Biomedical Engineering: Applications, Basis and Communications, Vol. 31, No. 1 (2019) 1950002 (10 pages)
DOI: 10.4015/S1016237219500029

USING DIT-FFT ALGORITHM FOR IDENTIFICATION OF PROTEIN CODING REGION IN EUKARYOTIC GENE

Subhajit Kar, Madhabi Ganguly* and Saptarshi Das
Department of Electronics, West Bengal State University
Barasat, Kolkata 700126, India

Accepted 8 August 2018
Published 13 December 2018

ABSTRACT

The new research platform on biomedical engineering by Digital Signal Processing (DSP) is playing a vital role in the prediction of protein coding regions (Exons) from genomic sequences with great accuracy. We can determine the protein coding area in DNA sequences with the help of period-3 property. It has been seen that in order to find out the period-3 property, the DFT algorithm is mostly used but in this paper, we have tested FFT algorithm instead of DFT algorithm. DSP is basically concerned with processing numerical sequences. When digital signal processing used in DNA sequences analysis, it requires conversion of base characters sequence to the numerical version. The numerical representation of DNA sequences strongly impacts the biological properties mirrored through the numerical genre. In this work, the proposed technique based on DIT-FFT algorithm has been used to identify the exonic area with the help of integer value representation for transforming the DNA sequences. Digital filters are used to read out period 3 components from the output spectrum and to eliminate the unwanted high frequency noise from DNA sequences. To overcome background noise means to suppress the non-coding regions, i.e., Introns. Proposed algorithm is tested on four nucleotide sequences having single or multiple numbers of exons.

Keywords: Biomedical Signal Processing; DNA; DIT-FFT; FIR filter; Kaiser window.

INTRODUCTION

Biological experiments are analysed through biomedical signal processing to provide useful information. Thus necessary decisions can be made by clinicians based on that information. Engineers are discovering new ways to process these signals using a variety of mathematical formulae and algorithms. Bio-measurement software tools can provide physicians with real-time data and accurate insights to aid clinical assessments. Recently, DSP techniques have been applied in every research area of biomedical engineering for diagnosis and disease management. In our nature most of the signals and processes are presented as a continuous form but the genetic information stored in the human's genome in the form of protein, is actually a discrete form in nature.[1] DNA information is "discrete" both in terms of amplitude and time and it calls for probe and exploration by means of digital signal processing.

DNA, or deoxyribonucleic acid, is the material present in almost all organisms transferring heredity information from generation to generation. It is made up of genes and intergenic spaces — exons and introns. Each gene contains a particular set of instructions, usually coding for a particular protein or for a particular function. Only exonic area is responsible for making protein.

*Corresponding author: Madhabi Ganguly, Department of Electronics, West Bengal State University, Barasat, Kolkata 700126, India. E-mail: ray_madhabi@yahoo.co.in

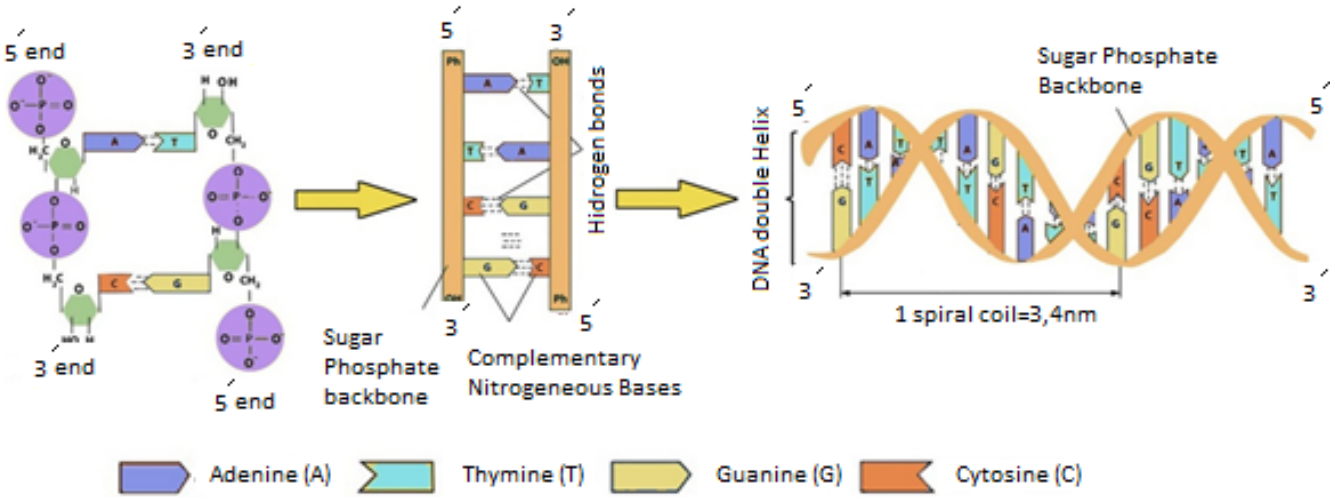


Fig. 1 DNA double helix structure.

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Four nucleotide bases take two bases to form base pairs that make up rungs in the spiral DNA ladder, forming 5′ end and 3′ end double helix structure of DNA molecule. Helical structure is stabilized by hydrogen bonding between complementary base pairs (Fig. 1). The base pair of DNA are A with T on the other hand C with G. Exons (or coding regions) coding for proteins are rich in nucleotides C and G whereas Introns (or non-coding regions) that do not code for proteins are rich in nucleotides A and T. In genes, the protein-coding sections of the DNA (exons) are interrupted by non-coding regions (introns). By splicing process introns are removed from the genes to produce the final set of instructions for the protein (Fig. 2).

During the protein synthesis 3-adjancent bases on a DNA molecule, known as "codon" determines the position of a specific amino acid in a protein molecule. There are 64 possible codons [(nucleotide alphabet size)$^{\text{word length}} = 4^3$], mapped into 20 amino acids that combine together to form proteins. All the amino acids except two (tryptophan and methionine) can be coded by more than one codon (i.e., code is said to be degenerate), those two has a unique codon. The degeneracy of the codons is not uniform. Non uniform codons are responsible for formation of the protein. In protein coding region, there is a type of property, which exists due to non-uniform codons biased in the translation of codons into amino acids during protein synthesis, known as "period-3 property". It can be shown by performing Fourier analysis on signals derived from segments of DNA sequences.[2]

Over the last few years several methods by Graphical representation techniques have been applied to reveal the pattern in DNA sequences in a compact graphical form in one, two or three dimensions that can be expanded as necessary to identify regions of extensive

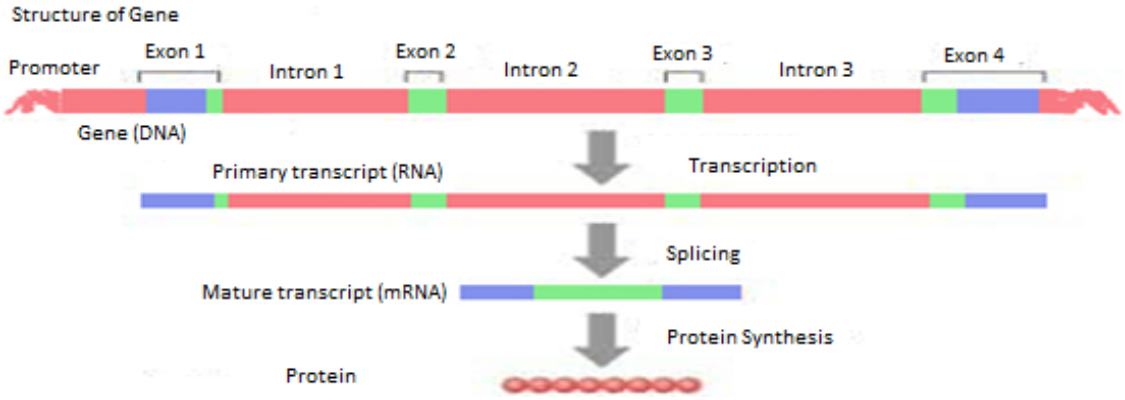


Fig. 2 Eukaryotic protein coding process involving gene.

repetitive sequences, distinguish between coding and non-coding regions of protein. Traditional graphical representation techniques do not reveal about genome structure to identify hidden periodicities and features in DNA sequences. But DSP techniques help to achieve varieties of goals such as detection of protein coding region in DNA sequences, abnormalities present in the coding region etc. That is why this technique is mostly preferred and used in genomic DNA research than conventional DNA symbolic and graphical representation techniques.[3]

DNA sequences are digital in nature therefore proper numerical values can be assigned to represent DNA sequences for DSP-based analysis. The numerical representation is suitable for the quantitative analysis of the sequences.[4] Some possible desirable properties of DNA numerical representation include:

(1) Each nucleotide has equal weight (e.g., magnitude); since there is no biological evidence to suggest that one is more important than another.
(2) Distance between all pairs of nucleotides should be equal, since there is no biological evidence to suggest that any pair is closer than another.
(3) Representations should be compact, in particular, redundancy should be minimized.
(4) Representations should allow access to a range of mathematical analysis tools.

In genomic signal processing various useful tools of DSP technique will be applied when DNA character strings are mapped into one or more numerical sequences. When the mapping is done the resulting numerical sequences are mostly acceptable for DSP applications such as gene prediction which leads to identification of Protein coding regions (exons) in a long DNA sequences. Therefore, it is necessary to map character symbols into numerical sequences.[5–7] Fourier transform is an analytical tool that takes a signal and expresses it in terms of the frequencies of the waves that make up that signal which is important for many applications in digital signal processing. In many previous papers, authors have suggested several processes based on DFT algorithms to determine the coding region but it requires more processing time, and complexity than FFT. FFT is an accelerated version of the DFT, producing ultimately the same result. Instead of working with a long signal, it is divided into smaller signals and performs DFT of these smaller signals. At the end all smaller DFT are added to achieve the actual DFT of the big signal.

This paper is organized as follows: Section 2 describes proposed DIT-FFT algorithm for identification of exons in genes using digital filters. The nucleotide bases are mapped into Integer value DNA representation, and then DIT algorithm is applied on those integer values of the bases. After that the resulting values of the bases are passed through the designed filter with Kaiser Window function. The result of the proposed methods with respect to various evaluation measures, are explained in Sec. 3. Finally, conclusions are given in Sec. 4.

MATERIALS AND METHODS

Proposed Algorithm and Block Diagram

FFT is a powerful technique that is commonly used for detecting periodicity, patterns and tandem repeats in DNA sequences (Pasquier *et al.*, 1998). For the identification of protein coding regions of genes in DNA sequences, the use of this process was described by Fickett and Tung (1992).[6,7] The DIT–FFT algorithm is a DSP tool which can be implemented using MATLAB software module. The given input of DNA sequences is processed by this software module which performs the DSP operations on it, and provides the output power spectrum as a result.

The following block diagram in Fig. 3 shows that there are five steps in the entire process. The data is collected from HMR 195 datasets. For applying the DSP technique on DNA sequences, it is mapped into indicator sequences as integer form by integer number representation mapping technique. After that DIT-FFT algorithm is applied on the genomic sequences with the help of integer values. Hence, Output values of the bases will be applied on the DNA character strings. The period-3 property will be observed in exonic region and can be obtained by FIR filter with Kaiser Window functions. The output power spectrum of the designed filter will indicate the exons by certain peaks.

DNA Sequence Database

The different eukaryotic genes of DNA sequences were downloaded from HMR 195 dataset (accession number AF055080, AF058762, AF028233, AF013711) provide by SangaRogic.[8]

Here, the modeled genes were chosen on two conditions. The first condition was that the length of the sequence should not over run 10000 base pairs and the second one was that the exons number should be less than five exons.

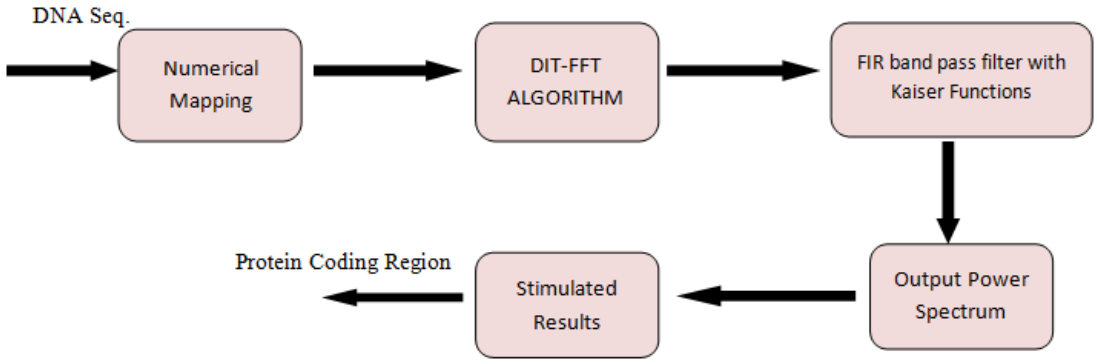


Fig. 3 Block diagram of the proposed method of exon identification.

Methods for Mapping of DNA Sequences

In mapping technique, nucleotides of DNA sequences are transformed into numerical signals. Many mapping techniques are used to convert the character symbols into numerical sequences. Some conventional methods of DNA numerical analysis are the real number,[9] integer number, complex number,[10] quaternary code,[11] EIIP,[12] atomic number,[13] paired numeric,[13] etc.

Assessing the four bases with binary representations, in this paper, we have applied Integer Number mapping methods on four nucleotide bases at initial stage. The integer number representation[11] is a 1D mapping of DNA bases. The four nucleotide bases are mapped by numerals {0, 1, 2, and 3}. We can choose a simple representation such as A = 0, T = 1, G = 2, C = 3 and use modulo operations, but this method implies a structure on the nucleotides such that T > A and C > G.[14] When DIT-FFT algorithm will be applied on the bases of the genomic sequences the integer values of the bases will be represented as their corresponding 2-bits binary no. i.e., A = 00, T = 01, G = 10, and C = 11.

DIT-FFT Algorithm

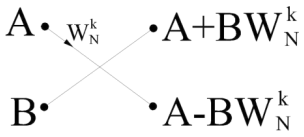
FFT is an algorithm for computing DFT. It is a highly efficient procedure for computing the DFT of a finite series and requires less number of computations than that of direct evaluation of DFT, which is defined on set of N samples $\{x_n\}$, as followed:

$$X_k = \sum_{n=0}^{N=1} x_n e^{-j2\pi kn/N}, \quad k = 0, 1, 2, \ldots, N = 1.$$

Decimation-in-time (DIT) algorithm is also known as Radix-2 DIT FFT algorithm which means the number of output points N can be expressed as a power of 2, i.e., $N = 2^M$, where M is an integer. In DIT algorithm, the N point sequence is divided into $N/2$ -point $X_0(n)$ odd number and $X_0(n)$ even number subsequences.

For DIT, it has been found that the input is a bit reversal while the output is in natural order [i.e., $X(k)$, $k = 0, 1, \ldots, N-1$] For a 4-point DFT algorithm, the input sequences are in the order of $x(0), x(2), x(1), x(3)$. In the process of FFT algorithms, a butterfly is a portion of the basic computation that combines the results of smaller DFTs into a large DFT or vice versa. The name "butterfly" comes from the shape of the data-flow diagram in the radix-2 case.[15] The FFT algorithm reduces the number of computations. The total number of complex multiplications required for calculating DIT-FFT $= N/2 \log_2 N$ and the total number of complex additions for evaluating a DFT using DIT-FFT is FFT $= N \log_2 N$.

Here, the nucleotide bases of the DNA sequences is computed by DIT-FFT algorithm as a 4-point DFT of a sequence $x(n) = \{0, 1, 2, 3\}$, By integer number mapping technique, A = 0, T = 1, G = 2, C = 3.

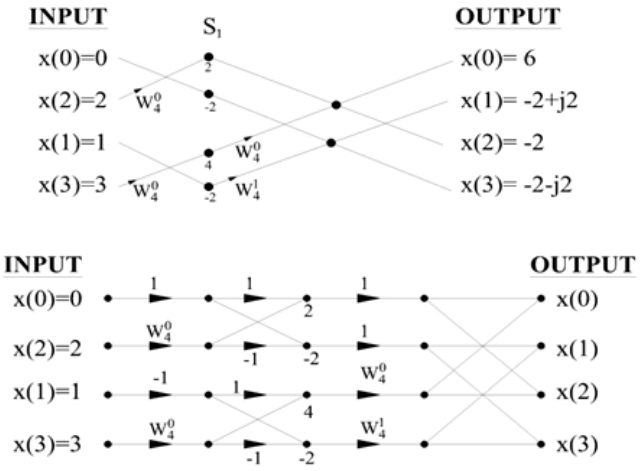


By DIT algorithm, Twiddle factors associated with butterflies are, $W_1^0 = 1$, $W_4^1 == e^{-j2\pi/4} = -j$

Bit reversal of input is given by,

Input Index	Binary Index	Bit Reversal	Bit Reversal Index
A = 0	00	00	0 = A
T = 1	01	10	2 = G
G = 2	10	01	1 = T
C = 3	11	11	3 = C

By Butterfly Operation, $X[k] = \{6, -2 + j2, -2, -2 - j2\}$. Therefore, the output values of four nucleotides are $A = 6, T = -2 + j2, G = -2, C = -2 - j2$.



DIT algorithm is an efficient way to evaluate the discrete convolution of a very long signal $x(n)$.

Design FIR Filter with Kaiser Window Function

The next step of the process is to pass the output of the nucleotide bases applied on the given DNA sequences (For example, TTAGCTGAC. . ..) through the FIR filter with Kaiser Window function having specification:

Filter order $N = 170$, the lower & upper cut off frequencies [0.662, 0.667].

Savitzky–Golay filter has been used for smoothing the filter response. Lack of distortions in FIR filters is one reason for their preferred use over IIR filters in biomedical applications.

There are many methods to design FIR filters. The essential methods for the design FIR filters are: (1) The window method, (2) The frequency sampling technique, and (3) optimal filter design method. The major advantage of using the window method is their relative simplicity as compared to other methods and ease of use. The Kaiser window is an approximation to the prolate spheroidal window, for which the ratio of the main lobe energy to the side lobe energy is maximized. For a Kaiser window of particular length, the parameter β controls the side lobe height. Kaiser estimates the minimum FIR filter order that will exactly meet the specifications. Kaiser converts the given filter specifications into pass band and stop band ripples and converts cut-off frequencies into the form needed for windowed FIR filter design. This window is a kind of adjustable window function which provides the independent control of the main lobe width and ripple ratio.[16]

Output power spectrum

Output power spectrum of the protein coding region is emphasized by the FIR filter based on property of 3-bases periodicity. The central frequency of the FIR filter is set to $2\pi/3$, to emphasize the period-3 property in the exonic regions. The period-3 property has been used as a preliminary indicator in gene prediction.

The period three property states that the spectral energy,

$$|S[k]|^2 = |A[k]|^2 + |T[k]|^2 + |G[k]|^2 + |C[k]|^2.$$

Derived from the DFT's of the four binary signals representing a DNA protein coding region of length N, exhibits a peak at discrete frequency $k = N/3$ (i.e., the spectrum of protein coding DNA has a peak at every third component at the frequency of $2\pi/3$).[17] By this process output spectral of the gene sequence can be exploited by to locate protein coding regions.

Evaluation Criterion

To measure the effect of various filtering methods to extract protein coding regions, we have estimated different evaluation parameters. These include sensitivity, specificity, precision, accuracy, F1 score and Matthews correlation coefficient (MCC). The definitions of these parameters are given in Table 1.

Here TP stands for True Positive, TN stands for True Negative, FN stands for False Negative, and FP stands for False Positive. P is the sum of TP and FP. Similarly N is the sum of TN and FN. Figure 5 describes different portions of annotated gene to find evaluation parameters at nucleotide level.

Here, Sensitivity (S_n) is the capability of the representation scheme to predict the true exons and specificity is the capability of the representation scheme to exclude the false exons. Matthews correlation coefficient (MCC) is the measure of accuracy and it ranges

Table 1. Definitions of Various Evaluation Parameters.

Parameters	Formula
Sensitivity	$\frac{TP}{(TP+FN)}$
Specificity	$\frac{TN}{(FP+TN)}$
Precision	$\frac{TP}{(TP+FP)}$
Accuracy	$\frac{(TP+TN)}{(P+N)}$
MCC	$\frac{TP*TN-FP*FN}{\sqrt{((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))}}$
Fl score	$\frac{2TP}{(2TP+FP+FN)}$

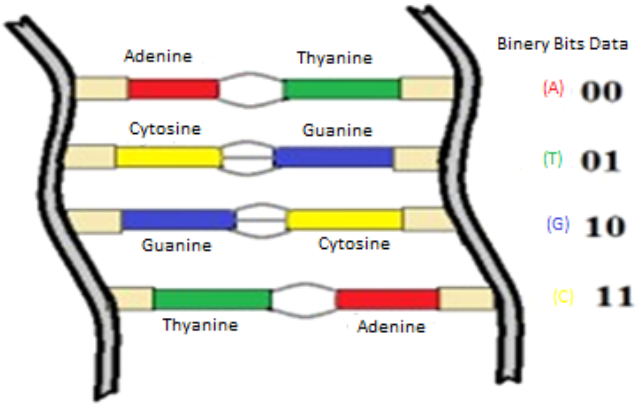


Fig. 4 Binary index of nucleotide bases.

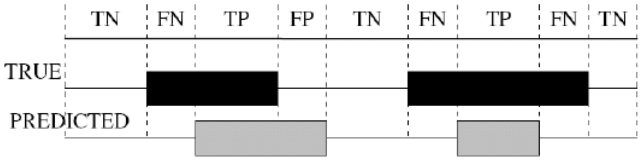


Fig. 5 Measures of prediction accuracy at the nucleotide level.

from −1 to 1. The precision (P) parameter is a measure of the system's correct identification of exon. Accuracy is a widely used compound measure which considers both sides of S_n and S_p from the global perspective. F1 Score is the weighted average of precision and sensitivity. Therefore, this score takes both false positives and false negatives into account. The value of F1 score is 1 for best case and 0 for worst case.

RESULTS

To reach our goal, the detection technique which is discussed in this paper is simulated using FIR band pass filter with Kaiser window functions on four human testing genes with single and multiple exons which were downloaded from HMR 195 dataset. In Table 2, the accession number, gene description, sequence length and true exon locations of the Genes are shown.[10] In Figs. 6–9, results of four sample genes with different exon numbers are shown.

Average CPU time is calculated for 1000 runs of the techniques for four gene sequences in order to compare

Table 2. Datasets of Human Genes Collected from HMR 195 Dataset.

Gene Accession Numbers	Name of the Organisms	Exon Types	Sequence Length	True Exon Locations
AF055080	Homo sapiens winged-helix transcription factor forkhead 5 gene.	One Exon Gene	2078 bp	964–1938
AF058762	Homo sapiens galanin receptor subtype 2 (GALNR2) gene	Two Exons Genes	3036 bp	115–482 and 1867–2662
AF028233	Homo sapiens distal-less homeobox protein (DLX3) gene	Three Exon Genes	4575 bp	68–392, 1483–1673 and 3211–3558
AF013711	Homo sapiens 22 kDa actin-binding protein (SM22) gene	Four Exon Genes	5388 bp	3643–3822, 3935–4112, 4410–4512 and 4843-4987

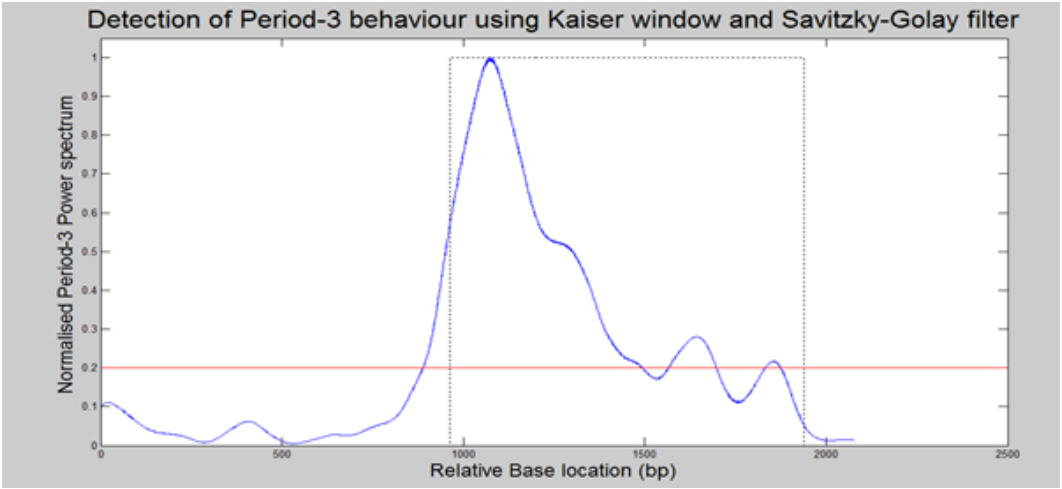


Fig. 6 Power spectrum of AF055080 gene.

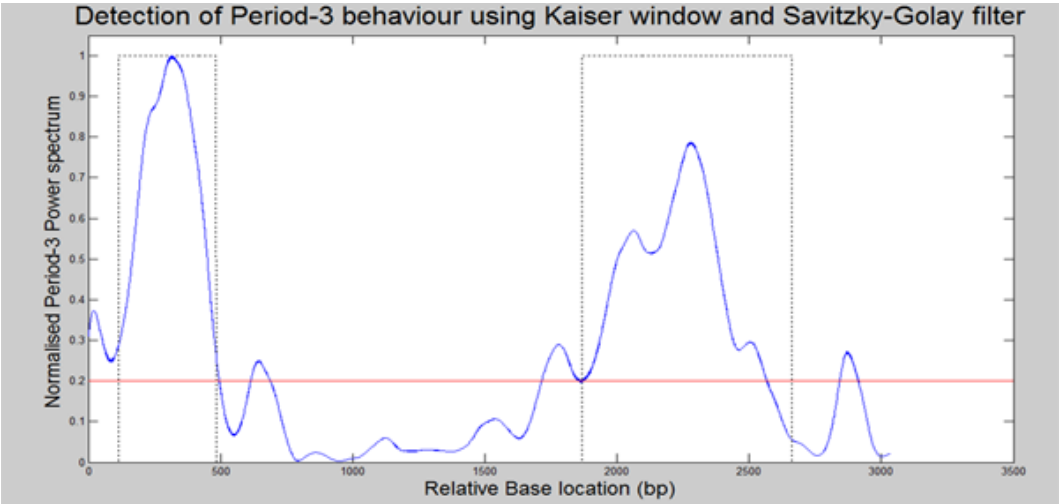


Fig. 7 Power spectrum of AF058762 gene.

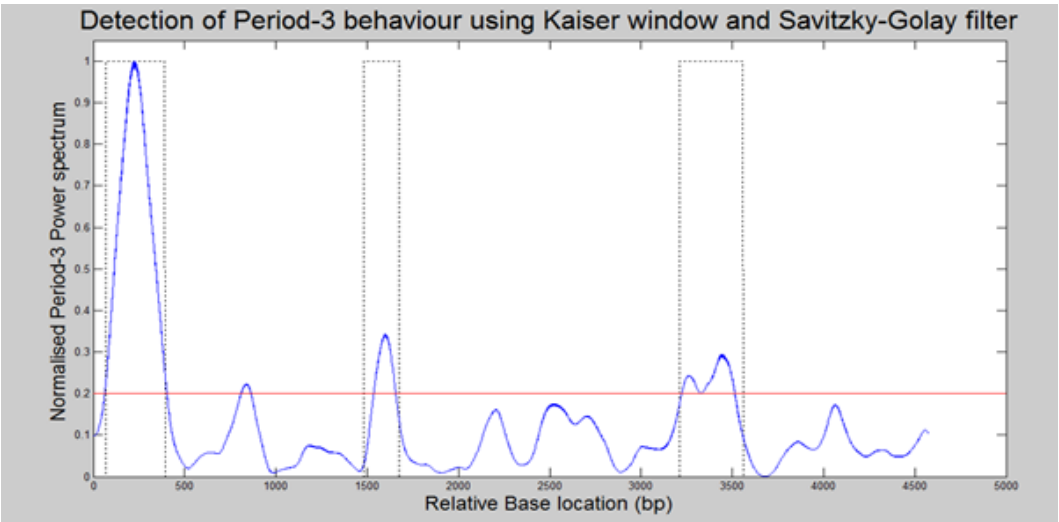


Fig. 8 Power spectrum of AF028233 gene.

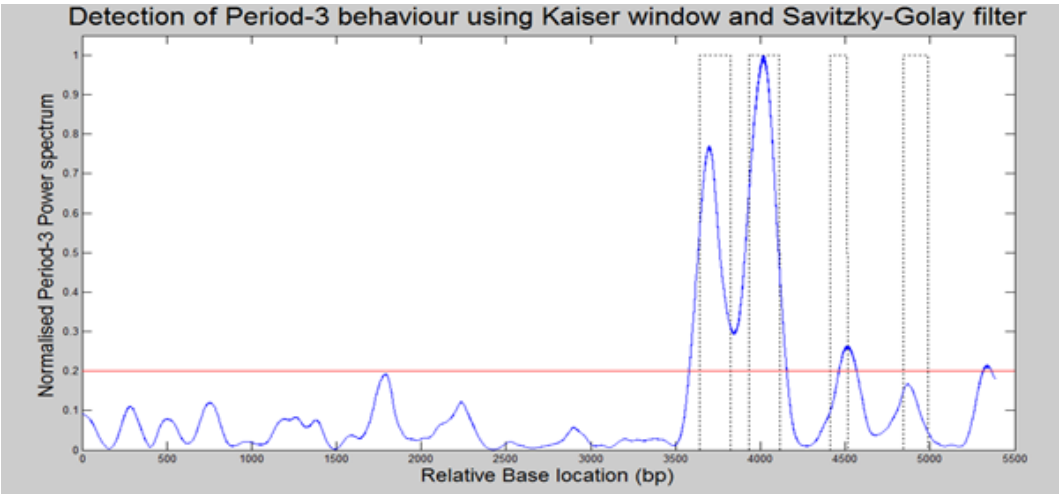


Fig. 9 Power spectrum of AF013711 gene.

Table 3. Average Computational Time Computed for the Different Algorithms.

Nucleotide Sequence	Length of the Sequence	Average Computational Time (in Sec)		
		Two Stage Filter	DFT	DIT-FFT
AF055080	2078 bp	0.0371	1.9422	0.000423
AF058762	3036 bp	0.0381	4.0172	0.000469
AF028233	4575 bp	0.0428	9.1209	0.000651
AF013711	5388 bp	0.0473	12.7632	0.000757

Table 4. Predicted Exon Ranges of the Sequences using DIT-FFT.

Nucleotide Sequence	Gene Description	Exon 1	Exon 2	Exon 3	Exon 4
AF055080	One Exon Gene	891–1875	—	—	—
AF058762	Two Exon Gene	1–497	1715–2568	—	—
AF028233	Three Exon Gene	63–404	1535–1657	3221–3515	—
AF013711	Four Exon Gene	3579–4155	4464–4569	—	—

the computational efficiencies of the algorithms with the previous methods. All the algorithms were run on a PC with a 2.2 Ghz processor (Intel Core 2 Duo) and 2 GB of RAM. Table 2 summarizes the average CPU times taken by previous algorithm and proposed algorithm. It is observed that the proposed algorithm has improved the average CPU times by the factor of 4591, 8565, 14010 and 16860 relative to the DFT in AF055080, AF058762, AF028233 and AF013711 gene sequences, respectively.

We have summarized the predicted introns and exons locations in Table 4. The adopted algorithm is very efficient to predict the intronic and exonic regions in various types of gene. The results suggest experimental findings are very close to NCBI ranges.

DISCUSSION

In this study, the effect of butterfly method in DIT-FFT algorithm on the accuracy of exon regions in DNA sequences was discussed for four different human gene using FIR filter. The power spectrum of genes by period-3 behavior has been shown in the above figure (Figs. 6–9). The dotted line describes the actual coding regions of the gene.

In this work it is shown that the proposed method has the advantage of better specificity and correlation coefficient when it comes in comparison with methods opted by others.[18,19] Most of the representation produces specificity of 75.9% and correlation coefficient of 0.8, in case of gene AF028233 whereas the proposed algorithm generates the specificity of 0.9811. Similarly the proposed algorithm produces specificity 0.9326 which is better than 0.8116 obtained from others method for the gene AF055080. The calculated result in this paper produces far better specificity and correlation coefficient when comparing with different method proposed by others. In case of an algorithm where the DNA sequence is converted into equivalent amino acid sequence to find exons regions showed an average specificity of 0.82 in case of HMR 195 dataset which is lower than result found in this work.[20] Although the sensitivity of the generated results found in the above experiments found to be 100% which is superior to our approach.

Table 5 shows the sensitivity, specificity and precision of the four nucleotide sequences in the proposed algorithm at thresholds T_h =0.2. The table predicts that, the proposed algorithm has the minimum nucleotides incorrectly identified as exons in all four gene sequences. The proposed method offers very high precision and accuracy compared to other algorithms.

Table 5. Gene Prediction Accuracy at Nucleotide Level.

Nucleotide Sequence	Sensitivity S_n	Specificity S_P	Matthews Correlation Coefficient MCC	Precision	Accuracy	F1 Score
AF013711	0.6705	0.9349	0.5635	0.5689	0.9040	0.6155
AF028233	0.8519	0.9811	0.8557	0.9132	0.9566	0.8814
AF058762	0.9174	0.7703	0.6690	0.7135	0.8268	0.8027
AF055080	0.7274	0.9326	0.6795	0.9052	0.8362	0.8066

Table 6. Gene Prediction Accuracy at Nucleotide Level.

Method	Average Sensitivity	Average Specificity	Average Accuracy
Hidden Markov Model	0.82	0.84	0.83
Antinotch Filter	0.81	0.82	0.82
Statistically Optimised Null Filter	0.89	0.86	0.84
S Transform	0.88	0.88	0.85
Multiresolution	0.89	0.88	0.86
DIT-FFT	0.80	0.90	0.88

The proposed technique conducted for both single and multiple exons from HMR 195 dataset. The results are shown in Table 6 and are compared with the methods found in literature.[21] The proposed method shows superior performance comparing to other established algorithms. The proposed method can be combined with autoregressive method to find very small exons.[22]

We have epitomized the prediction of the true exon locations in the Table 2. The results suggest that our experimental findings are very close to NCBI ranges.

CONCLUSION

We can clearly see that, in the case of Bit Reversal Operation in DIT Algorithm, all the input values are Integer numbers. We can find the output of the Corresponding Binary Bit of those Integer numbers by reversing it through the Butterfly Method. In the case of Mapping of the DNA Nucleotide Base, as the Integer Number Representation does not reflect directly to the DNA Sequence, the DNA nucleotide bases are converted into their numerical value by DIT algorithm. The resultant numerical sequences are passed through Kaiser Window based FIR filter in order to extract exonic regions.

The proposed algorithm is most efficient, requires less processing time, and provides high accuracy for available genomic data.

The entire process has been developed using MATLAB Software module which support bioinformatics tool box, Using tool box function, it can read genomic and proteomic data from standard file formats such as SAM, FASTA, CEL and CDF, as well as the NCBI Gene Expression Omnibus and Gene bank. It also provides statistical technique for detecting peaks. FDA tool is used in MATLAB to design all the digital filters. The peaks in the above figures give us the location of exons corresponding to base location, i.e., protein coding regions. The true Exon locations of different genes are covered by black dot shown in output power spectrums. Kaiser window is used to design FIR filter to give better output spectrum, because it has an adjusting parameter which increases the main lobe width and reduces the side lobe ratio.

REFERENCES

1. Ramachandran P, Antoniou A, Genomic digital signal processing, *Lectures, Genomic DSP04, Department of Electrical Engineering*, University of Victoria, BC, Canada.
2. Tsonis *et al.*, Periodicity in DNA coding sequences: Implications in gene evolution, *J Theor Biol*, 1991.
3. Roy A, Raychaudhury C, Nandy A, Novel techniques of graphical representation and analysis of DNA sequences: A review, *J Biosci* **23**:55, 1998.
4. Akhtar M, Epps J, Ambikairajah E, Signal processing in sequence analysis, Advances in eukaryotic gene prediction, *IEEE J Selected Topics Signal Process* **2**(3):310, 2008.
5. Anastassiou D, Genomic signal processing, *IEEE Signal Process Mag* **18**:8–20, 2001.
6. Pasquier CM *et al.*, A web server to locate periodicities in a sequence, *Bioinformatics* **14**(8):749, 1998.
7. Fickett JW, Finding genes by computer: The state of the art, *Trends Genetics* **12**(8):316, 1998.
8. Rogic S, Macworth AK, Evaluation of gene-finding programs on mammalian sequences, *Genome Res* 817, 2001.
9. Chakravarthy N, Spanias A, Lasemidis LD, Tsakalis K, Autoregressive modeling and feature analysis of DNA sequences, *EURASIP J Genomic Signal Process* **1**:13, 2004.
10. Cristea PD, "Genomic Signal representation and analysis," in *Proc. Society of Photo-Optical Instrumentation Engineers (SPIE) Conf*, Vol. 4623, pp. 77–84, January 2002.
11. Akhtar M, Epps J, Ambikairajah E, On DNA numerical representations for Period-3 based exon Prediction, in *Proc. IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, pp. 1–4, June 2007.
12. Nair AS, Pillai SS, A coding measure scheme employing electron-ion Interaction oseudo potential (EIIP), *Bioinformation* 1, pp. 197–202, October 2006.
13. Todd Holden, Subramaniam, R, Sullivan R, Cheng E, Sneider C, Tremberger G, Jr., Flamholz A, Leiberman DH, Cheung TD, "ATCG nucleotide fluctuation of Deinococus radiodurans radiation Genes, *Proc. Society of Photo-Optical Instrumentation Engineers (SPIE)*, Vol. 6694, pp. 669417-1–669147-10, August 2007.
14. Rosen GL, Signal processing for biologically-inspired gradient source localization and DNA sequence analysis, PhD thesis, Georgia Institute of Technology, August 2006.
15. Babu PR, Digital signal processing, 6th edn., Scitech Publications, April 2014.
16. Oppenheim AV, Schafer RW, Buck JR, *Discrete-Time Signal Processing*, 2nd edn., Upper Saddle River, NJ: Prentice Hall, 1989.

17. Chechetkin, Turygin VR, Yu A, Size-dependence of three-periodicity and long-range Correlations in DNA sequences, *Phys Letters A*, 1995.
18. Wassfy HM, Elnaby MMA, Salem ML, Mabrouk MS, Zidan A-AA, Advanced DNA mapping schemes for exon prediction using digital filters, *Am J Biomed Eng* **6**:25, 2016.
19. Mabrouk M, Advanced genomic signal processing methods in DNA mapping schemes for gene prediction using digital filters, *Am J Signal Process* **7**(1):12, 2017.
20. Meher JK, Dash GN, Meher PK, Raval MK, A reduced computational load protein coding predictor using equivalent amino acid sequence of DNA string with period-3 based time and frequency domain analysis, *Am J Mol Biol* **1**:79–86, 2011.
21. Inbamalar TM, Sivakumar R, Improved algorithm for analysis of DNA sequences using multi resolution transformation, *Sci World J* 2015.
22. Roy M, Barman S, Improved gene prediction by principal component analysis based autoregressive Yule-Walker method, *Gene* **575**(2):488, 2016.