Prediction of Coding Region in the DNA Sequences

Hitham Jleed and Sos Agaian
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, TX 78249

Abstract — In this paper, we propose a new algorithm based on Fourier spectral characteristics. This technique improves the coding prediction accuracy, boosting the protein coding region and suppressing the non-coding region within the DNA sequences. We also compare this algorithm using computer simulation with commonly used techniques showing that our technique for exon region prediction provides superior properties for the separation of exon and intron regions.

Keywords — DNA, Discrete Fourier transforms, indicator sequence, Receiver operating characteristic.

I. INTRODUCTION

Deoxyribonucleic Acid (DNA) is a double stranded structure made up of four nucleotides (bases) - Adenine, Cytosine, Guanine and Thymine, which are denoted with the letters A, C, G and T respectively [1]. The DNA consists of inter-genic regions. The genes are divided into small protein coding regions known as exons and non-coding spacers known introns, which is shown in Fig.1. The base sequences in the protein-coding regions of DNA molecules have a period-3 component. This periodicity can be attributed to the codon bias that refers to the unequal usage of codons in the coding regions (exons) and the triplet bias, and the periodicity is rarely existed within the introns or non-coding region in the gene [2]. With the exception of the extreme 5’ and 3’ ends of the first and last exons, which contain un-translated region (UTRs). Many researchers have used the 3-period prosperity to indicate the gene location by using bandpass digital filter or computing the discrete Fourier transform (DFT) that exhibits a peak at the frequency 2π/3 due to the periodicity known as spectral rotation (SR) measure proposed by Kotlar and Lavner [4]. The ability of Discrete Fourier transform to identify the accurate boundaries of the 3-period signal is limited by its requirement of chosen window size over which the spectrum is calculated.

II. BACKGROUND

DSP based techniques have been increasingly used to identify the protein coding region in the DNA sequences. That is because the symbols in protein coding regions exhibit a triplet periodicity (TP) and this periodicity is rarely existed in non-coding regions [2,3,8]. In all techniques, the DNA sequences are converted into numerical sequences suitable to digital signal processing such as the indicator sequences, where each set of nucleotide sequences is mapped into four indicator sequences.
\{ u_A \cdot u_C \cdot u_G \cdot u_T \} corresponding to four nucleotide characters A, C, G, and T. Each one of four types of nucleotides in the \( u[n] \) sequence whereby digit '1' in the presence of the character and digit '0' for its absence as defined below:

\[
U[n]=\begin{cases} 
1 & \text{if } u[n]=i \in \{A,C,G,T\} \\
0 & \text{otherwise}
\end{cases} \quad \ldots (1)
\]

Many digital signal processing methods have tried to analyze genomic data, considering the period-3 property as a good indicator for predicting the coding region locations. The example of such published work with proven results.

**DFT Methods [12]:** This method uses the binary indicator described above. As per the DFT definition in (2)

\[
X_i[k] = \sum_{n=0}^{N-1} U_i[n]e^{-j2\pi kn/N} \quad \ldots (2)
\]

\[0 \leq k \leq N-1 \quad \ldots \quad i \in \{A,C,G,T\}\]

The DFT coefficients are applied corresponding to \( N/3 \). Then the total power spectral is defined as:

\[
S[k]=|X_A[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 + |X_T[k]|^2 \quad \ldots (3)
\]

DFT is computed by sliding the window by one entry in the sequence.

**Filtering Methods [7, 13, 14]:** Digital filtering methods used for identification of exons make use of the period-3 behavior coding regions. The anti-notch filter is commonly used to detect the coding region because it has sharp gain at the frequency \( 2\pi/3 \). This filter has impulse response given by the definition in figure 2 [3].

![Image](image126x164to252x236)

Figure 2 impulse response of anti-notch filter

Given the indicator sequences \( U_i[n] \) as the input of the anti-notch filter, the output \( Y_i[k] \) of the filter has large energy in the passband filter, where \( i \in \{A,C,G,T\} \). The output should be comparatively large in the coding regions. With similar notation, defined as

\[
S[k]=|Y_A[k]|^2 + |Y_C[k]|^2 + |Y_G[k]|^2 + |Y_T[k]|^2 \quad \ldots (4)
\]

A plot of this function is used as a preliminary indicator of coding region.

### III. NEW METHOD

This method improves the use of Discrete Fourier transform method. Some articles have proved that the DNA can be spliced into coding and non-coding regions based on the Fourier spectrum of binary indicator sequences [5, 6, 7]. We used this indicator as well in this work. We first convert the DNA sequences into binary indicator sequence \( U_i[n] \), where \( i \in \{A,C,G,T\} \);

\[
\text{e.g.:}
\]

**Nucleotide Sequence:** A C T G T C A T C T A T C A

\[
\begin{align*}
U_A(0) &= 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 \\
U_C(0) &= 0 1 0 0 0 0 1 0 1 0 1 1 0 0 0 1 0 \\
U_G(0) &= 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 \\
U_T(0) &= 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 0
\end{align*}
\]

In order to implement the proposed algorithm and method, the DFT coefficients are applied for each indicator sequence. Because the DNA sequence is too long, a sliding window is used along the sequence by calculating the Fourier Transform of each subsequence. The window of size \( N \) is firstly applied to each binary indicator sequence starting B. The DFT of the windowed sequence of length \( N \) is then computed, which is taken to be multiplied of 3. Typically, the length of \( N \) is chosen to give the best accuracy of the prediction and \( B \) to be 7021. The window is then moved by sliding one or more points.

Based on Discrete Fourier transform, defined in (2), the sequence \( X_i[k] \) provides a measure of the frequency content at “frequency” \( k \) for each indicator sequence. To calculate the power spectrum, a new algorithm was developed, define

\[
S[n]=k_A|X_A[n]|^\rho + k_T|X_T[n]|^\rho + k_C|X_C[n]|^\rho + k_G|X_G[n]|^\rho \quad \ldots (5)
\]

\[
y[n]=\left[\frac{S[n]}{\max(S)}\right]^2 \times S[n] \quad \ldots (6)
\]

Where: \( \rho \), \( k_A \), \( k_T \), \( k_C \), and \( k_G \) are chosen to maximize the discriminatory capability between exons and introns. The values given by
\( S[n] \) defers between coding and non-coding regions. The peak values indicates the exon regions whereas the other values are considered introns. That is because the TP property of DNA sequence implies that the DFT coefficients are large in a single point at N/3 and reveals a peak for coding region and no such peak is observed for non-coding region. By calculating the DFT at that point in a sliding window, the output gives the final feature values for exon prediction. The block diagram in figure 3 illustrates the exon prediction steps.

**DNA Sequence**

- Nucleotide Representation
- Four Indicator Sequences
- Discrete Fourier Transform
- Combine 4 predictions
- Compute the proposed Algorithm
- Log | |
- END

**Figure 3** Exon Prediction Block diagram

### IV. DATABASES

The proposed algorithm described in this work applied on 8000 bp-long nucleotide sequence of the gene F56F11.5 of C elegans that can be retrieved directly from the Genbank database, maintained by the National Center for Biotechnology Information (NCBI) [9]. [GenBank access number AF0099922, positions 7021-15020]. This sequence has been used extensively as a test case for detection techniques. This gene holds five CDS of length located at positions 928-1135, 2528-2857, 4114-4377, 5465-5644, and 7255-7605.

### V. DISCUSSION

Using the simulated data, we have investigated different parameters in this work in order to obtain the best possible result. We first perform the investigation of the suitability of different window sizes for period-3 exon detection. Many existing signal processing-based gene prediction methods have used large windows [3,6,8,12]. Herein, we consider misclassification rate MR, the percentage of false prediction, as a measurement to choose the best window length. Across different window lengths we have tested, the optimum window length that gives less prediction error is 240. Figure 4 shows how window length affects the coding region prediction corresponding to MR.

![Figure 4](image)

**Figure 4** the effect of window lengths on misclassification rate. Window length of 240 gives less MR, so it leads to better coding region prediction

After utilizing the window’s length and to get better accuracy of prediction, we investigate the significance of the parameter ‘p’ in our algorithm as shown in table 1 and figure 5.

<table>
<thead>
<tr>
<th>P</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR %</td>
<td>6.58</td>
<td>6.73</td>
<td>6.20</td>
<td>6.04</td>
<td>6.16</td>
<td>6.70</td>
<td>7.07</td>
</tr>
</tbody>
</table>

**Table 1** The effect of p values

The purpose of this investigation is to enhance the predicted values in protein coding regions and suppress them in non-coding regions. From the chart, we can see that the value \( p=4 \) showed the least prediction error.

![Figure 5](image)

**Figure 5** the effect of power p on misclassification rate. The curve show that \( p=4 \) gives classification result
As can be seen from figure 6, the result gives a sharp boundary between coding and non-coding segments. We choose a cut-off PSD threshold of 0.2 (20%) . It can be seen from figure 7, this method gives a remarkable result. The five-regions of exons are classified

Sensitivity: probability that a test result will be positive when the exons are correctly predicted (true positive rate, expressed as a percentage).

\[ S_s = \frac{TP}{TP + FN} \]  \hspace{1cm} (7)

Specificity: probability that a test result will be negative when the introns are actually predicted (true negative rate, expressed as a percentage).

\[ S_p = \frac{TN}{FN + TN} \]  \hspace{1cm} (8)

Misclassification rate (MR): a proportion of all misclassified prediction, the sum of false negative and false positives, out of all predictions.

\[ MR = \frac{FN + FP}{TP + FN + FP + TN} \]  \hspace{1cm} (9)

Peak ratio (RP): a proportion of the peak value in the highest detected coding region, \( P_{HDCR} \), to the peak value of the lowest detected coding region, \( P_{LDCR} \), of a genomic sequence.

\[ RP = \frac{P_{HDCR}}{P_{LDCR}} \]  \hspace{1cm} (10)

The different fractions (TP, FP, TN, and FN) are represented in the following table II.
TABLE I
Reporting accuracy of the prediction [10]

<table>
<thead>
<tr>
<th>Test</th>
<th>Coding region</th>
<th>Non-coding region</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
<td>TP+FP</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
<td>FN+TN</td>
</tr>
<tr>
<td>Total</td>
<td>TP+FN</td>
<td>FP+TN</td>
<td></td>
</tr>
</tbody>
</table>

VII. RESULTS

We have proposed a spectral analysis method that gives good efficacy to discriminate the exons region in the DNA sequences. We have compared the results from the proposed method with two selected methods shown in figure 9 and 10 on the same databases, the gene F 56F11.5 of C elegans. When we use the 0.2 threshold, the result shows that our result has 6.04% prediction error, whereas the DFT method and Filtering method give 10.2% and 21.7% respectively. Furthermore, the proposed method gives relative improvement in both sensitivity and specificity.

Table III summarizes the comparative evaluation of the proposed gene prediction method with selected existing approaches, which have a feature value greater than threshold [Th=20%]

<table>
<thead>
<tr>
<th>Method</th>
<th>RP</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
<th>MR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>4.97</td>
<td>84.59</td>
<td>95.80</td>
<td>6.04</td>
</tr>
<tr>
<td>DFT Method</td>
<td>5.00</td>
<td>65.65</td>
<td>95.40</td>
<td>10.2</td>
</tr>
<tr>
<td>Filtering Method</td>
<td>4.99</td>
<td>27.51</td>
<td>84.37</td>
<td>21.7</td>
</tr>
</tbody>
</table>

Because the position of an arbitrary decision threshold is varied, we plot the percentage of the threshold as a function of the percentage of misclassification rate. Our result shows improvement in each value. As can be seen in figure 11, the new algorithm has lower error than others, especially in 10-20 % of the threshold.

VIII. CONCLUSION

In this paper, we have proposed a new method to identify the protein coding region in the DNA.
sequences based on the discrete Fourier transform. This technique improves the coding prediction accuracy by heightening the protein coding region and smothering the non-coding region within the DNA sequences. We have investigated the effects of window length based on misclassification rate. This reveals the optimum window length for our algorithm to be 240 bp. We have also reviewed two selected existing signal processing methods for gene and exon prediction in nucleotide sequences. We compared the results of our method with these methods based on the receiver operating characteristic curves. The experiment result shows that our technique for exon region prediction provides superior properties for the separation of exon and intron regions.

REFERENCES

[9]. “GenBank database,” NCBI