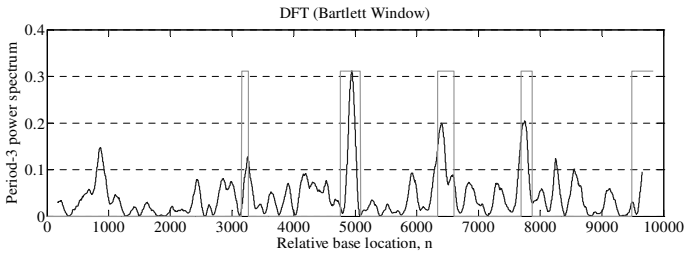


(c) AF019074.1

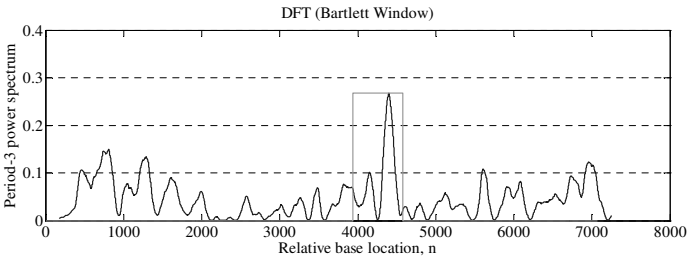


(d) AJ223321

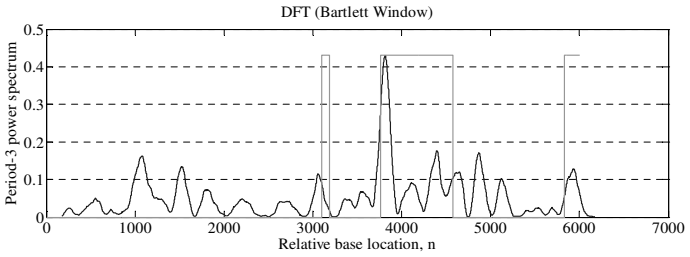
Figure 9: Identifying the coding regions in gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 by applying proposed method with the rectangular window size of 351 (bp).



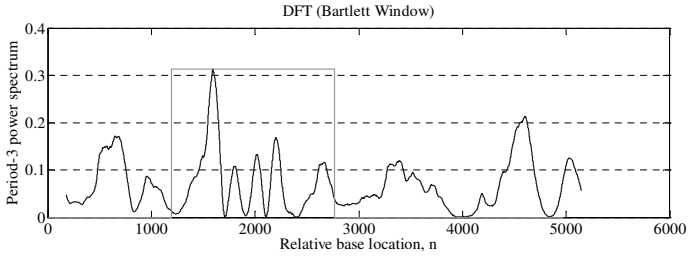
(a) F56F11.4a



(b) AF009962

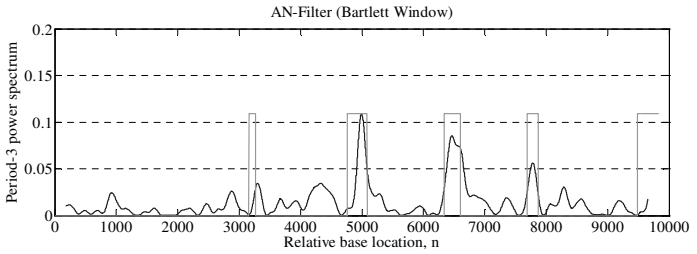


(c) AF019074.1

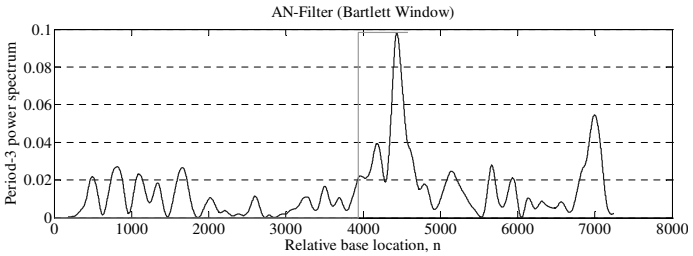


(d) AJ223321

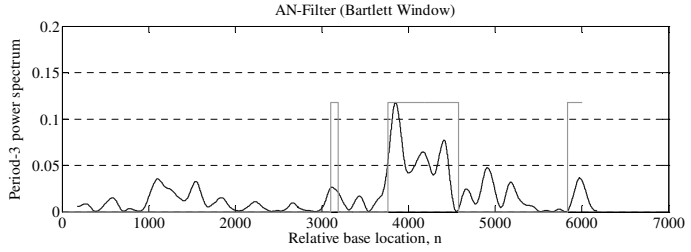
Figure 10: Identifying the coding regions in gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 by applying DFT method with the bartlett window size of 351 (bp).



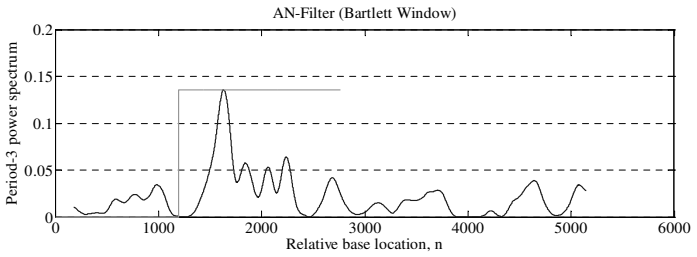
(a) F56F11.4a



(b) AF009962

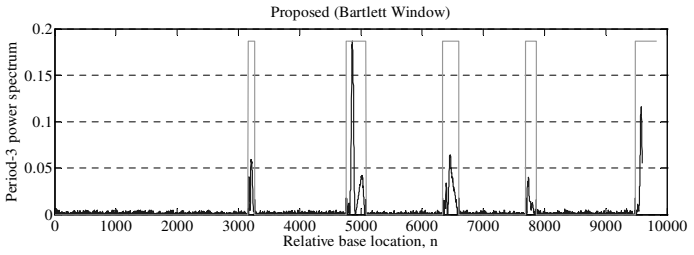


(c) AF019074.1

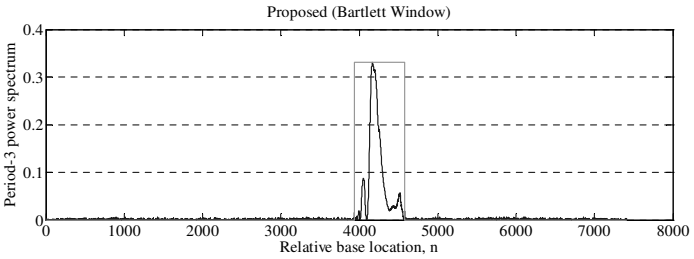


(d) AJ223321

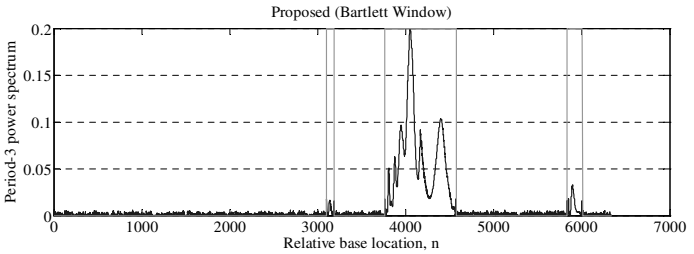
Figure 11: Identifying the coding regions in gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 by applying AN-filter method with the bartlett window size of 351 (bp).



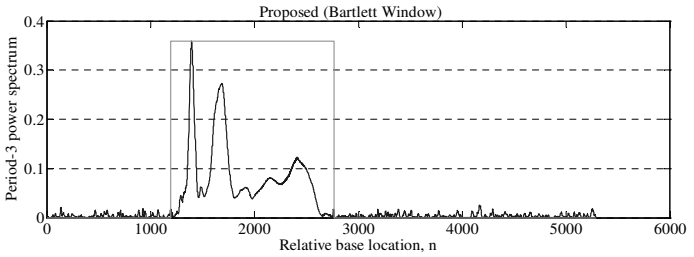
(a) F56F11.4a



(b) AF009962



(c) AF019074.1



(d) AJ223321

Figure 12: Identifying the coding regions in gene sequences F56F11.4a, AF009962, AF019074.1 and AJ223321 by applying proposed method with the bartlett window size of 351 (bp).

Table 1: Comparison of quantitative results of the proposed algorithm and with other applied methods which were mentioned, on sequence of the gene F56F11.4

Metod	Sp	AC
DFT	0.18	0.09
AN-filter	0.24	0.26
Asif	0.93	0.13
AMDF	0.21	0.20
TDP	0.50	0.54
Proposed	0.95	0.68

Table 2: Comparison of the quantitative results of the proposed algorithm with other methods applied on genes in BG570 database (with Sn =0.40).

Metod	AUC	FP	Sp	AC
DFT	0.6540	764	0.433	0.183
AN-filter	0.6765	499	0.497	0.174
Proposed	0.9365	35	0.925	0.874

Table 3: . Comparison of the quantitative results of the proposed algorithm with other methods applied on genes in HMR195 database (with Sn =0.40).

Metod	AUC	FP	Sp	AC
DFT	0.6782	1184	0.453	0.181
AN-filter	0.7615	562	0.574	0.324
Proposed	0.9656	96	0.895	0.752

Table 4: Average Computational Time computed for the different algorithms.

Gene Sequences	Sequence Length (bp)	Dissimilarity measures		
		Proposed	**AN-filter**	**DFT**
F56F11.4	9833	**22.8945**	629.231	718.4017
AF009962	7422	**17.2145**	650.321	391.0609
AF019074.1	6350	**12.5354**	580.5987	282.0491
AJ223321.1	5321	**9.3654**	520.3694	193.2907

7. CONCLUSIONS

Gene identification is a complicated problem, and the detection of the period-3 patterns is a first step towards gene and exon prediction. Due to the complex nature of the gene identification problem, we usually need a more powerful model that can effectively represent the characteristics of protein-coding genes. Many DSP techniques have been applied successfully for the identification task but still improvement in this direction is needed. In this paper, a fast model-independent algorithm is presented to exon detection in DNA sequences. Firstly, we used maximum homogeny estimation modeling method to convert the symbolic sequence into digital signal. Then, we used discrete wavelet transform to reduce the correlation between the numerical data and therefore reduce the high frequency noise. Finally, the minimum variance spectrum estimating technique was applied to the filtered sequence for the period-3 detection. Our proposed algorithm minimizes the number of nucleotides incorrectly predicted as coding regions which leads to increase the. Also area under the ROC curve is improved in our algorithm over the other tested methods. High speed characteristic in our algorithm is the major advantage which leads to increase the run process in it. Combination of advanced DSP techniques with the proposed algorithm can be used to identify the short exon regions in DNA sequences with low complexity and more efficiency which this issue is one of our goals in future works.

References

[1] D.P.Snustad, M.J.Simmons, Principles of genetics, John Wiley & Sons Inc.

[2] E.R.Dougherty, et al, Genomic signal processing and statistics.

[3] J.W.Fickett, C.S.Tung, Assessment of protein coding measures, Nucleic Acids Res (1992) 6441–6450.

[4] J.W.Fickett, The gene identification problem: an overview for developers, Comput (1996) 103118.

[5] P.P.Vaidyanathan, B.J.Yoon, The role of signal-processing concepts in genomics and proteomics, J. Franklin Inst (2005) 111–135.

[6] R.F.Voss, Evolution of long-range fractal correlations and 1/f noise in dna base sequences, Phy. Rev, Lett 85 (1992) 1342–1345.

[7] C.A.Chatzidimitriou-Dreismann, D.Larhammar, Long-range correlations in dna, Nature 361 (1993) 212–213.

[8] S.Tiwari, S.Ramachandran, A.Bhattacharya, S.Bhattacharya, R.Ramaswamy, Prediction of probable genes by fourier analysis of genomic sequences, Comput Appl Biosci 13 (1997) 263–270.

[9] H.Saberkari, M.Shamsi, M.H.Sedaaghi, Prediction of protein coding regions in dna sequences using signal processing methods, 2012 IEEE Symposium on Industrial Electronics and Applications (ISIEA 2012) (2012) 354–359.

[10] H.Saberkari, et al, 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) (2012) 27–28.

[11] S.Deng, et al, Prediction of protein coding regions by combining fourier and wavelet transform, Intarnational Conference on Image and Signal processing (ICISP).

[12] S.Datta, A.Asif, 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.

[13] M.Akhtar, J.Epps, E.Ambikairajah, Signal processing in sequence analysis: advanced in eukaryotic gene prediction, IEEE journal of selected topics in signal processing 2 (2008) 310–321.

[14] S.Haykin, Adaptive filter theory, in: Prentice Hall, 2001.

[15] M.Baoshan, Zh.Yi-Sheng, Kalman filtering approach for human gene identification, International Conference on Signal Processing Systems (ICSPS 2010).

[16] M.Baoshan, A novel adaptive filtering approach for genomic signal processing, IEEE 10th International Conference on Signal Processing (ICSP) (2010) 1805–1808.

[17] N.Chakravarthy, et al, Autoregressive modeling and feature analysis of dna sequence, EURASIP J. Appl. Sign. Proc (2004) 13–28.

[18] P.D.Cristea, Conversion of nucleotides sequences into genomic signals, J. Cell. Mol. Med 6 (2) (2002) 279–303.

[19] P.D.Cristea, Genetic signal representation and analysis, In SPIE Conference, InternationalBiomedical Optics Symposium, Molecular Analysis and Informatics (BIOS ?02) 4623 of Proceedings of SPIE (2002) 77–84.

[20] J.M.Claverie, Computational methods for the identification of genes in vertebrate genomic sequences, Hum. Mol.Genet 6 (10) (1997) 1735–1744.

[21] W.F.Doolittle, Phylogenetic classification and the universal tree, Science 284 (5423) (1999) 2124?2128.

[22] H. Herzel, E. N. Trifonov, O. Weiss, I. Gro?e., Interpreting correlations in biosequences physica a 249 (1998) 449–459.

[23] L. R. Rabiner, R. W. Schafer, Digital processing of speech signals, Prentice-Hall.

[24] M. Burset, R. Guigo, Evaluation of gene structure prediction programs, Genomics 34 (3) (1996) 353–367.

[25] T. Fawcett., Roc graphs: Notes and practical considerations for researchers hp laboratories.

[26] A. A. P. Ramachandran, W. S. Lu, 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).