scientific reports

Check for updates

OPEN

DNA sequences performs as natural language processing by exploiting deep learning algorithm for the identification of N4-methylcytosine

Abdul Wahab[1,5], Hilal Tayara[2,5], Zhenyu Xuan[3]✉ & Kil To Chong[1,4]✉

N4-methylcytosine is a biochemical alteration of DNA that affects the genetic operations without modifying the DNA nucleotides such as gene expression, genomic imprinting, chromosome stability, and the development of the cell. In the proposed work, a computational model, 4mCNLP-Deep, used the word embedding approach as a vector formulation by exploiting deep learning based CNN algorithm to predict 4mC and non-4mC sites on the *C.elegans* genome dataset. Diversity of ranges employed for the experimental such as corpus k-mer and k-fold cross-validation to obtain the prevailing capabilities. The 4mCNLP-Deep outperform from the state-of-the-art predictor by achieving the results in five evaluation metrics by following; Accuracy (ACC) as 0.9354, Mathew's correlation coefficient (MCC) as 0.8608, Specificity (Sp) as 0.89.96, Sensitivity (Sn) as 0.9563, and Area under curve (AUC) as 0.9731 by using 3-mer corpus word2vec and 3-fold cross-validation and attained the increment of 1.1%, 0.6%, 0.58%, 0.77%, and 4.89%, respectively. At last, we developed the online webserver http://nsclbio.jbnu.ac.kr/tools/4mCNLP-Deep/, for the experimental researchers to get the results easily.

DNA methylation is a mechanism that entails the chemical modification of DNA sequences, that changes hereditary performance without altering the DNA's nucleobases. DNA modification through methylation and demethylation plays a significant role in gene expression. DNA methylation can regulate various biological processes including genomic imprinting, chromosome stability, and cell development and extend the assortment of genes because of its structural changes in DNA[1]. Prokaryotic and eukaryotic genomes undergo three types of methylation; N4-methylcytosine (4mC)[2], 5-Methylcytosine (5mC)[3], and N6-methyladenine (6mA)[4].

Gene modifications are assembled by the distinct DNA methyltransferases (DNMTs) to transmit a methyl group to a particular exocyclic amino group[5]. 5mC is one of the most extensively studied types of cytosine methylation as a consequence of its widespread dissemination and complicated aspects[6]. 5mC plays an important role in numerous biological processes[7] associated with neurological diseases, diabetes, and cancer. 6mA, in contrast, takes place only on a very small-scale, and is only found in eukaryotes using high sensitivity methods. 4mC is considered a dynamic epigenetic modification because of the restriction-modification (R-M) method to protect restriction enzyme form deterioration of self-DNA. It was first discovered in 1983[8]. 4mC plays a significant role in the regulation of a number of processes intrinsic to the cell cycle, including gene expression, defining self and non-self-DNA , DNA replication, and correcting DNA replication errors[9,10]. Investigational studies related to 4mC have waned in part due to a lack of sufficient identification techniques. While there are several experimental procedures capable of identifying 4mC sites, including mass spectrometry, for the whole genome 4mC-Tet-assisted bisulfite, Single-Molecule of Real-Time (SMRT) sequencing, and methylation-precise PCR[11–14], these approaches are regarded as expensive and time-consuming when applied across an entire genome.

[1]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea. [2]School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea. [3]Department of Biological Sciences, The University of Texas at Dallas, Richardson 75080, USA. [4]Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea. [5]These authors contributed equally: Abdul Wahab and Hilal Tayara. ✉email: zhenyu.xuan@utdallas.edu; kitchong@jbnu.ac.kr

In recent years, various novel computational N4-methylcytosine site identification methods have been proposed and applied across a diverse number of species, including *Geoalkalibacter subterraneus*, *Arabidopsis thaliana*, *Geobacter pickeringii*, *Escherichia coli*, *Drosophila melanogaster*, and *Caenorhabditis elegans*[15–19]. These methods rely on state-of-the-art machine learning (ML) algorithms to make their predictions for 4mC sites. Each of these predictors used different encoding techniques such as binary encoding, nucleotide chemical properties, and nucleotide frequencies with various algorithms like support vector machine, random forest, and decision tree. Recently, a new predictor DNC4mC-Deep[20] was proposed to identify and analyze of Rosaceae genome, where they implemented a deep learning based Convolution Neural Network (CNN) algorithm. They used six different kinds of encoding methods binary encoding (BE), dinucleotide composition (DNC), trinucleotide composition (TNC), Nucleotide chemical property (NCP), nucleotide chemical property and nucleotide frequency (NCPNF), and multivariate mutual information (MMI). Another deep learning based model has been established named as 4mCDeep-CBI[21] to identify the N4-methylcytosine sites in the newly developed dataset of Caenorhabditis elegans, where they implemented 3-CNN and Bidirectional Long Short-Term Memory (BLSTM) to fetch the deep features for the prediction.

In this work, we developed a new tool, named 4mCNLP-Deep, to identify and analyze 4mC sites associated with *C. elegans* dataset which was recently expanded by increasing the number of samples. The Structure of the proposed model was built as follows; First, we used the encoding method word2vec, which has never been used before in N4-methylcytosine identification, to transform sequences into vectors form using word embedding. The word-embedding approach mostly operates on Natural Language Processing (NLP)[22], but thereafter executed efficaciously on wide-genome identification[23–28]. We obtained the final CNN model by applying the grid search algorithm with tuned hyper-parameters and fed the vectors of word embedding into it. We used a K-fold cross-validation method for different values of K. Then, we applied five evaluation metrics to assess the model. We also employed two applications, silico mutagenesis[29] and saliency[30] map to interpret the predictive deep learning model and influence of important features. The results of the predictive model showed outperformance when compared to the state-of-the-art model. 4mCNLP-Deep successfully achieves 0.9354, 0.8608, 0.8996, 09563, and 0.9731 for Accuracy (ACC), Mathew's correlation coefficient (MCC), Sensitivity (Sn), Specificity (Sp), and Area under the curve (AUC), respectively on C. elegan dataset by using 3-mer corpus word2vec and 3-fold cross-validation. Our model attained the increment of 1.1% on ACC, 0.6% on MCC, 0.58% on Sn, 0.77% Sp, and 4.89% on AUC, respectively.

Materials and methods

Benchmark datasets. The benchmark dataset of *Caenorhabditis elegans* (*C. elegans*) was attained from Feng Zeng et al.[21]. Where they extend the existing dataset of Ye et al.[31] by producing new samples. New samples were got from the MethSMRT database consisted of 4mC and non 4mC sites, where each had a length of 41 bp. Two steps were taken, first, in the Methylome Analysis Technical Note, it was shown that the modification QV (modQV) score for the IPD ratio had remarkably dissimilar from the estimated. With the modQV score of greater than 30, the samples were removed. Next, they used the CD-HIT[32] software to remove the redundancy of bias samples to make sure a biased dataset will not miscalculate the accuracy results. The cut off frequency was used 0.80.

Subsequently, newly acquired samples were integrated with the benchmark dataset which was used in several research works. A dataset formed with the number of 18747 samples, to reduce the similarity between the new and old samples a CD-HIT was used. After that, a new dataset of *C. elegans* prepared with a total of 17808 samples from which 11173 are 4mC samples and 6635 are non 4mC samples.

Distributed feature representation. The nature of raw genomic datasets is considered as complicated and noisy. With this necessity, we focused to apply the computational model for the instinctive feature representation learning approach on genomic data[33]. This method allows for inducing optimum features set and increases the performance of the computational model by reducing the model complexity.

Vector representation of words or word embedding is the most well-known technique in the natural language processing (NLP) operations. Theoretically, it transforms the 1-dimension per word into continuous N-dimension vectors. The first word2vec model was proposed by Mikolov et al.[22] Based on a neural network, resultant outcomes possessed distributed characterize sentences of linguistic words. The aforementioned, technique was much faster than the preceding methods, to train the model for continuous vector space and lower the dimensions. In recent years, the success of NPL has been shown caused by its advantageous applications for instance speech recognition, language assist, and translation devices, which made substantial progression on the word embedding methods. Furthermore, researchers revealed that genetic data can be used as language whether DNA or RNA samples that occur within the structure of the cell[34–36]. Additionally, several biological related complex problems have been successfully demonstrated through NLP approaches[25–28,37].

Corpus development is the first step for implementing the word2vec model, it divides the continuous biological sequence into k length (k-mer)[38] of nucleotides groups to formulate as a word and identify linguistic associations among them. In this work, we have carried out the preprocessing to compose the text corpus from the *C. elegans* genome and then trained the word2vec model. We have produced the corpus by operating the whole *C. elegans* genome assembly (WBcel235/ce11) which is downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/ce11/chromosomes/.

Firstly, the genome assembly was distributed into seven chromosomes (chrI, chrII, chrIII, chrIV, chrM, chrV, chrX) and then each chromosome split into the sequence of 41nt to shape the sentence. A continuous bag-of-words (CBOW) approach has been employed to train the word2vec model. CBOW determines the recent word $w(t)$ by surrounds the context word based on predefined window size which was set as 5. A biological sequence

Parameters	Word2vec learning model
Training approach	CBOW
Corpus	*C. elegans* (WBcel235)
Context words	(2-mer, 3-mer, 4-mer)
Vector size	100
Minimum count	5
Negative sampling	5
Window size	5
Number of epochs	20

Table 1. Parameters of word2vec model which were used in training.

Parameters	Range
Number of convolution layers	[1,2,3,4,5]
Filters in convolution layer	[8, 10, 12, 16, 18, 25, 32, 44, 64, 128]
Filter size	[2, 3, 4, 5, 6, 8, 10, 12]
Number of groups in GroupNormalization layer	[2, 4]
Pool-size in maxpooling layer	[2, 4]
Stride length in maxpooling	[2, 4]
Dropout values	[0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45]

Table 2. Demonstration of hyper-parameter tuning of proposed model.

which is mostly a combination of A, C, G, T nucleotides transforms into the sequence of words by setting the k-mer value. The k was set k = 3 with overlapping which forms a DNA sequence ACGTCAGT into words like ACG, CGT, GTC, TCA, CAG, AGT. Each 3-mer word is indicated by a 100-dimensional vector. We experimented with the word2vec model by different values of k-mer, such as k=2, k=3, k=4. Complete details of the parameters which were used are shown in Table 1.

The proposed model

In this work, adequate deep learning based CNN model was proposed for the prediction of N4-methylcytosine sites of the *C. elegans* genome. CNN has the capability to acquire leading quality features automatically for the classification prediction instead of manually handy crafted like traditional supervised learning methods. Whereas, an assorted CNN model can be made by using handcrafted features. Convolution Neural Network has been utilized in several research areas such as image processing[39,40], natural language processing[41], and computational biology[42–47]. A grid search algorithm was implemented with different hyper-parameters values to obtain the most optimal CNN model during its learning. The range of parameters is demonstrated in Table 2.

In the proposed work, the word2vec feature representation was introduced which is completely different from previous works of N4-methylcytosine. A CNN based word2vec model trained on the optimum model which got from the grid search. The input of the model is $(L - k + 1) \times 100$, where L is the length of the input sequence, k is the value of k-mer and 100 is a dimensional vector for each word in the sequence sentence. The model contains three blocks and each has several layers with diverse parameter ranges to construct the model. Each block comprises a convolution layer (Conv1D) having parameters as a filter number with values of 32, 32, 16, respectively, kernel-size with values of 5, 5, 4, correspondingly and stride with the value of 1 for all. The convolution layer has capability to extract the features by self-activating for the appropriate input. In all the convolution layers, L2 regularization weights and bias used to assure the model by overfitting. The values of both regularizations were set as 0.0001, for all three Conv1D, an exponential linear unit (ELU) utilized as an activation function. Each block Conv1D was followed by a group normalization layer (GN), which condensed the consequences of convolution layers. GN distributes the feature map into the desired numbers of groups and normalizes them within each group. The number of groups was fixed as 4, 4, 2, respectively in each block. Moreover, a max-pooling layer (MaxPooling1D) was applied to minimize the dispensable features after the GN layer and avail to turn down the dimensionality of the features. The pool-sizes were set as 4, 4, 2, correspondingly, and strides were set as 2 for all max-pooling layers in each block. Right after MaxPooling1D, dropout layers were used to avoid the overfitting problem while training the model. It supports by turning off the operations of some hidden nodes by regulating the neurons to zero at the learning process.

After the convolution blocks, a flatten layer was used to unstack the outcomes and squash the features vectors from preceding layers. Furthermore, a fully connected layer (FC) was implemented with the number of 32 neurons and L2 weights and bias regularization was utilized by setting the value at 0.001. In the FC layer, ELU activation was used. Finally, the last FC layer was employed with a sigmoid activation function for the bimodal classification. The sigmoid function helps to squeeze the outcome numbers on the scale of between 0 and 1 and

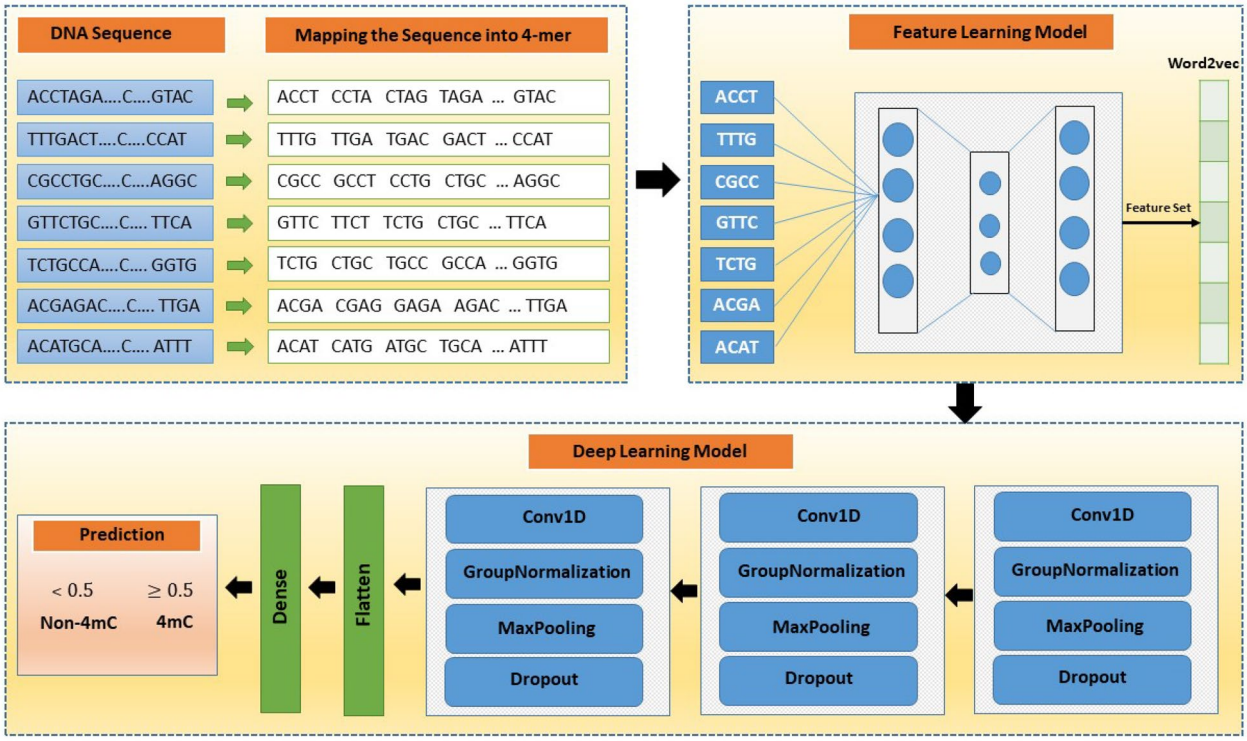

Figure 1. Illustration of complete architecture of 4mCNLP-Deep.

demonstrates the likelihood of acquiring the 4mC and non-4mC sites. Figure 1 shows the detailed architecture of presented CNN model and feature learning model.

The proposed model 4mCNLP-Deep was executed on Keras[48]. Stochastic gradient descent (SGD) optimizer was used with the value of 0.95 for momentum and 0.004 for the learning rate. Binary cross-entropy is deployed as a loss function. We fix the 100 epochs and 32 batch size for the fit function. The call back function was used to storing models and their corresponding weights by calling the checkpoint. While early stopping is used to stops the prediction accuracy at a certain point once validation puts an end to improve. The early stopping patience level was set as 20.

Performance evaluation metrics

The effective performance of the 4mCNLP-Deep model was measured by k-fold cross-validation, we used three different values for the k such as 3 fold, 5 fold, and 10 fold cross-validation to carry out the preeminent identification. Cross-validation is used to estimate the explicit achievement of the desired model by using the resampling method. The whole dataset merges and splits into k number of clusters, each cluster carries eight folds for training, one for validation, one for testing. The proposed CNN model was trained and tested k intervals. There are four metrics to evaluate the performance of the model such as Accuracy (ACC), Mathew's correlation coefficient (MCC), Specificity (Sp), and Sensitivity (Sn) with the given mathematical formulation[49–51].

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$SN = \frac{TP}{TP + FN} \tag{4}$$

Where TN and TP represent as true negative and true positive having the correct number of identified sequences related to 4mC and non-4mC, respectively. Whereas, FN and FP denote as false negative and false positive taking false number of identified sequences for 4mC and non-4mC, respectively. Moreover, the receiver operating characteristics curve (ROC) and area under the ROC curve (AUC) were also deployed to demonstrate the achievement of the presented deep learning model.

Value of k-mer	No. of folds	ACC	MCC	Sp	Sn	AUC
4mCNLP-Deep (2-mer)	3-fold	0.9323	0.8542	0.8973	0.9528	0.9730
	5-fold	0.9343	0.8585	0.9027	0.9528	0.9742
	10-fold	0.9382	0.8668	0.9034	0.9586	0.9758
4mCNLP-Deep (3-mer)	3-fold	0.9354	0.8608	0.8996	0.9563	0.9731
	5-fold	0.9427	0.8766	0.9141	0.9595	0.9764
	10-fold	0.9455	0.8825	0.9132	0.9643	0.9788
4mCNLP-Deep (4-mer)	3-fold	0.9328	0.8551	0.8974	0.9535	0.9732
	5-fold	0.9446	0.8805	0.9097	0.9650	0.9768
	10-fold	0.9500	0.8922	0.9209	0.9670	0.9798
4mC-Deep CBI	3-fold	0.9294	0.8498	0.8938	0.9486	0.9242

Table 3. Results demonstration of 4mCNLP-Deep with different experimental values compare to state-of-the-art model on benchmark dataset for *C. elegans*.

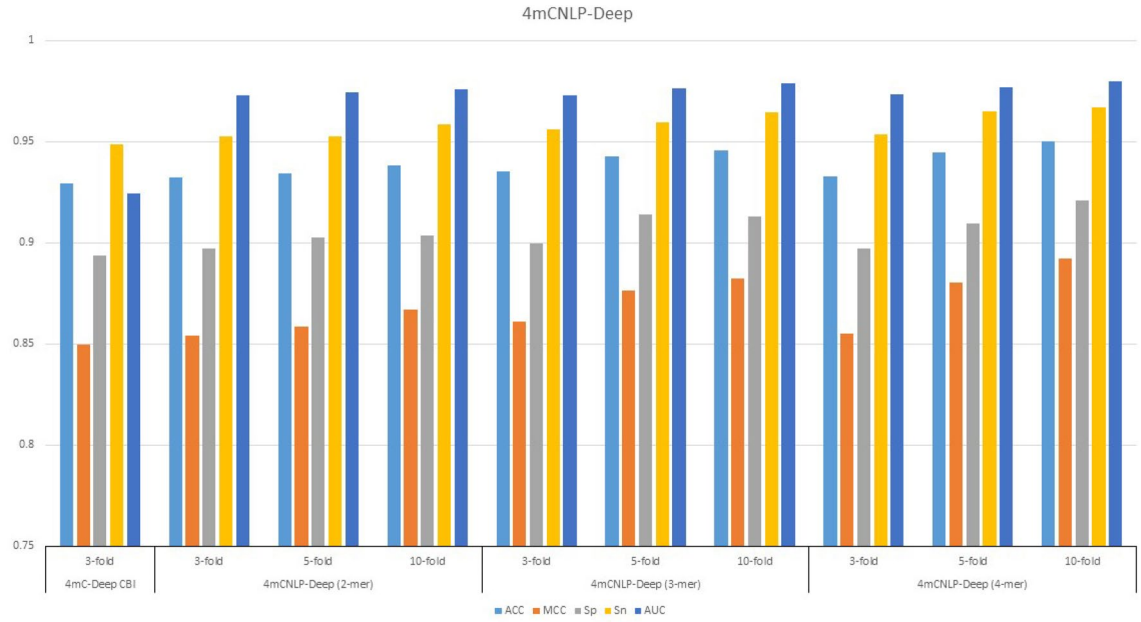

Figure 2. Depiction of performance comparison of 4mCNLP-Deep and existing model with various experimental values.

Results and discussion

A word2vec formulation technique was utilized with different ranges of k-fold cross-validation to predict the N4-methylcytosine by the implemented an optimal predictor to obtain the best performance.

Performance evaluation. In the proposed model 4mCNLP-Deep, we did diverse experiments with a distinct assortment of values for corpus k-mer (2, 3, 4) and k-fold (3, 5, 10) on the *C. elegan* dataset. Each model utilized word2vec with different k-mer values on a various number of folds to check the best performance. For example 2-mer word2vec was implemented with distinct folds of 3, 5, and 10 as cross-validation. As the value of k increases, the disparity in size among the training set and the resampling subset gets shorter. In the resulting model returns immeasurable results. If the difference increase, the bias of the procedure becomes larger and it affects the model outcomes as compared to a large difference. In contrast, the proposed model gives better results in 10 folds of each k-mer of word2vec.

For constructive comparison, we compared our predictor with state-of-the-art model 4mCDeep-CBI[21] and scrutinize the credibility of the model to identify the 4mC and non-4mC sites. The 4mcDeep-CBI applied 3-fold, our model outperformed with 3-mer word2vec using 3-fold and has reported as ACC of 0.93.54, MCC of 0.8608, Sp of 0.89.96, Sn of 0.9563, and AUC of 0.9731 and accomplished increment of 1.1%, 0.6%, 0.58%, 0.77%, and 4.89%, respectively. The detailed experimental results of 4mCNLP-Deep with all ranges have been shown in Table 3. The performance evaluation of 4mCNLP-Deep and state-of-the-art are demonstrated in Fig. 2.

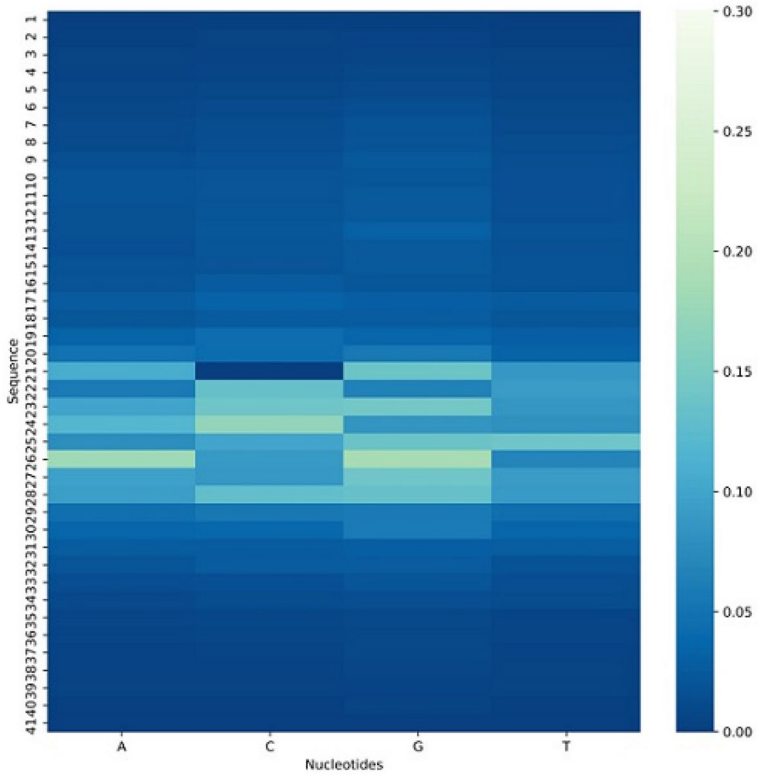

Figure 3. Visualization of heatmap of silico mutation, Where the C nucleotide in the center shows the weak effect on the final prediction as compare to other nucleotides.

Applications for interpreting deep learning model. Deep learning models have the capability to attain the pioneering results but it's laborious to interpret the algorithms as a standard statistical model. In the present study, there are two application established to understand and analyze the deep learning model by visualization techniques.

The first method to decipher a convolution neural network model for computational and statistical biology is silico mutagenesis which was used in various scientific works[29,52]. It is operated by mutating each nucleotide by a single base of sequences with a fixed length of four nucleotide A, C, G, T. In this systematical methodology, the model restore each outcome of resulted mutation and keeps the output as an absolute difference. Further, taken out the aggregated average of mutated predictions for the complete dataset.

For the mutated alterations, a heat map was implemented to show the impact of mutation. Figure 3, demonstrates the visualization of the mutation on the C. elegans dataset as an indigenous feature during the model's learning phase. As it can be shown that the influence of mutation is less in the center of sequence on ultimate identification due to C nucleotide which is symbolized as N4-methylcytosine modification. The recasting of C nucleobase can intimate the unique kind of gene modification. In comparison, more influence of mutation can be shown on the other sides of the heat map which leads to represent the modification of nucleotides can change the results of cytosine recognition.

The second technique to interpret the deep learning based CNN model is the saliency map which helps to identify the most influential features of the sequence by the help of the gradient of the model for final prediction. It points out the most significant characteristics in the samples to classify the class related to the modification, several investigators used in their work[30,53]. For the envision, the efficiency of each location was derived by pointwise product of the saliency map through the vector encoding to obtain the imitative values of actual nucleotide characters of sequences such as A, C, G, T. We experimented by splitting the samples into 3-mer chunks across all the sequence by the formulation of $L - K + 1$. The effect of tri-nucleotide letters at each place of the whole *C. elegans* dataset's outcome result can be shown in Figure 4. In the middle of the words, the CAA motif has a significant magnitude value which is illustrating the utmost vital features in the sample for the identification of the CNN model. The base 4mCNLP-Deep also specifying the current gene modification related to N4-methylcytosine.

Application of clinical research. *C. elegans* consider as a hereditary model organism used in the study of physiology, to sum up, the aspects of human disease. It is a widely applied non-mammalian animal model that is well proven for the highly versatile experiments for research on genetics, development, aging, muscle physiology, and radiobiology. The main purpose of the clinical research on *C. elegans* is to identify such type of genes which provides information about the mechanism of human disease development and also helps to enhanced diagnosis and treatment. Clinical experiments are costly and time-consuming when used for the whole genome. Therefore

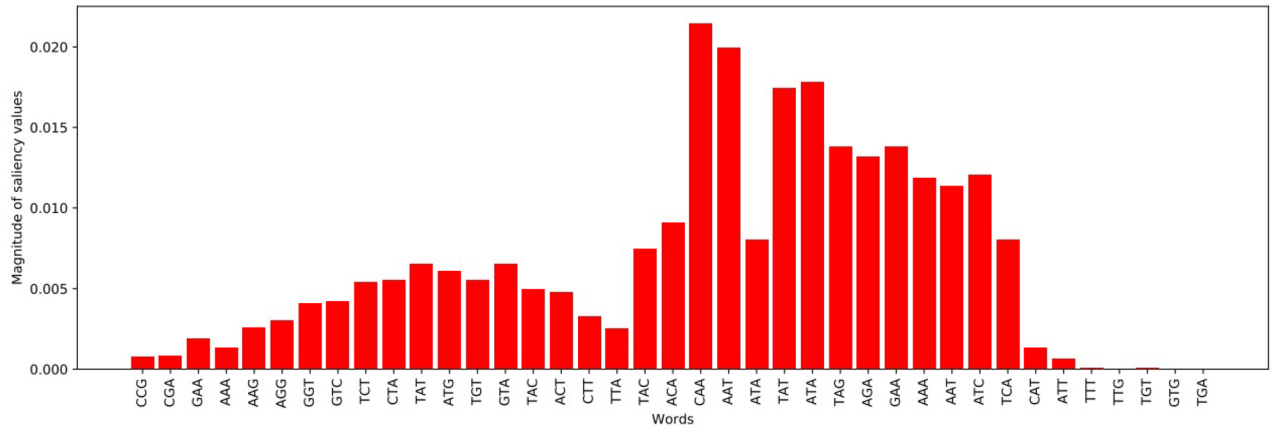

Figure 4. Illustration of saliency map of each 3-mer words having significant importance on the model outcome.

the computational approaches are utilized for several experiments that reduce the value and time. The presented model is made by considering this necessity, to ease clinical experimental biologist. They can easily detect the N4-methylcytosine site and then used further for the development of human disease identification or treatment. The proposed method already proved and contributing to the detection of different genetic sites in the genome, word2vec, and CNN made a big impact and utilized by several investigators[25–28,37] to contribute as a better solution. This type of applications helps the biologist by freely accessible online tools.

Conclusion

In the presented work, we introduced a persuasive computational biological model which is known as 4mCNLP-Deep for the prediction of 4mC and non-4mC sites. The expanded dataset of *C. elegans* was utilized for training and testing the deep learning model. Furthermore, a unique encoding technique was applied to transform sequences into the vectors representation by using a word embedding for the deep learning model. An optimal CNN algorithm was deployed after getting the best settings by exploiting hyperparameter tuning in a grid search. We performed several experiments for the values of k-mer in corpus and cross-validation for k-fold. All the experimental results are outperforming from the existed model. However, for rational comparison, 3-mer word2vec on 3-fold cross-validation has shown a prominent result which indicates the effective performance and high intelligence of the model for predicting the N4-methylcytosine sites. In the proposed work, five evaluation metrics were used like ACC, MCC, Sp, Sn, and AUC to measure the robustness and productivity of identification. Lastly, two diverse approaches named silico mutagenesis and saliency map were employed to interpret our deep learning based CNN model and understand the biological significance of gene modification. 4mCNLP-Deep can be appropriated by the biologists and create a high impact to identify a different kind of gene modification specifically N4-methylcytosine and specify the brain related diseases or development irregularities. In the future, we will expand the model complexity with a proper and efficient way for the prediction of all kinds of gene modification which will make a huge contribution in the field of bioinformatics and computational biology. Moreover, we developed the online webserver http://nsclbio.jbnu.ac.kr/tools/4mCNLP-Deep/, for the experimental researchers to get the results easily.

Received: 1 October 2020; Accepted: 14 December 2020
Published online: 08 January 2021

References

1. Chatterjee, A. & Eccles, M. R. Dna methylation and epigenomics: new technologies and emerging concepts (2015).
2. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
3. Breiling, A. & Lyko, F. Epigenetic regulatory functions of dna modifications: 5-methylcytosine and beyond. *Epigenet. Chromatin* **8**, 1–9 (2015).
4. Liang, Z. *et al.* Dna n6-adenine methylation in arabidopsis thaliana. *Dev. Cell* **45**, 406–416 (2018).
5. He, W., Jia, C. & Zou, Q. 4mcpred: machine learning methods for dna n4-methylcytosine sites prediction. *Bioinformatics* **35**, 593–601 (2019).
6. Suzuki, M. M. & Bird, A. Dna methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
7. Robertson, K. D. Dna methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
8. Janulaitis, A., Klimašauskas, S., Petrušyte, M. & Butkus, V. Cytosine modification in dna by bcni methylase yields n 4-methylcytosine. *FEBS Lett.* **161**, 131–134 (1983).
9. Cheng, X. Dna modification by methyltransferases. *Curr. Opin. Struct. Biol.* **5**, 4–10 (1995).
10. Chen, K., Zhao, B. S. & He, C. Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.* **23**, 74–85 (2016).
11. Doherty, R. & Couldrey, C. Exploring genome wide bisulfite sequencing for dna methylation analysis in livestock: a technical assessment. *Front. Genet.* **5**, 126 (2014).
12. Flusberg, B. A. *et al.* Direct detection of dna methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461 (2010).
13. Boch, J. & Bonas, U. Xanthomonas avrbs3 family-type iii effectors: discovery and function. *Annu. Rev. Phytopathol.* **48**, 419–436 (2010).

14. Buryanov, Y. I. & Shevchuk, T. Dna methyltransferases and structural-functional specificity of eukaryotic dna modification. *Biochemistry (Moscow)* **70**, 730–742 (2005).
15. Chen, W., Yang, H., Feng, P., Ding, H. & Lin, H. idna4mc: identifying dna n4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **33**, 3518–3523 (2017).
16. Wei, L., Chen, H. & Su, R. M6apred-el: a sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids* **12**, 635–644 (2018).
17. Wei, L., Zhou, C., Chen, H., Song, J. & Su, R. Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34**, 4007–4016 (2018).
18. Manavalan, B., Basith, S., Shin, T. H., Wei, L. & Lee, G. Meta-4mcpred: a sequence-based meta-predictor for accurate dna 4mc site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* **16**, 733–744 (2019).
19. Wei, L. *et al.* Iterative feature representations improve n4-methylcytosine site prediction. *Bioinformatics* **35**, 4930–4937 (2019).
20. Wahab, A., Mahmoudi, O., Kim, J. & Chong, K. T. Dnc4mc-deep: Identification and analysis of dna n4-methylcytosine sites based on different encoding schemes by using deep learning. *Cells* **9**, 1756 (2020).
21. Zeng, F., Fang, G. & Yao, L. A deep neural network for identifying dna n4-methylcytosine sites. *Front. Genet.* **11**, 209 (2020).
22. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv :1301.3781 (2013).
23. Duong, D., Ahmad, W. U., Eskin, E., Chang, K.-W. & Li, J. J. Word and sentence embedding tools to measure semantic similarity of gene ontology terms by their definitions. *J. Comput. Biol.* **26**, 38–52 (2019).
24. Habibi, M., Weber, L., Neves, M., Wiegandt, D. L. & Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**, i37–i48 (2017).
25. Hamid, M.-N. & Friedberg, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* **35**, 2009–2016 (2019).
26. Khanal, J., Tayara, H. & Chong, K. T. Identifying enhancers and their strength by the integration of word embedding and convolution neural network. *IEEE Access* **8**, 58369–58376 (2020).
27. Nazari, I., Tahir, M., Tayara, H. & Chong, K. T. in6-methyl (5-step): identifying rna n6-methyladenosine sites using deep learning mode via chou's 5-step rules and chou's general pseknc. *Chemometr. Intell. Lab. Syst.* **193**, 103811 (2019).
28. Oubounyt, M., Louadi, Z., Tayara, H. & Chong, K. T. Deep learning models based on distributed feature representations for alternative splicing prediction. *IEEE Access* **6**, 58826–58834 (2018).
29. Raimondi, D. *et al.* Large-scale in-silico statistical mutagenesis analysis sheds light on the deleteriousness landscape of the human proteome. *Sci. Rep.* **8**, 1–11 (2018).
30. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013).
31. Ye, P. *et al.* Methsmrt: an integrative database for dna n6-methyladenine and n4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkw950 (2016).
32. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
33. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
34. Searls, D. B. String variable grammar: a logic grammar formalism for the biological language of dna. *J. Logic Program.* **24**, 73–102 (1995).
35. Yandell, M. D. & Majoros, W. H. Genomics and natural language processing. *Nat. Rev. Genet.* **3**, 601–610 (2002).
36. MECHE, C. E. & Hoffmeyer, J. From language to nature: the semiotic metaphor in biology (1991).
37. Cohen, K. B. & Hunter, L. Natural language processing and systems biology. In *Artificial Intelligence Methods and Tools for Systems Biology*, 147–173 (Springer, 2004).
38. Ng, P. dna2vec: Consistent vector representations of variable-length k-mers. arXiv preprint arXiv:1701.06279 (2017).
39. Ilyas, T., Khan, A., Umraiz, M. & Kim, H. Seek: a framework of superpixel learning with cnn features for unsupervised segmentation. *Electronics* **9**, 383 (2020).
40. Khan, A., Ilyas, T., Umraiz, M., Mannan, Z. I. & Kim, H. Ced-net: crops and weeds segmentation for smart farming using a small cascaded encoder-decoder architecture. *Electronics* **9**, 1602 (2020).
41. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13**, 55–75 (2018).
42. Tahir, M., Tayara, H. & Chong, K. T. irna-pseknc (2methyl): Identify rna 2'-o-methylation sites by convolution neural network and chou's pseudo components. *J. Theor. Biol.* **465**, 1–6 (2019).
43. Tayara, H., Tahir, M. & Chong, K. T. iss-cnn: identifying splicing sites using convolution neural network. *Chemometr. Intell. Lab. Syst.* **188**, 63–69 (2019).
44. Wahab, A., Ali, S. D., Tayara, H. & Chong, K. T. iim-cnn: intelligent identifier of 6ma sites on different species by using convolution neural network. *IEEE Access* **7**, 178577–178583 (2019).
45. Mahmoudi, O., Wahab, A. & Chong, K. T. imethyl-deep: N6 methyladenosine identification of yeast genome with automatic feature extraction technique by using deep learning algorithm. *Genes* **11**, 529 (2020).
46. Rehman, M. U. & Chong, K. T. Dna6ma-mint: Dna-6ma modification identification neural tool. *Genes* **11**, 898 (2020).
47. Alam, W., Ali, S. D., Tayara, H. & to Chong, K. A. cnn-based rna n6-methyladenosine site predictor for multiple species using heterogeneous features representation. *IEEE Access* **8**, 138203–138209 (2020).
48. Chollet, F. *et al.* Keras: deep learning library for theano and tensorflow. https://keras. io/k **7**, T1 (2015).
49. Tayara, H. & Chong, K. T. Improving the quantification of dna sequences using evolutionary information based on deep learning. *Cells* **8**, 1635 (2019).
50. Tahir, M., Tayara, H. & Chong, K. T. ipseu-cnn: identifying rna pseudouridine sites using convolutional neural networks. *Mol. Ther. Nucleic Acids* **16**, 463–470 (2019).
51. Park, S., Wahab, A., Nazari, I., Ryu, J. H. & Chong, K. T. i6ma-dnc: prediction of dna n6-methyladenosine sites in rice genome based on dinucleotide representation using deep learning. *Chemometr. Intell. Lab. Syst.* **204**, 104102 (2020).
52. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
53. Lanchantin, J., Singh, R., Wang, B. & Qi, Y. Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. In *Pacific Symposium on Biocomputing 2017*, 254–265 (World Scientific, 2017).

Acknowledgements

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1A2C2005612), in part by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044816), and in part by Human Resources Program in Energy Technology" of the Korea Institute of Energy Technology Evaluation

and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20204010600470).

Author contributions

A.W. and H.T. Conceptualization, A.W., H.T. and Z.X. Methodology, A.W. Software Implementation, A.W., H.T. and Z.X. Validation, A.W., H.T. and Z.X. Investigation, A.W. Writing Original draft preparation, K.T.C. Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.X. or K.T.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.