Interdisciplinary Sciences: Computational Life Sciences (2020) 12:355–367
https://doi.org/10.1007/s12539-020-00370-y

SHORT COMMUNICATION

Identification of CpG Islands in DNA Sequences Using Short-Time Fourier Transform

Pardeep Garg[1] · Sunildatt Sharma[1]

Received: 18 December 2019 / Revised: 7 April 2020 / Accepted: 17 April 2020 / Published online: 11 May 2020

Abstract

In the era of big data analysis, genomics data analysis is highly needed to extract the hidden information present in the DNA sequences. One of the important hidden features present in the DNA sequences is CpG islands. CpG Islands are important as these are used as gene markers and also these are associated with cancer etc. Therefore, various methods have been reported for the identification of CpG islands in DNA sequences. The key contributions of this work are (i) extraction of the periodicity feature associated with CpG islands using Short-time Fourier transform (ii) a short-time Fourier transform-based algorithm has been proposed for the identification of CpG Islands in DNA sequences. The results of the proposed algorithm amply demonstrate its better performance as compared to other reported methods on CpG islands detection.

Keywords CpG islands (CGI) · DNA sequences · Numerical mapping · Short-time fourier transform (STFT)

1 Introduction

In the era of big data analysis, annotation and analysis of genomics data are highly needed to tackle current medical and societal problems. Genomics data contains deoxyribonucleic acid (DNA) sequences. The DNA sequences have four nucleotides: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). DNA sequences have the information about the protein-coding regions [1–5], tandem repeats [6–9], intron retentions [10], Helitrons [11], and CpG Islands (CGIs) [12] etc., which are useful in the genome annotation and associated with the biological functionalities of an organism. This study focuses on the CpG Islands (CGIs) present in the DNA sequences. CpG Islands are the regions in DNA sequences which consist of high-frequency CG dineucleotide as compared to the non-CGI regions. The '*p*' in CGI corresponds to the phosphodiester bond between C and G nucleotides [12]. CGIs act as a gene marker because these are useful to detect the first exonic regions, and promoter regions in DNA Sequences [13]. Also, the methylated CpG islands are associated with the important biological process like human malignancies, genome imprinting, X chromosome inactivation, aging, suppression of repetitive elements, and cancers. The Methylation is a process in which a methyl group (CH3) is added to the 5-position of the carbon in the pyrimidine ring of the cytosines of the CGI [14]. The first method for the identification of the CGIs in the DNA sequences has been developed by Gardiner-Garden and Frommer (GGF) [15], which is based on the following conditions:

(i) Length of CpG should be at least 200 bps,
(ii) Concentration C + G nucleotide should be minimum 50%,
(iii) Observed/Expected (O/E) ratio should be at least equal to 0.6.

Recently, various computational methods have been reviewed by Tahir et al. [12]. Some of the methods for the identification CGIs in DNA sequences have been discussed in this paper. These methods are CpG Cluster [16], IIR filter [17], FIR filter [18], Discrete Wavelet transform (DWT) [19], CpGcluster-TLBO [20] and CpGPNP [21]. Hackenberg et al. proposed a method in which the clustering has been done using the distance between CpG sites [16]. Vaidyanathan et al. developed an IIR filter-based method, in which the 40 filters have been used to calculate the weighted log score for the

✉ Pardeep Garg
garg.pardeep22@gmail.com

Sunildatt Sharma
sdsharma.juet@rediffmail.com

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

identification of the CGIs. The limitation of this method is its computational complexity due to the use of a large number of filters [17]. Rushdi and Tuqan proposed a method [18] in which Markov chain method and FIR filter together have been used for CGI detection. In this method, models for CGI and non-CGI have been developed and then FIR filter has been used to generate filtered likelihood ratio to detect the CGIs. Discrete Wavelet Transform (DWT) based CGI identification algorithm has been reported in [19]. It uses DWT along with adaptive filtering to identify CGIs. Park et al. proposed a sliding window-based method for CpG island detection [21]. Cheng et al. proposed a method CpGTLBO, in which the clustering method and teaching–learning-based optimization (TLBO) algorithm has been used. In this approach, clustering is used to detect the candidates CGIs and TLBO is used to optimize these candidates CGIs with respect to the actual CGIs [20]. In this paper, a short-time Fourier transform (STFT) based algorithm for CpG islands (CGIs) identification has been proposed. The performance of the proposed (STFT based) method has been compared with existing methods CpGTLBO, CpGPNP, and DWT based method. The remainder of the paper is organized as follows: materials and method have been explained in Sect. 2, data set and evaluation parameters have been described in Sect. 3, in Sect. 4 results have been discussed, and Sect. 5 presents the conclusion of the work.

2 Materials and Method

2.1 Periodicities in CpG Islands

It is reported that CpG islands are high-frequency recurring patterns of CG dineucleotide [18] in DNA sequences; therefore, we have considered small periodicities as a feature of CpG islands. To validate the periodicity feature first we converted the characters A, T, G, C into numeric sequences using integer mapping scheme [22] and then computed the STFT of all of the 17 CpG island sections present in the DNA sequence of L44140 [19, 23] individually. To compute STFT of the DNA sequence, DFT has been applied to get the power spectrum of windowed sequence with a sliding window approach [24]. The N-point DFT of a numeric sequence $x(n)$ at each nucleotide position “n” has been calculated as follows [24]:

$$X(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{\frac{-j2\pi nk}{N}}, \tag{1}$$

where, $w(n) = \left(1/\sigma\sqrt{2\pi}\right)\exp\left(-n^2/2\sigma^2\right)$, n is the length of Gaussian window, $\sigma = n/2\alpha$, α is a window shape parameter. In this work $n=210$, α is 2.5, FFT length $N=2520$, and $k=0\ldots N-1$ have been selected. Using (1), power spectrum of the windowed sequence is

$$S_1(k) = |X(k)|^2 \tag{2}$$

The value of power spectrum with respect to the periodicities i.e. frequency bins $k = N/p$ for periodicity $p = 2 - 10$ has been calculated from the windowed power spectrum $S_1(k)$ at each nucleotide position using the following equation

$$S(n,p) = S_1(n, N/p), \tag{3}$$

where n represents the nucleotide position at which window is centered and it varies from $n = 0 \ldots L$, where L is the full length of DNA sequence. Now, nucleotide position versus periodicity plots have been plotted for all 17 CpG islands segments and these are shown in Fig. 1.

The dominant periodicities from the nucleotide position-periodicity plots have been extracted using the following criterion:

- Minimum segment length should be twice of the periodicity.
- Minimum periodicity must be considered as dominating when the segments are overlapping.

Now, the segments of the detected dominant periodicities have been verified using two conditions of O/E ratio, and percentage of GC content as per GGF criterion. If the segments of detected periodicities satisfy these two GGF conditions for CpG Island then these are considered as verified dominating periodicities else rejected.

The detected and verified dominating periodicities in CpG islands segments of L44140 sequence have been summarized in Table 1.

From Table 1, it has been observed that periodicities 2–10 are associated with CpG islands. In the next section, an algorithm for CpG islands detection has been proposed using these verified dominating periodicities of the CpG islands.

2.2 Proposed Algorithm for CpG Island Detection

In this section, an algorithm has been proposed to identify the CpG islands in DNA sequences, which is based on dominant periodicities present in CpG islands. The flow chart of the proposed algorithm is shown in Fig. 2.

The DNA sequence having gene bank accession number L44140 is of Homo sapiens chromosome X region from filamin (FLN) gene to glucose-6-phosphate dehydrogenase gene. The length of this sequence is of 219,447 bp and it has been selected as an example sequence to describe the steps of the proposed algorithm, and these are described below

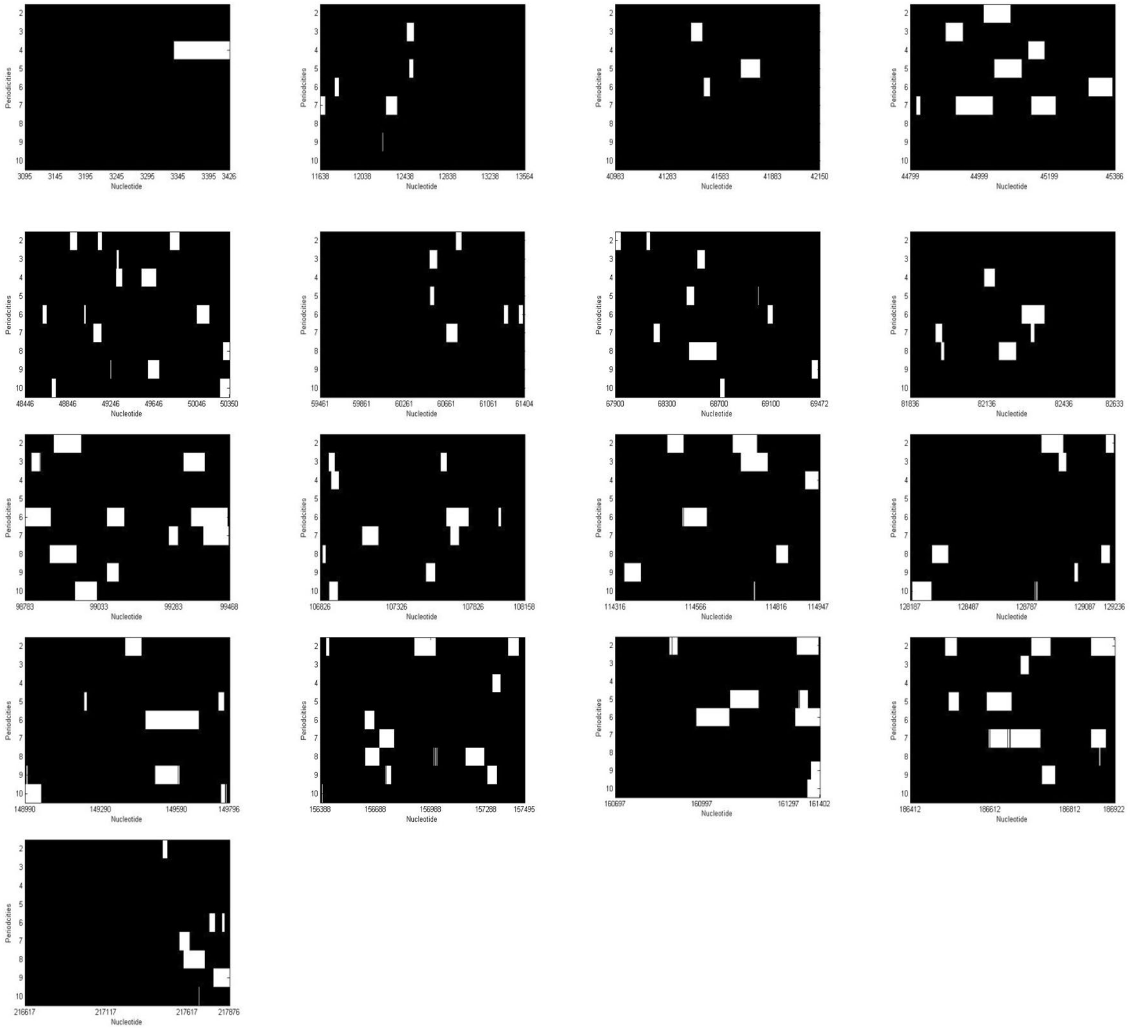

Fig. 1 Nucleotide position-periodicity plot of 17 CGIs of DNA sequence L44140

2.2.1 Numerical Mapping

The conversion of the character sequence into numeric sequence plays an important role in the digital signal processing (DSP) based algorithms to analyze the DNA sequences [22]. An example of integer numerical mapping scheme is shown in Table 2.

2.2.2 Calculate the Resultant Power Spectrum

By applying short-time discrete Fourier transform, we calculated the value of power spectrum components corresponding to each dominant periodicities i.e. periodicity 2–10 using Eq. 3. The power spectrums corresponding to dominating periodicities at each nucleotide position have been then combined linearly to get the resultant power spectrum for respective mapping scheme ‘a’. The resultant power spectrum $\mathrm{SR}_a(n)$ is calculated as follows

$$\mathrm{SR}_a(n) = \sum_{p=1}^{10} S(n,p) \tag{4}$$

The resultant power spectrum $\mathrm{SR}_a(n)$ has been plotted in Fig. 3.

2.2.3 Identify the Candidate CpG Islands

To identify the candidate CpG Islands from the resultant power spectrum, the 10% value of the maximum value of the resultant power spectrum $\mathrm{SR}_a(n)$ has been selected as a threshold empirically. The sections for which the peak value of the power spectrum is above the threshold have been considered as candidate CpG islands.

$$Y_a(n) = \begin{cases} \mathrm{SR}_a(n), \mathrm{SR}_a(n) > Th \\ \quad 0, \ \text{else}, \end{cases} \tag{5}$$

Table 1 Periodicities in CpG islands in L44140

S. No	Location of CGI as per NCBI website	Length of CGI (bps)	Detected periodicities in CpG Island Segments	Verified periodicities in CpG Island Segments
CGI 1	3095–3426	332	4	–
CGI 2	11,638–13,564	1927	3, 6, 7	3, 6
CGI 3	40,983–42,150	1168	3, 5, 6	3, 5, 6
CGI 4	44,799–45,386	588	2, 3, 4, 5, 6, 7	2, 3, 4, 5, 7
CGI 5	48,446–50,350	1905	2, 3, 4, 6, 8, 10	2, 3, 4, 6, 8, 10
CGI 6	59,461–61,404	1944	2, 3, 6, 7	3, 6, 7
CGI 7	67,900–69,472	1573	2, 3, 5, 6, 7, 9, 10	2
CGI 8	81,836–82,633	798	4, 6, 7, 8	4, 6
CGI 9	98,783–99,468	686	2, 3, 6, 7, 10	2, 3, 6, 7, 10
CGI 10	106,826–108,158	1333	3, 4, 6, 7, 8, 9	3, 6, 9
CGI 11	114,316–114,957	642	2, 3, 4, 6, 8, 9	2, 3, 4, 6, 8
CGI 12	128,187–129,236	1050	2, 3, 8, 9, 10	2, 3, 8
CGI 13	148,990–149,796	807	2, 5, 6, 10	2, 6, 10
CGI 14	156,388–157,495	1108	2, 4, 6, 7, 8	2, 6, 7, 8
CGI 15	160,697–161,402	706	2, 5, 6	2, 5, 6
CGI 16	186,412–186,922	511	2, 3, 5	2
CGI 17	216,617–217,876	1260	2, 6, 7	2, 6

where $Th = 0.1 \times \max(\mathrm{SR}_a(n))$.

$Y_a(n)$ has been plotted in Fig. 4 as a candidate CpG Island.

2.2.4 Verify the Candidate CpG Islands

The Segments corresponding to the detected candidate CpG Island have been verified using GGF criteria.

$$Z_a(n) = \begin{cases} Y_a(n), \text{if sections of } Y_a(n) \text{ satify GGF Criteria} \\ 0,\ \text{else} \end{cases} \tag{6}$$

The $Z_a(n)$ corresponding to the verified detected candidate CpG islands has been plotted in Fig. 5.

2.2.5 Combine Mapping Results

To select the appropriate mapping scheme, the performance of the proposed algorithm using 12 numerical mapping schemes has been compared in Table 3.

From Table 3, it has been observed that the Sn and AC of the proposed method for the 24 combinations of integer mapping are better as compared to other mapping schemes. Therefore, the final spectrum corresponding to CpG islands has been calculated by combining the verified spectrums of 24 mapping schemes and it is computed by the following equation

$$S_{CpG}(n) = \sum_{a=1}^{24} Z_a(n),\ a \in (1, 24) \tag{7}$$

The final spectrum corresponding to CpG islands for the proposed algorithm is shown in Fig. 6, where the horizontal axis represents the nucleotide position and the vertical axis represents the value of the power spectrum corresponding to nucleotide positions. To visualize the locations of detected CGIs, the final spectrum has been plotted in segments and these are plotted in Figs. 7, Fig. 8, Fig. 9, and Fig. 10.

The locations of CpG islands detected using the proposed algorithm have been shown in Table 4.

From Table 4, it has been clear that proposed algorithm identifies all 17 CpG islands present in the DNA sequence (acc. no. L44140) with some false positives. The performance of the proposed method has also been compared on the basis of the % coverage of the length of the true CpG Islands in Table 5.

In Table 5, it has been shown that the performance of the proposed algorithm is best amongst all methods with respect to percentage coverage of 80%, 90%, and 100% of the length of the true CpG Island.

3 Data Set and Evaluation Parameters

3.1 CpG Islands Data Set

To validate the performance of the proposed algorithm, we have made our own data set of CpG Islands of 100 DNA sequences for the species of human, mouse and fish [25]. The DNA sequence data set has been downloaded from the National Centre for Biotechnology Information (NCBI) [23].

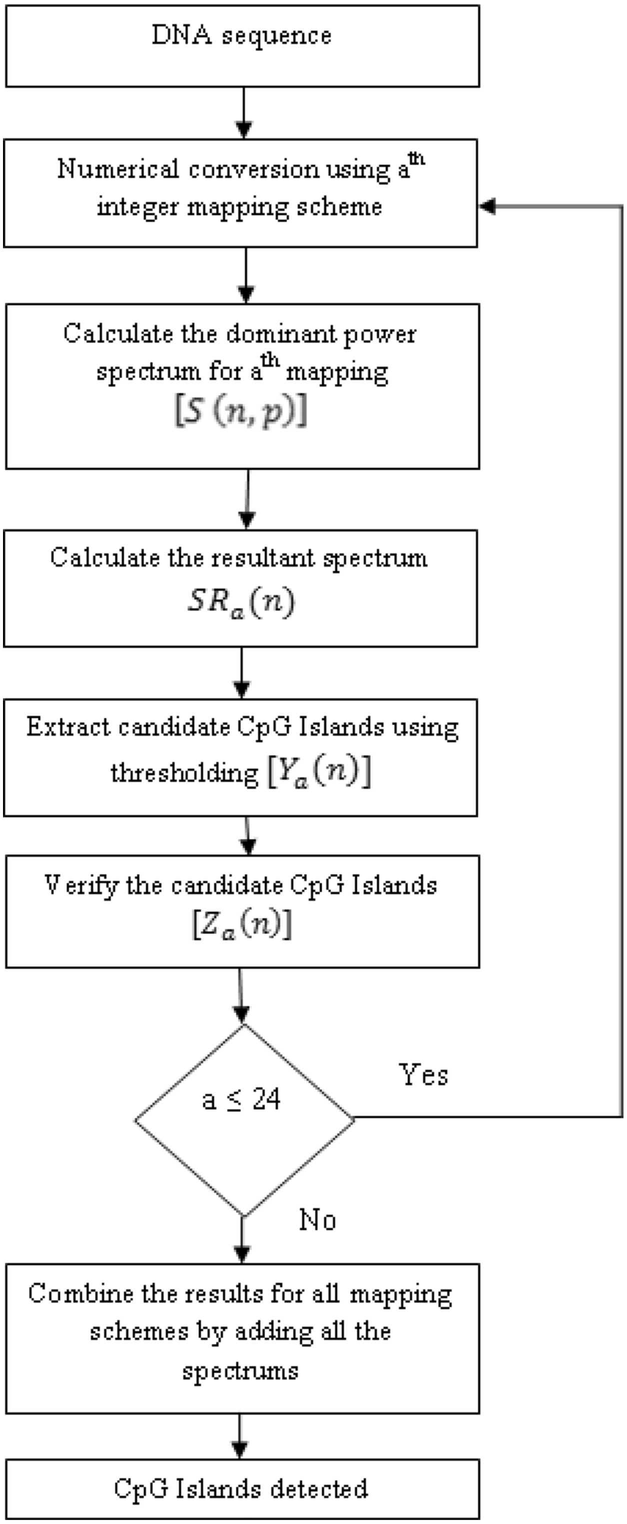
DNA sequence
Numerical conversion using a^{th} integer mapping scheme
Calculate the dominant power spectrum for a^{th} mapping $[S(n,p)]$
Calculate the resultant spectrum $SR_a(n)$
Extract candidate CpG Islands using thresholding $[Y_a(n)]$
V erify the candidate CpG Islands $[Z_a(n)]$
a ≤ 24
Yes
No
Combine the results for all mapping schemes by adding all the spectrums
CpG Islands detected

Fig. 2 Flow chart of the proposed algorithm

Table 2 Numerical Conversion

DNA Sequence	Numerical conversion using integer mapping
ATGCATG	[1432143 … …]

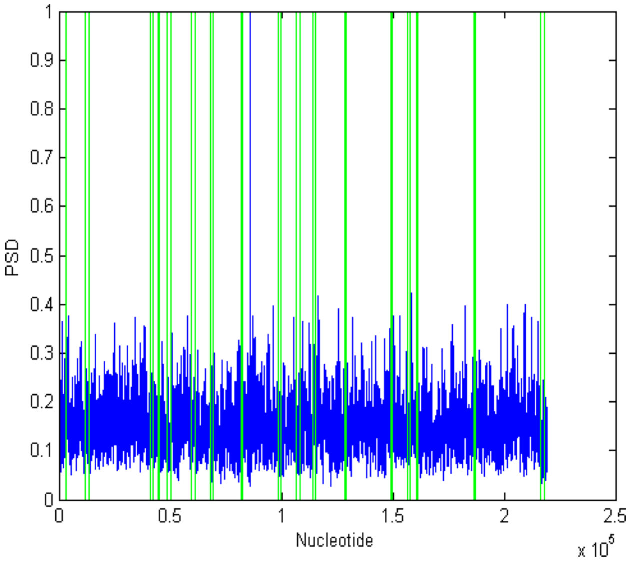

Fig. 3 Resultant power spectrum

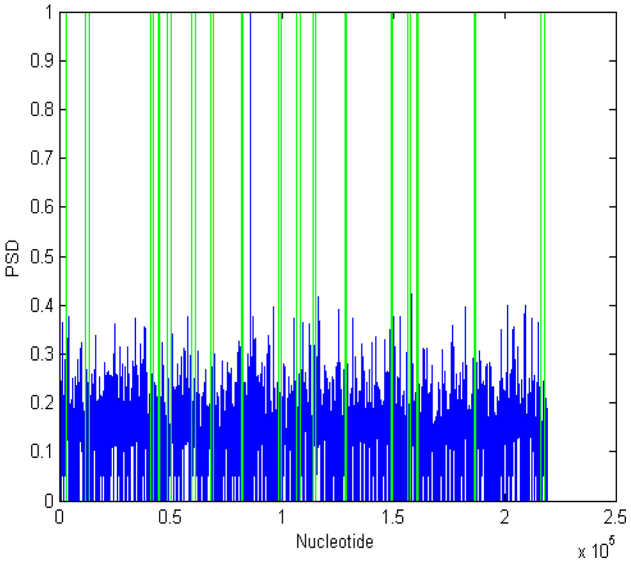

Fig. 4 Power spectrum corresponding to the candidate CpG islands

The detailed description of the data set with the accession number is shown in Table 6.

3.2 Evaluation Parameters

The performance analysis of the proposed algorithm has been carried out over the existing algorithms using the evaluation parameters, sensitivity (Sn), specificity (Sp), accuracy (AC) [3], and *F*-Measure [26]. These parameters are defined as follows:

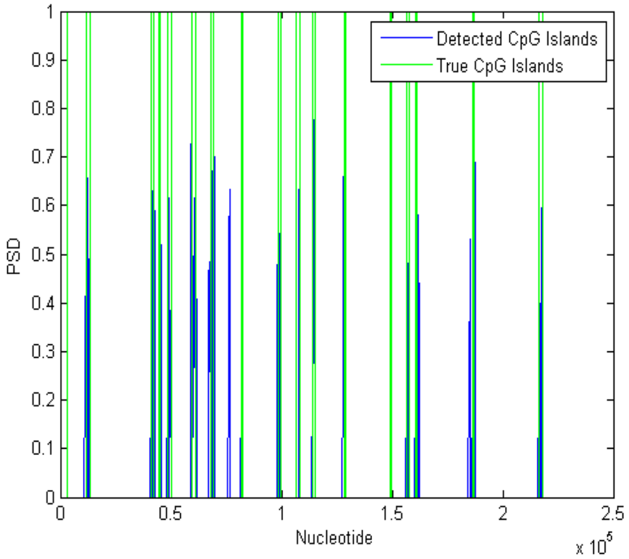

Fig. 5 Power spectrum corresponding to the verified candidate CpG islands

Table 3 Performance measures in L44140 using 12 Numerical Mappings

Mapping	Performance measure		
	Sn	Sp	AC
Atomic	0.0440	0.9767	0.5104
Complex	0.0295	1	0.5148
EIIP	0.4131	0.9538	0.6834
Fourbitbinary	0.0699	0.9888	0.5293
Threebitbinary	0.0154	0.9942	0.5048
Twobitbinary	0.5202	0.9618	0.7410
Integer	0.4758	0.9782	0.7270
Real Number	0.0336	0.9822	0.5079
Modified EIIP	0.5991	0.9492	0.7742
Molecular Mass	0.0440	0.9826	0.5133
Quaternary	0.4152	0.9689	0.6920
Pseudo EIIP	0.5464	0.9656	0.7560
Adding 24 mappings using integer mapping	**0.9590**	0.8285	**0.8938**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{9}$$

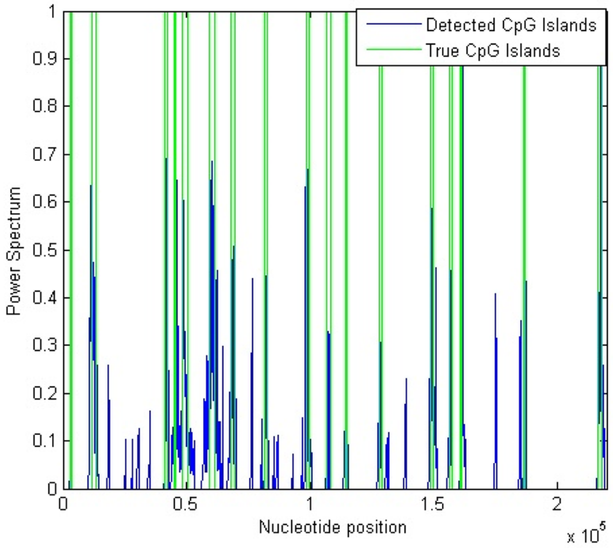

Fig. 6 Final power spectrum corresponding to the CpG islands

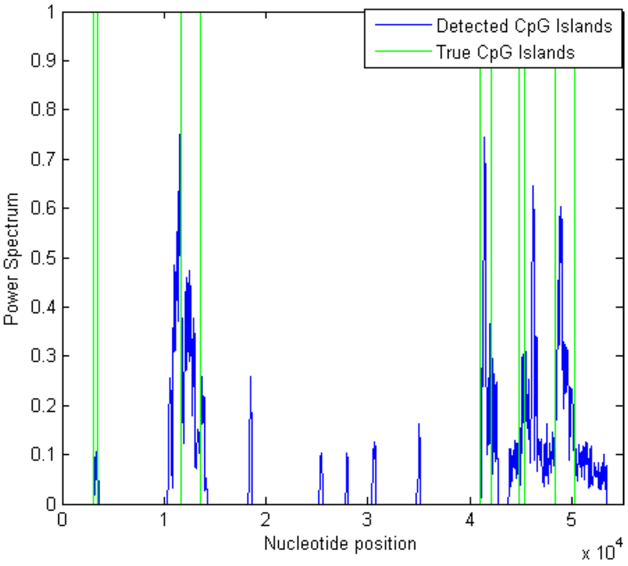

Fig. 7 Final power spectrum corresponding to the CpG islands (1–55,000 bps)

$$\mathrm{AC} = \frac{\mathrm{Sn} + \mathrm{Sp}}{2} \tag{10}$$

$$F\text{-measure} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \tag{11}$$

where

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},\ \text{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

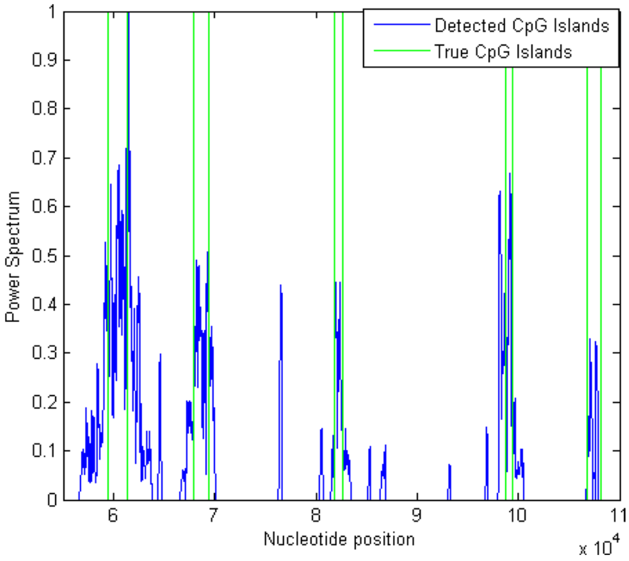

Fig. 8 Final power spectrum corresponding to the CpG islands (55,001–110,000 bps)

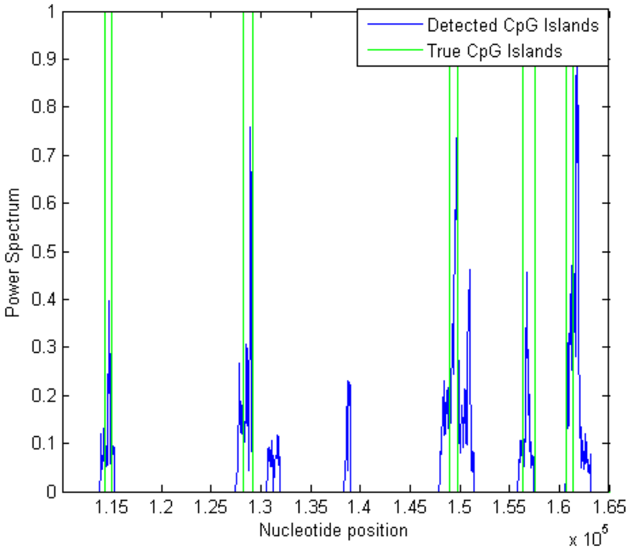

Fig. 9 Final power spectrum corresponding to the CpG islands (110,001–165,000 bps)

True positive (TP) is the correctly predicted locations where CpG islands are present, False positive (FP) is the falsely detected locations where CpG islands are not present, True negative (TN) is the correctly predicted locations where CpG islands are not present, and False negative (FN) is the missed locations where CpG islands are present. The value of both Sn and Sp vary between 0 and 1. The prediction result is considered to be perfect for the ideal case of value 1 of Sn and Sp. Accuracy which considers the effect of both Sn and Sp simultaneously has also been evaluated. The value of AC varies between 0 and 1. The *F*-measure is a measure of an algorithm's accuracy and it represents the harmonic mean of the precision and recall. It is used in place of receiver operating characteristics (ROC) if the analysis has been done on single threshold only. Its value lies between 0 and 1. For better performance, the value of *F*-measure should be 1.

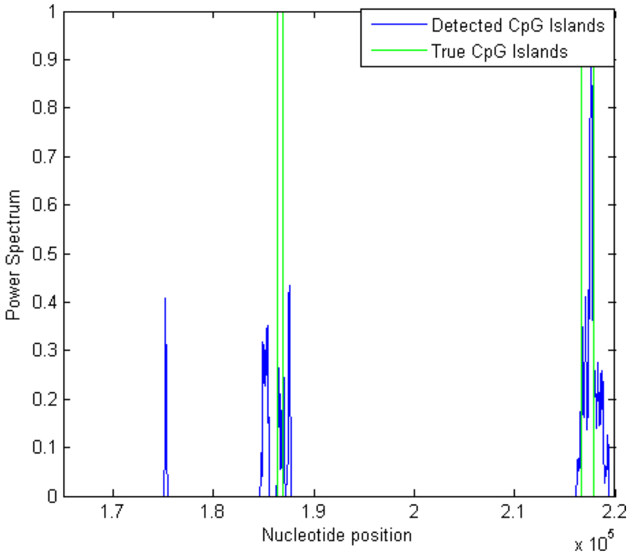

Fig. 10 Final power spectrum corresponding to the CpG islands (165,000–220,000 bps)

4 Results and Discussion

The performance of the proposed algorithm has been compared with the recently reported methods CpGclusterTLBO, DWT based algorithm and CpGPNP. The details of the performance parameters for human, fish and mouse species using CpGclusterTLBO, DWT, CpGPNP and proposed algorithm are shown in Tables 7, 8 and 9, respectively.

The performance of the proposed method for CpG island detection has been compared on the basis of the performance parameter Sn, Sp, AC, and *F*-measure with recently reported methods on the DNA sequences data set of human, fish and mouse. The comparison has been shown in Tables 7, Table 8, and Table 9. From this comparison, it has been found that the performance parameters Sn, Sp, AC, and *F*-measure of the proposed method are higher than other methods for human and mouse DNA Sequences whereas Sn is slightly less than the DWT based method for the DNA sequences of Fish. The performance of the proposed method for CpG island detection has also been compared with respect to the performance parameters Sn, Sp, AC, and *F*-measure on 100 DNA sequences of human, mouse and fish, and these are shown in Table 10.

Table 4 Detected CpG Islands

DNA sequence	True location of CpG Island		Locations detected by proposed algorithm	
	Start position	End position	Start position	End position
L44140				
1	3095	3426	3192	3576
2	11,638	13,564	10,470	14,217
			18,353	18,656
			25,277	25,597
			27,863	28,072
			30,464	30,766
			34,931	35,166
3	40,983	42,150	41,089	42,737
4	44,799	45,386	43,840	53,495
5	48,446	50,350	43,840	53,495
6	59,461	61,404	56,715	63,740
			64,457	64,720
			66,726	67,012
7	67,900	69,472	67,102	70,028
			76,336	76,687
			80,444	80,658
8	81,836	82,633	81,493	83,393
			85,176	85,394
			86,475	86,879
			93,080	93,286
			96,768	96,993
9	98,783	99,468	98,000	100,530
10	106,826	108,158	106,816	107,300
			107,345	107,583
			107,587	107,843
11	114,316	114,947	113,832	115,318
12	128,187	129,236	127,582	129,155
			130,652	131,218
			131,394	131,879
			138,508	139,016
13	148,990	149,796	147,981	151,460
14	156,388	157,495	155,887	157,400
15	160,697	161,402	160,653	163,220
			175,115	175,407
			184,658	185,511
16	186,412	186,922	186,327	187,110
			187,304	187,786
17	216,617	217,876	216,200	219,447

Table 5 Number of CGI detected in DNA sequence L44140

Methods	Number of CGIs detected at % coverage of true CGIs length		
	80%	90%	100%
CpGclusterTLBO	9/17	5/17	Nil
DWT	Nil	Nil	Nil
CpGPNP	4/17	3/17	2/17
Proposed algorithm	**15/17**	**15/17**	**12/17**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

It is observed from Table 10 that the performance of the proposed algorithm on 100 DNA sequences of human, mouse and fish is better in terms of Sn, Sp, AC and *F*-measure amongst all other recently reported methods.

The performance of the proposed method has also been compared using 100 DNA sequences of human, mouse and fish on the basis of the % coverage of the length of true CpG Islands in Table 11.

It has been observed from Table 11 that the performance of the proposed algorithm in the detection of CpG islands is the best amongst all methods. The 100 DNA sequences contain 194 CpG islands. Out of which the proposed algorithm has detected more number of CpG islands at 80%, 90%, and 100% portion coverage of the length of true CpG island as compared to existing methods.

5 Conclusion

In this paper STFT based algorithm for the identification of CpG islands has been studied. The algorithm has been tested on data set of 100 DNA sequences for human, mouse and fish. The performance of the proposed algorithm is better as compared to the reported methods in terms of Sn, Sp, AC, *F*-measure. The number of CGIs has also been detected at portion coverage of 80%, 90%, and 100% length of true CGIs and found that the proposed algorithm has identified more number of CGIs at portion coverage greater than 80%. Also, it has been studied that 24 combination of integer mapping schemes works better as compared to other mapping schemes. In future, the proposed algorithm for the CpG island detection in DNA sequences can be tested on non-human primates.

Table 6 Detailed description of the CpG islands data set

S. no	DNA sequence	Length of DNA sequence	Location of CpG Islands as per NCBI website	Number of CpG Islands	Gene Name/Gene ID
HUMAN species					
1	AL442638	188247	17,472–17,700, 22,868–23,148, 93,250–93,495, 163,847–164,132	4	LOC114827838/114827838
2	AC073335	68275	31,813–32,080, 33,619–34,458, 50,802–51,655	3	GTF2IP23/101929580
3	AC073517	67706	35,431–35,977	1	PRKRIP1/79706
4	AC127379	67291	30,060–30,318, 38,447–39,437	2	LRPPRC/10128
5	AC064843	66898	5531–5785	1	TRE-CTC7-1/100189491
6	AC129782	66860	38,868–40,898	1	BAC clone RP11-28O7
7	AC013270	66660	6075–6881, 25,374–26,035, 34,710–36,183, 48,185–48,621	4	ARID5A/10865
8	AC074386	66610	15,847–16,381, 16,593–16,830	2	OR2A1/346528
9	AC092103	66565	24,844–25,119	1	RNU6-1145P/ 106481541
10	AC124014	66552	56,936–57,769	1	IKZF1/10320
11	AL137791	66254	30,724–31,272, 46,196–46,906, 52,979–53,956, 61,007–62,096	4	Clone RP5-1079D1
12	AC096553	66229	11,867–12,256	1	PER3P1/168741
13	AC105413	65958	50,478–50,751	1	PTPN13/5783
14	AC005003	65750	38,374–41,067	1	PATZ1/23598
15	AC145546	65625	52,797–53,645	1	BAC clone RP11-1415P17
16	AC105402	65449	15,774–16,973, 28,628–28,925	2	KIF5C/3800
17	AC112698	65335	42,309–43,546	1	CDKN2AIP/55602
18	AC104129	65189	2966–3334, 8763–9020, 14,023–14,383, 20,695–20,991, 26,472–26,735, 28,330–29,188, 31,762–32,009, 55,671–55,878	8	MAD1L1/8379
19	BN000001	64961	895–1123	1	ELF3/1999
20	AC138782	64744	23,500–24,633	1	SEC24B/10427
21	AC005021	64607	24,663–25,225, 63,177–63,512	2	PON2/5445
22	AC093086	64601	58,914–59,518	1	CAMK2B/816
23	AC005233	64359	16,579–18,003	1	PAC clone RP5-1198O21
24	AC013436	63823	12,411–12,652, 21,066–21,331, 24,980–26,051, 26,467–26,807, 60,097–60,448	5	ZMIZ2/83637
25	AC131957	63780	45,526–45,799	1	BAC clone RP11-799G14
26	AC004694	63749	9107–9494, 54,481–54,756	2	BAC clone CTB-152H24
27	AC108463	63525	26,008–26,366, 26,575–26,982, 27,079–27,538	3	MIR4435-2HG/541471
28	AC080165	63279	8258–8531	1	LINC01789/105373536
29	AC010890	62764	11,407–11,926, 13,574–13,801, 53,142–53,415, 53,755–54,041	4	NCKAP5/344148
30	AC108142	62624	8864–11,837	1	TENM3/55714
31	AC080068	62623	535–774	1	LINC01162/104355138
32	AC093785	62466	31,397–31,665	1	LOC105373941/105373941
33	AC003079	62331	50,250–50,471	1	ASB4/51666
34	AC078937	62035	38,149–39,359	1	SLC26A4/5172
35	AC114803	61579	3256–4009, 18,815–19,353, 32,398–32,647, 33,247–33,659, 36,773–37,302, 39,696–39,964, 55,808–56,144	7	PTPRN/5798
36	AC093652	61340	48,156–49,072	1	FRAS1/80144
37	AC093377	61056	729–1003	1	ST13P2/344328
38	AC073201	60776	9738–11,862	1	BZW2/28969
39	AC113611	60597	8638–9514	1	HTRA3/94031
40	AC099394	60024	2826–4863, 10,806–11,866, 19,723–19,934, 25,482–25,769, 31,861–32,884, 36,728–36,931, 54,994–55,361	7	TFR2/7036
41	AC098831	59776	39,343–39,572, 51,406–51,689	2	ICA1L/130026

Table 6 (continued)

S. no	DNA sequence	Length of DNA sequence	Location of CpG Islands as per NCBI website	Number of CpG Islands	Gene Name/Gene ID
42	AC074013	59657	22,602–22,873, 51,602–52,508, 53,105–53,331	3	PUS7/54517
43	AC062028	59634	44,629–44,851	1	C2orf50/130813
44	AC106875	59580	4526–5382	1	LPIN1/23175
45	AC023670	59565	25,568–27,400	1	BAC clone RP11-457M7
46	AC079882	59427	39,153–39,736	1	RSPH10B2/728194
47	AC006008	57554	28,800–30,423	1	ACTR3C/653857
48	AC108222	21776	21,237–21,776	1	BAC clone RP11-1180N13
49	AH006464	21230	1187–2051	1	ATP12A/479
50	AC093609	20710	7857–8257	1	LINC02580/100506047
51	AL590794	18042	11,568–12,215	1	Clone RP11-148L13
52	AC136375	17863	16,369–17,534	1	BAC clone RP11-104P1
53	BD432859	14646	2762–2973, 4065–5181	2	TB7
54	AC111201	13470	4327–4727, 5323–5554, 12,500–13,455	3	ANO7/50636
55	NM005876	10782	6154–7734	1	SPEG/10290
56	NM053043	10168	9597–9820	1	RBM33
57	AC093460	10103	6951–7418	1	STARD3NL/83930
58	AC108032	9716	30–269	1	LOC101927687/101927687
59	X86012	9541	335–3853	1	F8A1/8263
60	AC106048	8594	7941–8180	1	SLC8A1/6546
61	AH008870	6797	341–1340	1	ICA1/3382
62	AC079401	6568	3086–3935	1	FAM3C/10447
63	AH007568	6513	543–803, 1212–1430, 1662–2474	3	CAV1/857
64	AC105385	5952	2844–3080	1	BAC clone RP11-115I2
65	AJ308559	5596	1228–1657	1	Plagl1/22634
66	M92844	3889	3198–3889	1	ZFP36/7538
67	AF196313	3700	2092–3580	1	ARHGAP26/23092
68	AF281043	3662	1611–2734	1	HMGB1/3146
69	U48937	3278	2588–3230	1	SCNN1G/6340
70	AF307776	3113	2334–2745, 2791–3064	2	ADRB1/153
71	AJ000757	3046	650–2840	1	GLI3/2737
72	AJ289875	2916	2325–2916	1	PRNP/5621
73	L07287	2704	1–1350	1	RPL26/6154
74	Z92546	73511	20,746–21,240	1	CABIN1/23523
75	AL591222	147211	54,605–55,080, 68,825–69,091	2	SLC24A2/25769
76	AL513502	174636	116,364–117,432	1	ADAMTSL1/92949
77	AL513498	155780	18,305–18,582	1	MLLT3/4300
78	AL357615	171446	56,753–57,030, 59,607–59,874	2	LOC107987055/107987055
79	AL353786	139565	19,000–19,400	1	SPATA7/55812
80	AL121926	139544	102,641–104,201, 126,562–127,299	2	CSTF3/1479
81	AL049547	129811	27,801–29,311, 37,094–37,773, 109,041–110,125, 113,196–114,024, 126,815–127,265	5	TNXB/7148
82	AL031706	13012	7–552	1	Clone LA16-305F3
83	AL031703	35098	15,319–17,699, 25,107–26,048, 30,327–30,736, 31,615–32,204	4	CACNA1H/8912
84	AJ006998	123521	11,140–11,417	1	LOC101927745/101927745
85	AL031707	28707	6050–6520, 6693–7445, 24,481–25,248, 28,059–28,669	4	Clone LA16c-313F9
FISH species					
86	AL603785	89874	4151–4634	1	Musk/334526

Table 6 (continued)

S. no	DNA sequence	Length of DNA sequence	Location of CpG Islands as per NCBI website	Number of CpG Islands	Gene Name/Gene ID
87	AL672065	82767	44,999–45,681	1	Rsu1/553276
88	AL672083	111516	88,040–88,588	1	Pknox1.2/170445
89	AL691521	109831	34,191–36,572	1	Men1/30130
90	AL672171	114103	50,521–51,167	1	clone BUSM1-270G24
91	AL713869	104577	6954–7435	1	Si:busm1-105l16.2/368709
MOUSE species					
92	AJ970309	7050	3025–4010	1	Apaf1/11783
93	AC149868	190971	38,226–39,751, 109,499–110,391, 114,105–114,977, 167,115–168,150	4	Slc17a7/72961
94	AC125063	194931	97,498–98,367, 99,058–100,402, 106,255–107,246, 144,134–145,047	4	Pilra/231805
95	AC124505	222439	36,111–37,119, 132,685–133,458, 139,610–140,565, 202,532–203,418	4	Mapk3/26417
96	AC145199	220892	29,996–30,867, 59,938–60,771, 114,341–115,758, 133,121–133,903, 204,198–205,934, 217,247–218,028	6	Dmpk/13400
97	AC122821	220013	43,295–44,322, 59,514–60,693, 122,943–123,697, 163,194–164,078, 185,979–186,978, 218,075–218,923	6	Srrm2/75956
98	AF073797	46872	9395–9666, 18,386–18,651, 32,350–32,477, 33,946–34,206	4	Aire/11634
99	AC126029	212472	5851–6810, 75,564–76,663, 82,722–84,043, 152,561–153,650, 195,134–196,503	5	Rela/19697
100	AF059580	36326	2076–3209, 2382–3017, 14,983–15,869	3	Zdhhc7/102193

Table 7 Performance comparison for 85 DNA sequences of human

Performance parameter	Methods			
	CpGcluster TLBO	DWT	CpGPNP	Proposed algorithm
TP	71,218	65,822	66,048	**78,338**
FP	136,172	2,814,422	228,024	**130,623**
TN	4,444,891	1,772,242	4,358,640	**4,456,041**
FN	27,735	37,938	37,709	**25,419**
Sn	0.7197	0.6344	0.6366	**0.7550**
Sp	0.9702	0.3864	0.9503	**0.9715**
Ac	0.8449	0.5104	0.7934	**0.8632**
F-measure	0.4650	0.0441	0.3320	**0.5010**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

Table 8 Performance comparison for 6 DNA sequences of fish

Performance parameter	Methods			
	CpGcluster TLBO	DWT	CpGPNP	Proposed algorithm
TP	2763	3555	3181	**3496**
FP	27,673	370,842	31,308	**12,020**
TN	579,762	236,594	576,127	**595,415**
FN	2464	1672	2046	**1731**
Sn	0.53	0.68	0.61	0.67
Sp	0.954	0.389	0.948	**0.98**
Ac	0.742	0.535	0.779	**0.825**
F-measure	0.1550	0.0187	0.1602	**0.3371**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

Table 9 Performance comparison for 9 DNA sequences of mouse

Performance parameter	Methods			
	CpGcluster TLBO	DWT	CpGPNP	Proposed algorithm
TP	25,985	17,192	11,155	**30,434**
FP	57,015	614,844	107,332	**55,750**
TN	1,260,968	703,139	1,210,651	**1,262,233**
FN	7989	16,782	22,819	**3540**
Sn	0.765	0.506	0.328	**0.896**
Sp	0.957	0.533	0.919	**0.958**
Ac	0.861	0.52	0.624	**0.927**
F-measure	0.4443	0.0516	0.1463	**0.5066**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

Table 10 Performance comparison for 100 DNA sequences

Performance parameter	Methods			
	CpGcluster TLBO	DWT	CpGPNP	Proposed algorithm
TP	99,966	86,569	80,384	**112,268**
FP	220,860	3,800,108	366,664	**198,393**
TN	6,285,621	2,711,975	6,145,418	**6,313,689**
FN	38,188	56,392	62,574	**30,690**
Sn	0.7236	0.6055	0.5623	**0.7853**
Sp	0.9661	0.4165	0.9437	**0.9695**
Ac	0.8448	0.5110	0.7530	**0.8774**
F-measure	0.4356	0.0430	0.2725	**0.4950**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

Table 11 Number of CGI detected

Methods	Number of CGIs detected at % coverage of true CGIs length		
	80%	90%	100%
CpGclusterTLBO	108/194	76/194	50/194
DWT	1/194	Nil	Nil
CpGPNP	60/194	46/194	39/194
Proposed algorithm	**112/194**	**101/194**	**93/194**

The bold values represent that the performance of the proposed algorithm is better as compared to the CpGclusterTLBO, CpGPNP, DWT based methods in terms of respective parameters

Compliance with ethical standards

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest in this work.

References

1. Shakya DK, Saxena R, Sharma SN (2013) An adaptive window length strategy for eukaryotic CDS prediction. IEEE/ACM Trans Comput Biol Bioinf 10(5):1241–1252. https://doi.org/10.1109/TCBB.2013.76
2. Meher JK, Panigrahi MR, Dash GN, Meher PK (2012) Wavelet based lossless DNA sequence compression for faster detection of eukaryotic protein coding regions. I. J Image, Graph Signal Process 4(7):47–53. https://doi.org/10.5815/ijigsp.2012.0.7.05
3. Das L, Nanda S, Das JK (2019) An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser window. Genomics 111(3):284–296. https://doi.org/10.1016/j.ygeno.2018.10.008
4. Das L, Das JK, Nanda S (2017) Identification of Exon location applying Kaiser window and DFT techniques. In: 2nd Conf for convergence in technology, pp. 211–216, DOI: 10.1109/I2CT.2017.8226123.
5. Das L, Nanda S, Das JK, (2017) A novel DNA mapping scheme for improved exon prediction using digital filters. In: 2nd Int Conf on man and machine interfacing, pp 1–6, 10.1109/MAMI.2017.8307889.
6. Sharma SD, Saxena R, Sharma SN (2015) Identification of microsatellites in DNA using adaptive S- transform. IEEE J Biomed Health Inf 19(3):1097–1105. https://doi.org/10.1109/JBHI.2014.2330901
7. Sharma SD, Saxena R, Sharma SN (2015) Short tandem repeats detection in DNA sequences using modified S-transform. Int J Adv Eng Technol 8(2):233–245
8. Sharma SD, Saxena R, Sharma SN (2017) Tandem repeats detection in DNA sequences using Kaiser window based adaptive S-transform. Bio-Algorithms Med Syst 13(3):167–173. https://doi.org/10.1515/bams-2017-0014
9. Garg P, Sharma SD, Sharma SN (2017) Tandem repeats detection in DNA sequences using P-spectrum based algorithm. In: Conference on Information and Communication Technology (CICT 2017), 2017 IEEE International Conference, pp. 1–5, 10.1109/INFOCOMTECH.2017.8340621.
10. Sharma SD, Sharma SN, Saxena R (2019) Identification of Short Exons Disunited by a Short Intron in Eukaryotic DNA Regions. IEEE/ACM Trans Compu Biol Bioinform. https://doi.org/10.1109/TCBB.2019.2900040
11. Touati R, Messaoudi I, Oueslati AE, Lachiri Z (2018) A combined support vector machine- FCGS classification based on the wavelet transform for Helitrons recognition in C. elegans. Multim Tools Appl 78:13047–13066. https://doi.org/10.1007/s11042-018-6455-x
12. Tahir RA, Zheng D, Nazir A, Qing H (2019) A review of computational algorithms for CpG islands detection. Indian Acad Sci 44(143):1–11. https://doi.org/10.1007/s12038-019-9961-8
13. Wang Y, Leung F (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20(7):1170–1177. https://doi.org/10.1093/bioinformatics/bth059
14. Feng P, Chen W, Lin H (2014) Prediction of CpG island methylation status by integrating DNA physicochemical properties. Genomics 104(4):229–233. https://doi.org/10.1016/j.ygeno.2014.08.011
15. Garden MG, Frommer M (1987) CpG Islands in vertebrate genomes. J Mol Biol 196(2):261–282. https://doi.org/10.1016/0022-2836(87)90689-9
16. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martinez-Aroza J, Oliver JL (2006) CpGcluster: a distance-based algorithm for CpG-Island detection. BMC Bioinform 7:446. https://doi.org/10.1186/1471-2105-7-446
17. Yoon B, Vaidyanathan P (2004) Identification of CpG Islands using a bank of IIR low-pass filters. Digital Signal Process Workshop. https://doi.org/10.1109/DSPWS.2004.1437966

18. Rushdi A, Tuqan J (2006) A new DSP-based measure for CpG Islands detection. In: 12th IEEE Signal Processing Education Workshop, pp. 561–565, 10.1109/DSPWS.2006.265486.
19. Mariapushpam IT, Rajagopal S (2017) Improved algorithm for the location of CpG Islands in genomic sequences using discrete Wavelet transforms. Curr Bioinform 12:57–65. https://doi.org/10.2174/1574893611666160805111825
20. Yang CH, Chiang YC, Chuang LY, Lin YD (2017) A CpGCluster-teaching-learning-based optimization for prediction of CpG Islands in the human genome. J Comput Biol 24:1–12. https://doi.org/10.1089/cmb.2016.0178
21. Park HC, Ahn ER, Jung JY, Park JH, Lee JW, Lim SK, Kim W (2018) Enhanced sensitivity of CpG Island search and primer design based on predicted CpG Island position. Forensic Sci Int Genet 34:134–140. https://doi.org/10.1016/j.fsigen.2018.02.013
22. Sharma SD, Shakya DK, Sharma SN (2011) Evaluation of DNA mapping schemes for exon detection. In: IEEE International Conference, pp. 71–74, 10.1109/ICCCET.2011.5762441.
23. National Centre for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/. Accessed 15 June 2019.
24. Akhtar M, Epps J, Ambikairajah E (2008) Signal processing in sequence analysis: advances in eukaryotic gene prediction. IEEE J Select Topics Signal Process 2(3):310–321. https://doi.org/10.1109/JSTSP.2008.923854
25. Barazandeh A, Mohammadabadi MR, Ghaderi-Zefrehei M, Nezamabadipour H (2016) Predicting CpG Islands and Their relationship with genomic feature in cattle by hidden Markov Model Algorithm. Iran J Applied Anim Sci 6(3):571–579
26. Touati R, Oueslati AE, Messaoudi I, Lachiri Z (2019) The Helitron family classification using SVM based on Fourier transform features applied on an unbalanced dataset. Med Biol Eng Comput 57:2289–2304. https://doi.org/10.1007/s11517-019-02027-5