Front. Comput. Sci., 2022, 16(5): 165903
https://doi.org/10.1007/s11704-021-1015-3

RESEARCH ARTICLE

Towards a better prediction of subcellular location of long non-coding RNA

Zhao-Yue ZHANG, Zi-Jie SUN, Yu-He YANG, Hao LIN (✉)

Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

Abstract The spatial distribution pattern of long non-coding RNA (lncRNA) in cell is tightly related to their function. With the increment of publicly available subcellular location data, a number of computational methods have been developed for the recognition of the subcellular localization of lncRNA. Unfortunately, these computational methods suffer from the low discriminative power of redundant features or overfitting of oversampling. To address those issues and enhance the prediction performance, we present a support vector machine-based approach by incorporating mutual information algorithm and incremental feature selection strategy. As a result, the new predictor could achieve the overall accuracy of 91.60%. The highly automated web-tool is available at lin-group.cn/server/iLoc-LncRNA(2.0)/website. It will help to get the knowledge of lncRNA subcellular localization.

Keywords lncRNA, subcellular localization, support vector machine, mutual information, Web server

1 Introduction

In mammalian, most of RNAs transcribed from genomic DNA sequences do not encode proteins, but participate in the regulation of protein translation mechanism as functional molecules. lncRNA is a kind of non-coding RNA with a length of more than 200 nucleotides. Most of its functions are unknown or uncertain. It is worth noting that lncRNA-regulation modality is dependent on their cellular localization [1]. lncRNA located in the nucleus usually modulates the RNA transcription, and the lncRNA in the cytoplasm participates in post-transcriptional regulation [2−9]. And the localization of lncRNA is also related to some diseases [10]. Thus, the knowledge about the subcellular region of lncRNA will be helpful for the function annotation of lncRNA.

Development of RNA fluorescence in situ hybridization (RNA FISH) and RNA-sequencing (RNA-seq) for large scale detection of RNA provides a lot of data for the study of RNA subcellular localization [11−13]. For the convenience of researchers, three RNA subcellular localization databases have been constructed. RNALocate [14] is the first database designed for RNA subcellular location. The database manually collects various RNA subcellular localization entries and experimental evidence. At present, there are 2382 lncRNA subcellular localization entries involving nine organisms in RNALocate. By collecting RNA-seq data from GENCODE and quantitative analysis, lncATLAS [15] provides the tendency of subcellular location of 6768 *Homo sapiens* lncRNAs according to the relative concentration index. lncSLdb [16] is another platform for the collection of lncRNA subcellular location. The subcellular localization of more than 11000 lncRNAs from human, mouse and fruit fly was manually curated from literatures. lncSLdb classifies lncRNAs into three basic types: nucleus, cytoplasm or both (nucleus/cytoplasm) according to fluorescence in situ hybridization image or relative expression level in different cell compartments. However, the localization annotation of lncRNA is extremely incomplete, which will prevent scholars to further study lncRNA.

The establishment of RNA subcellular localization databases provides a possibility to develop computational methods for the identification of lncRNA subcellular location. The cytoplasmic/nuclear ratio based on RNA-seq data has been used to infer the preference location of lncRNA by combining with machine learning methods. DeepLncRNA [17] used deep neural network [18−22] to construct model and achieved the accuracy of 72.4%. Another model [23] was built based on random forest [24]. Its recall rates range from 53.8% to 84.1% in different cell lines. Due to the technique limitations, the two RNA-Seq-based works can only infer the enrichment of lncRNA in the nucleus or cytoplasm. As far as we know, lncLocator [25], iLoc-lncRNA [26], Locate-R [27] and lncLocation [28] are four only predictors for lncRNA subcellular localization based on RNALocate. lncLocator is a neural network-based model which could produce the accuracy of 0.598. iLoc-lncRNA is a computational tool by combining optimal eight-tuple oligonucleotides with support vector machine (SVM) [29] to predict the lncRNA subcellular location. The overall accuracy reached 86.72%. Locate-R used *n*-gapped *l*-mer composition and *l*-mer composition features to build model. Its accuracy is 90.69%. lncLocation integrate

Received January 12, 2021; accepted March 31, 2021

E-mail: hlin@uestc.edu.cn

multi-source features to construct a sequence-based computational tool that obtain an 87.78% accuracy. The good performance of these tools indicates that sequence motifs are the main driving force for subcellular localization of lncRNA. However, we noticed that there are still room to improve the prediction performance of lncRNA subcellular localization. In this work, an optimizing model based on iLoc-lncRNA was developed (Fig. 1(a)), which could greatly improve the performance and efficiency of lncRNA subcellular location prediction.

2 Materials and methods

2.1 Benchmark datasets

The new version of iLoc-lncRNA was built based on the same benchmark dataset of iLoc-lncRNA. Subcellular location information of lncRNA in mammals was retrieved from RNALocate [14]. The corresponding nucleotide sequences were download from RefSeq [30]. After removing the redundant sequences with more than 80% similarity, 655 samples were used to constructed the benchmark dataset, in which 64.12% samples come from *Mus musculus*, 34.5% from *Homo sapiens* (Fig. 1(b)). These lncRNAs unique located in nucleus, cytoplasm, ribosome and exosome with average length of 3462, 2199, 1890 and 1657, respectively. The number of samples and the length distribution of each subcellular location can be seen in Fig. 1(c).

2.2 Feature encoding

A lot of studies indicated that *k*-tuple (also called *k*-mer) nucleotide composition[31−36] performs well in the description of RNA sequences in subcellular location prediction [26,37,38]. In this study, we adopt the eight-tuple oligonucleotides feature encoding method as in iLoc-lncRNA. Let the lncRNA sequence $\boldsymbol{S}$ expressed as following:

$$\boldsymbol{S} = R_1R_2R_3R_4R_5\cdots R_iR_{i+1}\cdots R_L, \tag{1}$$

where R_i is the ith base and $R_i \in \{A,G,C,T\}$, L is the length of $\boldsymbol{S}$.

The normalized frequency of the ith eight-tuple nucleotide component occurring in $\boldsymbol{S}$ and can be calculated by

$$f_i^{8-tuple} = \frac{n_i}{\sum_{i=1}^{4^8} n_i} = \frac{n_i}{L-8+1}, \tag{2}$$

where n_i means the number of occurrences of the ith eight-tuple nucleotide component in the lncRNA sequence $\boldsymbol{S}$.

The primary sequence $\boldsymbol{S}$ can be transferred into a vector $\boldsymbol{V}$ with 4^8 elements as following:

$$\boldsymbol{V} = [f_1^{8-tuple}\cdots f_i^{8-tuple}\cdots f_{4^k}^{8-tuple}]^{\mathrm{T}}, \tag{3}$$

where the symbol T means the transposition of a vector.

2.3 Feature selection

In previous version, 4107 optimal octamers were selected as the final features for iLoc-lncRNA by using binomial distribution score with incremental feature selection strategy (IFS) [37,39,40]. First, all eight-tuple oligonucleotides features were sorted according to the binomial distribution score. The high binomial distribution score indicates the presence of the octamer in a subcellular location is not accidental. Then, IFS based on the wrapper method was performed to determine the optimal feature set. IFS is a kind of sequential search strategies that added features one by one to feature set from higher to lower ranked score. Owing to its low time and space complexity, IFS has been widely used in feature selection[26,41−43]. However, it should be noted that the feature set selected by the combination of binomial distribution score and IFS could have rich redundancy. In current work, minimal-redundancy-maximal-relevance crite-

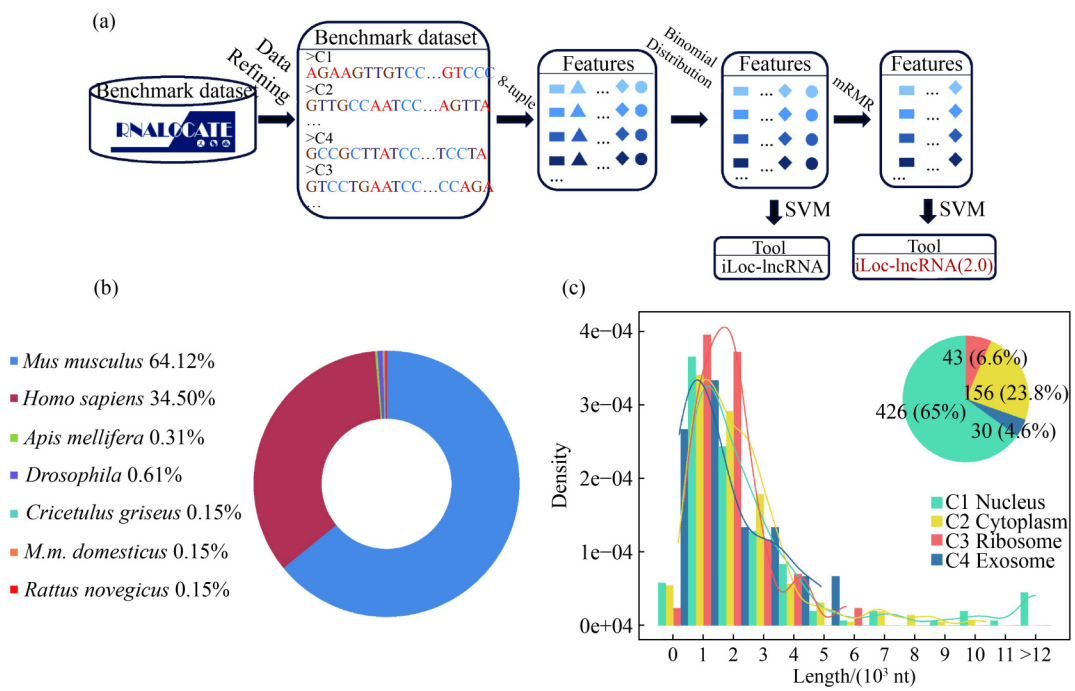

Fig. 1 (a) The workflow diagram of developing the iLoc-lncRNA(2.0); (b) species composition in the benchmark dataset; (c) length distribution of lncRNA sequences and the number of samples in each subcellular location

rion (mRMR) [44,45] was implement on the 4107 optimal octamers to winnow out the redundant features.

2.3.1 mRMR

mRMR is a mutual-information-based feature selection algorithm. For two random variables x and y, their mutual information means the information obtained about y (or x) after the knowledge about x (or y), which can be calculated as:

$$I(x,y)=\int\int p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}\mathrm{d}x\mathrm{d}y, \tag{4}$$

where $p(\cdot)$ represents probability density.

Suppose we already have a feature set O_{m-1} with $m-1$ features, mRMR select the mth feature by maximizing relevance to the target class and minimizing correlation between features with incremental search methods. The relevance to the target class can be accumulated by the mean value of all mutual information values between individual feature x_i and the class c:

$$D=\frac{1}{|O|}\sum\nolimits_{x_i\in O}I(x_i;c). \tag{5}$$

The correlation with the existing features can be accumulated by

$$R=\frac{1}{|O|^2}\sum\nolimits_{x_i,x_j\in O}I\left(x_i;x_j\right). \tag{6}$$

Finally, the mth feature will be selected according to the following criteria:

$$\max\,\phi(D,R),\phi=\frac{D}{R}. \tag{7}$$

After sorted the 4107 octamers according to their mRMR score, IFS was used to perform a search in the space of feature subsets. The optimal feature subset was determined by the classification performance on a cross-validation

2.4 Support vector machine (SVM)

SVM is one kind of non-linear classification models [46], which were widely applied to the field of bioinformatics [47]. It is effective in cases where number of dimensions is greater than the number of samples. For linearly inseparable samples, SVM maps the samples into a high-dimension feature space so that different categories of examples can be divided by a maximum-margin hyperplane. C-SVC is a classic SVM model, where C represents the penalty coefficient of soft margin SVM. The large value of C implies the small margin, there is a tendency to overfit the training model. RBF is a typical kernel function in the calculation of the inner product to avoid the explicit computation in the feature space. In this study, C-support vector classification(C-SVC) with radial basis function (RBF) was performed by using library LIBSVM [48]. LIBSVM provides a cross validation via parallel grid search tool for parameter selection. The searching space of the best penalty coefficient C and width parameter γ of RBF is as following:

$$\begin{cases} C\in\left[2^{-5},2^{15}\right],\ step=2, \\ \gamma\in\left[2^{-15},2^{3}\right],\ step=2^{-1}. \end{cases} \tag{8}$$

2.5 Performance evaluation metrics

Six evaluation metrics: overall accuracy (OA), sensitivity (Sn) (also known as recall), specificity (Sp), precision (Pre), Matthew's correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUC) were used to evaluate the performance of the model [21,49−58]. Receiver operating characteristic (ROC) curve shows the relationship between sensitivity and specificity. In general, higher AUC values indicate better test performance [59−63]. The other five indexes were formulated as

$$Sn=\frac{TP}{TP+FN}, \tag{9}$$

$$Sp=\frac{TN}{TN+FP}, \tag{10}$$

$$Pre=\frac{TP}{TP+FP}, \tag{11}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}, \tag{12}$$

$$OA=\frac{TP+TN}{TP+TN+FN+FP}, \tag{13}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.

3 Results

3.1 Performance evaluation metrics

The evaluation in feature selection was conducted through a 5-fold cross-validation with LIBSVM. Firstly, 4107 octamers composition features were ranked by mRMR score. By using the feature selection technique, each feature subset was used to train the model by using LIBSVM. Finally, we obtained 1409 optimal features which could produce the maximum prediction accuracy of 92.06% with the parameter C of 2^5 and γ of 2^{-9} (Fig. 2). Thus, a new version iLoc-lncRNA(2.0) was constructed based on the 1409 features with the best penalty coefficient of 2^5 and width parameter of 2^{-9}. For comparison

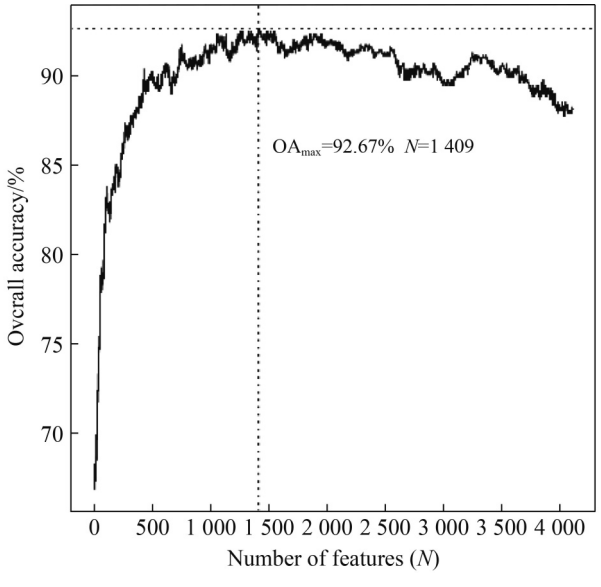

Fig. 2 Incremental feature selection strategy accuracy curve for mRMR feature selection

Table 1 The performance of the SVM-based subcellular location prediction model

Method	Feature dimension	Subcellular location	*Sn* %	*Sp* %	*Pre* %	*MCC*	*AUC*	*OA* %
iLoc-lncRNA(2.0) [a]	1409	Nucleus	91.03	95.59	86.59	0.852	0.969	91.60
		Cytoplasm	94.37	89.96	94.59	0.842	0.969	
		Ribosome	83.72	99.01	85.71	0.837	0.986	
		Exosome	66.67	99.36	83.33	0.735	0.949	
Locate-R [b]	857	Nucleus	66.92	95.15	/	0.66	0.900	90.69
		Cytoplasm	84.74	89.1	/	0.725	0.930	
		Ribosome	100	98.37	/	0.97	1.000	
		Exosome	100	99.17	/	0.978	1.000	
lncLocation	/	Nucleus	74.19	/	95.83	/	/	87.78
		Cytoplasm	100	/	85	/	/	
		Ribosome	55.56	/	100	/	/	
		Exosome	33.33	/	100	/	/	
iLoc-lncRNA [c]	4107	Nucleus	77.56	97.59	/	0.796	/	86.72
		Cytoplasm	99.06	67.68	/	0.742	/	
		Ribosome	46.51	99.83	/	0.652	/	
		Exosome	16.67	1	/	0.4	/	

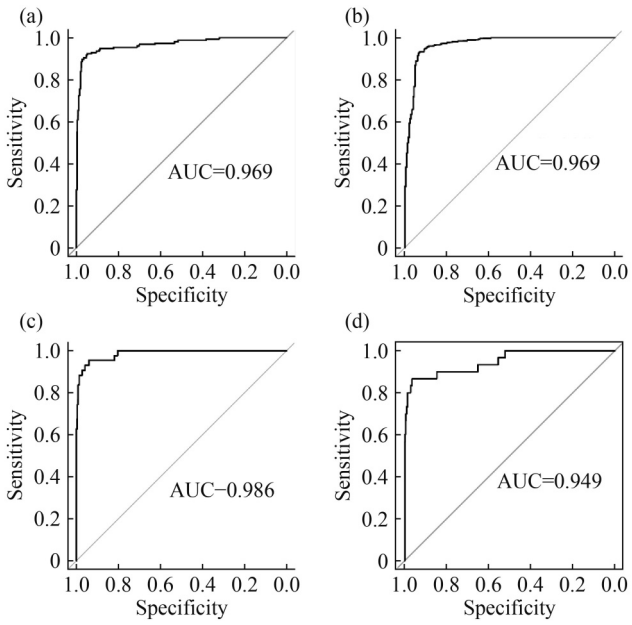

Fig. 3 ROC curves for the iLoc-lncRNA(2.0) for (a) nucleus, (b) cytosol, (c) ribosome, (d) exosome

with other works, the evaluation of iLoc-lncRNA(2.0) was performed by using 10-fold cross-validation. The performances of the iLoc-lncRNA(2.0) are shown in Table 1. In addition, ROC curves were also plotted in Fig. 3 to visually show the prediction capability of our model.

3.2 Comparison with other works

Compared with the old version, the new prediction tool shows good performance in both prediction results and running time. The *OA* of the new version is 91.60% which is 4.88% higher than that of the old one. The values of *OA*, *Sn*, *Sp* and *MCC* of iLoc-lncRNA(2.0) at four locations are generally better than those of iLoc-lncRNA (Table 1). We compared the running time of the two versions on the same sever with Intel(R) Core(TM) i5-4570 CPU and 4 GB of RAM. We found that the new predictor is three times faster than the original iLoc-lncRNA.

Recently, two predictors, Locate-R [27] and lncLocation [28] have been developed for predicting the subcellular location of lncRNAs. Locate-R was constructed on the same data as iLoc-lncRNA(2.0) but performed synthetic minority over-sampling technique (SMOTE) to balance the dataset. The model was built by locally deep SVM based on 867 k-mer and n-gapped k-mer features. It obtains the *OA* of 90.69% with the macro-average AUC of 0.960. It was found that our proposed model yielded *OA* of 91.60% with the macro-average AUC of 0.968 (Fig. 3), which are superior to the Locate-R. Besides, we noticed that the performance of Locate-R in nucleus class which is unprocessed by SMOTE is extremely poor. It indicates that over-sampling the minority may lead to the overfitting of Locate-R. lncLocation filtered k-tuple features and multi-scale structure features by using autoencoder and recursive feature elimination algorithms. The model obtains high precision of 95.83%, 100% and 100% but poor recall of 74.19%, 55.56% and 33.33% for nucleus, ribosome and exosome, respectively. Compared with lncLocation, iLoc-lncRNA(2.0) striking the balance between precision and recall (Table 1).

3.3 Feature analysis

Motif analysis and motif distribution analysis were performed to mining the subcellular localization signal information of lncRNA [64,65]. Firstly, all the 1409 octamers that were used to construct iLoc-lncRNA(2.0) were assigned to four subcellular classes according to the maximum binomial distribution *CL* value. As a result, 263, 395, 294 and 457 octamers were assigned to nucleus, cytoplasm, ribosome and exosome, respectively. Then, ungapped motif discovery was performed on the class-specific octamers by using DREME [66]. Six significant motifs were found in the class-specific

The tools are available at
a) http://lin-group.cn/server/iLoc-LncRNA(2.0)
b) http://locate-r.azurewebsites.net
c) http://lin-group.cn/server/iLoc-LncRNA

octamers (E-value < 0.05). They are [(A/T)C(A/T)] and [(A/C)ACCAA] for nucleus, [A(A/C/T)(A/T)] and [C(C/T)TA] for cytoplasm, [C(C/G)(C/G)] for ribosome and [CG(C/G/T)] for exosome (Fig. 4). In order to observe the distribution pattern of subcellular localization signals, class-specific motifs were mapped to corresponding sequences. The relative position of features on sequences is present with violin plot by using R package "ggplot2" [67] (Fig. 5). Motifs distribution in ribosome and exosome implied that 5' sequence is essential for the subcellular localization of lncRNA. The indiscriminate distribution pattern in nucleus and cytoplasm may cause by their complex subcellular components.

3.4 Web server

The new predictor iLoc-lncRNA(2.0) has been established online and can be freely available at lin-group.cn/server/iLoc-LncRNA(2.0)/website. Once the nucleotide sequence of target lncRNA is submitted to the new predictor, users could effectively obtain its potential subcellular location. We also provided a local tool for the forecast of the big data. All the data used in this study can be downloaded from the website.

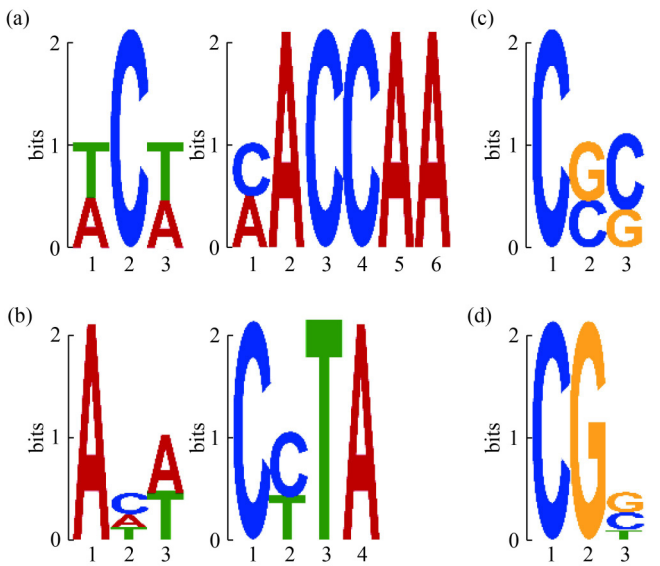

Fig. 4 Visualization of significant class-specific sequence motifs for (a) nucleus, (b) cytosol, (c) ribosome, (d) exosome by using DREME

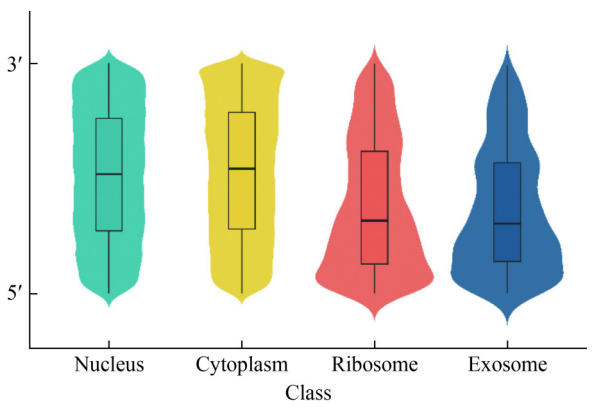

Fig. 5 The class-specific motifs distribution in four classes

4 Conclusions

The comprehensive analysis of the lncRNA subcellular location prediction demonstrated that information in lncRNA sequence has great influence on its subcellular localization. However, the performance of the available tools for lncRNA subcellular location prediction is inadequate due to the feature redundant or severe overfitting issue. Hence, we developed a new predictor named iLoc-lncRNA(2.0) on the basis of mutual information algorithm. The new tool will powerfully support the study of lncRNA subcellular localization. The feature analysis discovered six lncRNA subcellular localization associated motifs which are mostly concentrated in 5′ sequence in ribosome and exosome. A publicly accessible webserver has been established to provide the potential target subcellular location of mammalian lncRNA.

In the future, we will still focus on the lncRNA subcellular location prediction issues. On the one hand, we will discuss ways to handle the imbalance dataset and improve the prediction accuracy of the minority class. On the other hand, though deep learning algorithms such as deep neural network and locally deep SVM have been applied to the prediction of lncRNA subcellular location, their present performance are seen as having chance for future advancement. More efforts will be made to extract and identify features and patterns in the data with deep learning algorithms.

Acknowledgements This work was supported by the National Nature Scientific Foundation of China (Grant No. 61772119), Sichuan Provincial Science Fund for Distinguished Young Scholars (2020JDJQ0012).

References

1. Chiu H S, Somvanshi S, Patel E, Chen T W, Singh V P, Zorman B, Patil S L, Pan Y, Chatterjee S S, Cancer Genome Atlas Research N, Sood A K, Gunaratne P H, Sumazin P. Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. Cell Reports, 2018, 23(1): 297–312. e12
2. Ji J, Tang J, Xia KJ, Jiang R. LncRNA in tumorigenesis microenvironment. Current Bioinformatics, 2019, 14(7): 640–641
3. Guo C J, Xu G, Chen L L. Mechanisms of long noncoding RNA nuclear retention. Trends in Biochemical Sciences, 2020, 45(11): 947-960
4. Chowdhury M R, Basak J, Bahadur R P. Elucidating the functional role of predicted miRNAs in post-transcriptional gene regulation along with symbiosis in medicago truncatula. Current Bioinformatics, 2020, 15(2): 108–120
5. Cheng L, Hu Y, Sun J, Zhou M, Jiang Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics, 2018, 34(11): 1953–1956
6. Cheng L, Wang P, Tian R, Wang S, Guo Q, Luo M, Zhou W, Liu G, Jiang H, Jiang Q. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Research, 2019, 47(D1): D140–D144
7. Jiang Q, Ma R, Wang J, Wu X, Jin S, Peng J, Tan R, Zhang T, Li Y, Wang Y. LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. BMC Genomics, 2015, 16(3): 1–11
8. Jiang Q, Wang J, Wu X, Ma R, Zhang T, Jin S, Han Z, Tan R, Peng J, Liu G, Li Y, Wang Y. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. Nucleic Acids Research, 2015, 43(Database issue): D193–196
9. Jiang Q, Wang J, Wang Y, Ma R, Wu X, Li Y. TF2LncRNA:

identifying common transcription factors for a list of lncRNA genes from ChIP-Seq data. Biomed Research International, 2014, 2014: 317642

10. Ning L, Cui T, Zheng B, Wang N, Luo J, Yang B, Du M, Cheng J, Dou Y, Wang D. MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Research, 2021, 49(D1): D160–d164
11. Mora-Marquez F, Luis Vazquez-Poletti J, Chano V, Collada C, Soto A, Lopez de Heredia U. Hardware performance evaluation of de novo transcriptome assembly software in amazon elastic compute cloud. Current Bioinformatics, 2020, 15(5): 420–430
12. Hu B, Zheng L, Long C, Song M, Li T, Yang L, Zuo Y. EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. Open Biology, 2019, 9(6): 190054
13. Zhu X, Li H D, Guo L, Wu F X, Wang J. Analysis of single-cell RNA-seq data by clustering approaches. Current Bioinformatics, 2019, 14(4): 314–322
14. Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, Yang H, Hu Z, Zhang L, Hu C, Li C, Qian K, Zhang C, Huang Y, Li K, Lin H, Wang D. RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Research, 2017, 45(D1): D135–D138
15. Mas-Ponte D, Carlevaro-Fita J, Palumbo E, Hermoso Pulido T, Guigo R, Johnson R. LncATLAS database for subcellular localization of long noncoding RNAs. RNA, 2017, 23(7): 1080–1087
16. Wen X, Gao L, Guo X, Li X, Huang X, Wang Y, Xu H, He R, Jia C, Liang F. lncSLdb: a resource for long non-coding RNA subcellular localization. Database (Oxford), 2018, 2018: 1–6
17. Gudenas B L, Wang L. Prediction of LncRNA subcellular localization with deep learning from sequence features. Science Reports, 2018, 8(1): 16385
18. Zhao T, Hu Y, Peng J, Cheng L. DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. Bioinformatics, 2020, 36(16): 4466–4472
19. Zhao T, Hu Y, Cheng L. Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning Approaches. Briefings in Bioinformatics, 2020, 22(4): bbaa212
20. Wu B, Zhang H, Lin L, Wang H, Gao Y, Zhao L, Chen Y-P P, Chen R, Gu L. A similarity searching system for biological phenotype images using deep convolutional encoder-decoder architecture. Current Bioinformatics, 2019, 14(7): 628–639
21. Charoenkwan P, Nantasenamat C, Hasan M M, Shoombuatong W. Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. Journal of Computer-Aided Molecular Design, 2020, 34(10): 1105–1116
22. Liu K, Cao L, Du P, Chen W. im6A-TS-CNN: identifying the N(6)-methyladenine site in multiple tissues by using the convolutional neural network. Molecular Therapy-Nucleic Acids, 2020, 21: 1044–1049
23. Zuckerman B, Ulitsky I. Predictive models of subcellular localization of long RNAs. RNA, 2019, 25(5): 557–572
24. Dong Y M, Bi J H, He Q E, Song K. ESDA: an improved approach to accurately identify human snoRNAs for precision cancer therapy. Current Bioinformatics, 2020, 15(1): 34–40
25. Cao Z, Pan X, Yang Y, Huang Y, Shen H B. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics, 2018, 34(13): 2185–2194
26. Su Z D, Huang Y, Zhang Z Y, Zhao Y W, Wang D, Chen W, Chou K C, Lin H. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics, 2018, 34(24): 4196–4204
27. Ahmad A, Lin H, Shatabda S. Locate-R: subcellular localization of long non-coding RNAs using nucleotide compositions. Genomics, 2020, 112(3): 2583–2589
28. Feng S, Liang Y, Du W, Lv W, Li Y. LncLocation: efficient subcellular location prediction of long non-coding RNA-based multi-source heterogeneous feature fusion. International Journal of Molecular Sciences, 2020, 21(19): 7271
29. Wang Y, Shi F, Cao L, Dey N, Wu Q, Ashour A S, Sherratt R S, Rajinikanth V, Wu L. Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. Current Bioinformatics, 2019, 14(4): 282–294
30. Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 2007, 35(Database issue): D61–65
31. Lai H Y, Zhang Z Y, Su Z D, Su W, Ding H, Chen W, Lin H. iProEP: a computational predictor for predicting promoter. Molecular Therapy-Nucleic Acids, 2019, 17: 337–346
32. Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics, 2020, 36(11): 3336–3342
33. Hasan M M, Basith S, Khatun M S, Lee G, Manavalan B, Kurata H. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Briefings in Bioinformatics, 2020, 22(3): bbaa202
34. Manavalan B, Basith S, Shin T H, Wei L, Lee G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Molecular Therapy-Nucleic Acids, 2019, 16: 733–744
35. Basith S, Manavalan B, Shin T H, Lee G. SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Molecular Therapy-Nucleic Acids, 2019, 18: 131–141
36. Zheng L, Huang S, Mu N, Zhang H, Zhang J, Chang Y, Yang L, Zuo Y. RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. Database (Oxford), 2019
37. Zhang Z Y, Yang Y H, Ding H, Wang D, Chen W, Lin H. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. Briefings in Bioinformatics, 2021, 22(1): 526–535
38. Zhang J, Liu B. A review on the recent developments of sequence-based protein feature extraction methods. Current Bioinformatics, 2019, 14(3): 190–199
39. Liang P F, Yang W R, Chen X, Long C S, Zheng L, Li H S, Zuo Y C. Machine learning of single-cell transcriptome highly identifies mRNA signature by comparing F-score selection with DGE analysis. Molecular Therapy-Nucleic Acids, 2020, 20: 155–163
40. Liu K, Chen W, Lin H. XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. Molecular Genetics and Genomics, 2020, 295(1): 13–21
41. Guo X, Gao L, Wang Y, Chiu D K Y, Wang B, Deng Y, Wen X. Large-scale investigation of long noncoding RNA secondary structures in human and mouse. Current Bioinformatics, 2018, 13(5): 450–460
42. Zhang D, Xu Z C, Su W, Yang Y H, Lv H, Yang H, Lin H. iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics, 2021, 37(2): 171–177
43. Wang S P, Zhang Q, Lu J, Cai Y D. Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. Current Bioinformatics, 2018, 13(1): 3–13
44. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226–1238
45. Chen J, Zhao J, Yang S, Chen Z, Zhang Z. Prediction of protein

ubiquitination sites in arabidopsis thaliana. Current Bioinformatics, 2019, 14(7): 614–620
46. Charoenkwan P, Nantasenamat C, Hasan M M, Shoombuatong W. iTTCA-Hybrid: improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. Analytical Biochemistry, 2020, 599: 113747
47. Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microRNA-disease associations based on support vector machine. International Journal of Dato Mining and Bioinformatics, 2013, 8(3): 282–293
48. Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27
49. Wei L, He W, Malik A, Su R, Cui L, Manavalan B. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Briefings in Bioinformatics, 2021, 22(4): bbaa275
50. Hasan M M, Manavalan B, Shoombuatong W, Khatun M S, Kurata H. i4mC-Mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. Computational and Structural Biotechnology Journal, 2020, 18: 906–912
51. Charoenkwan P, Yana J, Schaduangrat N, Nantasenamat C, Hasan M M, Shoombuatong W. iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. Genomics, 2020, 112(4): 2813–2822
52. Charoenkwan P, Chiangjong W, Lee V S, Nantasenamat C, Hasan M M, Shoombuatong W. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Scientific Reports, 2021, 11(1): 1–13
53. Charoenkwan P, Kanthawong S, Nantasenamat C, Hasan M M, Shoombuatong W. iDPPIV-SCM: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase IV (DPP-IV) inhibitory peptides using a scoring card method. Journal of Proteome Research, 2020, 19(10): 4125–4136
54. Charoenkwan P, Kanthawong S, Nantasenamat C, Hasan M M, Shoombuatong W. iAMY-SCM: improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. Genomics, 2021, 113(1): 689–698
55. Charoenkwan P, Kanthawong S, Schaduangrat N, Yana J, Shoombuatong W. PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method. Cells, 2020, 9(2): 353
56. Charoenkwan P, Nantasenamat C, Hasan M M, Shoombuatong W. iTTCA-Hybrid: improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. Analytical Biochemistry, 2020, 599: 113747
57. Charoenkwan P, Shoombuatong W, Lee H C, Chaijaruwanich J, Huang H L, Ho S Y. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS ONE, 2013, 8(9): e72368
58. Charoenkwan P, Yana J, Nantasenamat C, Hasan M M, Shoombuatong W. iUmami-SCM: a novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. Journal of Chemical Information and Modeling, 2020, 60(12): 6666–6678
59. Long H, Sun Z, Li M, Fu H Y, Lin M C. Predicting protein phosphorylation sites based on deep learning. Current Bioinformatics, 2020, 15(4): 300–308
60. Cheng L. Computational and biological methods for gene therapy. Current Gene Therapy, 2019, 19(4): 210–210
61. Cheng L, Hu Y. Human disease system biology. Current Gene Therapy, 2018, 18(5): 255–256
62. Kuang L, Zhao H, Wang L, Xuan Z, Pei T. A novel approach based on point cut set to predict associations of diseases and LncRNAs. Current Bioinformatics, 2019, 14(4): 333–343
63. Chen W, Feng P, Song X, Lv H, Lin H. iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. Molecular Therapy Nucleic Acids, 2019, 18: 269–274
64. Liu D, Li G, Zuo Y. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. Briefings in Bioinformatics, 2019, 20(5): 1826–1835
65. Zheng L, Liu D, Yang W, Yang L, Zuo Y. RaacLogo: a new sequence logo generator by using reduced amino acid clusters. Briefings in Bioinformatics, 2021, 22(3): bbaa096
66. Bailey T L. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics, 2011, 27(12): 1653–1659
67. Ginestet C. ggplot2: elegant graphics for data analysis. Journal of the Royal Statistical Society Series a-Statistics in Society, 2011, 174: 245–245

Zhao-Yue Zhang received the MS degree in Biophysics from University of Electronic Science and Technology of China, China. She is a research assistant of Center for Informational Biology and the Key Laboratory for NeuroInformation of Ministry of Education in University of Electronic Science and Technology of China. Her research interests are bioinformatics, machine learning and RNA subcellular localization.

Zi-Jie Sun is a graduate student at the Center for Informational Biology, University of Electronic Science and Technology of China, China. Her research interests are bioinformatics, statistical analysis and drug repositioning.

Yu-He Yang is a graduate student at the Center for Informational Biology, University of Electronic Science and Technology of China, China. Her research interests are bioinformatics, machine learning and RNA methylation.

Hao Lin is a professor at the Center for Informational Biology, University of Electronic Science and Technology of China, China. His research is in the areas of bioinformatics and system biology.