A Better Way for Exon Identification in DNA Splicing

Muneer Ahmad¹, Azween Abdullah², Khalid Buragga³

¹, ³ King Faisal University, Department of Computer Science, Saudi Arabia
² University Technology PATRONAS, Department of Computer Science, Malaysia

¹ mmalik@kfu.edu.sa
² azweenabdullah@petronas.com.my
³ kburagga@kfu.edu.sa

Abstract— The strands of chromosomes are supposed to be split into genic and intergenic regions. Gene identification is an optimization problem in Eukaryotic DNA splicing. An optimal solution is essential which may help in better protein translation.

We have proposed a novel approach for gene identification by employing discrete wavelet transforms for noise reduction in DNA sequences and introducing a new indicator sequence for better signal generation. Wavelet transforms greatly reduced the background noise and visible peaks of genic regions were found in power spectral estimation. The comparative analysis of proposed and existing approaches showed significant results for novel approach over prevailing solutions over dataset Yersinia pestis (ACCESSION: NC_004088, 4000 bp, four genes and exons from location 5000 bp to 8999 bp). Similar outperformance was observed over many other datasets.

Keywords: Genic regions, indicator sequence, background noise, discrete wavelet, introns, DNA splicing

I. INTRODUCTION

Eukaryotes are cells with nucleus [3]. DNA contains chromosomes that are blend of genic and intergenic regions. It is well-known that exonic regions exhibit period three property [1, 2, 4, 6, 10] that is not visible in intergenic regions. Due to this property of exons, protein coding region prediction can be made that can help in better identification of these regions for proper and exact translation of DNA to RNA [11].

DNA sequence is composed of four nucleotide bases called Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) [12, 14]. The order of these bases in a gene determines the genetic variation, living, habits and nature of species. The nucleotides join together to form a triplet base molecule called codon (helpful for protein structure and function prediction). There are 64 possible combinations of nucleotide triplets arranged in order to distinguish genic and intergenic regions [13]. Protein is composed of small scale units called amino acids. There are 20 types of amino acids and the sequence of these units determines the type and function of individual protein molecule.

Gene is considered to be sectioned into two regions called genic and intergenic regions [3, 15] as depicted in Fig. 1. The genic regions (Exons) are sequences of nucleotide that actually code for protein. The sequences are translated into messenger RNA and then to protein. Intergenic regions called as interrupted regions (Introns) don't code for protein. Gene is diffused in both Exons and Introns.

In EIIP method, one indicator sequence is proposed as against four binary indicator sequences with numeric values of nucleotides A = 0.1260, T = 0. 1335, G = 0. 0806 and C = 0.1340.

As a replacement of Binary indicator sequence, Complex indicator sequence [8] uses one sequence of values namely X (A) = +1, X (T) = +j, X (G) = -1 and X (C) = -j. The discrete wavelet transform involves the concepts of discritization of continuous transform and discrete coefficients can be calculated using the equation

\[ X_{a,b} = X_{j,k} = \sum_{n \in Z} x[n] g_{j,k}[n] \]

Where \( a = 2^j \), \( b = k2^j \), \( j \in N \), \( k \in Z \).

The process of performing convolution with scaled wavelet can be repeated so that a set of approximate and detail coefficients can be obtained for each iteration. The discrete transform after normalization can be defined as

\[ P_k = \frac{1 - \alpha}{1 + \alpha^2 - 2\alpha \cos(2\pi k / N)} \]

Where \( k \) can be termed as a frequency index and alpha as noise index.

Many different approaches have been proposed [1-15] that addressed this open problem but still a better optimized solution is essential.

Fig. 1 Exons and Introns in Gene

Gene


Kakumani et al., [6] proposed a method by employing statistically optimal null filter for maximization of SNR (signal to noise ratio) and aided with least square optimization criteria. Akhtar et al., [7] have shown an optimized solution using Discrete Fourier transforms by monitoring the effect of window lengths for signal processing based coding regions identification.

A. Nucleotide Translation
Nucleotide translation is very important phase for both peak maximization and noise suppression. We have introduced a new indicator sequence with nucleotide values as Adenine (A) = X

\[
(A) = 0.260, \quad \text{Thymine (T)} = X(T) = 0.375, \quad \text{Guanine (G)} = X(G) = 0.125 \quad \text{and Cytosine (C)} = X(C) = 0.370 \text{ after a comprehensive analysis.}
\]

The above expression is a Kaiser Window with \( \beta = 3.5 \) (minimizes the leakage factor and enhances the main lobe width).

B. 1/f suppression
It is mandatory to reduce the ratio of noise in the sequence for achieving the optimal results. We have used discrete wavelet transforms for this purpose.

C. Clustering
Multiplication of signal chunks with appropriate window is another essential task for reduction of spectral leakage. We have tested variant windows and found that Kaiser Window of length 351 bp is a good choice for minimizing spectral leakage.

\[
w(n) = \begin{cases} 
I_o \left( \beta \left(1 - ((n - \alpha)/\alpha)^2 \right)^{\frac{1}{2}} \right) & 0 \leq n \leq M - 1 \\
0 & \text{otherwise}
\end{cases}
\]

The above expression is a Kaiser Window with \( \beta = 3.5 \) (minimizes the leakage factor and enhances the main lobe width).

D. Power spectral analysis
Power spectral analysis is performed as calculation of magnitude and power of individual frames along with normalization of frequencies to bring the graph at smooth scale.

\[
|\text{Frame}| = Ax(f) = |XI(f)| (Magnitude of frame)
\]

Also called absolute value \( Ax(f) = \sqrt{|X|^2 + iX^2} \)

Power of Frame = Absolute value of frame raised to the power of factor 2 = \( |\text{Frame}|^2 = Px(f) = |XI(f)|^2 \)

The frequency components are then normalized by \( Px(f) = |XI(f)|^2 \frac{1}{f_sL} \) Where, \( f_s \) is the sampling frequency and \( L \) is the length of original signal.
It is worth mentioning that the any increment in normalization factor beyond $fs*L$ creates a need for rescaling the frequency vector rather than any further improvement in spectral analysis.

E. Ratio Calculation
Discrimination measure is a ratio of lowest exonic peak height (in a set of exons) to the heights peak value of intron (in a set of introns) in estimation of power spectral density of frames. The calculated discrimination measure for proposed and existing approaches is shown in table 1 under results and discussions.

F. Bounds Analysis
The genic regions bounds are estimated from power spectral density estimation plots. The results for exonic boundaries for specie Yersinia with 4000 bp have been summarized in table 2.

III. RESULTS AND DISCUSSIONS
The proposed approach showed significant results as compared with existing techniques. Several datasets were used for the compact analysis. As a test case, we have included the results performed over dataset Yersinia pestis (ACCESSION: NC_004088). The analysis revealed sharp and higher exonic peaks for coding regions identification. Significant improvement in prediction was obtained in calculation of discrimination measure for power spectral density estimation.

Fig. 2(a) narrates the PSD for Binary indicator sequence method. There is a considerable difference of bounds 100 bp almost for nucleotide ranges than NCBI. The third exon carries promising difference of 200 bp for the first initiation with a slight difference in terminating region. The EIIP method in Fig. 2(b) shows the same behavior for the first and second gene but there is a variation in nucleotide range for third exon. Third exon carries a major gap of almost 400 bp which is obviously another revealing flip for this method.

Fig. 2(c) describes Complex indicator sequence method. The first gene carries the major gap in nucleotide ranges than NCBI standard range (almost 150 bp in initiation). There is a breakup of range for the second exon between 500-900 and then 900 to 2500 bp. Third gene is more close to the standard range than Binary and EIIP methods.

The proposed approach in Fig. 2(d) describes the more promising close range of nucleotide to the standard range. We can see a clear difference of closeness of bounds compared with the prevailing methods. We observed the sharp peaks for exons. The graphs were found as continuous without any disconnection. There are two important aspects in gene identification i.e. nucleotide range estimation and discrimination measure estimation. The proposed methodology achieved the tasks significantly.
Table 1: Exon Boundaries in Different Approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>E_1</th>
<th>E_2</th>
<th>E_3</th>
<th>E_4</th>
<th>NCBI Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Method</td>
<td>200-450</td>
<td>450-2370</td>
<td>2400-2950</td>
<td>3000-4000</td>
<td></td>
</tr>
<tr>
<td>EIIP Method</td>
<td>200-450</td>
<td>450-2250</td>
<td>2251-2950</td>
<td>2950-4000</td>
<td></td>
</tr>
<tr>
<td>Complex Method</td>
<td>150-500</td>
<td>500-900</td>
<td>2500-2950</td>
<td>2950-4000</td>
<td></td>
</tr>
<tr>
<td>Filter 1 (Antinoch)</td>
<td>210-260</td>
<td>450-2300</td>
<td>2400-2900</td>
<td>3000-4000</td>
<td></td>
</tr>
<tr>
<td>Filter 2 (Multistage)</td>
<td>200-450</td>
<td>400-2200</td>
<td>2300-4950</td>
<td>5200-6950</td>
<td></td>
</tr>
<tr>
<td>Proposed Method</td>
<td>260-470</td>
<td>500-2400</td>
<td>2550-2950</td>
<td>3000-4000</td>
<td></td>
</tr>
<tr>
<td>NCBI Range</td>
<td>301-573</td>
<td>574-2442</td>
<td>2647-3066</td>
<td>3117-4000</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 describes the exonic boundaries calculated against different approaches. Complex method contains a disconnection in second exon and exon peaks are far from the standard range. The proposed approach bestows the closer range comparable with NCBI range.

Table 2: Comparative Analysis of Various Methods

<table>
<thead>
<tr>
<th>Method Employed</th>
<th>Exons and Intron Peaks in PSD analysis</th>
<th>Discrimination Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Indicator Sequence Method</td>
<td>E1 = 29, E2 = 280, E3 = 340, E4 = 425, Intron = 23</td>
<td>1.26</td>
</tr>
<tr>
<td>EIIP Indicator Sequence Method</td>
<td>E1 = 0.8, E2 = 0.84, E3 = 1.22, E4 = 0.035, Intron = 0.03</td>
<td>1.16</td>
</tr>
<tr>
<td>Complex Indicator Sequence Method</td>
<td>E1 = 250, E2 = 1250, E3 = 2000, E4 = 2400, Intron = 100</td>
<td>2.5</td>
</tr>
<tr>
<td>IIR antinoch Filter (Filter 1)</td>
<td>E1 = 21, E2 = 270, E3 = 335, E4 = 420, Intron = 15</td>
<td>1.4</td>
</tr>
<tr>
<td>Multistage (Filter 2)</td>
<td>E1 = 0.6, E2 = 0.75, E3 = 1.2, E4 = 0.030, Intron = 0.02</td>
<td>1.5</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td>E1 = 20, E2 = 16, E3 = 30, E4 = 1.5, Intron = 0.5</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2 presents the comparative analysis of proposed and existing approaches for discrimination measure. We can see a larger value for this factor for proposed approach. There is a gain of 100% than Filter 2, 114% than Filter 1, 20% than Complex indicator sequence method, 159% than EIIP method and 138% than Binary method. This significant improvement in results depicts the outperformance of proposed approach.

Fig. 3 shows the PSD of Complex method against the proposed approach. X and Y axis represent the nucleotide locations and power spectral density estimation respectively. We can monitor larger exon peaks of proposed approach against the promising Complex indicator method (with discrimination measure of 2.5). The discontinuity seen in Complex method for second exon was removed in proposed method. The third exon contains a larger peak in Complex indicator method while peaks for the fourth exons are almost similar. First exon carries high peak against a comparatively high peak of intron in Complex indicator method.

Fig. 4 depicts the discrimination measure factors for all approaches. It is worth mentioning that proposed method contains the most significant factor for distinguishing genic peaks as against the existing approaches. There is a gain of almost 20% than the second highest measure calculated in this prediction.
CONCLUSION

A novel method for gene identification is proposed in this paper. The method reduces the background noise in DNA signal by employing the discrete wavelet transforms along with mapping nucleotide with a new indicator sequence. The power spectral analysis and discrimination measures calculated over Yersinia pestis (ACCESSION: NC_004088) showed a gain of 20% to 158% in coding regions identification compared with existing techniques. The computational overhead is also reduced to 75% than traditional Binary indicator method. The significant improvement in prediction may help in understanding cell growth, function and protein transcription and drug design.

ACKNOWLEDGMENT

We are obliged for the kind assistance and support from College of Computer Sciences and Information Technology King Faisal University Saudi Arabia.

REFERENCES