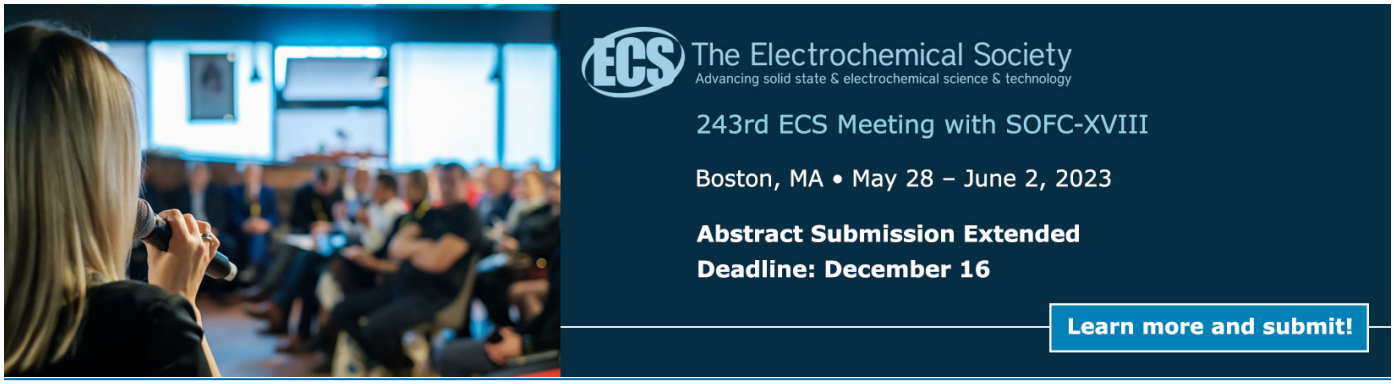
CpG islands identification in DNA sequences using modified P-spectrum based algorithm



CpG islands identification in DNA sequences using modified P-spectrum based algorithm

P Garg[1] and S D Sharma[1]

[1]Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, 173234, India

garg.pardeep22@gmail.com, sdsharma.juet@rediffmail.com

Abstract: The identification of CpG Islands play a major role in the analysis of DNA sequences because of association of CpG Islands with many epigenetic events. Some of these events are promoter activity and consequently gene prediction, chromosome inactivation, and for early detection of cancer etc. And hence the exact identification of CpG Islands in the stretch of DNA sequences has always remained a challenging task. Numerous computational algorithms have been developed for the identification of CpG Islands in DNA sequences. Recently various transform based methods have been reported for the CpG Islands detection in literature. In these transform based methods, there is a requirement of transforming the signal from time domain to frequency domain and correspondingly there is probability of transform biasing. Hence to overcome this issue, a modified P-spectrum based algorithm has been proposed in this paper. Also the performance of the proposed method has been compared with recently reported methods of CpG Islands detection using standard evaluation metrics. The performance of proposed method has been proved to be much better than the other methods and hence the proposed approach is an efficient method in detecting CpG Islands.

Keywords: P-spectrum, discrete wavelet transform (DWT), CpG Islands (CGI), DNA sequences, numerical mapping.

1. Introduction

Genomic signal processing is gaining a lot of popularity after the completion of human genome sequencing project. Tremendous amount of genomic data was generated after this project and there is a need to extract the very useful biological information hidden in this data. Numerous computational methods have been developed and reported in literature in recent years for the interpretation of this genomics data. Genomics data like as deoxyribonucleic acid (DNA) sequences is comprised of Adenine (A), Guanine (G), Thymine (T), and Cytosine (C) characters. Most of the DNA data generally consists of recurring patterns. The examples of such patterns which have recurring nature are like exons [1-3], retention of introns [4], tandem repeats [5-6], CpG Islands [7-11] etc. CpG Islands identification is the major area of emphasis in this paper. CpG Islands are the regions inside DNA sequences where the DNA character ‘C’ is followed by character ‘G’ and the concentration of CG dinucleotides is high in this region. CpG Islands are considered as a very important region inside the stretch of DNA sequences

because the identification of CpG Islands can help in the prediction of promoter regions and consequently genic regions, some human malignancies, and also can be useful for the early prediction of cancers [12]. These are some of the motivations for the researchers working in the field of genomics signal processing to develop algorithms for the identification of CpG Islands.

Recently many transform based algorithms have been reported in literature for CpG Islands identification. It is an essential part of transform based methods to transform the signal from time domain to frequency domain. It is quite probable to occur transform biasing because of domain transformation which may consequently lead to loss of information. Hence to overcome this issue, a modified P-spectrum based algorithm has been proposed in this paper for CpG Islands identification which is devoid of domain transformation. The key contributions of proposed method are as follows:

i)Modified P-spectrum has been employed to overcome the domain transformation effect to identify the CpG Islands in DNA sequences.
ii) The proposed method's performance has been proved to the best amongst the recently reported methods.

The proposed method's performance has been tested on the CpG Islands dataset downloaded from national centre for biotechnology information (NCBI) and compared with the recently reported methods. It has been interpreted from the results obtained that the proposed method's performance is the best amongst the other methods. The organization of the remainder of the paper is as following. The methods employed for CpG islands identification in this paper have been described in section 2. The CpG islands data set and the performance metrics used in the paper have been discussed in section 3. In section 4 results and discussion is presented and the conclusion is presented in section 5.

2. CpG Islands Identification Methods Used

The methods used for CpG islands detection in this paper are discussed as following:

2.1 DWT based CpG Islands identification method

The method proposed by Mariapushpam et al. [13] uses DWT based algorithm for the CpG Islands detection. In this method, the DNA characters are converted to numerical values using modified EIIP mapping scheme reported in [13]. In the modified EIIP mapping scheme, the numerical values assigned to DNA characters are: A=0.1260, C= 1, G=1, T= 0.1335. Then the numerical version of DNA sequence is pre-processed to filter out the range of frequencies to required range using band-pass filters. The discrete wavelet transform in which Symlet 11 wavelet function has been used is then applied to the pre-processed signal and then the recursive least squaring (RLS) algorithm based adaptive filtering is done to obtain the detected CpG Islands.

2.2 DWT using combination of 24 mappings of integer mapping based CpG Islands identification method

Garg et al. [14] proposed a method for CpG Islands detection in which the authors have enhanced the sensitivity of DWT based method of CpG Islands identification. In this paper the authors have analysed the performance of 13 existing mapping schemes, for DWT based CpG Islands identification method and proposed the combination of 24 mappings of integer mapping scheme. Using the proposed approach of combination of 24 mappings of integer mapping scheme, the authors have enhanced the sensitivity of DWT based algorithm of CpG Islands identification method.

2.3 Proposed modified P-spectrum based method for CpG Islands identification

P-spectrum has been used in literature [15-18] to detect the periodicities present in signals and hence it is also known as periodicity spectrum. It is reported in literature that CpG Islands are the segments in

DNA sequences which generally have high frequency recurring pattern of CG dinucleotides [12]. As the high frequency recurring pattern correspond to microsatellites in DNA sequences and the periodicity range of microsatellites is generally 2-10 periods [19]. Therefore, to capture these periodicities and hence to identify the CpG Islands in DNA sequences; the modified P-spectrum based algorithm has been proposed in this paper. The proposed algorithm's flow chart is depicted in figure 1.

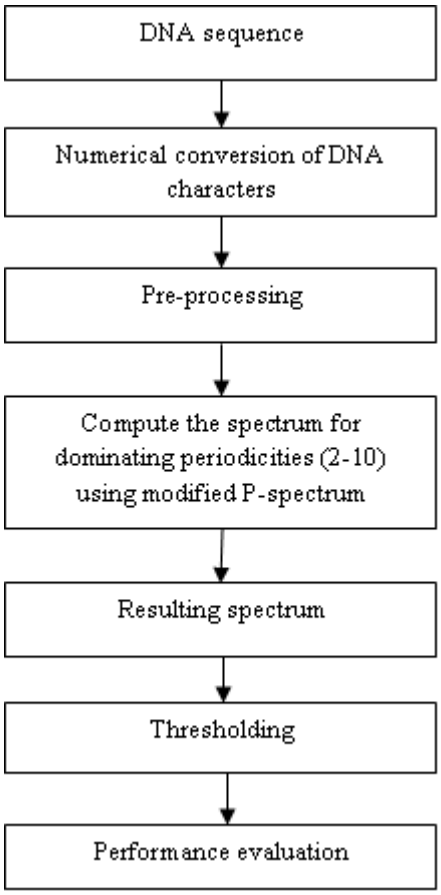


Figure 1. Proposed algorithm's flow chart.

The steps of the proposed algorithm are discussed as following:

i) The DNA sequence is obtained from database.

ii) The DNA characters are converted to numerical values using electron ion interaction potential (EIIP) mapping scheme. In EIIP mapping scheme, the numerical values T= 0.1335, G= 0.0806, A= 0.1260, C= 0.1340 are assigned to DNA characters.

iii) The numerical sequence data is passed through an anti-notch filter centered at angular frequency $2\pi/3$ to filter out the noise.

iv) The spectrum of 2-10 dominating periodicities is then computed using modified P-spectrum.

The overview of modified P-spectrum for a period p is as following [15-18]:

Let's assume a signal B in discrete-time is represented as:

$$\boldsymbol{B} = [\boldsymbol{b_1} \;\; \boldsymbol{b_2} \;\; \boldsymbol{b_3} \ldots \;\; \boldsymbol{b_M}] \tag{1}$$

It is required that the given signal's length should be a strict multiple of period p to compute the p-spectrum. To achieve this, zeros have to be inserted in the end of the signal B equivalent to difference between period p and the remainder; where remainder has to be computed by the division of signal B with period p. Hence the signal B becomes:

$$\boldsymbol{B} = [\boldsymbol{b_1} \;\; \boldsymbol{b_2} \;\; \boldsymbol{b_3} \ldots \;\; \boldsymbol{b_M}\, \mathbf{0\ 0\ 00} \ldots \ldots] \tag{2}$$

In the next step, a matrix X_p is calculated by dividing signal B into 'n' non-overlapping regions of length of period p. The matrix X_p is represented as:

$$\mathbf{X_p} = \begin{bmatrix} \mathbf{b_1} & \mathbf{b_2} & \mathbf{b_3} & \cdots \mathbf{b_i} & \mathbf{b_{i+1}} \cdots & \mathbf{b_p} \\ \mathbf{b_{p+1}} & \mathbf{b_{p+2}} & \mathbf{b_{p+3}} & \cdots \mathbf{b_{p+i}} & \mathbf{b_{p+i+1}} \cdots & \mathbf{b_{2p}} \\ \mathbf{b_{np+1}} & \mathbf{b_{np+2}} & \mathbf{b_{np+3}} & \mathbf{b_M} & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{3}$$

Now, we compute the singular value decomposition (SVD) of X_p as:

$$\boldsymbol{X_{svd}} = \boldsymbol{SVD}(\boldsymbol{X_p}) \tag{4}$$

The largest singular value of X_{svd} has been then selected to calculate the signal max_sing as shown in equation (5):

$$\mathbf{max}\_\boldsymbol{sing} = \boldsymbol{max}(\boldsymbol{X_{svd}}) \tag{5}$$

Then the signal s is computed which is obtained by summing all the values of X_p as represented in equation (6):

$$\boldsymbol{s} = \boldsymbol{sum}(\boldsymbol{X_p}) \tag{6}$$

Now, the signal mod_spec is calculated as shown in equation (7):

$$\boldsymbol{mod\_spec} = \boldsymbol{max}(\boldsymbol{s}/\mathbf{2}) \tag{7}$$

And finally the resultant spectrum signal, res_spec is calculated by row-wise multiplication of max_sing and mod_spec as represented in equation (8):

$$\boldsymbol{res\_spec} = \mathbf{max}\_\boldsymbol{sing} \times \boldsymbol{mod\_spec} \tag{8}$$

v) The resultant spectrum is calculated by adding the spectrums of 2-10 periodicities.

vi) An empirically selected value as mean of the spectrum is then chosen as threshold for the proposed method and the other methods as well.

vii) And finally the performance metrics are computed.

An example DNA sequence having accession number AL031724 [20] has been taken to show the applicability of the proposed algorithm for CpG Islands detection and also for the comparison of proposed algorithm's performance with the other two CpG Islands detection methods used in the paper. The length of DNA sequence AL031724 is 21419 base pairs (bps) and the locations of two CpG Islands in this sequence are 1456-2506, and 4427-6491. The result obtained using the three methods for example DNA sequence is shown in figure 2-4.

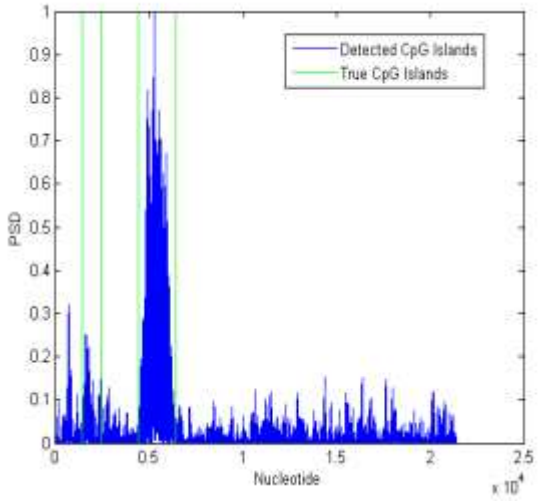


Figure 2. Result obtained using DWT method.

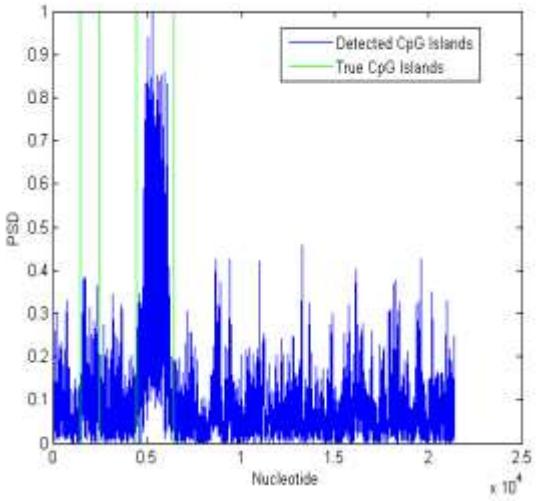


Figure 3. Result obtained using DWT using combination of 24 mappings of integer mapping method.

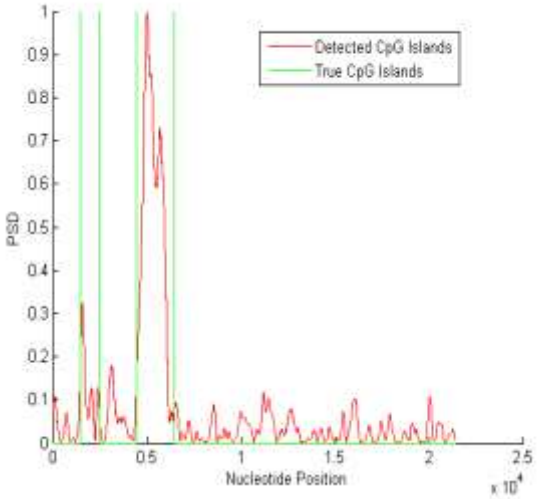


Figure 4. Result obtained using proposed method.

The 4 possible detection outcomes which are true positive (TP), true negative (TN), false positive (FP), false negative (FN), and the performance metrics namely sensitivity (Sn), specificity (Sp), accuracy (AC), F-measure obtained using the three methods for example DNA sequence are presented in table 1.

Table 1. Performance parameters for example DNA sequence AL031724

Performance metric	Methods		
	DWT	DWT using combination of 24 mappings of Integer mapping	Proposed algorithm
TP	1784	2198	**2395**
FP	1974	4669	**1418**
TN	1332	13633	**16884**
FN	16328	918	**721**
Sn	0.5725	0.7054	**0.7686**
Sp	0.8921	0.7449	**0.9225**
AC	0.8456	0.7391	**0.9001**
F-measure	0.5191	0.4403	**0.6913**

The ability of proposed method to identify the CpG islands has been interpreted from table 1and also its performance is much higher amongst other methods in terms of all performance metrics.

3. CpG Island's Data Set and Performance Metrics

3.1 Data set of CpG Islands

The publically available database NCBI has been used to download the data set of CpG islands for the performance evaluation of proposed method with the other existing methods. The DNA sequences used in this paper are Z68274, AL031718, and D13370 [20]. The detailed description of these DNA sequences is as following. The DNA sequence Z68274 has length of 20587 bps and the CpG Island in this sequence is located at 13249-13799 [20]. The length of DNA sequence AL031718 is 20612 bps and 2 CpG islands in this sequence are located at 4620-8266 & 16033-18297 locations [20]. The DNA sequence D13370 comprises of 3730 bps and CpG Island is located at 226-1645 [20].

3.2 Performance metrics

The performance of the proposed algorithm has been compared with other CpG Islands detection methods on the basis of performance metrics namely sensitivity (Sn), specificity (Sp), accuracy (AC) [12], F-measure [10] These performance metrics are described as following:

$$Sn = \frac{TP}{TP + FN} \tag{9}$$

$$Sp = \frac{TN}{TN + FP} \tag{10}$$

$$AC = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

$$F - measure = \frac{2 \times (precision \times recall)}{precision + recall}, \tag{12}$$

$$where\ precision = \frac{TP}{TP + FP} \& recall = \frac{TP}{TP + FN}$$

True positive (TP) represents those locations which are detected correctly by algorithm where true CpG Islands are available, False positive (FP) represents those locations which are falsely predicted by algorithm where true CpG Islands are not available, True negative (TN) corresponds to the locations which are correctly identified where true CpG Islands are not available, and False negative (FN) gives the unidentified locations where true CpG Islands are available. The term sensitivity (Sn) emphasizes the details regarding the proportion of TP which are truly detected. Specificity (Sp) depicts the proportion of TN which is correctly identified. The range of Sn and Sp can be between 0 to 1. It is desired to have an ideal value of 1 for Sn and Sp for a perfect algorithm. Another evaluation metric used in the paper is Accuracy (AC) which varies in the range of 0 to 1. It is expected to have the value of AC near to 1 for an algorithm to be considered perfect. The F-measure is an assessment of an algorithm's accuracy and it describes the harmonic mean of the precision and recall. Its value ranges between 0 to 1. It is preferred to have the value of F-measure as 1 for better performance.

4. Results and discussion

Two recently reported methods of CpG Islands detection namely DWT based method and DWT using combination of 24 mappings of integer mapping based method have been examined for the performance comparison of the proposed algorithm. The result obtained on DNA sequence Z68274 using the three methods is presented in figure 5-7.

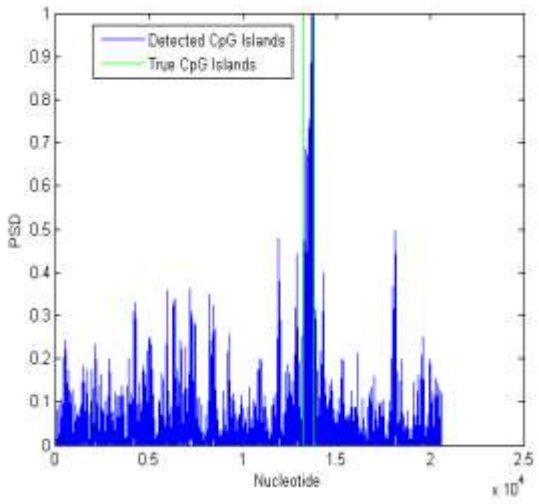


Figure 5. Result obtained using DWT method.

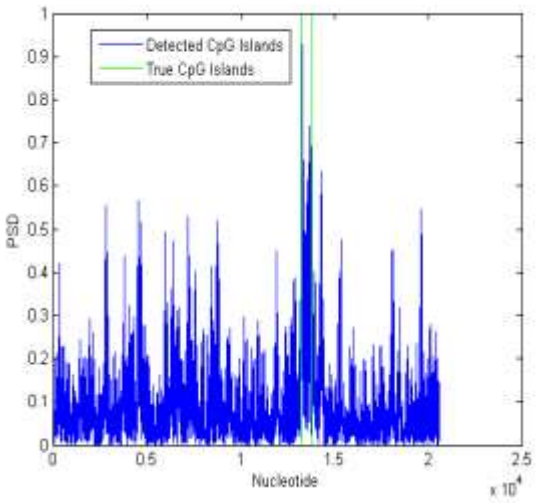


Figure 6. Result obtained using DWT using combination of 24 mappings of integer mapping method.

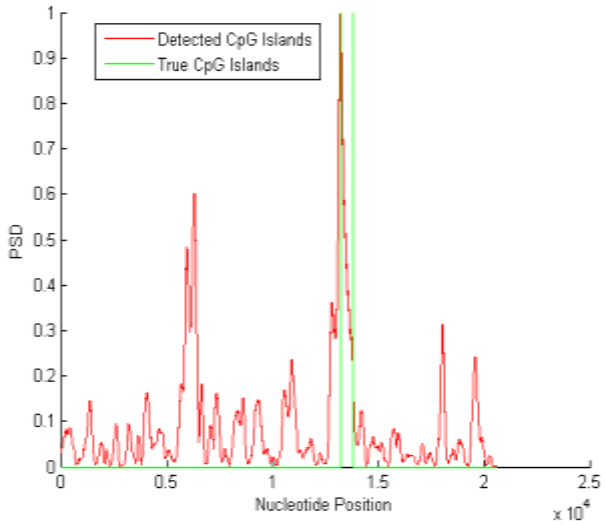


Figure 7. Result obtained using proposed method.

Table 2 depicts the detection outcomes and the performance metrics obtained using the three methods for DNA sequence Z68274.

Table 2. Performance parameters for DNA sequence Z68274

Performance metric	Methods		
	DWT	DWT using combination of 24 mappings of Integer mapping	Proposed algorithm
TP	380	469	**551**
FP	5298	6707	**4998**
TN	14737	13328	**15037**
FN	171	82	**0**
Sn	0.6897	0.8512	**1**
Sp	0.7356	0.6652	**0.7505**
AC	0.7343	0.6702	**0.7572**
F-measure	0.1220	0.1214	**0.1807**

The result obtained on DNA sequence AL031718 using the three methods is depicted in figure 8-10.

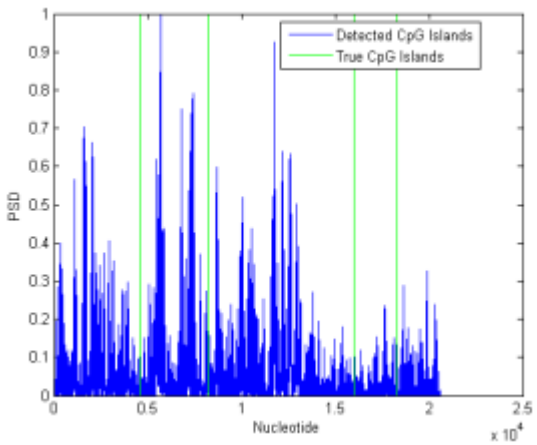


Figure 8. Result obtained using DWT method.

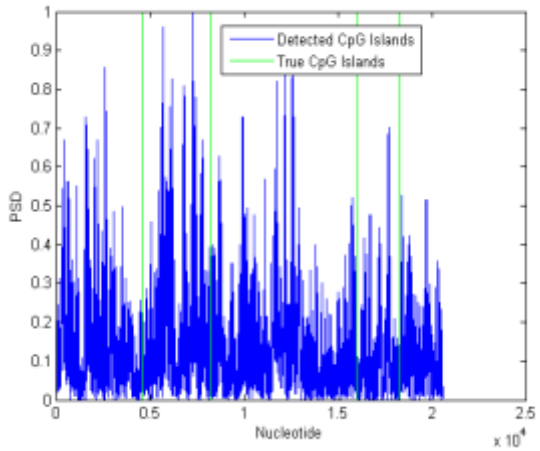


Figure 9. Result obtained using DWT using combination of 24 mappings of integer mapping method.

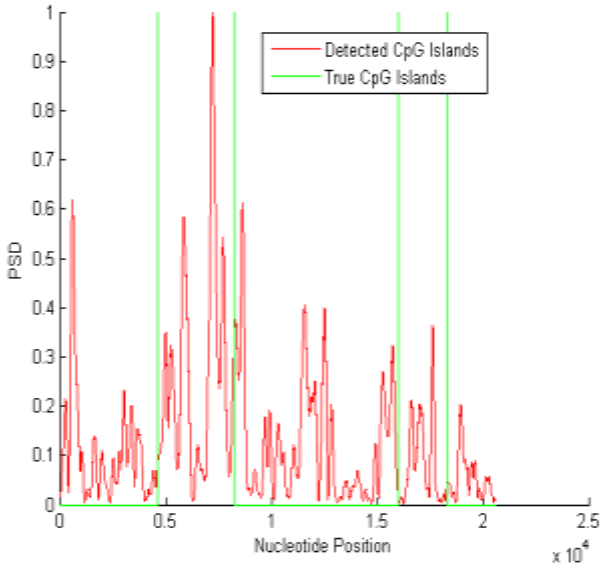


Figure 10. Result obtained using proposed method.

The detection outcomes and the performance metrics obtained using the three methods for DNA sequence AL031718 are presented in table 3.

Table 3. Performance parameters for DNA sequence AL031718

Performance metric	Methods		
	DWT	DWT using combination of 24 mappings of Integer mapping	Proposed algorithm
TP	1556	2266	**2915**
FP	4226	5122	**4224**
TN	10473	9577	**10475**
FN	4356	3646	**2997**
Sn	0.2632	0.3833	**0.4931**
Sp	0.7125	0.6515	**0.7126**
AC	0.5836	0.5746	**0.6497**
F-measure	0.2661	0.3408	**0.4467**

The result obtained on DNA sequence D13370 using the three methods is depicted in figure 11-13.

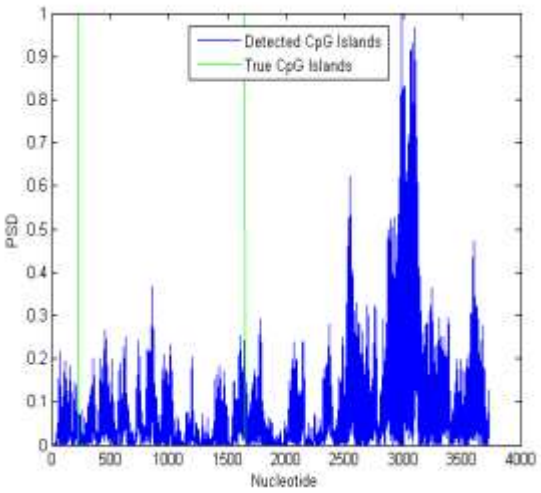


Figure 11. Result obtained using DWT method.

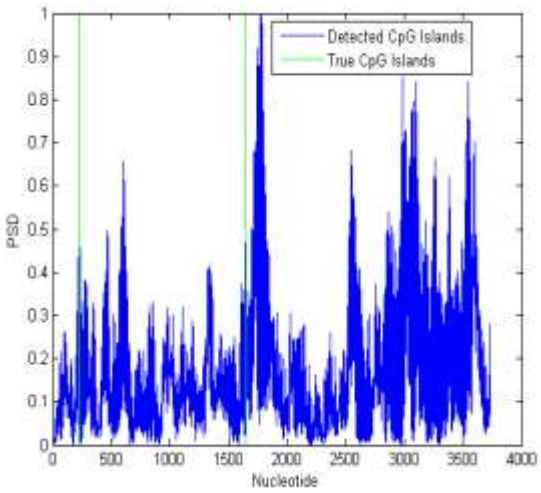


Figure 12. Result obtained using DWT using combination of 24 mappings of integer mapping method.

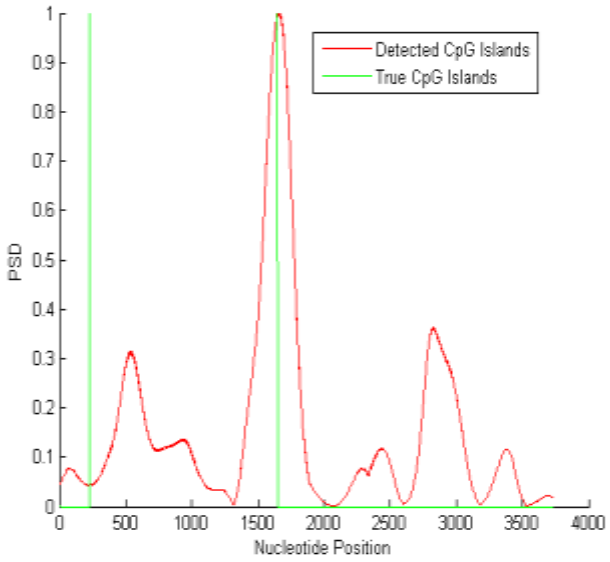


Figure 13. Result obtained using proposed method.

The detection outcomes and the performance metrics obtained using the three methods for DNA sequence D13370 are presented in table 4.

Table 4. Performance metrics for DNA sequence D13370

Performance metric	Methods		
	DWT	DWT using combination of 24 mappings of Integer mapping	Proposed algorithm
TP	270	378	**482**
FP	901	984	**509**
TN	1408	1325	**1800**
FN	1150	1042	**938**
Sn	0.1901	0.2662	**0.3394**
Sp	0.6098	0.5738	**0.7796**
AC	0.4500	0.4567	**0.6120**
F-measure	0.2084	0.2717	**0.3998**

The superiority of the proposed method has been observed from the tables 2, 3, and 4. The value of all the performance metrics obtained using proposed method on the individual dataset of DNA sequences is much higher amongst the other methods of CpG Islands detection. Also, the combined detection outcomes and the performance metrics obtained using the three methods for the whole dataset of 3 DNA sequence has been computed and presented in table 5.

Table 5. Performance metrics for whole dataset of 3 DNA sequences

Performance metric	Methods		
	DWT	DWT using combination of 24 mappings of Integer mapping	Proposed algorithm
TP	2206	3113	**3948**
FP	10425	12813	**9731**
TN	26618	24230	**27312**
FN	5677	4770	**3935**
Sn	0.2798	0.3949	**0.5008**
Sp	0.7186	0.6541	**0.7373**
AC	0.6416	0.6086	**0.6958**
F-measure	0.2151	0.2615	**0.3662**

The proposed method's performance is the best amongst other methods and it has been interpreted from table 5. The value of Sn, Sp, AC, and F-measure obtained using the proposed method is much higher than the other methods of CpG Islands detection.

5. Conclusion

In this paper a modified P-spectrum based method has been proposed for CpG Islands detection in DNA sequences. The proposed method's performance has been tested on publically available database. Also, the proposed method's performance has been compared with other recently reported methods of CpG Islands detection. The value of performance metrics Sn, Sp, AC, and F-measure obtained using proposed method is much higher than the other methods of CpG Islands detection. Hence it can be concluded that the proposed method is able to detect the CpG Islands in DNA sequences and is superior to other methods in all performance metrics. In future, the performance of proposed method can be tested on large dataset of DNA sequences.

References

[1] Mena-Chalco JP, Carrer H, Zana Y, Cesar Jr. RM. Identification of protein coding regions using the modified Gabor-wavelet transform. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 2008; 5(2), pp. 198-207.

[2] Kumar RM, Vaegae NK.Walsh code based numerical mapping method for the identification of protein coding regions in Eukaryotes. Biomedical Signal Processing and Control (BSPC) 2020;58, pp. 1-11.

[3] Das L, Nanda S, Das JK. An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser Window. Genomics 2019; 111(3), pp. 284-96.

[4] Sharma, SD, Sharma SN, Saxena R. Identification of Short Exons Disunited by a Short Intron in Eukaryotic DNA Regions. IEEE/ACM transactions on computational biology and bioinformatics. 2020; 17(5) pp. 1660-70.

[5] Sharma SD, Saxena R, Sharma SN. Identification of microsatellites in DNA using adaptive S-transform. IEEE Journal of Biomedical and Health Informatics. 2015;19(3), pp. 1097-1105.

[6] Garg P, Sharma SD, Sharma SN. MGWT based algorithm for tandem repeats detection in DNA sequences. 2019 5th IEEE International Conference on Signal Processing, Computing and Control, 10-12 October 2019, JUIT Waknaghat. 2019; pp. 196-99.

[7] Tahir RA, Zheng D, Nazir A, Qing H. A review of computational algorithms for CpG islands detection. Indian Academy of Sciences. 2019; 44:143, pp. 1-11.

[8] Takai D, Jones PA. Comprehensive analysis of CpG islands in Human chromosomes 21 and 22. Proc Natl. Acad. Sci. 2002; 99(6), pp. 3740-45.

[9] Yu N, Guo X, Zelikovsky A, Pan Y. GaussianCpG: A Gaussian model for detection of CpG island in human genome sequences. BMC Genomics. 2017;18 (Suppl 4), 392.

[10] Garg P, Sharma, SD. Identification of CpG Islands in DNA sequences using Short-Time Fourier Transform. Interdiscip Sci Comput Life Sc*i*. 2020;12(3) 355-67.

[11] Rushdi A, Tuqan J. A New DSP-based measure for CpG islands detection. In Digital Signal Processing Workshop, 12th- Signal Processing Education Workshop, IEEE. 2006; pp. 561-65.

[12] Kakumani R, Ahmad O, Devabhaktuni V. Identification of CPG islands in DNA sequences using statistically optimal null filter. Eurasip J on Bioi. And Sys Biol. 2012; 2012(1):12.

[13] Mariapushpam IT, Rajagopal S. Improved algorithm for the location of CpG Islands in genomic sequences using discrete Wavelet transforms. Current Bioinformatics, 2017;12, pp. 57-65.

[14] Garg P, Sharma SD. Sensitivity Enhancement of DWT based Algorithm for CpG islands detection in DNA sequences. Procedia Computer Science. 2020;167 (2020), pp. 1829-38.

[15] Kanjilal PP, Bhattacharya J, Saha G. Robust method for periodicity detection and characterization of irregular cyclical series in terms of embedded periodic components. Phy. Rev. E. 1999; 59(4), pp. 4013-25.
[16] Qui P and Liu KJR. A robust method for QRS detection based on modified P-spectrum. ICASSP. 2008; pp. 501-4.
[17] Garg P, Sharma SD, Sharma SN. Tandem repeats detection in DNA sequences using P-spectrum based algorithm. In IEEE CICT 2017; pp. 1-5.
[18] Liscombe M and Asif A. A new method for instantaneous signal period identification by repetitive pattern matching. IEEE 13th Int. Multitopic Conf. 2009; pp. 1-5.
[19] Sharma SD, Saxena R, Sharma SN. Short tandem repeats detection in DNA sequences using modified S- transform. Int. J of Adv. in Engg. & Tech. 2015;8(2), pp. 233-45.
[20] National Centre for Biotechnology Information: Available at: https://www.ncbi.nlm.nih.gov/nuccore/. Accessed: 30 October 2020.