Molecular Therapy
Nucleic Acids
Original Article

AMERICAN SOCIETY of GENE & CELL THERAPY

# A Linear Regression Predictor for Identifying N$^6$-Methyladenosine Sites Using Frequent Gapped K-mer Pattern

Y.Y. Zhuang,[1,4] H.J. Liu,[2,4] X. Song,[3] Y. Ju,[1] and H. Peng[1]

[1]School of Informatics, Xiamen University, Xiamen 361005, China; [2]College of Information Technology and Computer Science, University of the Cordilleras, Baguio 2600, Philippines; [3]School of Computer and Information Technology, Nanyang Normal University, Nanyang 473000, China

**N6-methyladenosine (m$^6$A) is one of the most common and abundant modifications in RNA, which is related to many biological processes in humans. Abnormal RNA modifications are often associated with a series of diseases, including tumors, neurogenic diseases, and embryonic retardation. Therefore, identifying m$^6$A sites is of paramount importance in the post-genomic age. Although many lab-based methods have been proposed to annotate m$^6$A sites, they are time consuming and cost ineffective. In view of the drawbacks of the intrinsic methods in RNA sequence recognition, computational methods are suggested as a supplement to identify m$^6$A sites. In this study, we develop a novel feature extraction algorithm based on the frequent gapped k-mer pattern (FGKP) and apply the linear regression to construct the prediction model. The new predictor is used to identify m$^6$A sites in the *Saccharomyces cerevisiae* database. It has been shown by the 10-fold cross-validation that the performance is better than that of recent methods. Comparative results indicate that our model has great potential to become a useful and effective tool for genome analysis and gain more insights for locating m$^6$A sites.**

## INTRODUCTION

Over 100 modifications occur in RNA.[1] The functions of internal modifications of mRNA are used to keep the stability of mRNA, and the most common internal modifications of mRNA include N$^6$-methyladenosine (m$^6$A), N$^1$-methyladenosine (m$^1$A), 5-methylcytosine (m$^5$C). Among them, global scientists have verified many enzymes that m$^6$A engages, such as histone demethylases, methylase, and methylation recognition enzyme.[2] Abnormal m$^6$A modifications are often related to a series of diseases, including tumors, neurogenic diseases, and embryonic retardation.[3] RNA m$^6$A was first observed in 1970s.[4] Since then, m$^6$A is found in a wide spectrum of all living organisms and linked to many important roles of biological activities, including mRNA splicing, stability, nuclear processing, and immune response.[5–8] Therefore transcriptome-wide annotation of m$^6$A sites will be helpful to understand its biological functions.

In the past few years, high-throughput sequencing techniques such as MeRIPSeq[9] and m$^6$A-seq[10] have identified m$^6$A peaks in *Saccharomyces cerevisiae*, *Mus musculus*, and *Homo sapiens*. At the same time, the miCLIP technique[11] was proposed to provide the recognition method of m$^6$A sites in the human transcriptome. However, in consideration of the biological inherent reliance of the techniques,[12] they are still neither budget nor time efficient in performing transcriptome-wide analysis.

Although lab-based technologies have been widely applied to identify m$^6$A, some cost-effective computational methods are developed in assisting the process as well. To identify methylated m$^6$A sites, building a high-resolution database is of paramount importance in predicting m$^6$A sites. Using the high-resolution database of *Saccharomyces cerevisiae* constructed by Schwartz et al.,[13] Chen et al.[14–18] proposed a series of predictors such as "iRNA-Methyl," "M6ATH," "MethyRNA," "iRNA-3typeA" and "iRNA(m6A)-PseDNC," which formulated RNA sequences by using different combinations of feature extractions and classifiers to make predictions. Feng et al.[19] used a method called "iRNA-PseColl," which incorporated collective features of the RNA sequence elements into PseKNC to make predictions. Jaffrey et al.[11] built a single-nucleotide resolution map of m$^6$A sites across *Homo sapiens*. More recently, Chen et al.[20] proposed a support-vector-machine-based method to predict m$^6$A sites in *Arabidopsis thaliana*. As mentioned in some references, well-established ensemble classifiers have been proven to outperform single classifiers.[21–23] Based on this, Wei et al.[24] thus proposed an m$^6$A predictor by constructing an ensemble classifier based on the support vector machine (SVM) to successfully improve the predictive performance. Wei et al.[25,26] have also done a lot of research with the ensemble classifier, which has great significance for reference in our study.

In this article, we propose a novel method for the identification of m$^6$A sites within RNA sequences. As for feature representation, we use the frequent gapped k-mer pattern (FGKP) discovery algorithm

Received 9 May 2019; accepted 3 October 2019;
https://doi.org/10.1016/j.omtn.2019.10.001.

[4]These authors contributed equally to this work.

**Correspondence:** Xiao Song, School of Computer and Information Technology, Nanyang Normal University, Nanyang 473000, China.
**E-mail:** sxyoland@foxmail.com

**Table 1. Comparison of Different Feature Extractions**

| Feature | SP (%) | SN (%) | ACC (%) | MCC | AUROC |
|---|---|---|---|---|---|
| Triplet | 56.92 | 63.85 | 59.92 | 0.20 | 0.6669 |
| Pse-SSC | 78.77 | 64.66 | 72.52 | 0.44 | 0.7284 |
| Frequent gapped k-mer | 71.92 | 83.62 | 77.10 | 0.55 | 0.8307 |

to mainly capture the properties in RNA sequences. In the predictive model, we use the linear regression to discriminate the positive and negative samples. Experimental results show that our model outperformed other existing methods in the literature under the 10-fold cross-validation test.

## RESULTS

Several diseases have their underlying causes in RNA,[27,28] including cancers.[29–31] In our study, we combined the advantage of effective extraction of frequent gapped k-mer (FGK) and the strong ability of classification of the linear predictive model to create a powerful predictive tool in order to discriminate the positive and negative samples of m$^6$A. The learning machine that we used was logistic regression (LR). We have experimented with our predictor in the *Saccharomyces cerevisiae* genome using 10-fold cross-validation. It turns out that our model is superior to M6A-HPCS, the recent classifier in this area, and also has a better performance than other feature extractions and different parameters within our model. We anticipate that it will shed some light on genome analysis in future practice.

### Four Evaluation Metrics

In general, the following four metrics are used to measure the quality of a predictor:[32] sensitivity (*SN*), specificity (*SP*), accuracy (*ACC*), and Matthews correlation coefficient (*MCC*). These metrics were first introduced by Chou[33] and then they were widely applied to a wide range of biological areas (see Liu et al.,[34–37] Ehsan et al.,[38] Feng et al.,[19] Song et al.,[39] Lin et al.,[40] and Xu et al.[40,41]). Their definitions are as follows:

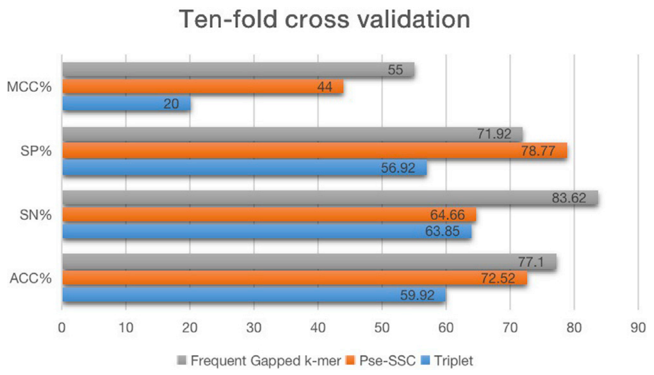


**Figure 1. Performance of Different Feature Extractions Using 10-Fold Cross-validation**

Here, we compare the effect of our feature extraction (FGK) with Pse-SSC and Triplet methods.

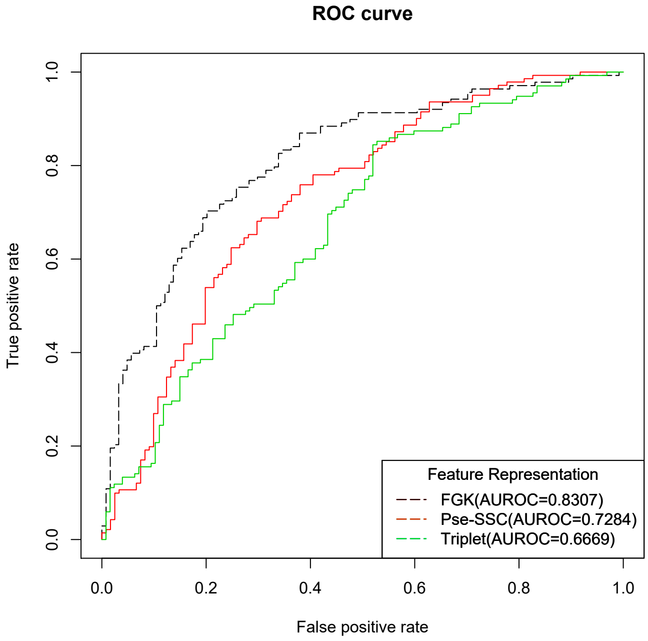


**Figure 2. ROC Curves of Frequent Gapped K-mer, Pse-SSC, and Triplet and Their AUROC Values**

$$SP = \frac{TN}{TN+FP} \times 100\% \qquad \text{(Equation 1)}$$

$$SN = \frac{TP}{TP+FN} \times 100\% \qquad \text{(Equation 2)}$$

$$ACC = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \qquad \text{(Equation 3)}$$

and

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}} \qquad \text{(Equation 4)}$$

where *TP*, *TN*, *FP*, and *FN* are true positive, true negative, false positive, and false negative, respectively. In this research, *TP* represents the true m$^6$A site predicted correctly, *TN* represents the non-m$^6$A site predicted incorrectly, *FP* represents the non-m$^6$A site predicted incorrectly as the true m$^6$A site, and *FN* represents the non-m$^6$A site predicted correctly as the non-m$^6$A site. The values of *SN*, *SP*, *ACC*, are between 0 and 1. The closer to 1 they get, the more accuracy our model achieves; the value of *MCC* is between −1 and 1. The larger the value that *MCC* gets, the better performance our prediction model obtains.

### Cross-Validation

Normally, three types of validation are used to derive the metric values: independent test sets, subsampling (or K-fold cross-validation), and

**Table 2. Performance Comparison of Different Classifiers**

| Classifier | SP (%) | SN (%) | ACC (%) | MCC |
|---|---|---|---|---|
| SVM | 80 | 46.83 | 48.09 | 0.10 |
| RF | 75.51 | 72.56 | 73.66 | 0.47 |
| LR | 71.92 | 83.62 | 77.10 | 0.55 |

the jackknife test (or LOOCV). Although the jackknife test can fully train the data we already have to acquire a more accurate classifier, and it has definite sampling and error estimation based on the specific dataset, the jackknife test is not a time-efficient method compared with the other two types of validation. In this article, we adopted the 10-fold cross-validation method used by many researchers[42–44] in this area.

### ROC Curve

ROC curve (also called the sensitivity curve) is the abbreviation for receiver operating characteristic curve. Every point on the curve reflects the same sensitivity. They react to the same signal simulation in the different judgment standards. Therefore, the ROC curve can be generally treated as the overall performance in the binary classification problems. The ROC curve is normally plotted with the x-axis true-positive rate (TPR) and the y-axis false-positive rate (FPR) in the different thresholds of the classification. We can understand the TPR as the sensitivity as described earlier, and the FPR can be computed as $1 - $ specificity. The area under the ROC curve (AUROC) can also be calculated. The AUROC is the indicator of the performance of a predictor. The AUROC ranges from 0.5 to 1. The closer the AUROC score of a predictor to 1, the better and more robust the predictor we can reckon, and we can deem the AUROC score of 0.5 of a predictor as a random predictor.

## DISCUSSION

### Comparison among Different Feature Extractions

To justify our feature extraction technique, we make comparisons with two of the most commonly used feature representation techniques, Triplet and Pse-SSC, and this shows that the FGK method gets the much better performance than the other two feature representations. We show the result in Table 1, and from Figure 1, we can see the graphical comparisons from four different evaluation metrics. The FGK leads Pse-SSC by 4% and Triplet by 17% for the *ACC*, and for the *MCC* metric, FGK outnumbers its counterparts by over 10%. From Figure 2, we can see the effects of three different feature extractions from their ROC curves. The larger areas under the curve we get, the better performance the method achieves. Also, we can also see from Table 1 that our feature representation is 63.2% and 16.4% higher than features Pse-SSC and Triplet, respectively.

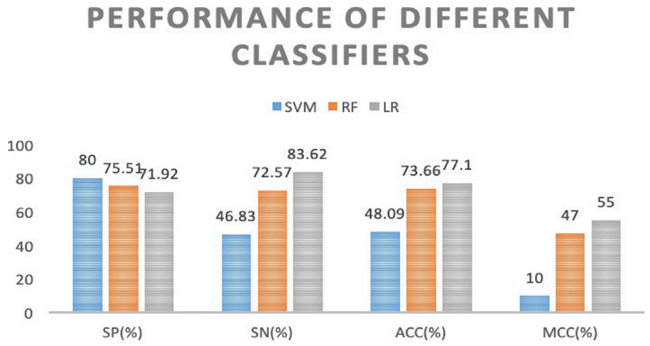


**Figure 3. Comparison of Performances among the LR Classifier and Other Popular Classifiers (SVM and RF) with the Same Learning Feature Representations on the *S. cerevisiae* Dataset**

**Table 3. Performance Comparison of Different Parameters in Our Model**

| Classifier | SP (%) | SN (%) | ACC (%) | MCC |
|---|---|---|---|---|
| LR (k = 5, γ = 0.05) | 73.53 | 73.81 | 73.66 | 0.47 |
| LR( k = 4, γ = 0.025) | 71.92 | 83.62 | 77.10 | 0.55 |

### Comparison with Other Classifiers

In Table 2, we compare LR with SVM and random forest (RF). The reason for choosing SVM and RF for comparison is because SVM[20,21,45,46] and RF[5,47–50] are two of the most widely used classifiers in bioinformatics. Although the *SP* of the proposed method is lower than those of SVM and RF, its *SN*, *ACC*, and *MCC* are higher than those of SVM and RF, indicating that the performance of the LR-based model can effectively discriminate the $m^6A$ sites in *Saccharomyces cerevisiae*. We can see the overall performance of three classifiers in Figure 3. In this figure, we can see that, although the *SP* of LR performs poorly compared to that of the other two classifiers, the other three metrics are much better than the rest for the two predictors. The *ACC* of LR is far better than that of SVM, topping by almost 30% and slightly exceeding by 3.5% the *ACC* of RF.

### Comparison with Different Parameters

In Table 3, we compared the model prediction performance of linear regression using different parameters and found that, with parameters $k = 4$ and $\gamma = 0.025$, we get the most desirable result. The classifier with parameters $k = 5$ and $\gamma = 0.05$ is almost 25% higher than its counterpart in *ACC*.

### Comparison with Existing Predictors

To evaluate the performance of our proposed predictor, we compared our predictor with two existing predictors, iRNA-Methyl[14] and M6A-HPCS.[51] The reason to choose these two predictors for comparison is that they have been reported to achieve outstanding performance in

**Table 4. Comparison of M6APred-FG with Other Well-Known Classifiers**

| Prediction Method | SP (%) | SN (%) | ACC (%) | MCC |
|---|---|---|---|---|
| iRNA-Methyl | 60.63 | 70.59 | 65.59 | 0.29 |
| M6A-HPCS | 62.89 | 71.77 | 67.33 | 0.35 |
| iRNA-Freq | 71.92 | 83.62 | 77.10 | 0.55 |

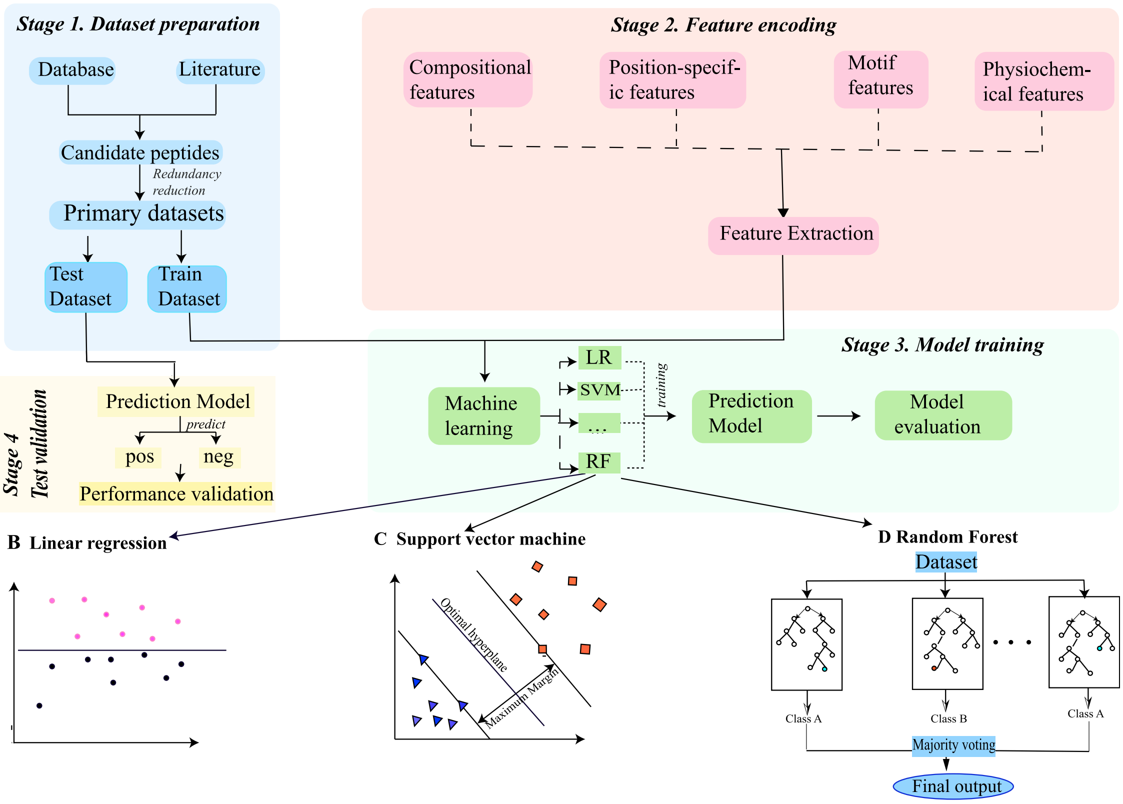


**Figure 4. Flowchart of the Proposed Predictor**

Stage 1 shows the procedure of dataset preparation. We chose a benchmark database and used updated literature to obtain candidate peptides. Since the candidate peptides have imbalanced positive and negative samples, we needed to balance the samples (or reduce redundancy) to get the primary dataset. Then, we divided the dataset into the test dataset and the train dataset. Stage 2 shows the feature encoding or feature extraction. In our sample sequences, there is information hidden. We needed to find a way to extract their features to best represent the original samples and digitalize them. Stage 3 shows how we used the train dataset and chose the appropriate model to gain a prediction model and evaluate it. Stage 4 shows how we tested and validated our predictive model. In our article, we combined stages 3 and 4 together using 10-fold cross-validation to evaluate our model.

m$^6$A site identification. For fairness of comparison, all compared predictors are trained and validated on the same benchmark dataset. The results are summarized in Table 4. It can be observed that, among the compared predictors, the proposed model obtains the best performance in terms of *ACC* and *MCC*, with 77.10% and 55%, respectively. Compared with the best of the existing predictors, M6A-HPCS, our classifier performance is about 10% higher for *ACC* and 20% higher for *MCC*.

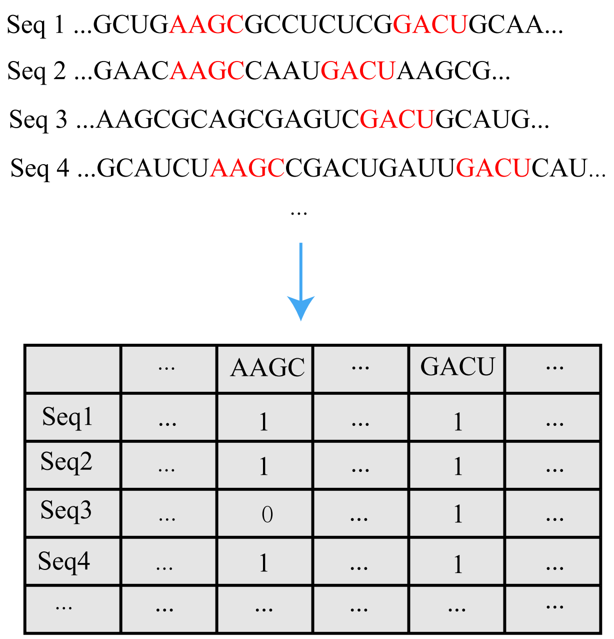
## MATERIALS AND METHODS

### Framework of the Proposed Predictor

Figure 4 shows the flowchart of the proposed predictor. The first stage is to collect data from verified databases and relevant literature.[14,15,52] In this research, we use the organized dataset from Chen et al.'s[14] work. The second stage is feature encoding. This stage includes feature representation and feature optimization. Feature representation means extracting characteristics of RNA sequences using various feature descriptors, including composition features like Dinucleotide-based auto covariance (DAC), physicochemical features like PC-PseDNC-General, and our newly found FGKP. The final stage is to train the machine learning model (i.e., SVM, RF, and linear regression) using the feature extraction from the last stage. The predictive model constructed is based on the feature extraction mentioned earlier and validated through validation methods. In this study, we used the 10-fold cross-validation test.

### Datasets

m$^6$A sites have been widely identified in *Saccharomyces cerevisiae*,[13] *Homo sapiens*,[10,11] *Mus musculus*,[10] and *Arabidopsis thaliana*.[53] In this work, we used the dataset from *Saccharomyces cerevisiae*. In the *Saccharomyces cerevisiae* genome, m$^6$A sites have the same motif, GAC, and they are more easily methylated.[13] Since RNA sequences in *Saccharomyces cerevisiae* have different lengths, we used the organized dataset from Chen et al.'s[14] work. There are 1,307 positive samples and 1,307 negative samples, where the negative samples were randomly collected from 33,280 sequences with non-m$^6$A sites. All sequences in the dataset are 51 nt long (25 nt on each side of the m$^6$A/non-m$^6$A sites), with the sequence similarity less than 85%.



Seq 1 ...GCUGAAGCGCCUCUCGGACUGCAA...

Seq 2 ...GAACAAGCCAAUGACUAAGCG...

Seq 3 ...AAGCGCAGCGAGUCGACUGCAUG...

Seq 4 ...GCAUCUAAGCCGACUGAUUGACUCAU...

...

|  | ... | AAGC | ... | GACU | ... |
|---|---|---|---|---|---|
| Seq1 | ... | 1 | ... | 1 | ... |
| Seq2 | ... | 1 | ... | 1 | ... |
| Seq3 | ... | 0 | ... | 1 | ... |
| Seq4 | ... | 1 | ... | 1 | ... |
| ... | ... | ... | ... | ... | ... |

**Figure 5. The Transformation from the Original Samples to 0–1 Sequences**

### Representation of RNA Sample

The RNA samples in our dataset can be generally expressed as the following pattern:

$$R = M_1 M_2 M_3 \cdots M_{51} \qquad \text{(Equation 5)}$$

where

$$M_i \in \{A(adenine), C(cyto\sin e), G(guanine), U(uracil)\} i = 1, 2, 3, \cdots, 51.$$

The first thing we would need to do is to transform the RNA sequence in Equation 5 to a vector. However, a vector might lose its sequential information and pattern. In order to solve the problem, we introduce the FGKP discovery algorithm that we recently found. In this method, we can separate our algorithm into four steps and elaborate each step accordingly:

(1) Search all the FGK sub-sequences from each sequence in the dataset.

We find all FGK sub-sequences from each sequence in the dataset and calculate the frequency of gapped k-mer sub-sequences, and we can set the frequency threshold here. Here, the parameter *k* means the matching length of the sub-sequences, and we denote the frequency threshold as $\gamma$.

(2) Build a set for the frequent sub-sequences.

FGK are subjects and whose lengths over a threshold is an attribute clause which modifies the subjects. We can map each FGK sub-sequence into a column of the table as shown in Figure 5.

(3) Utilize the frequent k-mer sub-sequence set as features to generate vectors.

First of all, we define the following functions:

$$c\left(S_i, FkM_j\right) = \begin{cases} 1, & \textit{if } FkM_j \textit{ exactly matches } S_i \\ 0, & \textit{Otherwise} \end{cases} \qquad \text{(Equation 6)}$$

$$\phi(S_i) = \begin{pmatrix} c(S_i, FkM_1) \\ c(S_i, FkM_2) \\ \cdots \\ c(S_i, FkM_n) \end{pmatrix}. \qquad \text{(Equation 7)}$$

Here, $S_i$ denotes the sequence that is predicted, and $FkM_j$ denotes the j-element of the frequent k-mer sub-sequence set. As you can see from the function in Equation 6, we define a function c, which compares the predicted sequence $S_i$ and the j-element of the frequent k-mer sequence set, and we discriminate the perfect matching between $S_i$ and $FkM_j$ using 1 and 0 otherwise. After this procedure, we map the sequence $S_i$ using the function $\phi$ to a 0–1 vector as shown in the function in Equation 7.

### Linear Predictive Model

Although a huge amount of literature is related to classification methods such as SVM[21,52,54–62] and RF,[5,47–50] as we can see from the feature representation algorithm of RNA sample, a series of sparse data is produced. Therefore, the need to deal with a large amount of sparse data is imperative. The linear predictive model is a linear classifier for processing a large amount of sparse data with a large number of examples and features. It is a general term for supervised models, including LR, SVM, and support vector regression (SVR). In this study, we used the packages LIBSVM[63] and LIBLINEAR.[64] They support the multiple types of linear classifiers that we mentioned earlier. In this study, we used LR and achieved a good result. LR uses the optimal decision boundary to construct regression formula and fitted parameter sets. The main idea is as follows:

1. Construct the prediction function $h_\theta$, where $\theta$ represents the parameter sets of eigenvalue X.

As far as we know, $h_\theta$could have a linear relationship or non-linear relationship with X, as we can see from Figure 6. Normally, we can represent the linear relationship between $h_\theta$and X using the formula

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \qquad \text{(Equation 8)}$$

and the non-linear relationship using the formula

$$h_\theta(x) = g\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2\right). \qquad \text{(Equation 9)}$$

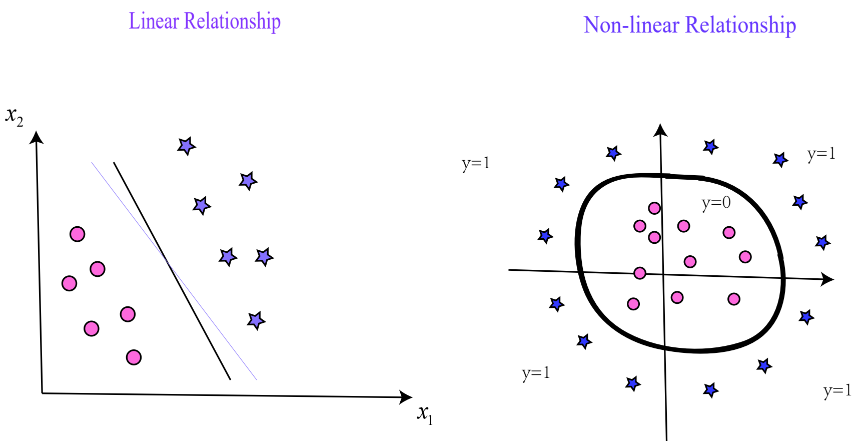


**Figure 6. The Linear and Non-linear Relationships between $h_\theta$ and X**

For details, see the Linear Predictive Model section in Materials and Methods.

In linear programming, the idea of cost function is to minimize the difference of predictive result $h_\theta$and actual y; i.e.,

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2. \qquad \text{(Equation 10)}$$

Then in LR, we can represent $J(\theta)$ as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}Cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right). \qquad \text{(Equation 11)}$$

2. Use gradient descent to calculate the maximum of $J(\theta)$.

We can achieve the maximum of $J(\theta)$ through fitting parameters using the gradient of the function. For simplicity, we can consider the following cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}Cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) \qquad \text{(Equation 12)}$$

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y=1 \\ -\log(1-h_\theta(x)) & if\ y=0 \end{cases} \qquad \text{(Equation 13)}$$

and we can renew the parameter $\theta_j: = \theta_j + \alpha(\partial J(\theta)/\partial\theta_j)$; that is,

$$\theta_j : = \theta_j - \alpha\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}. \qquad \text{(Equation 14)}$$

## AUTHOR CONTRIBUTIONS

Y.Z. conceived the project, designed the experiments, and edited the final version of the paper. H.L. performed the experiment. X.S. wrote the paper and drafted the figures. H.P. contributed to materials and data analysis.

## CONFLICTS OF INTEREST

The authors declare no competing interests.

## ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (grant no. U1504605). Y.Z. is supported by the Xiamen University (grant no. 030/510037).

## REFERENCES

1. Acharjee, A., Kloosterman, B., Visser, R.G.F., and Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. BMC Bioinformatics *17* (*Suppl 5*), 180.

2. Belk, A., Xu, Z.Z., Carter, D.O., Lynne, A., Bucheli, S., Knight, R., and Metcalf, J.L. (2018). Microbiome data accurately predicts the postmortem interval using random forest regression models. Genes (Basel) *9*, E104.

3. Cantara, W.A., Crain, P.F., Rozenski, J., McCloskey, J.A., Harris, K.A., Zhang, X., Vendeix, F.A., Fabris, D., and Agris, P.F. (2011). The RNA modification database, RNAMDB: 2011 update. Nucleic Acids Res. *39*, D195–D201.

4. Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. Proc. Natl. Acad. Sci. USA *71*, 3971–3975.

5. Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. Cell *169*, 1187–1200.

6. Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

7. Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y.G., and He, C. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat. Chem. Biol. *7*, 885–887.

8. Nilsen, T.W. (2014). Molecular biology. Internal mRNA methylation finally finds functions. Science *343*, 1207–1208.

9. 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.

10. 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 m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201–206.

11. 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.

12. Meyer, K.D., and Jaffrey, S.R. (2017). Rethinking m$^6$A readers, writers, and erasers. Annu. Rev. Cell Dev. Biol. *33*, 319–342.

13. Schwartz, S., Agarwala, S.D., Mumbach, M.R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T.S., 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.

14. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2015). iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. *490*, 26–33.

15. Chen, W., Feng, P., Ding, H., and Lin, H. (2016). Identifying N $^6$-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. Mol. Genet. Genomics *291*, 2225–2229.

16. Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N$^6$-methyladenosine sites. J. Biomol. Struct. Dyn. *35*, 683–687.

17. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2018). iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol. Ther. Nucleic Acids *11*, 468–474.

18. Chen, W., Ding, H., Zhou, X., Lin, H., and Chou, K.C. (2018). iRNA(m6A)-PseDNC: identifying N$^6$-methyladenosine sites using pseudo dinucleotide composition. Anal. Biochem. *561–562*, 59–65.

19. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.C. (2017). iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol. Ther. Nucleic Acids *7*, 155–163.

20. Chen, J., Long, R., Wang, X.L., Liu, B., and Chou, K.C. (2016). dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. Sci. Rep. *6*, 32333.

21. Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., and Chou, K.C. (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics *30*, 472–479.

22. Zou, Q., Guo, J., Ju, Y., Wu, M., Zeng, X., and Hong, Z. (2015). Improving tRNAscan-SE annotation results via ensemble classifiers. Mol. Inform. *34*, 761–770.

23. Chen, W., Xing, P., and Zou, Q. (2017). Detecting N$^6$-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. Sci. Rep. *7*, 40242.

24. Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. Mol. Ther. Nucleic Acids *12*, 635–644.

25. Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. *83*, 67–74.

26. Wei, L., Wan, S., Guo, J., and Wong, K.K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. *83*, 82–90.

27. Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 905–915.

28. Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2019). Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. IEEE/ACM Trans. Comput. Biol. Bioinform. *16*, 396–406.

29. Tang, W., Wan, S., Yang, Z., Teschendorff, A.E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics *34*, 398–406.

30. Moridikia, A., Mirzaei, H., Sahebkar, A., and Salimian, J. (2018). MicroRNAs: potential candidates for diagnosis and treatment of colorectal cancer. J. Cell. Physiol. *233*, 901–913.

31. Zhao, R., Zhang, Y., Zhang, X., Yang, Y., Zheng, X., Li, X., Liu, Y., and Zhang, Y. (2018). Exosomal long noncoding RNA HOTTIP as potential novel diagnostic and prognostic biomarker test for gastric cancer. Mol. Cancer *17*, 68.

32. Chen, J., Liu, H., Yang, J., and Chou, K.C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids *33*, 423–428.

33. Chou, K.-C. (2001). Prediction of signal peptides using scaled window. Peptides *22*, 1973–1979.

34. Liu, B., Wang, S., Long, R., and Chou, K.C. (2017). iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics *33*, 35–41.

35. Liu, L.M., Xu, Y., and Chou, K.C. (2017). iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med. Chem. *13*, 552–559.

36. Liu, B., Yang, F., Huang, D.S., and Chou, K.C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.

37. Liu, B., Yang, F., and Chou, K.C. (2017). 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Mol. Ther. Nucleic Acids *7*, 267–277.

38. Ehsan, A., Mahmood, K., Khan, Y.D., Khan, S.A., and Chou, K.C. (2018). A novel modeling in mathematical biology for classification of signal peptides. Sci. Rep. *8*, 1039.

39. Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I., and Chou, K.C. (2019). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief. Bioinform. *20*, 638–658.

40. Lin, H., Deng, E.Z., Ding, H., Chen, W., and Chou, K.C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. *42*, 12961–12972.

41. Xu, Y., Shao, X.J., Wu, L.Y., Deng, N.Y., and Chou, K.C. (2013). iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ *1*, e171.

42. Li, W.C., Deng, E.Z., Ding, H., Chen, W., and Lin, H. (2015). iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemometr. Intell. Lab. Syst. *141*, 100–106.

43. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics *35*, 1469–1477.

44. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics *35*, 2796–2800.

45. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics *33*, 3518–3523.

46. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.C. (2019). iDNA6mA-PseKNC: Identifying DNA N$^6$-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics *111*, 96–102.

47. Adams, J.M., and Cory, S. (1975). Modified nucleosides and bizarre 5′-termini in mouse myeloma mRNA. Nature *255*, 28–33.

48. Naue, J., Hoefsloot, H.C.J., Mook, O.R.F., Rijlaarsdam-Hoekstra, L., van der Zwalm, M.C.H., Henneman, P., Kloosterman, A.D., and Verschure, P.J. (2017). Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. Forensic Sci. Int. Genet. *31*, 19–28.

49. Sarkar, R.K., Rao, A.R., Meher, P.K., Nepolean, T., and Mohapatra, T. (2015). Evaluation of random forest regression for prediction of breeding value from genomewide SNPs. J. Genet. *94*, 187–192.

50. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. *43*, 1947–1958.

51. Zhang, M., Sun, J.W., Liu, Z., Ren, M.W., Shen, H.B., and Yu, D.J. (2016). Improving N(6)-methyladenosine site prediction with heuristic selection of nucleotide physical-chemical properties. Anal. Biochem. *508*, 104–113.

52. 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.

53. Luo, G.Z., MacQueen, A., Zheng, G., Duan, H., Dore, L.C., Lu, Z., Liu, J., Chen, K., Jia, G., Bergelson, J., and He, C. (2014). Unique features of the m6A methylome in *Arabidopsis thaliana*. Nat. Commun. *5*, 5630.

54. Chen, X., Yan, C.C., Zhang, X., You, Z.H., Deng, L., Liu, Y., Zhang, Y., and Dai, Q. (2016). WBSMDA: within and between score for miRNA-disease association prediction. Sci. Rep. *6*, 21106.

55. Tang, H., Chen, W., and Lin, H. (2016). Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Mol. Biosyst. *12*, 1269–1275.

56. Yang, H., Tang, H., Chen, X.X., Zhang, C.J., Zhu, P.P., Ding, H., Chen, W., and Lin, H. (2016). Identification of secretory proteins in *Mycobacterium tuberculosis* using pseudo amino acid composition. BioMed Res. Int. *2016*, 5413903.

57. Lin, H., Liang, Z.Y., Tang, H., and Chen, W. (2019). Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans. Comput. Biol. Bioinform. *16*, 1316–1321.

58. Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K.C. (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE *10*, e0121501.

59. Xiao, Y., Zhang, J., and Deng, L. (2017). Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. Sci. Rep. *7*, 3664.

60. Lai, H.Y., Chen, X.X., Chen, W., Tang, H., and Lin, H. (2017). Sequence-based predictive modeling to identify cancerlectins. Oncotarget *8*, 28169–28175.

61. Chou, K.C., and Cai, Y.D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. *277*, 45765–45769.

62. Cai, Y.D., Zhou, G.P., and Chou, K.C. (2003). Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. *84*, 3257–3263.

63. Chang, C.C., and Lin, C.J. (2011). LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 27.

64. Fan, R.E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. *9*, 1871–1874.