Australian Journal of Basic and Applied Sciences, 5(3): 158-170, 2011
ISSN 1991-8178

DNA Splicing in Eukaryotes by an Enhanced and Robust Technique

[1]Muneer Ahmad, [2]Azween Abdullah and [3]Khalid Buragga

[1,3]Department of Computer Science, King Saud University P.O. Box 51178, Riyadh 11543,
Saudi Arabia
[2]Department of Computer Science, University Technology PETRONAS, Malaysia

Abstract: DNA splicing in genomes is a critical and important task to pay attention. Several approaches have been proposed for genic region prediction accuracy in Eukaryotes that mostly rely on digital signal processing techniques with incorporation of indicator sequence methods but an optimal solution is required that could provide the best minimization of 1/f noise and maximization of discrimination between exons and introns. In this paper, we have proposed an enhanced and robust technique for exonic prediction by introducing a novel UTP (University Technology PETRONAS) indicator sequence with denoising target DNA signal using discrete wavelet of third order. We have calculated the percentage improvement in accuracy as compared to existing approaches over traditional (commonly used in literature) datasets PHIX174 (accession number V01128 from location 101 containing 1284 base pairs), gene TUP1 of DNA sequence S.cerevisiae chromosome III (Accession number NC001135) from location 309 and gene F56F11.5 of C elegans (Accession number AF099922) from location 7021. The results obtained from proposed criterion reduced the computational complexity to 75% than Binary indicator sequence method and showed an improvement of 166% as compared to IIR antinoch filter, 133% to Binary indicator sequence method, 130% to filter 2 (multistage filter), 65% to EIIP indicator sequence method and 36% to Complex indicator sequence method. We tested our approach over four other datasets and found results very close to NCBI range.

Key words: DNA splicing, UTP indicator sequence, exonic region, wavelet, power spectral density, STFT

INTRODUCTION

Nucleotide base pairs are considered as the alphabets of life. The basic function of cell is the production of concerned protein. DNA is composed of chromosomes that contain genes. Exons and introns are regions that may or may not translate to RNA. The identification of protein coding regions is an important and challenging task (Roy *et al.*, 2009; Hazrina and Siti, 2008; Shuo and Yi-Sheng, 2008; Hota and Srivastava, 2008 and Akhtar *et al.,* 2007). Many statistical and signal processing approaches have been proposed claiming the improvement in the prediction accuracy for genic regions (Akhtar *et al.,* 2008 and Hang *et al.,* 2005). Optimal solution is strongly desired to predict the protein coding regions accurately as it helps in determination of cell function and drug design. Digital signal processing techniques (Shuo and Yi-Sheng, 2009; Kakumani *et al.,* 2008 and Akhtar *et al.,* 2008) involve the concept of DNA signal generation and analysis of signal using statistical approaches. Indicator sequence methods (Hota and Srivastava, 2008) are considered to be powerful methods with application of discrete time Fourier transforms (Grandhi and Vijay, 2008) short time Fourier transforms (Roy *et al.*, 2009).

(Roy *et al.*, 2009) has presented an approach concerned with genetic algorithm that is claimed to have applicability to variant datasets. The frequency distribution analysis outperformed the proposed method over performance metrics. This approach (Roy *et al.*, 2009) is flexible and provides some better glimpses over test datasets. (Shuo and Yi-Sheng, 2009) have used the concept of STFT short time Fourier transforms with support vector machine to reveal the exonic peaks in gene. The time frequency graphs attributed some rich flavors of peaks for clear discrimination between the two regions under discussions. (Kakumani *et al.*, 2008) described a statistical optimal null filter for maximization of signal to noise ratio. A rectangular window of length 351 was applied in this experiment. C Allegan (Chromosome III) was used as test dataset for revelation of regions. A local visualization criterion was used to distinguish intron from exons. (Akhtar *et al.*, 2008) presented a power modulus criterion depending DFT with windowing technique (window length 351). Matters of

Corresponding Author: Muneer Ahmad, Department of Computer Science, King Saud University P.O. Box 51178, Riyadh 11543, Saudi Arabia
E-mail: mmalik@kfu.edu.sa

computational complexity and prediction accuracy have been tried to be resolved using an improved PWSR procedure along with GENSCAN for learning and testing.

(Hota and Srivastava, 2008) proposed a complex indicator sequence method. Power spectral estimation of predicted regions was performed as against binary indicator sequence and IIIP method. Proposed complex values are namely 1, -1, j and -j for A, G, T and C respectively. This indicator sequence was tested over statistical digital filters and Fourier transforms with windowing data.

(Grandhi and Vijay, 2008) has proposed 2-simplex mapping method for identification of exon regions in DNA. It is a triangle based mapping in which four bases of nucleotides have been assigned to the vertex and centre of a triangle with values as distances from origin to respective vertex. Each base has been replaced with the corresponding vector value. The indicator sequence was passed through an IIR filter whose pass band was located at place where period 3 property remained conserve.

(Gupta *et al*., 2007) have shown a time series approach. Wavelet and feature extraction concepts were used. The variance information is utilized for construction of feature vector and finally a pattern recognition approach is applied for determination of bounds for exon and intron. (Mena-Chalco, 2008) have used Modified Gabor-Wavelet Transform for exonic analysis. The schematic analysis involves the numerical mapping of DNA sequence, application of MGWT to the binary indicator sequences, projection of spectral values to the position axis and threshold to coefficient values. Authors claimed the good performance of system with scale variations.

Wavelet transforms is another powerful tool for signal analysis and processing (Mena-Chalco, 2008 and Hang *et al*., 2005). Wavelet provides a better time scale multi-resolution analysis that is lacked in Fourier analysis of signal. These have been used for signal denoising in literature and this paper.

MATERIALS AND METHODS

Indicator Sequence:

Indicator sequence is used to transform a DNA nucleotide signal (consisting of alphabets A (adenine), G (guanine), Thymine (T) and Cytosine (C)) into some numeric equivalent for revealing the period three component of signal for exonic prediction. The equivalent values of these characters play an important role in discriminating the boundaries between genic and intergenic regions. The indicator sequences proposed in the literature are described below.

Binary Indicator Sequence:

The gene data is expressed in the form of nucleotides A, T, G, C. indicator sequence methods help us in translation of this data into numeric format that later can be used for spectral analysis of DNA signal. Binary indicator sequence method prices 1 and 0 for the existence or non existence of a specific nucleotide in strand, For example x[n] = [T T A G G T C C T] translates to [0 0 1 0 0 0 0 0 0], similarly other binary indicator sequences are formed and then DFT of individual sequences is calculated. Sum of all binary indicator sequences is 1,

$$uA[n] + uG[n] + uC[n] + uT[n] = 1$$
$$\text{for } n=0,\ 1,\ 2,\ldots.N\text{-}1.$$

Let UA[k], UG[k], UC[k] and UT[k] be DFT's of the binary sequences, then

$$U_x[k] = \sum_{n=0}^{N-1} u_x[n] e^{-j2\pi kn/N} \qquad k = 1,\ 2, \ldots, N$$

and Ux may be one of indicator sequences. After the calculation of DFT,

$$S\,[k] = \sum \left| \; U_x[k] \; \right|^2$$

We need to calculate the absolute value of frequency vector with exponent power 2. This transformation gives us the power spectral density or power spectra of the desired DNA signal. The power in the form of magnitude can be plotted against the frequency vector to identify the peaks of exonic regions.

Electron-Ion Interaction Potential (EIIP) with windowed DFT:

In this method, one indicator sequence is proposed as against four binary indicator sequences which computationally reduce the overhead by 75 %.

$$Y_{EIIP} = W_A X_A + W_T X_T + W_C X_C + W_G X_G$$

Where numerical values are

A = 0.1260
T = 0. 1335
G= 0. 0806
C = 0.1340

And the transform becomes

$$X_{EIIP}[k] = \sum_{n=0}^{N-1} x_{EIIP}[n] e^{-j2\pi kn/N}$$

$$k = 1,\ 2, ..., N$$

Where k is bound in sample space, 0 ≤ k ≤ N

2.1.3. Complex Indicator Sequence with windowed DFT

As a replacement of binary indicator sequences, complex indicator sequence uses one sequence of values [7] namely

X (A) = +1
X (T) = +j
X (G) = -1
X (C) = -j

And the corresponding transform becomes

$$X_C[k] = \sum_{n=0}^{N-1} x_c[n] e^{-j2\pi kn/N}$$

$$k = 1,\ 2, ..., N$$

Where value of k remains between the sample space bounds.

The method of Complex Indicator Sequence reduces the computational overhead by 75 % and provides more accurate prediction of genic regions.

UTP Indicator Sequence method

The authors have minutely examined the clusters of variant exonic datasets of different species with statistical functions to estimate the occurrences of nucleotide clusters. The inter-correlation in nucleotide clusters proposed a following optimal indicator sequence,

Adenine (A) = X (A) = 0.260
Thymine (T) = X (T) = 0.375
Guanine (G) = X (G) = 0.125
Cytosine (C) = X (C) = 0.370

The corresponding transform becomes

$$X_{UTP}[k] = \sum_{n=0}^{N-1} x_{UTP}[n] e^{-j2\pi kn/N}$$

$$k = 1,\ 2, ..., N$$

Digital filter methods:

Finite Impulse Response Filter (FIR):

The filters that carry a finite response to impulse signals are called FIR filters. The FIR filter of length k can be described as

$$y[n] = \sum_{k=0}^{K-1} a_k x_{[n-k]}$$

Y is the transformed data and x is the input data. The filter takes a summation over input vector multiplied by a constant factor. The output vector has the same length as input vector. K is called the order of this filter.

$$A(z) = \frac{Y(z)}{X(z)}$$

A (z) is a transfer function for this filter. It is obtained by dividing the output vector values by the input vector. We can also term this as

$A(z) = \sum_{k=0}^{K-1} a_k z^{-k} = a_0 + a_1 z^{-1} + ... + a_{(K-1)} z^{-(K-1)}$ Which shows a polynomial equation in z-transform and defines the

same FIR filter? These filters are widely used because of their stability.

Infinite Impulse response Filter (IIR):

This filter carries an infinite response to signal.

$$y[n] = -\sum_{k=1}^{N-1} a_k y[n-k] + \sum_{k=0}^{M-1} b_k x[n-k]$$

y represents a vector of length n that contains the transformed values for IIR filter. The filter used two kinds of coefficients, feed forward and feed backward represented by a_k and b_k.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^{M-1} b_k z^{-k}}{1 + \sum_{k=1}^{N-1} a_k z^{-k}}$$

H is the transform function over z-transform when output vector is divided by input vector. The main difference between the two filters is stability, band width and order of filter. IIR filter with its extension is widely used in DSP techniques for DNA signal analysis.

Discrete Wavelet Transforms:

Discrete Wavelet transforms provide the best time scale localization of DNA signal. We have used DWT for denoising our signals.

A Wavelet transform can be presented as

$$WT_f(a,b) = \langle f(t), \psi_{a,b}(t) \rangle = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} f(t)\ \psi^* \left(\frac{t-b}{a} \right)\ dt,$$

Where $\psi(t)$ is mother wavelet and b is shift parameter, the Discrete coefficients after choosing values of a

(initial) =2 and b (initial) = 1 can be written as

$$C_{j,k} = \int_{-\infty}^{+\infty} f(t) \psi^*{}_{j,k}(t) dt = \langle f, \psi_{j,k} \rangle$$

Proposed Method:

We have proposed a novel indicator sequence defined in section 2.1.4 along with discrete wavelet transforms of order 3. The Daubechies transforms (discrete wavelet transforms defined in 2.2.3) denoised the target DNA signal obtained from application of UTP indicator sequence. We used Kaiser Window described in section 5 for signal frames generation.

Absolute value of Frame = |Frame| = $Ax(f) = |Xl(f)|$

Where Xl (f) calculates the absolute value $Ax(f) = \sqrt{Xl^2 + iXl^2}$

Power of Frame = Absolute value of frame raised to 2 = $|\text{Frame}|^2$ = $Px(f) = |Xl(f)|^2$

Normalization of frequencies is done by $Px(f) = |Xl(f)|^2 \frac{1}{f_s L}$, where fs is the sampling frequency and L is the length of original signal. Any increment in normalization factor after the product of fs and L normally doesn't guarantee a good impact for power spectral measure rather it makes for rescaling the plot.

The discrimination measure "D" is defined as a ratio of the lowest peak value of exon in the exon set to the highest peak value of intron in intron set. This estimation determines the strength of distinguishing factors between genic and intergenic regions.

RESULTS AND DISCUSSIONS

Power spectral analysis over PHIX174 Dataset:

The proposed UTP indicator sequence reveals a clear difference in revealing peaks as compared to existing sequences. The prediction was made over genic dataset PHIX174 (accession number V01128 from location 101 containing 1284 base pairs) as a standard because the same data has been used by other researchers for prediction.

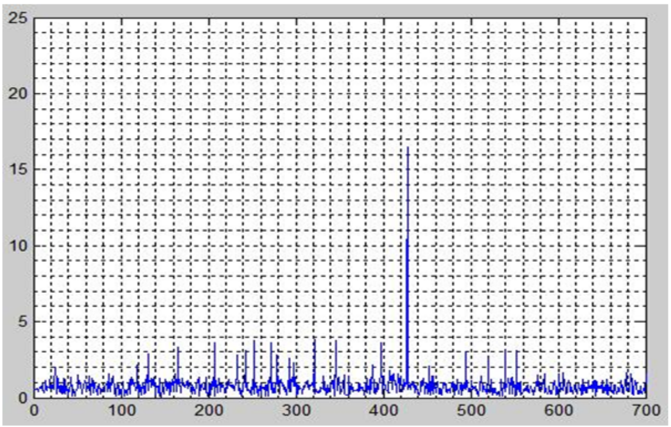

Fig. 1: PSD using Binary Indicator Sequence

Fig. 1 describes the spectral contents of data vector using binary indicator sequence. The period 3 property of DNA can be viewed at sample N/3 (slightly more than 400 points in graph). The spectral peak is prominent but we may notice the other peaks around this larger peak which represent the spectral leakage in the form of noise in DNA signal.

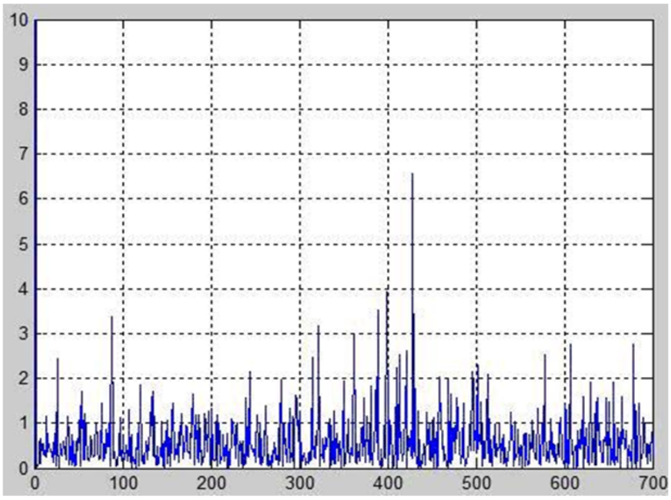

Fig. 2: PSD using EIIP indicator Sequence

The power spectral density graph for EIIP indicator sequence reveals the exonic peak at k=N/3 with DFT leakage. There are high peaks of noise along with central genic region.

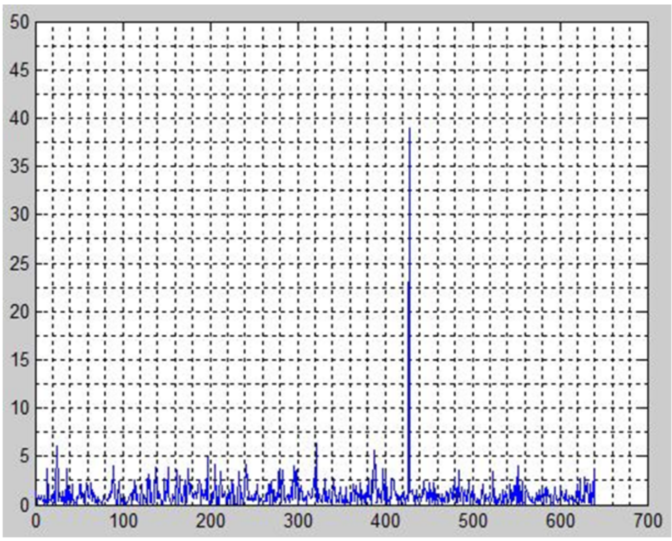

Fig. 3: PSD using Complex indicator Sequence

Fig. 3 clearly narrates the outperformance of complex indicator sequence for exonic prediction. The prediction peak is very high as compared to leakage factor.

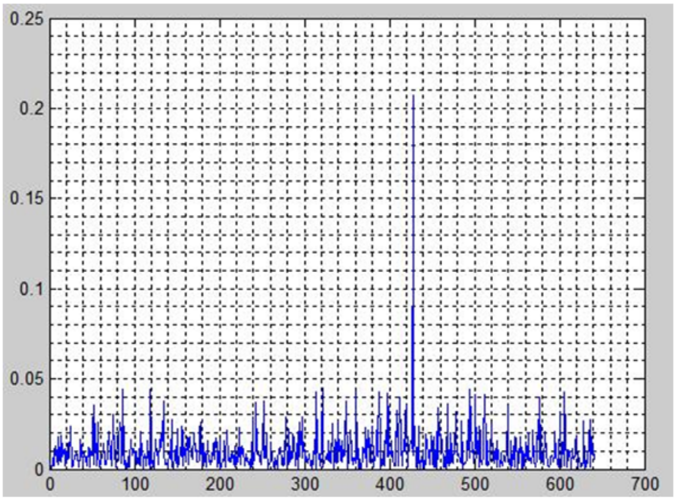

Fig. 4: PSD using UTP indicator Sequence

We can observe a larger peak at N/3 as compared to other indicator sequence in Fig. 4. There is larger peak for DNA period-3 property which is distinguishable from smaller peaks. There is strength of component between 400 to 500 in the form of major peak value and minor side lobes. This claims the outperformance of UTP indicator sequence.

Power spectral analysis over PUT1 Dataset:

We performed our second initial analysis on gene TUP1 (accession number NC001135.4) of S.cerevisiae chromosome III from location 309 with 2142 base pairs. This specific gene has also been used for spectral analysis in past literature.

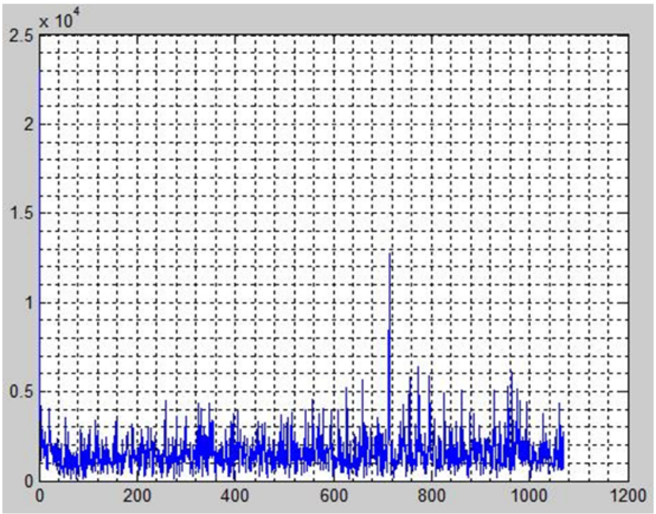

Fig. 5(a): Binary indicator sequence

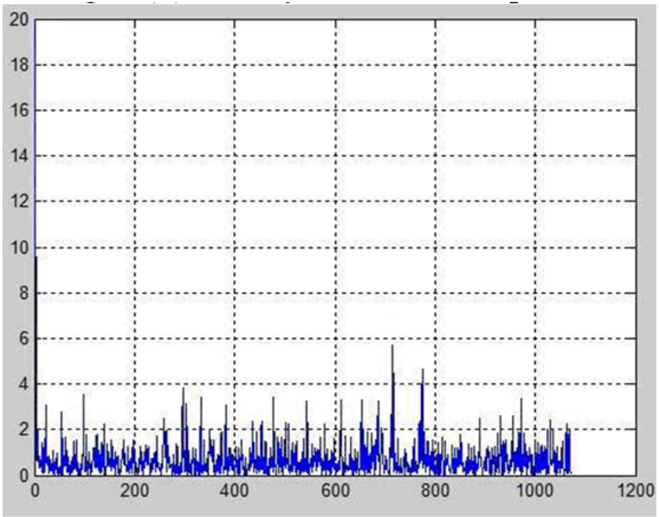

Fig. 5(b): EIIP indicator sequence

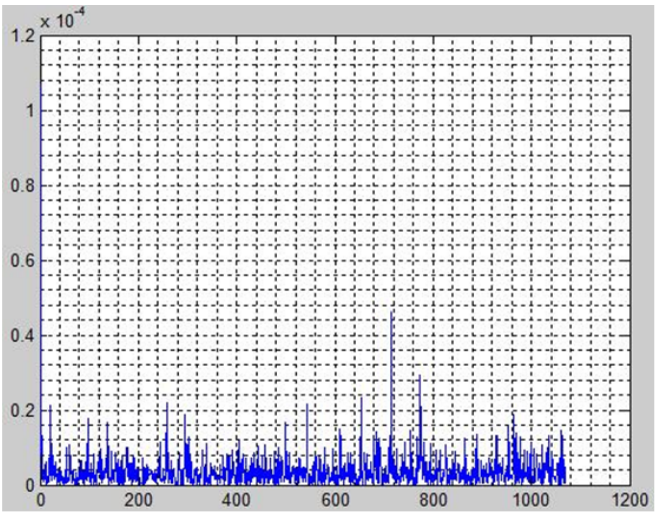

Fig. 5(c): Complex indicator sequence

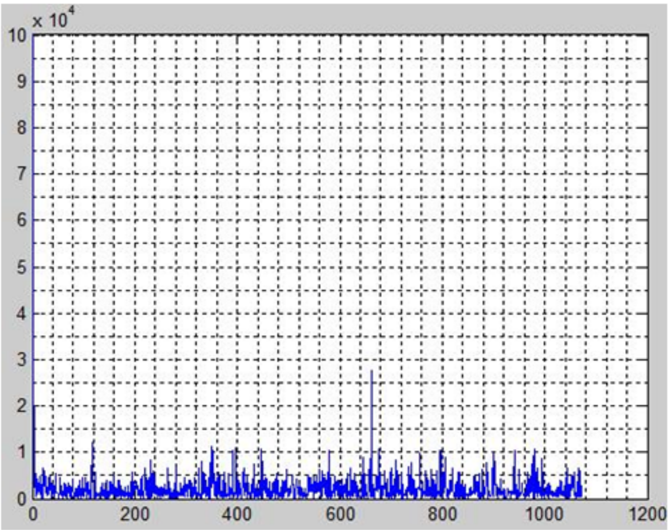

Fig. 5(d): UTP indicator sequence

Fig. 5(a, b, c, d) depict the exonic peaks for gene PUT1. X-axis and Y-axis represent the nucleotides and PSD respectively. We can estimate the relative differentiation on behalf of peak length and 1/f noise. UTP indicator sequence generates best optimal result for spectral analysis. We can view a strong long peak with minimized noise level. The peak is more distinguished from surrounding peaks and exhibits the period-3 property in optimal way as compared to EIIP spectral analysis.

Power spectral analysis over S.cerevisiae Dataset (STFT and Wavelet):

All sections of Fig. 6 depict various methods employed for spectral analysis of gene S.cerevisiae chromosome III (AF099922). X-axis and Y-axis represent the nucleotides and PSD respectively. We have calculated the Discrimination factor D for all methods. The Discrimination factor is the ratio of lowest peak in set of exonic peaks to the highest peak in set of intronic peaks. Greater the value of D, greater is the prediction accuracy and clear differentiation can be made between introns and exons. Numeric value of D is another picture of minimization of 1/f noise and maximization of genic peak values.

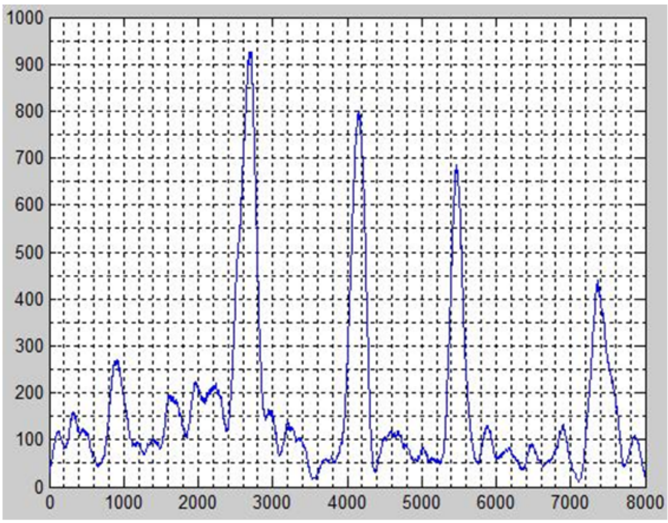

Fig. 6(a): Binary indicator sequence method

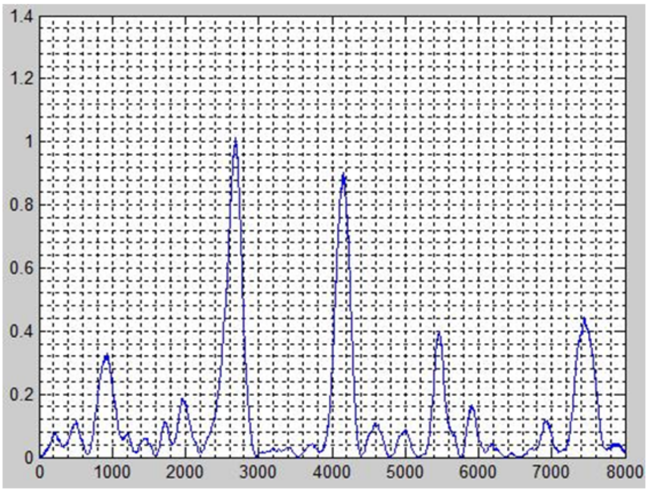

Fig. 6(b): EIIP indicator sequence method

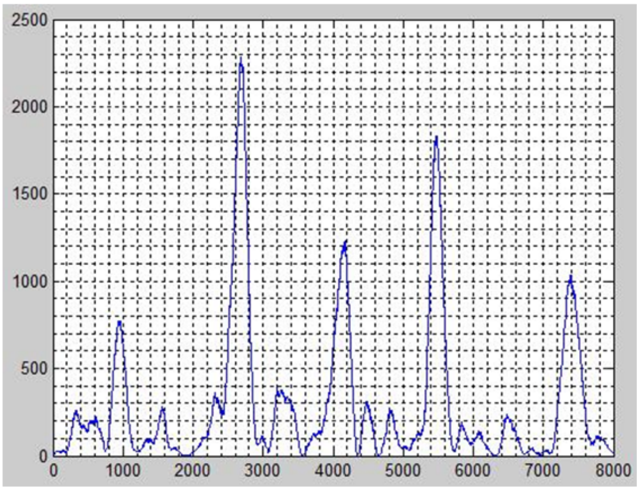

Fig. 6(c): Complex indicator sequence method

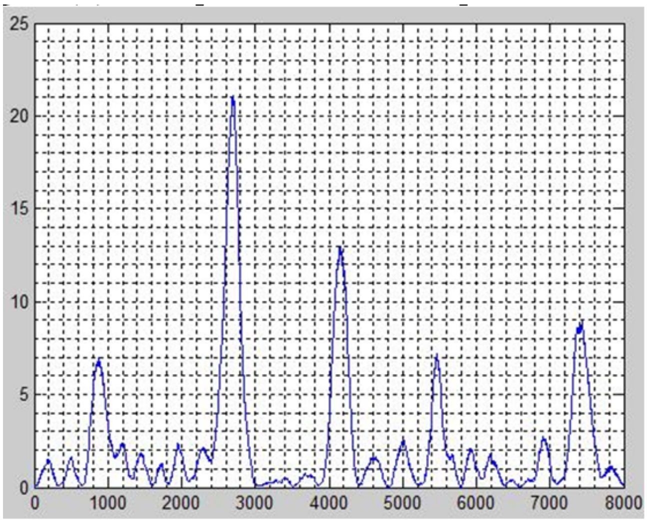

Fig. 6(d): UTP indicator sequence method

Table 1: Comparative analysis of various methods

Method Employed	Exons and Intron Boundaries	Discrimination Measure	Percentage Improvement in Prediction

Binary Indicator Sequence (STFT with Kaiser window of length 351)	E1 = 270
	E2 = 925
	E3 = 800
	E4 = 685
	E5 = 445
	Intron = 220
		1.2	133%

EIIP Indicator Sequence (STFT with Kaiser window of length 351)	E1 = 0.32
	E2 = 1
	E3 = 0.92
	E4 = 0.4
	E5 = .44
	Intron = 0.18
		1.7	65%

Complex Indicator Sequence (STFT with Kaiser window of length 351)	E1 = 775
	E2 = 2260
	E3 = 1230
	E4 = 1830
	E5 = 1030
	Intron = 375
		2.06	36%

Filter 1(IIR antinoch Filter)	E1 = 23.7
	E2 = 63
	E3 = 53.2
	E4 = 47.8
	E5 = 37.1
	Intron = 22.4
		1.05	166%

Filter 2 (Multistage Filter)	E1 = 34.80
	E2 = 113
	E3 = 88.20
	E4 = 77.8
	E5 = 48.3
	Intron = 28.25
		1.22	130%

Proposed Approach	E1 = 7
	E2 = 21
	E3 = 12
	E4 = 7
	E5 = 9
	Intron = 2.5
		2.5
		2.8
			More than 36 % improvement in prediction accuracy than the highest discrimination factor (2.06)

Table 1 describes the comparative analysis of various methods for power spectral analysis performed over S.cerevisiae chromosome III. We can see that Complex indicator sequence method generates D as 2.06 which was the highest discriminant value over all existing techniques. The UTP indicator sequence with wavelet transforms generates D as 2.8 which provide 36% more prediction accuracy.

We obtained a gain of 130% prediction accuracy than Filter 2, 166% than Filter 1, 65% than EIIP indicator sequence method and 133% than Binary indicator sequence method.

Discussions:

Section 3 describes the improvement in results for protein coding regions prediction using wavelet and employing a novel UTP indicator sequence. Satisfactory large exonic peaks with reduced 1/f noise are visible in all three tests and comparative analysis with existing solutions present the significant value of discrimination measure D.

We have used Matlab for all our dataset analysis and processing and Java for indicator sequence generation and translation.

Fig. 7 shows the values of Discrimination Factor D against various methods employed for spectral analysis of exonic prediction in DNA sequence. We can see a clear high peak of UTP indicator method as compared to other existing approaches. The outperformance of proposed approach is due to innovation of new UTP sequence with denoising characteristics of Discrete Wavelet transforms.

Discrete Fourier transforms provide the frequency domain analysis of DNA signal. We have used STFT (Short time Fourier transforms) for all our experiments in this paper with Kaiser Window of size 351 bp. The window is defined as

$$w(n)=\begin{cases} I_0\left(\beta\left(1-((n-\alpha)/\alpha)^2\right)^{\frac{1}{2}}\right)/I_0(\beta) & 0\le n\le M-1\\ 0 & otherwise\end{cases}$$

Kaiser window contains two very important parameters that are α and β, normally α is chosen over three sample spaces ranging from 0 to 21, 22 to 50 and 51 to entire length of signal. The default value for β is 0.5 and it determines the leakage factor and side lobes attenuations. We observed in testing different combination that optimal value of β may range from 10 to 60 against window size of 120 to 350 samples.

Third order discrete wavelet used for denoising DNA signals is describes as,

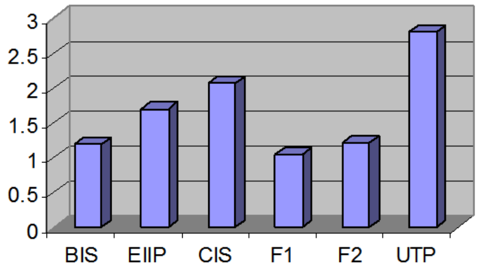

Fig. 7: Discrimination Factor against Methods

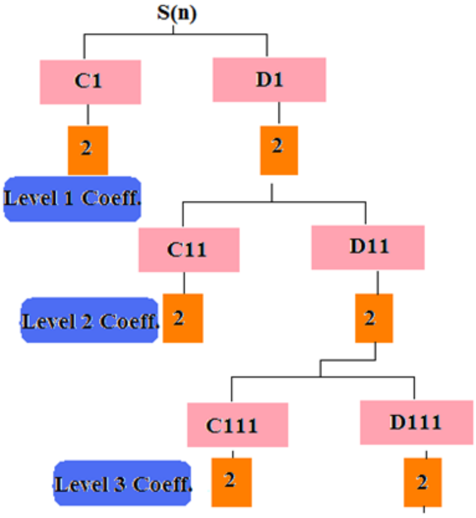

Fig. 8: Third level DWT Transform

In Fig. 8, coefficients C's represent the approximate coefficients and D's as detail coefficients. This DWT is down sampled by a factor of 2. We used discrete wavelet of order three for our DNA signal denoising and analysis. The good time resolution at higher frequencies and good scale resolution at lower frequencies is the best property and choice of wavelet to be used for signal analysis.

There is 36% improvement in exonic prediction as compared to Complex indicator sequence method (with latest higher value of discrimination measure).

We calculated the nucleotide range for exons and summarized as follows,

Table 2: Nucleotide range for exons

Method	E_1	E_2	E_3	E_4	E_5
Binary Method	650-1200	2400-3100	3800-4400	5300-5800	7100-7700
EIIP Method	700-1200	2200-2900	3900-4400	5200-5800	7200-7700
Complex Method	750-1100	2600-2900	3600-4400	5200-5700	7100-7600
Filter 1 (Antinoch)	650-1200	2450-3100	3800-4450	5300-5850	7100-7750
Filter 2 (Multistage)	700-1250	2200-2950	3900-4450	5200-5850	7200-7700
UTP Method	750-1050	2450-2900	3950-4380	5200-5600	7220-7680
NCBI Range	928-1039	2528-2857	4114-4377	5465-5644	7255-7605

Table 2 summarizes the nucleotide range of five exons. We can monitor clear differences as a comparative analysis of various approaches. Binary and EIIP methods glimpse more or less wide range difference than standard NCBI results. Complex method results are better than the first two approaches. Filter 1 and 2 behave accordingly while there is significant improvement in prediction of exons range with proposed approach.

Conclusion:

The period-3 property of DNA sequence can be exploited to reveal the genic regions that can best provide a way for DNA to RNA translation. Various methods and approaches have been proposed in literature for exonic prediction. We have proposed a novel robust methodology that places an improved solution to this problem. The UTP indicator sequence with discrete wavelet generates a discrimination factor D = 2.8 which is much higher than the previously proposed approaches. This method also reduces the computational complexity of 75% as compared to Binary indicator sequence with prediction accuracy more than 133%. The comparative analysis with other existing methods revealed 130% to 166% more prediction accuracy than Filter methods, 65% than EIIP method and 36% more accuracy than complex indicator sequence method. There is also significant improvement in prediction of exons concerning nucleotide range in DNA sequence.

ACKNOWLEDGEMENT

We are obliged for the kind assistance and support from Departments of Computer Sciences at King Faisal University Saudi Arabia and University Technology PETRONAS Malaysia.

Conflict of Interest:

We state that we have no competing and conflict of interest in the proposed research.

REFERENCES

Akhtar, M., E. Ambikairajah and J. Epps, 2008. Optimizing period-3 methods for eukaryotic gene prediction, IEEE International Conference on Acoustics, Speech and Signal Processing., pp: 621-624.

Akhtar, M., E. Ambikairajah and J. Epps, 2008. Advances in Eukaryotic Gene Prediction, IEEE Journal of Signal Processing in Sequence Analysis, Selected Topics in Signal Processing, 2(3): 310-321.

Akhtar, M., E. Ambikairajah and J. Epps, 2007. On DNA numerical representation of period-3 based exon prediction, IEEE International Workshop on Genomic Signal Processing and Statistics., pp: 1-4.

Changchuan Yin and S.-T. Yue Stephen, 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence, Journal of theoretical Biology., 247: 687- 694.

Datta, S. and A. Asif, 2005. A fast DFT based gene prediction algorithm for identification of protein coding regions, IEEE International Conference on Acoustics, Speech, and Signal Processing, 5: 653-656.

Gupta, R., A. Mittal, K. Singh, P. Bajpai and S. Prakash, 2007. A Time Series Approach for Identification of Exons and Introns, 10th International Conference on Information Technology, pp: 91-93.

Grandhi, D.G. and C. Vijay Kumar, 2008. 2-Simplex mapping for identifying the protein coding regions in DNA, IEEE region conference (TENCON) pp: 1-3.

Hota, M.K. and V.K. Srivastava, 2008. DSP technique for gene and exon prediction taking complex indicator sequence, IEEE Region 10 Conference (TENCON) pp: 1-6.

Hang Chen, 2005. Fei Gu and Feng Liu, Predicting protein secondary structure using continuous wavelet transform and Chou-Fasman method, 27th Annual International Conference of the Engineering in Medicine and Biology Society, pp: 2603-2606.

Hazrina Yusof Hamdani and Siti Rohkmah Mohd Shukri, 2008. Gene prediction system, International Symposium on Information Technology, 2: 1-7.

Kakumani, R., V. Devabhaktuni and M.O. Ahmad, 2008. Prediction of protein-coding regions in DNA sequences using a model-based approach, IEEE International Symposium on Circuits and Systems, pp: 1918-1921.

Mena-Chalco, J.P., H. Carrer, Y. Zana and R.M. Cesar, 2007. Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5(2): 198-207.

Roy, M., S. Biswas and S. Barman, 2009. Identification and Analysis of Coding and Noncoding Regions of a DNA Sequence by Positional Frequency Distribution of Nucleotides (PFDN) Algorithm, 4th International Conference on Computers and Devices for Communication, pp: 1-4.

Shuo Guo Zhu Yi-Sheng, 2009. Prediction of Protein Coding Regions by Support Vector Machine, International Symposium on Intelligent Ubiquitous Computing and Education, pp: 185-188.

Shuo Guo and Yi-Sheng Zhu, 2008. An integrative algorithm for predicting protein coding regions, IEEE Asia Pacific Conference on Circuits and Systems., pp: 438-441.