Neurocomputing 422 (2021) 214–221

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A Convolutional Neural Network Using Dinucleotide One-hot Encoder for identifying DNA N6-Methyladenine Sites in the Rice Genome

Zhibin Lv [a], Hui Ding [b], Lei Wang [c], Quan Zou [a,b,d,*]

[a] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China
[b] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China
[c] College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, China
[d] Hainan Key Laboratory for Computational Science and Application, Hainan Normal University, Haikou, China

ARTICLE INFO

Article history:
Received 13 January 2020
Revised 28 May 2020
Accepted 2 September 2020
Available online 14 October 2020

Keywords:
convolution neutral network
deep learning
rice
epigenetic
N6-methyladenine

ABSTRACT

N6-methyladenine (m^6A) is one of the crucial epigenetic modifications and is related to the control of various DNA processes. Carrying out a genome-wide m^6A analysis via wet experiments is fundamental but takes a long time. As complementary methods, computing tools, especially those based on machine learning, are urgently needed. A new protocol, iRicem6A-CNN, for identifying m^6A sites in the rice genome was developed. This protocol was designed to use dinucleotide one-hot encoding to generate input tensors for predictions by convolutional neutral networks, and achieved five-fold cross-validation and independent testing accuracy values of 93.82% and 96.19%, respectively, performing better than those of other available predictors. The experiment results demonstrates that only the ability of iRicem6A-CNN based on 2-mer one-hot encoding is to display high performance but also to perform more stably and robustly than models using 1-mer one-hot encoding. A webserver is accessible at http://iRicem6A-CNN.aibiochem.net

1. Introduction

As the product of a critical chemical modification of DNA, N6-methyladenine (m^6A) is found ubiquitously in various organisms ranging from eukaryotes to prokaryotes [1] and is evidently involved in DNA replication, DNA repair and transcription regulation [2–5]. Genomic analysis of DNA methylation has emerged as the next-generation sequencing technology, and especially single-molecule real-time (SMRT) sequencing detection has shown increasing application [6–8]. The genome-wide distribution of N6-methyladenine sites has become better characterized, which has led to a better understanding of its biological functions [9–12]. For example, genome-wide m^6A sites studies have revealed different regulatory functions of m^6A in different eukaryotes, and have shown m^6A in prokaryotes functioning as a marker to distinguish invading foreign DNA from host DNA [13].

In 2018, Zhou et al. [13] used SMRT to show that about 0.2% of all adenines in the rice genome are m^6A methylated, and since this discovery, all sorts of machine-learning-based computing methods for m^6A in the rice genome have been springing up. In 2019, Chen et al. [11] developed im6A-Pred, a support vector machine (SVM)-based method, that was trained on a benchmark dataset consisting of 1760 samples and achieved an accuracy of 83.13%. Subsequently, other methods based on the strengths of traditional machine learning algorithms like SVM, random forest (RF) and Markov chain models were developed for identifying m^6A sites in the rice genome; these methods included im6A-DNCP [14], MM-m6Apred [15] , SDM6A [16], iN6-methylat [17] and iDNAm6A-rice [18], and of them, iDNAm6A-rice showed the best accuracy at 91.7%.

Deep learning is a sub-discipline of machine learning and has shown great effectiveness in vision computing, natural language processing and even genomics computational modeling [19–24]. The major difference between deep learning and other machine learning methods is that deep learning does not need to handcraft features before modeling; instead it automatically learns the features embedded in the input data and could discover high-level abstract features to yield end-to-end models itself [25]. Although deep learning can generally achieve a better predictive performance than can the other traditional methods, it is not perfect. Traditional machine learning methods involve direct feature engineering and are easy to interpret and understand. They are also easier to adjust the hyperparameters and change the model design for traditional machine learning [1,26–31]. In contrast, the

[*] Corresponding author.
E-mail address: zouquan@nclab.net (Q. Zou).

https://doi.org/10.1016/j.neucom.2020.09.056
0925-2312/© 2020 Elsevier B.V. All rights reserved.

deep learning model is usually a "black box" and is hard to interpret. Moreover, deep learning involves the use of a data-driven model demanding large amounts of input data for training and a relatively high computational capacity [32]. Nevertheless, given the above-described advantages of deep learning, Alipanahi et al. [33] in 2015 set out to demonstrate the application of deep learning to DNA sequence analysis, and were the first to do so successfully. So since then, deep learning has been rapidly and increasingly applied to DNA, RNA and protein data mining [33–47]. In 2019, Tahir et al. [48] used a deep learning approach named iDNAm6A for predicting m^6A sites in the rice genome and obtained an accuracy of ~87%. Then Yu et al. [49] developed another deep-learning-based rice m^6A predictor, SNNRicem6A, and achieved with it a 5-fold cross-validation accuracy of 92.04%, making it one of the top rice m^6A predictors. Both methods adopted a single-nucleotide one-hot encoder and convolution neural network (CNN) for modeling.

In this study, we developed a new protocol, called iRicem6A-CNN, to improve the accuracy of predicting rice genome-wide m^6A sites. Distinguishing from reported protocols, our protocol employed a dinucleotide one-hot encoder to transfer DNA sequences into tensors, which were then input into the well-designed CNN for model optimization. It realized a 5-fold cross-validation accuracy of 93.82% and an independent testing accuracy of 96.19%, by values 1.97% and 12.5% better than those of the previously considered the best-performing predictor SNNRicem6A. Evidences indicate iRicem6A-CNN using 2-mer one-hot encoder is more robust and accurate than that utilized 1-mer one-hot encoding. The metrics comparison showed that the better performance of iRicem6A-CNN was not only due to its stable ability to recognize positive samples but also to its more precise identification of negative samples. A webserver for this protocol is available at http://iRicem6A-CNN.aibiochem.net.

2. Methods and materials

2.1. Datasets

Two widely used m^6A benchmark datasets for the rice genome were set up by Chen [11] and by Lv [18] separately, and were labeled as Chen-rice-m6A and Lv-rice-m6A. The Chen-m6A-rice dataset, consisting of 1760 samples with half of them being positive and other samples negative, has been extensively utilized by reported models based on non-deep-learning algorithms. The Lv-m6A-rice dataset was made of 154000 DNA sequences with m^6A sites and 154000 sequences without m6A sites, and has been adopted by Lv for iDNAm6A-rice [18] and by Yu for SNNRicem6A [49]. Sequences in both datasets were chosen to be 41 bp in length and with each having an adenine at the center position. Considering the data-hungry nature of the CNN model, we took the Lv-m6A-rice dataset for model training and the Chen-m6A-rice dataset for independent testing as doing in ref [18] for convenient comparison.

2.2. Data representation

Of the reported methods, some such as iDNAm6A-rice [18], SDM6A [16] and im6A-pred [11] involved the manual conversion of the sequence data into feature vectors via several predefined features, including but not limited to pseudo-nucleic acid composition, accumulated nucleotide frequency, nucleotide chemical property, etc. Instead of requiring prior knowledge of features to be extracted from the sequences, methods based on deep learning such as iDNAm6A [48] and SNNRicem6A [49] could automatically extract features from encoded DNA sequences. iDNAm6A and SNNRicem6A both applied one-hot encoding of A,C,G,T to convert DNA sequences into a matrix for feeding into a following CNN. Generally, the DNA kmer one-hot encoding process is shown in Figure 1 using AATTCAG as instance. Assuming that the length of the DNA sequence is L, and a windows with kmer sliding along with the sequence, then it will get L-k+1 kmer nucleotides and the i^{th} kmer $s_i$will be encoded as following:

$$s_i = [0, \cdots, 0, 1, 0, \cdots, 0] \tag{1}$$

The length of the vector$s_i$is 4^k, and the value of i^{th} position is 1 and values of the other positions are 0. As a result, a DNA sequence converts into a $R^{(L-k+1)\times 2^k}$ matrix composing of vectors like s_i.

As comparison to iDNAm6A and SNNRicem6A, in this study, we used one-hot encoding methods, but here to code dinucleotides (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT). A window two nucleotides long was made to slide along a DNA sequence and each dinucleotide in the window was encoded into a one-hot encoding vector with 16 dimensions, and thus a DNA sequence with 41 nucleotides was transformed into a matrix with 40 rows and 16 columns. Each such matrix was further turned into a tensor fed into the built CNN model for training. The single nucleotides (A, C, G, T) one-hot encoding methods are also used for comparison.

2.3. Model construction

Figure 2 shows the framework of iRicem6A-CNN. First, the input layer is prepared. The DNA sequences were encoded via a 2mer one-hot coding method into matrixes each with 40 rows and 16 columns. Second, the matrixes were fed into a CNN. Here, the CNN consisted of one 1D convolutional layer with 200 filters and a rectified linear unit (ReLU) as the activation function. The filter kernel size was 8. After carrying out the convolution operation, a maxpooling layer was added with a pooling size of 4. A dropout layer was used next to the pooling layer, with a dropout rate of 0.16. Third, the intermediate feature units were flattened to fully connect them to the hidden dense layer. Finally, it is the output layer with 2 units fully connected to the previous layer. The softmax function was used for predicting whether or not a specified DNA sequence would contain a m^6A site: output probabilities greater than and less than a given threshold value were considered to indicate, respectively, the presence and absence of a m^6A site in the input sequence. The given threshold values include 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99. When the model was training, it compiled an adaptive moment estimation optimizer with a learning rate of 0.001, and categorical cross-entropy as a loss function.

2.4. Evaluation metrics and methods

Accuracy (ACC), sensitivity (Sn), specificity (Sp) and Matthew's correlation coefficient (MCC) are commonly and generally used valuation metrics for machine learning binary classification models [13,50–71]. These four metrics were calculated using the equations shown in supplementary materials Part 1. The receiver operating characteristic (ROC) curve and the area under ROC (auROC) were also used for model evaluation [72]. Three performance evaluation protocols are commonly applied in evaluating bioinformatics predictors [18,49,73–75]. They are resubstituting test, cross-validation and independent dataset test. K-Fold validation and jackknife test are two types of cross-validation methods. In this paper, we employed the 5-fold cross-validation and independent testing method to test the efficiency of the classification for comparison. The details please see the supplementary materials Part 2.

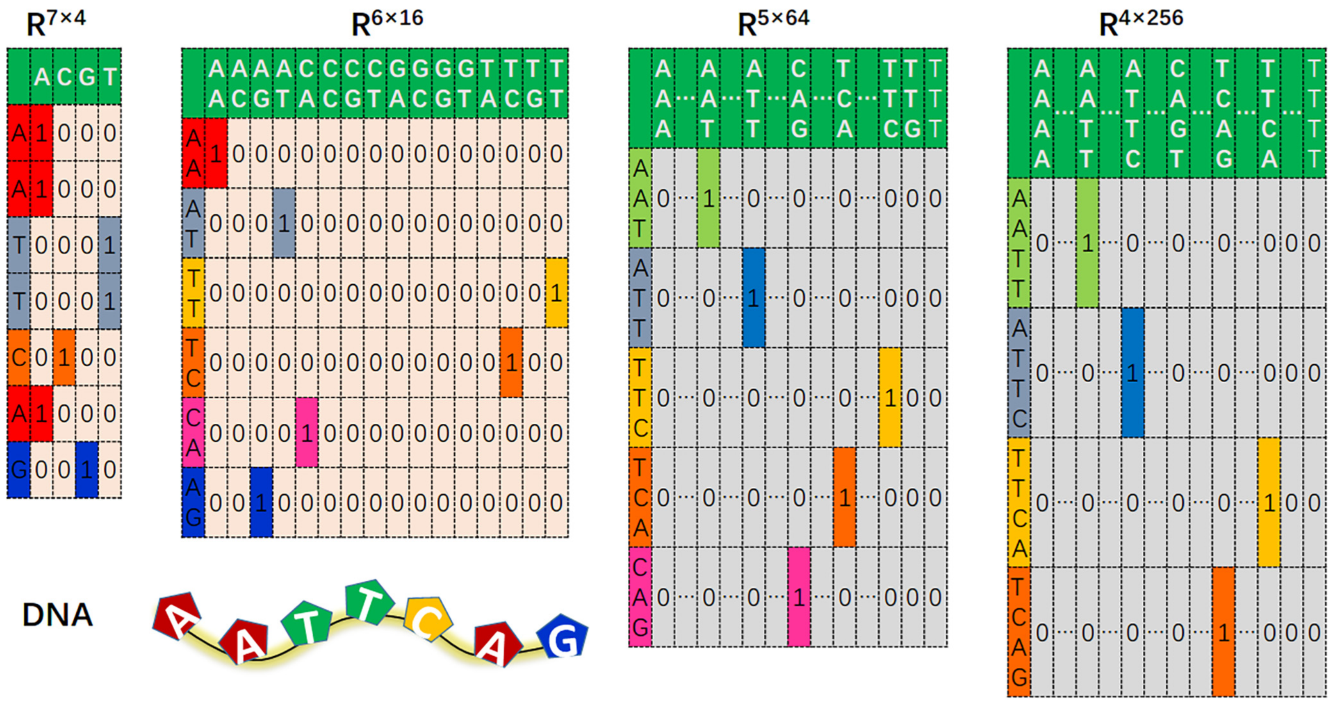


Figure 1. The schematic diagram of 1mer, 2mer, 3mer and 4mer one-hot encoding for the DNA sequence, taking AATTCAG as example. The AATCAG is converted into a 0-1R$^{m\times2^k}$ matrix, while m= sequence length-k+1 and k=1, 2, 3,4.

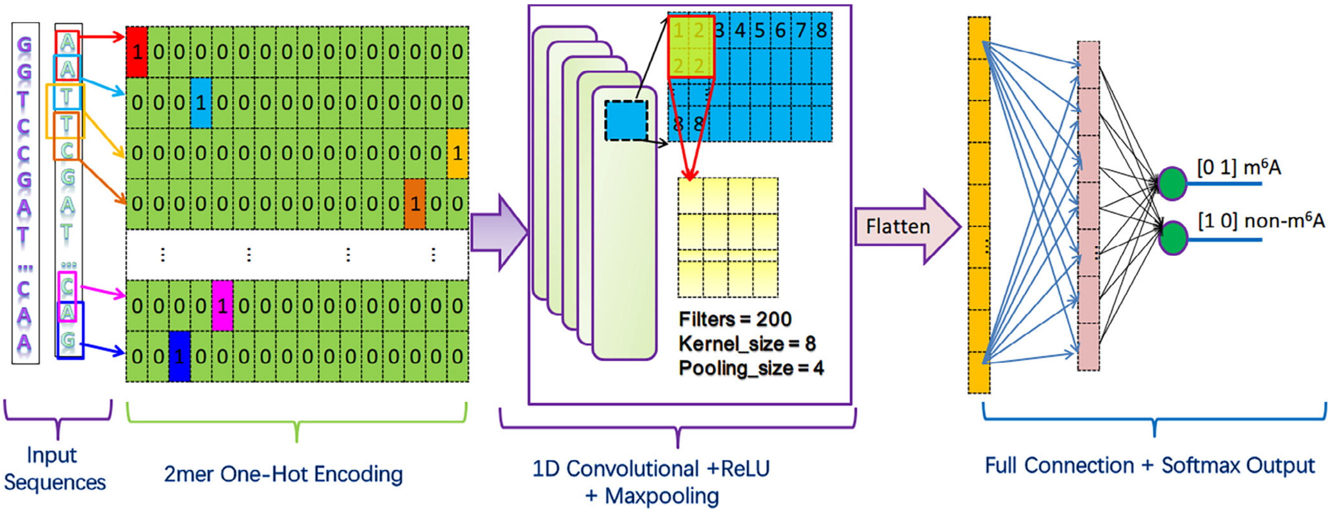


Figure 2. The architecture of the iRicem6A-CNN model used 2mer one-hot encoding.

3. Results and discussion

3.1. Comparison of models based on the different encoder

We constructed two models. One is based on 1-mer encoding, i.e. iRicem6A-CNN (1mer) and the other one is based on 2-mer encoding, i.e. iRicem6A-CNN (2mer). As shown in Figure 3A, the final evaluation results of the 5-fold cross-validation [76–81] for iRicem6A-CNN (1mer) were found to be ACC = 93.41%, MCC = 0.868, auROC = 0.975, Sn = 94.10%, Sp = 92.72%. While the 5-fold cross-validation score of iRicem6A-CNN (2mer) were that ACC = 93.82%, MCC = 0.877, auROC = 0.979, Sn = 94.34%, Sp = 93.31%, which increased by value of 0.4%, 1.0%, 0.4%, 0.3% and 0.6% over those of iRicem6A-CNN (1mer). In the view of model training cross-validation results, iRicem6A-CNN (2mer) has only a slight advantage in performance over the iRicem6A-CNN (1mer). In consideration of independent testing, iRicem6A-CNN (2mer) shows better prediction accuracy and more robustness than iRicem6A-CNN (1mer). For 3mer or 4mer one-hot encoding CNN models, their overall five-fold cross-validation metrics are inferior to iRicem6A-CNN (2mer)(see supplementary materials Figure 1. Also, if kmer≥3, the required memory for computing grew exponentially (supplementary materials Figure 2). Thus it is not a wise strategy to use 3mer or even more for one-hot encoding if not to take steps to reduce memory requirements.

Figure 3B and supplementary materials Table 1 display independent testing scores of both models under various predictions probability threshold. Overall, iRicem6A-CNN (2mer) shows higher scores at every predictions probability threshold. As predictions probability threshold is increased from 0.50 to 0.99, the

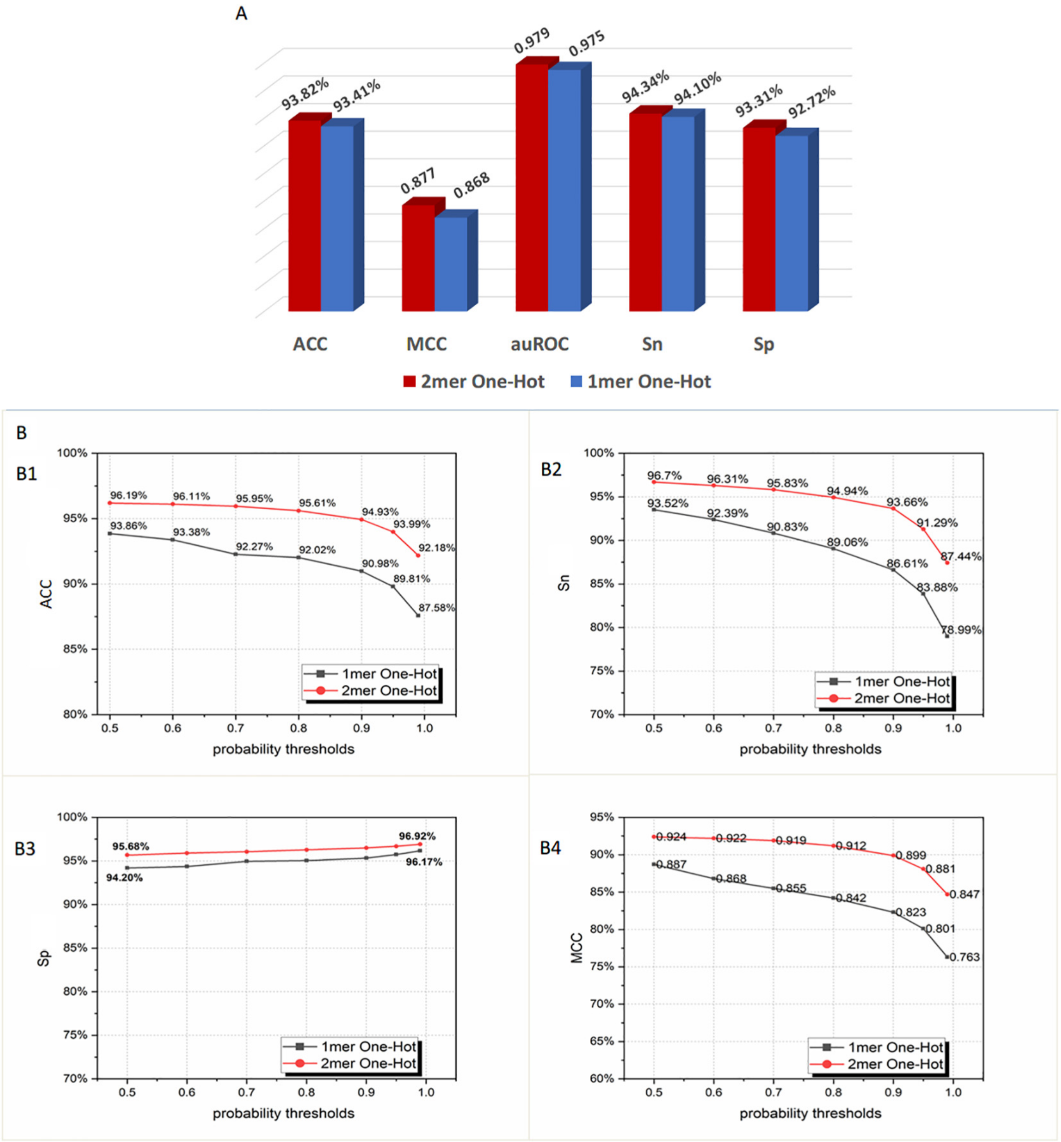


Figure 3. A) Five-fold cross validation and B) Independent testing results for models with different encoding methods under various prediction probability thresholds.

independent accuracy value of iRicem6A-CNN (2mer) declines from 96.19% to 92.18%, decreasing by 4.16%, while the value of iRicem6A-CNN (1mer) goes down from 93.86% to 87.58%, reducing by 6.69%. The sharper reduction of independent testing accuracy of iRicem6A-CNN (1mer) is mainly due to its lower positive samples identification ability as predictions probability threshold rising as displayed in Figure 3**B**. For instance, when predictions probability threshold increases from 0.5 to 0.99, the 15.5% declining magnitude of Sn for iRicem6A-CNN(1mer) is nearly 1.62 times of the Sn value (9.57%) of iRicem6A-CNN(2mer). These results provided evidence for not only the ability of iRicem6A-CNN(2mer) to display high performance but also to perform stably and robustly. It means that convolution neutral network using 2-mer one-hot encoding in rice DNA m6A analysis could automatically learn more effective latent features for positive samples than that based 1-mer one-hot encoding.

3.2. Comparison with previously reported models

Multiple predictors for rice DNA m^{6}A sites have been implemented and their cross-validation scores and independent testing results are **listed in supplementary materials Table 2 and 3** respectively. The predictions probability threshold is 0.5, which is widely used in reported predictors.

As shown in Figure 4A, in comparison to the previously considered best-performing predictor, i.e., SNNRicem6A, iRicem6A-CNN showed increased values of the ACC, MCC, auROC and Sp metrics, by 1.93%, 4.40%, 0.92% and 3.97%, respectively. With almost the same value of Sn, the Sp improvement showed that iRicem6A-CNN could predict negative samples more correctly than that SNNRicem6A did, which led to better accuracy performance of iRicem6A-CNN.

As shown in Figure 4B, we further comparatively analyzed the independent testing scores of three models trained on the Lv-m6A-rice dataset. The scores of iDNAm6A-rice were obtained from reference [18]. The data for SNNRicem6A were obtained from the model yielding the largest auROC value of the five cross-validation models according to reference [49], the values of which were computed via the source code offered in reference [49]. Evidently, although the cross-validation results of SNNRicem6A were not found to be significantly different from the accuracy of iRicem6A-CNN, the independent testing metrics values of iRicem6A-CNN were found to be considerably greater than those of SNNRicem6A. The ACC, MCC, auROC, Sn and Sp values for

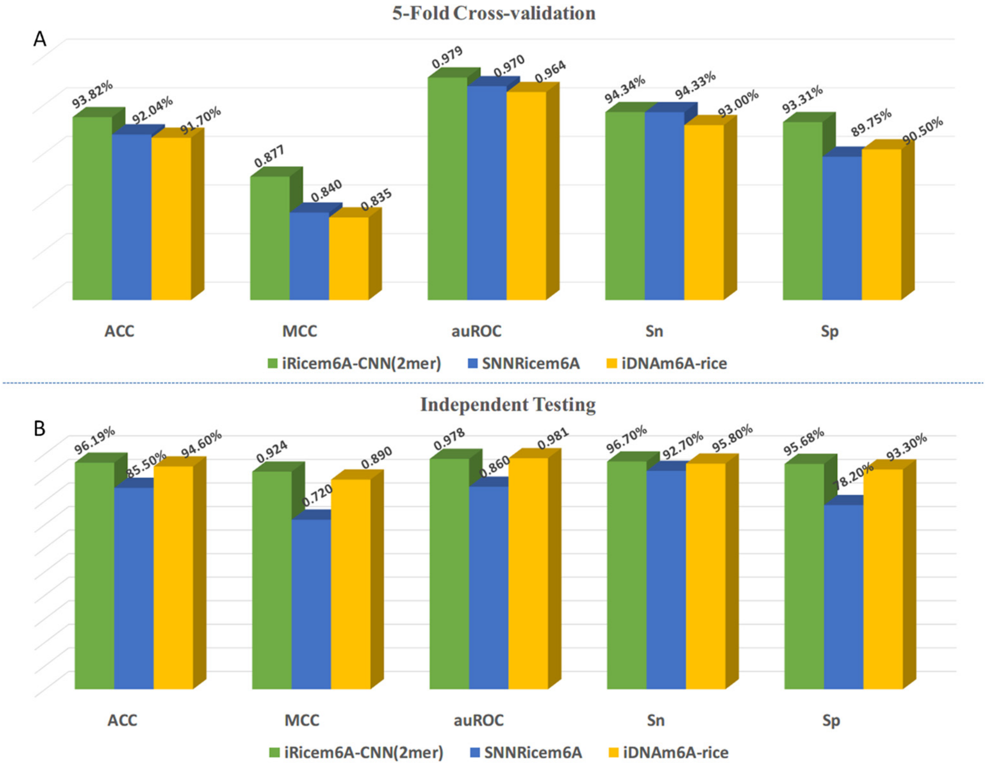
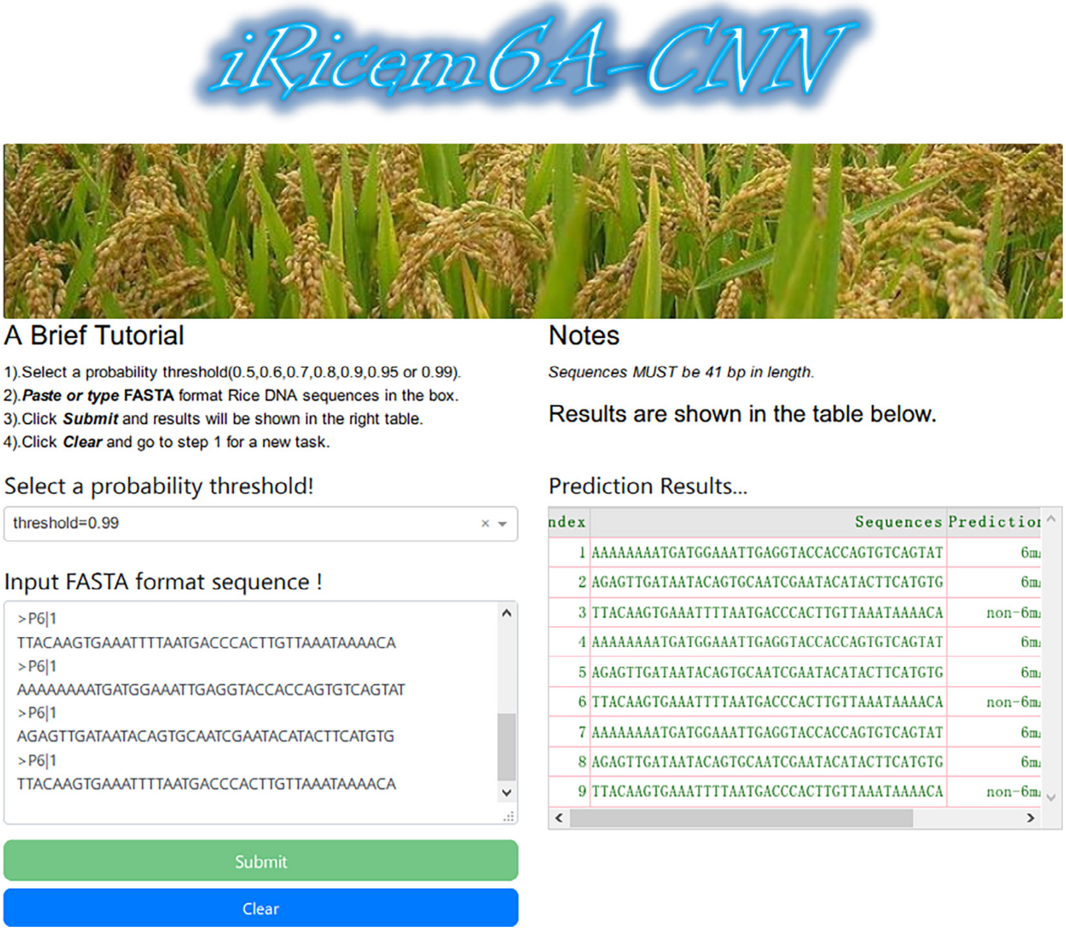


Figure 4. A) Five-fold Cross-validation scores of iRicem6A-CNN and the other two state-of-the-art models based on Lv-rice-m6A training dataset (probability threshold=0.50) and B) independent testing scores for iRicem6A-CNN and other models based on Chen-rice-m6A dataset.



A Brief Tutorial

1).Select a probability threshold(0.5,0.6,0.7,0.8,0.9,0.95 or 0.99).
2).***Paste or type*** **FASTA** format Rice DNA sequences in the box.
3).Click ***Submit*** and results will be shown in the right table.
4).Click ***Clear*** and go to step 1 for a new task.

Select a probability threshold!

threshold=0.99

Input FASTA format sequence !

```
>P6|1
TTACAAGTGAAATTTTAATGACCCACTTGTTAAATAAAACA
>P6|1
AAAAAAAATGATGGAAATTGAGGTACCACCAGTGTCAGTAT
>P6|1
AGAGTTGATAATACAGTGCAATCGAATACATACTTCATGTG
>P6|1
TTACAAGTGAAATTTTAATGACCCACTTGTTAAATAAAACA
```

Notes

Sequences MUST be 41 bp in length.

Results are shown in the table below.

Prediction Results...

ndex	Sequences	Predictio
1	AAAAAAAATGATGGAAATTGAGGTACCACCAGTGTCAGTAT	6m
2	AGAGTTGATAATACAGTGCAATCGAATACATACTTCATGTG	6m
3	TTACAAGTGAAATTTTAATGACCCACTTGTTAAATAAAACA	non-6m
4	AAAAAAAATGATGGAAATTGAGGTACCACCAGTGTCAGTAT	6m
5	AGAGTTGATAATACAGTGCAATCGAATACATACTTCATGTG	6m
6	TTACAAGTGAAATTTTAATGACCCACTTGTTAAATAAAACA	non-6m
7	AAAAAAAATGATGGAAATTGAGGTACCACCAGTGTCAGTAT	6m
8	AGAGTTGATAATACAGTGCAATCGAATACATACTTCATGTG	6m
9	TTACAAGTGAAATTTTAATGACCCACTTGTTAAATAAAACA	non-6m

Figure 5. Webserver interface.

iRicem6A-CNN all outperformed those of SNNRicem6A, by 12.5%, 28.3%, 13.7%, 4.3% and 22.3% respectively. Particularly, the remarkable difference between their Sp values indicated a greater ability of iRicem6A-CNN than of SNNRicem6A to accurately distinguish negative samples from positive samples, and hence indicated a better general performance of iRicem6A-CNN for rice-gnome m^6A sites discrimination. Compared with iDNAm6A-rice based on the random forest method, iRicem6A-CNN showed improvements of 1.68%, 3.82%, 0.94% and 2.55% in ACC, MCC, Sn and Sp respectively, with the other two metrics with tiny decline (0.30% of auROC). The better Sp and means its better recognition of negative and positive samples for iRicem6A-CNN compared to those of iDNAm6A-rice, which led to the greater ACC and MCC values for iRicem6A-CNN.

3.3. Webserver implementation

A webserver for iRicem6A-CNN is easily accessed at http://iRicem6A-CNN.aibiochem.net. The user firstly select a probability threshold (e.g. threshold = 0.99) and just needs to type or paste the DNA sequence in FASTA format into the text box as shown in the webpage (Figure 5) and then click the submit button for a prediction. After allowing a while for computing, the prediction results are displayed in a tabular format. To start a new query task, clicking the clear button resets the model, and repeating the above-mentioned operation allows iRicem6A-CNN to automatically handle the new query task.

4. Conclusions

Here, a new deep-learning-based protocol for m^6A sites in the rice genome, iRicem6A-CNN, was developed. The iRicem6A-CNN model was designed with a special feature, involving DNA sequences fed into the model becoming first converted into a 2-mer one-hot encoding tensor. The experiment results shown that model with 2-mer one-hot encoding outperforms model with 1-mer one-hot encoding and demonstrates more robust under various prediction probability threshold values. Application of the model showed a good 5-fold cross-validation accuracy (93.82%) and independent testing accuracy (96.19%), making it one of the best predictors of the m^6A site in the rice genome. Our analysis and comparisons showed iRicem6A-CNN to be capable of not only accurately predicting m^6A-positive samples but also reducing the error rate of negative sample identification, and our results indicated it to run better than the state-of-the-art predictors. Moreover, a user-friendly webserver for iRicem6A-CNN has been made available. We expect iRicem6A-CNN to become a useful tool for computationally discerning m6A sites. Note, moreover, that with regards to iRicem6A-CNN, there remains room for improvement: we expect new methods such as Gen2Vec [82], auto encoder [83], MOEA [84,85] or long- and short-term memory networks [86] to be explored for the purpose of further optimizing the model.

Funding

The work is funded by the National Natural Science Foundation of China (No. 62001090, 91935302, 61922020, 61771331), and by the China Postdoctoral Science Foundation (No.2020M673184).

CRediT authorship contribution statement

Zhibin Lv: Conceptualization, Methodology, Data curation. **Hui Ding:** Investigation. **Lei Wang:** Visualization, Investigation. **Quan Zou:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. Zhang et al., N-6-Methyladenine DNA Modification in Drosophila, Cell 161 (4) (2015) 893–906.

[2] S. Kumar et al., N-4-cytosine DNA methylation regulates transcription and pathogenesis in Helicobacter pylori, Nucleic Acids Research 46 (7) (2018) 3429–3445.

[3] Koziol, M.J., et al., Identification of Methylated Deoxyadenosines in Genomic DNA by dA6m DNA Immunoprecipitation. Bio-protocol, 2016. 6(21): p. 10.21769/BioProtoc.1990.

[4] K.G. Au, K. Welsh, P. Modrich, Initiation of Methyl-directed Mismatch Repair, Journal of Biological Chemistry 267 (17) (1992) 12142–12148.

[5] B. Li et al., NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res 45 (W1) (2017) W162–W170.

[6] A. Ameur, W.P. Kloosterman, M.S. Hestand, Single-Molecule Sequencing: Towards Clinical Applications, Trends in Biotechnology 37 (1) (2019) 72–85.

[7] F.J. Sedlazeck et al., Accurate detection of complex structural variations using single-molecule sequencing, Nature Methods 15 (6) (2018) 461.

[8] Y. Wang et al., Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics, Nucleic Acids Res (2019) p, https://doi.org/10.1093/nar/gkz981.

[9] B.A. Flusberg et al., Direct detection of DNA methylation during single-molecule, real-time sequencing, Nature Methods 7 (6) (2010) 461–U72.

[10] C.-L. Xiao et al., N-6-Methyladenine DNA Modification in the Human Genome, Molecular Cell 71 (2) (2018) p. 306-+.

[11] J. Xiong et al., N-6-Hydroxymethyladenine: a hydroxylation derivative of N-6-methyladenine in genomic DNA of mammals, Nucleic Acids Research 47 (3) (2019) 1268–1277.

[12] J. Yin et al., VARIDT 1.0: variability of drug transporter database, Nucleic Acids Res (2019) p, https://doi.org/10.1093/nar/gkz779.

[13] L. Cheng et al., DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function, Bioinformatics 34 (11) (2018) 1953–1956.

[14] L. Kong, L. Zhang, i6mA-DNCP: Computational Identification of DNA N6-Methyladenine Sites in the Rice Genome Using Optimized Dinucleotide-Based Features, Genes 10 (10) (2019) 828.

[15] C. Pian et al., MM-6mAPred: Identifying DNA N6-methyladenine sites based on Markov Model, Bioinformatics (Oxford, England) (2019), https://doi.org/10.1093/bioinformatics/btz556.

[16] S. Basith et al., SDM6A: A Web-Based Integrative Machine-Learning Framework for Predicting 6mA Sites in the Rice Genome, Molecular therapy. Nucleic acids 18 (2019) 131–141.

[17] N.Q.K. Le, iN6-methylat (5-step): identifying DNA N-6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule, Molecular Genetics and Genomics 294 (5) (2019) 1173–1182.

[18] H. Lv et al., iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice, Frontiers in Genetics (2019) p, https://doi.org/10.3389/fgene.2019.00793.

[19] G. Eraslan et al., Deep learning: new computational modelling techniques for genomics, Nature Reviews Genetics 20 (7) (2019) 389–403.

[20] L. Yu et al., Conserved Disease Modules Extracted From Multilayer Heterogeneous Disease and Gene Networks for Understanding Disease Mechanisms and Predicting Disease Treatments, Frontiers in Genetics 9 (2019) 745.

[21] J. Tang et al., Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains, Mol Cell Proteomics 18 (8) (2019) 1683–1699.

[22] P. Wang et al., Differentiating Physicochemical Properties between Addictive and Nonaddictive ADHD Drugs Revealed by Molecular Dynamics Simulation Studies, ACS Chem Neurosci 8 (6) (2017) 1416–1428.

[23] L.L. Huang et al., Learning deep representations for semantic image parsing: a comprehensive overview, Frontiers of Computer Science 12 (5) (2018) 840–857.

[24] J. Zhang et al., Robust feature learning for online discriminative tracking without large-scale pre-training, Frontiers of Computer Science 12 (6) (2018) 1160–1172.

[25] J. Zou et al., A primer on deep learning in genomics, Nature Genetics 51 (1) (2019) 12–18.

[26] A. L'Heureux et al., Machine Learning With Big Data: Challenges and Approaches, Ieee Access 5 (2017) 7776–7797.

[27] Y. Ding, J. Tang, F. Guo, Identification of drug-target interactions via multiple information integration, Information Sciences 418–419 (2017) 546–560.

[28] Y. Ding, J. Tang, F. Guo, Identification of Protein-Protein Interactions via a Novel Matrix-Based Sequence Representation Model with Amino Acid Contact Information, International Journal of Molecular Sciences 17 (10) (2016) 1623.

[29] Y. Ding, J. Tang, F. Guo, Predicting protein-protein interactions via multivariate mutual information of protein sequences, Bmc Bioinformatics 17 (1) (2016) 398.

[30] W. Xue et al., What Contributes to Serotonin-Norepinephrine Reuptake Inhibitors' Dual-Targeting Mechanism? The Key Role of Transmembrane Domain 6 in Human Serotonin and Norepinephrine Transporters Revealed by Molecular Dynamics Simulation, ACS Chem Neurosci 9 (5) (2018) 1128–1140.

[31] J. Tang et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief Bioinform (2019) p, https://doi.org/10.1093/bib/bby127.

[32] W.B. Liu et al., A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.

[33] B. Alipanahi et al., Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nature Biotechnology 33 (8) (2015) p. 831-+.

[34] B.H. Tang et al., Recent Advances of Deep Learning in Bioinformatics and Computational Biology, Frontiers in Genetics 10 (2019) 10.

[35] B.J. Ainscough et al., A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data, Nature Genetics 50 (12) (2018) 1735.

[36] Z. Zhang et al., Deep learning in omics: a survey and guideline, Briefings in Functional Genomics 18 (1) (2019) 41–57.

[37] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Briefings in Bioinformatics 18 (5) (2017) 851–869.

[38] Z. Lv, C. Ao, Q. Zou, Protein Function Prediction: From Traditional Classifier to Deep Learning, Proteomics 19 (2019) 1900119.

[39] Q. Zou et al., An approach for identifying cytokines based on a novel ensemble classifier, BioMed research international 2013 (2013) (2013) 686090.

[40] F.G.C. Cabarle et al., On solutions and representations of spiking neural P systems with rules on synapses, Information Sciences 501 (2019) 30–49.

[41] H. Tan et al., Multi-omic analyses of exogenous nutrient bag decomposition by the black morel Morchella importuna reveal sustained carbon acquisition and transferring, in: Environmental Microbiology, ED, MediumSize, 2019, pp. 3909–3926.

[42] L. Yu, J. Zhao, L. Gao, Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity, International Journal of Biological Sciences 14 (8) (2018) 971–980.

[43] L. Yu, J. Zhao, L. Gao, Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome, Artificial Intelligence in Medicine 77 (2017) 53–63.

[44] Y. Shen, J. Tang, F. Guo, Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC, Journal of Theoretical Biology 462 (2019) 230–239.

[45] L. Yu et al., Drug and Nondrug Classification Based on Deep Learning with Various Feature Selection Strategies, Current Bioinformatics 13 (3) (2018) 253–259.

[46] L. Peng et al., The Advances and Challenges of Deep Learning Application in Biological Big Data Processing, Current Bioinformatics 13 (4) (2018) 352–359.

[47] L.L. Nie et al., Prediction of Protein S-Sulfenylation Sites Using a Deep Belief Network, Current Bioinformatics 13 (5) (2018) 461–467.

[48] M. Tahir, H. Tayara, K.T. Chong, iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule, Chemometrics and Intelligent Laboratory Systems 189 (2019) 96–101.

[49] Yu, H. and Z. Dai, SNNRice6mA: A Deep Learning Method for Predicting DNA N6-Methyladenine Sites in Rice Genome. Frontiers in Genetics, 2019. 10 (1071): p. 10.3389/fgene.2019.01071.

[50] H. Lv et al., Evaluation of different computational methods on 5-methylcytosine sites identification, Briefings in Bioinformatics (2019) p, https://doi.org/10.1093/bib/bbz048.

[51] L. Wei et al., A novel hierarchical selective ensemble classifier with bioinformatics application, Artificial Intelligence in Medicine 83 (2017) 82–90.

[52] L. Wei et al., Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier, Artificial Intelligence in Medicine 83 (2017) 67–74.

[53] L. Wei et al., CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency, Journal of Proteome Research 16 (5) (2017) 2044–2053.

[54] B. Liu, X. Gao, H. Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches, Nucleic Acids Research 47 (20) (2019) e127.

[55] Liu, B., et al., iRO-PsekGCC: identify DNA replication origins based on Pseudo k-tuple GC Composition. Frontiers in Genetics. 10: p. 842.

[56] L. Cheng et al., InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk, BMC Genomics 19 (Suppl 1) (2018) 919.

[57] L. Yu et al., Prediction of Novel Drugs for Hepatocellular Carcinoma Based on Multi-Source Random Walk, Ieee-Acm Transactions on Computational Biology and Bioinformatics 14 (4) (2017) 966–977.

[58] L. Xu et al., An Efficient Classifier for Alzheimer's Disease Genes Identification, Molecules 23 (12) (2018) 3140.

[59] L. Xu et al., k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification, Frontiers in Genetics 10 (33) (2019).

[60] L. Xu et al., SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins, International Journal of Molecular Sciences 19 (6) (2018) 1773.

[61] L. Xu et al., A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides, Genes 9 (3) (2018) 158.

[62] L. Jiang et al., FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association, BMC Genomics 19 (911) (2019) 11–25.

[63] Y. Ding, J. Tang, F. Guo, Identification of drug-side effect association via multiple information integration with centered kernel alignment, Neurocomputing 325 (2019) 211–224.

[64] L. Jiang et al., MDA-SKF: Similarity Kernel Fusion for Accurately Discovering miRNA-Disease Association, Frontiers in Genetics 9 (618) (2018) 1–13.

[65] X. Zhu et al., A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. Brief Funct, Genomics (2019) p, https://doi.org/10.1093/bfgp/elz018.

[66] Wang, X., et al., STS-NLSP: A Network-Based Label Space Partition Method for Predicting the Specificity of Membrane Transporter Substrates Using a Hybrid Feature of Structural and Semantic Similarity. Frontiers in Bioengineering and Biotechnology, 2019. 7(306): p. 10.3389/fbioe.2019.00306.

[67] X. Shan et al., Prediction of CYP450 Enzyme-Substrate Selectivity Based on the Network-based Label Space Division Method, J Chem Inf Model 59 (11) (2019) 4577–4586.

[68] Y. Xiong et al., PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method, Front Microbiol 9 (2018) 2571.

[69] X. Zeng et al., deepDR: a network-based deep learning approach to in silico drug repositioning, Bioinformatics (2019), https://doi.org/10.1093/bioinformatics/btz418.

[70] Y. Liu et al., Inferring MicroRNA-Disease Associations by Random Walk on a Heterogeneous Network with Multiple Data Sources, IEEE/ACM Transactions on Computational Biology and Bioinformatics 14 (4) (2017) 905–915.

[71] X. Zhang et al., Meta-path methods for prioritizing candidate disease miRNAs, IEEE/ACM Transactions on Computational Biology and Bioinformatics 16 (1) (2019) 283–291.

[72] J.X. Tan et al., Identification of hormone binding proteins based on machine learning methods, Mathematical Biosciences and Engineering 16 (4) (2019) 2466–2480.

[73] M. Zhang et al., Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble, Analytical Biochemistry 550 (2018) 41–48.

[74] K. Chen et al., WHISTLE: a high-accuracy map of the human N-6-methyladenosine (m(6)A) epitranscriptome predicted using a machine learning approach, Nucleic Acids Research 47 (7) (2019) e41.

[75] Y. Jiao, P. Du, Performance measures in evaluating machine learning based bioinformatics predictors for classifications, Quantitative Biology 4 (4) (2016) 320–330.

[76] C.Q. Feng et al., iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators, Bioinformatics 35 (9) (2019) 1469–1477.

[77] Liu, B., BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. Briefings in Bioinformatics. DOI: 10.1093/bib/bbx165.

[78] L. Cheng et al., OAHG: an integrated resource for annotating human genes with multi-level ontologies, Scientific Reports 6 (2016) 1–9.

[79] X.X. Zeng et al., Prediction of potential disease-associated microRNAs using structural perturbation method, Bioinformatics 34 (14) (2018) 2425–2432.

[80] X. Zeng et al., Prediction and Validation of Disease Genes Using HeteSim Scores, IEEE/ACM Transactions on Computational Biology and Bioinformatics 14 (3) (2017) 687–695.

[81] L. Cheng et al., gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions, Nucleic Acids Res (2019) p, https://doi.org/10.1093/nar/gkz843.

[82] Q. Zou et al., Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA, Rna 25 (2) (2019) 205–218.

[83] L. Wei et al., Prediction of human protein subcellular localization using deep learning, Journal of Parallel and Distributed Computing 117 (2018) 212–217.

[84] H. Xu et al., An Evolutionary Algorithm Based on Minkowski Distance for Many-Objective Optimization, IEEE Transactions on Cybernetics 49 (11) (2019) 3968–3979.

[85] H. Xu et al., MOEA/HD: A Multiobjective Evolutionary Algorithm Based on Hierarchical Decomposition, IEEE Transactions on Cybernetics 49 (2) (2019) 517–526.

[86] R. Halder et al., DNA methylation changes in plasticity genes accompany the formation and maintenance of memory, Nature Neuroscience 19 (1) (2016) 102.

Zhibin Lv majored in materials and chemistry and the research area was polymers, nano-materials and solar cell from 2004 to Jul. 2013. He worked as an assistant professor and senior engineer of nuclear materials and testing in China Academy of Engineering Physics from Jul. 2013 to Apr. to 2017. Then he joined fine chemicals department of Chengrand Institute of Chemical Industry Co., Ltd as a department deputy responsible for epoxy resin R&D. Since May.2019, he is a postdoctoral in University of Electronic Science and Technology of China. His interest is to apply artificial intelligent to cheminformatics and bioinformatics.

Ding Hui is currently an associate professor in the School of Life Science and Technology of the University of Electronic Science and Technology of China. She received a Bachelor of Science degree from the Department of Physics of Inner Mongolia University in 2000, a Master of Science degree in Biophysics from the Institute of Technology of Inner Mongolia University in 2003, and a PhD degree in theoretical physics from the Institute of Technology of Inner Mongolia University in 2009. Since March 2009, she has been an assistant professor in the School of Life Science and Technology, University of Electronic Science and Technology of China.

Lei Wang is now a professor of College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, China. He received his PH.D. from Hunan University, P.R.China in 2005. From 2005 to 2007, he worked as postdoc in Tsinghua University for computer application technology. From 2008 to 2009, he visited Lakehead University in Canada. The he had been a professor of Xiangtan University until Apr. 2019. His research areas are bioinformatics and computer science.

Quan Zou is a professor of Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China. He received his PH.D. from Harbin Institute of Technology, P.R.China in 2009, and worked at Xiamen University and Tianjin University from 2009 to 2018. His research is in the areas of bioinformatics, machine learning and parallel computing. Now he is putting the focus on protein classification, genome assembly, annotation and functional analysis from the next generation sequencing data with parallel computing methods. Several related works have been published by Briefings in Bioinformatics, Bioinformatics, PLOS Computational Biology and IEEE/ACM Transactions on Computational Biology and Bioinformatics. Google scholar showed that his more than 100 papers have been cited more than 4000 times. He is also a reviewer for many impacted journals and NSFC(National Natural Science Foundation of China).