International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Mar-2022)

http://dx.doi.org/10.12785/ijcds/110188

CpG Islands Detection in Human DNA Sequences using Wavelet Transform

Pardeep Garg[1] **and Sunil Datt Sharma**[2]

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

Received 19 Apr. 2021, Revised 23 Feb. 2022, Accepted 9 Mar. 2022, Published 31 Mar. 2022

Abstract: In the era of the big data analysis, genomic signal processing (GSP) is gaining popularity to analyze the genomics data. GSP is used to extract the useful or hidden information from the genomics data such as DNA sequences using digital signal processing tools. The hidden information is closely associated with the different biological functionalities in the living organisms. CpG Island is one of such hidden information in DNA sequences, which is associated with the gene silencing, cancers and many other epigenetic diseases. Therefore, the extraction of the information about the CpG islands is highly needed to serve the people. So, in this work an algorithm based on wavelet transform has been proposed to overcome the fixed window length limitation of short-time Fourier transform based method for the detection of CpG Islands. The performance assessment of proposed method has been carried out on hundred DNA sequences of human species and detection performance has been improved over other state of art methods in terms of sensitivity, accuracy, and F-measure.

Keywords: CpG Islands (CGIs), DNA Sequences, Wavelet Transform, Numerical Representation

1. Introduction

The completion of human genome sequencing project provided tremendous opportunities for researchers working in the area of genomic signal processing, big data analysis, and bioinformatics. The genomics data such as deoxyribonucleic acid (DNA) sequences include a lot of concealed information which needs to be analyzed to extract important biological information. DNA sequences are made up of four nucleotides: thymine (T), guanine (G), cytosine (C), and adenine (A). It is reported in literature that in DNA sequences different patterns are present such as three-base periodicity based protein coding region [1], [2], [3], [4], [5], [6], [7], tandem repeats [8], [9], [10], [11], introns retention [12], Helitrons [13], [14], splice sites [15], CpG islands (CGI) [16] and many more. In this paper the emphasis is on CpG islands detection using a signal processing based algorithm. CGI regions in DNA sequences are those segments which have high frequency CG dineucleotide as opposed to other regions which are considered as non CGIs [16]. It has been reported in literature that various biological processes are associated with CGIs which make the detection of CGIs in DNA sequences essential [17]. It is reported that CGIs are associated with promoter regions and hence these find application in the identification of the promoter regions and consequently to predict the genes in DNA sequences [18]. Also, gene silencing, cancers and many other epigenetic issues [19] are caused by the process of methylation of CGIs which happens by the addition of methyl group (CH_3) to the 5-position of the carbon. These are some of the reasons which make the detection of CGIs in DNA sequences necessary and therefore various algorithms have been proposed so far and are reported in literature which is discussed in detail in section 2 of the paper. The organization of the rest of the paper is as follows: in Section 2 detailed discussion of related work has been presented, materials and methods have been discussed in Section 3, description of data set and evaluation parameters has been given in Section 4, in Section 5 results have been discussed, and the paper has been concluded in Section 6.

2. Related Work

It is known that the results provided by the biologists for CGI detection obtained using experimental methods are accurate but these methods are highly time consuming because of vast amount of genomic data [20]. But the computational methods developed by researchers for the detection of CGIs are effective and efficient [21]. The first computational method for CGI detection was proposed by Gardiner-Garden and Frommer (GGF) [22], according which a particular DNA segment is termed as CGI if it

E-mail address: garg.pardeep22@gmail.com, sdsharma.juet@rediffmail.com

satisfies the following three conditions: (i) minimum 200 nucleotides are contained in the segment (ii) Concentration of C+G nucleotide should be at least equal to 50 %, and (iii) minimum value required for observed/expected (O/E) ratio is 0.6. Later the ensuing method developed by Takai and Jones [23] gave more firm conditions for a DNA segment to be classified as CGI.

Recently, Tahir *et al.* [16] reviewed various computational methods of CGI detection. It is reported that the computational algorithms for CGI identification are classified as window based, Hidden Markov Model (HMM) based, density based, and distance-/length based algorithms [16], [24]. In window based methods, a moving window is applied to examine the genome using predefined statistical conditions of CGI. Some of the methods developed based on this approach are discussed in [23], [25], [26], [27]. These window based methods are very much used because these strictly follow the given statistical parameters for classification of a section of DNA as CGI. But these methods have a major limitation in terms of their dependency on window size which plays a significant role in correctly prediction of CGI. The larger window size has the advantage of increase in predictive granularity but computationally slower. Whereas the smaller window size is computationally faster but has the drawback of probably missing a potential CGI [16], [24].

Hidden Markov model based CGI detection methods are discussed in [20], [28], [29], [30]. These HMM based methods utilize two separate models based on Markov chains for CGI and non CGI and then compute log-score of the sequences for the two models. These methods are basically data dependent as the transition probability tables vary according to data and also these are computationally inefficient [16], [24].

The principle of density based CGI detection methods is to find out the density of CpG sites [31], [32]. In these methods, the ratio of number of CpG sites in CGI and the total span of CGI is calculated to compute the density of CGI. The basic operation of density based methods is initialization of low threshold value of density to capture the approximate boundary of CGI and then subsequently a high threshold value is applied to finally capture the CGI borders where the DNA sequence within that border satisfies the density requirement. The dependency on the thresholds of density is considered as a major limitation of these density based methods [16], [24].

The distance-/length based approach of CGI detection is discussed in [33] and is considered as a faster approach for prediction of CGI. This approach is basically formulated on the clustering of data according to the distance between CpG sites. This method provided a new direction for the understanding of CGI by studying the sequence property of any two adjoining CpG sites. The authors criticized this method because of its dependency on sequence composition which results in dissimilar results for same CGI in different circumstances [16], [24]. A method called CpGclusterTLBO has been developed by Cheng *et al.* in which the clustering approach and teaching-learning-based optimization (TLBO) algorithm has been used. In this method, the use of clustering is to identify the probable CGIs and TLBO has been used for the optimization of probable CGIs with respect to the actual CGIs [34].

Currently, digital signal processing based CGI detection methods have also been developed [35], [36], [37], [38].A method has been developed by Rushdi and Tuqan [35]in which FIR filter and Markov chain method altogether are used for CGI detection. In this method, two different models have been developed out of which one model is for CGI another model is for non CGI; and then filtered likelihood ratio test measure is generated with the help of FIR filter. Mariapushpam *et al.* proposed discrete Wavelet transform (DWT) based CGI identification algorithm [36]. In this algorithm DWT based filtering along with adaptive filtering has been utilized to identify CGI. Recently, an algorithm has been proposed in which modified P-spectrum has been employed for the identification of CGIs in the DNA sequences [37].

Short-time Fourier transform (STFT) based CGI detection algorithm has been presented, in which the spectrums of the dominant periodicities have been utilized to detect the CGIs [38]. As it has been known that STFT based algorithm's performance may suffer because in STFT fixed window length criteria has been utilized. Therefore, to avoid the problem of fixed window length, an algorithm based upon wavelet transform has been proposed for the identification of the CGIs. In the proposed algorithm, spectrums corresponding to the dominating periodicities present in the CGIs have been calculated using wavelet transform. The sum of these spectrums has been calculated to find the resultant spectrum of the CGIs. An appropriate threshold has been selected to get the resultant spectrum of candidate CGIs, and then it has been verified using GGF criteria to remove the falsely detected spectrum of candidate CGIs. The verified resultant spectrums of candidate CGIs have been calculated for 24 combinations of integer mappings. Finally, the sum of these 24 verified spectrums of candidate CGIs has been calculated to get the final spectrum of CGIs. The key contributions of the proposed algorithm are:

i) Wavelet transform has been used to overcome the fixed window length limitation of the STFT,

ii) Selection of optimal threshold,

iii) Detection performance has been improved.

The proposed algorithm has been tested on the data set of 100 human DNA sequences. The performance assessment of proposed algorithm has been done with state of art CGI detection algorithms. The results specify that the approach proposed in this paper is better than the other

reported algorithms.

3. Materials and Methods

A. Characteristic Feature in CpG Islands

Characteristic features associated with CGIs in DNA sequences have already been reported in [38] as dominant periodicities of CGIs and these periodicities have been utilized in this paper to detect the CGIs in human DNA sequences.

B. Proposed Algorithm for CpG Island Detection

The algorithm for the detection of CpG Islands using wavelet transform is represented in Table I.

The major steps of the proposed algorithm are described in detail using following points:

1. Conversion of A, T, C, G characters of DNA sequence to numerical values.

2. Calculate the spectrums corresponding to the dominant periodicities 2 to 10 using wavelet transform.

3. Compute the sum of the spectrums of the dominant periodicities to get the resultant spectrum.

4. Select an appropriate threshold to get the resultant spectrum of candidate CGIs.

5. GGF criteria have been used to verify the resultant spectrum of candidate CGIs and to remove the falsely detected spectrum of candidate CGIs.

6. Combine the 24 verified resultant spectrum of candidate CGIs to compute the final spectrum of CGIs.

Above steps of the proposed algorithm have been explained below with the help of a benchmark DNA sequence having accession number L44140 [39] and this sequence has been considered as an example DNA sequence:

1. Numerical Conversion

The conversion of four characters of DNA sequence into numerical values has to be performed for the digital signal processing techniques to be applied. In this work, the A, T, G, C characters of DNA are converted to numerical values using all 24 representations of integer mapping scheme [38] to avoid the bias due to mapping. One of the representation of integer mapping scheme has been shown which assigns the numerical values to DNA characters as A = 1, T = 4, G = 3, C = 2. The representation of 24 mappings of integer mapping to the DNA characters is depicted in Table II.

2. Modified Gabor Wavelet Transform (MGWT)

As it has been already reported that CpG islands are associated with periodicities 2-10 base pairs (bps) [38]. So, in this work the Gabor wavelet based transform has been tuned to identify the spectrums corresponding to periodicities 2-10 bps; and this transform is called as modified Gabor wavelet transform (MGWT). And it is calculated for a numeric sequence z(x) using (1):

$$Z(n,b)_p = \int z(x)\, e^{-\frac{(x-n)^2}{2b^2}}\, e^{j\omega_0(x-n)} dx \tag{1}$$

(1) has been used to the capture the spectrums of different periodicity, the value of w_O = L/p has been fixed for the detection of periodicity "p" component, where L is considered as the length of the DNA segment which has to be analyzed and 2 to 10 values of periodicity have been considered of variable p. To obtain the spectrum of the sequence, squared complex modulus of the MGWT coefficients has been calculated as:

$$M(n,p)_p = \left|Z(n,b)_p\right|^2 \tag{2}$$

In the work proposed in this paper, 40 analyzing functions corresponding to 40 scale values have been used and these are exponentially separated between 0.1 and 0.7 for each p-periodic periodicity. The spectrums obtained corresponding to periodicity 2 to 10 have been added linearly to compute the resulting spectrum RM_m(n) employing corresponding mapping scheme 'm'.

$$RM_m(n) = \sum_{p=2}^{10} M(n,p) \tag{3}$$

RM_m(n) for example DNA sequence L44140 has been shown in Fig. 1.

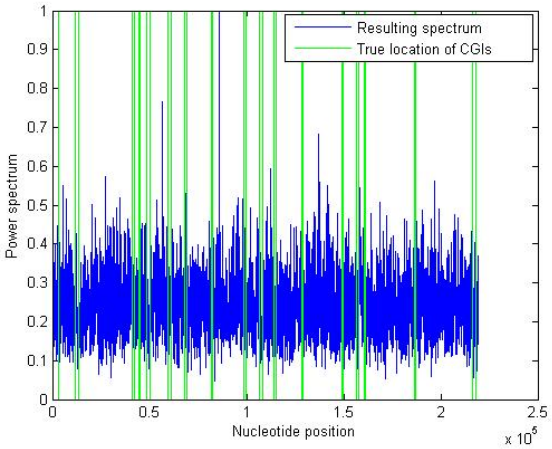

Figure 1. Resulting spectrum RM_m(n)

3. Thresholding

A suitable threshold value has been chosen experimentally to get the spectrum of candidate CGIs from resulting spectrum. The experiment has been conducted for DNA

TABLE I. Wavelet transform based algorithm for CpG islands detection

Input: DNA sequence
1) For nr = 1:24

2) For periodicities = 2:10,

Calculate spectrums of periodicities using wavelet transform.

End (loop end for periodicities).

Calculate the addition of spectrums of periodicities.

Apply suitable thresholding to select the candidate CpG Islands.

Apply the GGF criteria to verify the CpG islands.
3) Store the final spectrum for each nrth iteration.

4) End (loop end for nr) and calculate sum of final spectrums of all 24 iterations.

Output: CpG Islands are detected

TABLE II. 24 combinations of integer mapping

Numeric values to DNA characters				
	A	C	G	T
m=1	1	2	3	4
m=2	1	3	4	2
m=3	1	4	2	3
m=4	1	2	4	3
m=5	1	3	2	4
m=6	1	4	3	2
m=7	2	3	4	1
m=8	2	4	1	3
m=9	2	1	3	4
m=10	2	3	1	4
m=11	2	4	3	1
m=12	2	1	4	3
m=13	3	1	2	4
m=14	3	2	4	1
m=15	3	4	1	2
m=16	3	1	4	2
m=17	3	2	1	4
m=18	3	4	2	1
m=19	4	1	2	3
m=20	4	3	1	2
m=21	4	2	3	1
m=22	4	1	3	2
m=23	4	3	2	1
m=24	4	2	1	3

sequence L44140 considered as an example sequence by varying the threshold values from 10 % to 50 % is depicted in Table III.

The proposed algorithm's performance for example DNA sequence L44140 at threshold value 15 % is better compared to other threshold values in terms of Sn, AC and it has been observed from Table III. Hence, in this paper the threshold value has been selected as 15 % for all analysis work of the proposed MGWT based CGI detection algorithm.

The sections of the spectrum where the peak value is above the threshold value of 15 % have been then chosen as candidate CGIs.

$$Q_m(n) = \begin{cases} RM_m(n), & RM_m(n) > Thr \\ 0, & else \end{cases} \quad (4)$$

where Thr= 15% of max (RM_m(n))

Q_m(n) is the spectrum of the candidate CGIs, and for example DNA sequence L44140 it has been shown in Fig. 2.

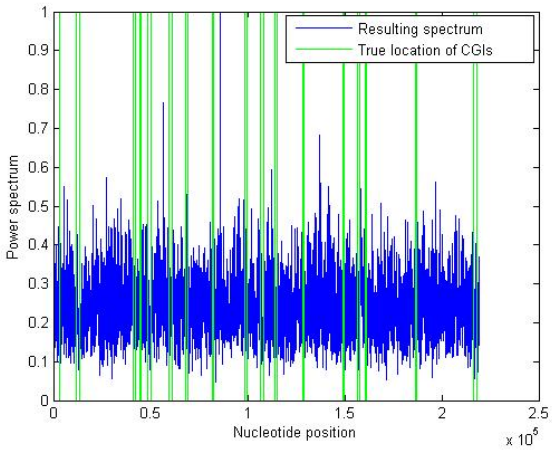

Figure 2. Candidate CpG Island's spectrum

4. Verification of Candidate CpG Islands

The GGF criterion has been applied to the respective segments of the corresponding spectrum of the candidate CGIs, to get verified spectrum of the candidate CGIs. It is also used to reduce the falsely detected spectrum of the candidate CGIs. Verified spectrum of the candidate CGIs has been calculated using (5):

$$V_m(n) = \begin{cases} Q_m(n), & Seg.of\ Q_m(n)\ meeting\ GGF\ Criteria \\ 0, & else \end{cases} \quad (5)$$

V_m(n) is the verified spectrum of the candidate CGIs, and for example DNA sequence L44140 it has been shown in Fig. 3.

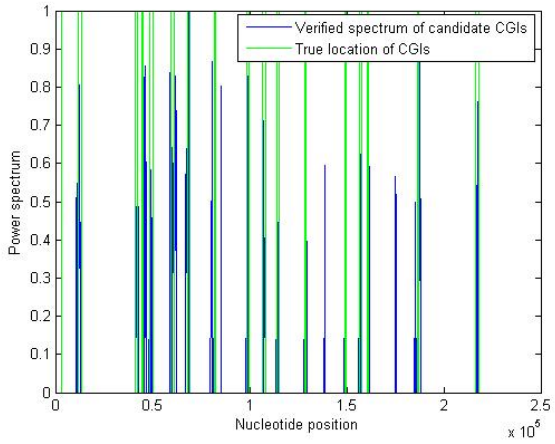

Figure 3. Verified CpG Island's spectrum

5. Combination of 24 spectrums of verified CpG Islands to compute final CpG Islands

Using steps 1-5, verified spectrums of candidate CGIs have been calculated using integer mapping scheme m=1 to 24. These 24 verified spectrums of candidate CGIs are then added to find the final CGI spectrum using (6):

$$F_{CGI}(n) = \sum_{m=1}^{24} V_m(n) \quad (6)$$

F_{CGI}(n) is the final spectrum of the CGIs and for example DNA sequence L44140 it has been shown in Fig. 4. The locations of CGIs detected for example DNA sequence

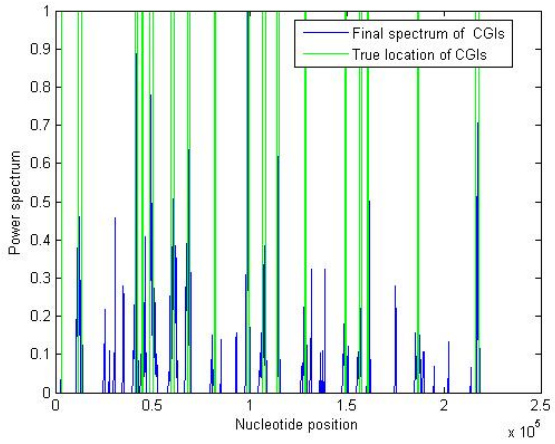

Figure 4. Final CpG Island's spectrum

L44140 using proposed MGWT based algorithm have been

TABLE III. Evalution parameters with varying thresholds for example DNA sequence L44140

	Thresholds								
Evaluation parameter	10%	15%	20%	25%	30%	35%	40%	45%	50%
TP	16659	17844	17414	15349	12088	8015	4318	1939	961
FP	24539	30516	27052	18419	12592	6702	2039	570	235
TN	175679	170602	174066	182699	188526	194416	199079	200548	200883
FN	1669	484	914	2979	6240	10313	14010	16389	17367
Sn	0.909	0.974	0.95	0.837	0.66	0.437	0.236	0.106	0.052
Sp	0.874	0.848	0.865	0.908	0.937	0.967	0.99	0.997	0.999
AC	0.891	0.911	0.908	0.873	0.798	0.702	0.613	0.551	0.526

shown in Table IV.

There are 17 CGIs present in DNA sequence L44140 and the location of these 17 CGIs has been presented in Table IV under the column CGI's true location as per NCBI. The detection outcome of proposed MGWT based algorithm has been shown in Table IV under the column CGI locations identified by proposed algorithm. It has been seen from Table IV that the MGWT based algorithm has detected all 17 CGIs present in DNA sequence L44140; however the algorithm has detected some false positives. Based on the %age coverage of the length of true CGIs which are 17, the performance assessment of the MGWT based algorithm with state of art CGI detection methods is shown in Table V.

The MGWT based algorithm's performance in terms of %age coverage varying from 80 % to 100 % of span of the actual CGI is the best compared to other state of art methods and it has checked from Table V; however the performance of MGWT based algorithm and STFT based algorithm is same at 90 % and full coverage of the span of actual CGI.

The applicability of proposed MGWT based algorithm in the context of the identification of CGIs has been understood from Table V as the proposed algorithm has been able to successfully identify all the CGIs present in benchmark example DNA sequence L44140. Now the performance of proposed method has been evaluated on a large data set of hundred DNA sequences using standard performance metrics and it has been presented in results section.

4. Data Set and Performance Metrics

A. CpG Islands Data Set

The CpG Island data set used in this paper for the validation of performance evaluation of the proposed MGWT based algorithm consists of hundred DNA sequences. These DNA sequences belong to human species. The data set has been collected from publically available database provided by National Centre for Biotechnology Information (NCBI) [39]. The total number of CGIs in this data set of hundred DNA sequences is 181. The description comprising of accession number of DNA sequences, their length, and the location of actual CGI as per NCBI in the length of the data set is presented in the supplementary material.

B. Performance Metrics

The comprehensive assessment of the proposed algorithm and the other existing algorithms has been carried out with the help of the evaluation metrics, sensitivity (Sn), accuracy (AC) [40], specificity (Sp), F-Measure [38]. The explanation of evaluation parameters used is as follows:

$$Sn = \frac{TP}{TP + FN} \tag{7}$$

$$Sp = \frac{TN}{TN + FP} \tag{8}$$

$$AC = \frac{Sn + Sp}{2} \tag{9}$$

$$F - measure = \frac{2 * (precision * recall)}{precision + recall} \tag{10}$$

where;

$$precision = \frac{TP}{TP + FP} \tag{11}$$

&

$$recall = \frac{TP}{TP + FN} \tag{12}$$

TP which is called as true positive corresponds to sections which are predicted accurately by algorithm where actual CGIs are present, FP known as false positive corresponds to erroneously identified regions by the algorithm where actual CGIs are not located, TN termed as true negative represents the appropriately predicted portions where actual CGIs are not located, and FN called as false negative shows the missed sections where actual CGIs are located. Sn abbreviated as sensitivity describes the details concerning the share of TP correctly captured by the algorithm. Sp abbreviated as specificity emphasizes the share of truly predicted TN. The outcome of Sn and Sp is in the range from 0 to 1. An algorithm is considered as perfect if is able to acquire the theoretically desired ideal value of 1 for Sn

TABLE IV. Detected CpG Islands

CGI's true location as per NCBI	CGI locations identified by proposed algorithm
Start position-End position	Start position-End position
3095 - 3426	2935-3207, 10427-10641
11638-13564	10869-13116, 13164-13979, 25224-25530, 27588-28063, 30456-30983, 34927-35171
40983-42150	40115-41829, 41897-42650
44799-45386	43882-46611
48446-50350	48352-52688
59461-61404	58509-62772, 66747-67133
67900-69472	67144-69752, 80359-80681
81836-82633	81542-82710, 85130 85420, 93049-93277
98783-99468	98027-100529
106826-108158	105118-108768
114316-114947	114159-115794, 127000-127238
128187-129236	127348-129369,131543-131904, 136367-136718, 137652-137905,138525-138994
148990-149796	148000-150470, 150764-151072
156388-157495	155288-157715
160697-161402	160782-162048, 162334-162550, 175076-175541
186412-186922	185089-188115, 189537-189740, 194873-195169,202511-202849,214080-214337
216617-217876	216668-218479

TABLE V. Number of CGIs identified in DNA sequence L44140

	No. of CGIs based on detection at % coverage of actual length of CGIs (total 17 CGIs)		
Methods	80%	90%	100%
CpGclusterTLBO	9	5	Nil
CpGPNP	4	3	2
DWT	Nil	Nil	Nil
STFT	15	15	12
Proposed algorithm	16	15	12

and Sp metrics. Accuracy which combines the outcomes of Sn and Sp altogether varies between 0 to 1. Its value should be as close to 1 as achievable for a perfect algorithm. The F-measure metric is a measure of accuracy which calculates the harmonic average of the recall and precision. The range of value of this metric is from 0 to 1. The value of F-measure is desired to be achieved as 1.

5. Results and Discussion

Four state of art methods of CpG island identification, STFT based algorithm, CpGclusterTLBO based CGI detection algorithm, CpGPNP based algorithm and DWT based algorithm have been assessed for the examination of proposed method's performance. The value of evaluation metrics TP, FP, FN, and TN obtained for hundred DNA sequences of human species using all methods considered in the paper is depicted in Table VI.

It has been observed from Table VI that the number of true positives (TPs) obtained using the proposed algorithm is the highest amongst all methods and the number of false negatives (FNs) obtained using proposed algorithm is the least compared to all methods. This feature is always desired theoretically for an algorithm to be considered as perfect that TPs should be as large as possible and correspondingly FNs should be the least. However, FPs which is desired to be as low as possible is little higher for proposed algorithm and CpGclusterTLBO method's FPs are the least amongst all methods. Correspondingly the value of TNs obtained of proposed algorithm which should be as high as possible is little lesser than STFT and CpGTLBO methods but higher than CpGPNP and DWT based methods.

The proposed method's performance for CGI detection is compared utilizing the evaluation metrics sensitivity (Sn), specificity (Sp), accuracy (AC), and F-measure with state of art methods on complete data set of human species comprising of hundred DNA sequences and the obtained results are depicted in Fig. 5-8.

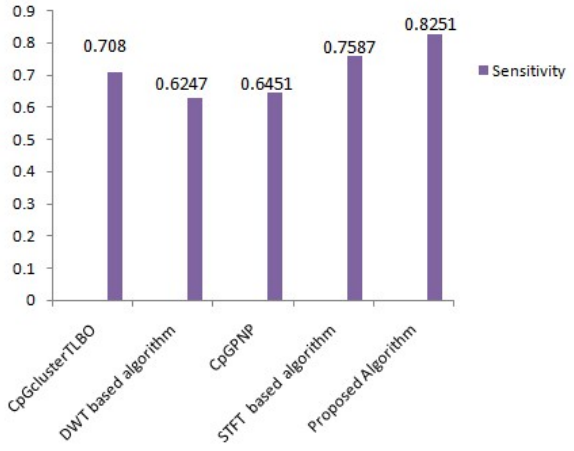

Figure 5. Graph of Sensitivity of all methods

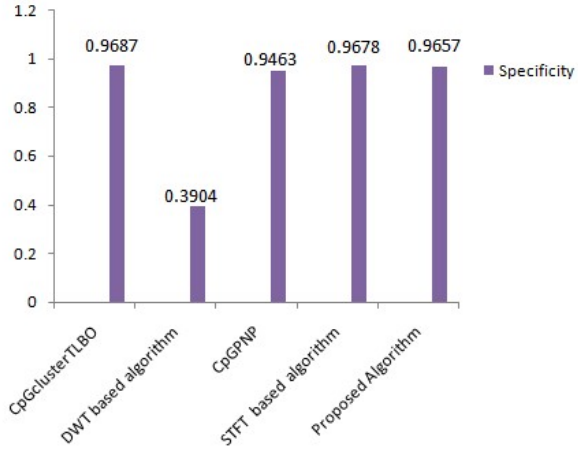

Figure 6. Graph of Specificity of all methods

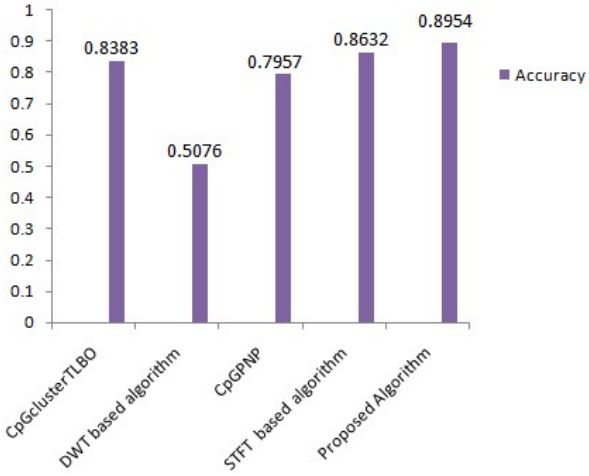

Figure 7. Graph of Accuracy of all methods

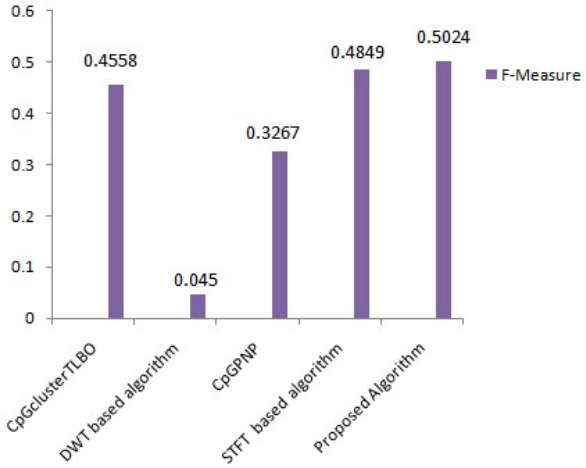

Figure 8. Graph of F-Measure of all methods

TABLE VI. Evaluation metrics for hundred DNA sequences of human species

	Methods				
Evaluation parameter	STFT	CpGclusterTLBO	CpGPNP	DWT	Proposed method
TP	94193	83584	79444	76934	102443
FP	170094	165139	283775	3220837	181252
TN	5112809	5109201	5000128	2063066	5101651
FN	29961	34480	43710	46223	21714

In Fig. 5-8, the methods used for comparison are depicted on x-axis and the values obtained for performance metrics Sn, Sp, AC, and F-measure respectively are presented on y-axis. The superiority of the proposed algorithm in performance parameters over other algorithms is examined in Fig. 5-8. The proposed algorithm's performance for CGI identification has been assessed via the evaluation metrics sensitivity (Sn), specificity (Sp), accuracy (AC), and F-measure from state of art methods on the hundred DNA sequences data set of human species. It has been proved based on the comparison depicted in Fig. 5-8 that the CGI detection performance of proposed MGWT based algorithm is better compared to state of art methods. As the number of TPs of the proposed algorithm is the highest amongst all methods and the number of FPs of the proposed algorithm is the least; correspondingly, evaluation parameters sensitivity (Sn), F-measure and accuracy (AC) of the proposed method are higher than state of art methods for human species with value 0.8251, 0.8954, and 0.5024 respectively. However, as the proposed algorithm has detected false positive little higher compared to CpGclusterTLBO & STFT based algorithm and true negative little lower than CpGclusterTLBO & STFT based algorithm. Correspondingly, the specificity Sp with value 0.9657 of proposed algorithm is almost same as the Sp of CpGclusterTLBO with value 0.9687 and STFT based algorithm value 0.9678.

The percentage improvement of the proposed algorithm over the existing methods in terms of evaluation parameters Sn, AC, and F-Measure has been computed and is represented in Table VII.

As seen from Table VII, the proposed MGWT based algorithm's performance in terms of percentage improvement over the state of art methods of CGI detection for evaluation metrics Sn, AC, and F-Measure is better.

The total number of CGIs in 100 DNA sequences data set is 181. The comparison of the proposed MGWT based algorithm's performance in terms of identification of number of CGIs out of 181 based on percentage coverage of the length of actual CGIs from state of art methods has also been carried out. The comparison result has been tabulated and depicted in Table VIII and Fig. 9 respectively.

The proposed algorithm's performance is better than the state of art methods in context of number of CGIs identified

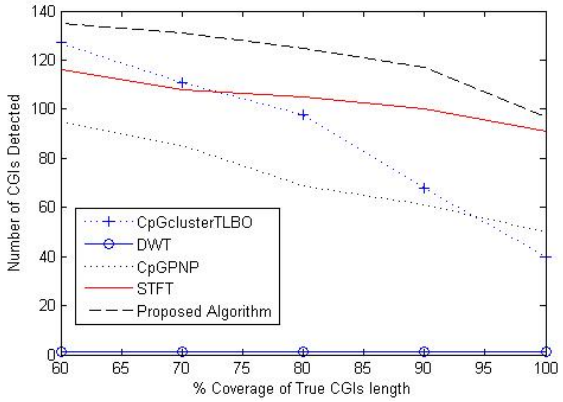

Figure 9. Number of CGIs detected out of total 181

at various percentage coverage of actual CGIs length and it has been observed from Table VIII and Fig. 9. The detection of number of CGIs of proposed MGWT based algorithm at percentage coverage of actual CGI length varying from 60 % to 100 % is much higher as compared to state of art methods.

6. Conclusion

In this paper, MGWT based algorithm for the detection of CGIs is proposed. The algorithm's assessment has been carried out on data set of hundred DNA sequences comprising of human species obtained from NCBI. The performance of the proposed algorithm is better as compared to the state of art methods of CGI detection in terms of sensitivity, accuracy, and F-measure. The specificity of proposed algorithm is almost same as that of CpGclusterTLBO and STFT based methods. Also, the proposed algorithm has detected more number of CGI at 60 % to 100 % coverage of true CGI length. Hence it has been concluded that the proposed MGWT based algorithm is an effective and efficient method for CpG islands detection in DNA sequences. In future this work can be extended to reduce the number of false positives and hence improve the specificity. Also, machine learning based approaches can be explored and employed in future work.

7. SUPPLEMENTARY MATERIAL

The details of data set of hundred DNA sequences of human species is presented in Table IX and X.

TABLE VII. Percentage improvement of proposed algorithm over STFT, CpGclusterTLBO, DWT, CpGPNP

	Methods			
Evaluation metric	STFT	CpGclusterTLBO	CpGPNP	DWT
Sn	8.05%	14.2%	21.82%	24.29%
AC	3.6%	6.38%	11.13%	43.31%
F-Measure	3.48%	9.28%	34.97%	91.04%

TABLE VIII. Number of CGIs detected

Number of CGIs detected at percentage coverage of true CGIs Length					
Methods	60%	70%	80%	90%	100%
CpGclusterTLBO	127/181	111/181	98/181	68/181	40/181
CpGPNP	95/181	85/181	69/181	61/181	50/181
DWT	1/181	1/181	1/181	1/181	1/181
STFT	116/181	108/181	105/181	100/181	91/181
Proposed algorithm	135/181	131/181	125/181	117/181	97/181

References

[1] J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar Jr, "Identification of protein coding regions using the modified gabor-wavelet transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 198–207, 2008.

[2] N. K. Vaegae *et al.*, "Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes," *Biomedical Signal Processing and Control*, vol. 58, p. 101859, 2020.

[3] L. Das, S. Nanda, and J. Das, "An integrated approach for identification of exon locations using recursive gauss newton tuned adaptive kaiser window," *Genomics*, vol. 111, no. 3, pp. 284–296, 2019.

[4] L. Das, J. Das, and S. Nanda, "Identification of exon location applying kaiser window and dft techniques," in *2017 2nd International Conference for Convergence in Technology (I2CT)*, pp. 211–216.

[5] D. K. Shakya, R. Saxena, and S. N. Sharma, "An adaptive window length strategy for eukaryotic cds prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 5, pp. 1241–1252, 2013.

[6] S. Putluri, M. Z. U. Rahman, C. S. Amara, and N. Putluri, "New exon prediction techniques using adaptive signal processing algorithms for genomic analysis," *IEEE Access*, vol. 7, pp. 80 800–80 812, 2019.

[7] A. M. Dessouky, A. El-Samie, E. Fathi, H. Fathi, and G. M. Salama, "Statistical dna sequence modeling and exon detection using non-parametric methods," *International Journal of Computing and Digital Systems*, vol. 9, no. 4, pp. 581–589, 2020.

[8] S. D. Sharma, R. Saxena, and S. N. Sharma, "Identification of microsatellites in dna using adaptive s-transform," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 1097–1105, 2014.

[9] S. D. Sharma, R. Saxena, and S. N. Sharma, "Tandem repeats detection in dna sequences using kaiser window based adaptive s-transform," *Bio-Algorithms and Med-Systems*, vol. 13, no. 3, pp. 167–173, 2017.

[10] S. Sharma, R. Saxena, S. Sharma, and A. Singh, "Short tandem repeats detection in dna sequences using modified s-transform," *International Journal of Advances in Engineering & Technology*, vol. 8, no. 2, p. 233, 2015.

[11] P. Garg and S. Sharma, "Mgwt based algorithm for tandem repeats detection in dna sequences," in *2019 5th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2019, pp. 196–199.

[12] S. Sharma, S. N. Sharma, and R. Saxena, "Identification of short exons disunited by a short intron in eukaryotic dna regions," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 5, pp. 1660–1670, 2019.

[13] R. Touati, I. Messaoudi, A. E. Oueslati, and Z. Lachiri, "Distinguishing between intra-genomic helitron families using time-frequency features and random forest approaches," *Biomedical Signal Processing and Control*, vol. 54, p. 101579, 2019.

[14] R. Touati, I. Messaoudi, A. E. Oueslati, Z. Lachiri, and M. Kharrat, "Classification of intra-genomic helitrons based on features extracted from different orders of fcgs," *Informatics in Medicine Unlocked*, vol. 18, p. 100271, 2020.

[15] S. D. Sharma, S. N. Sharma, and R. Saxena, "Model independent method for acceptor splice sites prediction in dna sequences," in *2019 IEEE Conference on Information and Communication Technology*, 2019, pp. 1–5.

[16] R. A. Tahir, D. Zheng, A. Nazir, and H. Qing, "A review of computational algorithms for cpg islands detection," *Journal of biosciences*, vol. 44, no. 6, pp. 1–11, 2019.

[17] S. Erkek, M. Hisano, C.-Y. Liang, M. Gill, R. Murr, J. Dieker, D. Schübeler, J. Van Der Vlag, M. B. Stadler, and A. H. Peters, "Molecular determinants of nucleosome retention at cpg-rich sequences in mouse spermatozoa," *Nature structural & molecular biology*, vol. 20, no. 7, pp. 868–875, 2013.

[18] Y. Wang and F. C. Leung, "An evaluation of new criteria for cpg islands in the human genome as gene markers," *Bioinformatics*,

TABLE IX. Detailed description of the data set as per NCBI website

S.No.	Acc. No.	Length	Locations
1	AL442638	188247	17472-17700, 22868-23148, 93250-93495, 163847-164132
2	AC073335	68275	31813-32080, 33619-34458, 50802-51655
3	AC073517	67706	35431-35977
4	AC127379	67291	30060-30318, 38447-39437
5	AC064843	66898	5531-5785
6	AC129782	66860	38868-40898
7	AC013270	66660	6075-6881, 25374-26035, 34710-36183, 48185-48621
8	AC074386	66610	15847-16381, 16593-16830
9	AC092103	66565	24844-25119
10	AC124014	66552	56936-57769
11	AL137791	66254	30724-31272, 46196-46906, 52979-53956, 61007-62096
12	AC096553	66229	11867-12256
13	AC105413	65958	50478-50751
14	AC005003	65750	38374-41067
15	AC145546	65625	52797-53645
16	AC105402	65449	15774-16973, 28628-28925
17	AC112698	65335	42309-43546
18	AC104129	65189	2966-3334, 8763-9020, 14023-14383, 20695-20991, 26472-26735, 28330-29188, 31762-32009, 55671-55878
19	BN000001	64961	895-1123
20	AC138782	64744	23500-24633
21	AC005021	64607	24663-25225, 63177-63512
22	AC093086	64601	58914-59518
23	AC005233	64359	16579-18003
24	AC013436	63823	12411-12652, 21066-21331, 24980-26051, 26467-26807, 60097-60448
25	AC131957	63780	45526-45799
26	AC004694	63749	9107-9494, 54481-54756
27	AC108463	63525	26008-26366, 26575-26982, 27079-27538
28	AC080165	63279	8258-8531
29	AC010890	62764	11407-11926, 13574-13801, 53142-53415, 53755-54041
30	AC108142	62624	8864-11837
31	AC080068	62623	535-774
32	AC093785	62466	31397-31665
33	AC003079	62331	50250-50471
34	AC078937	62035	38149-39359
35	AC114803	61579	3256-4009
36	AC093652	61340	48156-49072
37	AC093377	61056	729-1003
38	AC073201	60776	9738-11862
39	AC113611	60597	8638-9514
40	AC099394	60024	2826-4863, 10806-11866, 19723-19934, 25482-25769, 31861-32884, 36728-36931, 54994-55361
41	AC098831	59776	39343-39572, 51406-51689
42	AC074013	59657	22602-22873, 51602-52508, 53105-53331
43	AC062028	59634	44629-44851
44	AC106875	59580	4526-5382
45	AC023670	59565	25568-27400
46	AC079882	59427	39153-39736
47	AC006008	57554	28800-30423
48	AC108222	21776	21237-21776
49	AH006464	21230	1187-2051
50	AC093609	20710	7857-8257
51	AL590794	18042	11568-12215
52	AC136375	17863	16369-17534
53	BD432859	14646	2762-2973, 4065-5181
54	AC111201	13470	4327-4727, 5323-5554, 12500-13455

TABLE X. Detailed description of the data set as per NCBI website continued

S.No.	Acc. No.	Length	Locations
55	NM005876	10782	6154-7734
56	NM053043	10168	9597-9820
57	AC093460	10103	6951-7418
58	AC108032	9716	30-269
59	X86012	9541	335-3853
60	AC106048	8594	7941-8180
61	AH008870	6797	341-1340
62	AC079401	6568	3086-3935
63	AH007568	6513	543-803, 1212-1430, 1662-2474
64	AC105385	5952	2844-3080
65	AJ308559	5596	1228-1657
66	M92844	3889	3198-3889
67	AF196313	3700	2092-3580
68	AF281043	3662	1611-2734
69	U48937	3278	2588-3230
70	AF307776	3113	2334-2745, 2791-3064
71	AJ000757	3046	650-2840
72	AJ289875	2916	2325-2916
73	L07287	2704	1-1350
74	Z92546	73511	20746-21240
75	AL591222	147211	54605-55080, 68825-69091
76	AL513502	174636	116364-117432
77	AL513498	155780	18305-18582
78	AL357615	171446	56753-57030, 59607-59874
79	AL353786	139565	19000-19400
80	AL121926	139544	102641-104201, 126562-127299
81	AL049547	129811	27801-29311, 37094-37773, 109041-110125, 113196-114024, 126815-127265
82	AL031706	13012	7-552
83	AL031703	35098	15319-17699, 25107-26048, 30327-30736, 31615-32204
84	AJ006998	123521	11140-11417
85	AL031707	28707	6050-6520, 6693-7445, 24481-25248, 28059-28669
86	AL024496	27210	1284-1927, 9755-10674, 13099-13615, 15578-16126, 21132-21595
87	AL109743	96006	31713-33048, 56464-57695
88	AC027644	188207	27115-27651, 51380-51705, 130590-131909
89	AC110076	105211	93622-94410
90	AC073271	117930	102756-103541
91	AC005282	98219	8323-9168, 79507-80293
92	AC110787	7335	11-1165
93	L47124	6996	3226-4068
94	AC010990	6708	2347-2685, 4079-4357
95	AF129290	6324	2026-2238, 2436-2679, 2730-3021, 3033-3353, 3355-3637, 4479-4891
96	D13370	3730	226-1645
97	AH004914	5426	1018-1636
98	AC079588	4249	1137-2422
99	AH009772	4240	1-555, 656-1588
100	AL132818	38860	33379-33940

vol. 20, no. 7, pp. 1170–1177, 2004.

[19] P. Feng, W. Chen, and H. Lin, "Prediction of cpg island methylation status by integrating dna physicochemical properties," *Genomics*, vol. 104, no. 4, pp. 229–233, 2014.

[20] H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg, "Redefining cpg islands using hidden markov models," *Biostatistics*, vol. 11, no. 3, pp. 499–514, 2010.

[21] C. Bock, J. Walter, M. Paulsen, and T. Lengauer, "Cpg island mapping by epigenome prediction," *PLoS computational biology*, vol. 3, no. 6, p. e110, 2007.

[22] M. Gardiner-Garden and M. Frommer, "Cpg islands in vertebrate genomes," *Journal of molecular biology*, vol. 196, no. 2, pp. 261–282, 1987.

[23] D. Takai and P. A. Jones, "Comprehensive analysis of cpg islands in human chromosomes 21 and 22," *Proceedings of the national academy of sciences*, vol. 99, no. 6, pp. 3740–3745, 2002.

[24] N. Yu, X. Guo, A. Zelikovsky, and Y. Pan, "Gaussiancpg: a gaussian model for detection of cpg island in human genome sequences," *BMC genomics*, vol. 18, no. 4, pp. 1–9, 2017.

[25] L. Ponger and D. Mouchiroud, "Cpgprod: identifying cpg islands associated with transcription start sites in large genomic mammalian sequences," *Bioinformatics*, vol. 18, no. 4, pp. 631–633, 2002.

[26] P. Rice, I. Longden, and A. Bleasby, "Emboss: the european molecular biology open software suite," *Trends in genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[27] H.-C. Park, E.-R. Ahn, J. Y. Jung, J.-H. Park, J. W. Lee, S.-K. Lim, and W. Kim, "Enhanced sensitivity of cpg island search and primer design based on predicted cpg island position," *Forensic Science International: Genetics*, vol. 34, pp. 134–140, 2018.

[28] M. J. Sippl, "Biological sequence analysis. probabilistic models of proteins and nucleic acids, edited by r. durbin, s. eddy, a. krogh, and g. mitchinson. 1998. cambridge: Cambridge university press. 356 pp.£ 55.00 (80.00)(*hardcover*); £19.95(34.95)," *Protein Science*, vol. 8, no. 3, pp. 695–695, 1999.

[29] L.-Y. Chuang, C.-H. Yang, M.-C. Lin, and C.-H. Yang, "Cpgpap: Cpg island predictor analysis platform," *BMC genetics*, vol. 13, no. 1, pp. 1–9, 2012.

[30] B.-J. Yoon and P. Vaidyanathan, "Identification of cpg islands using a bank of iir lowpass filters [dna sequence detection]," in *3rd IEEE Signal Processing Education Workshop. 2004 IEEE 11th Digital Signal Processing Workshop, 2004.* IEEE, 2004, pp. 315–319.

[31] Y. Sujuan, A. Asaithambi, and Y. Liu, "Cpgif: an algorithm for the identification of cpg islands," *Bioinformation*, vol. 2, no. 8, p. 335, 2008.

[32] N. Elango and S. V. Yi, "Functional relevance of cpg island length for regulation of gene expression," *Genetics*, vol. 187, no. 4, pp. 1077–1083, 2011.

[33] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver, "Cpgcluster: a distance-based algorithm for cpg-island detection," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[34] C.-H. Yang, Y.-C. Chiang, L.-Y. Chuang, and Y.-D. Lin, "A cpgcluster-teaching–learning-based optimization for prediction of cpg islands in the human genome," *Journal of Computational Biology*, vol. 25, no. 2, pp. 158–169, 2018.

[35] A. Rushdi and J. Tuqan, "A new dsp-based measure for cpg islands detection," in *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*. IEEE, 2006, pp. 561–565.

[36] I. T. Mariapushpam and S. Rajagopal, "Improved algorithm for the location of cpg islands in genomic sequences using discrete wavelet transforms," *Current Bioinformatics*, vol. 12, no. 1, pp. 57–65, 2017.

[37] P. Garg and S. Sharma, "Cpg islands identification in dna sequences using modified p-spectrum based algorithm," in *Journal of Physics: Conference Series*, vol. 1921, no. 1. IOP Publishing, 2021, p. 012042.

[38] P. Garg and S. Sharma, "Identification of cpg islands in dna sequences using short-time fourier transform," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, no. 3, pp. 355–367, 2020.

[39] "National centre for biotechnology information: Available at: https://www.ncbi.nlm.nih.gov/nuccore/," 2021.

[40] P. Garg and S. D. Sharma, "Sensitivity enhancement of dwt based algorithm for detection of cpg islands in dna sequences," *Procedia Computer Science*, vol. 167, pp. 1829–1838, 2020.

Pardeep Garg completed his M. Tech in Electronics Communication Engineering from Jaypee Institute of Information Technology University Noida, Uttar Pradesh in 2009. He obtained his B. Tech in Electronics Communication Engineering affiliated from Kurukshetra University, Kurukshetra in 2004. Currently, he is pursuing a Ph. D. in Genomics Signal Processing from Jaypee University of Information Technology Waknaghat, Solan, Himachal Pradesh. He is working as an Assistant Professor in Electronics Communication Engineering department at Jaypee University of Information Technology Waknaghat, Solan, Himachal Pradesh, India since July 2010. His research interest includes genomics signal processing, bioinformatics, big data, internet of things (IoT), machine learning.

Sunil Datt Sharma received the BE degree in Electronics Instrumentation Engineering, the M. Tech degree in Embedded systems and VLSI, and the Ph. D. degree in ECE, in 2005, 2010, and 2016, respectively. He is working as an Assistant Professor in Electronics Communication Engineering department at Jaypee University of Information Technology Waknaghat, Solan, Himachal Pradesh, India. His research interest is in genomics signal processing, bioinformatics, internet of things (IoT), and time-frequency analysis of non stationary signals.