International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 9, No.4 (July-2020)

http://dx.doi.org/10.12785/ijcds/090406

Statistical DNA Sequence Modeling and Exon Detection Using Non-Parametric Methods

Ahmed M. Dessouky [1], Fathi E. Abd El-Samie [2], Hesham Fathi [3] and Gerges M. Salama [3]

[1] Department of Information Systems, Al Alson Academy, Cairo, Egypt
[2] Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf, 32952, Egypt
[3] Department of Electrical Engineering, Electronics and Communications Engineering, Faculty of Engineering, Minia University, Minia, 61111, Egypt

Received 13 Mar. 2020, Revised 28 May 2020, Accepted 20 Jun. 2020, Published 1 Jul. 2020

Abstract: This paper presents a hybrid approach based on digital bandpass filtering with non-parametric estimation techniques for the analysis of deoxyribonucleic acid (DNA) sequences. These spectral estimation techniques improve the analysis of DNA sequences and enable the extraction of some desirable information about them. The electron-ion interaction pseudopotential (EIIP) numerical representation method is used to convert a DNA sequence to numerical values through a mapping function. Also, mathematical modelling is used to create closed formulas for the represented DNA data sequences with different studied methods. The importance of this process is that the mathematical models can be used for any further processing or identification when applied to DNA sequences. The metrics used for performance evaluation are root mean square error (RMSE) and correlation coefficient (R) metrics. Also, the objective of this paper is investigating and predicting the location of the coding region (exon) in DNA sequences using the proposed approach. The results of gene prediction from DNA sequences for the original and modelled DNA sequences coincide and ensure the success of the proposed sum-of-sinusoids method for modelling of DNA sequences.

Keywords: DNA representation; mathematical modelling; RMSE; Correlation Coefficient; Non-Parametric Spectral Estimation Techniques.

1. INTRODUCTION

Bioinformatics is the science of how information is generated, transmitted, received, and interpreted in biological systems. It comprises the application of information technology in the field of biology [1-3].

Genomic information is encoded in the form of DNA inside the nuclei of cells. A DNA molecule is a long linear polymeric chain, composed of four types of sub-units. Each sub-unit is called a base. The four bases in DNA are adenine (A), thymine (T), guanine (G), and cytosine (C). DNA is a pair of strands. Bases pair up across the two strands. A always pairs with T, and G always pairs with C. Hence, the two strands are complementary [4-6].

Genomic information nature is digital, and it is represented with the shape of sequences in which each element can be one out of a finite number of entities. DNA and proteins have been described by character strings in which each character is a letter of an alphabet. In the DNA case, the alphabet consists of 4 letters, while in the case of proteins, the alphabet size is 20 [4-6]. DNA sequence representation is a vital topic in several fields of bioinformatics. DNA sequences generally comprise four different symbols with different periodicities that convey very vital information in fields such as gene prediction.

Before applying different computational methods, it is necessary to convert the DNA sequences (A, T, C, and G) into numeric sequences. The method called electron-ion interaction pseudopotential (EIIP) is used to convert the DNA sequences into numeric sequences [5, 6]. It is possible to create such mathematical models with polynomial, exponential, Gaussian, and sum-of-sinusoids closed formulas. Accuracy is an essential factor to be maximized through any modelling process.

This paper demonstrates the analysis of DNA sequences to predict the protein-coding regions called exons. The exact locations and positions of exons are determined. Some spectral analysis techniques are developed to serve these purposes. Protein coding regions

E-mail: ahmed_moawad7@yahoo.com, fathi_sayed@yahoo.com, hfah66@yahoo.com, gerges.salama@mu.edu.eg

and non-coding regions from DNA sequences are examined and predicted using a proposed hybrid approach based on a digital bandpass filtering with non-parametric spectral estimation techniques. Furthermore, this paper presents different modelling methods, including polynomial, exponential, Gaussian, and sum-of-sinusoids models. It is possible to create such models as closed formulas. Accuracy is studied based on statistical analysis with RMSE and correlation coefficient (R) metrics.

The rest of the paper is organized as follows. Section 2 represents the related work. Section 3 describes the DNA dataset. Section 4 gives the DNA numerical representation with the EIIP method. Section 5 presents the proposed solution methodology. Section 6 illustrates the proposed mathematical modelling method. Section 7 presents the performance evaluation metrics. Section 8 introduces non-parametric spectral estimation techniques. Section 9 presents the prediction of the exon region using the proposed hybrid approach. Section 10 introduces the result discussion and a comparison study. Finally, section 11 gives the concluding remarks.

2. RELATED WORK

Digital Signal Processing (DSP) techniques have been exploited in the analysis of DNA sequences. Some attempts have been reported in the literature for the application of spectral analysis for exon detection. In [3, 7-9], different Fourier-based methods have been used for the analysis of DNA sequences. These methods depend on the direct application of the Fourier transform on the DNA sequences and the estimation of the peaks in the spectral domain to detect exons. Some other attempts adopt the digital filtering for pre-processing of DNA sequences for noise removal [6].

It has been described in the literature that exons give specific peaks in the spectra that can lead to easy detection of them [10]. The detectability of the exons depends on the used efficiency and resolution of the spectral estimation method [10, 11].

Besides, in [12], the similarity/dissimilarity analysis of DNA sequences is performed using a 3D dynamic representation. The usefulness of the measurement of the similarity/dissimilarity ensures that it may be taken as a convenient mathematical tool in computational biology. In [13], different aspects of similarity, such as the asymmetry of the gene structure, have been studied either using new similarity measures related to four-component spectral representations of the DNA sequences or using alignment techniques with adjustments. In [14], the concepts of graphical bioinformatics have been introduced. These concepts emphasize the distinction between the branch of bioinformatics concerned with comparative studies of bio-sequences and the branch of bioinformatics that depends on arrangements with graphical representations of DNA and proteins.

The authors of [15, 16] considered a diversity of non-parametric spectral estimation techniques and compared between them for the application of exon detection for actual DNA sequences.

This paper considers the non-parametric spectral estimation techniques which include the periodogram, average periodogram (Bartlett), modified average periodogram (Welch), and Blackman and Tukey methods [17-21] in the proposed hybrid approach for exon prediction. The sensitivity of the exon detection process to the used spectral estimation technique is studied. Moreover, bandpass filtering is considered in conjunction with spectral estimation to enhance the detectability of exons.

The objective of these spectral estimation techniques is to investigate the locations of exons in DNA sequences for gene prediction from DNA sequences for both actual and optimum mathematical modelling. A comparison study is presented in this paper between the suggested spectral estimation techniques from the exon prediction perspective.

3. DNA DATASET DESCRIPTION

There are different types of dataset for DNA sequences, such as fasta, text, afa, ann, ftb, msf…... In this paper, we are concerned with the type of datasets called fasta [22]. It consists of a header and the DNA sequence. MATLAB using (fastaread) reads the subsequent file.

[>gi|1293613|gb|U49845.1|SCU49845:1-5028
Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds

GATCCTCCATATACAACGGTATCTCCACCTCAGG
TTTAGATCTCAACAACGGAACCATTGCCGACAT
GAGACAGTTAGGTATCGTCGAGAGTTACAAGCT
AAAACGAGCAGTAGTCAGCTCTGCATCTGAAGC
CGCTGAAGTTCTACTAAGGGTGGATAACATCAT
CCGTGCAAGACCAAGAACCGCCAATAGACAACA
TATGTAACATATTTAGGATATACCTCGAAAATAA
TAAACCGCCACACTGTCATTATTATAATTAGAAA
CAGAACGCAAAAATTATCCACTATATAATTCAA
AGACGCGAAAAAAAAAGAACAACGCGTCATAG
AACTTTTGGCAATTCGCG…..]

4. DNA NUMERICAL REPRESENTATION (EIIP) METHOD

In DNA numerical representation, each nucleotide of the DNA sequence is converted to a numerical value through a mapping function. Digital signal processing techniques can be applied to the numerically converted sequences to extract some desirable features and information from the DNA sequences.

The EIIP method is a coding method based on replacing the four binary indicator sequences by just one sequence. It is called the "EIIP indicator sequence". The EIIP values of amino acids are used to exchange the corresponding amino acids in protein sequences. In the current work, the EIIP values of nucleotides have been used instead of the values of amino acids. The EIIP values of nucleotides are defined as, A= 0.1260, G=0.0806, C=0.1340 and T=0.1335 [4-6, 23-26].

The EIIP representation method can improve the discrimination capability of gene finding techniques. Also, this method reduces the computational overhead by 75%. [23-26].

5. THE PROPOSED APPROACH

Figure 1 illustrates the block diagram of the proposed hybrid approach to estimate the PSD of the DNA sequences and to identify the coding regions.

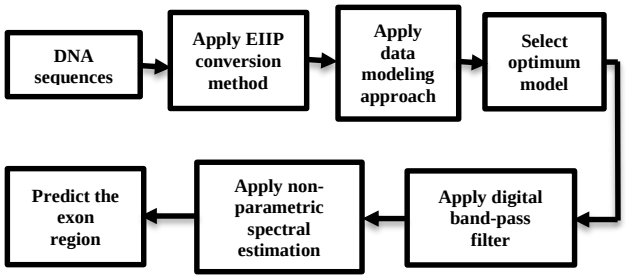

Figure.1 Proposed approach block diagram

The main steps of the proposed approach are summarized as follows:

1- Reading the DNA sequence using MATLAB function fastaread.
2- Encoding the DNA sequence into a numeric sequence using the EIIP method.
3- Applying the data modelling approach to represent the DNA sequence with a mathematical formula.
4- Utilization of a digital bandpass filter for extraction of the coding region and noise suppression.
5- The output of the digital filter is passed through the non-parametric spectral estimation methods for estimating the power spectra of both the actual data and the mathematical model.
6- On the estimated spectra, determine the peak to locate the exon region in the DNA sequence.

Note: for all illustrated figures given in the next section, all real (actual) data is indicated by a solid line, while a thin line shows the obtained theoretical result.

6. THE PROPOSED MATHEMATICAL MODELING APPROACH

This section presents the proposed DNA mathematical modelling approach. This approach is used to perform the transformation of DNA sequences to mathematical functions. The selected optimum mathematical model is based on minimizing the error and maximizing the correlation coefficient between the actual data and the obtained mathematical form. The data modelling approach is used to perform the representation of DNA sequences as different mathematical functions. In this section, the simulation model is presented for different cases using the simulation capabilities of the MATLAB software package.

A. DNA Polynomial Modeling

The mathematical representation using a polynomial function with order 9 is represented as:

$$F(x) = 3.279\text{e-}30\, x^9 - 7.676\text{e-}26\, x^8 + 7.599\text{e-}22 x^7 - 4.131\text{e-}18\, x^6 + 1.336\text{e-}14\, x^5 - 2.583\text{e-}11\, x^4 + 2.817\text{e-}08\, x^3 - 1.472\text{e-}05\, x^2 + 0.002403\, x + 0.3125 \quad (1)$$

B. DNA Exponential Modeling

The mathematical equation using the exponential function is given as:

$$F(x) = -2871\, e^{-0.0007404\, x} + 2872\, e^{-0.0007402\, x} \quad (2)$$

C. DNA Gaussian Modeling

The mathematical form using the general Gaussian model with 8 coefficients is represented as:

$$F(x) = 0.8818 \exp(-((x- 1583)/ 197.3)^2) +0.6943 \exp(-((x- 2361)/ 187.3)^2) +0.1777 \exp(-((x- 1788)/ 117.9)^2) +0.3109 \exp(-((x- 1111)/ 156.8)^2) +0.1641 \exp(-((x- 2154)/ 97.3)^2) +0.1266 \exp(-((x- 2638)/ 2838)^2) +0.4249 \exp(-((x- 153.1)/ 179.7)^2) -0.03567 \exp(-((x- 1340)/ 81.88)^2) \quad (3)$$

Figure 2-a, 2-b, and 2-c present the results of the actual DNA sequence using the given database and the obtained mathematical representations using polynomial, exponential and Gaussian models with a number of terms = 8, respectively.

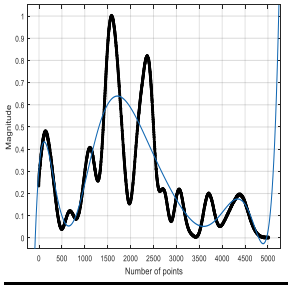

Figure. 2-a polynomial function results

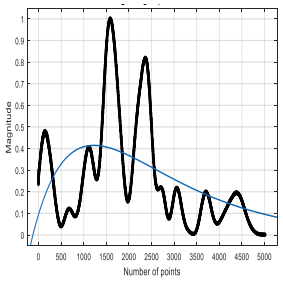

Figure.2-b exponential function results

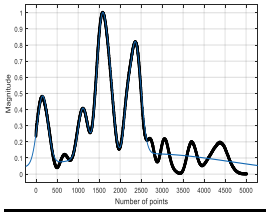

Figure. 2-c Gaussian function results

D. DNA Sum of Sinusoids (SoS) Modeling

This section gives a brief discussion of the data modelling using the SoS model. Figures (3-a) to (3-d) present the representations of the actual DNA sequence and the representation using SoS model with a number of terms equal to 2, 4, 6, and 7, respectively. Also, the mathematical equation for each model is presented.

- *Using No. of Sinusoidal Terms = 2*

The mathematical equation for this model is represented as:

$F(x) = 0.4358 \sin(0.000548x + 0.7682) + 0.2014 \sin(0.002473x + 3.012)$ (4)

- *Using No. of Sinusoidal Terms = 4*

The mathematical equation for this model is represented as:

$F(x) = 0.4319 \sin(0.000585x + 0.6188) + 0.1954 \sin(0.002427x + 3.194) + 0.1457 \sin(0.008924x - 0.3117) + 0.09679 \sin(0.005505x + 0.1261)$ (5)

- *Using No. of Sinusoidal Terms = 6*

The mathematical equation for this model is represented as:

$F(x) = 0.4488 \sin(0.0006096x + 0.5916) + 0.2126 \sin(0.002671x + 2.623) + 0.1465 \sin(0.00921x - 1.034) + 0.07928 \sin(0.003845x + 2.672) + 0.1219 \sin(0.007594x + 2.176) + 0.1094 \sin(0.005662x - 0.03721)$ (6)

- *Using No. of Sinusoidal Terms = 7*

The mathematical equation for this model is represented as:

$F(x) = 0.4415 \sin(0.0006193x + 0.5318) + 0.1868 \sin(0.002505x + 3.07) + 0.1429 \sin(0.00928x - 1.089) + 0.05778 \sin(0.00443x + 1.207) + 0.126 \sin(0.007762x + 1.779) + 0.1196 \sin(0.005715x - 0.03376) + 0.05054 \sin(0.01135x - 5.128)$ (7)

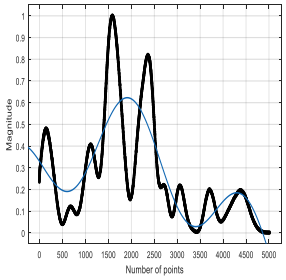

Figure. 3-a SoS model with 2 terms results

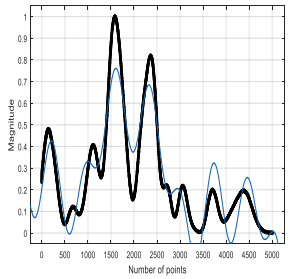

Figure. 3-b SoS model with 4 terms results

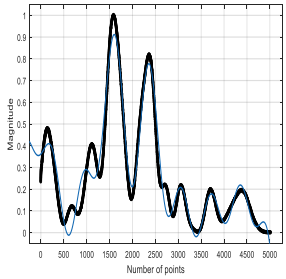

Figure. 3-c SoS model with 6 terms results

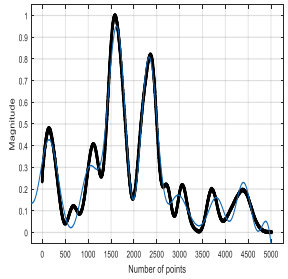

Figure. 3-d SoS model with 7 terms results

Figure 4 illustrates the signal representation for 5000 samples of DNA sequences using the EIIP method for actual data, Gaussian model, and SoS Model.

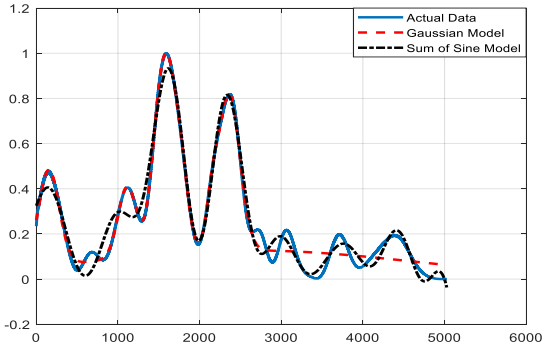

Figure. 4 signal representation results

7. PERFORMANCE EVALUATION METRICS

This section defines the metrics used for measuring the performance of the proposed mathematical modelling approach.

Root Mean Square Error (RMSE)

The mean square error is one of the most important parameters used to examine and measure the performance of the mathematical data model. It is defined as the average error between the actual data and the mathematical model.

If X is the actual data, and $\hat{X}$ is the mathematical model, the MSE will be defined as:

$$MSE = E[e(n)^2] = E[(X - \hat{X})^2] = E[(\hat{X} - X)^2] \quad (8)$$

If the value of the *MSE* tends to zero, there is a complete coincidence and matching between the actual data and the mathematical model. If the value of the *MSE* increases more than zero, the error between the actual data and the mathematical model also increases. Root Mean Square Error *(RMSE)* is the square root of the *MSE*.

Correlation Coefficient (R)

The correlation coefficient *R* is used to analyze how difference in one variable can be explained by a difference in a second variable. The correlation coefficient *(R)* measures the similarity between the actual data DNA sequence and the mathematical model. The variation range of *R* is 0 to 1 (i.e. 0% to 100%).

Table 1 represents the performance evaluation metrics comparison study between different data modelling forms. From these obtained results, it is seen that as the order of sine terms increases, the *RMSE* decreases, and the correlation coefficient R increases. This tends to the best matching between the actual data and using the Gaussian model function (with the number of terms = 8) and the SoS modelling function (with the number of terms=7). All these selected models have minimum *RMSE* and large *R* values as compared to other models.

TABLE 1. PERFORMANCE EVALUATION METRICS

Actual data and mathematical model data (5028 points)	RMSE	Correlation Coefficient (R)
Polynomial function (degree = 9)	0.1437	0.8126
Exponential function (number of terms = 2)	0.2113	0.5153
Gaussian function (number of terms = 6)	0.08095	0.9447
Gaussian function (number of terms = 7)	0.04592	0.9825
Gaussian function (number of terms = 8)	0.04591	0.9826
SoS function (number of terms = 2)	0.1557	0.7737
SoS function (number of terms = 4)	0.1001	0.9139
SoS function (number of terms = 6)	0.05724	0.9727
SoS function (number of terms = 7)	0.04655	0.9821

Statistical Analysis

Figures 2, 3, and 4 illustrate the obtained results using a selected data modelling approach. The different mathematical models examined in this paper are polynomial function with order 9, the exponential model expressions, the general Gaussian function with 8 terms, and the SoS with different terms. The numbers of sine terms tested using this method are 2, 4, 6 and 7 terms. From the obtained results listed in Table 1, we conclude the following:

A. The *RMSE* between the actual sequence and the mathematical form using the polynomial function and the exponential function is large, and *R* is small.

B. The *RMSE* values using the Gaussian model with the number of terms 8 and the SoS with the number of terms 7 are very small and minimum, while the correlation coefficient *R* increases to a value approximately = 1.

C. Increasing the number of terms of the SoS model reduces the *RMSE* between the actual sequence and the mathematical form. Also, the correlation coefficient *R* increases, which realizes the matching.

D. There is nearly complete, optimum coincidence and matching between the actual data and the proposed data modelling form using the two mathematical representation models (Gaussian model with number of terms = 8 & SoS with the number of terms = 7).

8. NON-PARAMETRIC TECHNIQUES FOR POWER SPECTRUM ESTIMATION

A simple way to estimate the power spectrum of the DNA sequence is to use non-parametric techniques, which are classical. These classical spectrum estimation techniques are based on the direct computation of the Fourier transform for the available data record, that is, the periodogram spectrum or its improved versions. The non-parametric power spectrum estimation techniques are Bartlett, Welch, and Blackman and Tukey. These techniques are called non-parametric because they do not depend on how the data were generated [17-21].

A. Periodogram

This technique for estimating the power spectrum of a process is to find the discrete-time Fourier transform of the samples of the process and take the magnitude squared of the result. The discrete Fourier transform of the real finite-length sequence $x(n)$, $0<n<N-1$ is $X(f)$. The quantity $|X(f)|^2$ represents the distribution of the signal energy as a function of frequency. It is called the power spectral density of the signal.

The periodogram is defined as [17]:

$$I_N(f) = \frac{1}{N}|X(f)|^2 \quad (9)$$

B. Bartlett Technique

The Bartlett technique used for reducing the variance in the periodogram involves 3 steps [17]:

For N points, they are divided into K segments with length M, i.e., $N = KM$. For each segment of the DNA sequence, a periodogram is estimated. This is followed by averaging the periodograms for K blocks $B(f)$. Thus,

$$Var[B(f)] = \frac{1}{k}Var[I_M(f)] \quad (10)$$

Therefore, the variance of the Bartlett power spectrum estimate has been reduced by the factor k.

C. Welch Technique

An improved estimator of the PSD is the one proposed by Welch [17]. This technique consists of dividing the time series data into possibly overlapping segments with 25% and 50% overlap between successive data segments. Then, windowing is applied to the data segments before computing the periodogram of each segment. Thus, averaging is performed on all segments to decrease the variance of the estimated spectrum.

D. Blackman and Tukey Technique

The steps of Blackman and Tukey technique based on autocorrelation function are summarized as follows [17]:

1. Estimate the sample auto-correlation sequence $r_{xx}(m)$.
2. Window the sample auto-correlation $r_{xx}(m)$ with $w(m)$.
3. Compute the Fourier transform to yield the power spectrum as:

$$P_{xx}(f) = \sum_{-(M-1)}^{M-1} r_{XX}(m)\ w(m)\ e^{-j2\pi fn} \qquad (11)$$

9. PREDICTION OF EXON REGION USING THE PROPOSED HYBRID APPROACH

This section examines the application of the proposed approach for the analysis of the DNA sequences to identify and predict the protein-coding regions. The proposed approach reduces the background noise in the DNA sequences using the digital filter. Calculating the power spectrum allows us to determine the exon regions.

A digital filter with inverse Chebyshev approximation has been chosen due to its high selectivity, which can be achieved using a low-order transfer function. Inverse Chebyshev filter does not exhibit ripples in its passband amplitude response, which is highly needed for the prediction of the exon region. The numerical DNA sequence is filtered using inverse Chebyshev bandpass filter with the following specifications:

Filter order N = 3, the lower and upper passband edge frequencies are [0.663, 0.669], the lower and upper stopband edge frequencies are [0.66, 0.672], the maximum passband attenuation = 1 dB and the minimum stopband attenuation = 30 dB [6, 15-16].

In order to identify gene region in the whole DNA sequence using the proposed hybrid approach, the DNA sequence is passed through that digital filter with the above specifications. All non-parametric spectral estimation methods based on Discrete Fourier Transform exhibit sharp peaks at the cut-off frequency. Exons are isolated within the genes of eukaryotic cells by using the proposed approach.

A. Exon Region Prediction Results Using the Periodogram Technique

Figure 5a to 5c illustrates the exon region prediction results using the periodogram technique as the number of points N increases from 512 to 2048 for actual data, Gaussian model, and the SoS model. From these obtained results, it is seen that a sharp peak is detected at the normalized cut-off frequency (0.667) above -70 dB threshold level for both actual data and SoS model, while no peak can be detected for the Gaussian model. These results indicate that the resolution is improved as the number of points N increases.

B. Exon Region Detection Results Using the Bartlett Technique

Figure 6a to 6c illustrates the exon region detection results using the Bartlett technique for k =2, 4, and 8 for actual data, Gaussian model, and the SoS model. The obtained results indicate that a sharp peak is detected at the normalized cut-off frequency (0.667) for both actual data and the SoS model. In addition, no peak can be detected using the Gaussian model. Also, these results indicate that the frequency resolution is decreased as k is increased, and the variance is reduced.

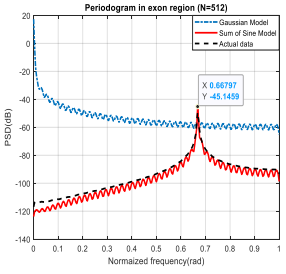

Figure. 5-a

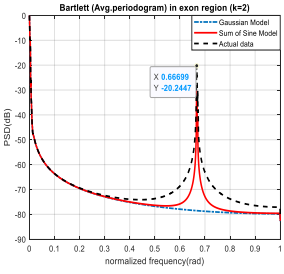

Figure. 6-a

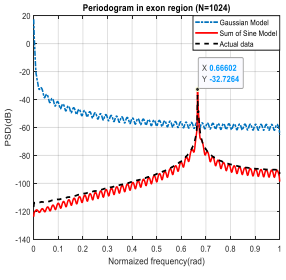

Figure. 5-b

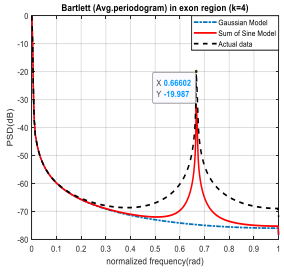

Figure. 6-b

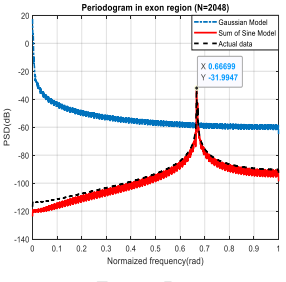

Figure. 5-c

Figure. 5 Periodogram results

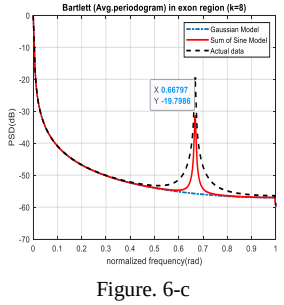

Figure. 6-c

Figure. 6 Bartlett results

C. Exon Region Prediction Results Using the Welch Technique

Figure 7 gives the exon region prediction results using the Welch technique for different overlap ratios. For the case of (25%) overlap, the obtained results are shown in Fig. 7a, 7b, and 7c. For the case of an overlap ratio of (50%), the results are given in Fig. 7d, 7e, and 7f for actual data, Gaussian model, and the SoS model. From these obtained results, it is seen that above -70 dB threshold level, a sharp peak is detected at the normalized cut-off frequency (0.667) for actual data, and the SoS model, while no peak can be detected using the Gaussian model. These results indicate that the frequency resolution is decreased as the number of points is increased from 512 to 2048 points, and the variance is also reduced. The data segment overlapping improves the characteristics of the spectrum estimate as compared to the obtained spectrum using the Bartlett technique.

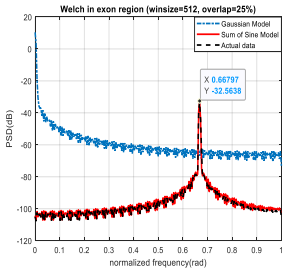

Figure. 7-a

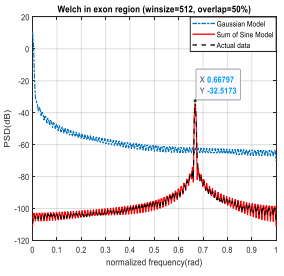

Figure. 7-d

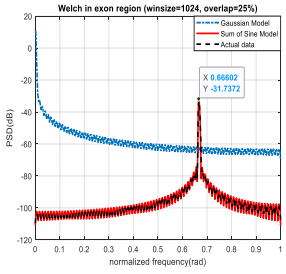

Figure. 7-b

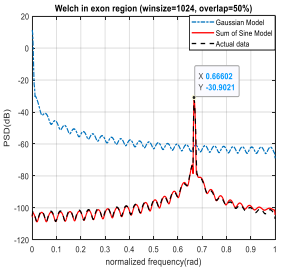

Figure. 7-e

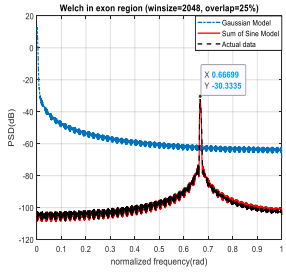

Figure. 7-c

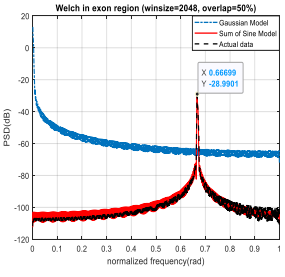

Figure. 7-f

Figure. 7 Welch results

D. Exon Prediction Results Using the Blackman and Tukey Technique

Figure 8a to 8c depicts the exon region prediction results using the Blackman and Tukey technique for actual data, Gaussian model, and the SoS model. From these obtained results, it is seen that a sharp peak is detected at the normalized cut-off frequency (0.667) above -70 dB threshold level for both actual data and the SoS model, while no peak can be detected with the Gaussian model. These results indicate that the resolution is improved as the number of points *N* increases from 512 to 2048 points.

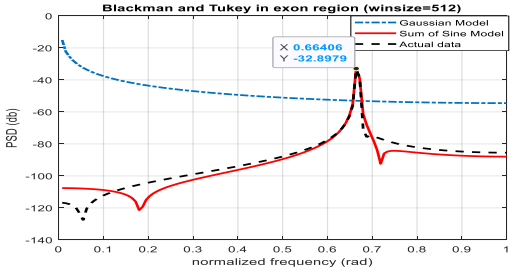

Figure. 8-a

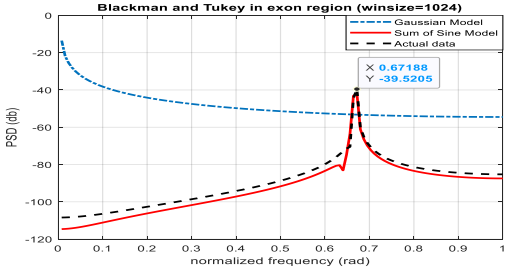

Figure. 8-b

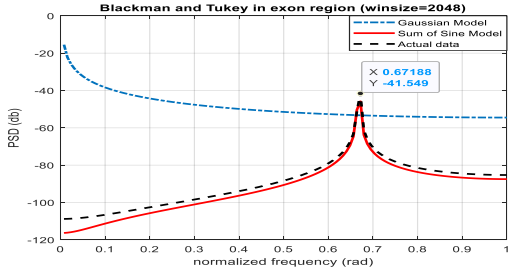

Figure. 8-c

Figure. 8 Blackman and Tukey results

10. RESULT DISCUSSION AND COMPARISON STUDY

The periodogram is computationally efficient when the FFT algorithm is employed. The variance of the periodogram decreases as *N* increases and does not approach zero as *N* tends to infinity. Thus, the spectrum is said to be consistent.

From these obtained results, we have concluded that:

- The variance of the Bartlett spectrum estimate is reduced at the expense of increased bias and decreased spectrum resolution.
- Welch and Blackman and Tukey power spectrum estimates are better than Bartlett estimate, but the difference in performance is relatively small.
- The Welch method requires a little more computational power than do the other two methods.
- When the DNA sequences are mapped to EIIP indicator sequences, and by applying the power spectrum techniques on these sequences reveals a sharp peak for the exon region, which provides good discrimination between exon areas and non-coding areas of several genomes.
- A sharp peak is detected at the normalized cut-off frequency (0.667) for both actual data and the SoS model. In addition, no peak can be detected using the Gaussian model.
- An efficient mechanism for the identification of exon region based on digital filtering and non-parametric spectral estimation techniques has been presented. The output of the digital filter gives a sharp peak at the normalized cut-off frequency (0.667).

11. CONCLUSION

In this paper, a proposed hybrid approach using non-parametric spectral estimation techniques have been presented for the analysis of DNA sequences and exon region detection in DNA sequences. EIIP method has been used to convert the DNA symbolic sequence to numerical values. Also, different closed-form mathematical equations have been presented using a data modelling approach. The best mathematical expressions which represent the DNA sequences are the Gaussian model with the number of terms = 8 and the SoS model with the number of terms = 7. The statistical *RMSE* and correlation coefficient *R* metric parameters are optimum for the Gaussian model and the SoS model. Thus, each DNA sequence can be represented mathematically for each one, which can be used to discriminate between different DNA sequences. The proposed DNA mathematical modelling methods have been used to build dictionaries for the DNA of different persons. Such dictionaries can be used for further exploration of the characteristics of various diseases. Besides, this paper presented a useful track for doing various analyses of DNA sequences and exon region prediction. The proposed hybrid approach improves the detectability of the peak in the exon region that is used for finding genes. The results of gene prediction from DNA sequences in exon regions based on original and modelled DNA sequences coincide to a great extent, which ensures the success of the proposed SoS method for optimum modelling of DNA sequences, while the Gaussian model is not appropriate for this task.

REFERENCES

[1] Luscombe, N.M., D. Greenbaum, and M. Gerstein, What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine, 2001. 40(4): p. 346-358.

[2] Attwood, T. and C. Miller, Progress in bioinformatics and the importance of being earnest, Biotechnology Annual Review 8, 2002, P: 1-54.

[3] Anastassiou, D., Genomic signal processing. Signal Processing Magazine, IEEE, 2001. 18 (4): p. 8-20.

[4] I. M. El-Badawy, S. Gasser, and A. M. Aziz, on the use of pseudo-EIIP mapping scheme for identifying exons locations in DNA sequences. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2015.

[5] Abo-Zahhad, M., S.M. Ahmed, and S.A. Abd-Elrahman, Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques. International Journal of Information Technology and Computer Science (IJITCS), 2012. 4(8): p. 22-36.

[6] H. M. Wassfy, M. L. Salem, M. M. Abdelnaby, M. S. Mabrouk, A. Zidan, Advanced DNA Mapping Schemes for Exon Prediction Using Digital Filters. American Journal of Biomedical Engineering, Vol 6(1), 2016: p. 25-31.

[7] Inbamalar, T. and R. Sivakumar, Study of DNA sequence analysis using DSP Techniques. Journal of Automation and Control Engineering Vol, 2013. 1(4): P. 336-342.

[8] Saberkari, H., et al. prediction of protein-coding regions in DNA sequences using signal processing methods, in Industrial Electronics and Applications (ISIEA), IEEE Symposium on. 2012 :p 355-360.

[9] Berger, J., S. Mitra, and J. Astola. Power spectrum analysis for DNA sequences, in Signal Processing and Its Applications, Proceedings. Seventh International Symposium, IEEE, 2003: p.29-32.

[10] Saberkari, H., et al., A fast algorithm for exon regions prediction in DNA sequences. Journal of medical signals and sensors, 2013. 3(3): p. 139-147.

[11] Saberkari, H., et al., A fast algorithm for exon regions prediction in DNA sequences. Journal of medical signals and sensors, 2013. 3(3): p. 139-147.

[12] P. Wąż, D. Bielińska-Wąż, Non-standard similarity/dissimilarity analysis of DNA sequences, Genomics 104, 2014, PP 464-471.

[13] D. Bielińska-Wąż, Graphical, and numerical representations of DNA sequences statistical aspects of similarity, J. Math. Chem. 49 (2011) 2345-2407.

[14] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, International Journal of Quantum Chemistry (2013), vol. 113, 2413-2446.

[15] Ahmed M. Dessouky, Taha E.Taha, Mohamed M. Dessouky, Ashraf A. Eltholth, Emadeldeen Hassan, and Fathi E. Abd El-Samie, "Non-Parametric Spectral Estimation Techniques for DNA Sequence Analysis and Exon Region Prediction," Computers and Electrical Engineering Journal, Vol.73, 2019, 334-348.

[16] Ahmed M. Dessouky, Taha E.Taha, Mohamed M. Dessouky, Ashraf A. Eltholth, Emadeldeen Hassan, and Fathi E. Abd El-Samie, " Visual Representation of DNA Sequences for Exon Detection Using Non-Parametric Spectral Estimation Techniques," Nucleosides, Nucleotides and Nucleic acid journal, 2019. (Taylor and Francis).

[17] J. G. Proakis, C. M. Rader, F. Ling, and C. L. Nikias, Advanced Digital Signal Processing, Maxwell Macmillan International, New York, USA. 1992.
[18] S. J. Orfanidis, Optimum Signal Processing: An Introduction, Mcgraw-Hill Book Company, 1990.
[19] N. Kalouptsidis, Signal Processing Systems, Theory and Design, John Wiely And Sons, Inc. 1997.
[20] G. Zelniker and F. J. Taylor, Advanced Digital Signal Processing: Theory And Applications, Marcel Dekker, Inc., New York 1994.
[21] E. A. Robinson, A Historical Perspective of Spectrum Estimation, Proc. Of The IEEE, Vol.70, No.9, Sept.1982, P.P.885-907.
[22] https://srogic.wordpress.com/datasets/hmr195-dataset/. [accessed at 6 / 4 / 2020]
[23] Mabrouk, M., S., A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques, American Journal of Biomedical Engineering, Vol, 2012. 2 (2): p. 17-22.
[24] Nair, A.S., and S.P. Sreenadhan, A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation, 2006. 1(6): p. 197-202
[25]. Yin, C., and S.S.-T. Yau. Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes. In Computational Intelligence in Bioinformatics and Computational Biology, IEEE Symposium on. 2008: p. 223-227.
[26] Bai Arniker, S. and H.K. Kwan. Advanced numerical representation of DNA sequences. In International Conference on Bioscience, Biochemistry and Bioinformatics IPCBEE. 2012, 31: p. 2-5.

Ahmed Moawad Dessouky received the B.Sc., (Hons.), and M.Sc degree from Menuofia University, Menouf, Egypt in 2013 and 2019 respectively. Since 2016, he has been a Teaching Staff Member in AL-ALSON Academy, Cairo, Egypt. His research interest is in communication engineering, computer science and engineering and medical signal processing.

Fathi E. Abd El-Samie received the B.Sc.(Hons.), M.Sc., and PhD degrees from Menoufia University, Menouf, Egypt, in 1998, 2001, and 2005, respectively. Since 2005, he has been a Teaching Staff Member with the Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, menoufia University. He worked as a researcher at KACST-TIC in Radio Frequency and Photonics for the e-Society (RFTONICs) from 2013 to 2015. His current research interests include image enhancement, image restoration, image interpolation, super-resolution reconstruction of images, data hiding, multimedia communications, medical image processing, optical signal processing, and digital communications. He was a recipient of the Most Cited Paper Award from the Digital Signal Processing journal in 2008.

Hesham F. A. Hamed received the B.Sc. degree in electrical engineering, the M.Sc. and PhD degrees in electronics and communications engineering from EL-Minia University, EL-Minia, Egypt, in 1989, 1993, and 1997 respectively. He currently is the dean of faculty of engineering, Minia University. He was a Visiting Researcher at Ohio University, Athens, Ohio. From 1989 to 1993 he worked as a Teacher Assistant in the Electrical Engineering Department, EL-Minia University. From 1993 to 1995 he was a visiting scholar at Cairo University, Cairo, Egypt. From 1995 to 1997 he was a visiting scholar at Texas A&M University, College Station, Texas (with the group of VLSI). From 1997 to 2003 he was an Assistant Professor in the Electrical Engineering Department, EL-Minia University. From 2003 to 2005 he was Associate Professor in the same University. He has published more than 80 papers. His research interests include analog and mixed-mode circuit design, low voltage low power analog circuits, current-mode circuits, nano-circuits design, and FPGA.

Gerges Mansour Salama received the B.Sc. degree in electrical engineering and the M.Sc. degrees in electronics and communications engineering from EL-Minia University, EL-Minia, Egypt, in 1999 and 2006 respectively. He received the Ph.D from Faculty of Telecommunication Networks, Switching Systems, and Computer Technology (FTN, SS, and CT) ST. PETERSBURG STATE UNIVERSITY OF TELECOMMUNICATIONS NA. PROF. MA BONCH-BRUEVICH. Ministry of Communications and Mass Media of the Russian Federation Federal Communications Agency in 2012. Now, he is an assistant professor at the Faculty of Engineering, Minia University, Egypt.