Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 3, 2025

5mC Detection Based on Deep Learning from Nanopore Sequencing

Jiongjiong Teng[1, *], Wei He[2, a] and Zhihao Zhang[1, b]

[1] School of Control and Computer Engineering, North China Electric Power University; Beijing, 100096, China
[2] School of Computer Science, Key Laboratory of High Confidence Software Technologies, Peking University, Beijing, 100871, China
[*] Corresponding author: 18736442503@163.com, [a]heweibright@gmail.com, [b]zzhao132@outlook.com

Abstract: DNA methylation is a key regulator in diverse biological processes, particularly in mammals where 5-methylcytosine (5mC) methylation is the predominant form. Nanopore sequencing has attracted considerable interest due to its capacity for direct detection of 5mC modifications. Nonetheless, its accuracy in detecting 5mC methylation remains inferior to bisulfite sequencing. In this study, we introduce a novel deep learning algorithm, "single-nano," which aims to improve the precision of 5mC methylation detection. Our method segments the detection task into several subtasks based on motifs, which enhances both the accuracy of predictions and the efficiency of computational processes. Evaluations on publicly available datasets have shown that single-nano outperforms existing algorithms in terms of effectiveness.

Keywords: Deep learning, nanopore sequencing, 5-methylcytosine (5mC), DNA methylation, motif-based detection.

1. Introduction

In the realm of epigenetics, 5-methylcytosine (5mC) methylation stands out as the most significant methylation modification in mammals. DNA methylation is pivotal in key life processes within organisms, with aberrant patterns linked to a spectrum of human diseases [1-3]. Bisulfite sequencing, considered the gold standard for 5mC detection [4], is not without flaws; its process involving chemical DNA treatment and PCR amplification can introduce extraneous errors. In contrast, long-read sequencing technologies like Oxford Nanopore enable direct sequencing of modified nucleotides, circumventing the issues associated with bisulfite treatment. Tools such as Nanopolish [5] leverage hidden Markov models to identify CpG methylation, while others like Megalodon, DeepSignal [6], DeepMod [7], and Rockfish [8] rely on neural network algorithms for detection. Despite the availability of multiple deep learning algorithms for 5mC detection, their reliance on a singular model to assess all motifs can lead to lower accuracy in specific contexts, hindering the improvement of overall detection precision.

This study introduces a motif-centric strategy for segmenting the 5mC detection model into discrete, smaller models. This approach simplifies the problem by decomposing it into more manageable parts, thereby enhancing detection accuracy and diminishing computational demands in the prediction phase. Historically, tools have been constructed using genomic data, such as that from Escherichia coli. However, repetitive sequences within genomes often lead to alignment errors during dataset assembly, compromising dataset quality. To mitigate these issues, this research employs synthetic DNA datasets, thereby achieving a new level of precision in model accuracy.

2. Artificial Dataset Construction

2.1. DNA Sequence Design

In this research, we developed an artificial dataset consisting of four extended DNA strands. These strands were constructed by piecing together short sequences centered around the motif (XXCGXX), ensuring that each strand contains a specific number of motifs covering all those needed for training.

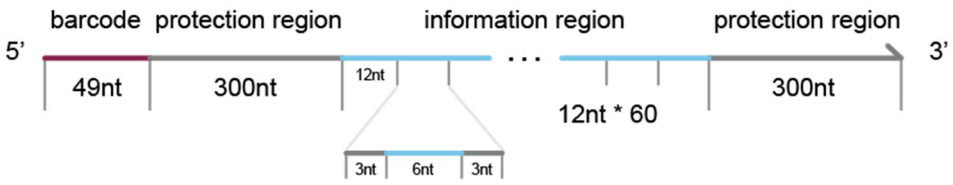

Figure 1. The structure of long DNA strands

As illustrated in the figure1, the structure of these strands is segmented into three distinct sections: a barcode, a protection zone, and an information zone. The barcode is a unique DNA sequence that uniquely identifies each strand, spanning 49 nucleotides. Throughout experimentation, fully methylated and non-methylated DNA strands are differentiated using distinct barcodes. The information zone incorporates every possible motif, separated by 4-6 nucleotide DNA random sequences, including all motifs necessary for database construction, with each motif appearing only once to maintain equilibrium in the training dataset. Considering the elevated error rates at sequence termini in nanopore sequencing, we incorporated protection zones at both ends of the strands to safeguard the central information region, thereby ensuring that the nanopore signals from this zone are adequately reliable.

2.2. Methylation Experiment

We utilized the enzyme M.SssI (NEB, M0226S) for the preparation of fully methylated DNA samples. This enzyme converts all cytosines at CG sites to 5-methylcytosine (5mC), effecting DNA methylation. The long DNA strands intended for methylation were synthesized by Tsingke Bio and were embedded in plasmids. The required DNA strands were first amplified via PCR. While the positive samples underwent M.SssI treatment for complete methylation, the negative samples remained untreated. These samples were then combined and analyzed using nanopore sequencing with the R10.4.1 kit, with distinct barcodes used to differentiate between negative and positive samples (Figure 2). This approach ensures a clear distinction and analysis of methylated and non-methylated regions within the DNA samples, providing valuable data for epigenetic studies.

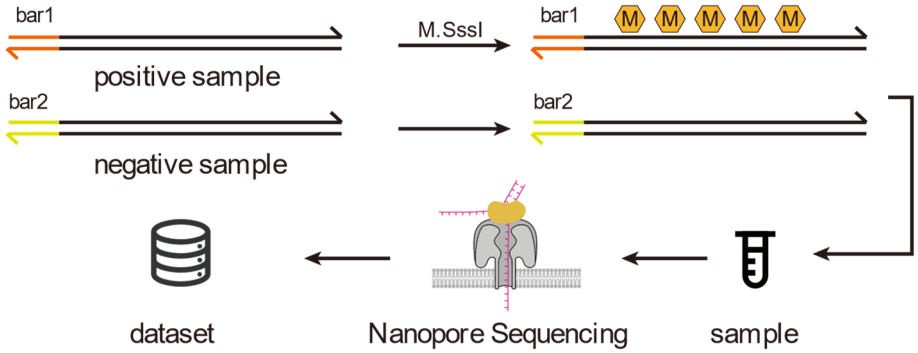

Figure 2. Sample preparation

2.3. Feature Extraction

Our artificial dataset includes a comprehensive set of features: base sequences, raw current signals, along with their respective lengths, averages, medians, standard deviations, base qualities, and basecalling results characterized by insertions, deletions, and substitutions(table 1).Upon acquiring the nanopore detection signals, we utilized Dorado for conducting basecalling and mapping of the nanopore sequencing current signals. This approach facilitated the establishment of a mapping relationship between the current signals and the corresponding bases. This refined version enhances clarity by specifying the types of features included in the dataset and clearly describing the process of using Dorado for basecalling and mapping. The academic style is maintained throughout, with a focus on precision and readability.

Table 1. Three Scheme comparing

Name	Size
Kmer	k
Signal	k*100
Length	k*1
Mean	k*1
Median	k*1
Std	k*1
Quality	k*1
Mismatch	k*1
Insertion	k*1
Detection	k*1

Each entry in our dataset corresponds to a CpG site, with a span of 10 bases on either side of the central cytosine, encompassing a total of k=21 bases. For each base within these 21 bases, there is an associated set of raw current signals denoted as $I_1^{raw} \sim I_m^{raw}$. To enrich our signal features, we standardize the length of the current signal L_{sig} to 100. If the length of the current signal for a base does not equal 100, we apply linear interpolation to adjust the signal. The original time axis for the current signal is represented as $t^{raw}=\{0,1,\ldots,m-1\}$, consisting of m points. For the new time axis t^{new}, each point t_j^{new} (with $j \in \{0,1,\ldots, L_{sig}\}$) is calculated using Equation 2-1.

$$t_j^{new} = j \cdot \frac{m-1}{L_{sig}-1} \tag{2-1}$$

For each index j, once t_j^{new} is computed, we extract the integer part $k = \lfloor t \rfloor$ and the decimal part $\alpha = t - k$. If $k \geq m-1$, we use the value of the last data point I_m^{raw}. In other cases, the interpolated value is derived from a linear combination of the two nearest points(Equation 2-2):

$$I_j^{new} = I_k^{raw} \cdot (1-\alpha) + I_{k+1}^{raw} \cdot \alpha \tag{2-2}$$

For each base, the current signal I_j^{new} ($j \in \{0,1,\ldots, L_{sig}\}$).The mean, median, and standard deviation of these signals are calculated based on the original raw signals $I_1^{raw} \sim I_m^{raw}$. Base quality assessments and basecalling error metrics — insertions, deletions, and substitutions — are extracted from alignments in the bam file using samtools (v1.19.2). Errors are coded as binary values: 0 for no error and 1 for the presence of an error.

2.4. Dataset Construction

Following feature extraction, the dataset is segmented into sub-datasets according to various motifs. To enhance the precision of the training process, superfluous data is eliminated, ensuring an equal distribution of negative and positive samples, each comprising 50% within the sub-datasets. Each sub-dataset is then further divided, allocating 90% for training purposes and the remaining 10% for testing.

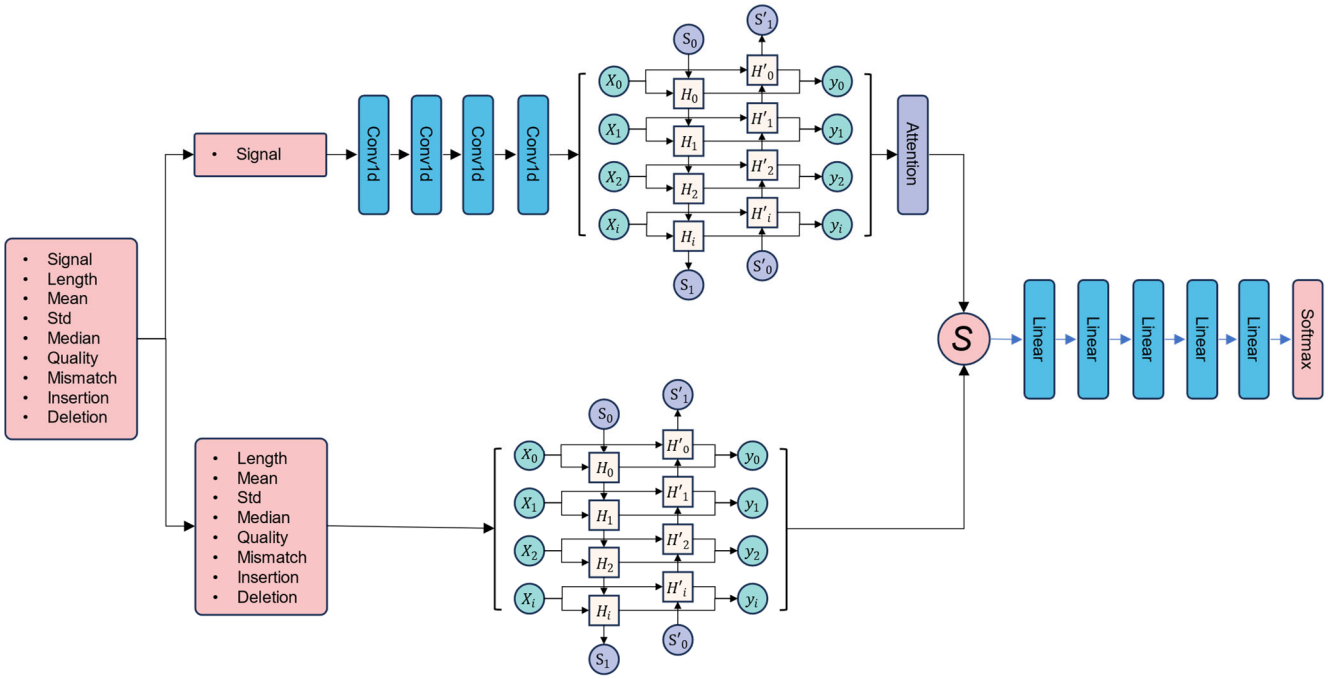

Figure 3. Model Framework

3. Model Framework

After categorizing the training set by motifs, sequence features are rendered ineffective within this context. The remaining nine features are classified into two groups based on their informational density: sparse features, such as current signals; and dense features, including the original current signal's length, average value, median, standard deviation, base quality, and error metrics from basecalling (insertions, deletions, and substitutions).

All input features being sequential in nature prompts the use of a Bidirectional Long Short-Term Memory (BiLSTM) network for feature extraction. Given the voluminous nature of current signal data, a three-layer 1D Convolutional Neural Network (1DCNN) is first applied for local feature extraction to alleviate computational demands, succeeded by BiLSTM for extracting sequential features. To account for methylation's position-dependent impact on current signals, an Attention mechanism is employed to assign weights to these features. Sparse features proceed directly into the BiLSTM.The two types of features are then integrated, followed by five fully connected layers for classification. The final step involves employing a Softmax function to transform predictions into probabilities indicative of methylation.

4. Single Motif Model

Previous research typically employed a single model to predict methylation across all motifs. However, since methylation affects the current signals differently for each motif, this approach often resulted in models favoring certain motifs in terms of accuracy [9]. A case in point is Megalodon's performance on the human genome (NA12878) [10], where it was benchmarked against bisulfite sequencing data (ENCODE, ENCFF798RSS and ENCFF113KRQ). Notably, Megalodon's prediction accuracy was significantly lower for the TTCGTC motif compared to other motifs.

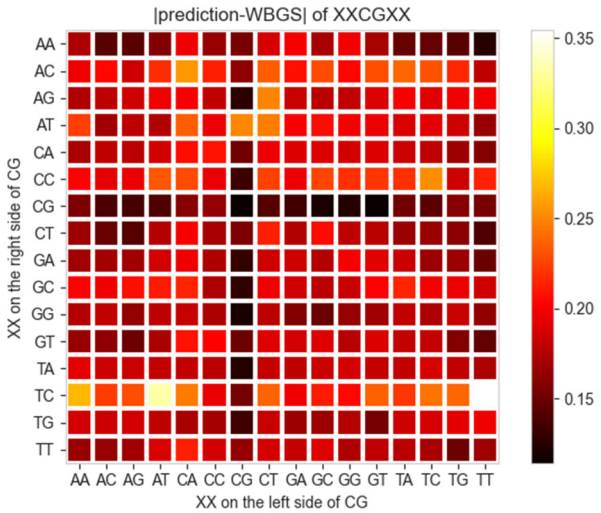

Figure 4. The accuracy of Megalodon on different motifs

In the single-nano framework, a distinct mini-model is designated for each motif. To minimize training duration, a

base model is first trained on an extensive dataset. Thereafter, each mini-model leverages transfer learning with its specific training set to achieve more accurate predictions.

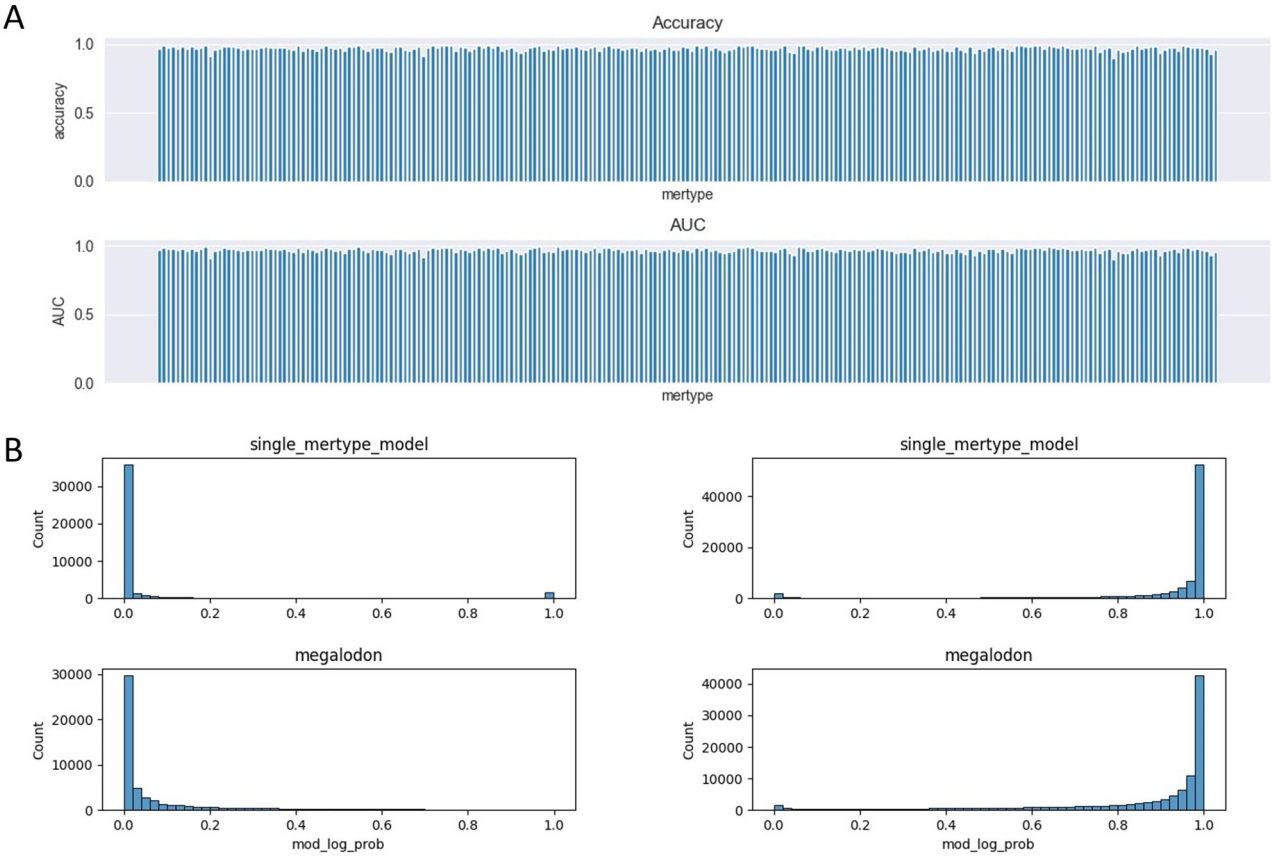

Figure 5. Model performance evaluation

5. Model Validation

5.1. Model Training

This section initiates with the training of a foundational model utilizing the most frequent kmer dataset post-data balancing. The model parameters are fine-tuned by minimizing cross-entropy on the training dataset. The formula for the cross-entropy loss function is presented below(Equation 5-1):

$$L(\theta) = -\frac{1}{N}\sum_{n=1}^{N}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \quad (5\text{-}1)$$

In this chapter, we utilize the Adam optimizer with a learning rate of 0.001 to train the model parameters. The optimization targets minimizing the cross-entropy loss function, where y_i and $\hat{y}_i$ correspond to the actual and predicted labels for the i-th sample, respectively. The batch size of samples is denoted by N, and θ represents the trainable parameters during training. Furthermore, we implement an early stopping mechanism to prevent overfitting; training is terminated if there is no reduction in the validation loss for 10 successive epochs.

The mini-models adopt a training strategy in the transfer learning phase that closely mirrors the base model's approach, with the learning rate adjusted to 0.0001.

5.2. Model Evaluation

We determined the accuracy (ACC) and the area under the ROC curve (AUC) using actual and predicted labels. Figure X illustrates that 93% of the mini-models obtained an AUC above 0.95, with even the least accurate mini-model attaining a precision rate of 0.901(Figure 5a).

We conducted validation of the mini-models using a publicly available dataset [11]. The predictive outcomes for both negative and positive sites are depicted in Figure 5b.

6. Conclusion

In this study, we present a novel single-motif model termed "single-nano," which integrates 1D Convolutional Neural Networks (1DCNN) and Bidirectional Long Short-Term Memory (BiLSTM) architectures for the detection of CpG methylation in DNA sequences. This model innovatively segments the traditional 5mC methylation detection task based on motifs, thereby overcoming the limitations of conventional models that struggle with accurate identification on specific motifs. As a result, single-nano enhances the accuracy of methylation detection.

The natural world is replete with various DNA and RNA modifications, many of which present significant challenges in detection. Single-nano stands out as a versatile model capable of training on individual motifs, even those that are rare. Moreover, it offers the potential for transfer learning across different types of modifications.

In conclusion, single-nano is well-suited for detecting a variety of modifications on nanopores. Its motif-based segmentation approach and adaptability to diverse modifications render it a valuable asset for methylation detection and potentially broader applications in epigenetic research.

References

[1] Woods D., Doty D., Myhrvold C., et al. Diverse and robust molecular algorithms using reprogrammable DNA self-assembly[J]. Nature 567, 2019, 366-372

[2] Zhang D., Hariadi R., Choi H., et al. Integrating DNA strand-displacement circuitry with DNA tile self-assembly[J]. Nature Communications 4, 2013, 1965

[3] Song T., Eshra A., Shah S., et al. Fast and compact DNA logic circuits based on single-stranded gates using strand-displacing polymerase[J]. Nature Nanotechnology 14, 2019, 1075-1081

[4] Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl Acad. Sci. USA 89, 1827–1831 (1992).

[5] Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods 14, 407–410 (2017).

[6] Ni, P. et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics 35, 4586–4595 (2019).

[7] Liu, Q. et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat. Commun. 10, 2449 (2019).

[8] Stanojević, D., Li, Z., Bakić, S. et al. Rockfish: A transformer-based model for accurate 5-methylcytosine prediction from nanopore sequencing. Nat Commun 15, 5580 (2024).

[9] Yuen, Z.WS., Srivastava, A., Daniel, R. et al. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat Commun 12, 3438 (2021).

[10] Jain, M., Koren, S., Miga, K. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338–345 (2018).

[11] Zhang, C., Wu, R., Sun, F. et al. Parallel molecular data storage by printing epigenetic bits on DNA. Nature 634, 824–832 (2024).