Detection of Protein Coding Regions using Goertzel Algorithm

Sanjay Verma
Department of Electronics and Comm. Engg.
Samarat Ashok Technological Institute, Vidisha,
M.P., India

Devendra Kumar Shakya
Department of Biomedical Engg.
Samarat Ashok Technological Institute Vidisha,
M.P., India

ABSTRACT

Processing and interpretation of genomic sequence by DSP (digital signal processing) tools has attracted many researchers in last two decades particularly, the protein coding regions (exons) detection have been a challenging task in bioinformatics. The three base periodicity (TBP) or period-3 property of exonic regions form basis for most researchers for identification purpose. Many DSP based model dependent and model independent techniques have been applied for identification but still improvement is needed. In this article, a simple model independent technique using Goertzel algorithms proposed for exonic regions detection. The potential of the proposed method have been evaluated on the basis of performance parameters like sensitivity, specificity and correlation coefficient and found that the proposed method provides better performance than conventional DFT methods.

General Terms

DNA (deoxyribonucleic acid), Fourier transform, DSP, Exons, TBP (Three base periodicity).

Keywords

Protein coding regions, DNA (deoxyribonucleic acid) sequence, Goertzel algorithm, Period-3 property, Digital signal processing (DSP).

1. INTRODUCTION

Genomics research is of fundamental importance because it provides complete genetic information about living organisms. A DNA sequence is a long molecular chain that carries the genetic information of living organisms and it is composed of four types of different nucleotides namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) [1]. The DNA sequence is divided into genic regions and intergenic regions. In eukaryotic cells the genic region is divided into exonic regions and intronic regions as shown in figure 1. The exonic regions also called protein coding regions contain information about protein formation. It has been observed that the exonic regions strongly exhibit period-3 property because of non-uniform codon usage in the translation of codons into amino acids [2]. The period-3 property has been formed basis for most researchers for exonic regions prediction. The background noise (1/*f*) present in DNA sequence due to long range correlation of bases makes the task of exons prediction more complex [3]. The first step in processing the DNA sequence data by DSP based tools is to convert it into numeric form by suitable existing mapping technique. Today there are about seventeen numerical mapping techniques exists. In this paper we will discuss some popular and recently reported mapping schemes. In last two decades many DSP based model dependent and model independent methods of exons prediction have been successfully applied in this area. Model dependent methods like neural network, hidden markov model, GENSCAN, etc. have been used for identification [4]. The model dependent methods require prior knowledge about genome database of organizes, so it prefer for to predict same type of genome. Model independent methods do not require prior knowledge of genome organizes, so most researchers were attracted towards them. Most of the model independent methods are based on spectral content measures or spectral characteristics of genomic sequence. In this work, Goertzel filter [4] based algorithm is proposed to extract the period-3 components. Teager energy operator is applied to calculate the spectral energy of period-3 components. The remainder of this paper is organized as follows: In section 2 numerical mapping and period-3 power spectrum is discussed. Proposed algorithm discussed in section 3. In the section 4 simulation results are presented and in section 5 discussed the conclusion remarks.

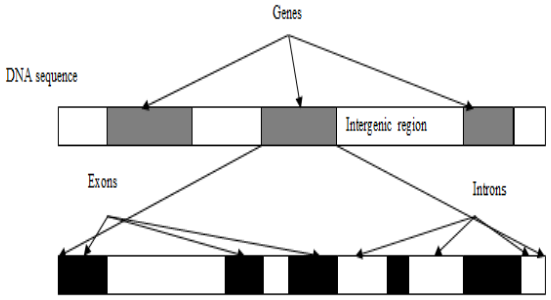


Fig 1: Classification of various regions of DNA sequence.

Table 1: Code 13 values of nucleotides

Nucleotide	Code13 representation
A	1
T	J
G	-1
C	-j

2. NUMERICAL MAPPING OF DNA SEQUENCE AND PERIOD-3 SPECTRUM

The conversion of character string of the DNA sequence into suitable numeric form is main requirement to apply any DSP tool. The main purpose of each numerical mapping method is to enhance the hidden biological information for analysis. At present there are about seventeen mapping methods exists but Voss mapping [6] and electron ion interaction potential (EIIP) mapping are two popularly used [7]. According to Voss mapping, assign numeral '1' to indicate the presence of particular nucleotide otherwise'0'. For a random gene sequence $x(n)$ = [ATCGAGCTAA….] for each nucleotide (A , T , C, and G) the numerical mapped signals are $x_A(n)$=[1000100011….], $x_T(n)$=[0100000100…], $x_C(n)$=[0010001000….] and $x_G(n)$=[000101000….] respectively. There is one drawback with the Voss mapping is that, it provides four numerical mapped sequences of a single DNA sequence, so it is more time consuming. In case of electron-ion interaction potential (EIIP) mapping scheme the complete DNA sequence is encoded into single numerical sequence and each nucleotide is represented by their particular EIIP values. The EIIP values for corresponding nucleotides are A=0.1260, C=0.1340, G=0.0806 and T=0.1335. For the above mentioned DNA sequence the EIIP mapped numerical code is $x_{EIIP}(n)$ = [0.1260 0.1335 0.1340 0.0806 0.1260 0.0806 0.1340 0.1335….]. In this work, a newly reported mapping scheme is used i.e. Code13 mapping [8] to convert DNA sequence into numerical mapped sequence. Code13 mapping technique is also one dimensional mapping like EIIP but in case of code13 mapping representation assign the values 1 , -1 , j, and –j to nucleotides A , C , T and G respectively, where j is imaginary unit. Code13 mapping technique provides good results for exons detection as compare to Voss and EIIP mapping techniques. Table-1 shows the code13 values for each nucleotide.

The DFT of a length N block of $x(n)$ is defined as X[k] and its energy spectrum is defined as S[k] given by equation (1) and (2) respectively.

$$X[K]=\sum_{n==0}^{N-1}x(n).w(n)e^{-\frac{j2\pi nk}{N}} \quad ,0\le n\le N-1 \qquad (1)$$

$$S[K]=\left|X[K]\right|^{2} \qquad (2)$$

Where $w(n)$ is a window function.

In the proposed work, teager energy operator is used to plot the period-3 power spectrum of output of Goertzel filter. The discrete form of teager energy operator is given by equation (3).

$$\Psi[X[K]]=X[K]^{2}-X[K-1]X[K+1] \qquad (3)$$

Where X[K] is output of Goertzel filter.

3. PROPOSED ALGORITHMS

The procedure discussed in proposed method for the detection of exons is compress of various steps as following-

1. Convert the DNA character string of interest into numeric sequence using Code13 mapping scheme.
2. Choosing appropriate window function and sliding it over the interested numerical sequence.
3. Apply Goertzel filter to extract period-3 components of the sequence.
4. Using teager energy operator (TEO) estimate the energy Y[x[n]] of the output of Goertzel filter. The peaks in the energy plot indicate the presence of exons.

The block diagram of the proposed method is shown in figure 2. The Goertzel algorithm [4] is a DSP technique used to compute the DFT spectrum of short sequences efficiently. This algorithm is mainly used in dual tone multi-frequency (DTMF) signals, digital multi frequency receivers, very small aperture terminals (VSAT) etc. The transfer function of second order Goertzel IIR filter is given by equation (4). Where, N is length of input sequence and K is index of computed DFT. To extract period-3 components take K=N/3. The realization of Goertzel filter is shown in figure 3. Goertzel algorithm has an advantage over other DFT computing techniques is that it reduced computational complexity and resolves N real multiplications and only single complex multiplication to compute a sample, whereas DFT requires N^2 complex multiplications. It is faster than other DFT computing techniques. Teager energy operator was first defined by J.F. Kaiser in 1991 as a non-linear energy operator and used to estimate signal energy and it provides high resolution energy plots with less circuit complexity as compare to other energy estimators [10]. The discrete version of the operator is defined by equation (3), where $X[K]$ is the output of Goertzel filter and $Y[X[K]]$ is its period-3 power spectrum.

$$H_K(Z)=\frac{1-e^{-J\left(2\pi\frac{K}{N}\right)Z^{-1}}}{1-2\cos\left(2\pi\frac{k}{N}\right)Z^{-1}+Z^{-2}} \qquad (4)$$

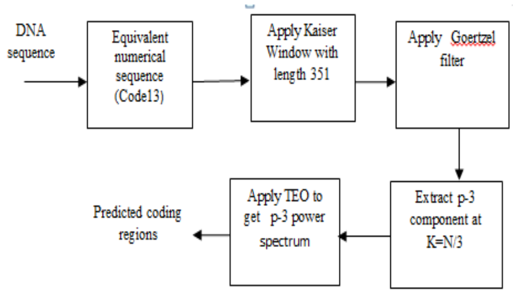

DNA sequence
Equivalent numerical sequence (Code13)
Apply Kaiser Window with length 351
Apply Goertzel filter
Extract p-3 component at K=N/3
Apply TEO to get p-3 power spectrum
Predicted coding regions

Fig 2: Block diagram of proposed algorithm.

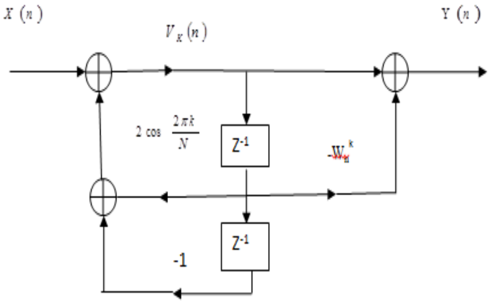


Fig 3: Filter realization of Goertzel algorithm.

4. SIMULATION RESULTS

In order to estimate the performance of proposed method, the period-3 plots and performance parameter results are compared with DFT spectral content method [13]. For the analysis purpose National Centre of Biotechnology Information (NCBI) provides public access to DNA sequences for analysis purpose [5]. The test is particularly done for the DNA sequence of Gene F56F11.4 of Caenorhabditis elegans Chromosome-III (gene bank Accession no.: FO081497 AF099922). According to NCBI data base, Gene F56F11.4 is 8100 bases long sequence contains five exonic regions range from 928-1039, 2528-2857, 4114-4377, 5465-5644, and 7265-7605. The comparative results for gene F56F11.4 and F56F11.4a are shown in figure-4 and figure-5 respectively and it is observed that the proposed method is better than DFT spectral content method of gene identification.

In order to compare the accuracy in exons detection, some performance parameters are introduced which are defined as Sensitivity (S_n), Specificity (S_P) and Correlation coefficient (CC) and are computed for threshold level of 20% ,40%, 60% and 80% [11],[13]. The formulation of these parameters is given below, Where TP, FP, TN and FN is parameters calculated by observing the NCBI data set (Actual region**s**) and predicted region**s** by proposed method. For example true positive (TP) is number of coding nucleotides correctly predicted as coding and false positive (FP) is number of non-coding nucleotides predicted as coding. True negative (TN) is number of non-coding nucleotides correctly predicted as non-coding and false negative (FN) is number of coding nucleotides predicted as non-codlings [12]. The definition of these parameters is illustrated by figure-6. Since the exonic regions exhibits strong period-3 property, the energy in these regions has higher peaks as compare to non-exonic regions.

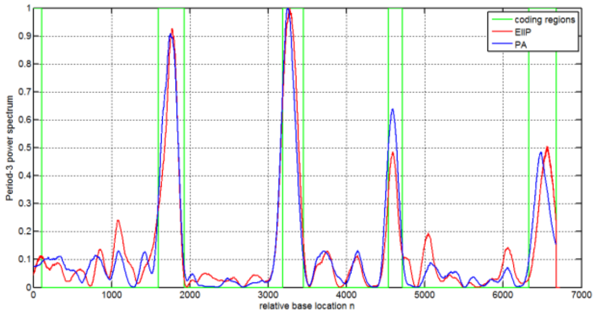


Fig 4: Comparative analysis for gene F56F11.4a.

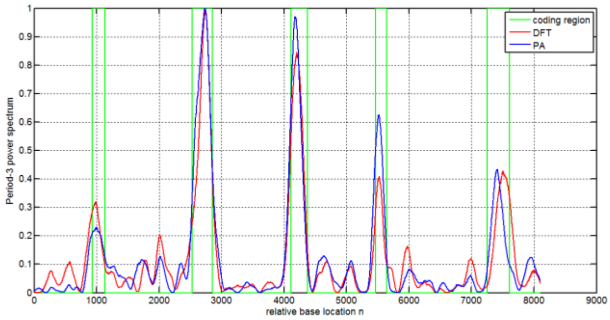
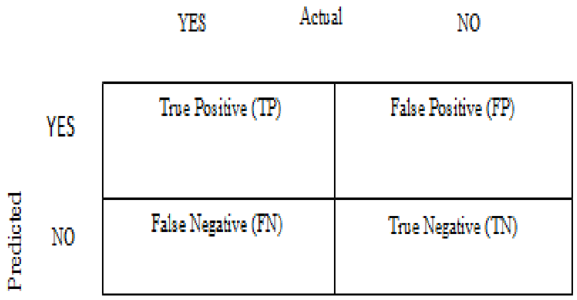


Fig 5: Comparative analysis for gene F56F11.4.



Predicted	Actual: YES	Actual: NO
YES	True Positive (TP)	False Positive (FP)
NO	False Negative (FN)	True Negative (TN)

Fig 6: Definition of observing parameters.

Table 2: Comparative analysis of performance parameters for Gene F56F11.4.

Threshold	Proposed Algorithm			DFT (EIIP mapping)		
	Sn	Sp	CC	Sn	Sp	CC
0.2	0.8266	0.8247	0.7913	0.6188	0.6721	0.5481
0.4	0.5188	0.9152	0.6494	0.3182	0.7851	0.5255
0.6	0.3333	0.9752	0.5348	0.1695	0.8120	0.3125
0.8	0.1839	1.0000	0.3981	0.0829	0.8049	0.2142

5. CONCLUSION

In case of exons detection the accuracy and speed are always desired. In this paper, a simple and fast model independent method based on Goertzel algorithm has been proposed. In order to demonstrate the performance of proposed algorithm, test data is adopted from NCBI (National Centre for Biotechnology information). The test is performed over gene sequence F56F11.4 and F56F11.4a of C. elegans Chromosome-III and results are compared with DFT based spectral content method. Simulation is performed in MATLAB and it is observed that proposed method provides better results for exons detection. The potential of the proposed method is that it is speedy and accurate.

6. ACKNOWLEDGMENT

The authors would like to express their sincere thanks to Dr. S.N. Sharma, Head of department of electronics and Communication Engineering, for his timely support for this research work. Authors would also like to express their sincere thanks to all the authors of the reference papers.

7. REFERENCES

[1] J. Tuqnan and A. Rushdi, "A DSP Approach for finding the codon bias in DNA sequence," IEEE Journal of Selected Topics in Signal Processing, vol. 2, no. 3, pp. 343-356, June 2008

[2] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, vol. 18,no. 4, pp. 8-20, July 2001.

[3] R.F. Voss ''Evolution of long-range fractal correlation and 1/f noise in DNA base Sequences, 'Physical Review Letters, vol. 68, no. 25,pp.3805-3808,june 1992.

[4] H. Sabarkari, M. Shamsi, H. Heravi. and M.H. sedaaghi "A novel fast algorithm forexon prediction in eukaryotic genes using linear predictive coding model and.Goertzel algorithm based on the Z-curve,"Journal of Medicalsignal and sensors, vol.3,PP.139-149,2013

[5] National Center for Biotechnology Information, Available: http://ncbi.nlm.nih.gov/.

[6] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction," IEEE Journal of Selected Topics in Signal Processing, vol. 2,no. 3, pp. 310-321, June 2008.

[7] K. D. Rao and M. N. S. Swamy "Analysis of genomics and proteomics using DSP techniques, " IEEE Transactions on Circuits and Systems-1, vol. 55, no. 1, pp. 370-378,February 2008.

[8] W. F. Zhang and H. Yan, “Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences,” Pattern Recognition, vol. 45, no. 3, pp. 947–955, 2012.

[9] F. J. Harris, “On the use of windows for harm.onic analysis with the discrete fourier transform,” Proc. IEEE, vol. 66, pp. 51–83, 1978.

[10] J.F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal”, Proceedings of the IEEE ICASSP-90, Albuquerque, NM, pp-381-384, April 1990.

[11] C. Burge, “Identification of genes in human genomic DNA”, Ph.D. dissertation, Stanford University, Stanford, CA, 1997.

[12] D. K. Shakya, Rajiv Saxena, and S. N. Sharma, “A Simple Algorithm for Gene Prediction with Improved Noise Suppression”, Proceedings of the 10th IEEE International Conference on Signal Processing, Beijing, China, 2010, pp.1765-1768.

[13] D.K. Shakya, Rajiv Saxena and S.N. Sharma, “A DSP Based Approach for Gene Prediction in Eukaryotic Genes”, IJEEI, vol 3, no.4, 2011.

8. AUTHOR PROFILE

Sanjay Verma received the B.E. degree in Electronics and Communication Engineering from Rajiv Gandhi Prodyogiki Vishwavidyalaya Bhopal, M.P., India, in 2012, and currently working as M.Tech. Scholar in Electronics and Communication Engineering department in Samrat Ashok Technological Institute, vidisha, M.P., India. His research interest includes Genomic signal processing and Image processing.

D.K. Shakya received degree in Electronics and Instrumentation Engineering from Barkatullah University, Bhopal, M.P., India, in 1999, and the M.E. in Digital Techniques and Instrumentation from Rajiv Gandhi Prodyogiki Vishwavidyalaya Bhopal, M.P., India, in 2002 and has received Ph.D. degree from Rajiv Gandhi Prodyogiki Vishwavidyalaya Bhopal, M.P., India in 2013. He is currently working as an Assistant Professor in the Department of Biomedical Engineering, Samrat Ashok Technological Institute, Vidisha, M.P., India. His teaching and research interests include Genomic and Proteomic Signal Processing, Digital Filter Design, Bio-signal and Image Processing.