Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA

QUAN ZOU,[1,2] PENGWEI XING,[2] LEYI WEI,[2] and BIN LIU[3]

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 610051 Chengdu, China
[2]School of Computer Science and Technology, Tianjin University, 300350 Tianjin, China
[3]School of Computer Science and Technology, Harbin Institute of Technology, 150001 Shenzhen, China

ABSTRACT

N^6-Methyladenosine (m^6A) refers to methylation modification of the adenosine nucleotide acid at the nitrogen-6 position. Many conventional computational methods for identifying N^6-methyladenosine sites are limited by the small amount of data available. Taking advantage of the thousands of m^6A sites detected by high-throughput sequencing, it is now possible to discover the characteristics of m^6A sequences using deep learning techniques. To the best of our knowledge, our work is the first attempt to use word embedding and deep neural networks for m^6A prediction from mRNA sequences. Using four deep neural networks, we developed a model inferred from a larger sequence shifting window that can predict m^6A accurately and robustly. Four prediction schemes were built with various RNA sequence representations and optimized convolutional neural networks. The soft voting results from the four deep networks were shown to outperform all of the state-of-the-art methods. We evaluated these predictors mentioned above on a rigorous independent test data set and proved that our proposed method outperforms the state-of-the-art predictors. The training, independent, and cross-species testing data sets are much larger than in previous studies, which could help to avoid the problem of overfitting. Furthermore, an online prediction web server implementing the four proposed predictors has been built and is available at http://server.malab.cn/Gene2vec/.

Keywords: N^6-methyladenosine; machine learning; deep learning; RNA word embedding; mRNA

INTRODUCTION

N^6-Methyladenosine (m^6A) is an RNA methylation modification at the nitrogen-6 position of the adenosine base. It has been identified as the most commonly modified base in the messenger RNA of most eukaryotes. Research has shown that m^6A modification is involved in numerous biological activities, including the differentiation and reprogramming of stem cells (Yue et al. 2015), translation and alternative splicing (Geula et al. 2015; Xu et al. 2018), circadian clock (Fustin et al. 2013), and cerebellar development (Wang et al. 2018). Research in cancer biology has also shown that m^6A mRNA modification plays a critical role in glioblastoma stem cell self-renewal and tumorigenesis (Cui et al. 2017; Zhang et al. 2017), and m^6A modification was also shown to exert anti-leukemic activity in recent studies (Li et al. 2017b; Su et al. 2018). Moreover, Lichinchi et al. showed that viral infection triggers a massive increase in m^6A in both host and viral mRNAs (Lichinchi et al. 2016a), and similar regulatory mechanisms of several other viruses have also been confirmed (Gokhale et al. 2016; Lichinchi et al. 2016b; Hesser et al. 2018). Furthermore, high-throughput analysis of m^6A, for instance, using the RNA immunoprecipitation biotechnologies MeRIP-seq and m6A-seq (Dominissini et al. 2012; Meyer et al. 2012), has provided insights into the functions and topological patterns of m^6A mRNA modification. Based on information from MeRIP-seq experiments, comprehensive databases of m^6A modification have been built to help researchers determine the locations and effects of these modifications (Liu et al. 2017a; Xuan et al. 2017). The work of Wan et al. (2015) showed that m^6A patterns are similar between plants and mammals, with both being abundant near stop codons and 3′ untranslated regions (UTRs) and having similar consensus m^6A methylation motifs and similar frequencies of m^6A sites per transcript in the transcriptome. Following the profiling of m^6A distributions in mammalian transcriptomes (Dominissini et al. 2012; Meyer et al. 2012) and the mapping of the yeast m^6A methylome (Schwartz et al.

Corresponding authors: zouquan@nclab.net, bliu@hit.edu.cn
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.069112.118.

2013), transcriptome-wide profiling of m^6A in two accessions of *Arabidopsis thaliana* (Luo et al. 2014) was performed, which indicated that m^6A is a highly conserved mRNA modification in plants and that there is a positive correlation between m^6A deposition and mRNA abundance. In a recent study, it was also shown that m^6A is regulated by microRNA via complementary sequence motifs (Chen et al. 2015a), which suggested that sequence signaling is very important for m^6A sites. Recently, a novel, single-base-resolution technique, miCLIP-seq (Ke et al. 2015; Linder et al. 2015), was also established, which prompted a new wave of research on computational methods of m^6A identification (Xiang et al. 2016; Zhou et al. 2016).

Notably, m^6A sites have methylation-specific surroundings, with the topology of a DRACH (where D = A, G, or U; R = A or G; and H = A, C, or U) consensus motif and a GAC consensus motif localized near stop codons, in 3′-UTRs, within long internal exons, and at 5′-UTRs (Meyer et al. 2012; Schwartz et al. 2013; Li et al. 2014; Luo et al. 2014; Zhou et al. 2016). Furthermore, the conserved Pu[G > A]m^6AC[A/C/U] consensus motif dominates mammalian m^6A sites (Dominissini et al. 2012). However, only ∼15% of all methylation Pu[G > A] m^6AC[A/C/U] consensus motifs are m^6A sites (Yue et al. 2015). Identification of the actual methylated m^6A sites among these consensus motifs remains a problem. High-throughput sequencing and wet experiments could not solve this problem due to the cost and time-consuming nature of the research, as well as inaccuracy regarding the identified sites. Therefore, computational tools were required to guide the accurate prediction of modification sites and to help reduce the costs associated with high-throughput sequencing.

In the above context, computational tools were developed for detecting different modification sites, including protein methylation (Wei et al. 2018d), protein phosphorylation (Wei et al. 2017) and dephosphorylation (Jia et al. 2017), protein O-GlcNAcylation (Jia et al. 2018), histone crotonylation (Qiu et al. 2017), DNA N^4-methylcytosine (Chen et al. 2017c; Wei et al. 2018b), RNA pseudouridine (Chen et al. 2016b), and various RNA adenosine modifications (Chen et al. 2018). However, it has been proven that sequence alignment (e.g., PSI-BLAST) cannot accurately identify the modification sites. Instead, machine learning techniques are used in approaches that are currently popular. For these, a sliding window is selected around the candidate modification sites. Then, sequences in the sliding window are collected for standard machine learning training and testing processes. Related features are then proposed for representing the sliding window sequences with equal length (Liu et al. 2015, 2018; He et al. 2018). However, in this context, there is a major problem regarding construction of the training data set. Specifically, low-quality negative samples cause low generalizability of the model, resulting in poor performance when applied to novel data. In addition, the compatibility of prediction methods in different species also remains a problem, so researchers cannot currently be certain that cross-species predictions are accurate.

The yeast data set (Schwartz et al. 2013) and *Arabidopsis* data set (Luo et al. 2014) are two benchmark data sets for the computational prediction of m^6A. Focusing on the *Arabidopsis* data set, Chen and coworkers first proposed a support vector machine-based method to identify m^6A sites. Soon afterwards, they proposed a predictor called "iRNA-Methyl" (Chen et al. 2015b) on a near single-nucleotide resolution yeast data set. Chen and coworkers represented RNA sequences using "pseudo-dinucleotide composition," which focuses on the physiochemical properties of RNA. Other improved works in terms of prediction accuracy on the two data sets were also presented (Chen et al. 2017b; Xing et al. 2017). Recently, Zhou et al. established an m^6A data set from published single-nucleotide-resolution maps of human and mouse m^6A sites (Ke et al. 2015; Linder et al. 2015). They then developed an m^6A predictor named SRAMP (Zhou et al. 2016), which simply uses three sequence-derived features with Random Forest classifiers. Following this work, Xiang et al. (2016) improved the predictive performance by integrating multiple sequence features, including positional binary nucleotide sequence encoding, nucleotide pair spectrum encoding, position-specific encoding, and *k*-mer nucleotide frequency encoding. These methods were used with conventional features in position and frequency statistics, with the sliding window being limited to a narrow region and focusing on only a single species. Additionally, there was a lack of independent testing, which resulted in overfitting.

To solve the above problems, we used convolutional neural network (CNN) for m^6A prediction. CNNs have been applied in various fields of bioinformatics (Zhang et al. 2018), including regulatory genomics (Angermueller et al. 2016; Xu et al. 2017), drug discovery (Stephenson et al. 2018), protein subcellular localization (Almagro Armenteros et al. 2017; Wei et al. 2018a), protein function prediction (Cao et al. 2017), and single-cell DNA methylation states (Angermueller et al. 2017; Xu and Zhou 2018). These networks directly trained predictor models without predefined features and outperformed conventional predictors with larger sequence sliding windows. The combination of big data and larger sliding windows together with deep learning techniques appeared to solve the m^6A overfitting problem. It has also become common to represent sequences with word embedding algorithms, instead of the sparse one-hot encoding (Dai et al. 2017; Min et al. 2017; Wei et al. 2018c).

In this paper, we report a neural embedding predictor named Gene2vec (gene subsequence to embedding vector). We extended the sliding window length to the thousand level, used word embedding to represent mRNA subsequences, which were parts of long sliding window sequences, and performed classification with CNNs. We

progressively experienced various gene sequence representation schemes with their most effective CNN structures with library Keras (https://keras.io/) and proved that the gene-subsequence-based neural embedding method is the best option for various large-scale sequence data. It appeared that CNN together with Gene2vec can address the problem of overfitting in m^{6}A prediction. To our knowledge, this is the first time to test ~1000 nt length sliding frame, and we used much more training and testing sequences than before, which could help to avoid overfitting of deep learning techniques.

RESULTS AND DISCUSSION

Optimization of parameters

Large sequence windows confer more contextual sequence information and greater GAC/AAC site coverage. It is an important ingredient of the contributions to effective performance. We evaluated the effect of different sequence windows on the prediction results with simple one-hot encoding using the two convolutional cell structures mentioned above (Fig. 1). The results in Figure 1 demonstrated that the AUROC and MCC have growth trend integrally with the increase of the sequence window length and no longer change drastically when it is up to 1001 nt.

The selection of RNA word split length is very important for embedding and Gene2vec prediction mode. Too short a length will lead to a small number of nucleotide letter combinations, which in extreme conditions will degenerate into one-hot encoding with assigning four types of a unique integral index to each individual nucleotide, while an excessive length will result in too many combinations, leading to complex vector representation and high computing costs. To determine the optimal RNA word split length, we compared the performance on a validated set with different slice lengths of RNA word from two to five nucleotides (Fig. 2). As shown in Figure 2, three nucleotides is an appropriate length in this range with metrics of both AUROC and MCC. We also compared the results of different embedding output dimensions of word embedding methods with two convolutional cell structures on a validation set (Fig. 3). We used AUROC and MCC metrics to compare performances of different embedding output dimensions of word embedding. As shown in Figure 3, the performances of 128 output dimensions reach a peak with both AUROC and MCC metrics in an appropriate range of [64,256], though the performance of 256 output dimensions with MCC metrics is slightly higher than the 128 output dimensions. As a result, we chose 128 as an appropriate word embedding output dimension parameter.

The main network hyperparameters of the four prediction modes are the convolution number of filters and size of the convolution kernel. For the convolution number of filters, we first empirically set this as decreasing power

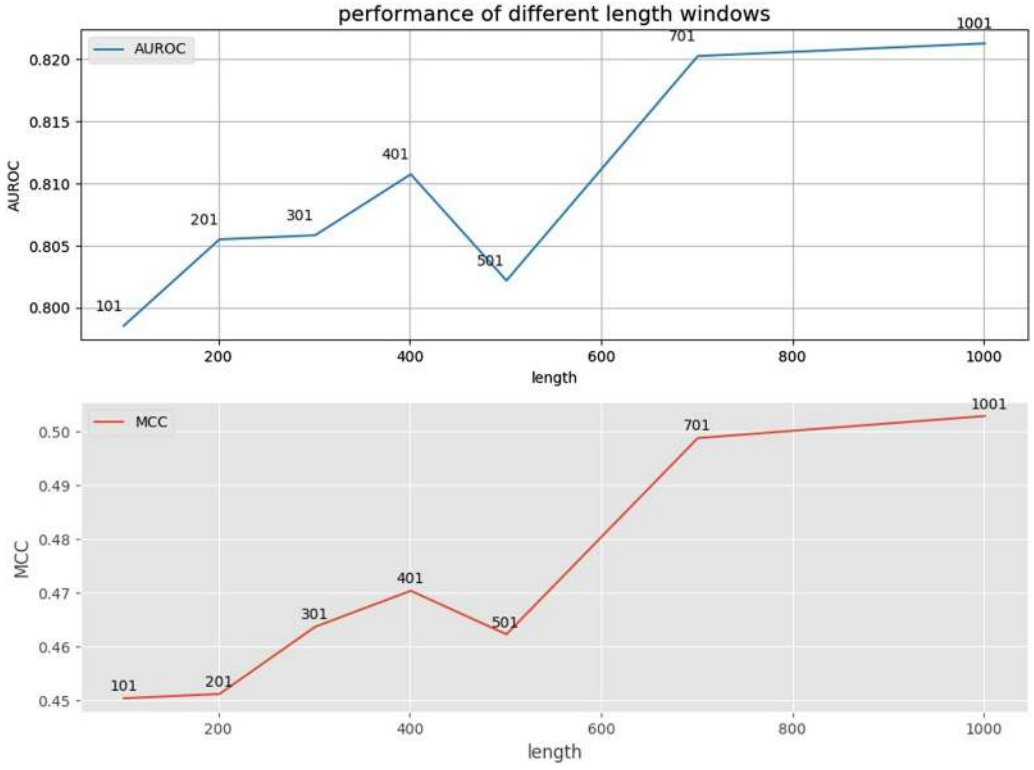

FIGURE 1. Performance of different length windows with one-hot encoding on the validation set.

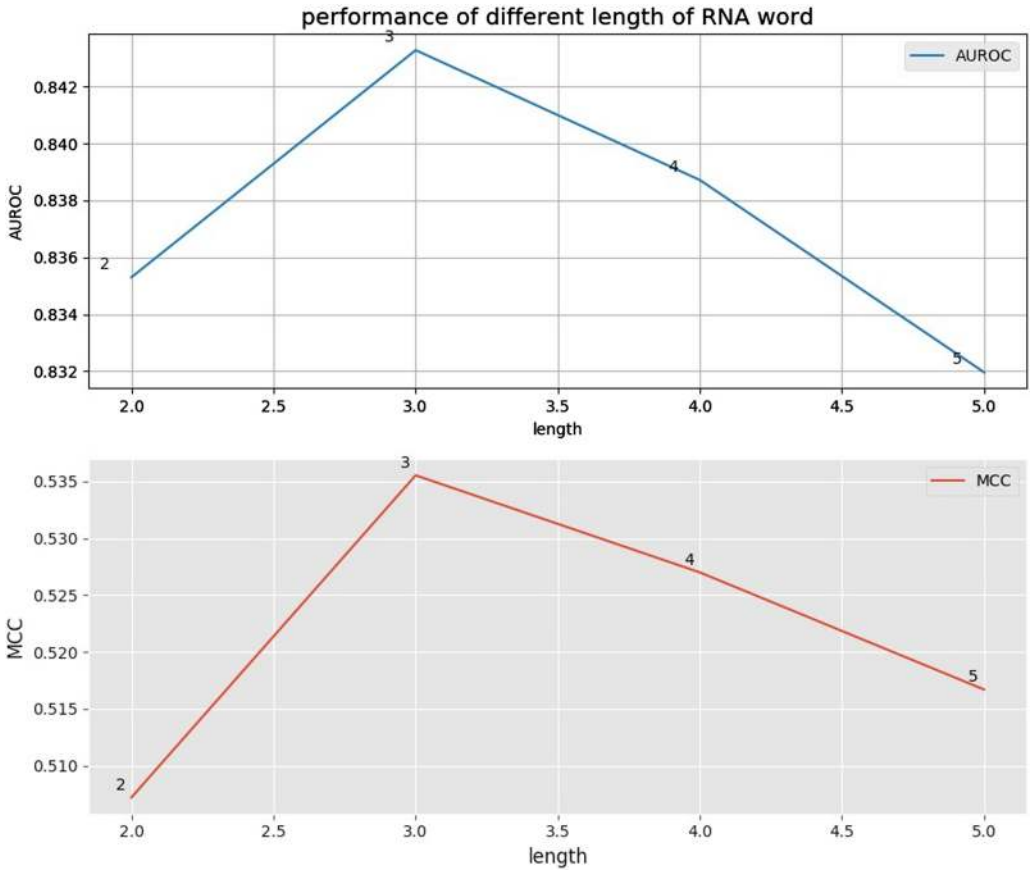

FIGURE 2. Performance of different lengths of RNA words with word embedding on the validation set.

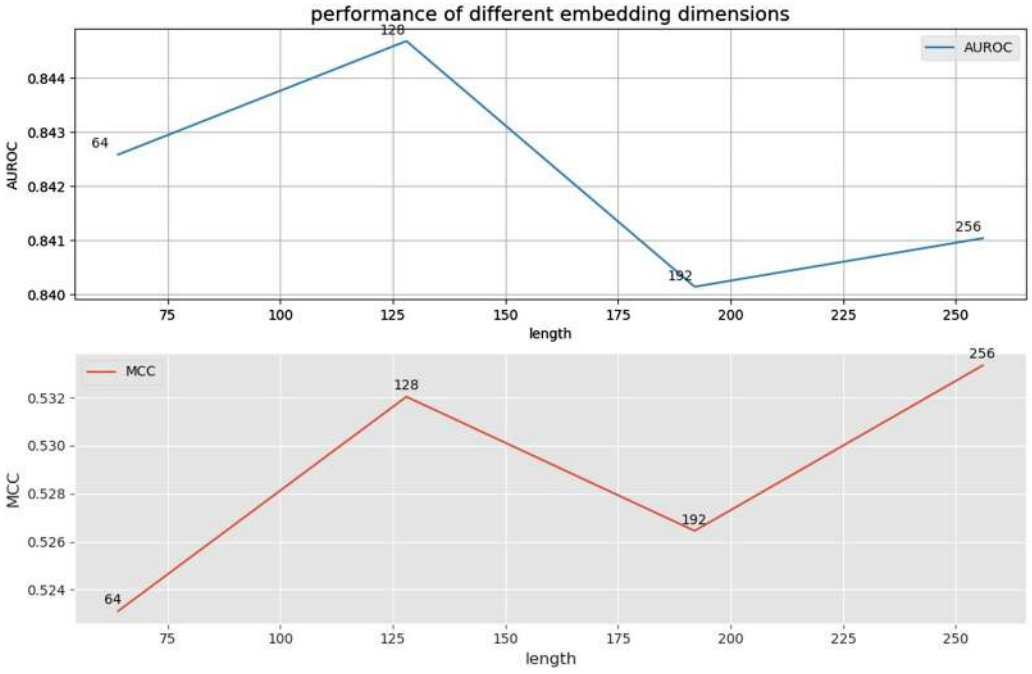

FIGURE 3. Performance of different embedding dimensions with word embedding on the validation set.

values of 2, and for the size of the convolution kernel, we set it as some decreasing odd values. The convolution network structures of all four prediction schemes are similarly constituted in terms of cell structures, the detailed hyperparameters of which are supplied in the Supplemental Material.

Learned and analyzed motifs from CNN kernel

Deep learning is to a certain extent a black box with the difficulty in tracing a prediction back to which features are important. And it is very meaningful to explain biological meaning in the process of training of CNN with visualization. Recently, many studies on biological computing prediction classifications (Liu and Li 2018; Li et al. 2017a; Liu et al. 2017c; Zeng et al. 2018) involving CNNs have used convolution kernels of the first layer to extract informative motifs from massive sequence data sets. These followed on from the heuristic work from Deepbind (Alipanahi et al. 2015) that generated a position weight matrix (PWM) by aligning all matched sequence segments and calculating the frequency for each kernel. We here apply this new method to achieve conversion from convolution kernels to PWMs on single training sets from humans and mice. For compatibility regarding representation of the four nucleotides in PWMs, we retrained a CNN model with 4-nt binary representation by randomly transforming the padding character into one of the four nucleotides.

We used 64 convolution kernels with a length of 11 binary representations for each kernel on the first CNN layer and generated 64 motifs after transforming kernels to PWMs. For further analysis of these motifs, we used the TOMTOM (Gupta et al. 2007) motif comparison tool to compare one or more motifs against a database of known RNA motifs. We picked out several representative examples of similar known motifs found in previous research (Ray et al. 2013) on humans and mice, as shown in Figure 4. The matching metrics shown are *E*-value and the number of overlaps, with *E*-value being the expected number of false positives in the matches up to this point. The comparative results in Figure 4 show that the CNN kernel or filter weight learning from an abstract representation of a deep neural network by scanning convolution operation corresponds to response function of consensus motifs in biological sequences. For instance, the 54th kernel in 64 convolution kernels is very similar to the RNCMPT00041 motif, which we analyzed from RNA-binding protein Musashi homolog 1 (MSI1), encoded by the MSI1 gene. On the one hand, the known consensus motifs matched

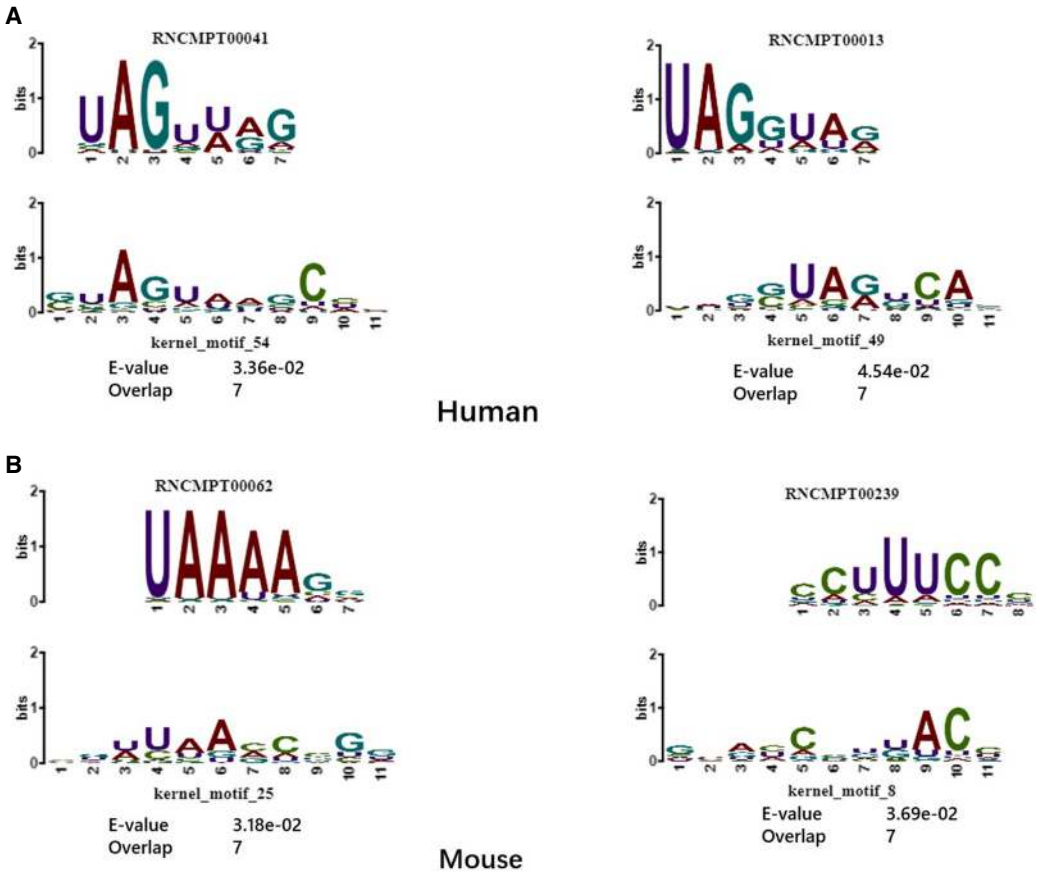

FIGURE 4. Comparison of motifs from CNN kernel and known motifs for (*A*) humans and (*B*) mice.

with convolution kernels help researchers to study the relationship between them in biological experiments. On the other hand, the unmatched kernels may help to uncover new motifs. All 64 motifs of PWM logos and a detailed motif comparison report can be accessed in the Supplemental Material or on our web server.

Performance of the predictors

We compared four prediction modes with four different data preprocessing and encoding schemes on 1001-nt sequence data extracted from the unbalanced independent test set transcript ID of Zhou et al. (2016) (Fig. 5). For building a prediction model with an unbalanced training set with a positive-to-negative ratio of 1:10, the same numbers of negative samples and positive samples were randomly selected as training data sets. We used 80% of the training set for building a model and used the other 20% of the training set to verify the model and optimize the neural network parameters, so we evaluated the performance of our predictor on an unbalanced independent test set. With the same unbalanced independent test set, neural networks based on all four predictors achieved better prediction than SRAMP, with the Gene2vec method achieving the best results (Table 1).

To use these four learning algorithms to obtain better predictive performance, we also established a soft vote with average predicted probabilities. We compared the ensemble result using the metrics of accuracy (Acc), *Sn*, *Sp*, and *MCC* in the unbalanced independent test set. As shown in Figure 6, all four metrics of the soft vote showed better results than for any of the single prediction methods. Note that AUROC and AUPR metrics were not considered here due to their step-wise character.

Cross-validation (10-fold) was performed with mature mRNA data from two different species and the corresponding trained Gene2vec model. The different predictive AUROC values are shown in Figure 7. We also added a mixed model built using the above-mentioned human and mouse data to predict independent species test data. As the figure shows, the model built using mouse data was less effective at prediction than the model built using human data due to the smaller amount of mouse training data available. We also found in the cross-species validation that the prediction results were poorer than when using the species-consistent data and model, which indicates the specificity of our method for a particular species. Furthermore, we obtained almost the same AUROC values of 0.8415 and 0.8414 for human and mouse data predicted using the mixed model, which are consistent with the mixed test data prediction result obtained with Gene2vec as shown in Table 1.

Further assessment of robustness of prediction model was performed in the YTHDF binding site data set. YTHDF proteins are m^{6}A readers. The above predictor could not only identify N^6-methyladenosine sites, but should also predict YTHDF protein binding sites that recognize m^{6}A-modified mRNAs by selection. Our method performed better with AUROC = 0.737 and AUPRC = 0.963 (Fig. 8) than SRAMP with AUROC = 0.720 and AUPRC = 0.251 as reported previously for the YTHDF binding site data set.

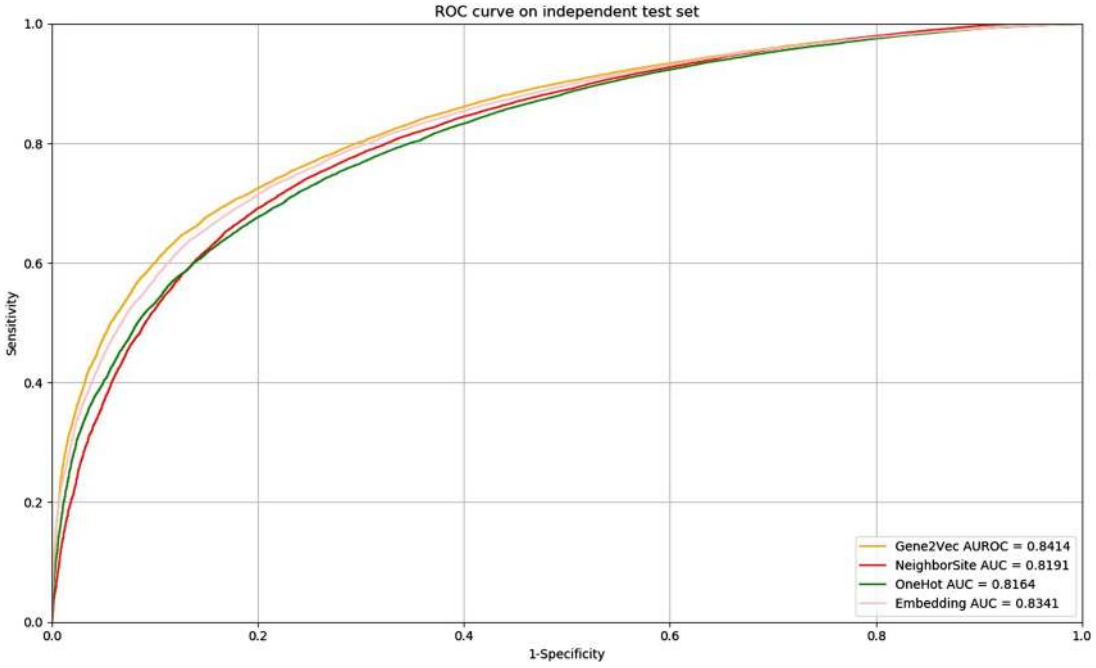

FIGURE 5. Performance of our four different predictors on an unbalanced independent test set.

TABLE 1. Comparison of different methods on an unbalanced independent test set

Predictors	AUROC	AUPR
OneHot	0.816	0.976
NeighboringSite	0.819	0.975
Embedding	0.834	0.979
Gene2vec	0.841	0.980
SRAMP	0.794	0.321

We also compared our method with the newest predictor of mammalian N^6-methyladenosine sites of which we are aware, RNAMethPre, for a remapped unbalanced independent test set. We used the balanced training set from RNAMethPre with Gene2vec encoding and the same CNN structure to build a prediction model for the purpose of establishing consistent comparison conditions. As Table 2 shows, RNAMethPre was better than SRAMP in terms of predictive performance with three stringency thresholds corresponding to 90%, 85%, and 80% specificity in the independent data set tests. Our method achieved more effective prediction than the RNAMethPre predictor with these four thresholds.

Conclusions

In this study, we built four prediction modes for predicting mammalian N^6-methyladenosine sites with various RNA sequence representations and optimized CNN structure. We achieved more effective prediction than in conventional methods with information on the context regarding flanking sequences for 1000 nucleotides. Specifically, we combined deep learning and word embedding technology to present an RNA N^6-adenosine methylation predictor based on gene-subsequence-based neural embedding algorithms by splitting the sequence into pseudo-RNA words. We also used these pseudo-RNA words to train a corpus model. Using this model, we analyzed and explained semantic equivalence and semantic symmetry phenomena of RNA sequences with vector space presentation. We also trained an embedding model and a network model, and optimized encoding parameters and network parameters on a validation set. We evaluated our method on a rigorous unbalanced independent mammalian m^{6}A site test set and YTHDF binding site test set and achieved better results than with the conventional method. We also established a user-friendly web server at http://server.malab.cn/Gene2vec/, where users can submit uncharacterized mRNA sequences for the prediction of potential m^{6}A sites. In the future, we will pay more attention to the genomics data (e.g., tRNA, rRNA) besides mRNA sequences if related training dates are released. Supplemental Material can be accessed on our web server.

MATERIALS AND METHODS

Data sets

Data sets were retrieved from *Homo sapiens* and *Mus musculus* complementary DNA (cDNA) FASTA data in Ensembl (Zerbino et al. 2017). The mature mRNA transcript IDs were derived from the work of Zhou et al. (2016), which involved annotation of the mammalian m^{6}A sites. In our work, RNA sequences were derived from and were equal to cDNA. Besides the data sets of Zhou and

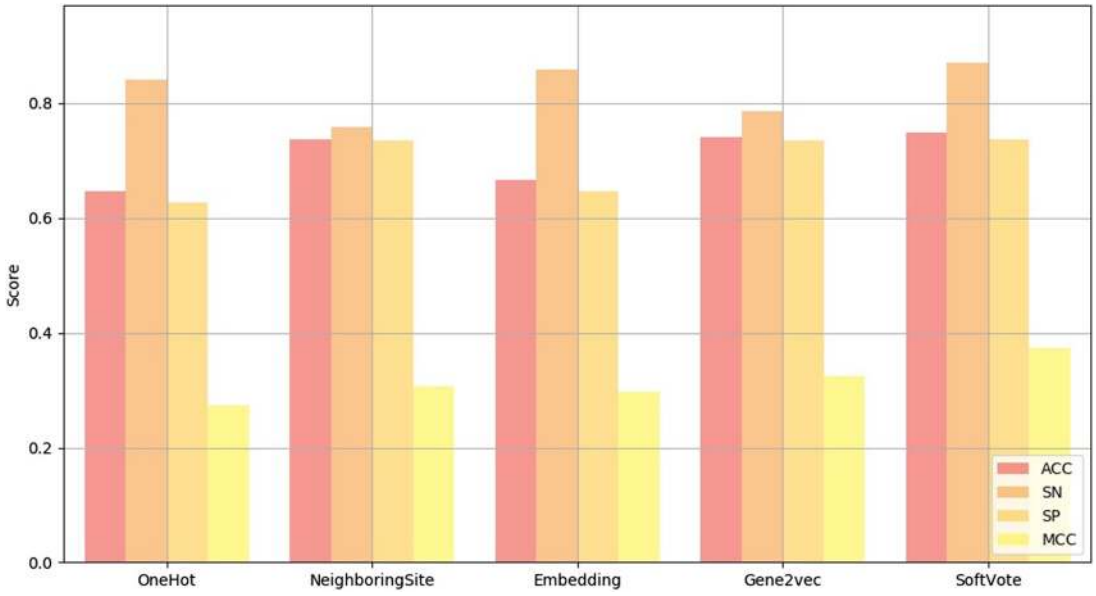

FIGURE 6. Comparisons of soft vote and single methods on an unbalanced independent test set.

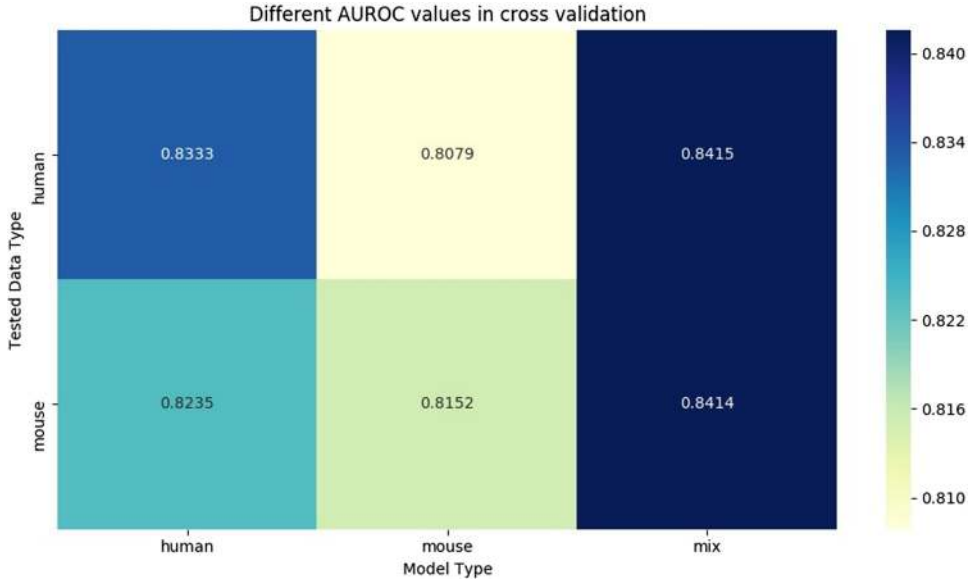

FIGURE 7. Heat map of different AUROC values in cross-species validation.

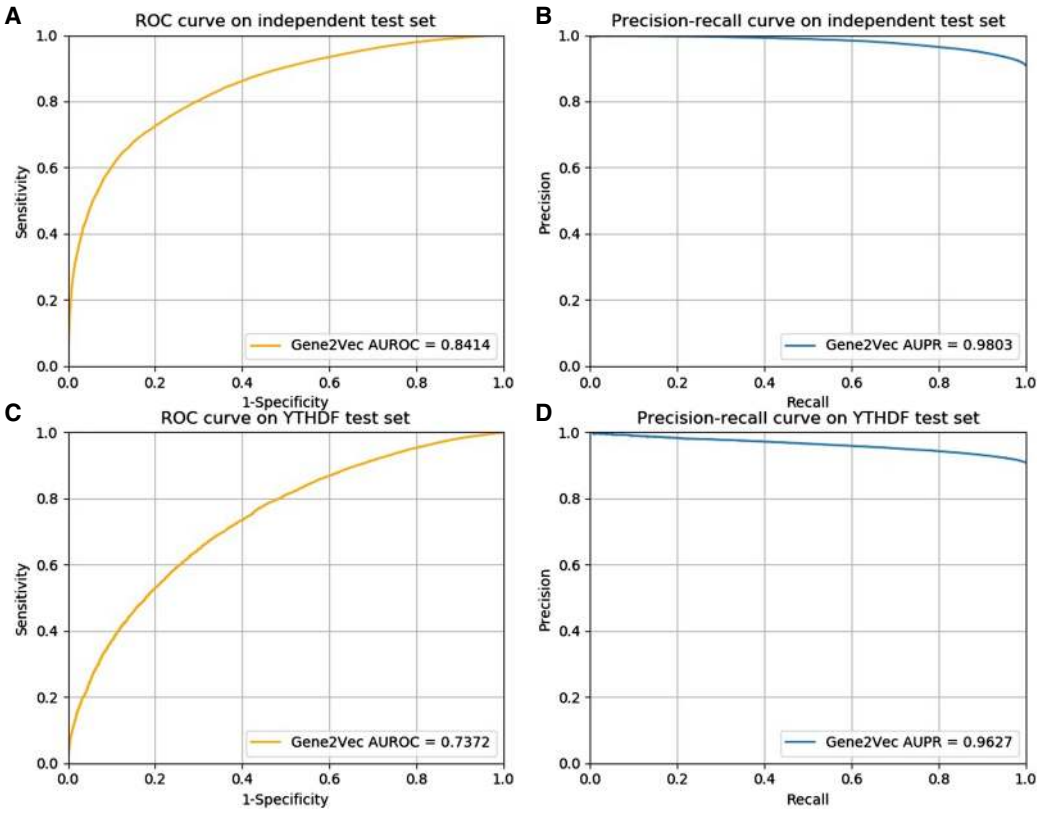

FIGURE 8. Performance of our method on an unbalanced independent test set and YTHDF binding site test set. (*A*) ROC curve illustrating the performance of the Gene2vec predictor on the independent test data set. (*B*) Precision–recall curve illustrating the performance of the Gene2vec predictor on the independent test data set. (*C*) ROC curve illustrating the performance of the Gene2vec predictor on the YTHDF binding site test set. (*D*) Precision–recall curve illustrating the performance of the Gene2vec predictor on the YTHDF binding site test set.

TABLE 2. Performance of various stringency thresholds with RNAMethPre and SRAMP

		Sensitivity			MCC		
Confidence	Specificity	SRAMP	RNAMethPre	Gene2vec	SRAMP	RNAMethPre	Gene2vec
High	0.90	0.44	0.47	0.63	0.29	0.31	0.50
Moderate	0.85	0.54	0.56	0.71	0.29	0.31	0.48
Low	0.80	–	0.64	0.77	–	0.30	0.45

coworkers, we built other data sets in accordance with the work of Xiang et al. (2016) for comparative experiments. For further confirming the validity of the prediction model, Zhou and coworkers also used YT521-B homology domain family (YTHDF) proteins which are m^{6}A readers. RNA sequences with m^{6}A sites would be bound by YTHDF. Therefore, YTHDF-binding RNAs could be selected as those with m^{6}A sites.

For all RNA nucleotide sequence data, we cut out a 1001-nt local sequence window centered at an m^{6}A/non-m^{6}A site, removing the sequences whose centers were not GAC or AAC consensus motifs, and filled in the rest with the character "X" when the sequence was shorter than 1001 nt. Finally, we obtained 495,572 sequences in the training set and 128,561 in the testing set. For both of these, the positive-to-negative ratio was approximately 1:10.

The YTHDF-binding data set contained 57,516 testing sequences. There were 88,579 training sequences (positive-to-negative ratio of 1:1) and 88,227 testing sequences (positive-to-negative ratio 1:10) in the work of Xiang et al. (2016). The sizes of the different data sets are shown in Table 3. All of these sequences can be accessed in the Supplemental Material.

Data preprocessing and encoding

In this section, we introduce four sequence-encoding schemes, namely, one-hot encoding, neighboring methylation state encoding, RNA word embedding, and Gene2vec. The DNA sequences were represented by the four encoding schemes.

One-hot encoding

In one-hot encoding, there are five characters, standing for the four types of nucleotide along with the padding character "X." One-hot encoding uses five-dimensional binary vectors, where A = [1, 0, 0, 0, 0], T = [0, 1, 0, 0, 0], G = [0, 0, 1, 0, 0], C = [0, 0, 0, 1, 0], and X = [0, 0, 0, 0, 1] and transforms the 1001 nt windows sequences centered at the prediction site into 5005-binary-long vector.

Neighboring methylation state encoding

m^{6}A sites appear to cluster in the chromosomes. Therefore, the positive sites may be close to other positive sites, instead of negative ones. For neighboring methylation state encoding, we counted the neighboring positive/negative site numbers as a kind of feature. We scanned GAC and AAC sites throughout the entire transcript sequences, and gave a label of 1 to known methylation (m^{6}A) sites and 0 to unknown GAC/AAC sites. For every positive (label 1) or negative (label 0) site, we listed the 250 upstream GAC/AAC sites and 250 downstream GAC/AAC sites. Therefore, we extracted 501 0/1 codes for every site. If the end

TABLE 3. The sizes of different data sets used in our work (positive:negative)

		Training set	Windows
Data sets built and used in our work	Zhou et al. (2016) *Homo sapiens* and *Mus musculus*	Training set + Validation set: 495572 (1:10) Test set 128561 (1:10)	1001 nt
	Zhou et al. (2016) YTHDF binding *Homo sapiens*	Test set: 57516 (1:10)	1001 nt
	Xiang et al. (2016) *Homo sapiens* and *Mus musculus*	Training set + Validation set: 88579 (1:1) Test set:88227 (1:10)	1001 nt
Other researchers' data sets	iRNA-Methyl (Chen et al. 2015b) *Saccharomyces cerevisiae*	Training set + Test set: 2614 (1:1)	51 nt
	M6ATH (Chen et al. 2016a) *Arabidopsis thaliana*	Training set + Test set: 788 (1:1)	25 nt
	MethyRNA (Chen et al. 2017a) *Homo sapiens* and *Mus musculus*	Training set + Test set: 3710 (1:1)	41 nt
	M6aPred (Chen et al. 2015c) *Saccharomyces cerevisiae*	Training set: 1664 (1:1) Test set: 5225 (1:10)	21 nt
	RFAthM6A (Wang and Yan 2018) *Arabidopsis thaliana*	Training set: 4200 (1:1) Test set: 836 (1:1)	101 nt

of the transcript was reached, we filled the vacancy with values of 0 in both sides to the 501-code-longvector.

RNA word embedding

For RNA word embedding, heuristically, we shifted a 3-nt-long window along 1001-nt sample sequences to generate RNA subsequences that can be analogized into gene words. Embedding encoding identified possible 3-nt combinations (105 different combinations in our training data) with a unique integral index, and transformed sample sequences into integral sequences with the corresponding integral index. Then, we input them into the Keras embedding layer to transform each integral sequence into a data table $S \in \mathbb{R}^{n \times e}$, where n is the length of the integral sequence and e is the dimension of the dense embedding.

Gene2vec

Word2vec (Church 2017) is a statistical method for learning word embedding from a text corpus with neural-network-based training via Skip-gram and continuous bag-of-words (CBOW) models. The Skip-gram model predicts the surrounding words from the current word, while the CBOW model predicts the current word from its surroundings. Both models are focused on learning about words given their local usage context, where the context is defined by a window of neighboring words.

For Gene2vec, similar to the processing of embedding encoding, we regarded lengths of three RNA nucleotides as an RNA word, in an overlapping manner, analyzed them for sequence content as the RNA corpus, and used Word2vec in the Gensim tool package (https://radimrehurek.com/gensim/models/word2vec.html) with a five-word-long window of neighboring words to learn the vector relationship of those RNA words and generate a 100-dimensional feature vector.

An RNA word co-occurrence network was built, as shown in Figure 9A, to find similarities between word pairs and the word patterns of latent structures and representations. This network was built using a square symmetric matrix, which was the result of multiplying a rectangular matrix by its transpose. Each row vector of this rectangular matrix represents a different RNA context unit (S_1, S_2, $S_3 \ldots S_n$) corresponding to a sentence or paragraph, and each column-vector represents a different word (W_1, W_2, $W_3 \ldots W_n$). Therefore, in the RNA word co-occurrence matrix, each row or column stands for an RNA word. Each cell stands for a context unit size, which is the number of RNA words W_i co-occurring with the RNA word W_j. Correlation function was computed to find the correlations, and then a t-test was used on the individual correlations using the following formulas in the Psych package (https://cran.r-project.org/web/packages/psych/):

$$t = \frac{r * \sqrt{(n-2)}}{\sqrt{(1-r^2)}},$$

$$\text{se} = \frac{\sqrt{(1-r^2)}}{\sqrt{(n-2)}},$$

where r is the matrix of correlations and n is the number of cases per correlation. We built the co-occurrence network by setting a correlation threshold of 0.7 and linked two points together with an edge when their correlation exceeded this threshold. For instance, the group [AGA, GAA, AAG] (green group in Fig. 9A) can be selected among 4^3 types of 3-nt-long RNA sequence permutations as an independent group, due to their strong correlation.

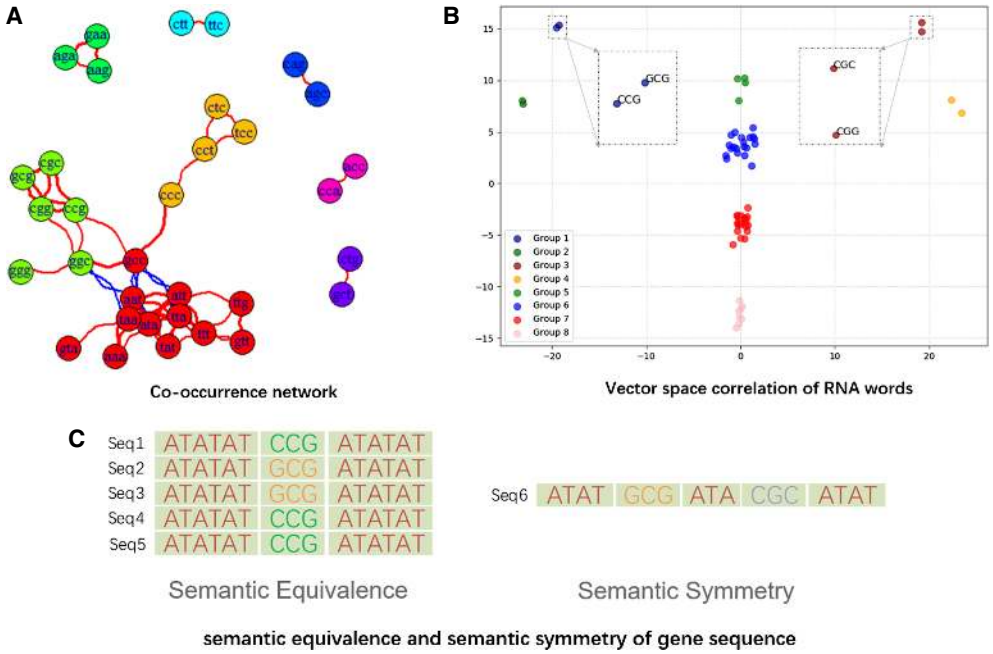

FIGURE 9. RNA word correlation analysis. (*A*) RNA words co-occurrence network. If there is an edge in the network, it means that the two RNA words would co-appear in the m^{6}A sentences. (*B*) RNA words were transformed into vector space. From this figure, we can conclude that GCG/CCG always appears as a pair-word. It is also the same as CGC/CGG. (*C*) The figure shows examples of semantic equivalence and semantic symmetry.

We also removed RNA words containing the character "X" and transformed this 100-dimensional vector (Gene2vec output) of 4^3 types of RNA words to two components utilizing principal component analysis (PCA) to reveal the two-dimensional spatial correlation of those RNA words (Fig. 9B). Spatial distance is representative of word similarity. All 2D and 3D graphs with detailed annotations can be obtained on our web server. As can be seen in Figure 9B, 64 RNA words cluster into eight groups with axial symmetry. For instance, Group 1 [CCG, GCG] and Group 3 [CGG, CGC] are symmetrical about the vertical, which first indicates that CCG has a similar "meaning" for sequence composition to GCG due to their cluster in vector space, and second reveals that CCG is to CGG as GCG is to CGC (Fig. 9B). This is just like "king" is to "queen" and "man" is to "woman" in natural language. We call the two phenomena gene semantic equivalence and gene semantic symmetry (Fig. 9C). Two RNA words are said to have semantic equivalence when they exist among different sequences that have almost the same gene context, while two RNA words are said to have semantic symmetry when they exist within one sequence that has a unique biochemical property of not only positional symmetry, but also nucleotide-order symmetry, just like a pair of hands where each finger represents a nucleotide letter.

Network structure

We used CNNs with multiple cell structures that have two one-dimensional convolution layers, one pooling layer and one dropout layer. Convolution layers are designed to extract features with high-dimensional abstract representation. The pooling layer limits the number of model parameters tractable by pooling operations. The dropout layer prevents overfitting of the model by randomly setting some of the input units to a value of 0. Four prediction methods had been established based on four different network structures composed by the cell structures mentioned above. One-hot encoding data were fed into the network with four cell structures and fully connected layers as input, while neighboring methylation state encoding data, RNA word embedding data, and Gene2vec processing data were fed into networks with two cell structures (Fig. 10). The final result was obtained by a voting strategy from the four prediction probabilities.

Taking an example of the one-hot coding sequence, the input data matrix X_n was first fed into a 1D-convolutional layer, which used a convolutional filter $W_f \in R^H$, where H is the length of the filter vector. The output feature A_i at the ith position was computed by

$$A_i = ReLU\left(\sum_{h=1}^{H} W_f X_{n,i+h} + b_f\right),$$

where $ReLU(x) = \max(0, x)$ is the rectified linear unit function and $b_f \in R$ is a bias (Mairal et al. 2014). These convolutional operations are similar to data block of H length in sequence filtered by a sliding filter window at each ith position.

Next, a max. pooling layer was used for reduction of the dimensions of output data generated by the multiple convolutional filter operations. A max. pooling layer is a form of nonlinear downsampling achieved by outputting the maximum of each subregion.

To reduce overfitting, we added a dropout layer in which individual nodes were either "dropped out" from the network with probability $1 - P$ or kept with probability P at each training stage. This not only prevented overfitting, but also led to integration of various deformed network structures to generate more robust features that are more generalizable to new data.

Finally, a flattening layer that "flattened" the input data was used, which transformed multidimensional data into a single dimension. Fully connected layers with an ReLU activation function and output layer predict the binary classification probability with activation function as follows (Han and Moraga 1995):

$$\hat{y}(x) = \text{sigmoid}(x) = \left(\frac{1}{1 + e^{-x}}\right).$$

Evaluation metrics

To assess the performance of our prediction model on an unbalanced independent test set, we used the following metrics (Liu

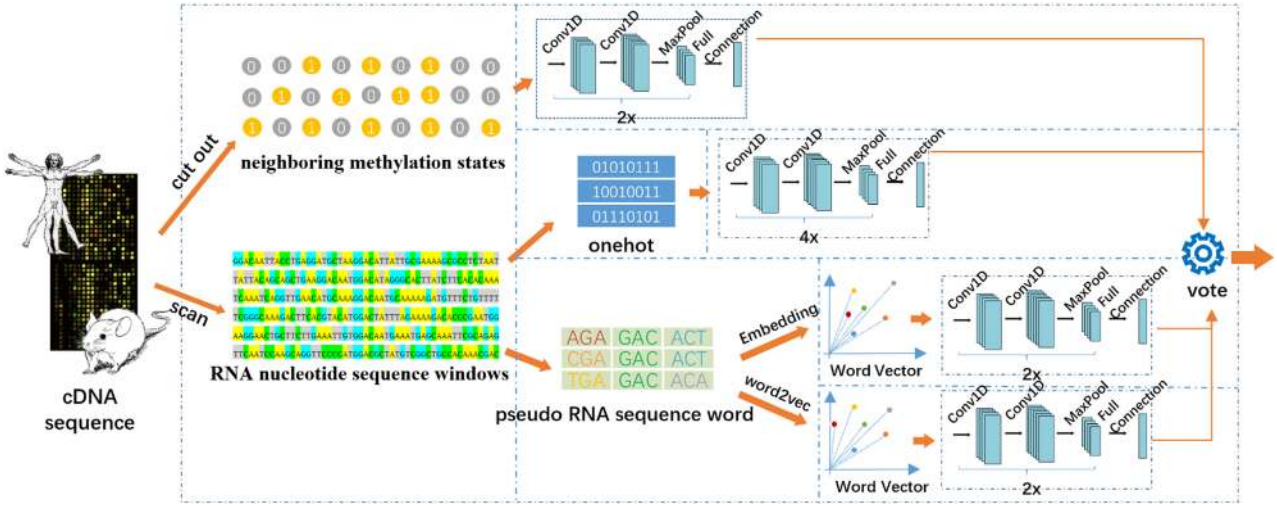

FIGURE 10. Workflow of multiple predictors. The figures showed the workflow of our method. The mRNA sequences were predicted by four different deep learning classifiers. Then they vote for the final results.

et al. 2017b): sensitivity (*Sn*), specificity (*Sp*), and Matthew's correlation coefficient (*MCC*), which are formulated as follows:

$$Sn = \frac{TP}{TP + FN} \times 100\%,$$

$$Sp = \frac{TN}{TN + FP} \times 100\%,$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) + (TN + FP) + (TP + FP) + (TN + FN)}},$$

where *TP*, *TN*, *FP*, and *FN* are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

The area under the ROC curve (AUROC) and the area under the precision–recall curve (AUPR) were calculated to evaluate the performance of the predictors. Receiver operating characteristic (ROC) curves can be plotted as sensitivity against 1 − specificity and precision–recall curves as precision (the proportion of true positives among all predicted positives) against recall (the proportion of relevant instances that have been retrieved among the total number of relevant instances). A precision–recall plot is more informative than an ROC plot when applied to unbalanced data sets (Song et al. 2014).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (SQ2018YFC090002) and the Natural Science Foundation of China (nos. 61771331, 61822306, 61672184). The authors thank Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

Received October 3, 2018; accepted November 1, 2018.

REFERENCES

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat Biotechnol* **33:** 831–838. doi:10.1038/nbt.3300

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33:** 3387–3395. doi:10.1093/bioinformatics/btx431

Angermueller C, Pärnamaa T, Parts L, Stegle O. 2016. Deep learning for computational biology. *Mol Syst Biol* **12:** 878. doi:10.15252/msb.20156651

Angermueller C, Lee HJ, Reik W, Stegle O. 2017. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18:** 67. doi:10.1186/s13059-017-1189-z

Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z. 2017. ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* **22:** E1732. doi:10.3390/molecules22101732

Chen T, Hao YJ, Zhang Y, Li MM, Wang M, Han W, Wu Y, Lv Y, Hao J, Wang L, et al. 2015a. m^{6}A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* **16:** 289–301. doi:10.1016/j.stem.2015.01.016

Chen W, Feng P, Ding H, Lin H, Chou KC. 2015b. iRNA-methyl: identifying N^6-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem* **490:** 26–33. doi:10.1016/j.ab.2015.08.021

Chen W, Tran H, Liang Z, Lin H, Zhang L. 2015c. Identification and analysis of the N^6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep* **5:** 13859. doi:10.1038/srep13859

Chen W, Feng P, Ding H, Lin H. 2016a. Identifying N^6-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol Genet Genomics* **291:** 2225–2229. doi:10.1007/s00438-016-1243-7

Chen W, Tang H, Ye J, Lin H, Chou KC. 2016b. iRNA-PseU: identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids* **5:** e332.

Chen W, Tang H, Lin H. 2017a. MethyRNA: a web server for identification of N^6-methyladenosine sites. *J Biomol Struct Dyn* **35:** 683–687. doi:10.1080/07391102.2016.1157761

Chen W, Xing P, Zou Q. 2017b. Detecting N^6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci Rep* **7:** 40242. doi:10.1038/srep40242

Chen W, Yang H, Feng P, Ding H, Lin H. 2017c. iDNA4mC: identifying DNA N^4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **33:** 3518–3523. doi:10.1093/bioinformatics/btx479

Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. 2018. iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. *Mol Ther Nucleic Acids* **11:** 468–474. doi:10.1016/j.omtn.2018.03.012

Church KW. 2017. Emerging trends: Word2Vec. *Nat Lang Eng* **23:** 155–162. doi:10.1017/S1351324916000334

Cui Q, Shi H, Ye P, Li L, Qu Q, Sun G, Sun G, Lu Z, Huang Y, Yang CG, et al. 2017. m^{6}A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep* **18:** 2622–2634. doi:10.1016/j.celrep.2017.02.059

Dai H, Umarov R, Kuwahara H, Li Y, Song L, Gao X. 2017. Sequence2Vec: a novel embedding approach for modeling transcription factor binding affinity landscape. *Bioinformatics* **33:** 3575–3583. doi:10.1093/bioinformatics/btx480

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. 2012. Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature* **485:** 201–206. doi:10.1038/nature11112

Fustin JM, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I, et al. 2013. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* **155:** 793–806. doi:10.1016/j.cell.2013.10.026

Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS, et al. 2015. Stem cells. m^{6}A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347:** 1002–1006. doi:10.1126/science.1261417

Gokhale NS, McIntyre ABR, McFadden MJ, Roder AE, Kennedy EM, Gandara JA, Hopcraft SE, Quicke KM, Vazquez C, Willer J, et al. 2016. N^6-methyladenosine in flaviviridae viral RNA genomes regulates infection. *Cell Host Microbe* **20:** 654–665. doi:10.1016/j.chom.2016.09.015

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8:** R24. doi:10.1186/gb-2007-8-2-r24

Han J, Moraga C. 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. International Workshop on Artificial Neural Networks. In *From natural to artificial neural computation* (ed. Mira J), pp. 195–201. Springer, Berlin, Germany.

He W, Jia C, Duan Y, Zou Q. 2018. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst Biol* **12:** 44. doi:10.1186/s12918-018-0570-1

Hesser C, Karijolich J, Dominissini D, He C, Glaunsinger BA. 2018. *N*^{6}-methyladenosine modification and the YTHDF2 reader protein play cell type specific roles in lytic viral gene expression during Kaposi's sarcoma-associated herpesvirus infection. *PLoS Pathog* **14:** e1006995. doi:10.1371/journal.ppat.1006995

Jia C, He W, Zou Q. 2017. DephosSitePred: a high accuracy predictor for protein dephosphorylation sites. *Comb Chem High Throughput Screen* **20:** 153–157.

Jia C, Zuo Y, Zou Q. 2018. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a *K*-means PCA oversampling technique. *Bioinformatics* **34:** 2029–2036. doi:10.1093/bioinformatics/bty039

Ke S, Alemu EA, Mertens C, Gantman EC, Fak JJ, Mele A, Haripal B, Zucker-Scharff I, Moore MJ, Park CY, et al. 2015. A majority of m^{6}A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev* **29:** 2037–2053. doi:10.1101/gad.269415.115

Li Y, Wang X, Li C, Hu S, Yu J, Song S. 2014. Transcriptome-wide *N*^{6}-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. *RNA Biol* **11:** 1180–1188. doi:10.4161/rna.36281

Li S, Chen J, Liu B. 2017a. Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinformatics* **18:** 443. doi:10.1186/s12859-017-1842-2

Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, Huang H, Nachtergaele S, Dong L, Hu C, et al. 2017b. FTO plays an oncogenic role in acute myeloid leukemia as a *N*^{6}-methyladenosine RNA demethylase. *Cancer Cell* **31:** 127–141. doi:10.1016/j.ccell.2016.11.017

Lichinchi G, Gao S, Saletore Y, Gonzalez GM, Bansal V, Wang Y, Mason CE, Rana TM. 2016a. Dynamics of the human and viral m^{6}A RNA methylomes during HIV-1 infection of T cells. *Nat Microbiol* **1:** 16011. doi:10.1038/nmicrobiol.2016.11

Lichinchi G, Zhao BS, Wu Y, Lu Z, Qin Y, He C, Rana TM. 2016b. Dynamics of human and viral RNA methylation during Zika virus infection. *Cell Host Microbe* **20:** 666–673. doi:10.1016/j.chom.2016.10.002

Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-nucleotide-resolution mapping of m^{6}A and m^{6}Am throughout the transcriptome. *Nat Methods* **12:** 767–772. doi:10.1038/nmeth.3453

Liu B, Li S. 2018. ProtDet-CCH: protein remote homology detection by combining long short-term memory and ranking methods. *IEEE/ACM Trans Comput Biol Bioinform* doi:10.1109/TCBB.2018.2789880

Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* **43:** W65–W71. doi:10.1093/nar/gkv458

Liu B, Wang S, Long R, Chou KC. 2017a. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* **33:** 35–41. doi:10.1093/bioinformatics/btw539

Liu H, Wang H, Wei Z, Zhang S, Hua G, Zhang SW, Zhang L, Gao SJ, Meng J, Chen X, et al. 2017b. MeT-DB V2. 0: elucidating context-specific functions of *N*^{6}-methyl-adenosine methyltranscriptome. *Nucleic Acids Res* **46:** D281–D287. doi:10.1093/nar/gkx1080

Liu Q, Xia F, Yin Q, Jiang R. 2017c. Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics* **34:** 732–738. doi:10.1093/bioinformatics/btx679

Liu Y, Wang X, Liu B. 2018. IDP–CRF: intrinsically disordered protein/region identification based on conditional random fields. *Int J Mol Sci* **19:** E2483. doi:10.3390/ijms19092483

Luo GZ, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, Liu J, Chen K, Jia G, Bergelson J, et al. 2014. Unique features of the m^{6}A methylome in *Arabidopsis thaliana*. *Nat Commun* **5:** 5630. doi:10.1038/ncomms6630

Mairal J, Koniusz P, Harchaoui Z, Schmid C. 2014. Convolutional kernel networks. Advances in Neural Information Processing Systems. *NIPS '14. Proceedings of the 27th International Conference on Neural Information Processing Systems*. Montreal, Canada, December 8–13, 2014. Vol. 2, pp. 2627–2635. MIT Press, Cambridge, MA.

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* **149:** 1635–1646. doi:10.1016/j.cell.2012.05.003

Min X, Zeng W, Chen N, Chen T, Jiang R. 2017. Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics* **33:** i92–i101. doi:10.1093/bioinformatics/btx234

Qiu WR, Sun BQ, Tang H, Huang J, Lin H. 2017. Identify and analysis crotonylation sites in histone by using support vector machines. *Artif Intell Med* **83:** 75–81. doi:10.1016/j.artmed.2017.02.007

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499:** 172–177. doi:10.1038/nature12311

Schwartz S, Agarwala SD, Mumbach MR, Jovanovic M, Mertins P, Shishkin A, Tabach Y, Mikkelsen TS, Satija R, Ruvkun G, et al. 2013. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **155:** 1409–1421. doi:10.1016/j.cell.2013.10.047

Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q. 2014. nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics* **15:** 298. doi:10.1186/1471-2105-15-298

Stephenson N, Shane E, Chase J, Rowland J, Ries D, Justice N, Zhang J, Chan L, Cao R. 2018. Survey of machine learning techniques in drug discovery. *Curr Drug Metab* doi:10.2174/1389200219666180820112457

Su R, Dong L, Li C, Nachtergaele S, Wunderlich M, Qing Y, Deng X, Wang Y, Weng X, Hu C, et al. 2018. R-2HG exhibits anti-tumor activity by targeting FTO/m^{6}A/MYC/CEBPA signaling. *Cell* **172:** 90–105.e23. doi:10.1016/j.cell.2017.11.031

Wan Y, Tang K, Zhang D, Xie S, Zhu X, Wang Z, Lang Z. 2015. Transcriptome-wide high-throughput deep m^{6}A-seq reveals unique differential m^{6}A methylation patterns between three organs in *Arabidopsis thaliana*. *Genome Biol* **16:** 272. doi:10.1186/s13059-015-0839-2

Wang X, Yan R. 2018. RFAthM6A: a new tool for predicting m^{6}A sites in *Arabidopsis thaliana*. *Plant Mol Biol* **96:** 327–337. doi:10.1007/s11103-018-0698-9

Wang CX, Cui GS, Liu X, Xu K, Wang M, Zhang XX, Jiang LY, Li A, Yang Y, Lai WY, et al. 2018. METTL3-mediated m^{6}A modification is required for cerebellar development. *PLoS Biol* **16:** e2004880. doi:10.1371/journal.pbio.2004880

Wei L, Xing P, Tang J, Zou Q. 2017. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans Nanobiosci* **16:** 240–247. doi:10.1109/TNB.2017.2661756

Wei L, Ding Y, Su R, Tang J, Zou Q. 2018a. Prediction of human protein subcellular localization using deep learning. *J Paral Distrib Comput* **117:** 212–217. doi:10.1016/j.jpdc.2017.08.009

Wei L, Luan S, Nagai LAE, Su R, Zou Q. 2018b. Exploring sequence-based features for the improved prediction of DNA *N*^{4}-methylcytosine sites in multiple species. *Bioinformatics* doi:10.1093/bioinformatics/bty824

Wei L, Su R, Wang B, Li X, Zou Q, Gao X. 2018c. Integration of deep feature representations and handcrafted features to improve the

prediction of N^6-methyladenosine sites. *Neurocomputing* **324:** 3–9. doi:10.1016/j.neucom.2018.04.082

Wei L, Xing P, Shi G, Ji ZL, Zou Q. 2018d. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans Comput Biol Bioinform* doi:10.1109/TCBB.2017.2670558

Xiang S, Liu K, Yan Z, Zhang Y, Sun Z. 2016. RNAMethPre: a web server for the prediction and query of mRNA m^6A sites. *PLoS One* **11:** e0162707. doi:10.1371/journal.pone.0162707

Xing P, Su R, Guo F, Wei L. 2017. Identifying N^6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci Rep* **7:** 46757. doi:10.1038/srep46757

Xu Y, Zhou X. 2018. Applications of single-cell sequencing for multiomics. *Methods Mol Biol* **1754:** 327–374. doi:10.1007/978-1-4939-7717-8_19

Xu Y, Wang Y, Luo J, Zhao W, Zhou X. 2017. Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res* **45:** 12100–12112. doi:10.1093/nar/gkx870

Xu Y, Zhao W, Olson SD, Prabhakara KS, Zhou X. 2018. Alternative splicing links histone modifications to stem cell fate decision. *Genome Biol* **19:** 133. doi:10.1186/s13059-018-1512-3

Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, Qu LH, Yang JH. 2017. RMBase v2. 0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res* **46:** D327–D334. doi:10.1093/nar/gkx934

Yue Y, Liu J, He C. 2015. RNA N^6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev* **29:** 1343–1355. doi:10.1101/gad.262766.115

Zeng W, Wu M, Jiang R. 2018. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics* **19:** 84. doi:10.1186/s12864-018-4459-6

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2017. Ensembl 2018. *Nucleic Acids Res* **46:** D754–D761. doi:10.1093/nar/gkx1098

Zhang S, Zhao BS, Zhou A, Lin K, Zheng S, Lu Z, Chen Y, Sulman EP, Xie K, Bögler O, et al. 2017. m^6A demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program. *Cancer Cell* **31:** 591–606.e596. doi:10.1016/j.ccell.2017.02.013

Zhang Z, Zhao Y, Liao X, Shi W, Li K, Zou Q, Peng S. 2018. Deep learning in omics: a survey and guideline. *Brief Funct Genomics* doi:10.1093/bfgp/ely030

Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q. 2016. SRAMP: prediction of mammalian N^6-methyladenosine (m^6A) sites based on sequence-derived features. *Nucleic Acids Res* **44:** e91. doi:10.1093/nar/gkw104

Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA

Quan Zou, Pengwei Xing, Leyi Wei, et al.

RNA 2019 25: 205-218 originally published online November 13, 2018
Access the most recent version at doi:10.1261/rna.069112.118