Correlation Between Different DNA Period-3 Signals: An Analytical Study for Exons Prediction

Ismail M. El-Badawy*, #, Ashraf M. Aziz†, Senior Member, IEEE, Zaid Omar* and M.B. Malarvili†

*Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
#Electronics and Communications Engineering Department, Arab Academy for Science and Technology, Egypt
†Electrical Engineering Department, Faculty of Engineering, Al-Baha University, Saudi Arabia
‡Faculty of Bioscience and Medical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
{ismail eb sadawy@aast.edu, amaziz64@ieee.org, zaidomar@utm.my, malarvili@biomedical.utm.my}

Abstract—The so-called DNA period-3 signals are significant indicators for exons locations in DNA sequences. However, the sole dependence on one period-3 signal to identify exons locations in a DNA sequence provides limited performance. It has been recently shown that exons prediction performance can be improved by employing multiple different period-3 signals. In this paper, we aim to justify the reason of this improvement by investigating the correlation between different DNA period-3 signals. For this purpose, we utilize Pearson’s correlation coefficient. Moreover, we examine another configuration of employing multiple period-3 signals for exons prediction that is not handled in previous work. MATLAB simulation is conducted on the HMR195 genomic dataset utilizing the ROC curve as a performance evaluation tool. The results reveal an average weak correlation of 12% between period-3 signals obtained by using different numeric mapping schemes. Consequently, employing those period-3 signals simultaneously provides better exons prediction performance as compared to employing either of them solely. On the other hand, the results show an average strong correlation of 80% between period-3 signals obtained by using different period-3 detection methods. Thus, the participation of those period-3 signals simultaneously does not provide a significant improvement in exons prediction performance as they carry similar information.

Index Terms—DNA, exons prediction, protein-coding regions, Pearson’s correlation, period-3 signals.

I. INTRODUCTION

Deoxyribonucleic (DNA) molecules are the hereditary information resources in living organisms [1]. A DNA strand consists of a linked sequence of nucleotides which are identified by their nitrogenous bases namely: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Each DNA sequence of nucleotides is mainly divided into genes separated by intergenic regions. In Eukaryotes, such as plants and animals, genes are subdivided into protein-coding regions (exons) and non-coding regions (introns). The nucleotides order in the exonic segments represents the code (i.e. set of instructions) for manufacturing proteins [1, 2]. This implies that exons hold important information which has a crucial role in determining proteins structures. Consequently, finding the locations of these information-bearing parts (exons) is considered a significant task and an important step towards genome annotation [2, 3]. The availability of large sequenced genomic databases makes it difficult to rely on manual exons localization methods, thereby giving an incentive for researchers to develop computerized algorithms for automatic prediction of exons locations in DNA sequences [4].

In literature, a number of computational methods for exons prediction in DNA sequences have been proposed [4-7]. These methods aim to detect the discriminating period-3 behavior that is prevalent for most of the exonic nucleotides arrangements [8]. The detected DNA period-3 signal (P3 signal) is subsequently used as an indicator for exons locations. In order to detect the P3 signal of a DNA sequence, two main steps are performed. Firstly, converting a DNA character sequence into a numeric one. Different DNA symbolic-to-numeric mapping schemes are previously reported in [9-11]. Secondly, tracking the three-base periodicity along the resulted DNA numeric sequence using a periodicity detection technique. Different digital signal processing (DSP) based techniques are previously proposed for this purpose. Those techniques are either frequency-domain such as discrete Fourier transform (DFT) and autoregressive (AR) method [4, 5], or time-domain such as time-domain periodogram (TDP) and average magnitude difference function (AMDF) [6].

To date, most of the existing approaches solely rely on one P3 signal in order to identify exons locations in a DNA sequence. This results in a limited exons prediction performance. Lately, we proposed in [12] a soft decisions fusion approach for exons prediction, in which multiple P3 signals are detected upon which exons locations are identified. The P3 signals were detected using different mapping schemes with the same period-3 detection method, and the results revealed a better performance as compared to that of the traditional approach. However, the reason of this improvement in exons prediction performance is not clearly justified. In addition, the feasibility of employing two P3 signals detected using the same mapping scheme with different period-3 detection methods is not examined. The objective of this study is to address the aforementioned issues. We conduct MATLAB simulation on the HMR195 benchmark genomic dataset [13]. Pearson’s correlation [14] is utilized to compute the correlation between different P3 signals. The results reveal
a weak correlation between P3 signals obtained by using different numeric mapping schemes, while those obtained by using different period-3 detection methods are strongly correlated. In addition, utilizing these strongly correlated P3 signals simultaneously shows no significant improvement in exons prediction performance.

The rest of the paper is organized as follows. Section II describes the steps of exons prediction based on single and multiple P3 signals. Section III discusses the Pearson’s correlation coefficient and its interpretation. Section IV presents results and discussion, and finally section V concludes the paper.

II. EXONS PREDICTION BASED ON SINGLE AND MULTIPLE PERIOD-3 SIGNALS

In order to predict exons locations in a DNA sequence based on its period-3 signal, two main steps should be performed. This section presents these steps as follows:

A. DNA Period-3 Signal Detection

A DNA period-3 signal is basically detected through two stages, as illustrated in Fig. 1. These stages are described in the following subsections.

1) Numeric Mapping: Analyzing a DNA sequence using signal processing tools is not possible until it is converted into numeric entities [3]. Thus, mapping the query DNA character sequence into numeric one is a primary mandatory step. DNA numeric mapping schemes are either multi-dimensional or one-dimensional [4, 5, 9-11].

A multi-dimensional scheme represents a DNA sequence by more than one numeric sequence. Voss is considered the most popular multi-dimensional mapping scheme [4, 5]. In Voss representation, four binary sequences: \( x_A(n) \), \( x_C(n) \), \( x_G(n) \) and \( x_T(n) \) are used to indicate the existence of the four nucleotides A, C, G and T, respectively, at position ‘n’ within a DNA sequence. For example, Voss representation of the DNA sequence ‘ATGCCGTAAGC’ is given by:

\[
\begin{align*}
x_A(n) &= 1 0 0 0 0 0 1 1 0 0 \\
x_C(n) &= 0 0 0 1 0 0 0 1 0 \\
x_G(n) &= 0 0 1 0 1 0 0 0 1 \\
x_T(n) &= 0 1 0 0 0 1 0 0 0 
\end{align*}
\]

A one-dimensional scheme, on the other hand, represents a DNA sequence by only one numeric sequence as follows [3]:

\[
x(n) = ax_A(n) + cx_C(n) + gx_G(n) + tx_T(n) \quad (1)
\]

Where the constants \( a \), \( c \), \( g \) and \( t \) are the numeric values assigned for the four nucleotides A, C, G and T, respectively. The choice of these four constants differs from one mapping scheme to another [9-11]. Although multi-dimensional representations provide better exons prediction accuracy, one-dimensional representations have the merit of reducing computational overhead in the subsequent period-3 detection stage [4, 5]. In this work, we consider the electron interaction pseudopotential (EIIP) [9] and complex indicator sequence (CIS) [10] numeric mapping schemes shown in Table I. The EIIP numeric sequence represents the four nucleotides in terms of the average energy of their valence electrons. The CIS represents the four nucleotides on the complex plane, projecting the purine bases (A and G) on the real axis (1 and -1, respectively) and the pyrimidine bases (C and T) on the imaginary axis (-j and j, respectively).

2) Period-3 Detection: The conversion of a DNA character sequence into a numeric one invites DSP-based methods to be employed for detecting the exonic period-3 behavior. These period-3 detection methods aim to extract the so-called DNA P3 signal which has the same length as the query DNA sequence. This implies that each nucleotide has a corresponding value in the P3 signal that indicates the relative strength of the three-base periodicity at this particular position. Hence, the peaks of a DNA P3 signal are expected to indicate exons locations in this DNA sequence. In this work, we consider the following period-3 detection methods.

a) Sliding Window Discrete Fourier Transform (DFT): The power spectrum of a finite windowed part of a DNA signal ‘\( x(m) \)’ is computed using DFT as follows [4, 5]:

\[
X[k] = \left| \sum_{m=0}^{M-1} x(m)w(m)e^{-j2\pi km/M} \right|^2, \quad 0 \leq k \leq M - 1 \quad (2)
\]

Where ‘\( w(m) \)’ is the window function of length ‘M’. Since the target is to extract the P3 signal, the 1/3 frequency component is considered the only significant component. Thus, calculating DFT at ‘\( k = M/3 \)’ is sufficient. Subsequently, sliding the window along the entire DNA sequence and calculating ‘\( X[M/3] \)’ for each window position provide the complete P3 signal. In our work, we use the simple rectangular window of length ‘M = 351’ since it is previously found to be the most suitable window length for capturing exons three-base periodicity [4, 5].

b) Time-domain Periodogram (TDP): Similar to sliding window DFT, the TDP also analyzes the query DNA sequence

<table>
<thead>
<tr>
<th>Numeric Mapping Scheme</th>
<th>Numeric value assigned for each nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIIP</td>
<td>0.1260 0.1340 0.0806 0.1335</td>
</tr>
<tr>
<td>CIS</td>
<td>1 -j -1 j</td>
</tr>
</tbody>
</table>
on a sliding window basis [4-6]. The DNA signal \( x(m) \) passes first through a second-order anti-notch filter with a center angular frequency of \( 2\pi/3 \) to suppress all spectral components except the 1/3 frequency. Subsequently, a window of length \( L \) is shifted along the filtered DNA signal \( y(m) \), where each \( L \)-points subsequence is rearranged in a 3-column matrix as shown below:

\[
\begin{bmatrix}
y(1) & y(2) & y(3) \\
y(4) & y(5) & y(6) \\
\vdots & \vdots & \vdots \\
y(L-2) & y(L-1) & y(L)
\end{bmatrix}
\]

Then the TDP vector is obtained by summing the values column-wise as follows [6]:

\[
\text{TDP}[c] = \sum_{n=1}^{L/3} y(3n - (3 - c)) , \quad c = 1, 2, 3
\]

Where, \( c \) indicates the column index. The maximum value in the TDP vector represents the P3 value corresponding to this window position. Shifting the window along the whole DNA sequence and repeating the aforementioned TDP computation steps give the complete P3 signal. In our work, we use \( L = 117 \) according to what is reported in literature [4-6].

Fig. 2 shows the normalized P3 signal for a sample gene downloaded from the genebank [15] using the mapping schemes and period-3 detection methods mentioned in this section. The actual exons locations are also illustrated on the figure.

**B. Exons Prediction Decision**

In order to predict exons locations, a decision is made for each nucleotide by comparing its corresponding P3 value to an arbitrary threshold [4]. Traditionally, exons prediction is performed using only one P3 signal by giving hard (binary) decisions according to whether the P3 values are above or below a threshold, as given below:

\[
D_n = \begin{cases} 
1 & \Delta_n \geq 0 \\
0 & \Delta_n < 0 
\end{cases}
\]

Where \( D_n \) is the prediction hard decision for a nucleotide at position \( 'n' \), and \( \Delta_n \) is the difference between the P3 value corresponding to this nucleotide and the decision threshold.

Lately, we proposed in [12] a soft (fuzzy) decisions fusion approach (shown in Fig.3) in which multiple different P3 signals are employed simultaneously to identify exons locations. A soft decision, \( d_n \), for a nucleotide at position \( 'n' \) is computed using sigmoid function as follows [12]:

\[
d_n = \frac{1}{1 + e^{-3\Delta_n}}
\]

Subsequently, those soft decisions \( d'_n \) and \( d''_n \) from upper and lower branches, respectively, are fed to a decisions fusion center (DFC) that makes a final hard decision as follows [12]:

\[
D_n = \begin{cases} 
1 & \frac{1}{2} (d'_n + d''_n) \geq 0.5 \\
0 & \frac{1}{2} (d'_n + d''_n) < 0.5 
\end{cases}
\]

Table II shows two different configurations for employing different P3 signals in the soft decisions fusion approach. The work in [12] investigated configuration I (P3EIIP−DFT and P3CIS−DFT), and the results revealed an improved performance over the traditional exons prediction approach that employs only one P3 signal. However, the reason of getting better performance in case of using P3 signals resulted from using different mapping schemes with the same period-3 detection method is not clearly justified. Moreover, the performance of the proposed approach in case of incorporating configuration II (P3EIIP−DFT and P3EIIP−TDP) is not examined. Thus, in the current study we investigate the correlation between P3 signals in both cases (configurations I and II) as well as the exons prediction performance in case of incorporating configuration II.
III. Correlation Between Different DNA Period-3 Signals

In this study, we employ the Pearson’s correlation coefficient [14] to measure the similarity between different DNA period-3 signals. This section presents the formula for calculating the correlation and discusses the interpretation of its value.

Consider two vectors, each of length \( N \), \( A = [a_1, a_2, \ldots, a_N] \) and \( B = [b_1, b_2, \ldots, b_N] \). The Pearson’s correlation coefficient ‘\( r \)’ (also called Pearson’s ‘\( r \)’) between them is the ratio of their covariance to the product of their standard deviations as follows [14]:

\[
    r_{AB} = \frac{\text{cov}(A, B)}{\text{std}(A) \times \text{std}(B)} = \frac{\sigma_{AB}}{\sigma_A \sigma_B} \tag{7}
\]

Where \( \sigma_{AB} \), \( \sigma_A \) and \( \sigma_B \) are calculated as shown below:

\[
    \sigma_{AB} = \frac{1}{N-1} \sum_{i=1}^{N} (a_i - \mu_A)(b_i - \mu_B) \tag{8}
\]

\[
    \sigma_A = \sqrt{\frac{\sum_{i=1}^{N} |a_i - \mu_A|^2}{N-1}}, \quad \sigma_B = \sqrt{\frac{\sum_{i=1}^{N} |b_i - \mu_B|^2}{N-1}} \tag{9}
\]

And the mean values \( \mu_A \) and \( \mu_B \) of \( A \) and \( B \), respectively, are given by:

\[
    \mu_A = \frac{1}{N} \sum_{i=1}^{N} a_i, \quad \mu_B = \frac{1}{N} \sum_{i=1}^{N} b_i \tag{10}
\]

Thus,

\[
    r_{AB} = \frac{\sum_{i=1}^{N} (a_i - \mu_A)(b_i - \mu_B)}{\sqrt{\sum_{i=1}^{N} |a_i - \mu_A|^2} \sqrt{\sum_{i=1}^{N} |b_i - \mu_B|^2}} \tag{11}
\]

Pearson’s ‘\( r \)’ is a statistical value, between -1 and 1, that measures the degree of linear association between two variables [14]. In other words, it evaluates the consistency between two groups of data. Low \( |r| \) values that are close to 0 reflect a weak correlation, while high \( |r| \) values that are close to 1 reflect a strong correlation. For example, consider the first two P3 signals (\( P3_{\text{CIS-DFT}} \) and \( P3_{\text{EIIP-DFT}} \)) shown in Fig. 2(a) and (b), respectively. The \( P3_{\text{CIS-DFT}} \) signal exhibits high values corresponding to the first exon’s location at which the \( P3_{\text{EIIP-DFT}} \) signal values are relatively lower. In addition, although the \( P3_{\text{CIS-DFT}} \) signal has low values corresponding to the second and third exons locations, the \( P3_{\text{EIIP-DFT}} \) signal has higher values at those locations. This implies that both P3 signals do not follow similar trends and consequently the correlation between them is \( r = 0.1577 \) indicating only 15.77% of similarity. On the other hand, consider the P3 signals (\( P3_{\text{EIIP-DFT}} \) and \( P3_{\text{EIIP-TDP}} \)) shown in Fig. 2(b) and (c), respectively. It can be seen that both P3 signals follow consistent trends. Consequently, the correlation between them is \( r = 0.7940 \) indicating 79.40% of similarity.

IV. Results and Discussion

The soft decisions fusion approach (shown in Fig. 3) is investigated, including both configurations I and II. For this purpose, MATLAB simulation is conducted on the HMR195 genomic dataset [13]. We utilized the receiver operating characteristics (ROC) curve [16] to assess the performance as compared to that of the traditional exons prediction that uses only one P3 signal.

The ROC curve is a plot that depicts the trade-off between false positive rate (FPR) and true positive rate (TPR) as the decision threshold is varied [16]. FPR and TPR are calculated using the components of the confusion matrix, illustrated in Table III, as follows:

\[
    \text{FPR} = \frac{FP}{FP + TN} \tag{12}
\]

\[
    \text{TPR} = \frac{TP}{TP + FN}
\]
FPR is the proportion of intronic nucleotides falsely predicted as positives, while TPR is the proportion of actual exonic nucleotides correct predicted as positives. Optimum prediction is the case at which ‘0’ FPR and ‘1’ TPR are achieved. Thus, the area under the ROC curve (AUC) [16] reflects exons prediction performance.

Fig. 4 illustrates the ROC curves and the corresponding TPRs at different FPRs that are listed in Table IV. For configuration I, the soft decisions fusion approach achieves higher TPRs at different FPRs as compared to the traditional approach that uses only one P3 signal. This can be also inferred from the AUC values shown in Table III. The participation of both P3\textsubscript{EIIP}−DFT and P3\textsubscript{CIS}−DFT signals in predicting exons locations provides an improvement of 24.5% and 13.9% over employing only P3\textsubscript{EIIP}−DFT or P3\textsubscript{CIS}−DFT, respectively, at 10% FPR. On contrary, no significant improvement is observed in case of configuration II. Employing both P3\textsubscript{EIIP}−DFT and P3\textsubscript{EIIP}−TDP signals provides exons prediction performance that is comparable to that provided in case of using P3\textsubscript{EIIP}−TDP only.

In order to justify the aforementioned results, 30 sample DNA sequences are selected from the HMR195 dataset to compute the Pearson’s $r$ value between their P3 signals pairwise as follows:

- P3\textsubscript{EIIP}−DFT and P3\textsubscript{CIS}−DFT (Configuration I)
- P3\textsubscript{EIIP}−DFT and P3\textsubscript{EIIP}−TDP (Configuration II)

Kernel density estimation [17] is employed, using MATLAB built-in function ‘ksdensity’, to plot the density functions (shown in Fig. 5) of the Pearson’s $r$ values. Fig. 5 also illustrates the correlation values between P3 signals for 5 sample DNA sequences. The correlation between P3\textsubscript{EIIP}−DFT and P3\textsubscript{CIS}−DFT has mean and standard deviation values of 0.12 and 0.18, respectively. This implies that using different mapping schemes with the same period-3 detection method provides weakly correlated P3 signals (12% ± 18% of similarity). In other words, those P3 signals carry complementary information about exons locations (See Figures 2(a) and (b)). On the other hand, the correlation between P3\textsubscript{EIIP}−DFT and P3\textsubscript{EIIP}−TDP has mean and standard deviation values of 0.80 and 0.06, respectively. This implies that using the same mapping scheme with different period-3 detection methods provides strongly correlated P3 signals (80% ± 6% of similarity). In other words, those P3 signals carry similar information about exons locations (See Figures 2(b) and (c)). These findings explain the reason of the successful and limited performances of configurations I and II, respectively.

### Table IV: ROC results summary for configurations I and II compared to the traditional exons prediction approach based on one period-3 signal.

<table>
<thead>
<tr>
<th>P3 Signals</th>
<th>Numeric Mapping Scheme</th>
<th>Period-3 Detection Method</th>
<th>AUC</th>
<th>True Positive Rate (%)</th>
<th>P3 Signals</th>
<th>Numeric Mapping Scheme</th>
<th>Period-3 Detection Method</th>
<th>AUC</th>
<th>True Positive Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>EIIP</td>
<td>DFT</td>
<td>0.7280</td>
<td>40.90</td>
<td>55.02</td>
<td>64.63</td>
<td>EIIP</td>
<td>DFT</td>
<td>0.7280</td>
</tr>
<tr>
<td>One</td>
<td>CIS</td>
<td>DFT</td>
<td>0.7398</td>
<td>44.71</td>
<td>57.06</td>
<td>65.53</td>
<td>EIIP</td>
<td>TDP</td>
<td>0.7470</td>
</tr>
<tr>
<td>Two</td>
<td>EIIP &amp; CIS</td>
<td>DFT</td>
<td>0.7862</td>
<td>50.93</td>
<td>63.75</td>
<td>72.47</td>
<td>EIIP</td>
<td>DFT &amp; TDP</td>
<td>0.7526</td>
</tr>
</tbody>
</table>

**Fig. 4:** ROC Curves (left) and corresponding TPRs at different FPRs (right) for: (a) Configuration I and (b) Configuration II, compared to the traditional approach that employ one P3 signal, applied on the HMR195 dataset.
V. CONCLUSION

This paper studied the correlation between different DNA P3 signals which are normally used for identifying exons locations in DNA sequences. According to the results, two cases are highlighted: 1) P3 signals obtained by using different numeric mapping schemes are weakly correlated (12% ± 18% of similarity), and 2) P3 signals obtained by using different period-3 detection methods are strongly correlated (80% ± 6% of similarity). In regard to the previously reported soft decisions fusion approach for exons prediction, this study investigates its performance employing multiple P3 signals of the foregoing two cases. The results show that for the first case, an average improvement of 19.2% in exons prediction accuracy at 10% FPR is achieved as the P3 signals are weakly correlated. For the second case incorporating both P3 signals provide complementary information about exons locations. On contrary, for the second case incorporating both P3 signals is not feasible as they carry similar information about exons locations. Findings of this study brings to light the significance of employing weakly correlated DNA P3 signals so as to locate exons in DNA sequences.

REFERENCES