METHODS
published: 03 May 2022
doi: 10.3389/fgene.2022.896925

i5hmCVec: Identifying 5-Hydroxymethylcytosine Sites of *Drosophila* RNA Using Sequence Feature Embeddings

*Hang-Yu Liu and Pu-Feng Du**

College of Intelligence and Computing, Tianjin University, Tianjin, China

5-Hydroxymethylcytosine (5hmC), one of the most important RNA modifications, plays an important role in many biological processes. Accurately identifying RNA modification sites helps understand the function of RNA modification. In this work, we propose a computational method for identifying 5hmC-modified regions using machine learning algorithms. We applied a sequence feature embedding method based on the dna2vec algorithm to represent the RNA sequence. The results showed that the performance of our model is better that of than state-of-art methods. All dataset and source codes used in this study are available at: https://github.com/liu-h-y/5hmC_model.

Keywords: 5-hydroxymethylcytosine, dna2vec, machine learning, cross-validation, i5hmcVec

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Wang-Ren Qiu,
Jingdezhen Ceramic Institute, China
Yan Xu,
University of Science and Technology Beijing, China

****Correspondence:***
Pu-Feng Du
pdu@tju.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 March 2022*
Accepted: *30 March 2022*
Published: *03 May 2022*

Citation:
Liu H-Y and Du P-F (2022) i5hmCVec: Identifying 5-Hydroxymethylcytosine Sites of Drosophila RNA Using Sequence Feature Embeddings. Front. Genet. 13:896925. doi: 10.3389/fgene.2022.896925

INTRODUCTION

Posttranscriptional modifications have been extensively studied over the last few years. More than 160 types of modification have been identified across all kingdoms of life (Boccaletto et al., 2018). Posttranscriptional modifications play important roles in various biological processes, such as RNA degradation (Sommer et al., 1978), RNA splicing (Lindstrom et al., 2003), and transcriptional regulations (Cowling, 2009). To understand the mechanism of RNA modifications, it is important to pinpoint the modification sites in the RNA sequences (Dominissini et al., 2012; Meyer et al., 2012).

With the rapid development of high-throughput technology, several experimental methods for identifying RNA modification sites have been developed, such as MERIP (Meyer et al., 2012) and m6A-seq (Dominissini et al., 2012). These methods are more capable of picking up the modified transcripts or regions on the transcripts, rather than accurately pinpointing the modification sites. With the advances in modern life sciences, especially the cross-linking technology, methods for identifying RNA modification sites at single-base resolution were also proposed, including miCLIP (Linder et al., 2015), PA-m6A-seq (Kai Chen et al., 2015), and m7G-MeRIP-seq (Zhang et al., 2019). However, these experimental methods are still costly and time-consuming. Therefore, computational methods have been proposed as alternative approaches. A series of bioinformatics tools using machine learning algorithms for predicting m6A (Wei Chen et al., 2015; Zhou et al., 2016; Huang et al., 2018; Kunqi Chen et al., 2019; Zou et al., 2019), m5C (Qiu et al., 2017; Sabooh et al., 2018; Akbar et al., 2020; Dou et al., 2020), m7G (Wei Chen et al., 2019, 7; Liu X. et al., 2020; Yang et al., 2020; Dai et al., 2021), and many others have been developed. A recent review article has elaborated on the differences between these studies, in the aspect of benchmarking datasets, feature encoding schemes, and the main algorithms (Chen et al., 2020).

5-Hydroxymethylcytosine (5hmC) plays a key role in various cellular processes. 5hmC modification exists on both RNA and DNA sequences (Zhang et al., 2016). Most of the existing studies focused on the DNA 5hmC modifications (Szwagierczak et al., 2010; Pastor et al., 2011; Yu et al., 2012; Bachman et al., 2014). The RNA 5hmC modifications were much less studied (Fu et al., 2014; Huber et al., 2015; Delatte et al., 2016; Miao et al., 2016). Fu et al. first found that the m5C site can be catalyzed by the Tet enzyme to form 5hmC sites with a ratio of about 0.02% *in vitro* in mammalian RNA (Fu et al., 2014). In addition, a discovery that Tet-mediated oxidation of m5C in RNA is much less efficient than that in DNA (Fu et al., 2014). Huber et al. verified that 5hmC is the result of m5C oxidation *in vivo* in a mouse model using an isotope-tracing methodology (Huber et al., 2015). They also found that in worms and plants, the formation of 5hmC in RNA does not require a Tet-mediated oxidation mechanism. Miao et al. (2016) found that 5hmC in RNA is rich in the mouse brain, which is potentially related to brain functions. Delatte et al. (2016) systematically identified 5hmC modifications in *Drosophila* transcriptome using the hMeRIP-seq method. Using the data from Delatte et al., Liu et al. developed a predictor iRNA5hmC for computationally identifying 5hmC modifications with machine learning algorithms (Liu Y. et al., 2020). Ahmed et al. also constructed a predictor iRNA5hmC-PS (Ahmed et al., 2020) by using position-specific binary indicators of RNA sequences. However, Delatte et al. did not provide the exact location of 5hmC modification sites in the transcriptome (Delatte et al., 2016). Liu et al. provided the exact location by randomly selecting cytosine sites within the peak region detecting by MeRIP-seq (Liu Y. et al., 2020). However, such a strategy may lead to many false-positive samples (Kunqi Chen et al., 2019). To avoid such uncertainty, we proposed a model based on low-resolution data.

The rapid development of deep learning has promoted natural language processing studies. Word2vec is a remarkable achievement in natural language processing technology (Mikolov et al., 2013). Distributed representation of word vector is the core idea of word2vec, which means the representation of a word can be inferred from its context. With the development of high-throughput sequencing technology, the sequencing quality of biological sequences can be guaranteed. Therefore, some researchers in bioinformatics regard the biological sequences as a sentence, and k-mers as words. The word2vec method can then be applied to represent the biological sequences. Asgari et al. proposed BioVec based on the skip-gram model for biological sequences representation (Asgari and Mofrad, 2015). Kimothi et al. developed a model named seq2vec based on doc2vec, which is an extension of the original word2vec (Kimothi et al., 2016). The dna2vec model is dedicated to representing variable-length words (Ng, 2017a). It has been applied to several topics in bioinformatics. For example, Deng et al. proposed D2VCB for predicting protein–DNA-binding sites based on k-mer embeddings (Deng et al., 2019). Hong et al. applied the pretrained k-mer embeddings to encode enhancers and promoters (Hong et al., 2020). We employed the dna2vec embeddings to represent k-mers of *Drosophila* genomic sequences.

In this study, we represent the RNA sequences by using feature embeddings. We applied an SVM classifier to create a model for predicting 5hmC modification sites. Our model was trained on the low-resolution modification datasets, which is more reliable than the 1-base resolution set. The result suggests that our model is effective in identifying 5hmC sites.

MATERIALS AND METHODS

Datasets

In this study, we constructed the benchmarking dataset according to the experimental result from Delatte et al. (2016). The result from Delatte et al. contains 3058 peak regions distributed on chromosomes, which contain chr2L, chr2R, chr3L, chr3R, chr4, chrX, chr2RHet, chr3LHet, chr3RHet, chrYHet, chrU, and chrUextra. According to Hoskins et al. (2015), the genome sequences are of high quality on chr2L, chr2R, chr3L, chr3R, chr4, and chrX, while the remaining chromosome sequences are of low quality. Therefore, we only used the sequence data from chr2L, chr2R, chr3L, chr3R, chr4, and chrX. We got 2616 peak regions containing 5hmC modification sites. Subsequently, we obtained the transcription direction of every region by querying the UCSC genome browser tracks (Karolchik et al., 2003). Finally, 2616 positive samples were curated, which are regions containing 5hmC modification sites. Non-peak regions within transcripts carrying peak regions are curated as negative samples. The non-peak regions were cropped to the same lengths as the peak regions in a one-vs.-one strategy. A total of 2616 positive samples and 2616 negative samples were finally curated. We plot the density distribution of sequence lengths in **Figure 1**.

K-Mer Embeddings

K-mer is a common and efficient way to represent RNA sequences, which divided the biological sequences into short segments of the length *k*. We employed the *k*-mer embeddings for representing the *k*-mer instead of one-hot encoding. *K*-mer embeddings can capture semantic and linguistic analogies and avoid the curse of dimensionality (Mikolov et al., 2013). The dna2vec model was used in this study for training *k*-mer embeddings (Ng, 2017a, 2). The corpus was collected from dm3 (Karolchik et al., 2003) genome assembly. We selected high-quality six chromosome sequences from dm3, including ch2L, chr2R, chr3L, chr3R, chr4, and chrX. The corpus was used as the input of the dna2vec. *K*-mer embeddings were obtained by training dna2vec. Let *p* (*k*, *i*) (*i* = 1, 2,. . .4^k) represent the *i*-th type *k*-mer fragment. The process of the dna2vec model can be expressed as follows:

$$p(k,i)\xrightarrow{h(\cdot)}\mathbf{v}\left(p(k,i)\right), \tag{1}$$

where *h*(.) is the mapping from a *k*-mer fragment to *k*-mer embedding and **v**(*p* (*k*, *i*)) is the embedding vector of the *i*-th type of *k*-mer. In this study, we chose *k* from 3 to 8. The dimension of **v**(*p* (*k*, *i*)) was set to 100.

Distribution Representation of RNA Sequences

Given an RNA sequence *r* with length *l*, it can be represented as follows:

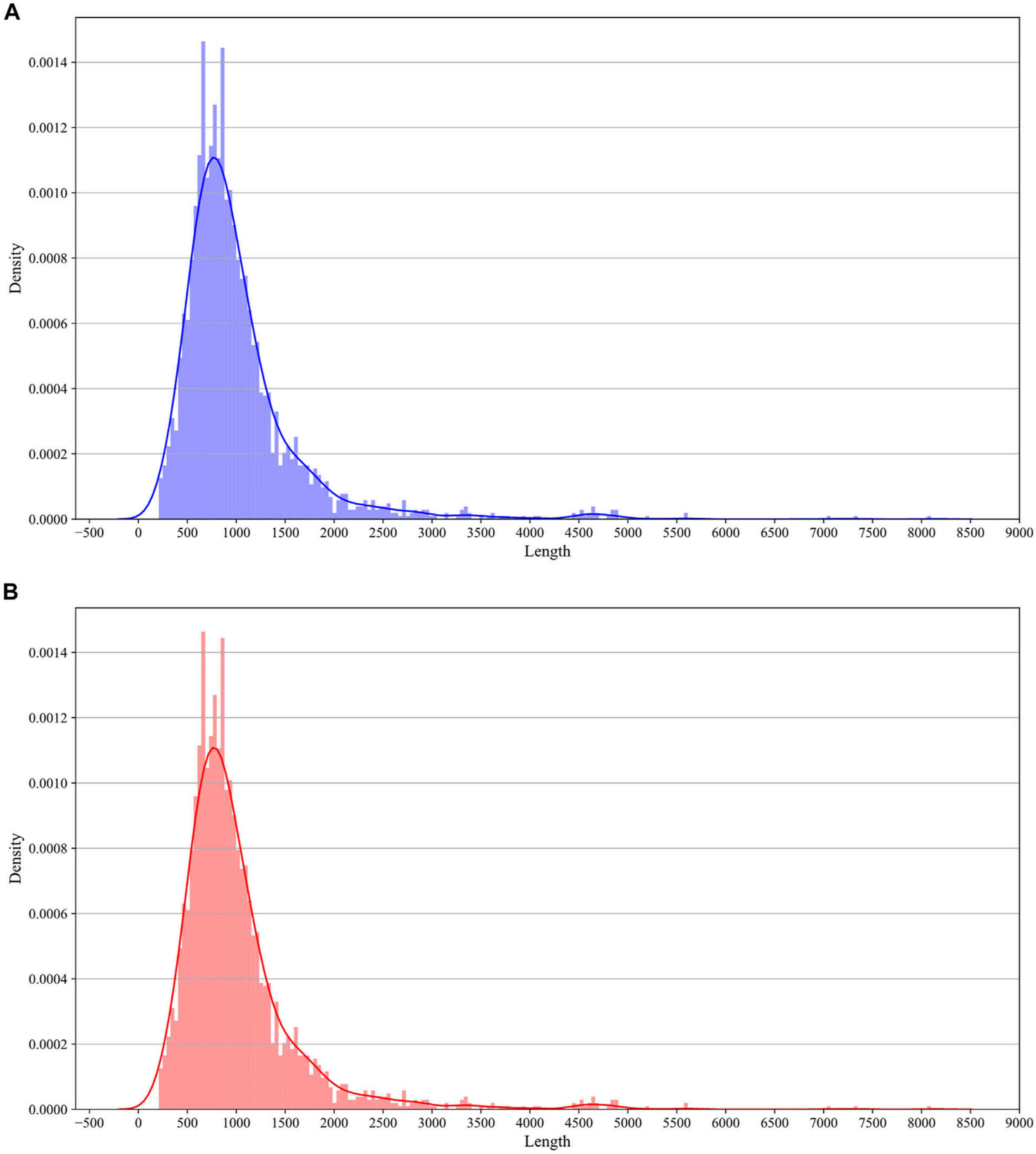

FIGURE 1 | Sequence length distribution. The X-axis represents the length of sequences. The Y-axis represents the density of distribution. **(A)** Histogram density for the distribution of the length of positive sequences. **(B)** Histogram density for the distribution of the length of positive sequences.

$$r = n_1 n_2 \cdots n_l, \tag{2}$$

where n_u (u = 1, 2,. . ., l) represents u-th nucleotide in RNA sequence. The RNA sequences are segmented into k-mers in an overlapping way. For example, we convert AUAGC into three 3-mers: “AUA,” “UAG,” “AGC.” Therefore, sequence r divided by k can be represented as follows:

$$r = \{w_1, w_2, \ldots, w_{l-k+1}\}, \tag{3}$$

where w_j (j = 1, 2,. . ., l–k+1) ∈ $\{p(k, i)\,|k = 3, 4,\ldots, 8, i = 1, 2,\ldots, 4^k\}$. The fragment of k-mer RNA sequence can be considered as an RNA word. With the mapping h(.) from dna2vec, w_i was converted into the corresponding embedding vector. Sequence r can be expressed in a matrix as follows:

$$\mathbf{E}(r,k) = \left[\mathbf{v}(w_1) \quad \mathbf{v}(w_2) \quad \dots \quad \mathbf{v}(w_{l-k+1})\right], \tag{4}$$

Since dna2vec was trained by a corpus of DNA sequences, the *k*-mers from dna2vec do not contain uracil. We replaced thymine with uracil on *k*-mers for using the mapping. Considering the sum of dna2vec embeddings along the sequence is related to concatenating *k*-mers (Ng, 2017b), we sum the embedding vector in *E(r, k)* for representing the sequence *r*, as follows:

$$\mathbf{e}(r,k) = \sum_{i=1}^{l-k+1} \mathbf{v}(w_i) \Big/ l - k + 1. \tag{5}$$

In this study, we chose $k = 3, 4, 5, 6, 7$, and 8. The final feature vector is formed by concatenating $\mathbf{e}(r, k)$ with different k, as follows:

$$\mathbf{e}(r) = \left[\mathbf{e}(r,3)^{\mathrm{T}} \quad \mathbf{e}(r,4)^{\mathrm{T}} \quad \dots \quad \mathbf{e}(r,8)^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{6}$$

Model Construction Algorithm

We evaluated three machine learning algorithms in this task, including SVM, CNN, and C4.5 classification tree. For the SVM classifier, we applied the radial basis function (RBF) kernel, as follows:

$$\kappa\left(\mathbf{e}_i, \mathbf{e}_j\right) = \exp\left(-\gamma\left\|\mathbf{e}_i - \mathbf{e}_j\right\|^2\right), \tag{7}$$

where γ is a parameter and ||.|| vector norm operator.

For the CNN classifier, the max-pooling layer and dropout layer are used to avoid the over-fitting problem. The sigmoid function followed by a fully connected network is applied for performing the output. We used stochastic gradient descent to optimize parameters (Bottou, 2012). The binary cross-entropy function is used as the loss function (de Boer et al., 2005), as follows:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} y_i log\left(h_\theta(\mathbf{e})\right) - \left(1 - y_i\right) log\left(1 - h_\theta(\mathbf{e})\right), \tag{8}$$

where y_i is the label of the i-th sample, $h_\theta(\mathbf{e})$ the output of the neural network, and N the number of samples.

For C4.5 algorithm, the information gain ratio for selecting appropriate features is defined as follows:

$$G_r(D, e_i) = \frac{G(D, e_i)}{IV(e_i)}, \tag{9}$$

where D is the whole dataset, G_r (D, e_i) the information gain, $IV(e_i)$ the intrinsic value of e_i (Salzberg, 1994), and e_i the i-th feature of feature $\mathbf{e}$.

Degree of Separation

To measure the degree of separation in the visualization analysis, we introduced the J-score. We first define the intra-class divergence sw and interclass divergence sb, as follows:

$$s_b = \left(\bar{\mathbf{e}}_+ - \bar{\mathbf{e}}_-\right)\left(\bar{\mathbf{e}}_+ - \bar{\mathbf{e}}_-\right)^T, \tag{10}$$

$$s_w = \sum_{j=1}^{m_+}\left(\mathrm{e}_+\left(r_j\right) - \bar{\mathrm{e}}_+\right)\left(\mathrm{e}_+\left(r_j\right) - \bar{\mathrm{e}}_+\right)^T + \sum_{j=1}^{m_-}\left(\mathrm{e}_-\left(r_j\right) - \bar{\mathrm{e}}_-\right)\left(\mathrm{e}_-\left(r_j\right) - \bar{\mathrm{e}}_-\right)^T, \tag{11}$$

where

$$\bar{\mathrm{e}}_+ = \frac{1}{m_+}\sum_{j=1}^{m_+}\mathrm{e}_+\left(r_j\right), \tag{12}$$

$$\bar{\mathrm{e}}_- = \frac{1}{m_-}\sum_{j=1}^{m_-}\mathrm{e}_-\left(r_j\right), \tag{13}$$

where $e_+(r_j)$ is the feature vector of the j-th positive sample, $e_-(r_j)$ is the feature vector of the j-th negative sample, and m+ and m- are the number of positive and negative samples, respectively.

The J-score can now be defined as follows:

$$J = \frac{s_b}{s_w}. \tag{14}$$

The higher J-score indicates a better degree of separation between positives and negatives.

Framework of This Study

The framework of i5hmcVec is illustrated in **Figure 2**. We obtained the *k*-mer embeddings using dna2vec (Ng, 2017a), which is trained by the *Drosophila* genome sequences version dm3. RNA sequences were encoded by the embedding vectors for variable-length *k*-mers. SVM was applied as a classifier to distinguish the positive and negative samples.

Parameter Calibration

In this section, we give a detailed introduction to optimizing parameters. SVM was implemented by the Python package scikit-learn. We chose to use the radial basis function (RBF) as the kernel function. A grid search strategy was applied to find the optimal parameters *c* and γ. The parameter *c* is the cost parameter in SVM, while γ is the parameter in the RBF kernel function. The range of parameter c is (2^{-5}, 2^{15}), while the range for parameter γ is (2^{-15}, 2^{-5}). The step for generating the logarithm searching grid is 2 and 2-1 for c and γ, respectively. The CNN algorithm is implemented by Keras. The batch size was set to 16. A logarithm grid search strategy was used to find the optimal parameters epoch e and learning rate a. The range of parameter a: 10^{-4}, 5 × 10^{-4}, 10^{-3}, 5 × 10^{-3}, 10^{-2}, and 5 × 10^{-2}. The range of parameters e is 100, 150, 200, 250, and 300. We used the weka package to implement C4.5. We evaluated the performance on different parameters C, which is the confidence threshold for pruning. The range of C is [0.2, 0.5] with a step of 0.05.

Performance Measures

Four statistics, including sensitivity (Sen), specificity (Spe), accuracy (Acc), and Matthews correlation coefficient (MCC), were used to measure the prediction performance of our method. These performance measures can be defined as follows:

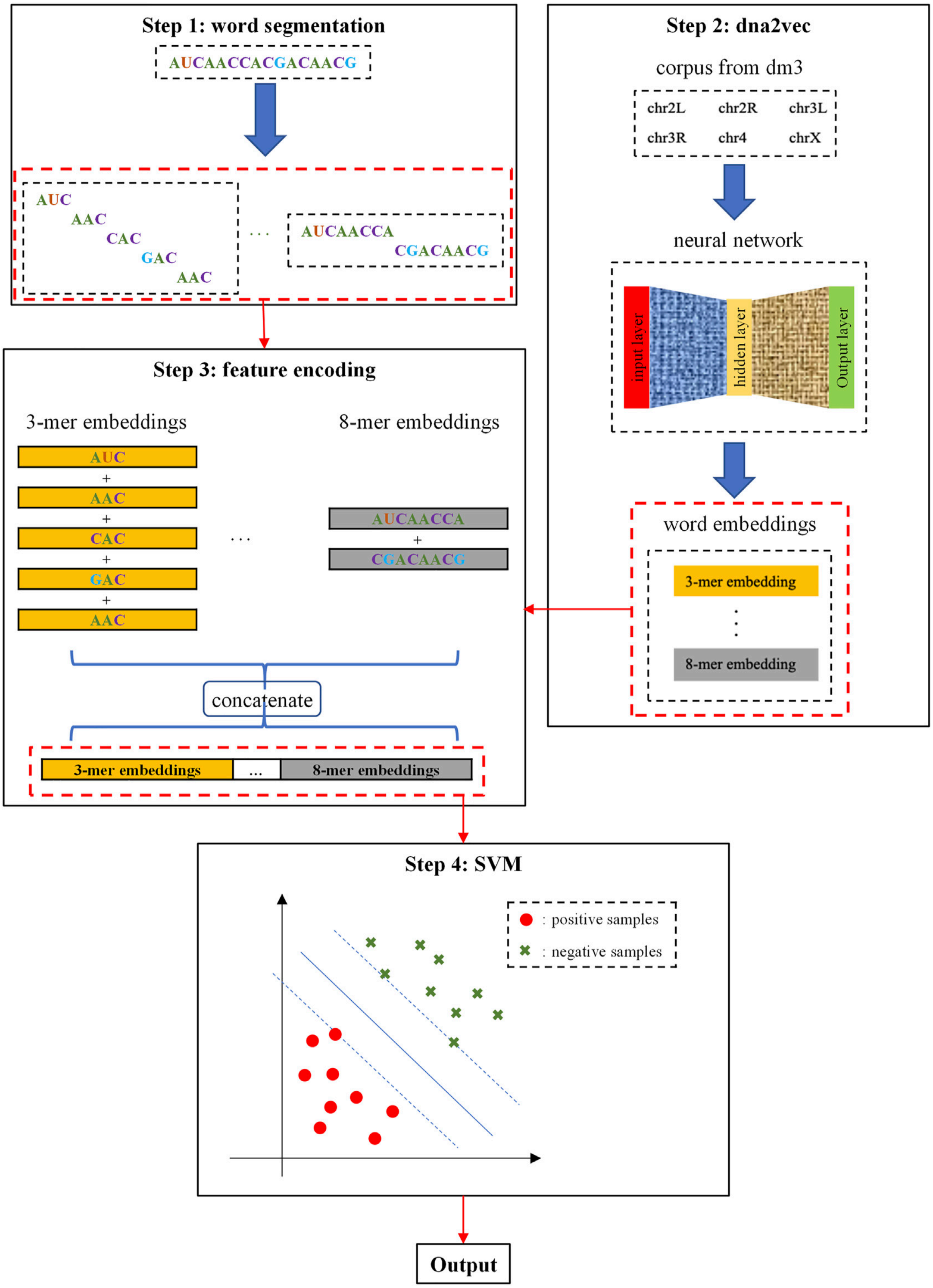

FIGURE 2 | Flowchart of this study. Step 1: RNA sequences are segmented into *k*-mers in the overlapping way, where *k* = 3, 4, 5, 6, 7, 8. Step 2: *k*-mers embeddings were trained by the dna2vec model with corpus from dm3. Step 3: We perform summation and concatenation on these *k*-mers embeddings to encode RNA sequences. Step 4: SVM is used as a classifier for distinguishing positive and negative samples.

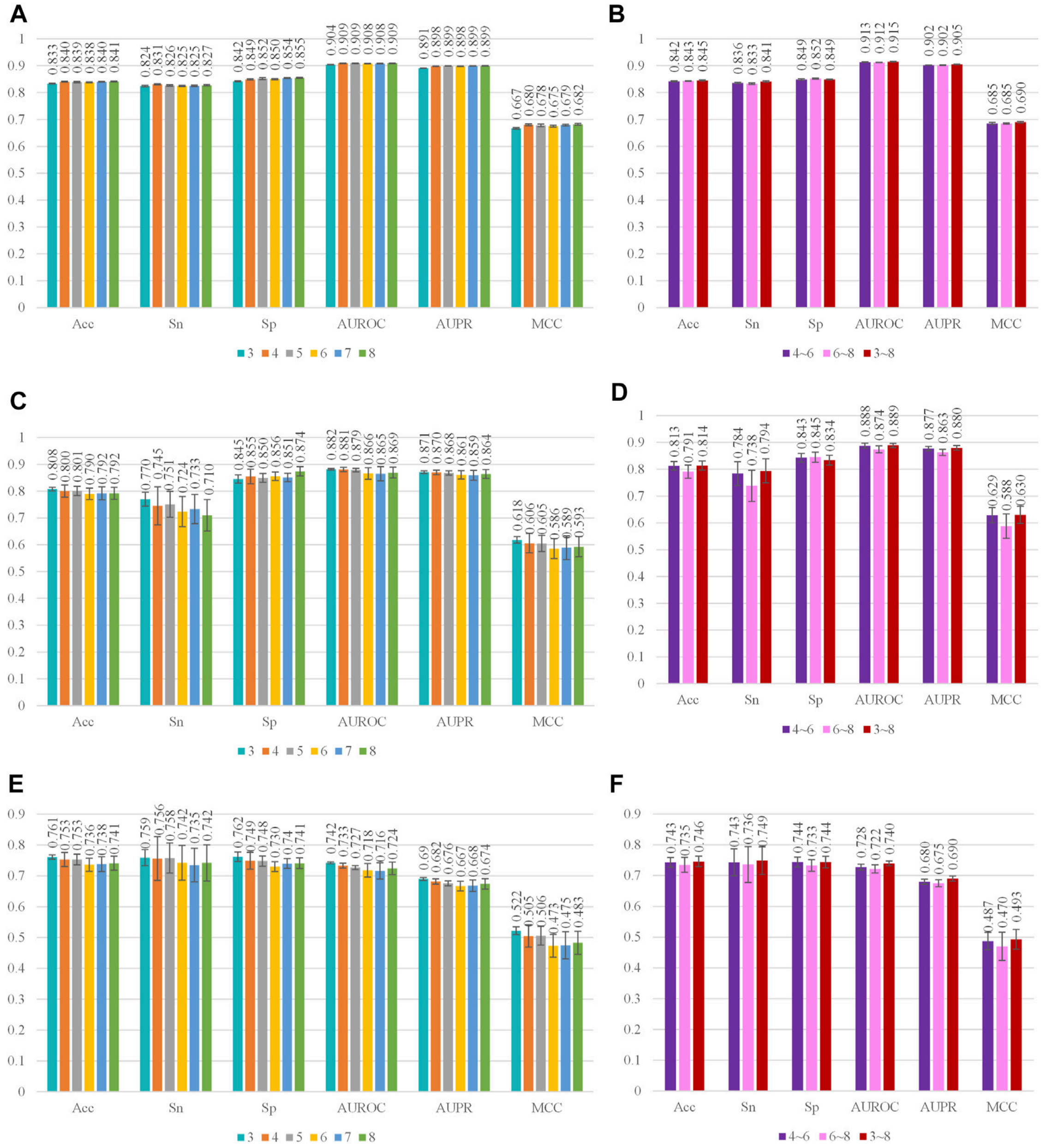

FIGURE 3 | Performance of different kinds features on SVM, CNN, and C4.5. Cyan, orange, gray, yellow, blue, and green, respectively, represent the performance of 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, and 8-mer embedding features. Purple, pink, and red, respectively, represent the performance of 4, 5, 6-mer concatenated embeddings, 6, 7, 8-mer concatenated embeddings, and 3, 4, 5, 6, 7, 8-mer concatenated embeddings. **(A,B)** Performance of different kinds of feature on SVM. The standard deviation of SVM on 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 4, 5, 6-mer, 6, 7, 8-mer, and 3, 4, 5, 6, 7, 8-mer is in the range (0.001, 0.003), (0.001, 0.003), (0.001, 0.003), (0.001, 0.003), (0.001, 0.003), (0.001, 0.004), (0.001, 0.004), (0.001, 0.003), and (0.001, 0.003); **(C,D)** Performance of different kinds of feature on CNN. The standard deviation of CNN on 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 4, 5, 6-mer, 6, 7, 8-mer, and 3, 4, 5, 6, 7, 8-mer is in the range (0.003, 0.026), (0.008, 0.071), (0.006, 0.049), (0.015, 0.056), (0.016, 0.055), (0.016, 0.059), (0.008, 0.044), (0.011, 0.058), and (0.008, 0.045); **(E,F)** Performance of different kinds of features on C4.5. The standard deviation of CNN on 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 4, 5, 6-mer, 6, 7, 8-mer, and 3, 4, 5, 6, 7, 8-mer is in the range (0.005, 0.545), (0.007, 0.531), (0.005, 0.049), (0.007, 0.685), (0.003, 0.440), (0.007, 0.489), (0.008, 0.630), (0.006, 0.567), and (0.005, 0.518).

$$Sen = \frac{TP}{TP + FN}, \tag{15}$$

$$Spe = \frac{TN}{TN + FP}, \tag{16}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \text{ and} \tag{17}$$

$$MCC = \frac{TPTN - FPFN}{\sqrt{(TP + FN)(TN + FN)(TP + FP)(TN + FP)}}, \tag{18}$$

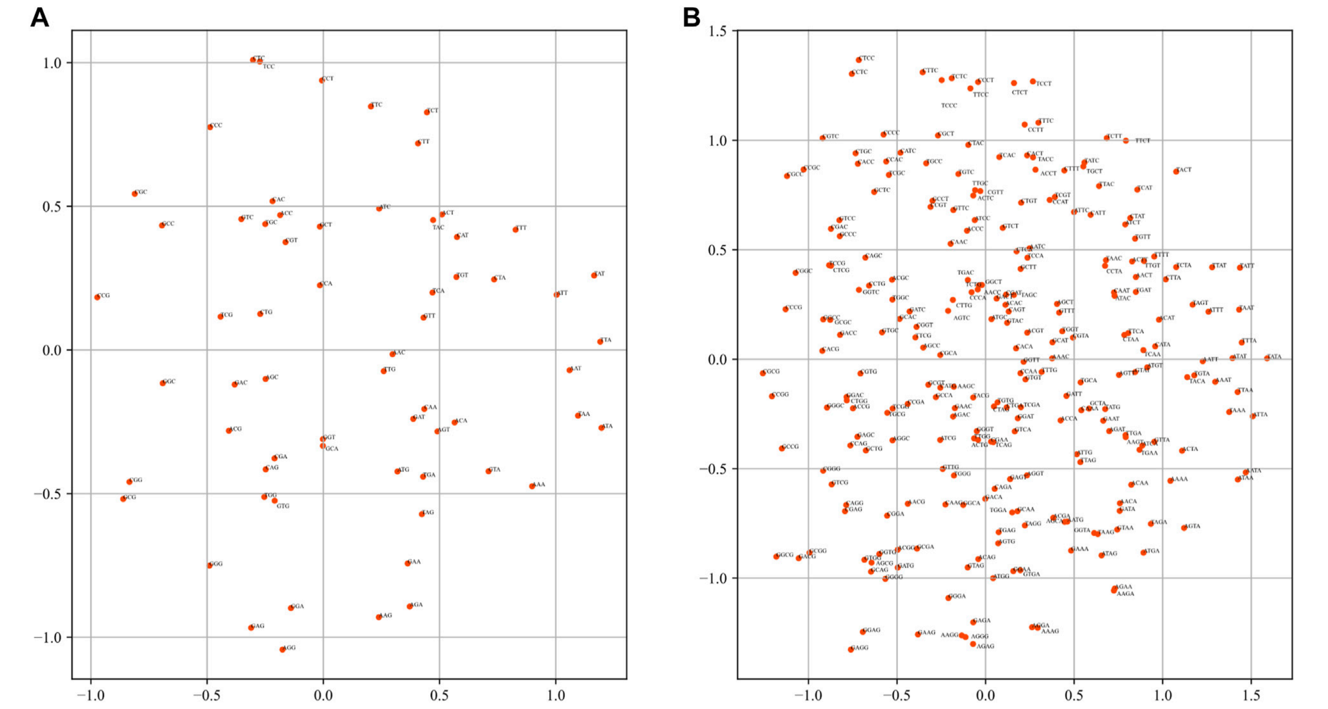

FIGURE 4 | Visualization of *k*-mer embeddings with PCA. Each dot represents a *k*-mer embedding vector. **(A)** 3-mer embedding. **(B)** 4-mer embedding.

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives in the cross-validation process, respectively.

In addition, we also draw the receiver operating characteristic (ROC) curve and precision–recall (PR) curve to describe the performance of our method. The area under the ROC curve (AUROC) and the area under the PR (AUPR) curve were also recorded as performance indicators.

RESULTS

Performance of Diffident Kind Features and Classifiers

In this study, nine kinds of *k*-mer embeddings were obtained, including six kinds of single *k* value embeddings and 3 kinds of multiple *k* value combinations. The single *k* values range from 3 to 8. The multiple *k* value combinations include the 4, 5, 6-mer combination, 6, 7, 8-mer combination, and 3, 4, 5, 6, 7, 8-mer combination. We first evaluate the performance of each single *k* value embedding. After that, we evaluate three multiple *k* value combinations.

Three machine learning-based classifiers were applied in this study. They are SVM, CNN, and C4.5. The parameters of these classifiers are optimized as in the method section. The optimization process is recorded as mesh surf plots in **Supplementary Figures S1–S3** in the supplementary materials. The data for quantitative analysis is recorded in **Supplementary Tables S1–S27**. The optimal parameters for different classifiers are: the c and γ of SVM on the 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 4, 5, 6-mer, 6, 7, 8-mer and 3, 4, 5, 6, 7, 8-mer are (29, 2–5), (27, 2–5), (27, 2–5), (27, 2–5), (27, 2–5), (27, 2–5), (25, 2–5), (25, 2–5), and (24, 2–5); the a and e of CNN on the 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 4, 5, 6-mer, 6, 7, 8-mer and 3, 4, 5, 6, 7, 8-mer are (5×10^{-2}, 200), (5×10^{-2}, 150), (5×10^{-2}, 150), (5×10^{-2}, 250), (5×10^{-2}, 150), (5×10^{-2}, 250), (5×10^{-2}, 150), (5×10^{-2}, 100), and (5×10^{-2}, 150); the C of C4.5 on the 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 4, 5, 6-mer, 6, 7, 8-mer and 3, 4, 5, 6, 7, 8-mer are 0.45, 0.3, 0.2, 0.5, 0.2, 0.45, 0.25, 0.3, and 0.3. The performances of all models are evaluated by 10 times 5-fold cross-validations. The optimal performance is recorded in **Figure 3** and **Supplementary Tables S28–S54**.

Semantic Symmetry of *K*-Mer Embeddings

One of the most important functions of word2vec is that the word embeddings can solve semantic and linguistic analogies (Mikolov et al., 2013). Therefore, the semantic relation of the *k*-mer embeddings from dna2vec needs to be discussed. Principal component analysis (PCA) was applied to reveal the relationship of *k*-mer fragments. For 5-mer embeddings, the number of words is 1024. To present the results clearly, we only plot the PCA results of 3-mer and 4-mer embeddings in **Figure 4**. As in **Figure 4**, many words show symmetry trends about the horizontal axis, such as (CGC, GCG), (CTT, AAG), and (TACT, AGTA). Many words with such property have the characteristics of complement or reverse complement. Zou et al. regarded this phenomenon as semantic symmetric in the human genome (Zou et al., 2019). We observe and confirm this phenomenon in *Drosophila* genome.

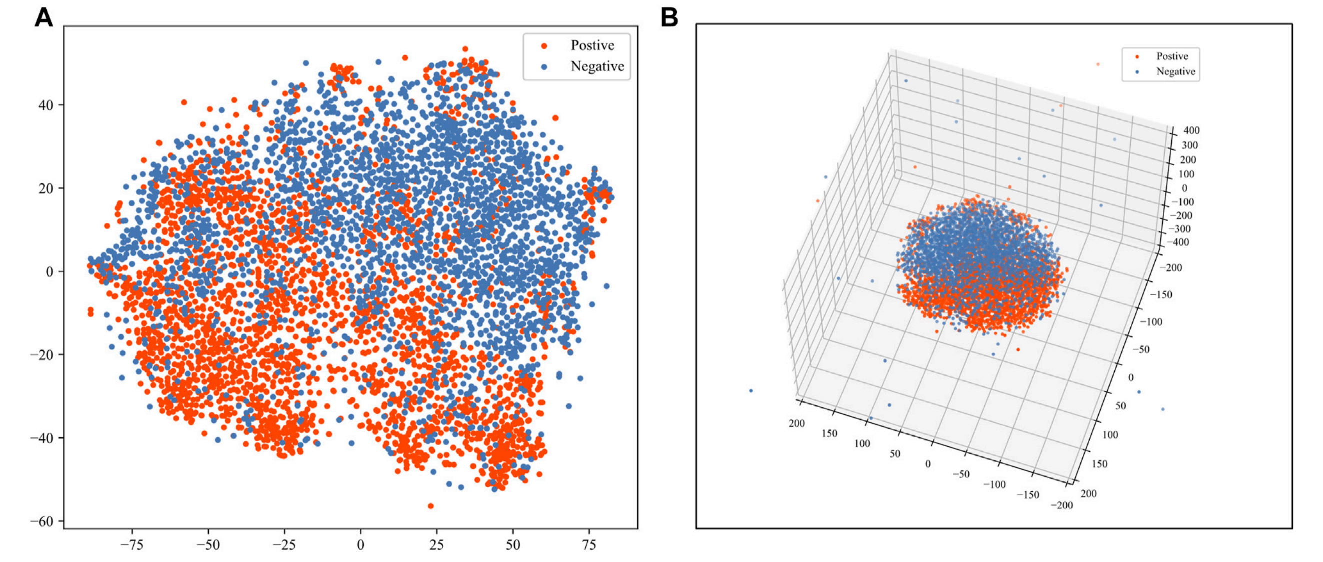

FIGURE 5 | Visualization of sequence features. The red dots represent the positive samples. The blue dots represent the negative samples. **(A)** Visualization of 2-dimensional t-SNE. **(B)** Visualization of 3-dimensional t-SNE.

TABLE 1 | Dataset distributions of i5hmcVec and WeakRM.

Method	Positive[a]	Negative[b]	Window size
i5hmCVec	2616	2616	209 nt~8097 nt
WeakRM (training)	1875	1875	210 nt~8090 nt
WeakRM (validation)	235	235	210 nt~8090 nt
WeakRM (testing)	234	234	210 nt~8090 nt

[a]Positive samples are sequences, which contain the 5hmC sites.
[b]Negative samples are sequences, which do not contain the 5hmC sites.

TABLE 2 | Performance of i5hmcVec and WeakRM on the dataset from WeakRM.

Method	Acc[a]	Sen[b]	Spe[c]	AUROR[d]	AUPR[e]	MCC[f]
WeakRM	0.790	0.617	**0.967**	0.892	0.905	0.619
i5hmCVec	**0.846**[g]	**0.838**	0.855	**0.920**	**0.908**	**0.692**

[a]Acc is short for accuracy.
[b]Sen is short for sensitivity.
[c]Spe is short for specificity.
[d]AUROC means the area under the ROC curve.
[e]AUPR means the area under the PR curve.
[f]MCC is short for Matthews correlation coefficient.
[g]Boldface indicates the best performance on each metric among methods.

Feature Visualization

We used the t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) method to help visualize the sequence features. The t-SNE algorithm is an effective way of reducing dimensions for visualization purposes. According to the visualization of t-SNE, we can judge whether the positive and negative samples are separable in the feature space. We applied the t-SNE for reducing the dimension of the feature to 2 and 3. We also calculated the J-score, which has been elaborated in the method section, as a quantitative separation measure in the reduced feature space. As shown in **Figure 5**, positive and negative samples are highly separable. The J-score of 2 and 3 dimensions of t-SNE are 0.202 and 0.165, indicating an acceptable level of separation.

Performance Comparison With Existing Methods

The i5hmCVec is constructed based on a low-resolution modification dataset. WeakRM (Huang et al., 2021) was also proposed for identifying the 5hmC modification sites on low-resolution data. We summarized the dataset distribution used in the i5hmCVec and WeakRM in **Table 1**.

We used the dataset from WeakRM for training the i5hmCVec model. We also reproduced WeakRM for obtaining more types of performance metrics. Due to inevitable randomness errors, our reproduced performances are slightly different from the original reports. The differences are so tiny that the comparison results would not change. As in **Table 2**, i5hmCVec achieved 0.846, 0.920, 0.908, and 0.692 on Acc, AUROC, AUPR, and MCC, respectively, which are higher than the performance values of WeakRM. In addition, we make a comparison of training time between i5hmcVec and WeakRM. Training WeakRM takes about 500 s, while i5hmCVec takes about 25 s. To describe the results more intuitively, we displayed the ROC curve and PR curve of two models, as in **Figure 6**. As in **Figure 6**, both the AUROC and AUPR of i5hmCVec are slightly better than the WeakRM. In total, iRNA5hmCVec achieved better performances than WeakRM on a low-resolution modification dataset.

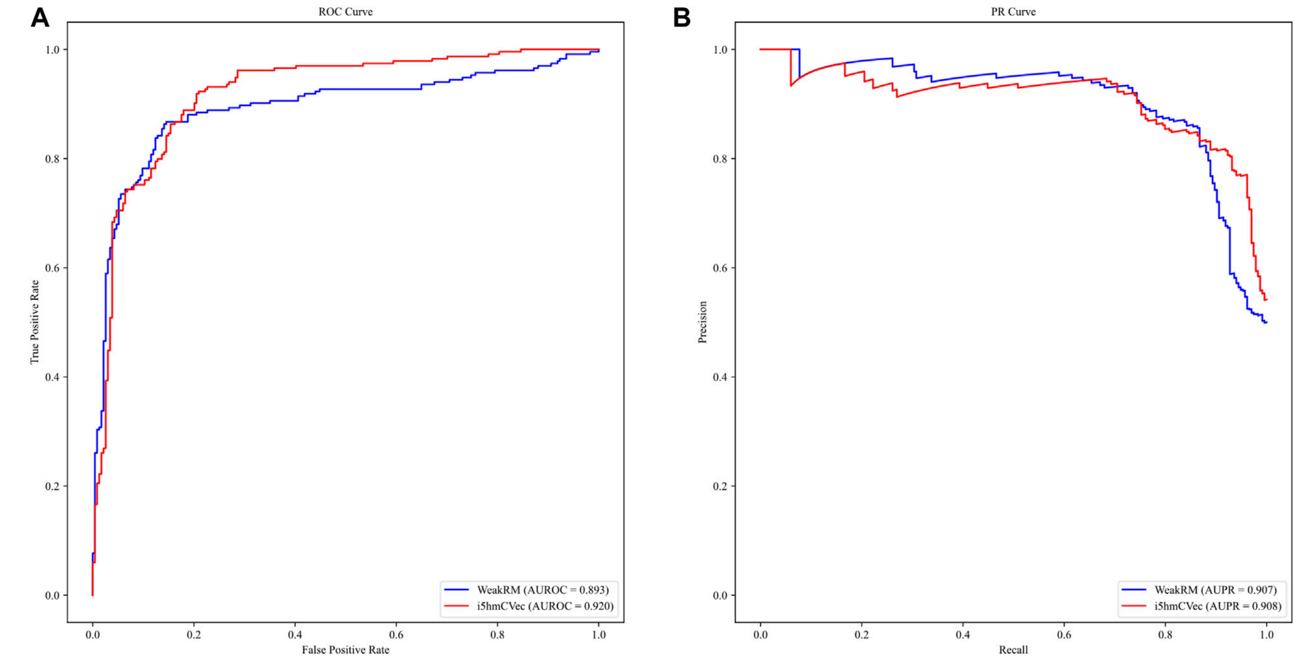

FIGURE 6 | ROC and PR curves of i5hmcVec and WeakRM on the dataset from WeakRM. **(A)** ROC curve. The X-axis is the false positive rate, and the Y-axis is the true positive rate. **(B)** PR curve. The X-axis is the recall, and the Y-axis is the precision.

DISCUSSION

Identifying modification sites is an important work for studying 5hmC modification. In this study, we used machine learning methods to construct the model. There are three key steps for a machine learning problem.

First, a high-quality dataset is essential for building an effective model. We constructed the low-resolution benchmarking dataset from experimental results (Delatte et al., 2016). We did not use the strategy of randomly selecting cytosine sites within peak regions like Liu Y. et al. (2020). Because such a strategy may lead to many false-positive samples (Kunqi Chen et al., 2019). In addition, to ensure high quality of sequences, we only employed the high-quality chromosomes sequences in the genome assembly.

Second, the samples from the dataset should be represented by an informative digital vector. We encode RNA sequences using the k-mer embeddings, which are derived from dna2vec. According to our results, the feature vector can effectively separate positive and negative samples. These results suggest that this encoding scheme is suitable for our study.

Finally, a suitable classifier should be used for constructing the model. We compared the performance of SVM, C4.5, and CNN. The SVM classifier has the best performance. In addition, we optimize the parameters using a grid search strategy.

Although our model was trained on low-resolution data, we tried to evaluate the performance of our model on high-resolution data. We performed 10 times 5-fold cross-validations on the benchmarking dataset from iRNA5hmC (Liu Y. et al., 2020). The sequence data in iRNA5hmC are 41 nt. The results are recorded in **Table 3**. According to the results, the i5hmCVec does not receive expected performance on a high-resolution modification dataset. We speculated that there may be two reasons for this phenomenon. One is the low quality of the high-resolution dataset. The high-resolution dataset of 5hmC modification was developed by Liu et al. with a random site picking strategy (Liu Y. et al., 2020), which may lead to many false positives.

TABLE 3 | Performance of i5hmcVec and iRNA5hmC on the benchmark dataset from iRNA5hmC.

Method	Acc[a]	Sen[b]	Spe[c]	AUROC[d]	AUPR[e]	MCC[f]
iRNA5hmC	**0.655**[g]	**0.677**	0.644	**0.697**	**0.685**	**0.310**
i5hmcVec[h]	0.642 ±0.008	0.636 ±0.010	**0.647** **±0.009**	0.684 ±0.007	0.676 ±0.007	0.284 ±0.016

[a]*Acc is short for accuracy.*
[b]*Sen is short for sensitivity.*
[c]*Spe is short for specificity.*
[d]*AUROC means the area under the ROC curve.*
[e]*AUPR means the area under the PR curve.*
[f]*MCC is short for Matthews correlation coefficient.*
[g]*Boldface indicates the best performance on each metric among different methods.*
[h]*Performance of i5hmcVec on the benchmark dataset from iRNA5hmC with 10 times 5-fold cross-validation. Results are expressed as the mean and standard deviation of 10 times experiments.*

The other is the limitation of resolution in our model. The length of low-resolution sequences is between 209 nt and 8097 nt, while the length of high-resolution sequences is 41 nt, which is much shorter than the lower bound of the low-resolution dataset. To estimate the resolution of our model, we evaluate the performance of the 5hmC on negative samples with different

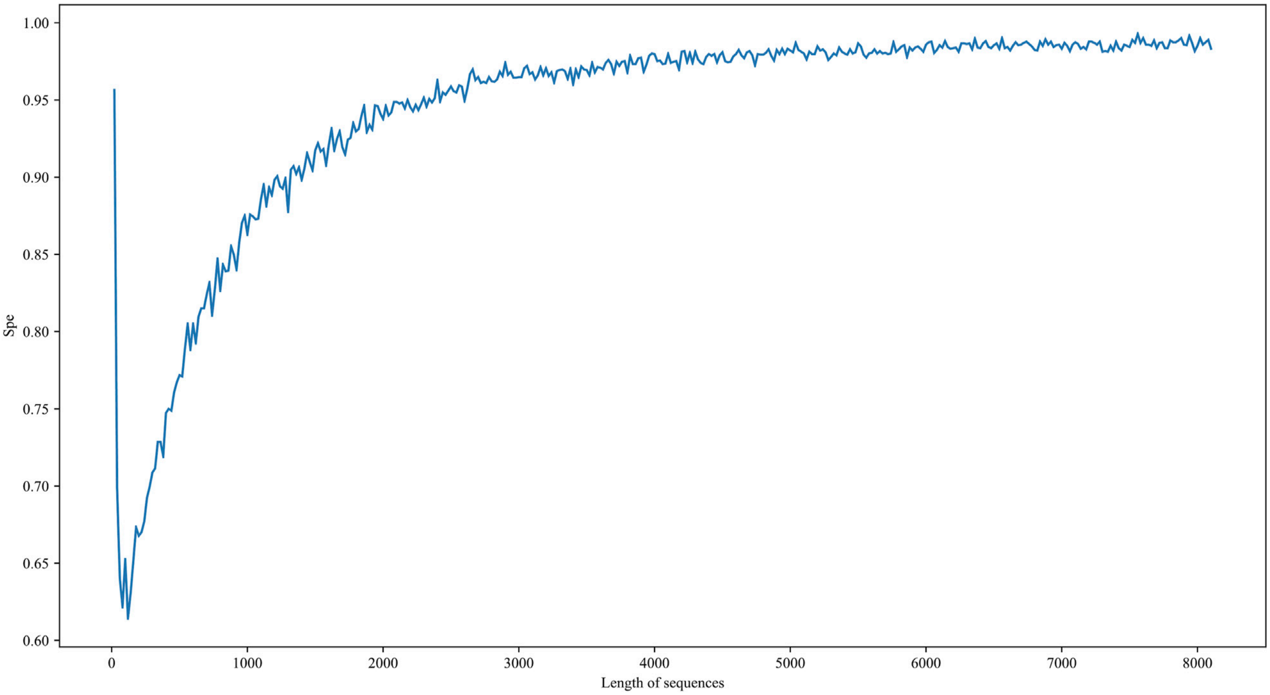

FIGURE 7 | Performance of i5hmCVec on the negative datasets of different sequence lengths with an independent test. The X-axis is the length of sequences in the negative dataset. The Y-axis is the performance of i5hmCVec on spe.

length restrictions. We re-select RNA sequences with sequence lengths ranging from 20 to 8100 on the non-peak region within the transcript carrying peak region as an independent testing dataset. It is worth noting that to prevent information leakage, there is no regional intersection between these negative samples and the negative samples in the benchmarking dataset. In addition, since there are only labels for negative samples, Spe is used as a performance metric. As shown in **Figure 7**, when the length of the sequence is less than 1000 nt, the performance of spe gradually drops. When the sequence length is around 100, the performance value takes a deep dive. Although the performance increases drastically when the sequence length is less than 100, we believe this is caused by over-fittings on negative samples. Therefore, the i5hmCVec model is not suitable for working on the high-resolution dataset.

CONCLUSION

In this study, we proposed a novel model named i5hmCVec for identifying 5hmC modification sites. We proposed a high-quality low-resolution 5hmC modification dataset. We construct the i5hmCVec based on dna2vec technology. The i5hmCvec achieved better performances than state-of-the-art methods on a low-resolution dataset. In addition, we analyze the semantic symmetric with the *Drosophila* genome. We hope our findings may be useful for future studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at: https://github.com/liu-h-y/5hmC_model.

AUTHOR CONTRIBUTIONS

H-YL collected the data, implemented the algorithm, performed the experiments, analyzed the results, and wrote the manuscript. P-FD directed the whole study, conceptualized the algorithm, supervised the experiments, analyzed the results, and wrote the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC 61872268) and the National Key R&D Program of China (2018YFC0910405).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.896925/full#supplementary-material

REFERENCES

Ahmed, S., Hossain, Z., Uddin, M., Taherzadeh, G., Sharma, A., Shatabda, S., et al. (2020). Accurate Prediction of RNA 5-hydroxymethylcytosine Modification by Utilizing Novel Position-specific Gapped K-Mer Descriptors. *Comput. Struct. Biotechnol. J.* 18, 3528–3538. doi:10.1016/j.csbj.2020.10.032

Akbar, S., Hayat, M., Iqbal, M., and Tahir, M. (2020). iRNA-PseTNC: Identification of RNA 5-methylcytosine Sites Using Hybrid Vector Space of Pseudo Nucleotide Composition. *Front. Comput. Sci.* 14, 451–460. doi:10.1007/s11704-018-8094-9

Asgari, E., and Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One* 10, e0141287. doi:10.1371/journal.pone.0141287

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., and Balasubramanian, S. (2014). 5-Hydroxymethylcytosine Is a Predominantly Stable DNA Modification. *Nat. Chem* 6, 1049–1055. doi:10.1038/nchem.2064

Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., et al. (2018). MODOMICS: a Database of RNA Modification Pathways. 2017 Update. *Nucleic Acids Res.* 46, D303–D307. doi:10.1093/nar/gkx1030

Bottou, L. (2012). "Stochastic Gradient Descent Tricks," in Neural Networks: Tricks of the Trade: Second Edition *Lecture Notes in Computer Science.* Editors G. Montavon, G. B. Orr, and K.-R. Müller (Berlin, Heidelberg: Springer), 421–436. doi:10.1007/978-3-642-35289-8_25

Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A. I., Webb, G. I., et al. (2020). Comprehensive Review and Assessment of Computational Methods for Predicting RNA post-transcriptional Modification Sites from RNA Sequences. *Brief Bioinform* 21, 1676–1696. doi:10.1093/bib/bbz112

Cowling, V. H. (2009). Regulation of mRNA Cap Methylation. *Biochem. J.* 425, 295–302. doi:10.1042/BJ20091352

Dai, C., Feng, P., Cui, L., Su, R., Chen, W., and Wei, L. (2021). Iterative Feature Representation Algorithm to Improve the Predictive Performance of N7-Methylguanosine Sites. *Brief Bioinform* 22, bbaa278. doi:10.1093/bib/bbaa278

de Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A Tutorial on the Cross-Entropy Method. *Ann. Oper. Res.* 134, 19–67. doi:10.1007/s10479-005-5724-z

Delatte, B., Wang, F., Ngoc, L. V., Collignon, E., Bonvin, E., Deplus, R., et al. (2016). Transcriptome-wide Distribution and Function of RNA Hydroxymethylcytosine. *Science* 351, 282–285. doi:10.1126/science.aac5253

Deng, L., Wu, H., and Liu, H. (2019). "D2VCB: A Hybrid Deep Neural Network for the Prediction of *In-Vivo* Protein-DNA Binding from Combined DNA Sequence," in *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*. Editors I. Yoo, J. Bi, and X. Hu (San Diego, CA, USA: IEEE), 74–77. November 18-21, 2019. doi:10.1109/BIBM47256.2019.8983051

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the Human and Mouse m6A RNA Methylomes Revealed by m6A-Seq. *Nature* 485, 201–206. doi:10.1038/nature11112

Dou, L., Li, X., Ding, H., Xu, L., and Xiang, H. (2020). Prediction of m5C Modifications in RNA Sequences by Combining Multiple Sequence Features. *Mol. Ther. - Nucleic Acids* 21, 332–342. doi:10.1016/j.omtn.2020.06.004

Fu, L., Guerrero, C. R., Zhong, N., Amato, N. J., Liu, Y., Liu, S., et al. (2014). Tet-mediated Formation of 5-hydroxymethylcytosine in RNA. *J. Am. Chem. Soc.* 136, 11582–11585. doi:10.1021/ja505305z

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying Enhancer-Promoter Interactions with Neural Network Based on Pre-trained DNA Vectors and Attention Mechanism. *Bioinformatics* 36, 1037–1043. doi:10.1093/bioinformatics/btz694

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., et al. (2015). The Release 6 Reference Sequence of the *Drosophila melanogaster* Genome. *Genome Res.* 25, 445–458. doi:10.1101/gr.185579.114

Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018). BERMP: a Cross-Species Classifier for Predicting m6A Sites by Integrating a Deep Learning Algorithm and a Random forest Approach. *Int. J. Biol. Sci.* 14, 1669–1677. doi:10.7150/ijbs.27819

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. doi:10.1093/bioinformatics/btab278

Huber, S. M., van Delft, P., Mendil, L., Bachman, M., Smollett, K., Werner, F., et al. (2015). Formation and Abundance of 5-hydroxymethylcytosine in RNA. *Chembiochem* 16, 752–755. doi:10.1002/cbic.201500013

Kai Chen, K., Luo, G.-Z., and He, C. (2015). High-Resolution Mapping of N6-Methyladenosine in Transcriptome and Genome Using a Photo-Crosslinking-Assisted Strategy. *Methods Enzymol.* 560, 161–185. doi:10.1016/bs.mie.2015.03.012

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., et al. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31, 51–54. doi:10.1093/nar/gkg129

Kimothi, D., Soni, A., Biyani, P., and Hogan, J. M. (2016). *Distributed Representations for Biological Sequence Analysis*. CoRR abs/1608.05949. Available at: http://arxiv.org/abs/1608.05949 (Accessed January 29, 2022).

Kunqi Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). WHISTLE: a High-Accuracy Map of the Human N6-Methyladenosine (m6A) Epitranscriptome Predicted Using a Machine Learning Approach. *Nucleic Acids Res.* 47, e41. doi:10.1093/nar/gkz074

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution Mapping of m6A and m6Am throughout the Transcriptome. *Nat. Methods* 12, 767–772. doi:10.1038/nmeth.3453

Lindstrom, D. L., Squazzo, S. L., Muster, N., Burckin, T. A., Wachter, K. C., Emigh, C. A., et al. (2003). Dual Roles for Spt5 in Pre-mRNA Processing and Transcription Elongation Revealed by Identification of Spt5-Associated Proteins. *Mol. Cel Biol* 23, 1368–1378. doi:10.1128/MCB.23.4.1368-1378.2003

Liu, X., Liu, Z., Mao, X., and Li, Q. (2020a). m7GPredictor: An Improved Machine Learning-Based Model for Predicting Internal m7G Modifications Using Sequence Properties. *Anal. Biochem.* 609, 113905. doi:10.1016/j.ab.2020.113905

Liu, Y., Chen, D., Su, R., Chen, W., and Wei, L. (2020b). iRNA5hmC: The First Predictor to Identify RNA 5-Hydroxymethylcytosine Modifications Using Machine Learning. *Front. Bioeng. Biotechnol.* 8, 227. doi:10.3389/fbioe.2020.00227

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (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

Miao, Z., Xin, N., Wei, B., Hua, X., Zhang, G., Leng, C., et al. (2016). 5-hydroxymethylcytosine Is Detected in RNA from Mouse Brain Tissues. *Brain Res.* 1642, 546–552. doi:10.1016/j.brainres.2016.04.055

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Editors Y. Bengio and Y. LeCun. Available at: http://arxiv.org/abs/1301.3781.

Ng, P. (2017a). *dna2vec: Consistent Vector Representations of Variable-Length K-Mers*. CoRR abs/1701.06279. Available at: http://arxiv.org/abs/1701.06279.

Ng, P. (2017b). *dna2vec: Consistent Vector Representations of Variable-Length K-Mers*. arXiv:1701.06279 [cs, q-bio, stat]. Available at: http://arxiv.org/abs/1701.06279 (Accessed January 23, 2022).

Pastor, W. A., Pape, U. J., Huang, Y., Henderson, H. R., Lister, R., Ko, M., et al. (2011). Genome-wide Mapping of 5-hydroxymethylcytosine in Embryonic Stem Cells. *Nature* 473, 394–397. doi:10.1038/nature10102

Qiu, W.-R., Jiang, S.-Y., Xu, Z.-C., Xiao, X., and Chou, K.-C. (2017). iRNAm5C-PseDNC: Identifying RNA 5-methylcytosine Sites by Incorporating Physical-Chemical Properties into Pseudo Dinucleotide Composition. *Oncotarget* 8, 41178–41188. doi:10.18632/oncotarget.17104

Sabooh, M. F., Iqbal, N., Khan, M., Khan, M., and Maqbool, H. F. (2018). Identifying 5-methylcytosine Sites in RNA Sequence Using Composite Encoding Feature into Chou's PseKNC. *J. Theor. Biol.* 452, 1–9. doi:10.1016/j.jtbi.2018.04.037

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach Learn.* 16, 235–240. doi:10.1007/BF00993309

Sommer, S., Lavi, U., and Darnell, J. E. (1978). The Absolute Frequency of Labeled N-6-Methyladenosine in HeLa Cell Messenger RNA Decreases with Label Time. *J. Mol. Biol.* 124, 487–499. doi:10.1016/0022-2836(78)90183-3

Szwagierczak, A., Bultmann, S., Schmidt, C. S., Spada, F., and Leonhardt, H. (2010). Sensitive Enzymatic Quantification of 5-hydroxymethylcytosine in Genomic DNA. *Nucleic Acids Res.* 38, e181. doi:10.1093/nar/gkq684

van der Maaten, L., and Hinton, G. (2008). Viualizing Data Using T-SNE. *J. Machine Learn. Res.* 9, 2579–2605.

Wei Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.-C. (2015). iRNA-Methyl: Identifying N6-Methyladenosine Sites Using Pseudo Nucleotide Composition. *Anal. Biochem.* 490, 26–33. doi:10.1016/j.ab.2015.08.021

Wei Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019). iRNA-m7G: Identifying N7-Methylguanosine Sites by Fusing Multiple Features. *Mol. Ther. - Nucleic Acids* 18, 269–274. doi:10.1016/j.omtn.2019.08.022

Yang, Y.-H., Ma, C., Wang, J.-S., Yang, H., Ding, H., Han, S.-G., et al. (2020). Prediction of N7-Methylguanosine Sites in Human RNA Based on Optimal Sequence Features. *Genomics* 112, 4342–4347. doi:10.1016/j.ygeno.2020.07.035

Yu, M., Hon, G. C., Szulwach, K. E., Song, C.-X., Zhang, L., Kim, A., et al. (2012). Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* 149, 1368–1380. doi:10.1016/j.cell.2012.04.027

Zhang, H.-Y., Xiong, J., Qi, B.-L., Feng, Y.-Q., and Yuan, B.-F. (2016). The Existence of 5-hydroxymethylcytosine and 5-formylcytosine in Both DNA and RNA in Mammals. *Chem. Commun.* 52, 737–740. doi:10.1039/c5cc07354e

Zhang, L.-S., Liu, C., Ma, H., Dai, Q., Sun, H.-L., Luo, G., et al. (2019). Transcriptome-wide Mapping of Internal N7-Methylguanosine Methylome in Mammalian mRNA. *Mol. Cel* 74, 1304–1316. e8. doi:10.1016/j.molcel.2019.03.036

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. doi:10.1093/nar/gkw104

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA* 25, 205–218. doi:10.1261/rna.069112.118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.