Briefings in Bioinformatics, 22(1), 2021, 526–535

doi: 10.1093/bib/bbz177
Advance Access Publication Date: 28 January 2020
Problem Solving Protocol

Design powerful predictor for mRNA subcellular location prediction in *Homo sapiens*

Zhao-Yue Zhang, Yu-He Yang, Hui Ding, Dong Wang, Wei Chen and Hao Lin

Corresponding authors: Dong Wang, Center for Information Biology, University of Electronic Science and Technology of China, Chengdu 610054; Department of Bioinformatics at Southern Medical University, Guangzhou 510091, China.
Tel: +86-15546005356; E-mail: wangdong79@smu.edu.cn; Wei Chen, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China. Tel:+86-15027549356 E-mail: chenweiimu@gmail.com; Hao Lin, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China. Tel: +86-13678168394; E-mail: hlin@uestc.edu.cn

Abstract

Messenger RNAs (mRNAs) shoulder special responsibilities that transmit genetic code from DNA to discrete locations in the cytoplasm. The locating process of mRNA might provide spatial and temporal regulation of mRNA and protein functions. The situ hybridization and quantitative transcriptomics analysis could provide detail information about mRNA subcellular localization; however, they are time consuming and expensive. It is highly desired to develop computational tools for timely and effectively predicting mRNA subcellular location. In this work, by using binomial distribution and one-way analysis of variance, the optimal nonamer composition was obtained to represent mRNA sequences. Subsequently, a predictor based on support vector machine was developed to identify the mRNA subcellular localization. In 5-fold cross-validation, results showed that the accuracy is 90.12% for *Homo sapiens* (*H. sapiens*). The predictor may provide a reference for the study of mRNA localization mechanisms and mRNA translocation strategies. An online web server was established based on our models, which is available at http://lin-group.cn/server/iLoc-mRNA/.

Key words: mRNA; subcellular location; feature selection; statistical analysis; web server

Introduction

Subcellular localization of mRNA is a critical step of gene expression, which controls where the proteins are synthesized [1]. The concentrations of mRNA indicated the high expression of the protein and have been demonstrated to be related to embryo development [2, 3], cell polarity [4], cell motility [5] and other biological regulatory mechanisms. Besides, since the interaction often occurs in the same location [6], the co-localization of mRNA and other biomolecules suggested that they are involved in the same regulatory mechanism. Thus, by identifying the mRNA location within cellular, a more accurate estimation of gene activity can be obtained. In addition, in the pharmaceutical industry, subcellular localization of mRNA is helpful to

Zhao-Yue Zhao is a master candidate of Center for Informational Biology at University of Electronic Science and Technology of China. Her research interests are RNA subcellular localization.
Yu-He Yang is a master candidate of Center for Informational Biology at University of Electronic Science and Technology of China. Her research interests are the application of machine learning in RNA modification.
Hui Ding is an associate professor of Center for Informational Biology at University of Electronic Science and Technology of China. Her research is in the areas of system biology and computational proteomics.
Dong Wang is a professor of Department of Bioinformatics at Southern Medical University. His research is in the areas of development and stem cell of computational systems biology.
Wei Chen is a professor of Innovative Institute of Chinese Medicine and Pharmacy at Chengdu University of Traditional Chinese Medicine. His research is in the areas of bioinformatics, computational epigenetics and epitranscriptome.
Hao Lin is a professor of Center for Informational Biology at University of Electronic Science and Technology of China. His research is in the areas of bioinformatics and system biology.
Submitted: 17 September 2019; **Received (in revised form):** 5 November 2019

silence the target genes such as oligonucleotide therapy [7] and macrophage-targeted therapy [8, 9].

The subcellular localization of RNA mainly involved in the following three mechanisms. First of all, RNA is protected from degradation in particular regions. Second, RNA is captured by anchoring protein localized to specific cellular sites during passively diffusing in cytoplasm. Third, RNA is actively transported by the combination of *cis*-elements and specific RNA binding proteins [10]. Specific *cis*-regulatory elements correlated with RNA subcellular localization are usually found in 3′UTR of mRNA and involved in RNA secondary structure [11]. It has been reported that mRNA localization was also affected by mRNA modifications, alternative splicing and polyadenylation [12, 13]. However, the mechanism by which motor proteins transport mRNA remains elusive. Therefore, researches about RNA subcellular localization are full of significance for further illustrating the mechanism of mRNA location.

RNA fluorescent in situ hybridization (RNA-FISH) has been widely used to identify the presence and location of a region of cellular nucleic acids [1, 14]. With the development of high-throughput sequencing technology, microarray [15] and RNA-Seq also provided lots of enrichment information about transcripts, which could be used to study RNA cellular localization [16, 17]. Based on the data generated by these experimental technologies, three databases have been constructed for RNA subcellular location. RNALocate [18] collected subcellular localization entries of all kinds of RNA, lncSLdb [19] and LncATLAS [20] collected localization data of long non-coding RNA (lncRNA). Although those traditional experiments have been successfully used in a variety of settings and provided lots of robust data, the experiments are both time consuming and expensive. Accordingly, new approaches are excepted to fill such a gap. In the past few years, machine-learning approaches have been frequently used to solve various biology problems. Recently, two machine-learning-based methods have been proposed to predict subcellular locations of lncRNA [21, 22], which demonstrated that machine-learning methods have become complements of experimental techniques to detect RNA subcellular location.

In this study, we focus on mRNA of *H. sapiens* due to their close correlation with diseases. Following the steps shown in Figure 1, we developed an effective and powerful model to predict mRNA subcellular location in *H. sapiens*. We firstly collected mRNA subcellular location information and mRNA sequences to construct a qualified benchmark dataset. Secondly, an optimal combination of features was selected by feature extraction and selection technique. Subsequently, the optimal features were inputted into machine-learning method to train, test and build a model. Finally, based on the proposed model, we established a user-friendly web server for mRNA subcellular location prediction.

Materials and methods

Benchmark datasets

Constructing a reliable benchmark dataset is the first important step for building a reliable predictor and understanding intrinsic mechanism of mRNA subcellular localization. The mRNA subcellular locations in *H. sapiens* with experimental evidence were retrieved from RNALocate [18] (RNALocate subcellular location data in EXCEL format were obtained from the RNALocate in September 2017; http://www.rna-society.org/rnalocate/). Assuming that the information gained from the sample with fewer labels is more focused, the samples that have already been detected in more than one location were removed. Figure 2 illustrates the main subcellular locations in the collected dataset. Corresponding mRNA sequences were downloaded from GenBank [23]. As the data might contain homologous sequences, CD-HIT-EST [24] was used for clustering sequences with a cut-off of 80% to reduce data redundancy. Finally, a total of 4901 mRNA sequences were obtained, which belong to four subcellular localizations. A brief summary of mRNA subcellular localization of all samples is displayed in Table 1. The length distribution of each subcellular location can be seen in Figure 3. The average length in the four classes is 3673, 2763, 3692 and 5238, respectively. Since 5-fold cross-validation is an objective way of examining the accuracy of a statistical prediction method, in this study, there is no need to artificially separate the benchmark into the training dataset and a testing dataset.

Feature encoding

Generally, the second step is to transfer sequences into vectors that could reflect the integrity information for samples [25–29]. Various feature-encoding techniques have been proposed for RNA sequence analysis [30, 31], for example, k-tuple (also called k-mer) nucleotide composition [32], pseudo K-tuple nucleotide composition [33], position-correlation scoring function [34] and binary encoding [35] have been used to formulate sequences. Let the mRNA $\mathbf{S}$ expressed as following:

$$\mathbf{S} = R_1R_2R_3R_4R_5 \ldots R_iR_{i+1} \ldots R_L \tag{1}$$

where L denotes the length of the mRNA sequence and R_i is the i-th base and $R_i \in \{A, G, C, T\}$.

By using k-tuple nucleotide composition, a primary sequence $\mathbf{S}$ can be transferred into a vector $\mathbf{V}$ with 4^kelements according to the following formula:

$$\mathbf{V} = \left[f_1^{k-tuple} f_2^{k-tuple} \cdots f_i^{k-tuple} \cdots f_{4^k}^{k-tuple}\right]^T \tag{2}$$

where the symbol $\mathbf{T}$ means the transposition of a vector and $f_i^{k-tuple}$ is the normalized frequency of the i-th k-tuple nucleotide component occurring in $\mathbf{S}$ and can be calculated by

$$f_i^{k-tuple} = \frac{n_i}{\sum_{i=1}^{4^k} n_i} = \frac{n_i}{L-k+1} \tag{3}$$

where n_i means the number of occurrences of the i-th k-tuple nucleotide component in the mRNA sequence $\mathbf{S}$.

Recent researches showed that conversed nonamer participates biological processes, such as DNA cleavage and DNA synapsis [36–38], suggesting that nonamer may have a unique evolutionary mechanism. Therefore, in this study, we used nonamer composition to represent mRNA sequence. Let the number of k equal to 9, the dimension of vector $\mathbf{V}$ is

$$\mathbf{\Gamma} = 4^k = 4^9 = 262\ \ 144 \tag{4}$$

Feature selection

To exclude noise and improve computational efficiency, feature selection is an indispensable step. As shown in Equation (4), the dimension of the vector is 262 144, which may lead to the large computation, overfitting and low robust of proposed model. The two-step feature selection method is a novel discriminated strategy to identify informative features. Several experiments have demonstrated that two-step feature selection could improve the prediction performance and reduce calculation

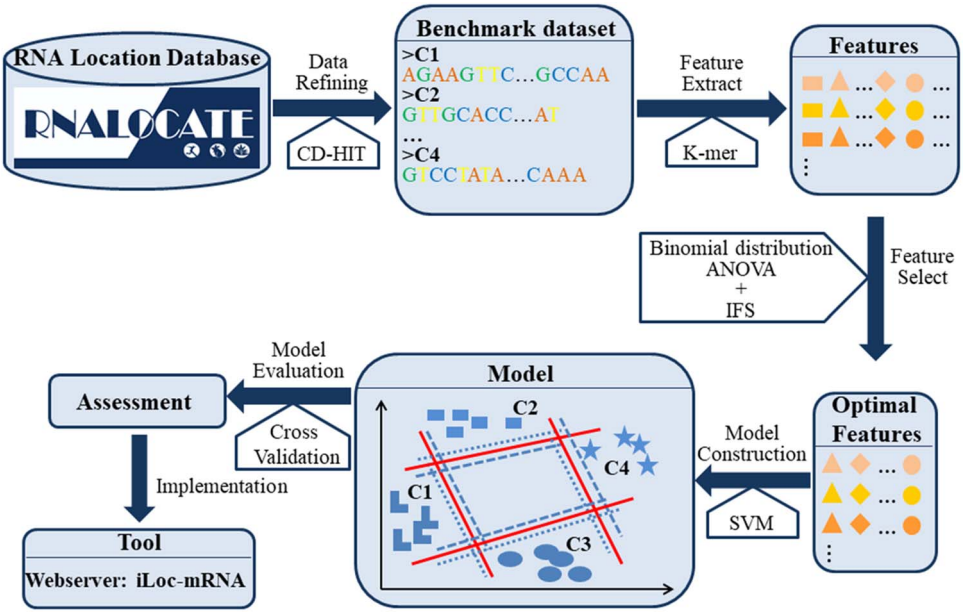

Figure 1. The flow chart of developing the model for subcellular location prediction of *H. sapiens*.

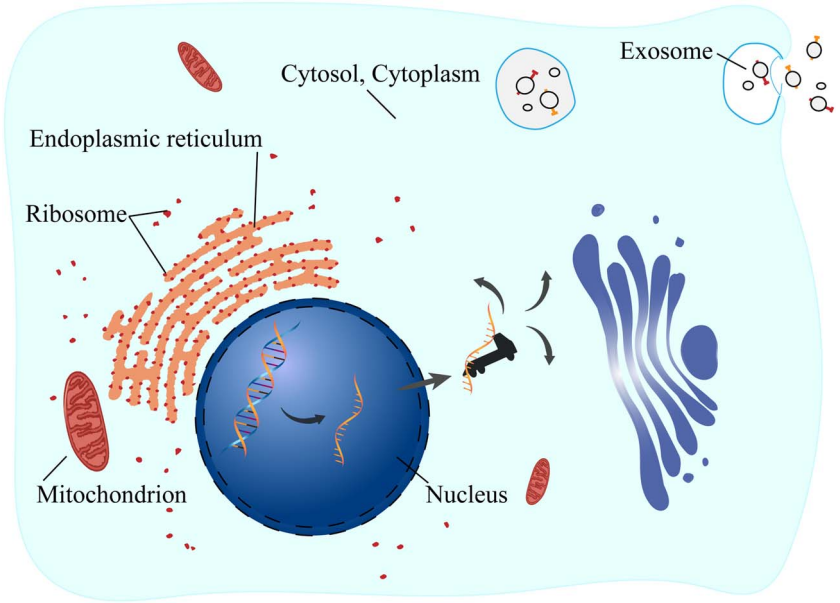

Figure 2. A schematic to show the locations of mRNA in a cell.

Table 1. The detail information of benchmark dataset for mRNA subcellular location prediction

Class	Subcellular locations	Before 80% cut-off	After 80% cut-off
C1, *n*(%)	Cytosol Cytoplasm	1992(39.6)	1954(39.9)
C2, *n*(%)	Ribosome	1740(34.6)	1666(34.0)
C3, *n*(%)	Endoplasmic reticulum	937(18.6)	924(18.9)
C4, *n*(%)	Nucleus Exosome Dendrite Mitochondrion	367(7.3)	357(7.3)
Total		5036	4901

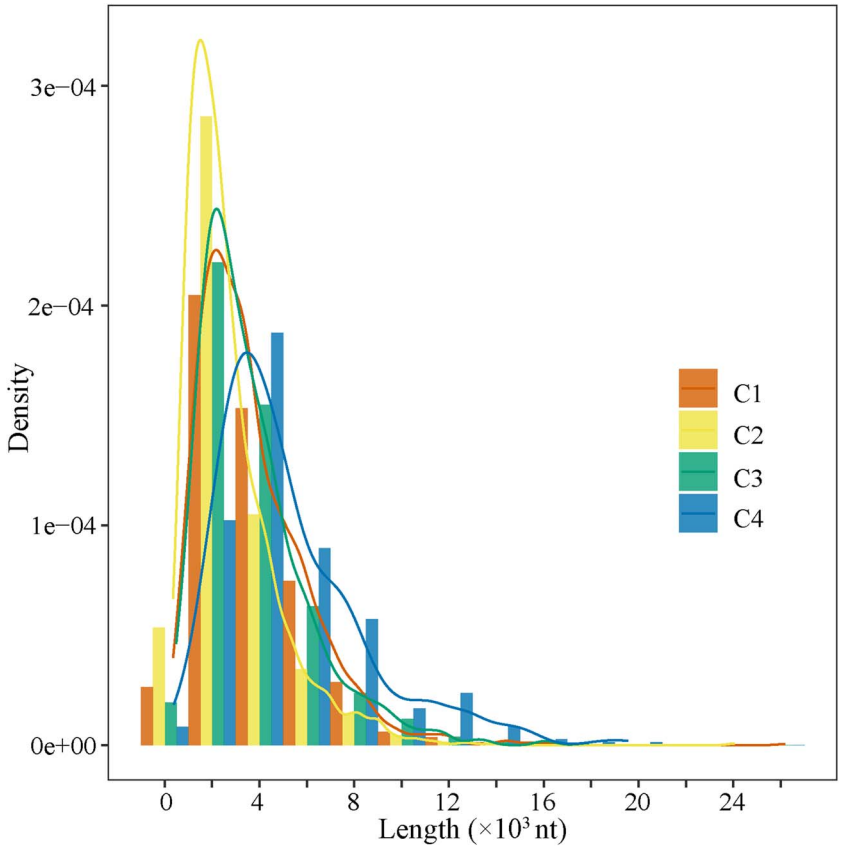

Figure 3. Length distribution of mRNA sequences.

time [39, 40]. The good performance of two-step feature selection method indicated that the two-step strategy could be extended to other fields. Binomial distribution is one of the wonderful feature selection techniques that have been successfully applied in many works [41–46]. In this study, binomial distribution combined with ANOVA was used to winnow out the irrelevant features.

Binomial distribution

The prior probability of nonamer appears in one type of class can be calculated by

$$q_j = m_j / M \tag{5}$$

where m_j is the total number of nonamers that appeared in the j-th type of class and M is the total occurrence frequency of all nonamers in the benchmark dataset.

The probability of the i-th nonamer occurring in the j-th class of location can be calculated by the following formula:

$$p\left(n_{ij}\right) = \sum_{m=n_{ij}}^{N_i} \frac{N_i!}{m!\left(N_i - m\right)!} {q_j}^m \left(1 - q_j\right)^{N_i - m} \tag{6}$$

where N_i is the total count of the i-th nonamer in the benchmark dataset, n_{ij} represents the number of occurrences of the i-th nonamer in the j-th type of location, and the sum in Equation (6) is taken from n_{ij} to N_i. The confidence level of the i-th nonamer in the j-th class can be given by

$$CL_{ij} = 1 - p\left(n_{ij}\right) \tag{7}$$

In this work, there are four subcellular locations, so each nonamer has four confidence level (CL) values. We assigned the largest one to be the CL value of the i-th nonamer

$$CL_i = \max\left(CL_{i1}, CL_{i2}, CL_{i3}, CL_{i4}\right) \tag{8}$$

ANOVA

ANOVA was used to analyze the differences among group means and their associated procedures. In ANOVA, the statistical significance is tested by the ratio of the variance between groups and the variance within the groups:

$$F(\mathrm{i}) = \frac{S_B^2(\mathrm{i})}{S_W^2(\mathrm{i})} \tag{9}$$

where $F(\mathrm{i})$ represents the F-score of the i-th feature, $S_B^2(\mathrm{i})$ denotes the sample variance between groups (also called means square between, MSB), $S_W^2(\mathrm{i})$ denotes the sample variable within groups (also called means square within, MSW). $S_B^2(\mathrm{i})$ and $S_W^2(\mathrm{i})$ are, respectively, expressed as:

$$S_B^2(\mathrm{i}) = \frac{\sum_{j=1}^{K} m_j \left(\sum_{s=1}^{m_j} f_{\mathrm{i}}\left(\mathrm{s}, j\right) / m_j - \sum_{j=1}^{K} \sum_{s=1}^{m_j} f_{\mathrm{i}}\left(\mathrm{s}, j\right) / \sum_{j=1}^{K} m_j\right)^2}{K - 1} \tag{10}$$

$$S_W^2(\mathrm{i}) = \frac{\sum_{j=1}^{K} \sum_{s=1}^{m_j} \left(f_{\mathrm{i}}\left(\mathrm{s}, j\right) - \sum_{s=1}^{m_j} f_{\mathrm{i}}\left(\mathrm{s}, j\right) / m_j\right)^2}{N - K} \tag{11}$$

where K and$N$, respectively, denote the number of classes and the total number of samples in the benchmark dataset. $f_{\mathrm{i}}(\mathrm{s}, j)$ represents the frequency of the i-th feature of the s-th sample

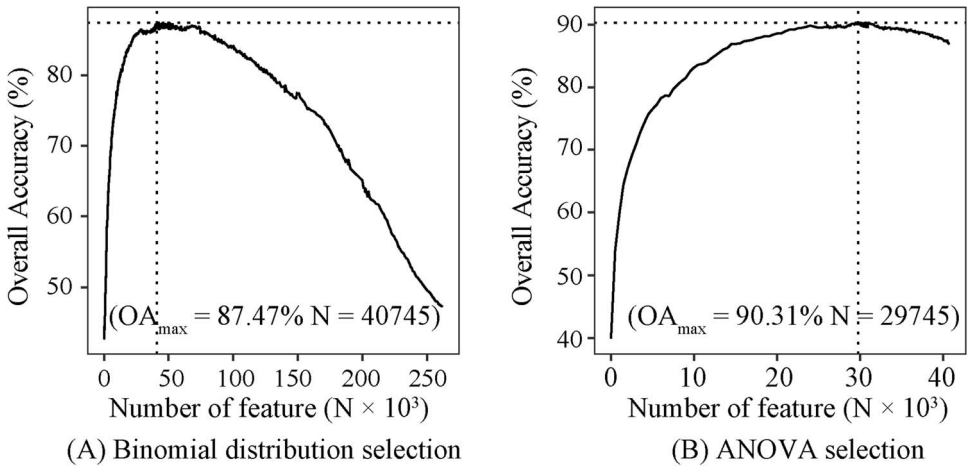

Figure 4. IFS accuracy curve for (**A**) binomial distribution feature selection. (**B**) ANOVA feature selection.

in the j-th location. $m_j$denotes the total number of samples in the j-th location. It is obvious that the larger the $F(i)$ value, the better the discriminative capability the i-th feature has.

Incremental feature selection

After ranking the features according to their statistical scores, incremental feature selection (IFS) strategy was adopted to determine the optimal feature set. The IFS strategy added features one by one to feature set from higher to lower ranked score. Once a new feature was added, a new feature set was composed. Given n ranked features, the j-th feature subset is formulated as:

$$F_i = \{1, 2, \ldots f_i\} \; i \in [1, n] \tag{12}$$

Classification algorithm is then performed on every feature subset to train and test prediction models. Finally, the optimal feature set is defined based on the principle that the prediction model based on such features could achieve the maximum accuracy. In this work, we used a two-step strategy, in which binomial distribution feature selection with IFS was initially used to roughly optimize features, and, subsequently, ANOVA with IFS was performed to pick out the optimal feature subset from selected features obtained by binomial distribution. For the convenience of researchers, a feature-ranking tool based on binomial distribution was developed, it can be freely downloaded at https://github.com/ZhaoyueZhang/BinomialDistribution.

Support vector machine

Supervised classification is one of the tasks frequently carried out by explosive growth of multiple data. **Support vector machine (**SVM) is one kind of non-linear models that can be applied to supervised classification or regression. It has solved many bioinformatics prediction problems successfully [47–51]. SVM maps the samples into a high-dimension feature space so that different categories of examples can be divided by a maximum-margin hyperplane. Despite SVM is highly insensitive to the curse of dimensionality, training and prediction are very expensive for large and complex problems. ThunderSVM exploits graphics processing units and multi-core central processing unit to help users easily and efficiently apply SVMs [52]. This study used ThunderSVM with radial basis function to perform multi-class classification. The ThunderSVM employed the grid search method with 5-fold cross-validation to seek the best penalty coefficient c of soft margin SVM and width parameter γ of Gaussian kernel function. The searching space is as following:

$$\begin{cases} c \in [\,2^{-5}, 2^{15}], step = 2 \\ \gamma \in [\,2^{-15}, 2^{3}], step = 2^{-1} \end{cases} \tag{13}$$

Performance evaluation metrics

Performance measurement in multiclass classification is different from traditional binary classification. In this study, four evaluation metrics was used to evaluate the performance of different approaches in the field of mRNA subcellular localization [53–55]. The four indexes namely the overall accuracy (OA), sensitivity (*Sn*), specificity (*Sp*) and Matthews correlation coefficient (*MCC*) were formulated as

$$Sn(j) = 1 - \frac{N_-^+(j)}{N^+(j)} \; 0 \le Sn(j) \le 1 \tag{14}$$

$$Sp(j) = 1 - \frac{N_+^-(j)}{N^-(j)} \; 0 \le Sp(j) \le 1 \tag{15}$$

$$MCC(j) = \frac{1 - \left(\frac{N_-^+(j)}{N^+(j)} + \frac{N_+^-(j)}{N^-(j)}\right)}{\sqrt{\left(1 + \frac{N_+^-(j) - N_-^+(j)}{N^+(j)}\right)\left(1 + \frac{N_-^+(j) - N_+^-(j)}{N^-(j)}\right)}} - 1 \le MCC(j) \le 1 \tag{16}$$

$$OA = \frac{1}{N}\sum_{i=1}^{K}\left[N^+(j) - N_-^+(j)\right] \; 0 \le OA \le 1 \tag{17}$$

where $N^+(j)$ is the total number of samples in the j-th class, $N_-^+(i)$ is the number of the samples in $N^+(j)$ that are incorrectly predicted to be of other classes, $N^-(j)$ is the total number of samples in any other class except the j-th class and $N_+^-(i)$ is the number of the samples in $N^-(j)$ that are incorrectly predicted to be of the j-th class.

Result

Performance evaluation metrics

As all samples were projected into a high-dimension feature space generated by 9-tuple feature extraction method, two-step feature selection strategies were performed to reduce feature

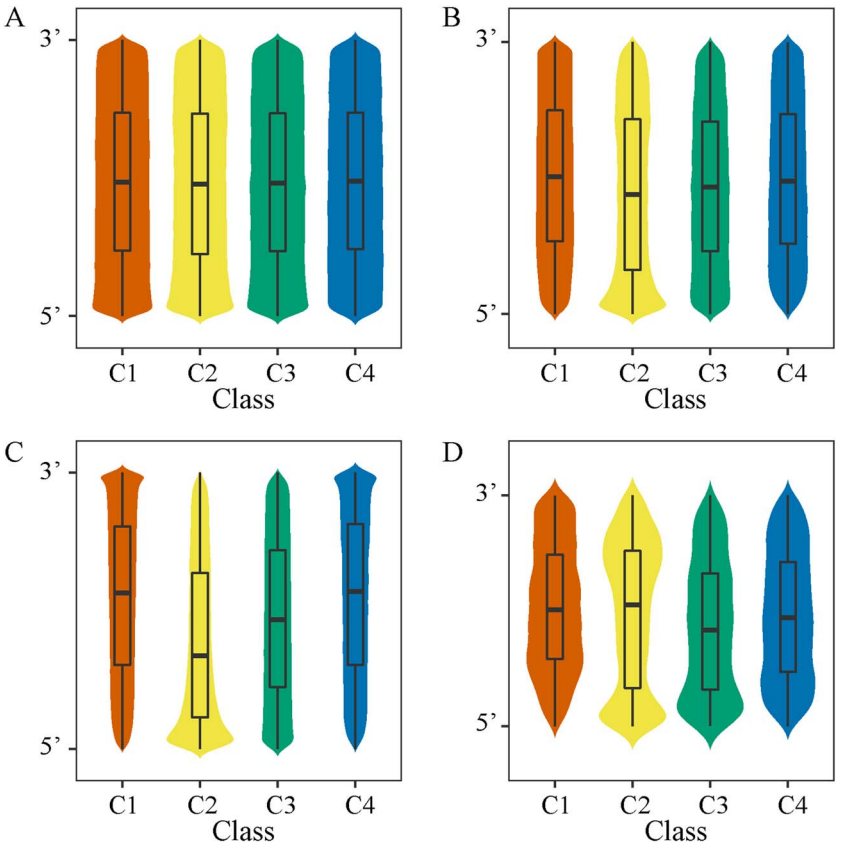

Figure 5. Analysis of distribution pattern of features. (**A**) The 29 745 nonamer distributions in four classes. (**B**) The class-specific nonamer distribution in four classes. (**C**) The class-specific motifs distribution in four classes. (**D**) The position-specific nonamer distribution in four classes.

Table 2. The performance of the SVM-based subcellular location prediction model

Class	iLoc-mRNA			
	Sn(%)	*Sp*(%)	*MCC*	*OA*(%)
C1	90.94	94.74	0.858	90.12
C2	90.58	94.37	0.846	
C3	89.72	97.56	0.872	
C4	84.59	98.90	0.840	

dimension. The evaluation was conducted through a 5-fold cross-validation with ThunderSVM. Firstly, 262 144 nonamer composition features were ranked by binomial distribution score. With the IFS technique, the maximum prediction accuracy of 87.47% was observed when the top 40 745 features were used to construct prediction model (Figure 4A). Subsequently, these 40 745 features were rearranged by ANOVA. When the top 29 745 ANOVA-ranked features were used to train and test the SVM model during IFS, the model could achieve the best accuracy of 90.31% (Figure 4B) with the parameter c of 2^{13} and γ of 2^{-15}. Thus, we constructed an mRNA subcellular location classifier based on the 29 745 features with the best penalty coefficient of 2^{13} and width parameter of 2^{-15}.

We called the classifier as iLoc-mRNA, which stands for 'identify or predict subcellular location of mRNAs'. The performances of the iLoc-mRNA with 5-fold cross-validation test are shown in Table 2.

Features analysis

Motif analysis was performed to mining the hidden information about the mRNA subcellular localization behind primary sequences. Firstly, all the 29 745 nonamers that were used to construct iLoc-mRNA were mapped to every mRNA sequences to get the relative position of mRNA sequences. The relative position of features on sequences is present with violin plot by using R package 'ggplot2' [56]. It is obvious that there is no special distribution pattern (Figure 5A). Then, 29 745 nonamers were assigned to four subcellular classes according to their binomial distribution *CL* value (a feature is classified to *j*-type feature if the CL_{ij} is the maximum among four CL_{ij} values), which indicated that those features may contribute to the *j*-type class subcellular localization. As a result, 6781, 7751, 7318 and 7895 nonamers were assigned to four classes, respectively. Though studies report that 3' UTR is of vital importance to mRNA subcellular localization [57], the relative position of nonamer has no obvious local enrichment trend in 3' UTR (Figure 5B). Finally, Discriminative Regular Expression Motif Elicitation (DREME) [58] was used to perform motif discovery on the class-specific nonamers of each class. DREME finds ten significant motifs in the class-specific nonamer data for cytosol/cytoplasm ([(A/T)(A/T/G)(A/T)], [GGGGGG(A/G)]), ribosome ([(T/G/C)(G/C)CG], [(G/C)C(G/C)], [CG(A/T)G]), endoplasmic reticulum ([G(T/G)(G/C)], [CC(A/T)], [CATC]) and nucleus/exosome/dendrite/mitochondrion ([(A/T)(A/T/G/C)(A/T)(A/T)], [(A/T)A(A/T)]) (Figure 6). Class-specific motifs were mapped to corresponding sequences (Figure 5C). The distribution pattern of the ten motifs reflected that although the *cis*-regulatory motifs modulating mRNA

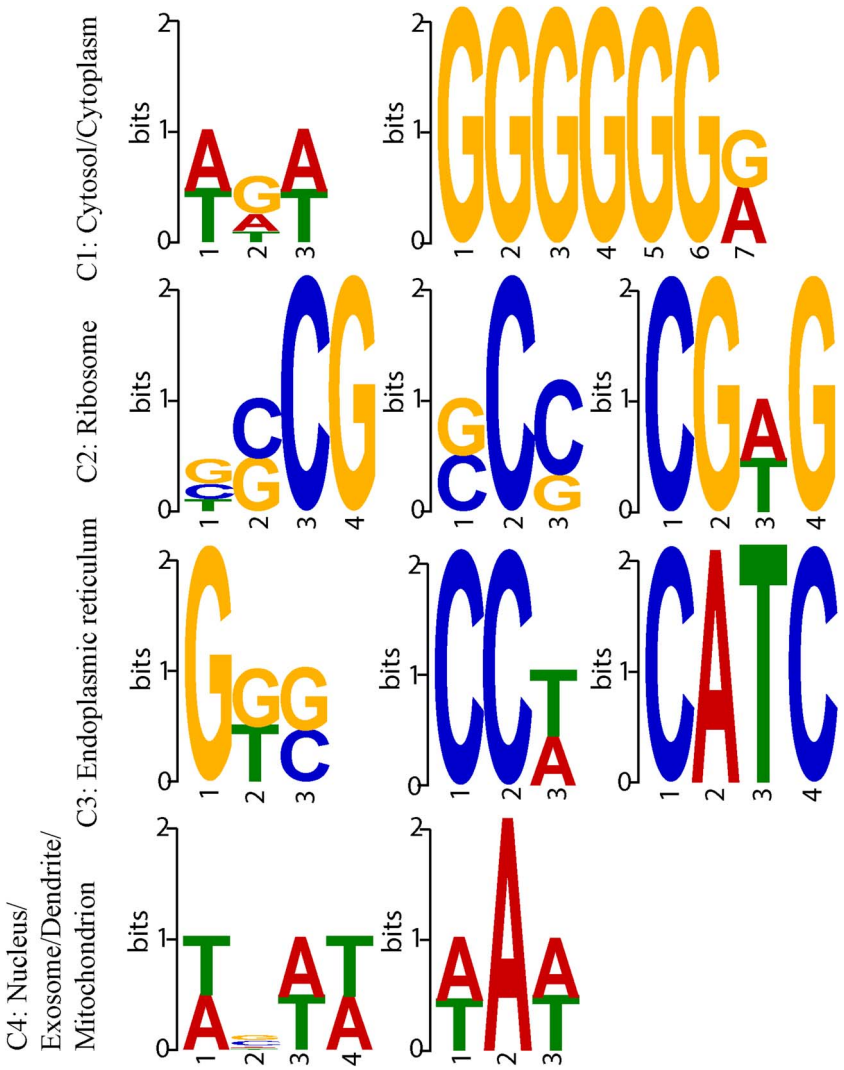

Figure 6. Visualization of class-specific sequence motifs.

localization have most often been characterized in the 3' UTR of transcripts, but also reside in the 5' UTR, coding regions of the mRNA [59].

The top 1% nonamers (297 nonamers) in ANOVA rank were selected to further analyze the distribution pattern of motifs along sequence. Each sequence was split to the first quartile region, inter-quartile range region and the third quartile region. The statistical differences of 297 nonamers in different regions were investigated by using binomial distribution with $CL > 95\%$. Results showed that a total of 140 nonamers display region bias, among which 46 nonamers prefer to locate in the first quartile region, 41 locate in inter-quartile range region and 53 in the third quartile region, respectively. The result indicated the scattered distribution pattern of *cis*-regulatory motifs modulating mRNA localization. The distribution pattern of these 140 sequence position-specific nonamers is shown in Figure 5D.

Comparison with other work

Recently, a predictor called RNATracker has been developed to predict the distributions of mRNA transcripts over subcellular compartments [60]. Relative expression value measured by RNA-Seq was used to annotate the mRNA subcellular location. Transcript of mRNA was transformed using one-hot encoding and then trained variants of RNATracker. Among the variants versions, $\text{RNATracker}_{\text{seq}}$, which used full-length sequences, achieved the best performance. The $\text{RNATracker}_{\text{seq}}$ get the AUC of 0.851, 0.787, 0.667 and 0.748 for cytosol, insoluble, membrane and nucleus by using 10-fold cross-validation. Due to the different datasets and modeling approaches, here, we just make a simple comparison. We performed 10-fold cross-validation and get the AUC of 0.983, 0.981, 0.987 and 0.987 for cytosol/cytoplasm, ribosome, endoplasmic reticulum and nucleus/exosome/dendrite/mitochondrion (Figure 7), respectively. It was noticed that the AUC of iLoc-mRNA for predicting cytosol/cytoplasm is superior to that of $\text{RNATracker}_{\text{seq}}$.

Web server

Although the RNATracker could perform mRNA subcellular location prediction [60], there is no web server based on the predictor, which will prevent scholars without computer or mathematics background from using the tool. For the convenience of researchers, a web server was established based on our work.

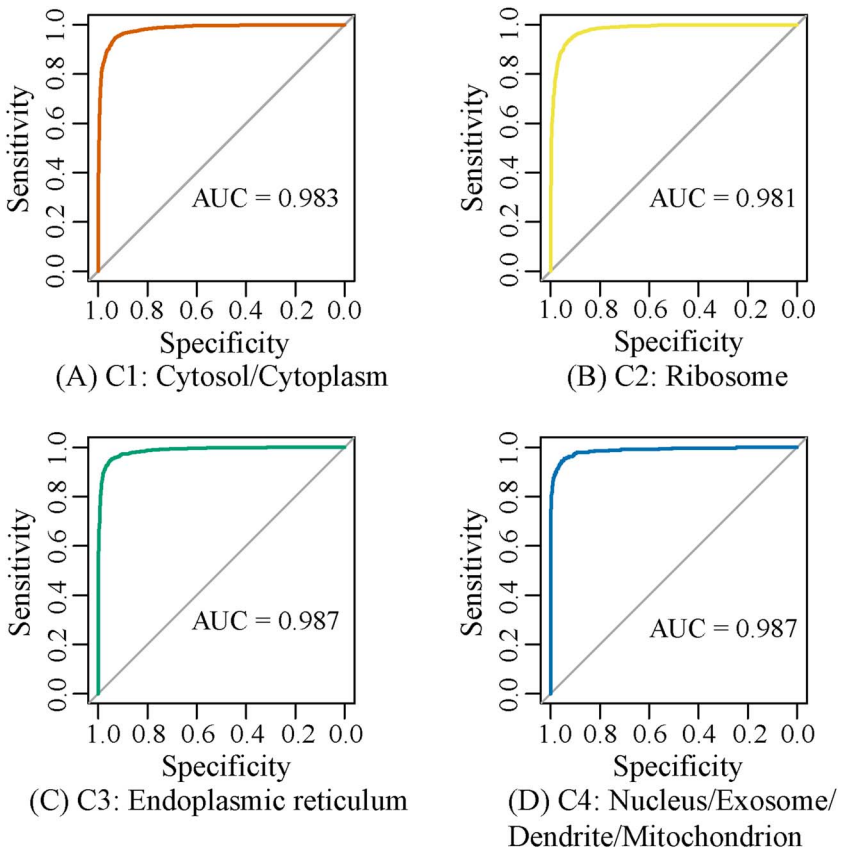

Figure 7. ROC curves for the iLoc-mRNA for each fraction.

The iLoc-mRNA provided online mRNA subcellular location prediction service, which only requires sequence in FASTA format. iLoc-mRNA is first web server for mRNA subcellular location prediction and is free available at http://lin-group.cn/server/iLoc-mRNA/.

Discussion

The subcellular spatial distribution pattern of mRNA in cell could provide much knowledge to understand the regulation and function of mRNA as well as related diseases [6]. In this study, we proposed a powerful predictor for identifying *H. species* mRNA subcellular location and established an online web server named iLoc-mRNA. The web server could give the potential target subcellular location of *H. species* mRNA only based on sequence information, which is very simple and flexible. The feature analysis discovered ten mRNA subcellular localization associated motifs, which are spread in the whole region of the mRNA sequence.

However, the prediction of subcellular localization of biomacromolecule is still a challenging task, which needs more reliable information source. Newly developed techniques like APEX-seq [17] will accelerate unveiling the biology progresses of RNA subcellular localization. We will stay focus on the RNA subcellular location issues and make more efforts on the prediction of RNA subcellular location and the identification of associated motifs. According to some latest RNA sequence classification works, we will consider to try deep learning [61–63], feature fusion [64] and ensemble learning [65] techniques in the future.

Key Points

- The background and identification methods for mRNA localization were comprehensively summarized.
- A high-quality data set was built to train a new prediction model for predicting mRNA subcellular location.
- Motif analysis was performed to investigate the *cis*-regulatory motifs modulating mRNA localization.
- A user-friendly web server was developed to for predicting mRNA subcellular location.

Funding

The National Nature Scientific Foundation of China (61772119, 31771471, 81770104) and the Natural Science Foundation of Guangdong Province (2019A1515010784).

References

1. Meyer C, Garzia A, Tuschl T. Simultaneous detection of the subcellular localization of RNAs and proteins in cultured cells by combined multicolor RNA-FISH and IF. *Methods* 2017;**118-119**:101–10.
2. Ephrussi A, Dickinson LK, Lehmann R. Oskar organizes the germ plasm and directs localization of the posterior determinant nanos. *Cell* 1991;**66**:37–50.

3. Liu D, Li G, Zuo Y. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief Bioinform* 2018;**06**:1–10.
4. Mili S. Macara IG. RNA localization and polarity: from a(PC) to Z(BP). *Trends Cell Biol* 2009;**19**:156–64.
5. Katz ZB, Wells AL, Park HY, *et al*. Beta-actin mRNA compartmentalization enhances focal adhesion stability and directs cell migration. *Genes Dev* 2012;**26**:1885–90.
6. Lin Y, Liu T, Cui T, *et al*. RNAInter in 2020: RNA interactome repository with increased coverage and annotation. *Nucleic Acids Res* 2020;**48**:D189–97.
7. Didiot MC, Ferguson CM, Ly S, *et al*. Nuclear localization of Huntingtin mRNA is specific to cells of neuronal origin. *Cell Rep* 2018;**24**:2553–2560 e2555.
8. Pelekanou V, Villarroel-Espindola F, Schalper KA, *et al*. CD68, CD163, and matrix metalloproteinase 9 (MMP-9) co-localization in breast tumor microenvironment predicts survival differently in ER-positive and -negative cancers. *Breast Cancer Res* 2018;**20**:154.
9. Liu H, Zhang W, Zou B, *et al*. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res* 2020;**48**:D871–81.
10. Taliaferro JM, Wang ET, Burge CB. Genomic analysis of RNA localization. *RNA Biol* 2014;**11**:1040–50.
11. Ciolli Mattioli C, Rom A, Franke V, *et al*. Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic Acids Res* 2019;**47**:2560–73.
12. Peer E, Moshitch-Moshkovitz S, Rechavi G, *et al*. The Epitranscriptome in translation regulation. *Cold Spring Harb Perspect Biol* 2018;**11**.
13. Taliaferro JM, Vidaki M, Oliveira R, *et al*. Distal alternative last exons localize mRNAs to neural projections. *Mol Cell* 2016;**61**:821–33.
14. Chen J, McSwiggen D, Unal E. Single molecule fluorescence in situ hybridization (smFISH) analysis in budding yeast vegetative growth and meiosis. *J Vis Exp* 2018, doi: 10.3791/57774.
15. Poon MM, Choi SH, Jamieson CA, *et al*. Identification of process-localized mRNAs from cultured rodent hippocampal neurons. *J Neurosci* 2006;**26**:13390–9.
16. Fagerberg L, Hallstrom BM, Oksvold P, *et al*. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 2014;**13**:397–406.
17. Fazal FM, Han S, Parker KR, *et al*. Atlas of subcellular RNA localization revealed by APEX-Seq. *Cell* 2019;**178**:473–90.
18. Zhang T, Tan P, Wang L, *et al*. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res* 2017;**45**:D135–8.
19. Wen X, Gao L, Guo X, *et al*. lncSLdb: a resource for long non-coding RNA subcellular localization. *Database (Oxford)* 2018;**2018**:1–6.
20. Mas-Ponte D, Carlevaro-Fita J, Palumbo E, *et al*. LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* 2017;**23**:1080–7.
21. Cao Z, Pan X, Yang Y, *et al*. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics* 2018;**34**:2185–94.
22. Su ZD, Huang Y, Zhang ZY, *et al*. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 2018;**34**:4196–204.
23. Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* 2017;**45**:D37–42.
24. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
25. Zuo Y, Li Y, Chen Y, *et al*. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 2017;**33**:122–4.
26. Song J, Li F, Leier A, *et al*. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 2018;**34**:684–7.
27. Manavalan B, Subramaniyam S, Shin TH, *et al*. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J Proteome Res* 2018;**17**:2715–26.
28. Manavalan B, Govindaraj RG, Shin TH, *et al*. iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front Immunol* 2018;**9**:1695.
29. Chen Z, Zhao P, Li FY, *et al*. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502.
30. Liu B, Gao X, Zhang H. 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 Res* 2019;**47**:e127.
31. Yang J, Chen X, McDermaid A, *et al*. DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics* 2017;**33**:2586–8.
32. Lai HY, Zhang ZY, Su ZD, *et al*. iProEP: a computational predictor for predicting promoter. *Mol Ther Nucleic Acids* 2019;**17**:337–46.
33. Feng CQ, Zhang ZY, Zhu XJ, *et al*. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 2019;**35**:1469–77.
34. Lin H, Liang ZY, Tang H, *et al*. Identifying Sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**:1316–21.
35. Lv H, Zhang ZM, Li SH, *et al*. Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief Bioinform* 2019, doi: 10.1093/bib/bbz048.
36. Yin FF, Bailey S, Innis CA, *et al*. Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. *Nat Struct Mol Biol* 2009;**16**:499–508.
37. Raveendran D, Raghavan SC. Biochemical characterization of Nonamer binding domain of RAG1 reveals its thymine preference with respect to length and position. *Sci Rep* 2016;**6**:19091.
38. Ru H, Zhang P, Wu H. Structural gymnastics of RAG-mediated DNA cleavage in V(D)J recombination. *Curr Opin Struct Biol* 2018;**53**:178–86.
39. Dao FY, Lv H, Wang F, *et al*. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* 2019;**35**:2075–83.
40. Song J, Wang Y, Li F, *et al*. iProt-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief Bioinform* 2019;**20**:638–58.
41. Yang H, Yang W, Dao FY, *et al*. A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief Bioinform* 2019, doi: 10.1093/bib/bbz123.
42. Zhu XJ, Feng CQ, Lai HY, *et al*. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl-Based Syst* 2019;**163**:787–93.

43. Long C, Li W, Liang P, *et al*. Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis. 2019;**7**:7794–802.
44. Yu L, Sun X, Tian SW, *et al*. Drug and nondrug classification based on deep learning with various feature selection strategies. *Curr Bioinforma* 2018;**13**:253–9.
45. Wei L, Xing P, Shi G, *et al*. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**:1264–73.
46. Liu B. BioSeq-analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief Bioinform* 2019;**20**:1280–94.
47. Liao ZJ, Li DP, Wang XR, *et al*. Cancer diagnosis through IsomiR expression with machine learning method. *Curr Bioinforma* 2018;**13**:57–63.
48. Chao L, Wei L, Zou Q. SecProMTB: a SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set. *Proteomics* 2019;**19**:e1900007.
49. Chao L, Jin S, Wang L, *et al*. AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front Bioeng Biotechnol* 2019;**7**:224.
50. Liu B, Li C, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief Bioinform* 2019, doi: 10.1093/bib/bbz098.
51. Deng L, Wang J, Zhang J. Predicting gene ontology function of human MicroRNAs by integrating multiple networks. *Front Genet* 2019;**10**:3.
52. Wen ZY, Shi JS, Li QB, *et al*. ThunderSVM: a fast SVM library on GPUs and CPUs. *J Mach Learn Res* 2018;**19**:1–5.
53. Manavalan B, Shin TH, PVP-SVM LG. Sequence-based prediction of phage Virion proteins using a support vector machine. *Front Microbiol* 2018;**9**:476.
54. Tang H, Cao RZ, Wang W, *et al*. A two-step discriminated method to identify thermophilic proteins. *Int J Biomath* 2017;**10**:1750050.
55. Liu B, Han L, Liu X, *et al*. Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**:1211–8.
56. Ginestet C. ggplot2: elegant graphics for data analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society* 2011;**174**:245–5.
57. Xu L, Peng L, Gu TL, *et al*. The 3' UTR of human MAVS mRNA contains multiple regulatory elements for the control of protein expression and subcellular localization. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms* 2019;**1862**:47–57.
58. Bailey TL. DREME motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011;**27**:1653–9.
59. Bergalet J, Lecuyer E. The functions and regulatory principles of mRNA intracellular trafficking. *Syst Bio of RNA Binding Proteins* 2014;**825**:57–96.
60. Yan ZC, Lecuyer E, Blanchette M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics* 2019;**35**:I333–42.
61. Zou Q, Xing P, Wei L, *et al*. Gene2vec: gene subsequence embedding for prediction of mammalian N6-Methyladenosine sites from mRNA. *RNA* 2019;**25**:205–18.
62. Stephenson N, Shane E, Chase J, *et al*. Survey of machine learning techniques in drug discovery. *Curr Drug Metab* 2019;**20**:185–93.
63. Cao R, Freitas C, Chan L, *et al*. ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 2017;**22**:1732.
64. Chen W, Feng P, Song X, *et al*. iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol Ther Nucleic Acids* 2019;**18**:269–74.
65. Ru X, Cao P, Li L, *et al*. Selecting essential MicroRNAs using a novel voting method. *Mol Ther Nucleic Acids* 2019;**18**:16–23.