Leverage the Explainability of Transformer Models to Improve the DNA 5-Methylcytosine Identification (Student Abstract)

Wenhuan Zeng, Daniel H. Huson

Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tuebingen, Germany
{wenhuan.zeng, daniel.huson}@uni-tuebingen.de

Abstract

DNA methylation is an epigenetic mechanism for regulating gene expression, and it plays an important role in many biological processes. While methylation sites can be identified using laboratory techniques, much work is being done on developing computational approaches using machine learning. Here, we present a deep-learning algorithm for determining the 5-methylcytosine status of a DNA sequence. We propose an ensemble framework that treats the self-attention score as an explicit feature that is added to the encoder layer generated by fine-tuned language models. We evaluate the performance of the model under different data distribution scenarios.

Introduction

DNA methylation happens when a methyl group (CH3) get added to a DNA sequence. The position of the methyl group added determines the type of methylation. In particular, the DNA modification on the fifth position of cytosine (5mC) plays a critical role in gene regulation and is involved in other important biological processes (Breiling and Lyko 2015), occurs in both bacteria and eukaryotes.

There is currently much interest in transformer-based language models. Models such as BERT (Devlin et al. 2018) and its variants perform very well on several natural language processing tasks. In addition to being adapted to a specific domain, like the medical field, transformer-based language model are also transferred to biological sequences, such as DNA sequences (Ji et al. 2021) and protein sequences (Teufel et al. 2022). In MuLan-Methyl (Zeng, Gautam, and Huson 2023), we presented several domain-specific fine-tuned language models for classifying the methylation status of short DNA sequences.

Here, our aim is to use such a model as an encoder to classify 5mC DNA methylation status in mammalian sequences. Previous studies (Abnar and Zuidema 2020) have shown that the self-attention mechanism of transformers can be used to interpret the model and quantify feature contribution, and our work on MuLan-Methyl demonstrated that attention score gives rise to a reasonable measure of feature importance.

Therefore, here we propose a study that uses the information given by attention weights generated by the encoder as an additional layer and adds it to the model structure to aid the model in generating the final prediction.

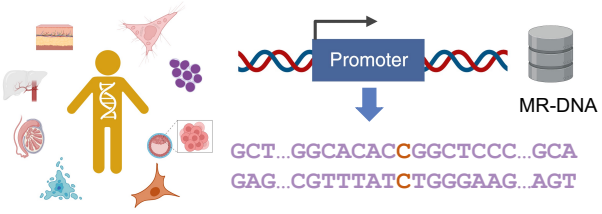

Figure 1: Data description. We used the MR-DNA database, which contains methylation data from eight human cell lines, to build a classification model.

Proposed Method

Dataset and Preprocessing

We downloaded the database presented by MR-DNA (Zeng and Huson 2023), which contains gene promoter regions of length 1000, annotated with 5mC methylation sites from eight human cell lines. We further extracted the methylated cytosine together with 20 bases both before and after the site from the MR-DNA database. These sequences of length 41 formed the positive samples in our dataset. The negative samples were formed in the same way, but centered on non-methylated cytosines (see figure 1). Each input for our ensemble model is a sequence of 3-mers, generated using a sliding window, of length 41-3+1=39. Using the training and test datasets provided by MR-DNA database, we used the described approach to generate a training and test dataset, respectively,

Transfer Learning for 5mC Prediction

Our framework uses two fine-tuned language models, obtained from MuLan-Methyl. MuLan-Methyl consists of five fine-tuned language models, training on a dataset that contains three types of methylation sites (6mA, 4mC, and 5hmC). We use MuLan-Methyl-DistiBERT, and MuLan-Methyl-ELECTRA as our encoders, fine-tuning each of them on our dataset, respectively.

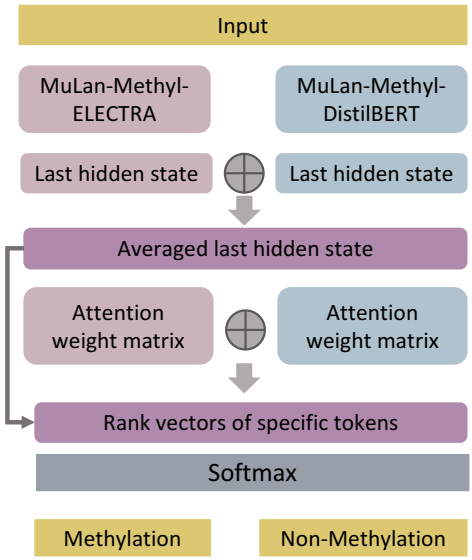

Figure 2: Model architecture.

Custom Layer Generation Using the Output of the Self-Attention Mechanism

On the given input, the fine-tuned language model gives the output of the hidden states and the output of the attention scores from the self-attention mechanism of the transformer architecture. We take the last hidden state as the representative vector of input.

Because the label of a DNA fragment is determined by the center cytosine, after tokenizing the DNA fragments into a sequence of 3-mers, and adding special tokens, the tokens in the 19^{th}, 20^{th} and 21^{st} positions contain the target cytosine.

To utlize all the information of the attention weight matrix, we rank the layer-wise attention weights that are assigned to the token [CLS] and extract the rank position of the above-mentioned three tokens, since both of the language models we employed contain multiple transformer layers, we then combine the results from each layer to a rank sequence of length $3 \times 12 = 36$. Finally, we customize a layer for generating the average rank sequences of attention weights from two models, then combine its output with the average of the last hidden state from both models to form our framework. We call our proposed framework EA-5mC, which is summarized in Figure 2.

Experimental Results

We trained and validated our model on the processed training dataset. The ratio of positive to negative labels during model training was 1:1. Model evaluation was conducted on the independent test dataset. We compared our proposed framework against a naive ensemble of the two models, which we considered the baseline. First, we evaluated model performance on the balanced dataset, where the ratio of the positive and negative samples is 1:1, EA-5mC outperforms the baseline regarding Accuracy, AUC and F1-score. Moreover, since the ratio of methylation cytosine and non-methylation cytosine is highly imbalanced in practice, we then adjusted the ratio of positive and negative samples to 1:10. Here, EA-5mC outperforms the baseline regarding accuracy and F1-score, (see Table 1)

Pos:Neg	Model	Accuracy	AUC	F1-score
1:1	EA-5mC	**0.9463**	**0.9674**	**0.9489**
	Baseline	0.9461	0.9669	0.9487
1:10	EA-5mC	**0.9033**	0.9647	**0.6514**
	Baseline	0.9025	**0.9661**	0.6502

Table 1: Model performance evaluation on multiple test datasets

Conclusion

This work introduces a framework that uses deep learning approaches to address an important biological problem. This framework suggests that extracting the information given by attention weights to construct a custom layer of the ensemble model enhances model performance to some extent. In future work, we intend to improve the performance of our framework by optimizing the custom layer construction.

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

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Breiling, A.; and Lyko, F. 2015. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & chromatin*, 8: 1–9.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Teufel, F.; Almagro Armenteros, J. J.; Johansen, A. R.; Gíslason, M. H.; Pihl, S. I.; Tsirigos, K. D.; Winther, O.; Brunak, S.; von Heijne, G.; and Nielsen, H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology*, 40(7): 1023–1025.

Zeng, W.; Gautam, A.; and Huson, D. H. 2023. MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction. *GigaScience*, 12: giad054.

Zeng, W.; and Huson, D. 2023. MR-DNA: Flexible 5mC-Methylation-Site Recognition in DNA Sequences using Token Classification. *bioRxiv*, 2023–06.