Applied Medical Informatics
Original Research Vol. 34, No. 2 /2014, pp: 13-22

A Hybrid Anti-notch/Goertzel Model for Gene Prediction in DNA Sequences

Hamidreza SABERKARI*, Mousa SHAMSI, Mohammad Hossein SEDAAGHI

Genomic Signal Processing Laboratory, Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran.
E-mail(s): {h_saberkari, shamsi, sedaaghi}@sut.ac.ir

* Corresponding Author, Tel. +98-911-2362692

Received: 15 June 2014 /Accepted: 23 June 2014/ Published online: 25 June 2014

Abstract
Accurate detection of the exon regions in DNA sequences using signal processing tools has become a challenge in bioinformatics. In this paper, a hybrid Anti-notch/Goertzel algorithm has been presented for gene selection in DNA sequences. The proposed algorithm has many advantages when compared to other conventional methods. Firstly, it leads to identify the coding protein regions more accurate due to using the Goertzel algorithm which is tuned at the desired frequency. Secondly, faster detection time is achieved. The proposed algorithm is applied on several genes, including genes available in databases BG570 and HMR195 and the results are compared to other methods based on the nucleotide level evaluation criteria. Implementation results showed excellent performance of the proposed algorithm in identifying protein coding regions, specifically in identification of small-scale gene areas.

Keywords: Protein Coding Regions; Period-3; Anti-notch filter; Goertzel Algorithm

Introduction

Deoxyribonucleic acid (DNA) (Figure 1) is of the most important chemical compounds in living cells, bacteria and some viruses [1]. It is composed of four types of different nucleotides which named 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 eukaryotes, DNA is divided into two regions: genes and inter-genic. Only the gene area carries the information for protein synthesis. Each gene in turn consists of exon and intron regions as shown in Figure 2. Therefore exons carry the necessary codes for protein synthesis and are called protein coding regions [3].

Protein coding regions have a period-3 property which has not been observed in other parts of the DNA molecule [4]. This phenomenon can be due to heterogeneous use of codons. This means that despite the fact that more than one codon may code a particular amino acid; all of them do not appear necessarily with equal probability in living organisms. For instance a *G* nucleotide takes specific positions in codons in the exon areas [5, 6]. Period-3 property can be used as a detector for defining gene areas.

Copyright ©2014 by the authors; Licensee SRIMA, Cluj-Napoca, Romania.

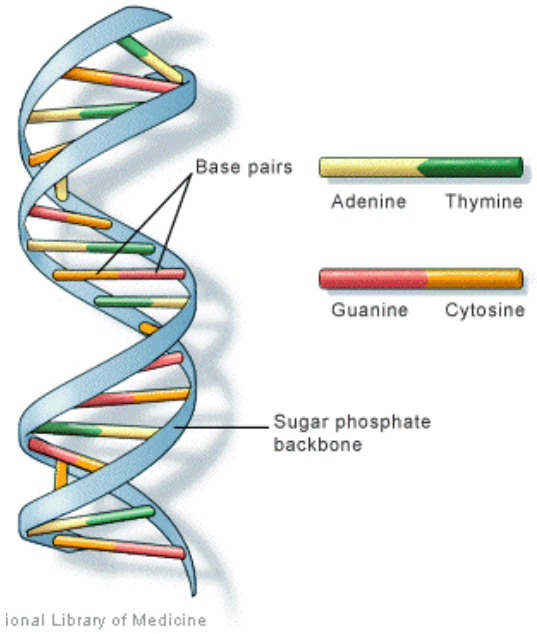

Figure 1. The structure of the DNA molecule [2]

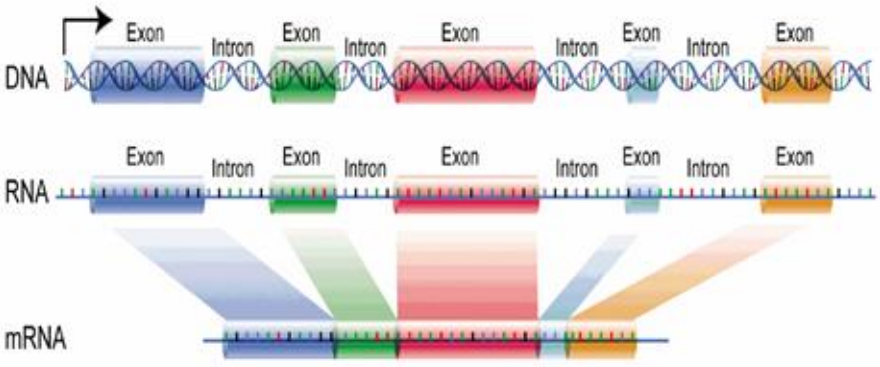

Figure 2. Exon-intron areas in eukaryotic DNA [2]

For a deeper understanding of the period-3 property in DNA sequence, we consider the periodic sequence $A--A--A--A--...,$ where the blanks '−' can be filled randomly by any one of the four bases *A*, *T*, *C* or *G*. This sequence gives a periodicity of three due to the repetition of base *A* at position 1 of each codon. *A* simple method to find the presence of a periodicity of three is to calculate the distance between the similar nucleotide frequencies. For a period of three, distance values of 2, 5, 8 and so on can be found between a particular at one position and other similar nucleotides in the sequence.

Presence of background noise in DNA sequences is a fundamental problem in determining the gene areas [7, 8]. Also, due to the complex nature of these areas usually a powerful tool is required to show the protein coding regions' characteristics effectively. Various methods are proposed to solve this problem. Generally, these methods can be divided into two main categories: Model-dependent algorithms or supervised methods, and model-independent algorithms or filter-based methods. Model-dependent algorithms like Hidden Markov model (HMM) [9] and neural networks (NN) [10], which are based on some primary information gathered from existing data, can be successfully used in prediction of exons in genes. However, the main problem with this method is that coding regions may not represented in the accessible datasets but exist in the sequenced

organism. To resolve this problem, filter-based methods which are based on spectral analysis [11, 12] have been turned into useful tools in gene detection. A variety of algorithms are proposed for determination of gene areas based on their period-3 properties. In [13] the Fourier transform has been implied for this purpose. In this method the gene areas are detected by selecting a fixed-length window and sliding it on the numerical sequence of the DNA and then applying the discrete Fourier transform and calculation of the resulting energy spectrum. In [14] the Anti-Notch Filter (ANF) with a $2\pi/3$ center frequency is used to remove non-coding regions. A new algorithm based on the Fourier transform and using the Bartlett window is proposed in [15] to predict gene areas. In [16], time series algorithms are used to determine the protein coding regions. Using the fixed-length window is the major restriction of the algorithms discussed above. In many cases, size of the selected window is not successful to predict the small size coding regions. For resolve this limitation, we have proposed an algorithm in [17] which is based on the Variable-Length Window (VLW) technique, but the major drawback of this method is high computational complexity due to applying S-Transform.

In this paper, a novel ensemble algorithm based on a combination of Goertzel Algorithm and Anti-notch filter has been presented to identify the protein coding regions in DNA sequences.

Material and Method

Databases

In early studies on genomic sequences, DNA sequences are directly extracted from the GenBank database [18]. This database is the most popular database in nucleotide sequence database and related documentations and was established in 1982 as part of the National Library of Medicine (NCBI). In total, there are more than 61 million sequences and more than 65 million base pairs in this database. In this paper, gene sequence with GeneBank reference number of F56F11.4 related to bases 1 to 9833 of chromosome III of *C. elegance* species is used. *C. elegance* is an intestinal parasite with a length of about 1mm which naturally lives in temperate soil environments. The five protein coding regions of the sequence are positioned in: 3156-3267, 4756-5085, 6342-6605, 7693-7872, and 9483-9833 bps [19].

In this article gene sequences from two other databases named HMR195 [20] and BG570 [21] are also used and these two databases are introduced as follows:

- **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.
- **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%.

The Proposed Algorithm

Figure 3 shows block diagram of the proposed algorithm for identification of protein coding areas. The main steps of the proposed algorithm, which will be completely described later, are as follows:

- Modeling the input symbolic strand using the Electron Ion Potential Interaction (EIIP) method,
- Using Anti-notch filter to remove the background noise from the numerical sequence, and

- Applying Goertzel Algorithm for detection of period-3 components.
- In Figure 3, $X_1(N/3)$ represents the period-3 magnitude component of numerical DNA data and it is calculated by choosing a Blackman window with the size of 351 (the length of the window should be multiple of 3). The Blackman window gives high weight to the codon positions that are center to the 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 such as rectangular. In order to capture the non-coding region noise and eliminate it from $X_1(N/3)$, the DNA sequence is first passed through a notch filter with the central frequency $2\pi/3$ and the filtered signal is then subjected to a similar sliding window to obtain the spectral output $|X_2(N/3)|$. In the difference signal SD(N/3), noise component in non-coding region is suppressed without affecting the coding region signals significantly; therefore this method can preserve the signal levels in the coding regions.

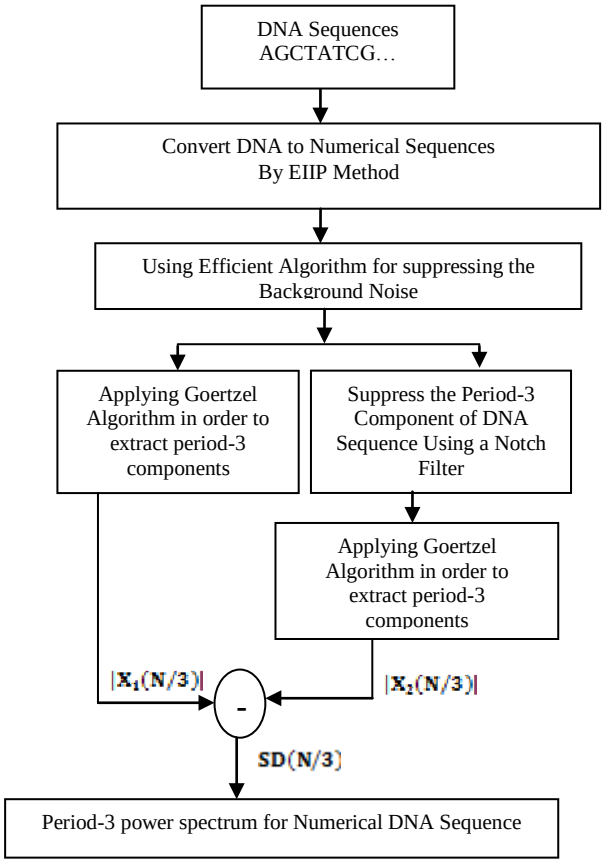

Figure 3. Block diagram of the proposed algorithm

Modeling the Input Symbolic Strand using EIIP Method

Modeling the input symbolic strand into digital signal is the first step in gene determining. In this paper, we have used EIIP method to convert DNA sequence into numerical signal. The EIIP values for the nucleotides are: A=0.1260, G= 0.0806, T=0.1335, C=0.1340.

Using Anti-notch Filter to Remove the Background Noise from the Numerical Sequence

In order to reduce the existence leakage in gene prediction techniques, it is necessary to utilize a window with high size. This phenomenon leads to increase the computational complexity and also reduce the resolution. To overcome this problem, we have used filters with infinite amplitude response which called 'Anti-notch filters'. The amplitude response of such filters has a sharp peak at the desired frequency ($2\pi/3$ in this study). The transfer function of such filter is as follow [22]:

$$G(z)=k\left[\frac{1-2R\cos\omega_0\, z^{-1}+z^{-2}}{1-2R\cos z^{-1}\, R^2\, z^{-2}}\right] \quad (1)$$

where

$$\cos\omega_0=\frac{2R\cos\theta}{1+R^2}$$

$$k=\frac{R^2+1}{2}$$

This filter has zeros at the frequency ω_0. We can see from Eq. (1) that the frequency ω_0 gets close to θ when R is close to unity. In this case, the pole and zero of the filter G (z) are very close to each other as illustrated in Figure 4. Therefore, at frequencies that are sufficiently away from ω_0, the magnitude response of G (z) will be close to unity. The notch filter is centered at frequency $2\pi/3$ and value of R is selected as 0.992 [22].

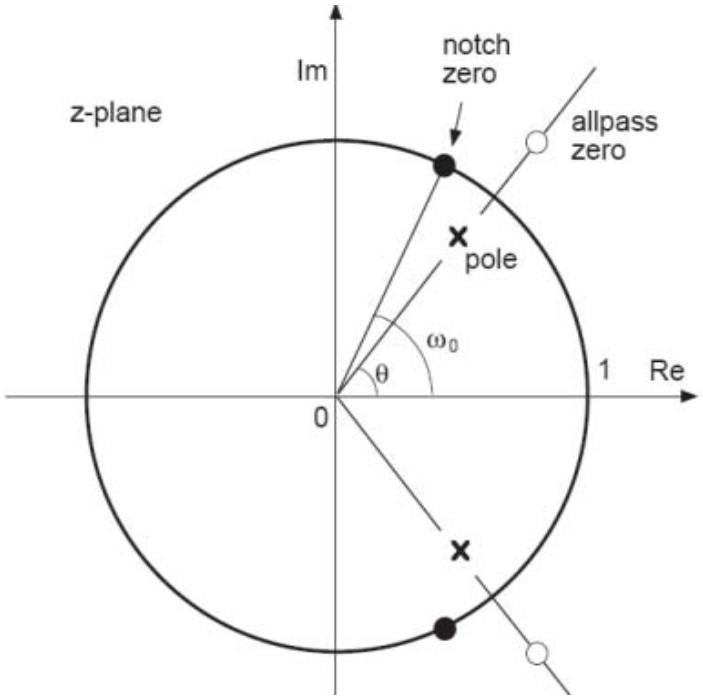

Figure 4. Poles and zeros of the notch filter G(z)

Using Goertzel Algorithm for Detection of Period-3 Components

The Goertzel algorithm is an efficient method to extract the single-tone components which has many applications in signal processing and communication systems . In this paper, this algorithm is used to extract the period-3 components by tuning the Goertzel algorithm at the frequency of $\frac{2\pi}{3}$. A second order realization of the Goertzel filter is also favoured over the direct DFT because it results in reduced computational burden. More details of Goetzel algorithm are explained in [23].

Assessment

In order to demonstrate the efficiency of the proposed algorithm, the evaluation criteria's at nucleotide level are used. These parameters are named False Positive (*FP*), Sensitivity (S_n), Specificity (S_p), Approximate Correlation (*AC*), and Area under the Receiver Operating Curve (*AUC*) which are explained in [23]. Also, in order to compare the computational efficiency of the proposed algorithm and other methods, the average CPU times over 1000 runs of the techniques for the sample gene sequence, F56F11.4 has been computed. We should mentioned 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.

Results and Discussion

In this paper, the Discrete Fourier Transform (DFT) and Multi-Stage Filter (MS) methods have been implemented in order to compare the performance of the proposed algorithm with these methods. All the simulation results are done in our designed user friendly package environment which is explained in [24]. Figures 5 (a)-(c) show results of implementation of these methods and the proposed algorithm which are applied on the gene sequences that were previously explained. As can be seen, the accuracy of the DFT method for detection of protein coding regions is not high because of the noise associated with the original signal. However, the MS-filter resulted a good spectral component compared to DFT and also reduced the computational complexity. In addition, the intron regions are relatively suppressed in it, but this method cannot detect the small size of protein coding regions. As can be seen in Figure 5(c), the large amount of noise is removed in the proposed method due to applying the AN-filter, and also small size of exons (For example, first exon in F56F11.4 gene sequence) can be identified because of using the Goertzel algorithm.

In Table 1, the values of *AC* and *Sp* for a fixed amount of *Sn* are presented in the proposed algorithm and other algorithms on gene sequences F56F11.4. As can be seen, the proposed algorithm has the highest value of the two parameters. The values of *Sp* and *AC* are 0.743 and 0.721, respectively. This implies that our proposed algorithm is superior to the other methods for identifying exonic gene regions.

Table 1. Comparison of the quantitative results of the proposed algorithm with other methods applied on gene sequence F56F11.4 (with *Sn* =0.30)

Method	Sp (Specificity)	AC (Accuracy)
DFT	0.182	0.091
MS-Filter	0.241	0.263
Proposed	**0.743**	**0.721**

Results of the average computational time are shown in Table 2. We observe that our algorithm has improved the average CPU times by the factor of 53.8 relative to the next-best performing method, MS-filter in F56F11.4 gene sequences.

Table 2. Average Computational Time computed for the different algorithms applied on gene sequence F56F11.4

Method	Average Computational Time (Second)
DFT	700.502
MS-Filter	652.245
Proposed	**12.124**

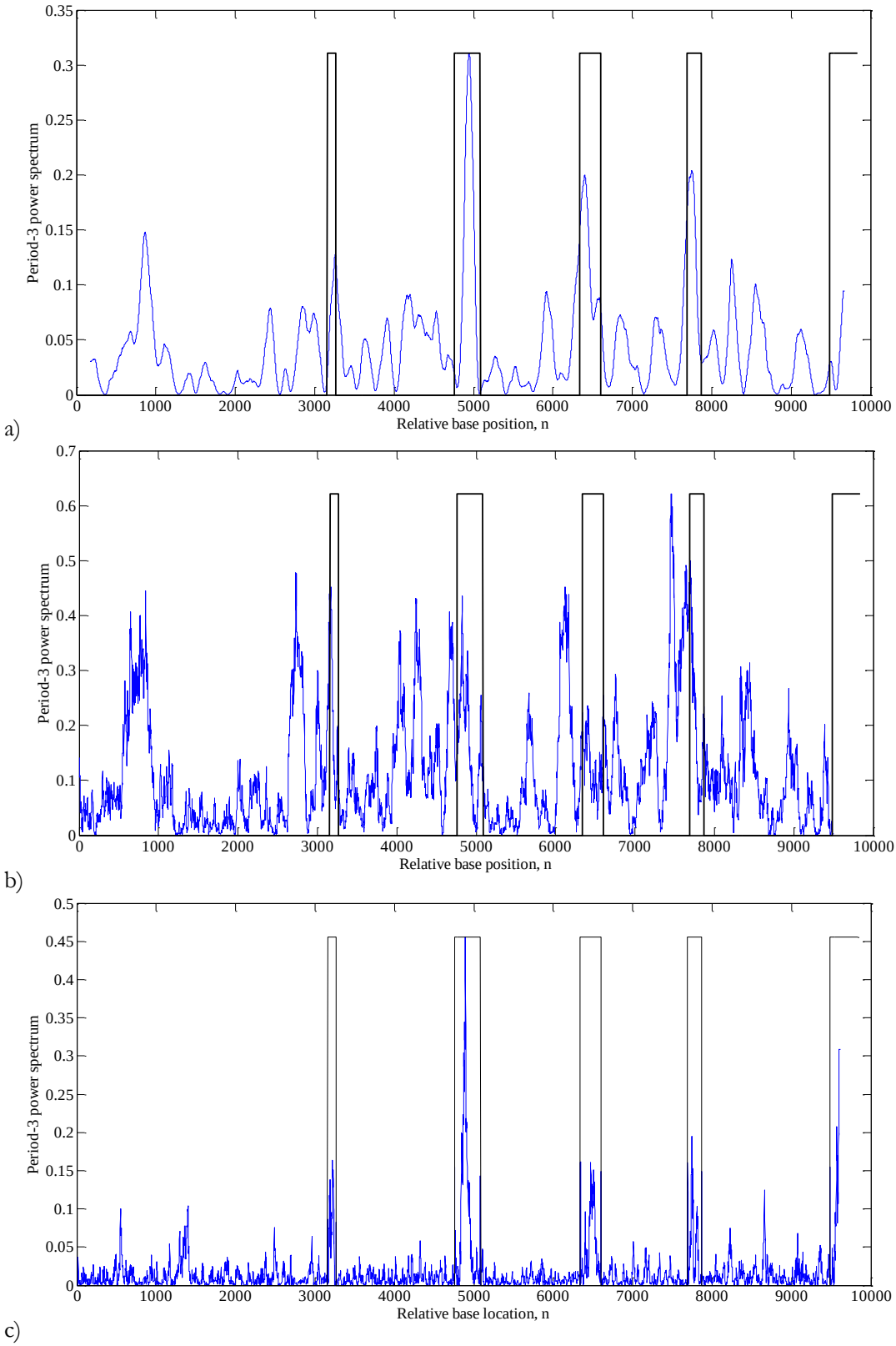

Figure 5. The algorithm result for detection of the exon regions on the gene sequence F56F11.4. (a) DFT, (b) MS-filter and (c) Proposed algorithm.

The results of applying of the proposed algorithm and other methods on a set of genes from BG570 database is shown in Table 3. It should be noticed that 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.

Table 3. Comparison of the quantitative results of the proposed algorithm with other methods applied on genes in BG570 database (with Sn =0.30)

Method	AUC	FP (False positive)	Sp (Specificity)	AC (accuracy)
DFT	0.6540	764	0.433	0.183
MS-Filter	0.6765	499	0.497	0.174
Proposed	**0.8265**	**65**	**0.813**	**0.625**

As can be seen from Table 3, the proposed algorithm has the least amount of *FP*. In case of *Sn* equal to 0.30, the number of incorrect nucleotides at the proposed algorithm improves by the factor of 7.7 in comparison to the best method, MS-filter. Also, the area under the ROC curve of the proposed method is improved by the deviation rate of 26.4% and 22.2% compared to the DFT and MS-filter methods, respectively. A similar superiority of the proposed algorithm is shown in Table 4 which relates to the HMR195 database. The value of *AC* of the proposed algorithm for *Sn* =0.30 equals to 0.748 while its value for the MS-filter method is 0.324.

Table 4. Comparison of the quantitative results of the proposed algorithm with other methods applied on genes in HMR195 database (with Sn =0.30)

Method	AUC	FP (False positive)	Sp (Specificity)	AC (accuracy)
DFT	0.6782	1184	0.453	0.181
MS-Filter	0.7615	562	0.574	0.324
Proposed	**0.8923**	**90**	**0.854**	**0.748**

In this paper, by using the EIIP method, mapping of DNA symbolic sequence into a numerical signal was investigated and an algorithm based on the combination of Anti-notch filter and Goertzel algorithms is proposed and its performance in terms of accuracy of estimation is assessed. An important advantage of the proposed algorithm is represented by noise reduction due to the use of Anti-notch filter. The ability to detect short exons is another advantage of the proposed algorithm. By comparing the proposed algorithm with other existing methods, it is seen that this algorithm, for datasets HMR195 and BG570, improves the area under the ROC curve from 4% to 45%. Our proposed method also reduces the number of incorrect nucleotides which are estimated to be in the coding region. This reduction results in increase of the specificity. For example, for Sn=0.30, Sp recovery rate of the proposed algorithm relative to other methods is from 15% to 85%. Combination of advanced signal processing techniques with the proposed algorithm and apply them on the real datasets explained in [25] can result in more accurate identification of the locations of protein coding regions with the least expenditure of time and computational complexity for any DNA sequences.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Snustad DP, Simmons MJ. Principles of Genetics, John Wiley & Sons Inc., 2000.

2. Dougherty ER. Genomic signal processing and statistics, EURASIP Book Series on Signal Processing and Communications, 2005.
3. Vaidyanathan PP, Yoon BJ. The Role of Signal Processing Concepts in Genomics and Proteomics. J Franklin Institute (special issue on Genomics) 2004;341:111-135.
4. Trifonov EN, Sussman JL. The Pitch of Cromatin DNA is Reflected in its Nucleotide Sequence. Proc Nat Acad Sci 1980;77:3816-3820.
5. Wan XF, Xu D, Kleinhofs A, Zhou J. Quantitative Relationship between Synonymous Codon usage Bias and GC Composition across Unicellular Genomes. BMC Evolutionary Biol 2004;4.
6. Herzel H, Trifonov EN, Weiss O, Groβe I. Interpreting Correlations in Biosequences. Physica A 1998;249:449-459.
7. Voss RF. Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences. Phy Rev Lett 1992;85:1342-1345.
8. Chatzidimitriou-Dreismann CA, Larhammar D. Long-Range Correlations in DNA. Nature 1993;361:12-213.
9. Henderson J. Finding Genes in DNA with a Hidden Markov Model. J Comput Biol 1997;4:127-141.
10. Ding CH, Dubchak I. Multi-Class Protein Fold Recognition using Support Vector Machines and Neural Networks. Bioinformatics 2001;17:349-358.
11. Anastassiou D. Genomic Signal Processing. IEEE Sign Proc Mag 2001;18:8-20.
12. Fox TW, Carreira A. A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression. EURASIP J Appl Aign Proc 2004;12:108-114.
13. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of Probable Genes by Fourier Analysis of Genomic Sequences. Comput Appl Biosci 1997;13:263-270.
14. Saberkari H, Shamsi M, Sedaaghi MH. Prediction of Protein Coding Regions in DNA Sequences using Signal Processing Methods. IEEE Symposium on Industrial Electronics and Applications (ISIEA), Bandung, Indonesia September 2012;354-359.
15. Datta S, Asif A. A Fast DFT-Based Gene Prediction Algorithm for Identification of Protein Coding Regions. Proc of the 30th International Conference on Acoustics, Speech, and Signal Processing 2005.
16. Akhtar M, Epps J, Ambikairajah E. Signal Processing in Sequence Analysis: Advanced in Eukaryotic Gene Prediction. IEEE J Sel Top Sign Proces 2008;2:310-321.
17. Saberkari H, Shamsi M, Sedaaghi MH. A Punctual Algorithm for Small Gene Prediction in DNA Sequences Using a Time-Frequency Approach Based on the Z-Curve. GSTF Int J Eng Technol 2013;2(1):4. doi: 10.5176/2251-3701_2.1.25.
18. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucl. Acids Res 2006;34:16-20.
19. National Center for Biotechnology Information [online] National Institutes of Health, National Library of Medicine. Available from: http://www.ncbi.nlm.nih.gov/Genebank/index.html (accessed on Februery, 2014).
20. Burset M, Guigo R. Evaluation of Gene Structure Prediction Programs. Genomics 1996;34:353-367. (Dataset is available at: htto://www.imim.es/GeneIdentification/Evaluation/Index.html).
21. Rogic S, Mackworth AK, Ouellette BF. Evaluation of Gene-Finding Programs on Mammalian Sequences. Genomes Res 2001;11:817-832. (Dataset is available at: htto://www.cs.ubc.ca/~rogic/evaluation/dataset.html).
22. Shakya DK, Saxena R, Sharma SN. A DSP-Based Approach for Gene Prediction in Eukaryotic Genes. IJEEI 2011;5:480-487.
23. Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A Fast Algorithm for Exonic Regions Prediction in DNA. J Med Sign Sens 2013;3:139-149.
24. 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) February 27th & 28th 2013; Tehran, Iran.

25. Ahmadi F, Saberkari M, Abiri R, Mohammadi Motlagh H, Saberkari H. In Vitro Evaluation of Zn-Norfloxacin Complex as a Potent Cytotoxic and Anti-Bacterial Agent, Proposed Model for DNA Binding. Appl Biochem Biotechnol 2013;170:988-1009.