ISGm1A: Integration of Sequence Features and Genomic Features to Improve the Prediction of Human m\(^1\)A RNA Methylation Sites

LIAN LIU\(^1\), XIUJUAN LEI\(^1\), (Member, IEEE), JIA MENG\(^2\), AND ZHEN WEI\(^2\)

\(^1\)School of Computer Science, Shaanxi Normal University, Xi’an 710119, China
\(^2\)Department of Biological Sciences, Xi’an Jiaotong-Liverpool University, Suzhou 215123, China

Corresponding authors: Xiujuan Lei (xjlei@snnu.edu.cn) and Zhen Wei (zhen.wei@xjtlu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61902230, Grant 61972451, and Grant 31671373, in part by the China Postdoctoral Science Foundation under Grant 2018M640949, in part by the Fundamental Research Funds for the Central Universities under Grant GK201903083 and Grant GK201901010, and in part by the Xi’an Jiaotong-Liverpool University (XJTLU) Key Program Special Fund under Grant KSF-T-01.

ABSTRACT As a new epitranscriptomic modification, N1-methyladenosine (m\(^1\)A) plays an important role in the gene expression regulation. Although some computational methods were proposed to predict m\(^1\)A modification sites, all of these methods apply machine learning predictions based on the nucleotide sequence features, and they missed the layer of information in transcript topology and RNA secondary structures. To enhance the prediction model of m\(^1\)A RNA methylation, we proposed a computational framework, ISGm1A, which stands for integration sequence features and genomic features to improve the prediction of human m\(^1\)A RNA methylation sites. Based on the random forest algorithm, ISGm1A takes advantage of both conventional sequence features and 75 genomic characteristics to improve the prediction performance of m\(^1\)A sites in human. The results of five-fold cross validation and independent test show that ISGm1A outperforms other prediction algorithms (AUC = 0.903 and 0.909). In addition, through analyzing the importance of features, we found that the genomic features play a more important role in site prediction than the sequence features. Furthermore, with ISGm1A, we generated a high accuracy map of m\(^1\)A by predicting all adenines sites in the transcriptome. The data and the results of the study are freely accessible at: https://github.com/lianliu09/m1a_prediction.git.

INDEX TERMS Epitranscriptome, m\(^1\)A, site prediction, sequence features, genomic features.

I. INTRODUCTION
RNA epigenetic modification plays an important role in various stages of RNA life cycle, which occurs in all types of RNA. At present, more than 170 different RNA modifications have been identified. Among them, the N1-methyladenosine (m\(^1\)A) is a new epitranscriptomic modification, that is, the nitrogen at the first position of adenine is modified by a methyl group\(^1\). The study from the University of Chicago analyzed the methylation of m\(^1\)A in eukaryotic mRNA comprehensively by m\(^1\)A RNA methylation sequencing and RIP sequencing technologies, revealing that the presence of m\(^1\)A chemical modification could significantly enhance the protein translation of transcript. M\(^1\)A modification is further observed to be evolutionarily conserved and prevalent in humans, rodents and yeast, and its topological selectivity and biological mechanism on RNA expression suggest the existence of a novel layer of epigenetic regulation on RNA. Collectively, these findings could provide a new perspective for the RNA biology\(^2\). Using techniques of mA-MAP, some research identified the methylated sites of m\(^1\)A in nucleus and mitochondria RNA. Their results demonstrated that most of the methylation modification sites of m\(^1\)A were concentrated in 5’UTR of mRNA transcript, and a small portion of m\(^1\)A sites in accordance with the “GUUCA” sequence motif was created by a known methylase complex TRMT6 / 61A. The study also identified a large number of m\(^1\)A methylation modifications in mitochondrial encoded transcripts. Unlike the prevalent mRNA modification m\(^6\)A, m\(^1\)A has relatively low abundance, and it is predominantly distributed in the

The associate editor coordinating the review of this manuscript and approving it for publication was Bin Liu\(^3\).
5’UTR region of mRNA, especially in the first and second positions of transcripts starting sites. The m^1A within the 5’UTR of the mRNA transcript can promote the translation of protein, but the m^1A observed in the coding region of the transcript can lead to the inhibition of translation [3]. Additionally, a RNA repair enzyme ALKBH1 functions as the m^1A eraser by catalyzing demethylating reaction [4]. Last but not least, m^1A can affect ribosome biosynthesis [5] and mediate antibiotic resistance [6] in rRNA, and it can mediate the respond to the environmental stress in tRNA [7], [8].

With the rapid development of high-throughput sequencing technology, base-resolution technique such as miCLIP [9], m1A-seq [10], m1A-MAP [3], m1A-IP-seq [11] can map the precise location of m^1A sites at single base-resolution in a given cellular sample. However, due to the technical complexity and the high cost of the base-resolution experiments, it has not been widely used to study the m^1A epitranscriptome under different biological contexts. Nevertheless, the existing base-resolution datasets provide sufficient information for training the machine learning prediction model of RNA m^1A modification. Several computational methods have been proposed for the prediction of m^1A methylation sites. Specifically, RAMPred [12] was developed to predict the m^1A modification sites in H. sapiens, M. musculus and S. cerevisiae using SVM. The feature encoding method used by RAMPred is the 41-nt sequence based on features described by physical properties, chemical properties, and base cumulative frequency. iRNA-PseColl [13] applies the same sequence feature coding schema to predict the m^10A, m^1A and m^5C sites using machine learning classifiers. In addition, the same method was also applied to predict the m^6A, m^1A and A-I sites in the project of iRNA-3typeA [14], which targets both the Homo sapiens and Mus musculus transcriptome. Meanwhile, the prediction frameworks for RNA m^1A, where having many methods published, can be transferred to m^1A sites prediction. For example, MethyRNA [15] utilized the same coding schema as RAMPred to predict m^5A sites on human. iRNA-Methyl [16] applied the sequence feature representation method of PseDNC to capture the dinucleotide composition and physicochemical characteristics. The prediction model of iRNA-Methyl is conducted by SVM. M6AMRFS [17] construct the sequence feature by dinucleotide binary encoding and local position-specific dinucleotide frequency. The machine learning method realized in it was XGBoost. Last but important, several deep learning based methods were developed for m^6A prediction [18]–[21]. However, all of the above methods predict the RNA modification sites based on the input of the surrounding sequence of the modification sites, ignoring the topological information of the sites derived from the gene based on annotations. Therefore, WHISTLE [22] addressed this issue by predicting m^6A sites through the combination of sequence features and genomic features, and the large scale prediction was generated to describe the high accuracy map of the whole m^6A epitranscriptome.

Inspired by WHISTLE, in this paper, we propose a computational framework, the integration of sequence features and genomic features to improve the prediction of human m^1A RNA methylation sites (ISGm1A), to classify the m^1A methylation status in human transcriptome. ISGm1A predicts the m^1A methylome based on the random forest algorithm trained using base-resolution m^1A sites. The predictors used in ISGm1A involve traditional sequence features of nucleotide physiochemical properties and cumulative frequencies as well as the additional 75 features derived from hand-crafted genome annotations.

### II. METHODS

#### A. DATASETS CONSTRUCTION

**TABLE 1. Single-base resolution datasets in m^1A prediction.**

| ID | Coll  | Note             | Technique     | Source  |
|----|-------|------------------|---------------|---------|
| 1  | HEK293T |                 | miCLIP        | [9]     |
| 2  | HEK293T |                 | m1A-MAP       | [3]     |
| 3  | HEK293T | TRMT6/TRM61A overexpression | m1A-seq       | [10]    |
| 4  | HEK293T |                 | m1A-seq       | [10]    |
| 5  | HEK293T |                 | m1A-seq       | [2]     |
| 6  | HEK293T |                 | m1A-IP-seq    | [11]    |
| 7  | HEK293T |                 | m1A-quant-seq | [11]    |
| 8  | HeLa   |                 | RB5-Seq       | [23]    |
| 9  | HeLa   |                 | m1A-seq       | [2]     |
| 10 | HepG2  | heat_shock      | m1A-seq       | [2]     |
| 11 | HepG2  | Glucose_starvation_4h | m1A-seq       | [2]     |
| 12 | HepG2  |                 | m1A-seq       | [2]     |

For predicting the m^1A methylation sites in Homo sapiens, we collected the single based m^1A sites from 12 datasets of three cell types, including HEK293T, HeLa and HepG2, where there are seven datasets from HEK293T, two from HeLa and three from HepG2 (see Table 1). The positive m^1A sites were defined as the union of m^1A sites from 12 datasets. The negative m^1A sites were determined by taking the same number of random adenine sites on the exon regions containing positive samples. In addition, no sites are reported from the ambiguous regions which could be mapped to multiple genes. Finally, 39104 m^1A sites were collected, including 19552 positive sites and 19552 negative sites. We randomly selected four fifths of the sites for training, and the rest were retained for testing.

#### B. FEATURE REPRESENTATION

In this work, two types of features, sequence features and genomic features, were applied to represent m^1A sites.

1) **SEQUENCE FEATURES**

Each base in the adenine centered 41 sequence is represented by a four-dimensional vector using the same method RAMPred [12]. An mRNA sequence is consisted by four kinds of nucleotides, including adenine (A), guanine (G), cytosine (C), and uracil (U). In sequence feature representation, three features firstly were used to represent a nucleotide...
According to different structural properties, namely ring number, hydrogen bonds and chemical functions. Among the four kinds of nucleotides, purine has two rings, while pyrimidine has only one ring, so adenine and guanine have only one ring, while cytosine and uracil have two rings. Although RNA is a single stranded molecule, the biological function of RNA is closely related to the internal hybridization captured by RNA secondary structure. When the secondary structure is formed, guanine and cytosine form at most three hydrogen bonds, while adenine and uracil form at most two hydrogen bonds. Therefore, according to the strength of hydrogen bonds, guanine and cytosine have strong hydrogen bonds, while adenine and uracil have weak hydrogen bonds. In addition, adenine and cytosine can be divided into amino group, and guanine and uracil into ketone group according to their chemical functions. Through this definition, three chemical properties of each base can be obtained. The three chemical properties are defined as follows:

\[
x = \begin{cases} 
1 & \text{if } s \in \{A, G\} \\
0 & \text{if } s \in \{C, U\}
\end{cases}, \quad y = \begin{cases} 
1 & \text{if } s \in \{A, C\} \\
0 & \text{if } s \in \{G, U\}
\end{cases},
\]

\[
z = \begin{cases} 
1 & \text{if } s \in \{A, U\} \\
0 & \text{if } s \in \{C, G\}
\end{cases}
\]  

(1)

According to the above definition of chemical properties, A, G, C and U can be coded as A = (1,1,1), G = (1,0,0), C = (0,1,0), U = (0,1,0), respectively. In addition, in order to describe the distribution of bases in the sequence, the base frequency accumulation value was added to the sequence features as the fourth feature, that is, the frequency of the occurrence of the \(i\)-th base before \(i\) position. The density \(f_i\) of the \(i\)-th base is defined as follows:

\[
f_i = \frac{d_i}{i}
\]

(2)

where \(d_i\) is defined as the sum number of occurrences of the \(i\)-th base in the previous \(i\) bases. For example, for a sequence with a base distribution of “GAAUCCUGGA”, G appears in the first, eighth and ninth positions respectively, so the cumulative frequencies of A are 1/2, 2/3, 3/10. Based on this principle, we can calculate the cumulative value of frequency in each base in the input nucleotide sequence. According to the previous described physiochemical properties and frequency accumulation characteristics, each base in the sequence can be represented by a four-dimensional vector.

2) GENOMIC FEATURES

At present, most of the m\(^1\)A prediction methods only use sequence features, but the nucleotide sequence features can only represent the characteristics for each base in the sequence, not the topological characteristics of RNA methylation sites within the transcript annotation. Therefore, 75 additional genomic features were generated to explain the topological and transcriptomic characteristic of m\(^1\)A sites. Among these features, genomic features 1-17 are the dummy variables, which indicate whether the site is overlapped to the topological region on the major RNA transcript. Features 18-23 represent the relative position on the regions, and the value of features 18-23 is real on \([0,1]\). Genomic features 24-25 stand for the nucleotide distances toward the splicing junctions. Features 26-33 represent the length of transcript region containing the m\(^1\)A methylation sites. Features 34-51 indicate which motif the m\(^1\)A site belongs to. Genomic Features 52-55 are clustering indicators or motif clustering, which can reflect the clustering characteristics of the methylation site. Features 56-59 are the distance to the neighboring sites or the nearest A motif. Genomic features 60-63 are the scores related to evolutionary conservation, including phast cons (PC) score, average PC score, fitness consequence (FC) score and average FC score. Features 64-65 indicate RNA secondary structures information by RNAfold [24]. Feature 66 stands for RNA annotation related to m\(^1\)A biology. Genomic features 67-71 are the dummy variables, which indicate whether the site is overlapped with snRNA, lncRNA, lincRNA, housekeeping genes and microRNA. The z-score of transcript isomform and exon number are described by features 72-73. Genomic features 74-75 contain z-scores of GC content of genes and 101bp regions centered by the m\(^1\)A sites. The genomic features we considered are recorded in detail in Supplementary Table S1.

C. EVALUATION METRICS

To further measure the performance of the model, four kinds of performance metrics, namely Sn (sensitivity), Sp (specificity), ACC (accuracy), MCC (Matthews’s correlation coefficient), are applied. Among these metrics, the Sn indicates the accuracy of the model in predicting the positive samples. A higher Sn indicates the higher performance in predicting the positive samples. Meanwhile, the higher Sp indicates the higher performance in predicting the negative samples. The ACC represents the success rate of true positive and true negative samples prediction. A good prediction model should have high Sn and Sp at the same time. If the Sn is very high and the Sp is low, it will produce a high false-positive, while if the Sp is very high and the Sn is low, it will produce a high false-negative. Additionally, the MCC can reflect the reliability of the results which is robust to the imbalance of sample proportions. The four indicators are defined as follows:

\[
S_n = \frac{TP}{TP + FN}
\]

(3)

\[
S_p = \frac{TN}{TN + FP}
\]

(4)

\[
ACC = \frac{TP + TN}{TN + FP + TP + FN}
\]

(5)

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

(6)

where TP, TN, FP, FN indicate the true positive, true negative, false positive and false negative values, respectively. In addition, the Receiver Operating Characteristic (ROC) curves are
plotted, and the areas under the curves (as called “AUC”) are calculated for evaluating the performance of the prediction model.

III. RESULTS AND DISCUSSION

A. COMPARISON OF FIVE CLASSIFIERS THROUGH CROSS-VALIDATION

In order to compare the prediction results of different classifiers, we used five common machine learning classifiers, which are random forest (RF) [25], support vector machine (SVM) [14], K-nearest neighbor (KNN) [26], logistic regression (LR) [27] and eXtreme Gradient Boosting (XGBoost) [28]. RF is a widely used tree-based machine learning algorithm, which is applied by SRAMP [29] to predict the m6A sites in mammals. SVM is another machine learning method widely applied in the many prediction models of computational biology, base on which, iRNA-TypeA, RAMPred and iRNA-PseColl realized the prediction of m1A sites in human, mouse and yeast. KNN is one of the most basic and simple algorithms in machine learning algorithm. It can be used for both classification and regression. KNN is classified by measuring the distance between different eigenvalues. LR is another fundamental classification method in machine learning, which is a linear model with high computational efficiency. XGBoost [28] is widely used in data science competition and industry, which can be effectively applied to classification, regression and sorting problems for structured data. Both M6AMRFS [17] and HMpre [30] predicted the modification sites based on XGBoost. In order to compare the performance of five classifiers, a five-fold cross validation was applied to the training data, and the best classifier would be employed in the m1A prediction. The performance of the different classifiers in cross validation is shown in Table 2 and Fig.1, which exhibits that RF has the best performance in five-fold cross validation with AUC = 0.903 except in Sp.

B. COMPARISON OF FIVE CLASSIFIERS IN INDEPENDENT TEST

Next, we compared the prediction results of five classifiers in independent test. Unsurprisingly, RF also achieved the best performance in independent tests (see Table 3). In addition, the ROC curves were plotted in Fig.2. Based on the result of cross validation and independent test, we selected RF as the final model used in the m1A sites prediction.

C. FEATURE SELECTION

In order to obtain the most effective prediction results, we used feature selection to define the most effective feature subset. First, we ranked the importance of features in the five-fold cross validation, which is calculated by RF in R package. Then, we added one feature to the feature set each time according to the sorted feature set, and calculated the AUC in the five-fold cross validation. Finally, the optimal feature subset with the highest AUC was obtained. As shown in Fig.3A, the top 5 most important features are the distance from site to the downstream (3' end) splicing junction (dist_3sj_2k), whether the site is overlapped with internal exons (inter_exon), the length of the CDS overlapped by the site (len_CDS), the length of the gene overlapped by the site (len_gene_full), and the distance to the nearest neighboring sites (maximum to 200bp) (ndist_m2h), which all belong to the genomic features. It suggests that the genomic features play a more important role in the prediction of m1A sites.
Furthermore, we found that the highest AUC can be obtained when all features are used (See Fig. 3B). This may be due to the fact that the selected features are independent of the data.

Then we used different types of features, namely sequence features (Sequence), genomic features (Genomic) and the combination of sequence features and genomic features (Sequence+Genomic) to predict m\textsuperscript{1}A sites, respectively. The performance was summarized in Fig. 4. As it is demonstrated by Fig. 4, the best prediction performance can be achieved when using the combination of sequence features and genomic features. Meanwhile, when using only a single type of feature, the results for genomic features are much better than sequence features. This observation indicates that the genomic features are in general more effective than sequence features in m\textsuperscript{1}A sites prediction.

Although genomic features can get better results than sequence features when the single type of features was used, the integration of genomic features and sequence features can further improve the performance of the prediction. Therefore, we used the integration of two kinds of features to predict m\textsuperscript{1}A sites.

**D. COMPARISON WITH OTHER METHODS**

In order to further verify the effectiveness of the proposed algorithm, we compared our model with the existing m\textsuperscript{1}A prediction methods. Because the same features and classifiers were applied in RAMPred, iRNA-3typeA, iRNA-PseColl for prediction, we only choose RAMPred for comparison. The results were summarized in Table 4 and Fig. 5 by showing the ROC curves. It can be seen that compared with the existing methods, ISGm1A achieved superior performance for the prediction of m\textsuperscript{1}A sites on RNA.

**E. TRANSCRIPTOME-WIDE M1A SITES PREDICTION**

In order to generate the complete map of m\textsuperscript{1}A methylation sites in human with our prediction model, we searched the entire exon regions of the transcriptome and use all adenine sites as the candidates. The probabilities for m\textsuperscript{1}A positive are estimated with the proposed model for all the candidate sites. Finally, 147,762 out of the total 23,138,573 adenines were predicted as m\textsuperscript{1}A RNA methylation sites according to the probability greater than 0.8. The complete prediction results are freely accessed at: https://github.com/lianliu09/m1a_prediction.git.
IV. CONCLUSION
As a recently revealed internal mRNA modification, m\textsubscript{1}A serves as a significant layer of information in gene expression regulation. With the development of high-throughput sequencing technology, people can measure and predict the mRNA methylation status with high accuracy in the whole transcriptome range. Few studies have previously established the prediction methods for the RNA modification status of m\textsubscript{1}A in transcriptome.

In this paper, we proposed ISGm1A, a novel framework to predict m\textsubscript{1}A sites by integrating sequence features and genomic features. ISGm1A firstly extracted the the physiochemical properties and the cumulative frequency of bases to capture the information of the flanking sequence of m\textsubscript{1}A sites. Additionally, ISGm1A crafted 75 additional domain knowledge based genomic features to enhance the predictive power of m\textsubscript{1}A sites. Using random forest classifier, ISGm1A achieved superior performance under both five-fold cross validation and independent test compared of currently published methods. Additionally, with the trained prediction model, we scanned all adenosine sites on exon regions and predicted 147,762 candidate m\textsubscript{1}A sites in human. In conclusion, we reason that the prediction framework can be transferred to predict the m\textsubscript{1}A methylation sites for species other than human, as long as the genome is accompanied with high quality gene annotations.

REFERENCES
[1] D. B. Dunn, “The occurrence of 1-methyladenine in ribonucleic acid,” Biochimica Biophys. Acta, vol. 46, no. 1, pp. 198–200, Jan. 1961.
[2] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M. S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, and W. C. Clark, “The dynamic N\textsuperscript{6}-methyladenosine methylation in eukaryotic messenger RNA,” Nature, vol. 530, no. 7591, pp. 441–446, 2004.
[3] X. Li, X. Xiong, M. Zhang, K. Wang, Y. Chen, J. Zhou, Y. Mao, J. Li, D. Yi, X.-W. Chen, C. Wang, S.-B. Qian, and C. Yi, “Base-resolution mapping reveals distinct m\textsubscript{1}A methylome in nuclear- and mitochondrial-encoded transcripts,” Mol. Cell, vol. 68, no. 5, pp. 993–1005, Dec. 2017.
[4] F. Liu, W. Chen, G. Luo, X. Wang, Y. Fu, J. Wei, X. Wang, Z. Hao, Q. Dai, G. Zheng, H. Ma, D. Han, M. Evans, A. Klungland, T. Pan, and C. He, “ALKBH1-mediated RNA demethylation regulates translation,” Cell, vol. 167, no. 3, pp. 816–828, Oct. 2016.
[5] C. Peifer, S. Sharma, P. Watzinger, S. Bertham, P. Köttér, and K.-D. Entian, “YesrCPR8p, a novel methyltransferase responsible for m\textsubscript{1}A 645 base modification of 25S rRNA,” Nucleic Acids Res., vol. 41, no. 2, pp. 1191–1196, Jun. 2013.
[6] J. P. Ballesta and E. Cundiff, “Site-specific methylation of 16S rRNA caused by PCT, a Pectaminycin resistance determinant from the producing organism, streptomyces pactum,” J. Bacteriol., vol. 173, no. 22, pp. 7213–7218, 1991.
[7] C. T. Y. Chan, M. Dyavaiah, M. S. DeMott, K. Taghizadeh, P. C. Dedon, and T. J. Begley, “A quantitative systems approach reveals dynamic control of RNA modifications during cellular stress,” PLoS Genet., vol. 6, no. 12, 2010, Art. no. e1001247.
[8] M. Helm and J. D. Alfonzo, “Posttranscriptional RNA modifications: Playing metabolic games in a Cell’s chemical legoland,” Chem. Biol., vol. 21, no. 2, pp. 174–185, Feb. 2014.
[9] A. V. Grozhik, A. O. Olarerin-George, M. Sindelar, X. Li, S. S. Gross, and S. R. Jaffrey, “Antibody cross-reactivity accounts for widespread appearance of m\textsubscript{6}A in 5’UTRs,” Nature Commun., vol. 10, no. 1, pp. 1–13, Dec. 2019.
[10] M. Safrä, A. Sas-Chen, R. Nir, R. Winkler, A. Nachshon, D. Bar-Yaakov, M. Erlicher, W. Rossmanith, N. Stern-Ginossar, and S. Schwartz, “The m\textsubscript{6}A landscape on cytosolic and mitochondrial mRNA at single-base resolution,” Nature, vol. 551, no. 7679, pp. 251–255, 2017.
[11] H. Zhou, S. Rauch, Q. Dai, X. Cui, Z. Zhang, S. Nachtergaele, C. Sepich, C. He, and B. C. Dickinson, “Evolution of a reverse transcriptase to map N\textsuperscript{6}-methyladenosine in human messenger RNA,” Nature Methods, vol. 16, no. 12, pp. 1281–1288, Dec. 2019.
[12] W. Chen, P. Feng, H. Tang, H. Ding, and H. Lin, “RAMPred: Identifying the N\textsuperscript{6}-methyladenosine sites in eukaryotic transcriptomes,” Sci. Rep., vol. 6, no. 1, Aug. 2016, Art. no. 31080.
[13] W. Chen, H. Tan, and H. Lin, “MethyRNA: A Web server for identification of N6-methyladenosine sites,” J. Biomol. Struct. Dyn., vol. 35, no. 3, pp. 683–687, Feb. 2017.
[14] W. Chen, P. Feng, H. Yang, H. Ding, and H. Lin, “IRNA-TypeA: Identifying three types of modification at RNA’s adenosine sites,” Mol. Therapy Nucleic Acids, vol. 11, pp. 468–474, Jun. 2018.
[15] W. Chen, H. Tang, and H. Lin, “MethyRNA: A Web server for identification of N6-methyladenosine sites,” J. Biomol. Struct. Dyn., vol. 35, no. 3, pp. 683–687, Feb. 2017.
[16] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, “IRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC,” Mol. Therapy Nucleic Acids, vol. 7, pp. 155–163, Jun. 2017.
[17] W. Chen, P. Feng, H. Yang, H. Ding, and H. Lin, “IRNA-TypeA: Identifying three types of modification at RNA’s adenosine sites,” Mol. Therapy Nucleic Acids, vol. 11, pp. 468–474, Jun. 2018.
[18] Y. Huang, N. He, Y. Chen, Z. Chen, and L. Li, “BERMP: A cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach,” Int. J. Biol. Sci., vol. 14, no. 12, pp. 1669–1677, 2018.
[19] Z. Wei, H. Zhou, S. Rauch, Q. Dai, X. Cui, Z. Zhang, S. Nachtergaele, C. Sepich, C. He, and B. C. Dickinson, “Evolution of a reverse transcriptase to map N\textsuperscript{6}-methyladenosine in human messenger RNA,” Nature, vol. 551, no. 7679, pp. 251–255, 2017.
[20] H. Zhou, S. Rauch, Q. Dai, X. Cui, Z. Zhang, S. Nachtergaele, C. Sepich, C. He, and B. C. Dickinson, “Evolution of a reverse transcriptase to map N\textsuperscript{6}-methyladenosine in human messenger RNA,” Nature, vol. 551, no. 7679, pp. 251–255, 2017.
LIAN LIU received the Ph.D. degree in pattern recognition and intelligent system from Northwestern Polytechnical University, Xi’an, China, in 2018. She is currently a Postdoctoral Student at Shaanxi Normal University, Xi’an, where she is also an Assistant Research Fellow with the School of Computer Science. Her current research interests include bioinformatics, pattern recognition, and machine learning.

LIAN LIU

XIUJUAN LEI (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, Xi’an, China, in 2005. She was a Visiting Scholar with the Department of Computer Science and Engineering, State University of New York at Buffalo, USA, from 2009 to 2010. She is currently a Professor with the School of Computer Science, Shaanxi Normal University, Xi’an. Her current research interests include bioinformatics, intelligent computing, pattern recognition, and data mining.

XIUJUAN LEI

JIA MENG received the Ph.D. degree in electrical engineering from The University of Texas at San Antonio, in 2011. From 2012 to 2014, he served as an Associate Scientist for the Broad Institute of MIT and Harvard. He is currently a Senior Associate Professor at the Department of Biological Sciences, Xi’an Jiaotong-Liverpool University. He has previously worked on a wide variety of computational biology projects that aim at a system level understanding of gene regulation and the integration of multiple high-throughput data types and databases with advanced multivariate techniques, such as Bayesian generative modeling, sparse representation, factorization, and nonparametric approaches.

JIA MENG

ZHEN WEI received the B.S. degree in biological science from Xi’an Jiaotong Liverpool University, China, in 2015, where he is currently pursuing the Ph.D. degree in bioinformatics. He is also a Lecturer of bioinformatics with the Department of Biological Science, Xi’an Jiaotong Liverpool University. His current research interests include normalization methods in HTP sequencing data and feature engineering in genomic prediction models.

ZHEN WEI