MR-DNA: Flexible 5mC-Methylation-Site Recognition in DNA Sequences using Token Classification

WENHUAN ZENG and DANIEL H. HUSON, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Germany

The process of DNA 5-methylcytosine modification has been widely studied in mammals and and plays an important role in epigenetics. Several computational approaches have been developed to aid the identification of methylation sites. In this study, we introduce a novel deep-learning framework MR-DNR that aims at predicting specific methylation sites located in gene promoter regions. The idea is to adapt the name-entity recognition approach to methylation-site prediction. MR-DNA is trained on a stacked model architecture that consists of a pre-trained MuLan-Mehtyl-DistilBERT language model and conditional random field algorithms. The resulting fine-tuned model achieves an accuracy of 95.4% on an independent test dataset. A key advantage of this formulation of the methylation-site identification task is that the input DNA sequence can be of any length, unlike previous methods that predict methylation state on short, fixed-length DNA sequences. For training and testing purposes, we provide a database of DNA sequences containing verified 5mC-methylation sites, obtained from eight human cell lines in ENCODE. Data and code are available at https://github.com/husonlab/MR-DNA.

CCS Concepts: • **Computing methodologies → Information extraction**.

Additional Key Words and Phrases: DNA methylation, natural language understanding, transferring learning, bioinformatics

1 INTRODUCTION

DNA-methylation plays an important role in epigenetics. Much attention has been given to 5-methylcytosine (5-mC) modifications in which a cytosine is methylated at its fifth position. In mammals, this process is associated with many different biological processes, such as gene silencing, X-chromosome inactivation, genomic imprinting and human cancer development [1, 2, 3, 4].

Much experimental work has been done that aimed at quantifying 5mC methylation in mammalian gene promoters, in particular, using bisulfite sequencing [5, 6]. Recently, several machine-learning approaches have been developed to predict whether the central base of a short DNA sequence is methylated, all showing reliable performance [7, 8, 9, 10].

These computational approaches use classic machine learning techniques, such as deep feed-forward neural networks (deep FNN) [7], bidirectional long short-term memory (BiLSTM) networks [8], or an ensemble framework [9, 10], which are applied to a combination of features extracted from the DNA sequences. These methods are all designed to predict whether a presented DNA sequence of fixed length 41 is methylated at the middle, or not. Thus, these methods do not directly apply to the main practical problem of interest, namely the prediction of individual methylation sites in a longer sequence.

Hence, there is the need for a machine-learning approach that operates directly on full-length DNA sequences. To address this, here we model DNA sequences as "language", where each 3-mer is a token, and phrase methylation-site prediction as a token classification task, a fundamental task in natural language understanding (NLU). The named-entity recognization (NER) task is used for information extraction (IE), for example to locate characters, events, organizations, or locations [11, 12], or to recognize other types of entities [13, 14] in a text. The approach has been applied to biomedical problems [15, 16, 17].

Authors' address: Wenhuan Zeng; Daniel H. Huson, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, Tuebingen, Germany.

Traditional statistical models [18, 19, 20, 21, 22] are gradually being replaced by transformer-based language models or other neural networks [23, 17, 24]. The stacked model architecture, which allows stacking of additional layers, such as LSTM and CRF, to a pre-trained transformer-based language model, has significantly improved the ability to capture dependencies between entities and corresponding labels [25, 26, 27].

As mentioned above, previous approaches address the question whether the middle based of a DNA sequence of length 41 is methylated or not. In our approach, we take advantage of the fact that NER assigns a label to each token and thus allows to predict methylation for individual bases in longer sequences.

The MuLan-Methyl framework [28] consists of five transformer-based language models, each trained on the iDNA-MS database [29], a comprehensive dataset containing DNA methylation sequences for three methylation types and twelve taxonomic lineages. Here, we propose a novel model architecture, MR-DNA (Methylation-site Recognition in DNA) that consists of one the five MuLan models, MuLan-Methyl-DistilBERT, and a conditional random fields (CRF) algorithm, which enhances the model prediction ability. we use this particular language model because it outperforms the other constituent of MuLan-Methyl for 5hmC-methylation site recognition in terms of both prediction performance and computational efficiency.

To evaluate the performance of MR-DNA, we created a database of DNA sequences, each of length 1000 bp, that correspond to gene promoter regions from eight human cell line projects in the ENCODE database [30, 31], with differing numbers of experimentally-verified methylation sites. We further cut each such DNA sequence into small pieces according to a set stride during model training. The performance of the fine-tuned MR-DNA model was evaluated under the exact-match evaluation criterion [12].

Our main contributions are:

- This is the first study to phrase methylation-site recognition as a named entity recognition (NER) task.
- We obtained a new model architecture by combining the statistical modelling method CRF with the fine-tuned MuLan-Methyl-DistilBERT model, obtaining promising performance on the test dataset.
- We provide a database of DNA gene-promoter sequences (each 1000 bp length) of varying numbers of verified methylation sites.
- The MR-DNA method aims at recognizing methylated sites on the sequences without length limitation.

2 METHODS

2.1 Database construction

The MR-DNA database was setup from DNA methylation-site information of eight human cell lines (K562, HepG2, GM12878, HeLa-S3, A549, SK-N-SH, H1, GM23248) downloaded from the ENCODE portal [30, 31], with the following identifiers: ENCSR765JPC, ENCSR786DCL, ENCSR890UQO, ENCSR550RTN, ENCSR481JIW, ENCSR145HNT, ENCSR617FKV, ENCSR625HZA. Each experiment provided the methylation status at CpG, CHH, CHG locations, determined using whole-genome bisulfite sequencing (WGBS), in BED format. We filtered the BED files to keep only methylation sites that are on the same strand as the gene, have sequencing coverage of 10 or more, and for which 100% of the assembled reads are reported as methylated. For each of the human cell lines, the filtered BED files where split in a ratio of 9:1 to generate training and test sets, respectively.

The start and end positions of gene promoter regions were determined from the GRCH38 annotation file downloaded from NCBI as the 1000 bp upstream of a transcription start position, and the corresponding DNA sequences were

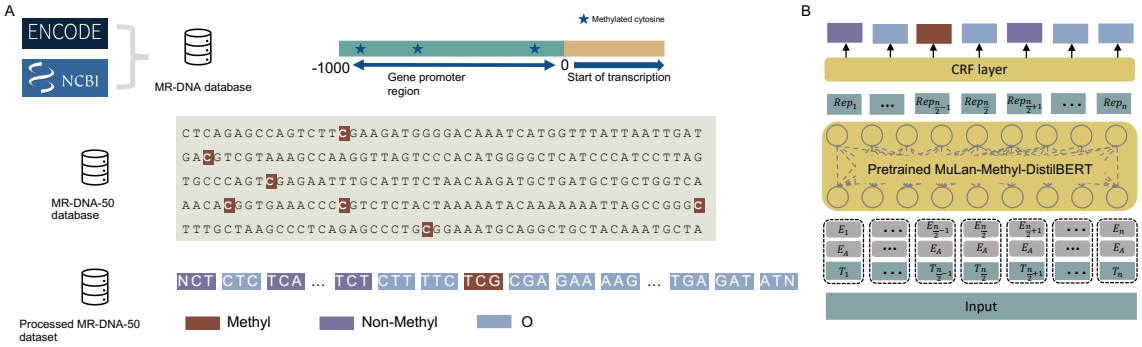

Fig. 1. A. Database creation. For eight human cell lines in ENCODE, we extracted 1000 bp upstream gene promoter regions, and annotated cytosines by their reported methylation state, and this data is referred to as the MR-DNA database. The MR-DNA-50 database is obtained from the sequences using a window size of 50 and stride of 25. The *processed* MR-DNA-50 database is obtained by extracting 3-mers from the each sequence in the MR-DNA-50 database and labeling each 3-mer as methylated, if its central nucleotide is so labeled.
B. Model structure. MR-DNA is a stacked model that uses a CRF layer on top of pretrained MuLan-Methyl-DistilBERT, assigning the most probable category to each token.

extracted from the GRCH38 genome reference. The filtered BED files where used to annotate the methylated sites in the promoter regions.

We first populated the MR-DNA database with the described annotated promoter sequences of length 1000 bp. Then, for training purposes, we constructed a second database, MR-DNA-50, in which all sequences where cut into short sequences of length 50 bp (50-mers), using a stride of 25. Any such 50-mer that is annotated with at least one methylation site was used to generate the MR-DNA-50 training dataset and the MR-DNA-50 test dataset, respectively (see Fig. 1A).

2.2 Entity category annotation

In preparation of the NER task, we processed and annotated all 50-mers as follows. First, each 50-mer was extended in each direction by a single "N" and each resulting 52-mer was converted into a sentence-like sequence consisting of 50 (=52-3+1) overlapping 3-mers obtained using a sliding window, separated by whitespace. Second, imitating the named-entity naming rule, for each 3-mer, we attached the label "Methyl", 'Non-Methyl", or "O", depending on whether the middle base is a methylated cytosine, a non-methylated cytosine, or neither, respectively.

In result, in both the processed MR-DNA-50 training dataset and test dataset, each sample is a sentence consisting of 50 3-mers separated by whitespace, where each element is tagged as "Methyl", "Non-Methyl", or "O", depending on its methylation status reported in the MR-DNA database (see Fig. 1A).

The median value of the number of methylation sites for each sample in the initial MR-DNA-50 training dataset is 1; thus the number of "Methyl" entities is smaller than the number of "Non-Methyl" entities, which in turn is much smaller than the number of "O" entities (see Fig. 2B-C), and so the distribution of labels is skewed.

2.3 Pre-trained MuLan-Methyl-DistilBERT

Our proposed framework, MR-DNA, consists of a token-level classifier based on pre-trained MuLan-Methyl-DistilBERT[28], followed by a Linear-Chain CRF as its decoder layer (see Fig. 1B).

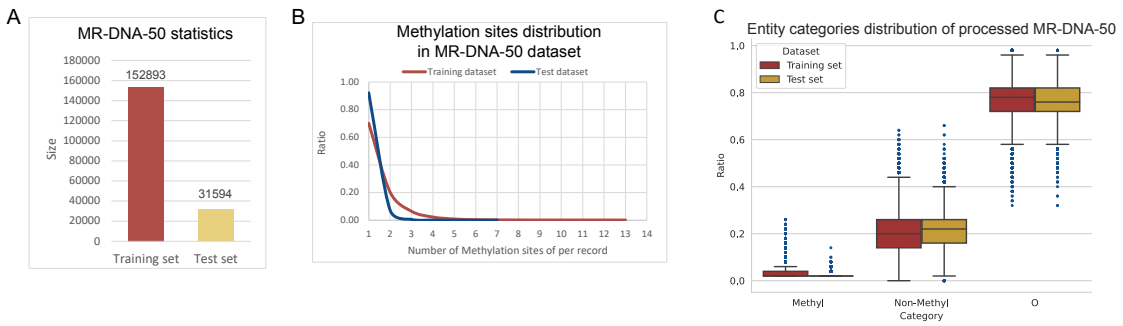

Fig. 2. For both the training and test dataset, we show (A) a comparison of their sizes, (B) the distribution of ratios of methylation sites per sample, and (C) a box-plot of the entity category distributions.

The MuLan-Methyl-DistilBERT classifier is an adaption of DistilBERT [32] to the problem of DNA-methylation identification, obtained by pre-training the masked language modelling (MLM) task on a custom corpus consisting of both DNA methylation and taxonomy data, and then fine-tuning on a DNA methylation-site dataset. The motivation here is that sequence representation vectors encoded by MuLan-Methyl-DistilBERT that were trained taking domain adaption in account can better capture the potential semantic relationships and information between tokens.

MuLan-Methyl-DistilBERT uses the same basic neural network architecture as DistilBERT, performing knowledge distillation of Bidirectional Encoder Representations from Transformers (BERT) [33], while reducing the number of transformer layers and adjusting pre-training tasks. The training objective of DistilBERT is a linear combination of distillation loss $\mathcal{L}_{ce}$, language modelling mask loss $\mathcal{L}_{mlm}$, and cosine-distance loss $\mathcal{L}_{cos}$ [32] as (1):

$$\mathcal{L} = \alpha_{ce} \times \mathcal{L}_{ce} + \alpha_{mlm} \times \mathcal{L}_{mlm} + \alpha_{cos} \times \mathcal{L}_{cos}. \tag{1}$$

We will now describe how to obtain representation vectors. Note that the pre-trained MuLan-Methyl-DistilBERT was trained on a custom corpus and that the vocabulary of the corresponding custom tokenizer contains all combinations of A, T, G, C of multiple lengths as well as words related to taxonomy lineages. However, the DNA sequences in the processed MR-DNA-50 dataset contain an additional character, N. So, the first step is to expand the vocabulary size of the original tokenizer from 25,000 to 25,032, where the additional 32 words are the combinations of A, T, G, C, N of length 3 that occur in the processed MR-DNA-50 training dataset. The embedding dimension of the pre-trained language model was expanded accordingly.

After tokenizing the MR-DNA-50 training dataset using the expanded tokenizer, we employ the updated pre-trained MuLan-Methyl-DistilBERT as an encoder to obtain a embedding representation vector in a three-dimensional array $W \in \mathbb{R}^{1\times t\times 768}$, for each input sequence $X = \{x_1, x_2, \cdots, x_t\}$, where t is the number of tokens.

2.4 Linear-chain CRF layer

For each input sequence, the embedding representation vectors W (generated by the encoder) serves as the input for the linear-chain CRF algorithm, for which the the posterior probability y of the input sequence is defined as (2):

$$p(y|W) = \frac{1}{Z(W)} \exp\left\{\sum_{n=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_n, y_{n-1}, W_n)\right\}, \tag{2}$$

where $\{f_k(y_n, y_{n-1}, W_n)\}_{k=1}^{K}$ is a set of real-valued feature functions, $\{\lambda_k\} \in \mathbb{R}^K$ is a parameter vector, and $Z(W)$ is a normalization factor that takes the form

$$Z(W) = \sum_{y} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_n, y_{n-1}, W_n)\right\}. \tag{3}$$

The objective for parameter learning is to maximize the conditional likelihood of the training data given by

$$l(\theta) = \log p(y, W) = \sum_{n=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_n, y_{n-1}, W_n) - \log Z(W), \tag{4}$$

and the most probable assignment for each element in the input sequence

$$y^* = \arg\max_{y} p(y|W) \tag{5}$$

is further decoded by the Viterbi algorithm.

2.5 Focal loss function

Because of the the skewed distribution of entity categories, the ability of the model to predict the minority entity is comparatively weak. To mitigate this, MR-DNA uses α-variant focal loss[34] as its loss function. This approach employs modulation terms γ>0 to reduce the relative loss for entities well-classified during model training and concentrate on challenging entity categories. The loss is defined as

$$\mathcal{L}_{Focal} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t), \tag{6}$$

where α_t is a hard hyper-parameter containing a list of weights corresponding to each category. MR-DNA achieves the best classification performance when the weights for the labels “O", “Methyl" and“Non-Methyl" are set to 0.1, 0.7, and 0.2, respectively.

3 EXPERIMENTS AND RESULTS

3.1 Training MR-DNA

MR-DNA is implemented in Python 3.10, primarily using the Pytorch, Huggingface[35], and pytorch-crf packages. It was run on a Linux Virtual Machine (Ubuntu 20.04 LTS) equipped with 4 GPUs provided by de.NBI (flavor: de.NBI RTX6000 4 GPU medium). MR-DNA was trained on the above-mentioned processed MR-DNA-50 training and validation dataset, configured with 32 training epochs, 64 batch sizes for each GPU, that is 280,352 training steps in total, 3e-5 learning rate with 100 steps warmup, and 0.01 decay weight.

3.2 Experiment evaluation

Our study uses the exact-match evaluation criterion to assess MR-DNA performance, instead of the seqeval [36] framework designed for sequence labelling evaluation. We report the following evaluation metrices based on the true and predicted labels:

$$
\begin{aligned}
Precision &= \frac{TP}{TP+FP}, \\
Recall &= \frac{TP}{TP+FN}, \\
F1-score &= \frac{TP}{TP+\frac{1}{2}(FP+FN)}, \\
Accuracy &= \frac{TP+TN}{TP+FN+TN+FP},
\end{aligned} \tag{7}
$$

where TP, FP, TN, and FN are the true positive, false positive, true negative and false negative counts, respectively.

3.3 Experimental results

We investigated the ability of MR-DNA to recognize methylation sites on the processed MR-DNA-50 test data, where each sample is a sentence of 50 whitespace-separated entities (3-mers), each with one of three possible annotations ("Methyl", "Non-Methly" or "O"). We used accuracy, F1-score, recall and precision to quantify the ability of MR-DNA to predict the label of each entity (3-mer). As this is a multiple classification task, we used macro averages in the calculation of the performance metrics.

In addition, to investigate the effect of sequence length on classification performance, we trained MR-DNA on a processed MR-DNA-100 and MR- DNA-200 training dataset, using sentences of length 100 and 200, respectively. As shown in Table. 1, MR-DNA trained on the MR-DNA-50 outperforms the models trained on the two datasets of longer sequence lengths regarding F1-score, recall, and precision. Confusions matrices (see Fig. 3) for MR-DNA-50, -100 and -200, show that MR-DNA trained on the former outperforms the model trained on the two latter, for the "Methyl" category.

Table 1. MR-DNA performance evaluation, MR-DNA is trained on the three datasets with different sequence length, and its performance is evaluated on the corresponding test dataset by accuracy, F1-score, recall and precision.

Dataset	Accuracy	F1-score	Recall	Precision
MR-DNA-50	0.9541	0.8286	0.8339	0.8306
MR-DNA-100	0.9585	0.7943	0.7963	0.7984
MR-DNA-200	0.9619	0.7488	0.7251	0.7778

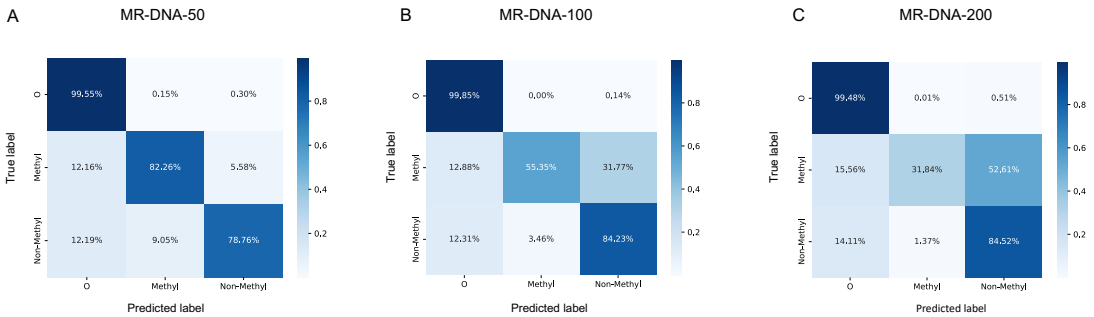

Fig. 3. Normalized confusion matrix visualization of the model trained on (A) MR-DNA-50 training dataset, (B) MR-DNA-100 training dataset and (C) MR-DNA-200 training dataset, and evaluated on the correpsonding test dataset.

4 CONCLUSIONS AND DISCUSSION

The study of 5mC modification in mammalian genomes helps the understanding of several biological processes, such as gene expression control. Previous computational approaches based on machine learning algorithms for identifying methylation sites are only able to categorize DNA sequences with a predetermined length rather than detect methylation sites on a single nucleotide, resulting in limited practical use.

Here, we presented MR-DNA, a novel framework based on natural language understanding technology, which transforms a naive classification task for DNA sequences into a NER problem with the ability to identify the methylation status of each entity. This method shows how machine learning approaches may be used to determine the methylation status of a specific cytosine, and allowing the application for methylation site identification on longer DNA sequences such as chromosomes or assembled contigs.

The performance of MR-DNA was evaluated using four difference metrics and all four indicate that MR-DNA-50, in particular, shows high reliability. We illustrated how one might combine a pretrained language model and a classic statistical algorithm to improve predictive ability. We used the Focal Loss approach to effectively alleviate the negative impact of skewed data on minority class prediction performance.

MR-DNA is the first study to transfer the research object of the NER concept from human language to biological sequences and to adapt the recognition target to the DNA methylation scenario, which aims to recognize the entity of “methylation”, “non-methylation” and “other nucleotide” in DNA sequences, instead of “persons”, “locations”, and “organizations”, in text.

Additionally, this study offers a custom MR-DNA database generated by obtaining methylation status from eight human cell lines in ENCODE database, each record consists of 1000 bp in the original MR-DNA database. MR-DNA was trained on the processed dataset by extracting different sequence lengths following self-define standards. The MR-DNA database is provided as a benchmark dataset for researchers interested in optimizing methylation-site detection.

We would like to emphasize that, for a sequence of arbitrary length n, a prediction is made by partitioning the sequence into $n - 50 + 1$ individual 50-mers and then applying MR-DNA to each such sequence.

ACKNOWLEDGMENTS

We acknowledge the support of the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

REFERENCES

[1] Chongyuan Luo, Petra Hajkova, and Joseph R Ecker. “Dynamic DNA methylation: In the right place at the right time”. In: *Science* 361.6409 (2018), pp. 1336–1340.

[2] En Li and Yi Zhang. “DNA methylation in mammals”. In: *Cold Spring Harbor perspectives in biology* 6.5 (2014), a019133.

[3] Marion Turpin and Gilles Salbert. “5-methylcytosine turnover: Mechanisms and therapeutic implications in cancer”. In: *Frontiers in Molecular Biosciences* 9 (2022).

[4] Christoph Schmutte, Peter A Jones, et al. “Involvement of DNA methylation in human carcinogenesis”. In: *Biological Chemistry* 379.4 (1998), pp. 377–388.

[5] Yuanyuan Li and Trygve O Tollefsbol. "DNA methylation detection: bisulfite genomic sequencing analysis". In: *Epigenetics Protocols* (2011), pp. 11–21.

[6] Nongluk Plongthongkum, Dinh H Diep, and Kun Zhang. "Advances in the profiling of DNA modifications: cytosine methylation and beyond". In: *Nature Reviews Genetics* 15.10 (2014), pp. 647–661.

[7] The-Anh Tran, Dinh-Minh Pham, Yu-Yen Ou, et al. "An extensive examination of discovering 5-Methylcytosine Sites in Genome-Wide DNA Promoters using machine learning based approaches". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.1 (2021), pp. 87–94.

[8] Xin Cheng et al. "BiLSTM-5mC: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide DNA promoters". In: *Molecules* 26.24 (2021), p. 7414.

[9] Lei Zhang, Xuan Xiao, and Zhao-Chun Xu. "iPromoter-5mC: a novel fusion decision predictor for the identification of 5-methylcytosine sites in genome-wide DNA promoters". In: *Frontiers in Cell and Developmental Biology* 8 (2020), p. 614.

[10] Xuan Xiao et al. "m5C-HPromoter: An Ensemble Deep Learning Predictor for Identifying 5-methylcytosine Sites in Human Promoters". In: *Current Bioinformatics* 17.5 (2022), pp. 452–461.

[11] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[12] Jing Li et al. "A survey on deep learning for named entity recognition". In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2020), pp. 50–70.

[13] Jeniya Tabassum et al. "Code and named entity recognition in stackoverflow". In: *arXiv preprint arXiv:2005.01634* (2020).

[14] Yuan Chang et al. "Chinese named entity recognition method based on BERT". In: *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*. IEEE. 2021, pp. 294–299.

[15] Leon Weber et al. "HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition". In: *Bioinformatics* 37.17 (2021), pp. 2792–2794.

[16] Bosheng Song et al. "Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison". In: *Briefings in Bioinformatics* 22.6 (2021), bbab282.

[17] Usman Naseem et al. "Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–7.

[18] Sudha Morwal, Nusrat Jahan, and Deepti Chopra. "Named entity recognition using hidden Markov model (HMM)". In: *International Journal on Natural Language Computing (IJNLC) Vol* 1 (2012).

[19] Li Zhang, Yue Pan, and Tong Zhang. "Focused named entity recognition using machine learning". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, pp. 281–288.

[20] Oliver Bender, Franz Josef Och, and Hermann Ney. "Maximum entropy models for named entity recognition". In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. 2003, pp. 148–151.

[21] Zhenfei Ju, Jian Wang, and Fei Zhu. "Named entity recognition from biomedical text using SVM". In: *2011 5th international conference on bioinformatics and biomedical engineering*. IEEE. 2011, pp. 1–4.

[22] Andrew McCallum and Wei Li. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons". In: (2003).

[23] Jason PC Chiu and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs". In: *Transactions of the association for computational linguistics* 4 (2016), pp. 357–370.

[24] Kai Hakala and Sampo Pyysalo. "Biomedical named entity recognition with multilingual BERT". In: *Proceedings of the 5th workshop on BioNLP open shared tasks*. 2019, pp. 56–61.

[25] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. "Portuguese named entity recognition using BERT-CRF". In: *arXiv preprint arXiv:1909.10649* (2019).

[26] Junlin Sun et al. "Deep learning-based methods for natural hazard named entity recognition". In: *Scientific reports* 12.1 (2022), p. 4598.

[27] Xiaojun Wu et al. "One improved model of named entity recognition by combining BERT and BiLSTM-CNN for domain of Chinese railway construction". In: *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*. IEEE. 2022, pp. 728–732.

[28] Wenhuan Zeng, Anupam Gautam, and Daniel H Huson. "MuLan-Methyl-Multiple Transformer-based Language Models for Accurate DNA Methylation Prediction". In: *bioRxiv* (2023), pp. 2023–01.

[29] Hao Lv et al. "iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes". In: *Iscience* 23.4 (2020), p. 100991.

[30] ENCODE Project Consortium Overall coordination (data analysis coordination) Dunham Ian 2 Kundaje Anshul 3 81 82, Writing group Bernstein Bradley E. 7 34 Birney Ewan Dunham Ian Green Eric D. 35 Gunter Chris 15 Snyder Michael 13, NHGRI project management (scientific management) Pazin Michael J. 36 Lowdon Rebecca F. 36 87 Dillon Laura AL 36 88 Adams Leslie B. 36 Kelly Caroline J. 36 Zhang Julia 36 89 Wexler Judith R. 36 90 Green Eric D. Good Peter J. 36 Feingold Elise A. 36, et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[31] Yunhai Luo et al. "New developments on the Encyclopedia of DNA Elements (ENCODE) data portal". In: *Nucleic acids research* 48.D1 (2020), pp. D882–D889.

[32] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019).

[33] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[34] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[35] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[36] Hiroki Nakayama. *seqeval: A Python framework for sequence labeling evaluation*. Software available from https://github.com/chakki-works/seqeval. 2018. URL: https://github.com/chakki-works/seqeval.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009