CpG Island Detection Using Transformer Model with Conditional Random Field

Md Jubaer Hossain
Dept. of EEE, BUET
Dhaka-1205, Bangladesh
1018062252@eee.buet.ac.bd

Mohammed Imamul Hassan Bhuiyan
Dept. of EEE, BUET
Dhaka-1205, Bangladesh
imamul@eee.buet.ac.bd

Zaowad Rahabin Abdullah
Dept. of EEE, BUET
Dhaka-1205, Bangladesh
rahabin48@gmail.com

***Abstract*—Detecting potential locations of CpG islands is one of the first steps for predicting promoter regions of many housekeeping and tissue-specific genes, which in turn, helps identify many epigenetic causes of cancer. Traditionally, finding potential CpG islands computationally involves calculating many manual-features and making different assumptions. Recently, in Natural Language Processing(NLP), transformer architectures incorporating mulit-head attention have surpassed many other sequence processing architectures such as RNN, GRU, LSTM etc. in terms of accuracy, speed, and computational efficiency. One of the major attributes of NLP is Named Entity Recognition(NER), which extracts the relevant information from a long sequence. In this study, CpG island identification is considered as an NER problem and transformer architecture is used for its detection. Conditional random field is further incorporated to include the dependencies of the associated labels. Additional attention mask is included on the input layer to give more importance to the regions relevant to DNA sequence. The publicly available EMBL human DNA database is used for experiments. It is observed that more than 96 % accuracy and 73 % F1-score can be achieved, a superior performance as compared to the existing results in the literature. The proposed approach can be utilized for identifying bio-markers for different important and disease-related genes efficiently. In addition, it may be used for other genome sequence analysis and processing tasks.**

***Index Terms*—CpG isalnd, Named Entity Recognition, Transformer, BERT, DistilBERT, Conditional Random Field, Epigenetics**

I. Introduction

In DNA sequence, CpG refers to pair of nucleotides, where C (cytosine) is immediately followd by G (Guanine). CpG islands indicate places where high profiles of CpG sites are densely contained in some genome regions.These islands can vary in length from a few hundred to a few thousand nucleotides long[1]. CpG islands are often located close to the transcription sites of genes[2]. They are often related with the promoters of most housekeeping genes and many tissue-specific genes[3]. They can be used as gene markers. In vertebrate, DNA methylation usually occurs in CpG islands and adds an additional methyl to Cytosine in a way that the gene silencing may be caused by the additional methyl. Methylation of promoter-related CGIs is very common in all types of cancer cells, so the hypermethylated CpG islands in promoter regions can be used as molecular tumor markers, making the early detection of cancer possible[4], [5].

Identification of potential CpG islands is essential for finding not only promoter regions of many specific genes but also locations of problematic sequence in DNA and therefore helps understand the epigenetic causes of many diseases.CpG island identification tasks can be approached either experimentally or computationally. Conventional and experimental bisulfate modification-based methods to determine CpG islands and methylation regions are time-consuming[6] and costly.Thus, computational approaches are regarded effective and used for many biological studies[7].These approaches can be divided into several categories, for example, window-based methods[8], [9], Hidden Markov Model(HMM)-based methods[6], [10], [11], gaussian model-based, [12], filter-based methods[13], density-based methods[14], [15], distance or length based methods [16], domain transformations [17], [18] etc.

Major limitations of the above mentioned methods are either they are rule-based or they depend on hand-crafted features, which don't generalize as these rules and features become data-specific for which they are developed. Also, these algorithms can't leverage large amount of data as deep learning algorithms do. Recent developments in Natural Language Processing(NLP), specially in transformer based architectures like BERT[19] , have created a new way of approaching other sequence processing tasks. Researchers have already used deep learning[20] and specifically transformer-based architectures in various tasks related to analysis of DNA sequences[21], [22] and other biomedical signals[23]–[25]. In this paper, the CpG islands detection has been modeled as Named Entity Recognition(NER) problem in NLP. For this, DistilBERT, which is a modified version of BERT, has been used as the base model. Then Conditional Random Filed(CRF) is used on top of the DistilBERT model to capture additional dependencies in the labels. Additional attention mask layer is also incorporated on the input to give more weights to the relevant regions. The overall contributions of this paper can be summarized as bellow:

- For the first time, a transformer-based architecture incorporating conditional random field is introduced for the detection of CpG islands.
- The proposed method provides a superior F1-score, sen-

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

sitivity, and specificity of 0.735, 0.852, and 0.959 respectively on a publicly available dataset as compared to the existing results reported in the literature.

The rest of the paper is organized as follows. A brief introduction of the transformer architecture is provided in Section II. Proposed methodology for the detection of CpG islands is described in Section III. Experimental results are discussed in Section IV and concluding remarks are provided in Section V.

II. Transformer Model Architectures

Until recently, RNN, LSTM, GRUs, etc were used to process sequence data. One of the limitations of these deep learning architectures is that they face difficulties processing longer sequence data due to vanishing-gradient problem. Then came attention based architectures which addressed this issue. Transformers, which rely on attention mechanism, have shown excellent performance in solving various natural language processing problems[26].

In transformers, self-attention, specifically "Scaled Dot-Product Attention" [26], is used so that every token can be aware of every other token in the sequence. After combining the token embedding and positional embedding of every individual token, three separate vectors are generated for every token called query(q) and key(k) of dimension d_k, and value(v) of dimension d_v. Then dot products of the query with all keys are computed; results are divided by $\sqrt{d_k}$ followed by the application of softmax function to get the weights on the values. The motivations for this partially come from the information retrieval system. In practice, the attention function on a set of queries are computed simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V . The matrix of outputs is computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

To learn multiple aspects of the sequence parallelly, multiple such attention heads are calculated, and then concatenated to create multi-head attention as

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2}$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

Here, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_k}$, and $W_i^Q \in \mathbb{R}^{hd_v \times d_{model}}$ are projection parameter matrices that are learned during training, d_{model} and h being the model output dimension and number of heads respectively. Each head learns different relations of the tokens from the sequence. The conceptual diagrams of Scaled Dot-Product Attention and Mulit-Head Attention are depicted in Fig. 1.

BERT and DistilBERT are two well-known transformer architectures[19], [27]. BERT is widely used in language model pre-training and many downstream tasks like sentiment analysis, named entity recognition, etc. Two unsupervised tasks are used to train a BERT model, Masked Language Modeling(MLM) and Next Sentence Prediction(NSP). MLM refers to the technique where some inputs are masked randomly and the model learns to predict those masked input. NSP is the technique where the model learns the relationship between the sentences. In this way, BERT can get the full context of the words or tokens and generate a robust embedding vector. When using BERT and its variants for text classification, a special token, [CLS] is added to the beginning of the sentence and the model learns to predict the class of that [CLS] token, as a representative of the whole sentence. But when doing named entity recognition, every token in a sentence has to be classified. So, the model is trained accordingly- it learns to classify every token in a sentence. DistilBERT is a variant and smaller version of BERT. It has the same general architecture of BERT while the number of layers is reduced by a factor of 2. It has been observed in other NLP domains that DistilBERT can achieve 97% performance of BERT keeping it's size and inference time about half of BERT [27]. In this work, DistilBERT is utilized as the base architecture.

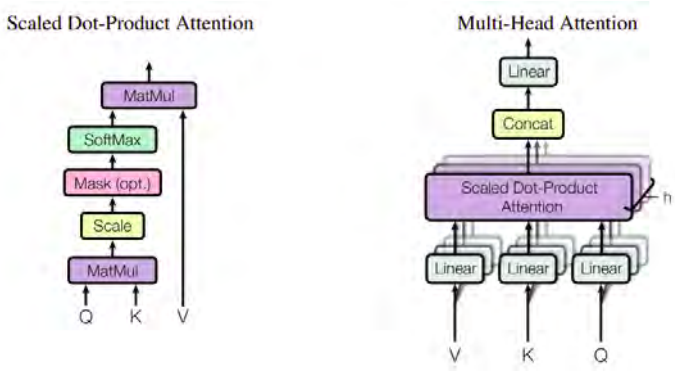

Figure 1. (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel [26]

Both of these architectures are used extensively in Named Entity Recognition(NER) tasks. For NER, every token in a sequence has to be labeled individually, either inside a tag of interest or outside. For example, let's take a sentence "New York is a state of the USA". If this sentence is tokenized based on space, then the tokens would be ["New", "York", "is", "a", "state", "of", "the", "USA"]. If one wants to extract the location entities from this, the labels will be like this, ["B_LOC", "I_LOC", "O", "O", "O", "O", "O", "I_LOC"]. "O" refers to outside of entity, B_LOC refers to the beginning of location entity, "I_LOC" refers to the inside of the location entity. During training, the model learns to classify every token according to its label in the sentence as well as the relationships between tokens using backpropagation. However, this labeling strategy varies according to the purpose of a particular task. Similar concepts can be used for identifying the CpG regions in a DNA sequence. To the best of our knowledge, nobody has applied similar techniques for finding CpG islands yet.

III. Methodology

A. Data Corpus

The dataset embl173hum is used in the present work. It has 233,004 sequences. Of them, 61051 contain total 142325 islands. The lengths of the islands vary from a couple of hundreds to thousands nucleotides. The mean island length was around 373. The data and the metadata about the CpG islands are available at European Bio-informatics Institute(EBI) https://www.ebi.ac.uk/ and ftp://ftp.ebi.ac.uk/pub/databases/cpgisle respectively.

B. Data Preprocessing

As CpG island detection is considered as NER problem in this paper, generating tokens itself requires experiments. After doing some analysis, Byte-Level Byte-Pair encoding(BPE) has been used for tokenizing the DNA sequences. BPE is an iterative algorithm that starts with a fixed vocabulary of individual nucleotides (A, T, C, G) and progressively merges the bytes into pairs based on which pairs occur most frequently in the data set of training sequences. The process is repeated to create larger-sized tokens until it reaches a target vocabulary size, which, in our case, was 5000.

For the case of CpG island detection, the starting and ending positions of every island are taken from the ground truth file . Then when tokens fall under these positions, they are labeled as "I_ISL" referring to the inside position. All other tokens are labeled as "O" (outside).

As the DNA sequences are very large, every large sequence is partitioned into smaller sub-sequence of different lengths. Sub-sequence lengths of 256 and 420 are used after experimentation.

C. Proposed Method

As DistilBERT is basically an encoder, task-specific softmax layer has to be used at the final layer for any downstream task. In this paper, the softmax layer is replaced by Conditional Random Field(CRF) to capture additional dependencies on the labels. CRFs model the conditional probability distribution of the labels. They have been very successful for named entity recognition in different domains [28]. For extracting CpG islands, it is important to take into account the relations between the labels for tokens that are side by side. CRFs don't make independent-label assumption and they take into account the previous labels and the whole sequence while decoding. In the end, an additional attention mask is activated, whereby the model pays more attention to the relevant regions ignoring the padding and other special tokens. The overall architecture is shown in Fig. 2, where E, T, and C refer to token embedding, output embedding, and class label.

IV. Results and Discussion

From the embl173hum data set, 5000 main sequences are taken. Then sub-sequences are created by splitting the main sequences based on token size and sequence length. For tokenization, fix token sizes of 2, 3 and byte level Byte Pair Encoding(BPE) have been tested. BPE is found to show better performance than the other two. Thus, experiments are conducted with BPE tokens. When BPE is used, total 190282 sub-sequences are generated, keeping the maximum sub-sequence length of 420 . 10% of the dataset are kept as test set, making sure about the integrity of the main and sub sequences.

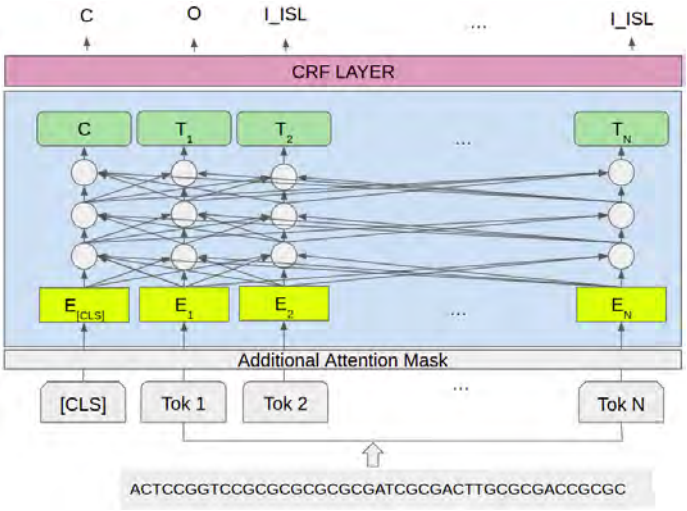

Figure 2. Proposed Transformer Architecture. The architecture has 6 hidden layers, 12 heads in each attention layer, 768 dimensions in the encoder and pooling layers.

Performance of the proposed approach is assessed in terms of sensitivity, specificity, positive predictive value (PPV), accuracy and F1-score as well as for the DistilBERT and DistilBERT in conjunction with CRFs. When CRF is used on top of DistilBERT, a significant improvement is observed as the model starts to capture inter-dependencies among the labels. After introducing an additional attention mask on the input, the model pays more attention to the relevant regions and as a result, another phase of improvement is observed. It is seen from Table I that the proposed method employing DistilBERT with CRF and additional attention mask provides significant improvement in terms of all the metrics over DistilBERT alone and DistilBERT+CRF.

TABLE I
Experiment Results

Architecture	Token Size	SEN	SPE	PPV	F1-score	Accuracy
DistilBERT	BPE	79.6	90.3	65.3	71.8	92.8
DistilBERT + CRF	BPE	82.4	91.2	67.8	72.6	93.4
DistilBERT + CRF + Attention Mask	**BPE**	**85.2**	**95.9**	**69.1**	**73.5**	**96.5**

Table II gives a comparison of the proposed method with that of [12], [17], and [18]. It should be mentioned that they used different datasets with a smaller number of whole DNA sequences. For example, [17] and [18] reported their results on 3 and 100 whole human DNA sequences respectively. On the other hand, the proposed method utilizes 500 human whole DNA sequences as test set. It can be seen that the performance of the proposed method is superior to that of [12], [17], and

[18] in terms of sensitivity and F1-score. CpG island datasets are highly imbalanced as only a very few portion of the sequences are island regions. For example, of the 500 test sequences that this paper has used, only around 1.5% of the sequences are island regions. In these cases, accuracy can give misleading results, as blindly predicting all of the nucleotides as non-island regions can give about 98.5% accuracy. For this kind of imbalanced datasets, F1-score gives more robust and accurate performance as it takes into account both precision and recall capability of the model.

One limitation of this paper is that only 5000 main sequences have been used for the experiment out of 61051 available sequences in the embl173hum dataset due to computational resource constraint. It is hoped that further improvement can be achieved if all of the available sequences can be utilized.

TABLE II
COMPARISON WITH OTHER METHODS

Method	SEN	SPE	F1-score	Accuracy
Yu et al., 2017[12]	83.9	**99.2**	58.2	**99.1**
Garg et al., 2020 [17]	78.5	96.9	49.5	87.7
Garg et al., 2021 [18]	50.1	73.7	36.6	69.5
Proposed Method	**85.2**	95.9	**73.5**	96.5

V. CONCLUSION

CpG islands are important genetic markers for many biological phenomenon and identifying them can help detect many diseases including cancers. Finding methylation level of DNA is a promising approach for cancer diagnosis and management[5], and efficient detection of CpG islands is one of the important steps for it. In this paper, it has been demonstrated that identification of CpG islands in DNA sequence can be regarded as Named Entity Recognition problem and superior detection-performance can be achieved using transformer architecture with CRF and additional attention mask on the input layer. A similar strategy can be employed to detect other genetic markers obviating the use of hand-crafted features, while leveraging the huge amount of genomic data available and superior classification ability of the proposed transformer-based deep neural architecture.

REFERENCES

[1] F. Antequera and A. Bird, "Cpg islands as genomic footprints of promoters that are associated with replication origins," *Current Biology*, vol. 9, no. 17, pp. 661–667, 1999.

[2] M. Gardiner-Garden and M. Fromer, "Comprehensive analysis of cpg islands in chromosomes 21 and 22," *Journal of Molecular Biology*, vol. 196, no. 2, 1987.

[3] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz, "Cpg islands as gene markers in the human genome," *Genomics*, vol. 13, no. 4, pp. 1095–1107, 1992.

[4] L. A. Uroshlev, E. T. Abdullaev, I. R. Umarova, *et al.*, "A method for identification of the methylation level of cpg islands from ngs data," *Scientific reports*, vol. 10, no. 1, pp. 1–5, 2020.

[5] J. Huang, A. C. Soupir, B. D. Schlick, *et al.*, "Cancer detection and classification by cpg island hypermethylation signatures in plasma cell-free dna," *Cancers*, vol. 13, no. 22, p. 5611, 2021.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] R. Kakumani, O. Ahmad, and V. Devabhaktuni, "Identification of cpg islands in dna sequences using statistically optimal null filters," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2012, no. 1, pp. 1–14, 2012.

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

[15] 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.

[16] 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.

[17] 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.

[18] P. Garg and S. Sharma, "Cpg islands identification in dna sequences using modified p-spectrum based algo-

rithm," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1921, 2021, p. 012 042.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[20] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle, "Deepcpg: Accurate prediction of single-cell dna methylation states using deep learning," *Genome Biology*, vol. 18, no. 1, pp. 1–13, 2017.

[21] G. De Waele, J. Clauwaert, G. Menschaert, and W. Waegeman, "Cpg transformer for imputation of single-cell methylomes," *Bioinformatics*, vol. 38, no. 3, pp. 597–603, 2022.

[22] J. Clauwaert, G. Menschaert, and W. Waegeman, "Explainability in transformer models for functional genomics," *Briefings in Bioinformatics*, vol. 22, no. 5, bbab060, 2021.

[23] D. M. D. Nguyen, M. Miah, G.-A. Bilodeau, and W. Bouachir, "Transformers for 1d signals in parkinson's disease detection from gait," *arXiv preprint arXiv:2204.00423*, 2022.

[24] M. D. Le, V. S. Rathour, Q. S. Truong, Q. Mai, P. Brijesh, and N. Le, "Multi-module recurrent convolutional neural network with transformer encoder for ecg arrhythmia classification," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2021, pp. 1–5.

[25] P. Lu, C. Wang, J. Hagenah, *et al.*, "Improving classification of tetanus severity for patients in low-middle income countries wearing ecg sensors by using a cnn-transformer network," *IEEE Transactions on Biomedical Engineering*, 2022.

[26] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. arXiv: 1910.01108. [Online]. Available: http://arxiv.org/abs/1910.01108.

[28] I. Mazharov and B. V. Dobrov, "Named entity recognition for information security domain.," in *DAMDID/RCDL*, 2018, pp. 200–207.