IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 24, 2020, accepted July 1, 2020, date of publication July 6, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007317

lncLocPred: Predicting LncRNA Subcellular Localization Using Multiple Sequence Feature Information

YONGXIAN FAN, MEIJUN CHEN, AND QINGQI ZHU
School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

Corresponding author: Yongxian Fan (yongxian.fan@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61762026 and Grant 61462018, in part by the Guangxi Natural Science Foundation under Grant 2017GXNSFAA198278 and Grant 2016GXNSFAA380043, and in part by the Innovation Project of GUET Graduate Education under Grant 2018YJCX47 and Grant 2019YCXS056.

ABSTRACT Determining the subcellular localization of long non-coding RNAs (lncRNAs) provides very favorable references to discover the function of lncRNAs. Instead of through time-consuming and expensive biochemical experiments, we develop a machine learning predictor based on logistic regression, lncLocPred, to predict the subcellular localization of lncRNAs. We adopt sequence features including k-mer, triplet, and PseDNC and systematically process feature selection through VarianceThreshold, binomial distribution, and F-score to obtain representative features. We observe that the top-ranked k-mers have a higher base content of G and C in the form of short repeats. Improving prediction accuracy on several subcellular localizations, our model achieves the highest overall accuracy of 92.37% on the benchmark dataset by jackknife, higher than the existing state-of-the-art predictors. Additionally, lncLocPred performs better for the prediction on an independent dataset collected by us as well. Related experimental data and source code are available at https://github.com/jademyC1221/lncLocPred.

INDEX TERMS Feature engineering, sequence features, subcellular localization of lncRNAs.

I. INTRODUCTION

Recently, long non-coding RNAs (lncRNAs) with more than 200 nucleotides [1] have become a research hotspot, whose number reaches about 20000 estimated by ENCODE [2] or FANTOM5 [3]. They make a variety of important biological functions affecting different biological processes [4]–[6]. Alterations in the expression level of lncRNAs and up-regulation or down-regulation of a novel lncRNA have been shown to be the diagnostic marker for several types of diseases and cancers, which contributes to the proposal of therapeutic strategies [7]–[9]. Most of their functions have yet to be discovered cause their detailed functional mechanisms remain unclear [10]. Of particular note is that different subcellular localization patterns of lncRNAs enable them to perform their assigned functions [11]–[13]. LncRNAs located in the nucleus perform their regulatory functions including chromatin organization, transcriptional and post-transcriptional gene expression, and act as structural scaffolds of nucleus domains [14], [15]. For example, morrbid, a nuclear-localized lncRNA, playing a regulatory role in the apoptosis of short-lived myeloid cells, can facilitate the diagnosis and treatment of inflammatory diseases [16]. Activity of lncRNA XIST affects the methylation level of TIMP-3 promoter and thus affects the degradation of collagen in osteoarthritis chondrocytes after tibial plateau fracture [17]. Cytoplasmic lncRNAs interfere with post-translational modification of proteins and affect gene regulation [18]–[20]. For example, as a tumor-promoting factor to enhance tumorigenesis of GBC cells, lncRNA-HGBC can be used as a target for GBC therapy [21]. LncRNA RP11-732M18.3 promotes glioma growth through the interaction with the protein of 14-3-3 β/α and provides clues for the treatment of glioma [22]. Obviously, the subcellular location of lncRNAs provides valuable clues for their biological functions.

Currently, databases related to lncRNA subcellular localization include RNALocate [23], LncATLAS [12], and lncSLdb [24]. To date, RNALocate has contained experimental results for 190,000 entries of RNA subcellular localizations. LncATLAS has collected 6,768 GENCODE-annotated

The associate editor coordinating the review of this manuscript and approving it for publication was Marcin Woźniak.

lncRNAs in different compartments of 15 cell lines with the help of relative concentration index. LncSLdb has collected lncRNA subcellular localization information for more than 11,000 transcripts from three species. Although methods to predict the subcellular localization of lncRNAs are limited compared with that of proteins [25]–[29], some achievements have been made so far. The existing methods to identify subcellular localization fall into two categories, biochemical experiments and computational methods. In spite of the fact that biochemical experiments such as fluorescence in situ hybridization (FISH) [12], [30], [31] report convincing results about the subcellular localization of lncRNAs, they have the disadvantages of being time-consuming, low in yield and requiring expensive reagents. To overcome these disadvantages, we need to develop more effective computational methods. There are, however, limited existing computational methods to predict the subcellular localization of lncRNAs. Cao *et al.* developed a predictor, lncLocator, to predict the subcellular location of lncRNAs through high level features from stacked autoencoder and an ensemble learning method [32]. Su *et al.* developed iLoc-lncRNA by integrating octamer into PseKNC and using SVM model to identify the subcellular location of lncRNAs [33]. The prior method used novel features derived from autoencoder but was not ideal in terms of accuracy. The latter method greatly improved the overall accuracy, but the prediction in the subcellular locations, such as nucleus and ribosome, needs further improvement. After that, Ahsan Ahmad *et al.* proposed Locate-R based on locally deep support vector machines with features of n-gapped l-mer composition and l-mer composition, which significantly improved the prediction performance on subcellular location of ribosome and exosome [34]. Unlike the former three predictors predicting multiple subcellular locations of lncRNAs on a small dataset, Gudenas et.al developed DeepLncRNA with a deep learning algorithm to predict two subcelluar locations, nucleus and cytosolic, on a big dataset [35]. Based on the foregoing methods, we are committed to proposing a method to predict the subcellular localization of lncRNA in multiclassification that can not only improve the overall accuracy, but also improve the accuracy in the subcellular locations with a smaller sample size.

In this study, we considered several sequence-derived features and proposed a logistic regression-based method to predict the subcellular localization of lncRNAs, named as lncLocPred. Figure 1 displays the overall framework of lncLocPred. To begin with, we collected the benchmark dataset from published work and sorted out an independent dataset from RNALocate. Subsequently, we adopted kinds of features to formulate lncRNA sequences, including pentamer, hexamer, octamer, the Series Correlation PseDNC (SC-PseDNC) and triplet. Next, according to the characteristics of different features, we proposed an effective feature selection process by systematically using several feature selection techniques to pick out the optimal feature set from different feature combinations. Finally, based on the logistic regression, we built the model lncLocPred for the subcellular localization prediction of lncRNAs, which outperformed previous methods [32]–[34]. What's more, in order to make the prediction method more practical, we developed a prediction software where people may find it useful in the prediction. Related data files and source code are available at https://github.com/jademyC1221/lncLocPred.

II. MATERIALS AND METHODS

A. DATASET

In order to build a reliable model, we first need to construct a reliable dataset. This paper referred to the dataset from previous studies. Cao *et al.* [32] constructed a dataset containing five subcellular localizations by using computational methods to predict the subcellular localization of lncRNA. Later, based on this, Su *et al.* [33] reorganized to obtain the dataset containing four subcellular locations. We used dataset from [33] as the benchmark dataset. As shown in Table 1, the benchmark dataset covers four subcellular locations of nucleus, cytoplasm, ribosome, and exosome. In addition, to better demonstrate the generalization capability of the model, we downloaded the raw sequence data file of lncRNAs from the RNALocate [23], and then by randomly selecting from the raw sequence data file, we sorted out a new independent dataset that does not appeared in the benchmark dataset. As listed in Table 1, the sample size of each subcellular localization in the independent dataset was allocated according to the original sequence data given by the database.

TABLE 1. The dataset of lncRNA subcellular localization.

Subcellular Locations	Benchmark Dataset	Independent Dataset
Nucleus	156	82
Cytoplasm	426	199
Ribosome	43	99
Exosome	30	16

B. FEATURE REPRESENTATION

1) K-MER NUCLEOTIDE COMPOSITION FEATURES

Recent studies have shown that similar k-mer profiles often appear in lncRNAs with related functions, and the protein-binding and subcellular localization of lncRNAs are correlated with k-mer content [36]. Different values of k formulate sequences into different feature spaces. According to Zhang *et al.*'s article [37], mutational analyses showed that a novel RNA pentamer sequence motif, AGCCC, mediated the nucleus localization of lncRNA BORG with a sequence limitation at positions -8 (T or A) and -3 (G or C) relative to the first nucleotide of the pentamer. On the report of Zuckerman and Ulitsky [38], the preference for C-rich hexamers correlated with nucleus enrichment. Moreover, the biological mechanism of CG-contented octamer trimodal spectrum was unique [39]. Therefore, we calculated the normalized frequency of pentamer occurrence ($k = 5$, 1024 dimensions), hexamer ($k = 6$, 4096 dimensions),

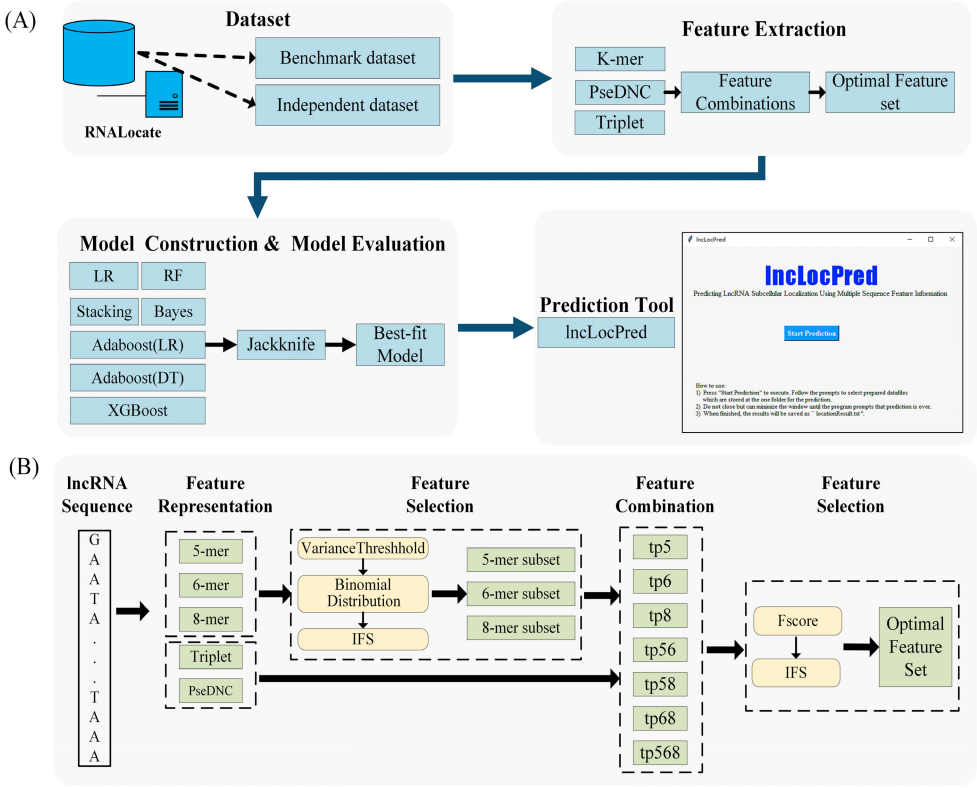

FIGURE 1. The overall framework. A shows the flow chart of our method. First to prepare the datasets and make feature extraction. Subsequent to compare different models based on the optimal feature set to construct the best-fit model and its tool. B presents the process of feature extraction. Step 1, formulate lncRNA sequences to five kinds of features. Step 2, select important features of 5-mer, 6-mer and 8-mer based on Variance Threshold and Binomial distribution. Step 3, combine triplet and PseDNC features with selected K-mer features. Step 4, perform feature selection on different feature combinations with F-score and choose the optimal feature set with the best performance.

and octamer ($k = 8$, 65536 dimensions) to make prediction for their unique properties in lncRNA subcellular localization.

2) PSEUDO-DINUCLEOTIDE COMPOSITION(PseDNC)

PseDNC extracts sequential information and physicochemical properties of the nucleotide sequence by a collection of auto-covariance and cross-covariance transformations [40]–[42]. It generates $16+\lambda\Lambda$ features, the first 16 of which reflecting the local information of dinucleotide and the other $\lambda\Lambda$ reflecting the global information of dinucleotide. λ is the total counted ranks of the correlation along the RNA sequence, and Λ is the number of physicochemical properties. ω is used to balance the local and global sequence effects, ranging from 0 to 1. The detailed definition of PseDNC could be referred to Chen *et al.* [43]. This paper chose the Series Correlation PseDNC (SC-PseDNC) for feature extraction [44], [45], in which ten physicochemical properties were adopted. We used three thermodynamics properties (enthalpy, entropy, and free energy) [46], [47], GC content [48], and six structural properties [43], [49], involving three angular parameters (twist, tilt, roll) and three translational parameters (shift, slide, rise). The Pse-in-One 2.0 web server [45] was used to calculate SC-PseDNC features.

3) LOCAL STRUCTURE-SEQUENCE TRIPLET ELEMENT(TRIPLET)

The function of lncRNAs is closely related to their secondary structure. Triplet [50]–[52] utilizes structural information of RNA sequences. According to the secondary structure [53] of an RNA sequence of length L, we express the sequence as [45]:

$$\mathrm{R} = [\psi_1, \psi_2, \psi_3, \ldots, \psi_L] \tag{1}$$

where ψ_1 represents the structural status of the first nucleotide of the sequence, ψ_2 represents the structural status of the second nucleotide of the sequence, and so forth.

Each nucleotide has two structural statuses. The brackets "(" or ")" and dots "." symbolize the paired and unpaired statuses, respectively. Here we use "(" to unify these two kinds of brackets. Therefore, given a three adjacent nucleotides, it can be expressed as 2^3 possible structural compositions, which are "(((", "((.", "(..", "(.(", ".((", ".(.", "..(", and "...". If we only consider the middle nucleotide in the three adjacent nucleotides, which can be A, C, G, and U, we obtained 32 $\left(4 \times 2^3\right)$ possible structural combinations. Therefore, triplet expresses an RNA sequence as follows.

$$\mathrm{R}=\left[\mathrm{f}_A\left(''(((''\right), \mathrm{f}_A\left(''((\cdot''\right), \ldots, \mathrm{f}_A\left(''\ldots''\right), \ldots, \mathrm{f}_U\left(''\ldots''\right)\right] \tag{2}$$

where f indicates the normalized frequency of the occurrence of the structure-sequence compositions. The Pse-in-One 2.0 web server was used to calculate triplet features.

C. FEATURE SELECTION

As stated by the feature representation, we obtained pentamers, hexamers, octamers, SC-PseDNC, and triplet features with a high dimension. Effective feature selection is required to be carried out to select important features and solve problems of information redundancy, over-fitting, and running time increasing causing by dimension-disaster [54], [55].

1) VARIANCE THRESHHOLD

VarianceThreshhold [56] follows a principle that the distinguishing ability of features with low variance is weak so that these features have little effect in the prediction. VarianceThreshold removes all features whose variances do not meet the threshold. As the value of k increases, the number of k-mers increases as well. Therefore, some k-mers may be not in the sequences of the dataset, or may appear the same number of times in all sequences. Consequently, to save calculation time, we first removed these k-mers with a zero variance for the reason that they have little influence on the prediction of subcellular localization of lncRNAs.

2) BINOMIAL DISTRIBUTION

The binomial distribution has been widely applied in bioinformatics to rank the importance of sequence-based features [33], [57], [58]. We formulate sequence samples in four subcellular localizations to k-mer features. Therefore, from the perspective of statistic, it may be a stochastic event that a certain k-mer occurs in a particular subcellular localization [59]–[61]. We first define the prior probability q_j for each class:

$$q_j = m_j / M \tag{3}$$

where m_j represents the number of a certain k-mer occurring in the jth class ($j = 1, 2, 3, 4$ corresponding to the four subcellular locations of the nucleus, cytoplasm, ribosome, and exosome, respectively), and M is the sum of the occurrence of all k-mers in the four classes.

If the ith k-mer does not occurs in the jth class randomly, the probability that the ith k-mer randomly occurs n_{ij} times and above in the jth class will be very small. Therefore, we can define the confidence level CL_{ij} to describe the propensity of a certain ith k-mer to appear in the jth class and choose the maximum confidence level CL_i for each k-mer as its final confidence level:

$$\begin{cases} CL_{ij} = 1 - \sum_{m=n_{ij}}^{N_i} \dfrac{N_i!}{m!\,(N_i - m)!} q_j^m \left(1 - q_j\right)^{N_i - m} \\ CL_i = \max\left(CL_{i1}, CL_{i2}, CL_{i3}, CL_{i4}\right) \end{cases} \tag{4}$$

where N_i and n_{ij} serves as the number of the ith k-mer occurring in the entire dataset and occurring in the jth class respectively.

3) F-SCORE

F-score is an easy-to-understand feature selection method originally proposed by Chen *et al.* to solve the problem of binary classification [62]–[64]. The larger the F-score, the stronger the distinguishing ability of the feature for the reason that it makes different classes sparse and makes the same class dense. Later, Xie *et al.* [65] improved this feature selection method and expanded it to the multi-classification problem. According to the improvement, the F-score of the features in this paper is defined as:

$$\mathrm{F} = \frac{\sum_{j=1}^{l} (\overline{x_l^{(j)}} - \overline{x_l})^2}{\sum_{j=1}^{l} \frac{1}{n_{j-1}} \sum_{k=1}^{n_j} \left(x_{k,i}^{(j)} - \overline{x_l^{(j)}}\right)^2} \tag{5}$$

where $\overline{x_l}$ and $\overline{x_l^{(j)}}$ denote the average value of the ith feature in the entire dataset and in the jth class respectively, $x_{k,i}^{(j)}$ represents the value of the ith feature of the kth sequence in the jth class, and n_j is the number of sequences in the jth class, $l = 4$.

4) FEATURE SELECTION PROCESS

Figure 1.B presents the process of feature selection. Firstly, we performed feature selection on k-mers for the reason that with the increase of k, the number of k-mers would be more than that of triplet and PseDNC. We used VarianceThreshhold to remove those k-mers ($k = 5$, 6, and 8) with a variance of zero. Only to find that none of pentamers and hexamers was removed and 681 octamers with a variance of zero were removed. Then, we reordered these three kinds of k-mers according to their confidence levels from the binomial distribution. Next, we performed IFS strategy on them based on logistic regression to determine the best dimensions of 5-mer, 6-mer and 8-mer with highest accuracies respectively. To improve efficiency and avoid over-fitting, other than adding features one by one, our IFS strategy was performed in the way of adding features by percentage. According to the different dimensions of k-mer, we gradually added the pentamer and hexamer by one percent each time, and gradually added the octamer by two percent each time. Figure 2 displays the IFS process of pentamer, hexamer and octamer. As a result, the top 205 pentamers with CL of 100% and accuracy of 67.51% was chosen, and the top 2008 hexamers with CL of higher than 96.06% and accuracy of 71.27% as well. Similarly, we obtained the top 9080 octamers with CL of higher than 96.91% and accuracy of 86.38%.

Subsequently, we combined triplet and SC-PseDNC features with the 205 pentamers, the 2008 hexamers, and the 9080 octamers in different ways to find the best combination. For each combination, we utilized the F-score method to reorder its features and the IFS strategy was utilized again. Consequently, we obtained the best dimensions of each combination. For the convenience of description, we used "tpk" to represent the way to combine, whose "tp" represents

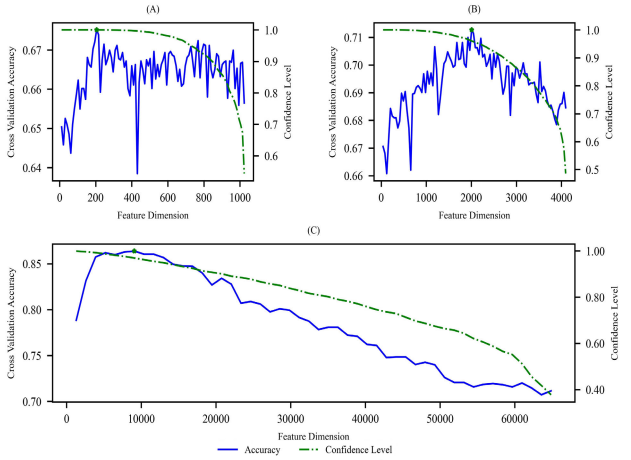

FIGURE 2. (A) Pentamer IFS selecting process. (B) Hexamer IFS selecting process. (C) Octamer IFS selecting process.

triplet and PseDNC, and "k"can be 5, 6, and 8 indicating pentamer, hexamer, and octamer respectively.

D. LOGISTIC REGRESSION

Logistic regression [66] is a nonlinear classification model in machine learning. It has been applied to solve many classification problems in bioinformatics [67]–[70]. Logistic regression converts the results of linear regression into a probability between 0 and 1 by a sigmoid function. In this paper, we used the logistic regression package in sklearn [56]. Adjustable parameters of the model include $\text{multi}_{\text{class}}$ for the multi-classification strategies, solver for the optimization algorithms of the loss function, and the regularization parameter C, which reduce overfitting of the model. The search ranges of the three parameters by the grid search were as follows:

$$\begin{cases} \text{multi}_{\text{class}} = \text{ovr, multinomial} \\ \text{solver} = \text{liblinear, lbfgs, newton} - \text{cg} \\ 10^{-3} \le C \le 10^{3} \end{cases} \tag{6}$$

E. PERFORMANCE EVALUATION

To provide a set of more intuitive and objective scales to evaluate the performance of the model, the criterion used in other methods about predicting the subcellular localization of lncRNAs was adopted in this study, including sensitivity (Sn), specificity (Sp), Matthews's correlation coefficient (MCC), and accuracy (Acc) [32]–[34]. Sensitivity measures the ability to correctly predict one sample belongs to a certain class while specificity measures the ability to correctly distinguish one sample does not belong to a certain class. MCC is a correlation coefficient describing the relationship between true classes and predicted classes, whose value of 1 represents a perfect prediction of the class while -1 represents a completely wrong prediction of the class. The following are

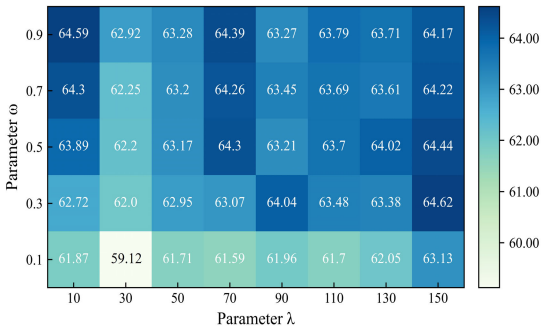

FIGURE 3. Influence of λ and ω in PseDNC for prediction of the subcellular localization of lncRNAs.

four metrics:

$$\begin{cases} \text{Sn}(i) = 1 - \dfrac{N_{-}^{+}(i)}{N^{+}(i)} & 0 \le \text{Sn}(i) \le 1 \\ \text{Sp}(i) = 1 - \dfrac{N_{+}^{-}(i)}{N^{-}(i)} & 0 \le \text{Sp}(i) \le 1 \\ \text{MCC}(i) = \dfrac{1 - \left(\dfrac{N_{-}^{+}(i)}{N^{+}(i)} + \dfrac{N_{+}^{-}(i)}{N^{-}(i)}\right)}{\sqrt{\left(1 + \dfrac{N_{+}^{-}(i) - N_{-}^{+}(i)}{N^{+}(i)}\right)\left(1 + \dfrac{N_{-}^{+}(i) - N_{+}^{-}(i)}{N^{-}(i)}\right)}} & -1 \le \text{MCC}(i) \le 1 \\ \text{Acc} = \dfrac{1}{\delta}\sum_{i=1}^{l}\left[N^{+}(i) - N_{-}^{+}(i)\right] & 0 \le \text{Acc} \le 1 \end{cases} \tag{7}$$

where $N^{+}(i)$ is the total number of class i, $N^{-}(i)$ is the number other than the class i. $N_{-}^{+}(i)$ is the number of samples in class i that are incorrectly predicted to other classes; $N_{+}^{-}(i)$ is the number of samples in other classes except i that are incorrectly predicted to class i. l is the number of classes and δ is the number of the total samples in the benchmark dataset.

III. RESULTS AND DISCUSSION

A. PARAMETERS OF PseDNC

In respect of PseDNC [43], we used the grid search to optimize two parameters, λ and ω. According to the definition of PseDNC, the value of λ should not exceed the difference between the sequence length and 2 (2 means dinucleotide). Besides, the larger λ is, the lower the cluster-tolerant capacity is [71]. Hence, we set the search range as formula (10). As shown in Figure 3, after carrying out 100 times of the five-fold CV on the benchmark dataset, we found that the best accuracy was obtained when $\lambda = 150$ and $\omega = 0.3$.

$$\begin{cases} 10 \le \lambda \le 150, & \Delta = 20 \\ 0.1 \le \omega \le 0.9, & \Delta = 0.2 \end{cases} \tag{8}$$

B. DIFFERENT FEATURE COMBINATIONS AND FEATURE ANALYSIS

Table 2 manifests details of the feature selection results of seven different feature combinations. We present the models trained by different feature combinations in the form

TABLE 2. Performance of different combinations of feature.

Model	Optimal dimension	Number in optimal dimension					Jackknife accuracy
		Triplet	PseDNC	Pentamer	Hexamer	Octamer	
lncLocPredtp5	526	25	316	181	-	-	69.16%
lncLocPredtp6	1103	23	160	-	920	-	76.64%
lncLocPredtp8	7228	26	468	-	-	6734	**92.37**%
lncLocPredtp56	1392	23	199	169	1001	-	74.66%
lncLocPredtp58	8017	27	562	193	-	7235	91.91%
lncLocPredtp68	8340	26	422	-	1425	6467	90.23%
lncLocPredtp568	8090	25	367	183	1365	6150	90.23%

of "lncLocPred+tpk", whose suffix "tpk" represents the combined features. We tested each model by jackknife cross-validation. By analyzing the optimal feature dimensions of those models whose accuracies were over 90%, we could know more details about the selected features. Regarding triplet feature, its number selected was relatively small but its utilization was over 78%(25/32). Regarding PseDNC, the number selected was in the interval of 367 to 562, with about 24% to 37% utilization in the total number of 1516. Regarding k-mer, not only the number of features selected was much more than triplet and PseDNC, but they also ranked higher in the F-score feature ranking. It is particularly worth noting that when the octamer was added into the feature set, the accuracy could be increased to over 90%, which suggested that octamer had better discriminating ability for subcellular localization of lncRNAs than other features. Figure 4 shows the frequency of the top 20 pentamers in the F-score ranking of lncLocPredtp58 occurring in four subcellular localizations, as well as the case of hexamers of lncLocPredtp68 and octamers of lncLocPredtp8. The frequency is the ratio of the times of the certain k-mer appeared in one subcellular localization and the times of all k-mers appeared in that subcellular localization. We observed that the top-ranked k-mers had a higher base content of G and C and often contained short repeats such as GG, CC, GGG, CCC, GGGG, and CCCC. This observation is consistent with the conclusion drawn in the previous study [37]–[39]. The supplementary material Table S1-S4 is available at our github address, which listed the details of the top 1000 features in the optimal feature sets of lncLocPredtp8, lncLocPredtp58, lncLocPredtp68, and lncLocPredtp568, demonstrating feature names and their F-scores.

From Table 2, we know that model lncLocPredtp8, lncLocPredtp58, lncLocPredtp68, and lncLocPredtp568 achieved accuracy of 92.37%, 91.91%, 90.23%, and 90.23% respectively, among which lncLocPredtp8 was the highest. Therefore, we chose the selected feature set of the combination of octamer, triplet and PseDNC as the optimal feature set and lncLocPredtp8 model as the best model named lncLocPred with the highest accuracy on benchmark dataset in our study.

FIGURE 4. (A)The frequency of the top 20 pentamers of lncLocPredtp58 in the F-score ranking on four subcellular locations. (B)The frequency of the top 20 hexamers of lncLocPredtp68 in the F-score ranking on four subcellular locations. (C)The frequency of the top 20 octamers of lncLocPredtp8 in the F-score ranking on four subcellular locations.

C. PERFORMANCE COMPARISON WITH DIFFERENT MACHINE LEARNING ALGORITHMS

To compare logistic regression with other machine learning algorithms, we repeated our experimental process, especially the feature selection process, based on other machine learning algorithms. As shown in Table 3, compared with other classifiers, lncLocPred performed most prominently on most indicators by jackknife. We considered that as long as the machine learning method was matched with a more appropriate feature selection process, it may obtain better prediction result like logistic regression done in this work.

D. COMPARISON WITH EXISTING STATE-OF-THE-ART METHODS

We compared lncLocPred with three published predictors, lncLocator, iLoc-lncRNA and Locate-R, on the benchmark

TABLE 3. Compared with different machine learning methods based on the jackknife.

	Location	lncLocPred	Adaboost(LR)	Adaboost(DT)	Stacking	XGBoost	Bayes	RF
Sn	Nucleus	0.968	0.904	0.308	0.909	0.314	0.942	0.160
	Cytoplasm	0.991	0.993	0.944	0.990	0.981	0.904	0.998
	Ribosome	0.605	0.465	0.116	0.233	0.163	0.093	0.140
	Exosome	0.2	0.2	0.067	0.433	0.067	0.067	0.1
Sp	Nucleus	0.968	0.974	0.942	0.967	0.982	0.816	0.994
	Cytoplasm	0.856	0.782	0.262	0.775	0.275	0.891	0.157
	Ribosome	0.998	0.997	1.0	0.998	0.995	1.0	1.0
	Exosome	1.0	1.0	1.0	1.0	0.998	1.0	1.0
MCC	Nucleus	0.915	0.882	0.330	0.873	0.444	0.671	0.325
	Cytoplasm	0.876	0.825	0.293	0.815	0.393	0.783	0.320
	Ribosome	0.751	0.635	0.331	0.445	0.319	0.296	0.363
	Exosome	0.439	0.439	0.253	0.649	0.201	0.253	0.310
Acc		92.37%	90.08%	68.86%	88.55%	72.67%	82.14%	70.23%

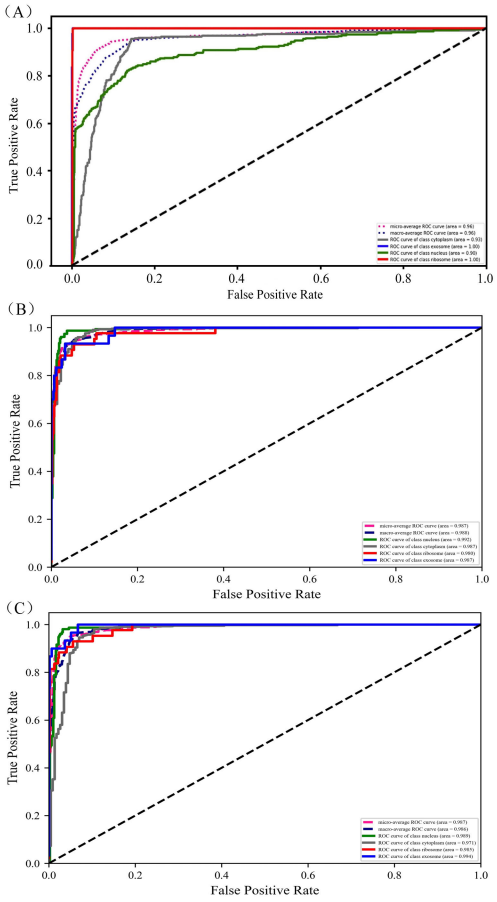

FIGURE 5. (A) ROC curve of Locate-R under 10-fold CV [34]. (B) ROC curve of lncLocPred under 10-fold CV. (C) ROC curve of lncLocPred under jackknife.

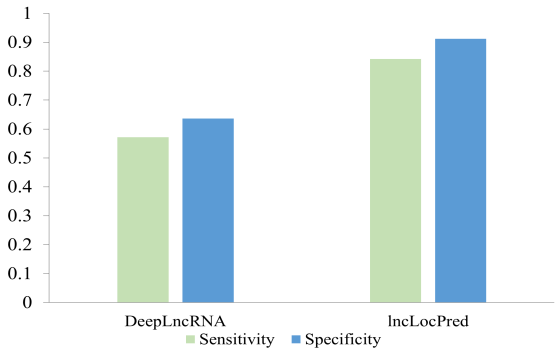

FIGURE 6. The performance comparison between lncLocPred and DeepLncRNA.

dataset by jackknife. Performance comparison results were presented in Table 4. In general, we have improved the overall accuracy by 2%. Regarding the subcellular location of the nucleus and cytoplasm, our method performed better in Sn and MCC. Regarding the subcellular location of ribosome and exosome, our method performed better than iLoc-lncRNA and lncLocator but worse than Locate-R in Sn and MCC. In order to better evaluate the performance of lncLocPred, we plotted the ROC curve under the jackknife cross-validation and 10-fold cross-validation methods, and calculated the AUC value. At the same time, we compared with the ROC curve of Locate-R under 10-fold cross-validation, which can be seen from Figure 5 that our method achieved better results in the overall AUC value, the AUC value on the subcellular localization of nucleus and the AUC value the subcellular localization of cytoplasm. In addition, we compared our method with DeepLncRNA [35]. The comparative data set came from 152 nuclear lncRNAs and 91 cytoplasmic lncRNAs in the lncLocator test set. In particular, lncLocPred is a four-category prediction method while DeepLncRNA is a binary one. Figure 6 shows that our methods have achieved satisfactory performance in terms of sensitivity and specificity.

E. COMPARISON IN THE INDEPENDENT DATASET

To compare the performance of the model more objectively, we input independent testset to the web servers provided by iLoc-lncRNA and Locate-R to obtain prediction results. Also, we input into our lncLocPred as well. What can be seen from Table 5 is that whether our models, iLoc-lncRNA and Locate-R, their performance were significantly different from the benchmark dataset. The lncLocPred, Locate-R and

TABLE 4. Comparison with existing state-of-the-art methods.

Metric	Location	lncLocPred	Locate-R	iLoc-lncRNA	lncLocator
Sn(%)	Nucleus	96.79	65.92	77.56	38.15
	Cytoplasm	99.06	84.74	99.06	88.01
	Ribosome	60.47	100	46.51	7.00
	Exosome	20.00	100	16.67	4.00
Sp(%)	Nucleus	96.79	95.15	97.59	92.17
	Cytoplasm	85.59	89.10	67.68	36.36
	Ribosome	99.84	98.37	99.83	97.53
	Exosome	100	99.17	100	97.27
MCC	Nucleus	0.915	0.658	0.796	0.357
	Cytoplasm	0.876	0.725	0.742	0.288
	Ribosome	0.751	0.970	0.652	0.070
	Exosome	0.439	0.978	0.400	0.015
Acc		92.37%	90.69%	86.72%	66.5%

TABLE 5. Comparing different methods on the independent dataset.

Method	Acc	Nucleus	Cytoplasm	Ribosome	Exosome
lncLocPred	44.44%	14	155	4	3
Locate-R	38.64%	9	121	20	3
iLoc-lncRNA	35.86%	14	115	9	4

iLoc-lncRNA differs by 0.4793 (0.9237 to 0.4444), 0.5205 (0.9069 to 0.3864) and 0.5086 (0.8672 to 0.3586), where the difference of our method is the smallest. It can be seen that our method has more advantages in the prediction of nucleus and cytoplasm, and the prediction effect of Locate-R on the ribosome is better than ours. However, the results still illustrate the advantages of our model, whose overall accuracy is higher and the difference between the benchmark dataset and independent dataset is smaller.

IV. CONCLUSION

In conclusion, we proposed an effective method, called lncLocPred, to predict the subcellular localization of lncRNAs based on logistic regression. In terms of feature selection, we put forward a process that combines multiple feature selection techniques to select different types of features. Feature analysis shows that the top k-mer features prefer repeated C bases or G bases, indicating that rich C-nucleotide and G-nucleotide impacted subcellular localization of lncRNAs. We admit that there are some limitations in lncLocPred. Due to the small number on the subcellular locations of ribosome and exosome on benchmark dataset, the model is more biased towards the subcellular locations of nucleus and cytoplasm. In the future, with the continuous increase of experimental data on these two locations, lncLocPred could be trained on a more balanced data set to improve our tools. In addition, we expect to collect a bigger dataset not only from RNALocate but also from other databases to develop a more effective bioinformatics tool.

REFERENCES

[1] J. T. Y. Kung, D. Colognori, and J. T. Lee, "Long noncoding RNAs: Past, present, and future," *Genetics*, vol. 193, no. 3, pp. 651–669, Mar. 2013. [Online]. Available: https://www.genetics.org/content/193/3/651

[2] I. Dunham, A. Kundaje, S. Aldred, P. Collins, C. Davis, F. Doyle, C. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. Lajoie, S. Landt, B.-K. Lee, F. Pauli Behn, K. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, and E. Birney, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[3] S. Djebali, C. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. Marinov, J. Khatun, B. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. Abdelhamid, and T. Gingeras, "Landscape of transcription in human cells," *Nature*, vol. 489, pp. 101–108, Sep. 2012.

[4] Y. Kondo, K. Shinjo, and K. Katsushima, "Long non-coding RNAs as an epigenetic regulator in human cancers," *Cancer Sci.*, vol. 108, no. 10, pp. 1927–1933, Oct. 2017.

[5] J. J. Quinn and H. Y. Chang, "Unique features of long non-coding RNA biogenesis and function," *Nature Rev. Genet.*, vol. 17, no. 1, pp. 47–62, Jan. 2016.

[6] J. M. Engreitz, N. Ollikainen, and M. Guttman, "Long non-coding RNAs: Spatial amplifiers that control nuclear structure and gene expression," *Nature Rev. Mol. Cell Biol.*, vol. 17, no. 12, pp. 756–770, Oct. 2016.

[7] X. Pan, L. J. Jensen, and J. Gorodkin, "Inferring disease-associated long non-coding RNAs using genome-wide tissue expression profiles," *Bioinformatics*, vol. 35, no. 9, pp. 1494–1502, Oct. 2018, doi: 10.1093/bioinformatics/bty859.

[8] Y. Meng, Y.-L. Liu, K. Li, and T. Fu, "Prognostic value of long non-coding RNA breast cancer anti-estrogen resistance 4 in human cancers: A meta-analysis," *Medicine*, vol. 98, no. 21, May 2019, Art. no. e15793.

[9] M. Xue, D. Shi, G. Xu, and W. Wang, "The long noncoding RNA linc00858 promotes progress of lung cancer through miR-3182/MMP2 axis," *Artif. Cells, Nanomedicine, Biotechnol.*, vol. 47, no. 1, pp. 2091–2097, Dec. 2019.

[10] D. Thiel, N. D. Conrad, E. Ntini, R. X. Peschutter, H. Siebert, and A. Marsico, "Identifying lncRNA-mediated regulatory modules via ChIA-PET network analysis," *BMC Bioinf.*, vol. 20, no. 1, p. 292, May 2019.

[11] L.-L. Chen, "Linking long noncoding RNA localization and function," *Trends Biochem. Sci.*, vol. 41, no. 9, pp. 761–772, Sep. 2016.

[12] D. Mas-Ponte, J. Carlevaro-Fita, E. Palumbo, T. H. Pulido, R. Guigo, and R. Johnson, "LncATLAS database for subcellular localization of long noncoding RNAs," *RNA*, vol. 23, no. 7, pp. 1080–1087, Jul. 2017.

[13] E. K. Robinson, S. Covarrubias, and S. Carpenter, "The how and why of lncRNA function: An innate immune perspective," *Biochimica Biophys. Acta (BBA)-Gene Regulatory Mech.*, vol. 1863, no. 4, Apr. 2020, Art. no. 194419. [Online]. Available: https://europepmc.org/articles/PMC7185634

[14] B. Yu and G. Shan, "Functions of long noncoding RNAs in the nucleus," *Nucleus*, vol. 7, no. 2, pp. 155–166, Apr. 2016.

[15] Q. Sun, Q. Hao, and K. V. Prasanth, "Nuclear long noncoding RNAs: Key regulators of gene expression," *Trends Genet.*, vol. 34, no. 2, pp. 142–157, Feb. 2018.

[16] I. Ahmad, A. Valverde, F. Ahmad, and A. R. Naqvi, "Long noncoding RNA in myeloid and lymphoid cell differentiation, polarization and function," *Cells*, vol. 9, no. 2, p. 269, Jan. 2020.

[17] H. Chen, S. Yang, and R. Shao, "Long non-coding XIST raises methylation of TIMP-3 promoter to regulate collagen degradation in osteoarthritic chondrocytes after tibial plateau fracture," *Arthritis Res. Therapy*, vol. 21, no. 1, p. 271, Dec. 2019.

[18] Y.-Y. Tseng, B. S. Moriarity, W. Gong, R. Akiyama, A. Tiwari, H. Kawakami, P. Ronning, B. Reuland, K. Guenther, T. C. Beadnell, J. Essig, G. M. Otto, M. G. O'Sullivan, D. A. Largaespada, K. L. Schwertfeger, Y. Marahrens, Y. Kawakami, and A. Bagchi, "PVT1 dependence in cancer with MYC copy-number increase," *Nature*, vol. 512, no. 7512, pp. 82–86, Aug. 2014.

[19] B. Liu, L. Sun, Q. Liu, C. Gong, Y. Yao, X. Lv, L. Lin, H. Yao, F. Su, D. Li, M. Zeng, and E. Song, "A cytoplasmic NF-κB interacting long noncoding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis," *Cancer Cell*, vol. 27, no. 3, pp. 370–381, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1535610815000549

[20] Y. Wang, K. Wang, L. Zhang, Y. Tan, Z. Hu, L. Dang, H. Zhou, G. Li, H. Wang, S. Zhang, F. Shi, X. Cao, and G. Zhang, "Targeted overexpression of the long noncoding RNA ODSM can regulate osteoblast function *in vitro* and *in vivo*," *Cell Death Disease*, vol. 11, no. 2, p. 133, Feb. 2020.

[21] Y.-P. Hu, Y.-P. Jin, X.-S. Wu, Y. Yang, Y.-S. Li, H.-F. Li, S.-S. Xiang, X.-L. Song, L. Jiang, Y.-J. Zhang, W. Huang, S.-L. Chen, F.-T. Liu, C. Chen, Q. Zhu, H.-Z. Chen, R. Shao, and Y.-B. Liu, "LncRNA-HGBC stabilized by HuR promotes gallbladder cancer progression by regulating miR-502-3p/SET/AKT axis," *Mol. Cancer*, vol. 18, no. 1, p. 167, Nov. 2019.

[22] C.-M. Kang, H.-L. Bai, X.-H. Li, R.-Y. Huang, J.-J. Zhao, X.-Y. Dai, L. Zheng, Y.-R. Qiu, Y.-W. Hu, and Q. Wang, "The binding of lncRNA RP11-732M18.3 with 14-3-3 β/α accelerates p21 degradation and promotes glioma growth," *EBioMedicine*, vol. 45, pp. 58–69, Jul. 2019.

[23] T. Zhang, P. Tan, L. Wang, N. Jin, Y. Li, L. Zhang, H. Yang, Z. Hu, L. Zhang, C. Hu, C. Li, K. Qian, C. Zhang, Y. Huang, K. Li, H. Lin, and D. Wang, "Rnalocate: A resource for rna subcellular localizations," *Nucleic Acids Res.*, vol. 45, pp. D135–D138, Aug. 2016.

[24] X. Wen, L. Gao, X. Guo, X. Li, X. Huang, Y. Wang, H. Xu, R. He, C. Jia, and F. Liang, "LncSLdb: A resource for long non-coding RNA subcellular localization," *Database*, vol. 2018, pp. 1–6, Jan. 2018, doi: 10.1093/database/bay085.

[25] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. BioSyst.*, vol. 8, no. 2, pp. 629–641, Dec. 2011.

[26] P.-F. Du, "Predicting protein submitochondrial locations: The 10th anniversary," *Current Genomics*, vol. 18, no. 4, pp. 316–321, Jul. 2017.

[27] F. Yang, Y. Liu, Y. Wang, Z. Yin, and Z. Yang, "MIC_Locator: A novel image-based protein subcellular location multi-label prediction model based on multi-scale monogenic signal representation and intensity encoding strategy," *BMC Bioinf.*, vol. 20, no. 1, p. 522, Oct. 2019.

[28] W. Long, Y. Yang, and H.-B. Shen, "ImPLoc: A multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images," *Bioinformatics*, vol. 36, no. 7, pp. 2244–2250, Apr. 2020, doi: 10.1093/bioinformatics/btz909.

[29] X. Pan, L. Chen, M. Liu, T. Huang, and Y. D. Cai, "Predicting protein subcellular location using learned distributed representations from a protein-protein network," *BioRxiv*, Sep. 2019, Art. no. 768739. [Online]. Available: https://www.biorxiv.org/content/10.1101/768739v1, doi: 10.1101/768739.

[30] E. Hacisuleyman, L. A. Goff, C. Trapnell, A. Williams, J. Henao-Mejia, L. Sun, P. McClanahan, D. G. Hendrickson, M. Sauvageau, D. R. Kelley, M. Morse, J. Engreitz, E. S. Lander, M. Guttman, H. F. Lodish, R. Flavell, A. Raj, and J. L. Rinn, "Topological organization of multichromosomal regions by the long intergenic noncoding RNA firre," *Nature Struct. Mol. Biol.*, vol. 21, no. 2, pp. 198–206, Feb. 2014.

[31] M. Woźniak, D. Połap, L. Kośmider, and T. Clapa, "Automated fluorescence microscopy image analysis of pseudomonas aeruginosa bacteria in alive and dead stadium," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 100–110, Jan. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197617302075

[32] Z. Cao, X. Pan, Y. Yang, Y. Huang, and H.-B. Shen, "The lncLocator: A subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier," *Bioinformatics*, vol. 34, no. 13, pp. 2185–2194, Feb. 2018, doi: 10.1093/bioinformatics/bty085.

[33] Z.-D. Su, Y. Huang, Z.-Y. Zhang, Y.-W. Zhao, D. Wang, W. Chen, K.-C. Chou, and H. Lin, "iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC," *Bioinformatics*, vol. 34, no. 24, pp. 4196–4204, Jun. 2018, doi: 10.1093/bioinformatics/bty508.

[34] A. Ahmad, H. Lin, and S. Shatabda, "Locate-R: Subcellular localization of long non-coding RNAs using nucleotide compositions," *Genomics*, vol. 112, no. 3, pp. 2583–2589, May 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888754319304100

[35] B. L. Gudenas and L. Wang, "Prediction of LncRNA subcellular localization with deep learning from sequence features," *Sci. Rep.*, vol. 8, no. 1, p. 16385, Nov. 2018, doi: 10.1038/s41598-018-34708-w.

[36] J. M. Kirk, S. O. Kim, K. Inoue, M. J. Smola, D. M. Lee, M. D. Schertzer, J. S. Wooten, A. R. Baker, D. Sprague, D. W. Collins, C. R. Horning, S. Wang, Q. Chen, K. M. Weeks, P. J. Mucha, and J. M. Calabrese, "Functional classification of long non-coding RNAs by *k*-mer content," *Nature Genet.*, vol. 50, no. 10, pp. 1474–1482, Oct. 2018.

[37] B. Zhang, L. Gunawardane, F. Niazi, F. Jahanbani, X. Chen, and S. Valadkhan, "A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA," *Mol. Cellular Biol.*, vol. 34, no. 12, pp. 2318–2329, Jun. 2014.

[38] B. Zuckerman and I. Ulitsky, "Predictive models of subcellular localization of long RNAs," *RNA*, vol. 25, no. 5, pp. 557–572, May 2019.

[39] Y. Jia, H. Li, J. Wang, H. Meng, and Z. Yang, "Spectrum structures and biological functions of 8-mers in the human genome," *Genomics*, vol. 111, no. 3, pp. 483–491, May 2019.

[40] K. K. Tan, Le, Yeh, and Chua, "Ensemble of deep recurrent neural networks for identifying enhancers via dinucleotide physicochemical properties," *Cells*, vol. 8, no. 7, p. 767, Jul. 2019.

[41] T. Fang, Z. Zhang, R. Sun, L. Zhu, J. He, B. Huang, Y. Xiong, and X. Zhu, "RNAm5CPred: Prediction of RNA 5-Methylcytosine sites based on three different kinds of nucleotide composition," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 739–747, Dec. 2019.

[42] S. Zhang, M. Chang, Z. Zhou, X. Dai, and Z. Xu, "PDHS-ELM: Computational predictor for plant DNase I hypersensitive sites based on extreme learning machines," *Mol. Genet. Genomics*, vol. 293, no. 4, pp. 1035–1049, Aug. 2018.

[43] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "IRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res.*, vol. 41, no. 6, p. e68, Jan. 2013.

[44] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: A Web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, May 2015.

[45] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-One 2.0: An improved package of Web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Sci.*, vol. 9, no. 4, pp. 67–91, 2017.

[46] W. Chen, H. Lv, F. Nie, and H. Lin, "I6 mA-pred: Identifying DNA N6-methyladenine sites in the rice genome," *Bioinformatics*, vol. 35, no. 16, pp. 2796–2800, Aug. 2019.

[47] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, and H. Lin, "Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique," *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, Jun. 2019.

[48] X. Zhang, J. Wang, J. Li, W. Chen, and C. Liu, "CRlncRC: A machine learning-based method for cancer-related long noncoding RNA identification using integrated features," *BMC Med. Genomics*, vol. 11, no. S6, p. 120, Dec. 2018.

[49] A. Pérez, A. Noy, F. Lankas, F. J. Luque, and M. Orozco, "The relative flexibility of B-DNA and A-RNA duplexes: Database analysis," *Nucleic Acids Res.*, vol. 32, no. 20, pp. 6144–6151, Jan. 2004, doi: 10.1093/nar/gkh954.

[50] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinf.*, vol. 6, p. 310, Dec. 2005.

[51] J. Wang and M. Gribskov, "IRESpy: An XGBoost model for prediction of internal ribosome entry sites," *BMC Bioinf.*, vol. 20, no. 1, p. 409, Jul. 2019.

[52] X. Ru, P. Cao, L. Li, and Q. Zou, "Selecting essential MicroRNAs using a novel voting method," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 16–23, Dec. 2019.

[53] B. Liu, L. Fang, F. Liu, X. Wang, and K.-C. Chou, "iMiRNA-PseDPC: MicroRNA precursor identification with a pseudo distance-pair composition approach," *J. Biomol. Struct. Dyn.*, vol. 34, no. 1, pp. 1–28, 2015.

[54] C. Ding, L.-F. Yuan, S.-H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *J. Proteomics*, vol. 77, pp. 321–328, Dec. 2012.

[55] M. Woźniak and D. Połap, "Soft trees with neural components as image-processing technique for archeological excavations," *Pers. Ubiquitous Comput.*, vol. 24, no. 3, pp. 363–375, Jan. 2020.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[57] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, and H. Lin, "ITerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators," *Bioinformatics*, vol. 35, no. 9, pp. 1469–1477, May 2019.
[58] H. Yang, W.-R. Qiu, G. Liu, F.-B. Guo, W. Chen, K.-C. Chou, and H. Lin, "IRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 883–891, May 2018.
[59] J. X. Tan, S. H. Li, Z. M. Zhang, C. X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosci. Eng.*, vol. 16, pp. 2466–2480, Mar. 2019.
[60] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowl.-Based Syst.*, vol. 163, pp. 787–793, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118304933
[61] P.-P. Zhu, W.-C. Li, Z.-J. Zhong, E.-Z. Deng, H. Ding, W. Chen, and H. Lin, "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Mol. BioSyst.*, vol. 11, no. 2, pp. 558–563, 2015, doi: 10.1039/C4MB00645C.
[62] Y. W. Chen and C. J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction* (Studies in Fuzziness and Soft Computing), vol. 207. Berlin, Germany: Springer, 2006. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-35488-8_13#citeas, doi: 10.1007/978-3-540-35488-8_13.
[63] W. He, Y. Ju, X. Zeng, X. Liu, and Q. Zou, "Sc-ncDNAPred: A sequence-based predictor for identifying non-coding DNA in saccharomyces cerevisiae," *Frontiers Microbiol.*, vol. 9, p. 2174, Sep. 2018.
[64] A. Porollo and J. Meller, "Prediction-based fingerprints of protein-protein interactions," *Proteins, Struct., Function, Bioinf.*, vol. 66, no. 3, pp. 630–645, Dec. 2006.
[65] J.-Y. Xie, C.-X. Wang, S. Jiang, and Y. Zhang, "Feature selection method combing improved F-score and support vector machine," *J. Comput. Appl.*, vol. 30, no. 4, pp. 993–996, Apr. 2010.
[66] C. Robert, "Machine learning, a probabilistic perspective," *CHANCE*, vol. 27, no. 2, pp. 62–63, Apr. 2014.
[67] C. Zeng and M. Hamada, "Identifying sequence features that drive ribosomal association for lncRNA," *BMC Genomics*, vol. 19, no. S10, p. 906, Dec. 2018.
[68] E. L. Xu, X. Qian, Q. Yu, H. Zhang, and S. Cui, "Feature selection with interactions in logistic regression models using multivariate synergies for a GWAS application," *BMC Genomics*, vol. 19, no. S4, p. 170, Mar. 2018.
[69] D. Ichikawa, T. Saito, W. Ujita, and H. Oyama, "How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach," *J. Biomed. Inform.*, vol. 64, pp. 20–24, Dec. 2016.
[70] J. Song, F. Li, A. Leier, T. T. Marquez-Lago, T. Akutsu, G. Haffari, K.-C. Chou, G. I. Webb, and R. N. Pike, "PROSPERous: High-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy," *Bioinformatics*, vol. 34, no. 4, pp. 684–687, Oct. 2017, doi: 10.1093/bioinformatics/btx670.
[71] K.-C. Chou, "A key driving force in determination of protein structural classes," *Biochem. Biophys. Res. Commun.*, vol. 264, no. 1, pp. 216–224, Oct. 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006291X99913256

YONGXIAN FAN received the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University. He is currently a Professor with the School of Computer Science and Information Security, Guilin University of Electronic Technology, China. His research interests include artificial intelligence, data analysis, machine learning, pattern recognition, and bioinformatics.

MEIJUN CHEN received the B.E. degree in information management and information system from the Guangdong University of Finance and Economics, Guangdong, China, in 2018. She is currently pursuing the M.E. degree in computer science with the Guilin University of Electronic Technology, Guilin, China. Her research interests include machine learning and bioinformatics.

QINGQI ZHU received the bachelor's degree in software engineering from Chaohu University, Hefei, China, in 2017. He is currently pursuing the master's degree with the Guilin University of Electronic Technology. His research interests include deep learning, machine learning, and bioinformatics.