Molecular Therapy
Nucleic Acids
Original Article

4acCPred: Weakly supervised prediction of N^4-acetyldeoxycytosine DNA modification from sequences

Jingxian Zhou,[1,4,7] Xuan Wang,[2,7] Zhen Wei,[2,5] Jia Meng,[2,3,6] and Daiyun Huang[2,4]

[1]Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China; [2]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China; [3]AI University Research Centre, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China; [4]Department of Computer Science, University of Liverpool, Liverpool L69 7ZB, UK; [5]Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool L69 7ZB, UK; [6]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

DNA methylation is one of the earliest epigenetic regulation mechanisms studied extensively, and it is critical for normal development, diseases, and gene expression. As a recently identified chemical modification of DNA, N4-acetyldeoxycytosine (4acC) was shown to be abundant in *Arabidopsis* and highly associated with gene expression and actively transcribed genes. Precise identification of 4acC is essential for studying its biological function. We proposed the 4acCPred, the first computational framework for predicting 4acC-carrying regions from *Arabidopsis* genomic DNA sequences. Since the existing 4acC data are not precise for a specific base but only report regions that are hundreds of bases long, we formulated the task as a weakly supervised learning problem and built 4acCPred using a multi-instance-based deep neural network. Both cross-validation and independent testing on the four datasets under different conditions show promising performance, with mean areas under the receiver operating characteristic curve (AUCs) of 0.9877 and 0.9899, respectively. 4acCPred also provides motif mining through model interpretation. The motifs found by 4acCPred are consistent with existing knowledge, indicating that the model successfully captured real biological signals. In addition, a user-friendly web server was built to facilitate 4acC prediction, motif visualization, and data access. Our framework and web server should serve as useful tools for 4acC research.

INTRODUCTION

DNA methylation is one of the major epigenetic mechanisms that critically influence a number of vital biological processes.[1] Aberrant DNA methylation patterns are frequently observed in diseases.[2] Currently, more than 17 types of modified bases have been identified in DNA.[3] Among them, N^6-methyladenine (6mA) and 5-methylcytosine (5mC) are the two most prevalent modifications, and their biological roles have been widely studied. 6mA usually plays critical roles in the regulation of gene expression,[4] DNA repairs,[5] and DNA replication[6] and is closely associated with cancer development[7] and neuro development,[7,8] and 5mC is responsible for the silencing of transposable elements,[9] atherosclerosis,[10] and aging.[11] Apart from 5mC to 6mA, other DNA methylations have not been extensively detected and explored. Recently, inspired by direct analogs of chemical modifications of RNA and DNA, such as m5C/5mC, hm5C/5hmC, and m6A/6mA, as well as the widely distributed and highly conserved RNA modification N^4-acetylcytosine (ac4C),[12,13] Wang et al. proposed 4acC immunoprecipitation followed by sequencing (4acC-IP-seq) to explore the presence and function of 4acC in DNA of *Arabidopsis thaliana*.[14] The protocol of 4acC-IP-seq is similar to methylated DNA IP (MeDIP) and 6mA-IP-seq,[15,16] applying the 4acC-specific antibody IP method to enrich DNA fragments containing 4acC modification and then constructing a high-throughput sequencing library. 4acC-IP-seq revealed that the enriched 4acC modification peaks were mostly distributed around the transcription start sites of protein-coding genes among euchromatin regions. Potential interactions of 4acC with 5mC and histones were also observed in the context of gene-expression regulation. Furthermore, the existence of 4acC in genomic DNA samples of rice, maize, mouse, and *Homo sapiens* was also confirmed by mass spectrometry. Therefore, precise identification of 4acC in eukaryotic DNA is crucial for exploring its biological function and its interplays with other epigenetic marks.

In practice, wet-lab experiments to detect modifiable sites, such as mass spectrometry and antibody-based sequencing method, are often time consuming with a high cost, and the specific antibodies used for IP sequencing restrict the accuracy of sequencing results.[17] To date, various computational models based on DNA or RNA sequences have been proposed to serve as useful alternatives. The existing algorithms can be roughly divided into two categories: feature-based algorithms and deep-learning-based algorithms. Examples of the former include iDNA6mA-Rice,[18] i6mA-Fuse,[19] SDM6A,[20] Methylator,[21] MethCGI,[22] iDNA-Methyl,[23] 4mCPred,[24] 4mCPred-SVM,[25] 4mCpred-EL,[26] 4mCpred-IFL,[27]

Received 9 August 2022; accepted 12 October 2022;
https://doi.org/10.1016/j.omtn.2022.10.004.

[7]These authors contributed equally

Corresponding author
E-mail: daiyun.huang@liverpool.ac.uk

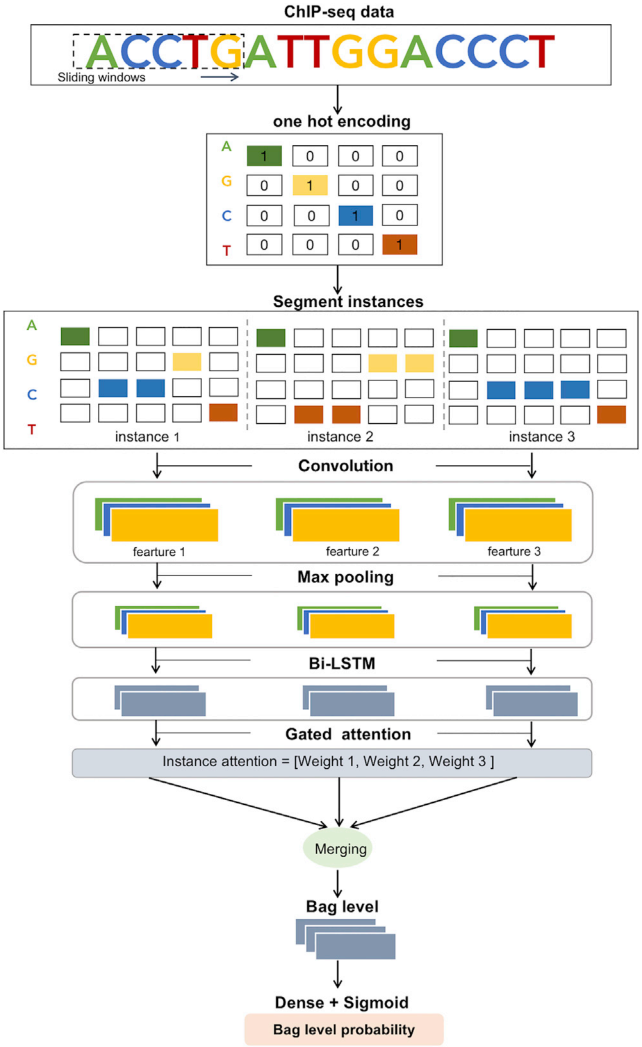


Figure 1. A simplified graphic illustration of the proposed 4acCPred framework

Meta-4mCpred,[28] and i4mC-ROSE, which rely on hand-crafted features to represent sequence context.[29] For instance, iDNA6mA-Rice applied multiple encoding schemes, including PseKNC, single-nucleotide binary encoding, and natural vectors, and predicted 6mA using random forest on rice DNA. While the performance of feature-based methods relies heavily on choosing the best representation in each case, many deep-learning-based methods have recently been proposed to learn from raw sequences and achieve superior performance, including iDNA6mA-Rice-DL,[30] DNA6mA-MINT,[31] 4mCPred-CNN,[32] Deep6mAPred,[33] BiLSTM-5mC,[34] and so on. For instance, 4mCPred-CNN is the first method based on a convolutional neural network (CNN) to identify 4mC sites in the mouse genome.[32] Similar machine-learning-based approaches have also been proposed to predict modifications in RNA sequences, which can be easily transferred to DNA. Examples include iRNA-Methyl,[35] SRAMP,[36] WHISTLE,[37] DeepPromise,[38] MultiRM,[39] NmRF,[40] BRPCA,[41] EDLm6APred,[42] REW-ISA V2,[43] m7GDisAI,[44] HN-CNN,[45] and m6Acomet.[46]

Most existing frameworks are based on strong supervision, which requires the precise location of the modified bases. However, such data are currently unavailable for 4acC. The only high-throughput-sequencing-based technology, 4acC-IP-seq, only allows the detection of 4acC-carrying DNA fragments of at least 200–400 bases in length.[14] Because there usually exist multiple cytosines in each fragment, it is unclear which ones are modified. Such data do not allow the model to learn modification-specific sequence contexts from fixed-length sequences centered on the target cytosine as strongly supervised learning does. Instead, the only label information that can be used for training is associated with DNA regions of different lengths, i.e., whether the region contains at least one 4acC site. To address the challenge of learning from these coarse-grained labels, here we consider a multiple-instance learning (MIL) framework, one of the weakly supervised learning algorithms.

Weakly supervised learning aims to construct predictive models by learning from noisy, limited, or imprecise sources. In genomics, weakly supervised learning, especially MIL, is widely considered in protein-DNA interaction prediction,[47–50] where the bound DNA sequence may contain multiple binding sites, and the exact location is unknown. WSCNN and its updated version, WSCNNLSTM, combine MIL with deep neural networks and have achieved superior results in *in vivo* and *in vitro* transcription factor binding site prediction.[49,50] In addition to DNA modeling, MIL has been applied to automate the annotation of protein functions,[51] protein splice variants,[52] specific functional binding sites in microRNA targets,[53] and proteome-wide interactions.[54] More recently, to predict RNA modifications from only low-resolution epitranscriptome data, WeakRM combined MIL and attention mechanisms and showed promising performance on three RNA modifications, including ac4C.[55] WeakRM divided the RNA sequence into multiple fixed-length subsequences. For peak regions called by bioinformatics tools, at least one subsequence contains a target modification and should cover specific sequence patterns. Thus, the integrated representation of all subsequences can be linked to a positive label. Whereas for regions from the same transcript but not detected as peaks, modification-specific sequence patterns should not be included in any subsequences, and their integration remains negative. Inspired by the WeakRM framework and considering the direct analog of ac4C and 4acC, we consider the 4acC prediction from 4acC-IP-seq data naturally and inherently an MIL task.

We propose 4acCPred, the first prediction framework for high-accuracy identification of 4acC-carrying regions from *Arabidopsis* genomic DNA sequences. Under the MIL framework, we combine CNN and bidirectional long short-term memory (LSTM) to exploit their advantages in local motif extraction and long-term interaction learning, respectively. A simplified graphic framework of 4acCPred is illustrated in Figure 1. Evaluation of all four conditions provided by 4acC-IP-seq (wild type, NH_2OH treatment, *met1* mutant, and

Table 1. Performance of 4acCPred under 10-fold cross-validation with standard deviations

Model	Group	Accuracy	AUC	AP	MCC
WSCNN	WT[a]	0.7089 (±0.0222)	0.8052 (±0.0438)	0.8041 (±0.0220)	0.4474 (±0.0270)
	NH_2OH	0.6860 (±0.0460)	0.8306 (±0.0595)	0.8271 (±0.0344)	0.4265 (±0.0616)
	met1	0.6788 (±0.0438)	0.8705 (±0.0704)	0.8695 (±0.0454)	0.4337 (±0.0625)
	Rdd	0.6668 (±0.0591)	0.8558 (±0.0738)	0.8584 (±0.0407)	0.4121 (±0.0891)
WSCNNLSTM	WT	0.8064 (±0.0099)	0.8801 (±0.0082)	0.8453 (±0.0123)	0.6175 (±0.0195)
	NH_2OH	0.8124 (±0.0077)	0.8881 (±0.0048)	0.8578 (±0.0068)	0.6263 (±0.0158)
	met1	0.8511 (±0.0088)	0.9226 (±0.0077)	0.9015 (±0.0115)	0.7027 (±0.0176)
	Rdd	0.8433 (±0.0079)	0.9216 (±0.0068)	0.8985 (±0.0122)	0.6884 (±0.0152)
4acCPred	WT	0.9485 (±0.0050)	0.9794 (±0.0018)	0.9695 (±0.0033)	0.8981 (±0.0098)
	NH_2OH	0.9506 (±0.0050)	0.9857 (±0.0021)	0.9800 (±0.0055)	0.9019 (±0.0098)
	met1	0.9679 (±0.0061)	0.9928 (±0.0024)	0.9875 (±0.0064)	0.9360 (±0.0122)
	Rdd	0.9706 (±0.0037)	0.9928 (±0.0030)	0.9890 (±0.0055)	0.9414 (±0.0071)

All methods were evaluated using the same datasets with a positive-to-negative ratio of 1:1. The performance is given as average ± standard deviation. The threshold for accuracy is 0.5.

[a] WT represents the condition of wild type.

ros1dml2dml3 [*rdd*] mutant) demonstrated the general effectiveness of our approach in predicting 4acC from DNA regions (average areas under the receiver operating characteristic curve [AUCs] in a 10-fold cross-validation test were 0.9877 and 0.9899 in independent testing). Furthermore, model interpretation showed that 4acCPred also captured motifs consistent with existing knowledge. The 4acCPred web server, accessible via http://www.rnamd.org/4accpred, is designed to help users predict 4acC modifications and visualize captured motif patterns on *Arabidopsis* genomic DNA. All data used in this study (the peak information called by MACS2 with annotation) and trained 4acCPred models have also been uploaded to the web server for user convenience. We anticipate that our newly proposed model and the web server can take full advantage of limited experimental data and facilitate the study of DNA 4acC modification by providing alternative computational prediction approaches.

RESULTS

We developed the first DNA 4acC modification predictor based on MIL. To demonstrate the model's stability, we formed four benchmark datasets based on four conditions provided in 4acC-IP-seq (GEO: GSE168538) and evaluated the model performance separately.[14] The initial learning rate, decayed learning rate, instance length, and instance stride are set as 5×10^{-4}, 1×10^{-5}, 40, and 5, respectively. Details can be found in the materials and methods. For each dataset, one-third of the data was selected as an independent test dataset. The proposed method was found to be robust on both cross-validation and independent testing.

The whole structure of the model contains a four-layer encoding module (see Figure 1): after the first convolutional layer, a max-pooling layer withdraws weak features in datasets to expand the receptive field. There also involves a dropout layer to prevent overfitting in training the model. Finally, a bidirectional LSTM layer captures the hidden long-term dependencies between sequential patterns. Each instance passes through the same encoding module (weights are shared) and outputs instance-level features. The network learns weights for each instance and sums all instance features as features for the entire input sequence for final classification.

Comparison with existing frameworks developed for protein-DNA binding prediction

We compared our algorithms 4acCPred with WSCNN and its updated version WSCNNLSTM using the same datasets,[49,50] which were originally developed for transcription factor binding site prediction. We split each dataset equally into ten partitions to perform 10-fold cross-validation. Table 1 shows the average performances in terms of the AUC, average precision (AP), Matthews correction coefficient (MCC), and accuracy of different algorithms under cross-validation. In the wild-type case, the AUC of 4acCPred is about 0.1 higher than that of WSCNN and WSCNNLSTM. 4acCPred also achieved an improvement of at least 0.07 AUC under the other three conditions. Under all four conditions, 4acCPred had an AUC of at least 0.97 and an AP of at least 0.96, indicating its promising performance in predicting 4acCPred-carrying regions.

Performance evaluation on independent test datasets

To further test the performance and robustness of our newly proposed predictor in finding 4acC-carrying regions, we apply the ten models obtained from the cross-validation for each condition to held-out independent test datasets and show their average results. As shown in Table 2, with accuracy, AUC, MCC, and AP as evaluation metrics, the results of our model are consistent with those in cross-validation. A significant improvement from the baseline model to 4acCPred can also be observed. The average performances in terms of AUC, AP, MCC, and accuracy of different algorithms under the ensemble of these ten models are shown in Table S1. It is pleasing that all metrics of 4acCPred are higher than WSCNN and WSCNNLSTM.

Table 2. Performance of 4acCPred on independent test datasets with standard deviations

Model	Group	Accuracy	AUC	AP	MCC
WSCNN	WT[a]	0.7101 (±0.0216)	0.8051 (±0.0483)	0.8044 (±0.0260)	0.4509 (±0.0251)
	NH_2OH	0.6831 (±0.0452)	0.8293 (±0.0652)	0.8250 (±0.0372)	0.4209 (±0.0601)
	met1	0.6766 (±0.0402)	0.8691 (±0.0740)	0.8694 (±0.0424)	0.4301 (±0.0524)
	Rdd	0.6728 (±0.0590)	0.8643 (±0.0744)	0.8647 (±0.0426)	0.4213 (±0.0898)
WSCNNLSTM	WT	0.8119 (±0.0032)	0.8866 (±0.0014)	0.8551 (±0.0020)	0.6288 (±0.0054)
	NH_2OH	0.8154 (±0.0033)	0.8893 (±0.0018)	0.8607 (±0.0035)	0.6324 (±0.0068)
	met1	0.8526 (±0.0036)	0.9241 (±0.0020)	0.9018 (±0.0032)	0.7061 (±0.0074)
	Rdd	0.8594 (±0.0053)	0.9284 (±0.0011)	0.9056 (±0.0020)	0.7200 (±0.0107)
4acCPred	WT	0.9512 (±0.0015)	0.9855 (±0.0006)	0.9796 (±0.0007)	0.9031 (±0.0031)
	NH_2OH	0.9530 (±0.0017)	0.9859 (±0.0005)	0.9779 (±0.0009)	0.9066 (±0.0033)
	met1	0.9728 (±0.0033)	0.9950 (±0.0004)	0.9926 (±0.0006)	0.9458 (±0.0065)
	Rdd	0.9726 (±0.0016)	0.9932 (±0.0002)	0.9891 (±0.0006)	0.9453 (±0.0031)

All methods were evaluated using the same datasets with a positive-to-negative ratio of 1:1. The performance is given as average ± standard deviation. The threshold for accuracy is 0.5.
[a]WT represents the condition of wild type.

Since an essential question in deep learning is to what extent the system learns peculiarities of a particular experimental setup (rather than the underlying biology), we designed an experiment for performance comparisons where training and evaluation data come from a different experiment to deal with this. 4acCPred also makes the best prediction (see Table S2).

Motifs identified by 4acCPred are consistent with existing knowledge

In 4acCPred, we quantified attribution scores for each input feature using the integrated gradient (IG) method.[56] Then, we extracted consensus motifs from instances by using TF-MoDISco.[57] TF-MoDISco first identifies subsequences (called seqlets) with significant importance scores, then clusters and aligns all seqlets to obtain representative motifs. After pruning through the overall letter frequency, we selected the consensus motif with the highest number of supporting seqlets. The top 1 result from wild-type models is shown in Figure 2. Due to the lack of strand information in the training data, it is natural to obtain two motifs separately from the two strands, and they should complement each other. According to Figure 2, our proposed model tends to assign high weights to CT-enriched or GA-enriched (complement) subsequences. Since 4acC is a kind of DNA modification, and we input sequence on both '+ ' and '− ' strands to train the model, 4acCPred provides two corresponding motifs on two strands, respectively.

Compared with the known motifs identified by HOMER with the same BED results from MACS2,[58,59] unsurprisingly, the motif learned by 4acCPred with the strictest score (a total of 229 seqlets were found) is consistent with the top 1 motif found by HOMER. To numerically measure the similarity between these two motifs, we applied the motif comparison tool MEME-Tomtom,[60] resulting in a significant p value of 0.0019. More results from the other three conditions can be found in Figures S1–S3, and the top 5 motifs found by 4acCPred for each condition are shown in Figures S4–S7. Together, we show that the motifs revealed by our model are consistent with existing knowledge, providing some evidence that our model learns true biological signals rather than technical biases and that attention weights have the potential to be used to predict acetylation at higher resolutions.

Web implementation

To facilitate the use of our model and assess of the used data, a web server has been developed using Hyper Text Markup Language (HTML), Cascading Style Sheets (CSSs), and JavaScript (JS) and is accessible at http://www.rnamd.org/4accpred (see Figure 3). The web server allows users to upload DNA sequences in FASTA format and provides predicted probability scores under user-specified conditions (wild type [WT], NH_2OH, *rdd*, and *met1* mutant). All the results could be downloaded in a CSV format file. The display of motifs is available upon request. In addition, all processed data originating from 4acC-IP-seq and analyzed using MACS2 can be freely downloaded from the web server.

DISCUSSION

4acC is a recently discovered abundant DNA chemical modification that is involved in various gene-expression functions. Precise prediction of 4acC modification-containing regions is vital for scientific investigations to understand its role in biological regulation. In this study, we designed the first 4acC predictor, 4acCPred, on *Arabidopsis* based on weakly supervised learning and LSTM. We collected data from all four conditions provided by 4acC-IP-seq and divided them into cross-validation and independent test datasets. 4acCPred achieved average AUCs of 0.9877 and 0.9899 in the two datasets, respectively. Additionally, the motif discovered by 4acCPred with the strictest score is consistent with motifs found in existing knowledge. Together, these results demonstrate the robustness of our model as a useful alternative to detect DNA 4acC acetylation. To facilitate

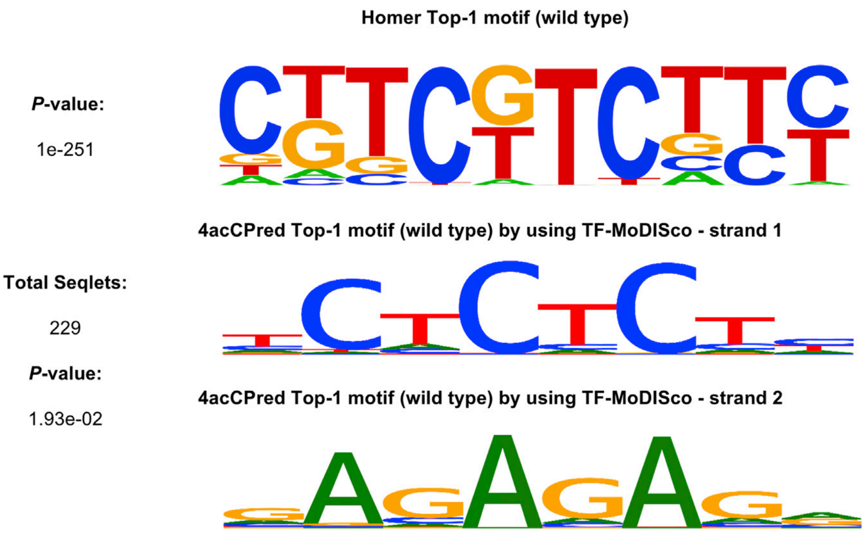

Homer Top-1 motif (wild type)
P-value:
1e-251
4acCPred Top-1 motif (wild type) by using TF-MoDISco - strand 1
Total Seqlets:
229
P-value:
1.93e-02
4acCPred Top-1 motif (wild type) by using TF-MoDISco - strand 2

Figure 2. Top 1 motif found in 4acCPred

The top motif found in 4acCPred matches the result of HOMER with a p value of 1.93×10^{-2} under the wild-type condition. The first p value (up) represents the significance of the HOMER motif. The second p value (down) represents the probability that a random motif of the same width has the same or better matching score as the reported motif.

the use of 4acCPred, we have also built a user-friendly web server for prediction and motif visualization, available at http://www.rnamd.org/4accpred. It is worth noting that 4acCPred is currently only constructed for *Arabidopsis*, and its performance in other species needs further experimental verification.

MATERIALS AND METHODS

Benchmark datasets

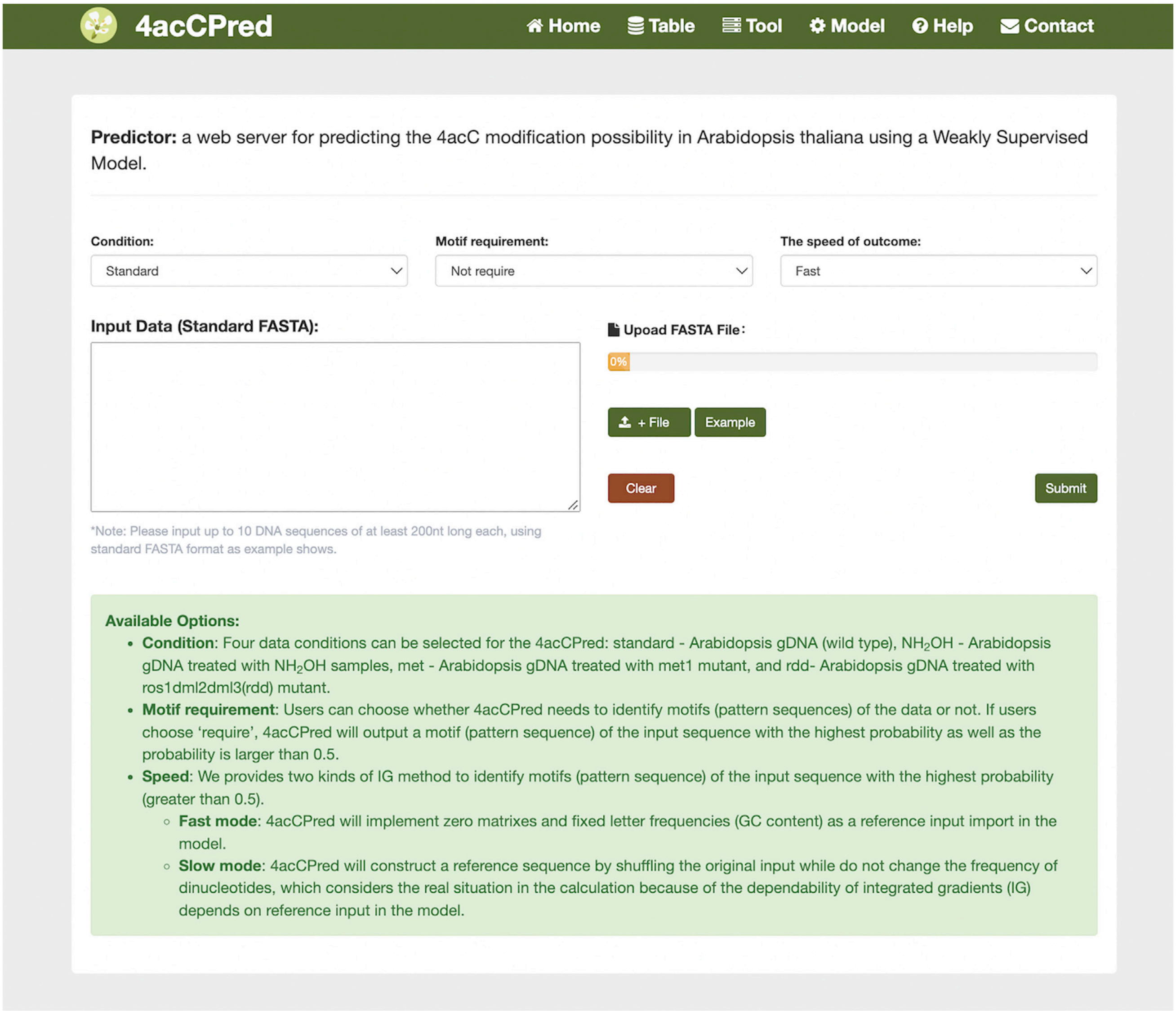
The high-throughput 4acC-IP-seq samples were collected from the recently published article.[14] All data were downloaded from NCBI Gene Expression Omnibus (GEO) under GEO: GSE168538 (see Table 3). Confident 4acC peaks were obtained through peak calling of MACS2 (https://github.com/macs3-project/MACS).[59] The sequencing results of four conditions were used to form four datasets separately: WT, NH_2OH treatment, *met1* mutant, and *rdd* mutant. Since there were two biological replicates for each condition, to improve data quality, we used the intersection of the two replicates as positive data. Considering that a single DNA fragment is 200-400 nt in length, we randomly supplemented regions that were too short after the intersection to 200-nt to ensure a consistent peak width distribution. Such supplemented sequences still contain the peak intersection regions and thus can be treated as positive. On the other hand, peaks longer than 2,000 nt in length were not used in this study to avoid potential false positives caused by peak calling software. Since such data only account for a very small proportion, treating them as outliers does not lose much information. For negative data, we select regions from the upstream and downstream regions of the called peak that do not have any overlap with the positive peak. The positive-to-negative ratio is set as 1:1 to avoid imbalance problems in training. Furthermore, for each positive peak, we extracted a negative sequence of the same length and left them in the same part of the dataset so that the model is not biased by input length. Conceivably, 4acC is one kind of DNA modification, so we reverse complemented the sequence and divided the results into instances as additional inputs. It is to be noted that we randomly retained two-thirds of the data as training sets and the rest of the data as independent valid sets to evaluate our model. The DNA sequence is represented using one-hot encodings (i.e., A: [1, 0, 0, 0], C: [0, 1, 0, 0], G: [0, 0, 1, 0], and T: [0, 0, 0, 1].).

Weakly supervised learning of 4acC

The 4acCPred framework treats each DNA sequence as a "bag" with multiple subsequences known as "instances." Specifically, a sliding window of length *c* moves along the bag sequence with stride *s*. This means that each time a subsequence of length *c* is extracted as an instance, the next instance starts *s*-bases downstream of where the previous instance started. If the length of a specific bag sequence is *L*, there will be $\frac{L-c}{s}+1$ instances in total. Window length *c* and stride *s* are two hyper-parameters. In our study, we chose a sliding window of length 50 and a stride of 10 for ac4C RNA modification in 4acCPred. The available label information is associated with the entire bag, not with each instance. The underlying logic of the MIL framework used in 4acCPred is that the network should highlight instances in positive data that contain target acetylation and capture their sequential patterns. Conversely, for negative data, the model should treat all instances as negative.

The network structure used to extract sequence features consists of one convolutional layer and one bidirectional LSTM layer. One of the greatest strengths of CNNs in genomics is that it naturally captures sequence motifs for a given target through its local receptive fields. However, it inevitably overlooks hidden long-term dependencies between sequential patterns, which can be addressed by using LSTMs. Between CNN and LSTM, we add a max-pooling layer to filter weak features and expand the receptive field and a dropout layer to prevent overfitting in model training. It is worth noting that the network uses shared weights to extract features for each instance.

A key step of the MIL framework is to merge the instance-level features to obtain the bag-level probability (i.e., the predicted value that a DNA sequence contains at least one 4acC acetylation). Unlike WSCNN and WSCNNLSTM,[49,50] which first let the network output a score for each instance and then use functions such as mean, max, and Noisy-and to aggregate scores into one value for the entire bag,[61,62] we assign weights to each instance using an attention mechanism and treat the weighted sum of all instance features as the final bag representation. Specifically, we use gated attention,



Predictor: a web server for predicting the 4acC modification possibility in Arabidopsis thaliana using a Weakly Supervised Model.

Condition: Standard

Motif requirement: Not require

The speed of outcome: Fast

Input Data (Standard FASTA):

Upoad FASTA File:

*Note: Please input up to 10 DNA sequences of at least 200nt long each, using standard FASTA format as example shows.

Available Options:

- **Condition**: Four data conditions can be selected for the 4acCPred: standard - Arabidopsis gDNA (wild type), NH_2OH - Arabidopsis gDNA treated with NH_2OH samples, met - Arabidopsis gDNA treated with met1 mutant, and rdd- Arabidopsis gDNA treated with ros1dml2dml3(rdd) mutant.
- **Motif requirement**: Users can choose whether 4acCPred needs to identify motifs (pattern sequences) of the data or not. If users choose 'require', 4acCPred will output a motif (pattern sequence) of the input sequence with the highest probability as well as the probability is larger than 0.5.
- **Speed**: We provides two kinds of IG method to identify motifs (pattern sequence) of the input sequence with the highest probability (greater than 0.5).
  - **Fast mode**: 4acCPred will implement zero matrixes and fixed letter frequencies (GC content) as a reference input import in the model.
  - **Slow mode**: 4acCPred will construct a reference sequence by shuffling the original input while do not change the frequency of dinucleotides, which considers the real situation in the calculation because of the dependability of integrated gradients (IG) depends on reference input in the model.

Figure 3. A web server of 4acCPred

We designed a user-friendly web server for prediction and motif visualization to facilitate the use of 4acCPred.

which consists of three fully connected layers. The attention weight a_k is calculated as follows:

$$a_k = \frac{exp\left\{w^T\left(tanh\left(Vb_k^T\right)\odot sigm\left(Ub_k^T\right)\right)\right\}}{\sum_{j=1}^{K} exp\left\{w^T\left(tanh\left(Vb_j^T\right)\odot sigm\left(Ub_j^T\right)\right)\right\}}$$

where K represents the number of instances in a bag; b_i is the hidden representation of instance i; w, U, and V are weights of three neural network layers; *tanh* and *sigm* are tanh and sigmoid activation functions; and T means transposition. The weights are normalized to guarantee that all weights sum to 1 so that the network can handle the input of any number of instances.

The advantage of this feature-merging approach is that the network can learn to assign high weights to instances with preferred sequence patterns. The attention weight can also be considered as an indicator to infer the subsequences most likely to contain acetylation. Conversely, score merging using a fixed function tends to lose information. For example, using the maximum value will force the network to make decisions based on only one instance, ignoring sequence context from upstream and downstream. Using the mean function is limited by the fact that acetylation is sparsely distributed across many instances. WSCNN and WSCNNLSTM demonstrated that Noisy-and is better than max and mean functions.[49,50] However, although Noisy-and allows a learnable threshold, it is still constructed based on the average score of the instances, which may suffer from the

Table 3. Positive peaks collected in 4acCPred

Modification	Technology	Condition	Size[a]	Species	Sample
4acC	4acC-IP-seq	Wild type (IP)	19,849	Tair10	GSM5145690[14] GSM5145691[14]
4acC	4acC-IP-seq	NH_2OH (WT IP)	16,682	Tair10	GSM5145692[14] GSM5145693[14]
4acC	4acC-IP-seq	met1 (mutate IP)	10,989	Tair10	GSM5145698[14] GSM5145699[14]
4acC	4acC-IP-seq	rdd (mutate IP)	9,844	Tair10	GSM5369300[14] GSM5369301[14]
4acC	4acC-IP-seq	Wild type (input)	–	Tair10	GSM5145694[14,b]

The ratio of positive and negative labels is 1:1.

[a]Size denotes the total number of sequences.

[b]Wild-type input sample GSM5145694 is used to call peaks by MACS2.

same drawbacks of mean function, i.e., the model is insensitive to positive instances due to a large number of negative instances.

Motif mining

In this study, we use the IG method to perform model interpretation and modification motif mining.[56] The IG method was developed based on backpropagation, a key design of neural networks. It uses the gradient value from the final output to each input multiplied by the input value itself as the contribution of that input (known as attribution score). To solve the problem of gradient saturation, the IG method first selects a reference for each input, performs linear interpolation from the reference to the input, calculates the attribute score of each interpolation point, and then averages all scores to get final values.[63] Its mathematical formula is given as

$$IG_i(x) = (x - x') \times \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x} \times \frac{1}{m}$$

where x is the input to be interpreted, x' is a selected reference of the same shape as x, and m is the number of steps in linear interpolation.

Two kinds of reference selection methods were explored in this study: fixed reference for all inputs such as zero matrices and dinucleotide-shuffled sequence for each specific input. The former is computationally efficient and can provide a cleaner view when interpreting individual sequences using one-hot encodings because, in this case, only one value remains for each base, directly corresponding to the contribution of that base. The dinucleotide-shuffled sequence refers to shuffling the sequence but maintaining the frequency of dinucleotides. Compared with the zero matrices, it takes more time but is more biologically interpretable and thus was used for motif discovery in our study.[64] Both references can be selected in the web server analysis.

After obtaining attribution maps for each input, we used TF-MoDISCO to obtain the consensus motif for DNA 4acC acetylation.[57] It firstly identifies input segments of user-specific length with high contribution scores, then clusters these segments based on continuous Jaccard similarity calculation, and finally aligns the segments in each cluster to form consensus motifs.

Web interface implementation

The web interface has been established by HTML, CSSs, and JS. MySQL database management systems were used to store our data. Datatables, a table plugin, was applied to show the data.

DATA AVAILABILITY STATEMENT

The raw data used in this study are publicly available in the GEO database under GEO: GSE168538. All processed data can be downloaded from the 4acCPred web server at http://www.rnamd.org/4accpred.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2022.10.004.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (32100519 and 31671373) and the XJTLU Key Program Special Fund (KSF-E-51 and KSF-P-02).

AUTHOR CONTRIBUTIONS

D.H., J.M., and Z.W. contributed to the conception of the study; J.Z., D.H., and J.M. contributed significantly to the analysis and manuscript preparation; J.Z. and D.H. performed the experiment; J.Z. and D.H. performed the data analyses and wrote the manuscript; X.W. and J.Z. designed and conducted the web server; D.H., J.M., and Z.W. helped perform the analysis with constructive discussions.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492. https://doi.org/10.1038/nrg3230.

2. Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. Nat. Struct. Mol. Biol. *20*, 274–281. https://doi.org/10.1038/nsmb.2518.

3. Raiber, E.-A., Hardisty, R., van Delft, P., and Balasubramanian, S. (2017). Mapping and elucidating the function of modified bases in DNA. Nat. Rev. Chem *1*, 0069. https://doi.org/10.1038/s41570-017-0069.

4. Ratel, D., Ravanat, J.L., Berger, F., and Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. Bioessays *28*, 309–315. https://doi.org/10.1002/bies.20342.

5. Au, K.G., Welsh, K., and Modrich, P. (1992). Initiation of methyl-directed mismatch repair. J. Biol. Chem. *267*, 12142–12148.

6. Campbell, J.L., and Kleckner, N. (1990). E. coli oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. Cell *62*, 967–979. https://doi.org/10.1016/0092-8674(90)90271-f.

7. Xiao, C.L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., Yu, G., Liu, J., Xie, S.Q., Luo, F., et al. (2018). N(6)-Methyladenine DNA modification in the human genome. Mol. Cell *71*, 306–318.e7. https://doi.org/10.1016/j.molcel.2018.06.015.

8. Yao, B., Li, Y., Wang, Z., Chen, L., Poidevin, M., Zhang, C., Lin, L., Wang, F., Bao, H., Jiao, B., et al. (2018). Active N(6)-methyladenine demethylation by DMAD regulates gene expression by coordinating with Polycomb protein in neurons. Mol. Cell *71*, 848–857.e6. https://doi.org/10.1016/j.molcel.2018.07.005.

9. Luo, C., Hajkova, P., and Ecker, J.R. (2018). Dynamic DNA methylation: in the right place at the right time. Science *361*, 1336–1340. https://doi.org/10.1126/science.aat6806.

10. Lund, G., Andersson, L., Lauria, M., Lindholm, M., Fraga, M.F., Villar-Garea, A., Ballestar, E., Esteller, M., and Zaina, S. (2004). DNA methylation polymorphisms precede any histological sign of atherosclerosis in mice lacking apolipoprotein E. J. Biol. Chem. *279*, 29147–29154. https://doi.org/10.1074/jbc.M403618200.

11. Gonzalo, S. (2010). Epigenetic alterations in aging. J. Appl. Physiol. *109*, 586–597. https://doi.org/10.1152/japplphysiol.00238.2010.

12. Sas-Chen, A., Thomas, J.M., Matzov, D., Taoka, M., Nance, K.D., Nir, R., Bryson, K.M., Shachar, R., Liman, G.L.S., Burkhart, B.W., et al. (2020). Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. Nature *583*, 638–643. https://doi.org/10.1038/s41586-020-2418-2.

13. Arango, D., Sturgill, D., Alhusaini, N., Dillman, A.A., Sweet, T.J., Hanson, G., Hosogane, M., Sinclair, W.R., Nanan, K.K., Mandler, M.D., et al. (2018). Acetylation of cytidine in mRNA promotes translation efficiency. Cell *175*, 1872–1886.e24. https://doi.org/10.1016/j.cell.2018.10.030.

14. Wang, S., Xie, H., Mao, F., Wang, H., Wang, S., Chen, Z., Zhang, Y., Xu, Z., Xing, J., Cui, Z., et al. (2022). N(4)-acetyldeoxycytosine DNA modification marks euchromatin regions in Arabidopsis thaliana. Genome Biol. *23*, 5. https://doi.org/10.1186/s13059-021-02578-7.

15. Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. *37*, 853–862. https://doi.org/10.1038/ng1598.

16. Fu, Y., Luo, G.Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Dore, L.C., et al. (2015). N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. Cell *161*, 879–892. https://doi.org/10.1016/j.cell.2015.04.010.

17. Jiang, J., Song, B., Tang, Y., Chen, K., Wei, Z., and Meng, J. (2020). m5UPred: a web server for the prediction of RNA 5-methyluridine sites from sequences. Mol. Ther. Nucleic Acids *22*, 742–747.

18. Lv, H., Dao, F.-Y., Guan, Z.-X., Zhang, D., Tan, J.-X., Zhang, Y., Chen, W., and Lin, H. (2019). iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. Front. Genet. *10*, 793. https://doi.org/10.3389/fgene.2019.00793.

19. Hasan, M.M., Manavalan, B., Shoombuatong, W., Khatun, M.S., and Kurata, H. (2020). i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. Plant Mol. Biol. *103*, 225–234. https://doi.org/10.1007/s11103-020-00988-y.

20. Basith, S., Manavalan, B., Shin, T.H., and Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Mol. Ther. Nucleic Acids *18*, 131–141. https://doi.org/10.1016/j.omtn.2019.08.011.

21. Bhasin, M., Zhang, H., Reinherz, E.L., and Reche, P.A. (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett. *579*, 4302–4308. https://doi.org/10.1016/j.febslet.2005.07.002.

22. Fang, F., Fan, S., Zhang, X., and Zhang, M.Q. (2006). Predicting methylation status of CpG islands in the human brain. Bioinformatics *22*, 2204–2209. https://doi.org/10.1093/bioinformatics/btl377.

23. Liu, Z., Xiao, X., Qiu, W.R., and Chou, K.C. (2015). iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. *474*, 69–77. https://doi.org/10.1016/j.ab.2014.12.009.

24. He, W., Jia, C., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics *35*, 593–601. https://doi.org/10.1093/bioinformatics/bty668.

25. Wei, L., Luan, S., Nagai, L.A.E., Su, R., and Zou, Q. (2019). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics *35*, 1326–1333. https://doi.org/10.1093/bioinformatics/bty824.

26. Manavalan, B., Basith, S., Shin, T.H., Lee, D.Y., Wei, L., and Lee, G. (2019). 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. Cells *8*, 1332. https://doi.org/10.3390/cells8111332.

27. Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., and Shi, X. (2019). Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics *35*, 4930–4937. https://doi.org/10.1093/bioinformatics/btz408.

28. Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol. Ther. Nucleic Acids *16*, 733–744. https://doi.org/10.1016/j.omtn.2019.04.019.

29. Hasan, M.M., Manavalan, B., Khatun, M.S., and Kurata, H. (2020). i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int. J. Biol. Macromol. *157*, 752–758. https://doi.org/10.1016/j.ijbiomac.2019.12.009.

30. He, S., Kong, L., and Chen, J. (2021). iDNA6mA-Rice-DL: a local web server for identifying DNA N6-methyladenine sites in rice genome by deep learning method. J. Bioinform. Comput. Biol. *19*, 2150019. https://doi.org/10.1142/s0219720021500190.

31. Rehman, M.U., and Chong, K.T. (2020). DNA6mA-MINT: DNA-6mA modification identification neural tool. Genes *11*, E898. https://doi.org/10.3390/genes11080898.

32. Abbas, Z., Tayara, H., and Chong, K.T. (2021). 4mCPred-CNN-Prediction of DNA N4-methylcytosine in the mouse genome using a convolutional neural network. Genes *12*, 296. https://doi.org/10.3390/genes12020296.

33. Tang, X., Zheng, P., Li, X., Wu, H., Wei, D.-Q., Liu, Y., and Huang, G. (2022). Deep6mAPred: a CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. Methods *204*, 142–150. https://doi.org/10.1016/j.ymeth.2022.04.011.

34. Cheng, X., Wang, J., Li, Q., and Liu, T. (2021). BiLSTM-5mC: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide DNA promoters. Molecules *26*, 7414. https://doi.org/10.3390/molecules26247414.

35. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2015). iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. *490*, 26–33. https://doi.org/10.1016/j.ab.2015.08.021.

36. Zhou, Y., Zeng, P., Li, Y.-H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. *44*, e91. https://doi.org/10.1093/nar/gkw104.

37. Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., Su, J., de Magalhães, J.P., Rigden, D.J., and Meng, J. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res. *47*, e41. https://doi.org/10.1093/nar/gkz074.

38. Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A.I., Webb, G.I., Akutsu, T., Baggag, A., Bensmail, H., and Song, J. (2020). Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief. Bioinform. *21*, 1676–1696. https://doi.org/10.1093/bib/bbz112.

39. Song, Z., Huang, D., Song, B., Chen, K., Song, Y., Liu, G., Su, J., Magalhães, J.P.d., Rigden, D.J., and Meng, J. (2021). Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat. Commun. *12*, 4011. https://doi.org/10.1038/s41467-021-24313-3.

40. Ao, C., Zou, Q., and Yu, L. (2022). NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. Brief. Bioinform. *23*, bbab480. https://doi.org/10.1093/bib/bbab480.

41. Ma, J., Zhang, L., Li, S., and Liu, H. (2021). BRPCA: bounded robust principal component analysis to incorporate similarity network for N7-methyguanosine(m7G) site-disease association prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. In preparation. https://doi.org/10.1109/TCBB.2021.3109055.

42. Zhang, L., Li, G., Li, X., Wang, H., Chen, S., and Liu, H. (2021). EDLm6APred: ensemble deep learning approach for mRNA m6A site prediction. BMC Bioinf. *22*, 288. https://doi.org/10.1186/s12859-021-04206-4.

43. Zhang, L., Chen, S., Ma, J., Liu, Z., and Liu, H. (2021). REW-ISA V2: a biclustering method fusing homologous information for analyzing and mining Epi-transcriptome data. Front. Genet. *12*, 654820. https://doi.org/10.3389/fgene.2021.654820.

44. Ma, J., Zhang, L., Chen, J., Song, B., Zang, C., and Liu, H. (2021). m7GDisAI: N7-methylguanosine (m7G) sites and diseases associations inference based on heterogeneous network. BMC Bioinf. *22*, 152. https://doi.org/10.1186/s12859-021-04007-9.

45. Zhang, L., Chen, J., Ma, J., and Liu, H. (2021). HN-CNN: a heterogeneous network based on convolutional neural network for m7 G site disease association prediction. Front. Genet. *12*, 655284. https://doi.org/10.3389/fgene.2021.655284.

46. Wu, X., Wei, Z., Chen, K., Zhang, Q., Su, J., Liu, H., Zhang, L., and Meng, J. (2019). m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network. BMC Bioinf. *20*, 223. https://doi.org/10.1186/s12859-019-2840-3.

47. Gao, Z., and Ruan, J. (2015). A structure-based Multiple-Instance Learning approach to predicting in vitro transcription factor-DNA interaction. BMC Genom. *16* (*Suppl 4*), S3. https://doi.org/10.1186/1471-2164-16-s4-s3.

48. Gao, Z., and Ruan, J. (2017). Computational modeling of in vivo and in vitro protein-DNA interactions by multiple instance learning. Bioinformatics *33*, 2097–2105. https://doi.org/10.1093/bioinformatics/btx115.

49. Zhang, Q., Zhu, L., Bao, W., and Huang, D.-S. (2020). Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. IEEE/ACM Trans. Comput. Biol. Bioinf. *17*, 679–689. https://doi.org/10.1109/TCBB.2018.2864203.

50. Zhang, Q., Shen, Z., and Huang, D.-S. (2019). Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. Sci. Rep. *9*, 8484. https://doi.org/10.1038/s41598-019-44966-x.

51. Wu, J.S., Huang, S.J., and Zhou, Z.H. (2014). Genome-wide protein function prediction through multi-instance multi-label learning. IEEE/ACM Trans. Comput. Biol. Bioinf. *11*, 891–902. https://doi.org/10.1109/tcbb.2014.2323058.

52. Panwar, B., Menon, R., Eksi, R., Li, H.D., Omenn, G.S., and Guan, Y. (2016). Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning. J. Proteome Res. *15*, 1747–1753. https://doi.org/10.1021/acs.jproteome.5b00883.

53. Bandyopadhyay, S., Ghosh, D., Mitra, R., and Zhao, Z. (2015). MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci. Rep. *5*, 8004. https://doi.org/10.1038/srep08004.

54. Mei, S., and Zhu, H. (2014). AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between Salmonella and human proteins. PLoS One *9*, e110488. https://doi.org/10.1371/journal.pone.0110488.

55. Huang, D., Song, B., Wei, J., Su, J., Coenen, F., and Meng, J. (2021). Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data. Bioinformatics *37*, i222–i230. https://doi.org/10.1093/bioinformatics/btab278.

56. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1703.01365.

57. Shrikumar, A., Tian, K., Avsec, Z., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. (2018). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco). Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.00416.

58. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

59. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

60. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24. https://doi.org/10.1186/gb-2007-8-2-r24.

61. Kraus, O.Z., Ba, J.L., and Frey, B.J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics *32*, i52–i59. https://doi.org/10.1093/bioinformatics/btw252.

62. Dauphin, Y.N., Fan, A., Auli, M., and Grangier, D. (2017). Language modeling with gated convolutional networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1612.08083.

63. Sotoudeh, M., and Thakur, A.V. (2019). Computing linear restrictions of neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.06214.

64. Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning important features through propagating activation differences. Preprint at arXiv. https://doi.org/10.48550/arXiv.1704.02685.