Mathematical Biosciences and Engineering

MBE, 20(12): 20648–20667.
DOI: 10.3934/mbe.2023913
Received: 08 September 2023
Revised: 24 October 2023
Accepted: 26 October 2023
Published: 15 November 2023

http://www.aimspress.com/journal/MBE

Research article

DlncRNALoc: A discrete wavelet transform-based model for predicting lncRNA subcellular localization

Xiangzheng Fu[1,2,4], Yifan Chen[2,4,*] and Sha Tian[3,*]

[1] Neher's Biophysics Laboratory for Innovative Drug Discovery, State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macao, China
[2] College of Information Science and Engineering, Hunan University, Changsha, Hunan, China
[3] Department of Internal Medicine, College of Integrated Chinese and Western Medicine, Hunan University of Chinese Medicine, Changsha, Hunan, China
[4] Department of Basic Biology, Changsha Medical College, Changsha, Hunan, China

* **Correspondence:** Email: cyf176@hnu.edu.cn, 003942@hnucm.edu.cn.

Abstract: The prediction of long non-coding RNA (lncRNA) subcellular localization is essential to the understanding of its function and involvement in cellular regulation. Traditional biological experimental methods are costly and time-consuming, making computational methods the preferred approach for predicting lncRNA subcellular localization (LSL). However, existing computational methods have limitations due to the structural characteristics of lncRNAs and the uneven distribution of data across subcellular compartments. We propose a discrete wavelet transform (DWT)-based model for predicting LSL, called DlncRNALoc. We construct a physicochemical property matrix of a 2-tuple bases based on lncRNA sequences, and we introduce a DWT lncRNA feature extraction method. We use the Synthetic Minority Over-sampling Technique (SMOTE) for oversampling and the local fisher discriminant analysis (LFDA) algorithm to optimize feature information. The optimized feature vectors are fed into support vector machine (SVM) to construct a predictive model. DlncRNALoc has been applied for a five-fold cross-validation on the three sets of benchmark datasets. Extensive experiments have demonstrated the superiority and effectiveness of the DlncRNALoc model in predicting LSL.

Keywords: lncRNA subcellular localization; discrete wavelet transform; physicochemical property matrix; synthetic minority over-sampling; local fisher discriminant analysis

1. Introduction

The sudden outbreak of coronavirus disease 2019 (COVID-19) has had a major impact on the economy, society and daily life [1–5]. Long non-coding RNAs (lncRNAs) have significant research implications for the prevention and treatment of viral epidemics. For instance, lncRNAs can serve as emerging regulators of COVID-19 [6].

LncRNA was once thought to be a transcriptional "noise" of genes with no biological function. Recently many important functions of lncRNAs, including the regulation of gene expression and the modulation of protein activity have been discovered. These functions are closely related to life activities such as cell structural integrity, cell differentiation, cell cycle, and immune response [7], which has caused lncRNAs to receive an increasing amount of attention in the field of life sciences. However, due to the complexity of the molecular mechanisms and functions themselves, lncRNA-related research has largely lagged behind the research on other types of non-coding RNAs (ncRNAs) and proteins.

The function of lncRNAs also depends on the cellular compartment in which they are located, which is similar to that of proteins; thus, localizing their information plays a vital role in understanding their function [8]. The computational prediction of subcellular localization (SL) has been an important topic in bioinformatics due to the fact that SL prediction is difficult to realize through biological experiments [9]. Consequently, there is an immediate necessity to adopt computational approaches in order to accelerate the research on lncRNAs, such as the research on the identification of drug targets [10–13], enhancers [14,15], interactions between ncRNAs and proteins [16,17], circular RNAs [18,19], miRNAs [20–23] and reducing dimensionality [24–27]. However, most prediction tools are constructed for proteins [28–35]. Sequence-based methods for predicting the SL of proteins can be generally categorized into statistical machine learning-based and homologous transfer. Homologous transfer attempts to determine annotated homologous proteins for query sequences from a large database, but failure may be predicted when no homologous proteins are found [36]. The relative slowness of lncRNA annotation and its sequence diversity is due to the difficulty in obtaining lncRNA sequences with definitive homology annotations. Therefore, the machine learning-based approach is suitable for the development of SL predictors for lncRNA at the current stage.

However, the machine learning-based method for predicting lncRNA subcellular localization (LSL) face several challenges. First, traditionally, it was assumed that most lncRNAs are primarily located only in the nucleus and regulate nuclear gene regulators [37]. Recent studies using fluorescence in situ hybridization techniques have revealed that lncRNAs exhibit diverse SLs, with many being found in the cytoplasm [38]. Some lncRNAs are also evenly distributed between the nucleus and cytoplasm. Second, the prediction of LSL via machine learning methods encounters difficulties due to insufficient relevant data.

With the deepening of LSL research, lncRNA subcellular-related databases and tools are constantly being proposed, providing a strong support for the study of LSL based on machine learning. Zhang et al. [39] constructed the RNALocate database, which collected more than 37,700 LSL data entries. Afterward, Mas-Ponte et al. [40] created the LncATLAS database, which is a specialized database for storing LSL data. Chen and Carmichael [37] systematically analyzed the distribution of LSL in gastric cancer, thus revealing its relationship with the existence of cancer development. Feng et al. [41] collected ncRNAs from mitochondrial, kinetoplast, and chloroplast genomes and proposed a model for predicting the location of ncRNA organelles. Cao et al. [8] developed a lncLocator

predictor to predict LSLs based on the k-mer frequency feature, and a deep learning model was proposed to allow the lncLocator predictor to predict LSL. Su et al. [42] proposed the model iLoc-lncRNAL for predicting LSL; it is based on lncRNA sequence octamers and demonstrated good prediction performance.

As mentioned above, although developing the calculation methods for LSL is very important, there are still relatively few studies on it, and there are three main problems: (i) data imbalance, the LSL dataset shows serious unbalanced distribution; for example, recently, some of the proposed methods have not considered the problem of data imbalance [42,43]; (ii) feature extraction of lncRNA sequences, unlike other short ncRNA sequences, due to the specificity of lncRNAs, it is more difficult to capture their feature information; (iii) the prediction accuracy is generally low. For instance, Cao et al. [8] proposed the lncLocator method, which addresses the issue of data imbalance. However, the method achieves an overall accuracy (OA) of only 0.59, suggesting that there is significant room for improvement.

To address those problems, we propose a novel discrete wavelet transform (DWT)-based predictive method, DlncRNALoc, to identify the LSL. Firstly, we introduce aDWT-based representation of lncRNA sequence features. Recently, the application of wavelet analysis in bioinformatics research has received much attention [44,45]. The Fourier coefficients contain only global information about the time domain name; therefore, there is a loss of characteristic information of the signal in the bit domain [46]. The decomposition of lncRNA sequence signals via DWT can remove the redundant information of features and yield the characteristic signals of each type of lncRNA. This is important for improving the prediction performance. Next, we employ the local Fisher discriminant analysis (LFDA) [47] to reduce the noisy information and feature dimensions and combine it with the synthetic minority over-sampling technique (SMOTE) [48] to mitigate the effects of data imbalance. Finally, we construct a support vector machine (SVM) to construct models to predict the LSL. Extensive experiments show that the DlncRNALoc model has excellent performance. The main highlights are as follows.

1) We propose the LP matrix, which converts an lncRNA sequence into an L*6-dimensional matrix by considering its 2-tuple physical structure's properties. The LP matrix can also be applied in other lncRNA research domains.
2) We introduce the DWT feature extraction method by combining the LP matrix with DWT. We utilize six discrete wavelet functions (WFs) (rbio2.4, coif2, db8, sym3, bior3.3, and dmey) for this purpose.
3) We employ the SMOTE to address the data imbalance caused by uneven distribution of lncRNA samples across different subcellular locations.
4) Extensive experimental results demonstrate that DlncRNALoc outperforms several state-of-the-art approaches in the area of LSL.

2. Materials and methods

2.1. Framework of the DlncRNALoc model

This study has yielded a DWT-based LSL prediction model termed DlncRNALoc. The DlncRNALoc consists of two major parts: feature extraction and model construction. We obtain the LP matrix of lncRNA based on its physicochemical properties. From this LP matrix, we extract the DWT

characteristics of each lncRNA. A total of six discrete WFs were used: rbio2.4, coif2, db8, sym3, bior3.3, dmey. Additionally, we use the SMOTE technology to synthesize new samples and the LFDA to mitigate the effects of the high-dimensional features. Furthermore, we adopt the one-to-one (OVO) strategy to construct the SVM classifier. The feature information is fed into the SVM to predict the LSL. Finally, extensive experimental analyses were performed to evaluate the performance of the DlncRNALoc model. Figure 1 shows the overall framework of the DlncRNALoc model.

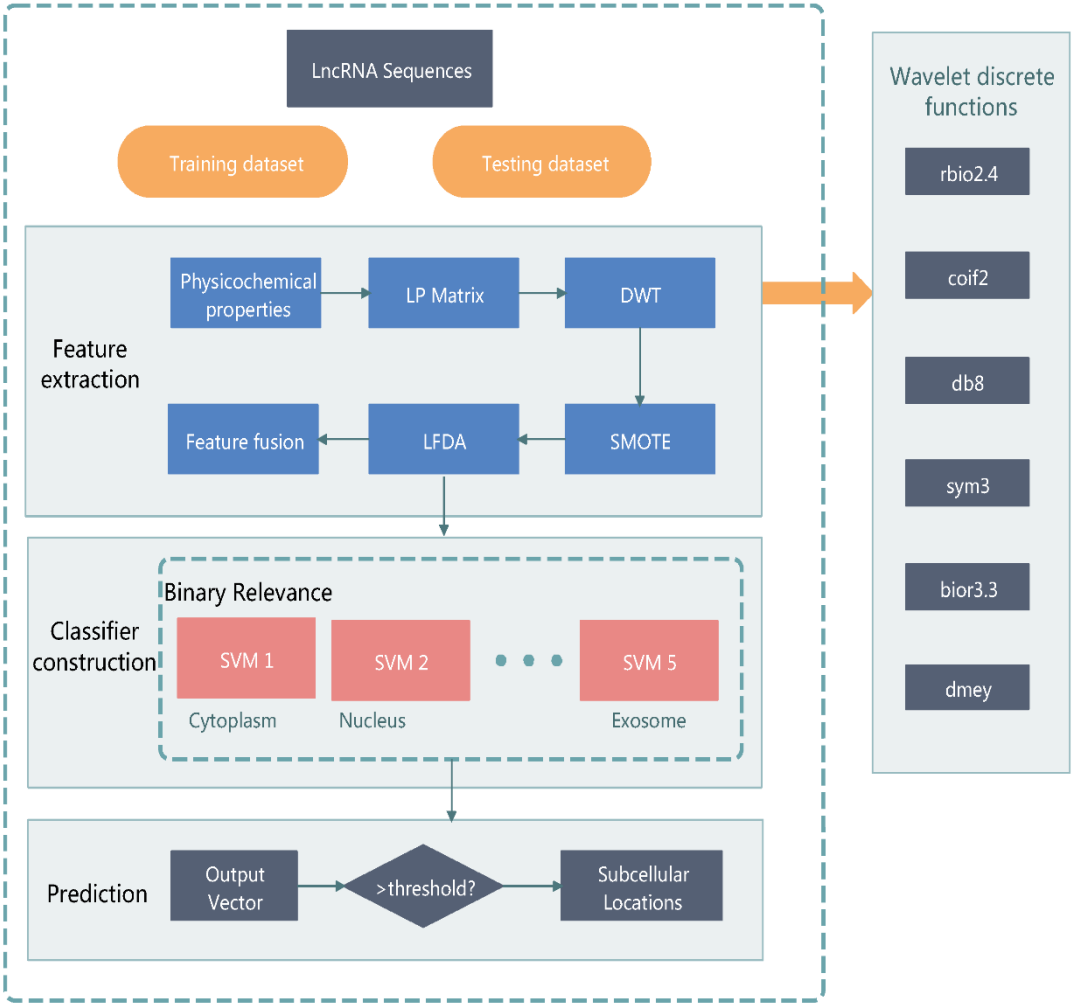

Figure 1. Framework for the DlncRNALoc model.

2.2. Datasets

We used two sets of lncRNA sequence benchmark datasets for SL experiments derived from the RNALocate database. This database serves as a repository for localization information on various RNA molecules, including mRNA, miRNA, and lncRNA molecules. The latest version of RNALocate encompasses over 37,700 manually-planned SL entries and experimental evidence for RNA molecules, and it covers more than 21,800 encoded and non-coding RNAs in 65 species [39]. Table 1 presents the two benchmark datasets for the SL of lncRNA sequences.

Table 1. Benchmark LSL datasets.

Subcellular Locations	Dataset 1	Dataset 2	Dataset 3
Cytoplasm	301	426	328
Cytosol	91	/	/
Exosome	25	30	28
Nucleus	152	156	325
Ribosome	43	43	8
Total	612	655	769

Dataset 1 was created by Cao et al. [8] using the RNALocate database. Dataset 1 consists of 612 lncRNA sequences that are categorized into five subcellular loci: cytoplasm (301), cytosol (91), exosome (25), nucleus (152), and ribosome (43).

Dataset 2 was established by Su et al. [42], and it contains 655 lncRNA sequences divided into four subcellular loci: cytoplasm (426), ribosome (43), exosome (30) and nucleus (156).

Dataset 3 was established by Li et al. [49], and it contains 769 lncRNA sequences divided into four subcellular loci: cytoplasm (328), ribosome (8), exosome (28) and nucleus (325).

To mitigate bias caused by redundant sequences, Datasets 1, 2 and 3 were constructed by using the cd-hit method [50,51] to remove redundant lncRNA sequences with a cutoff value of 80%. Detailed information about Datasets 1, 2 and 3 can be found in Table 1.

2.3. LP matrix

Table 2. The numerical values of the six physical structure-related properties of the 2-tuple of the lncRNA sequence.

2-Nucleotides	Twist	Tilt	Roll	Shift	Slide	Rise
N_1 (AA)	0.026	0.038	0.02	1.69	2.26	7.65
N_2 (AC)	0.036	0.038	0.023	1.32	3.03	8.93
N_3 (AG)	0.031	0.037	0.019	1.46	2.03	7.08
N_4 (AT)	0.033	0.036	0.022	1.03	3.83	9.07
N_5 (CA)	0.016	0.025	0.017	1.07	1.78	6.38
N_6 (CC)	0.026	0.042	0.019	1.43	1.65	8.04
N_7 (CG)	0.014	0.026	0.016	1.08	2	6.23
N_8 (CT)	0.031	0.037	0.019	1.46	2.03	7.08
N_9 (GA)	0.025	0.038	0.02	1.32	1.93	8.56
N_{10} (GC)	0.025	0.036	0.026	1.2	2.61	9.53
N_{11} (GG)	0.026	0.042	0.019	1.43	1.65	8.04
N_{12} (GT)	0.036	0.038	0.023	1.32	3.03	8.93
N_{13} (TA)	0.017	0.018	0.016	0.72	1.2	6.23
N_{14} (TC)	0.025	0.038	0.02	1.32	1.93	8.56
N_{15} (TG)	0.016	0.025	0.017	1.07	1.78	6.38
N_{16} (TT)	0.026	0.038	0.02	1.69	2.26	7.65

The LP matrix represents the physicochemical property matrix of the 16 2-tuples bases of the lncRNA sequence. The lncRNA sequence contains 16 2-tuples formed by combining the four bases: A, T, C, and G. These 2-tuples include AA, AC, AG, and so on, up to TT. Previous studies [52] have shown that each pair of 2-tuple bases contains six physicochemical properties: twist, tilt, slide, roll, shift and rise. The values for these six physicochemical properties corresponding to the 16 2-tuple bases were obtained from the literature [52], and they are presented in Table 2. For the sake of convenience, we represent the 16 2-tuple as N_1,…,N_{16} in Table 2.

Let S=s_1,…,s_L represent a lncRNA base sequence of length L. Then, the sequence S can be described as Eq (1):

$$\mathrm{S} = p_1, \ldots, p_i, \ldots, p_{L-1}; p_i \epsilon \{N_1, \ldots, N_{16}\} \tag{1}$$

Here, p_i represents any adjacent 2-tuples in the sequence S. Therefore, based on Table 1 and Eq (1), the given sequence can be transformed into an LP matrix of $(L-1) \times 6$, as given by Eq (2):

$$LP = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,6} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,6} \\ \vdots & \vdots & \vdots & \vdots \\ p_{\mathrm{i},1} & p_{\mathrm{i},2} & \cdots & p_{\mathrm{i},6} \\ \vdots & \vdots & \vdots & \vdots \\ p_{\mathrm{L}-1,1} & p_{\mathrm{L}-1,2} & \cdots & p_{\mathrm{L}-1,6} \end{bmatrix}_{L-1\times 6} \tag{2}$$

where $p_{i,\mathrm{j}}$ represents the *j*th (j = {1, 2, 3, 4, 5, 6}) physicochemical properties of the *i*th 2-tuple of the sequence S. For example, if S = "ACAGA", then $p_{1,1}$ represents the first physicochemical property "twist" of the first 2-tuple "AC". Using Table 1, we can determine that $p_{1,1}$ = 0.036. Similarly, $p_{2,2}$ represents the second physicochemical property "tilt" of the second 2-tuple "CA" of the sequence, and its value from Table 1 is $p_{2,2}$ = 0.025. These computations can be extended to derive the physicochemical properties for other 2-tuples in the sequence. The collection of all of these 2-tuple forms the LP matrix.

2.4. DWT

DWT [34] is a transformation that captures the discrete sampling of wavelets. It provides significant frequency and positional information for sequences. DWT is the decomposition of the physicochemical properties of lncRNA sequences into coefficients of different resolutions, as achieved by projecting the signal onto a wave function (WF). DWT effectively filters out noise from the high-pass curve. In our approach, we consider the LP matrix of each lncRNA sequence as a two-dimensional signal and utilize DWT for denoising purposes.

Mathematically, WT is the projection of the signal $f(t)$ onto a WF:

$$T(a,b) = \frac{1}{\sqrt{a}} \int_a^t f(t) \Psi\left(\frac{t-b}{a}\right) \mathrm{dt} \tag{3}$$

In this equation, a ($a>0$) is the scale factor and b is the translation factor, both of which are real numbers. $\Psi(\frac{t-b}{a})$ represents the analyzing WF, whereas $T(a,b)$ corresponds to the WT coefficient of the signal at a position ($t=b$) and a particular wavelet period defined by the scale factor a. By using

the DWT, lncRNA sequences can be decomposed into various dilation coefficients, facilitating the elimination of noise components. Nanni et al. [53,54] devised a strategy that utilizes the DWT. If we assume that $f(t)$ is a discrete signal represented by $x[n]$, it can be defined by the following equations.

$$y_{j,low}[n] = \sum_{k=1}^{N} x[k]g[2n-k] \tag{4}$$

$$y_{j,\text{high}}[n] = \sum_{k=1}^{N} x[k]\text{h}[2n-k] \tag{5}$$

The length of the discrete signal is denoted by N. $y_{\text{high}}[n]$ indicates the detailed coefficient, representing its high-frequency component. The approximation coefficients of the signal are denoted by $y_{low}[n]$, which represents the low-frequency component of the signal. The low-pass filter is denoted by g, while the high-pass filter is denoted by h. As the level of decomposition increases, more intricate signal characteristics become discernible. Figure 2 depicts four filters that are compatible with the "dmey" WF. These filters include the decomposition and reconstruction of the low-pass and high-pass filters.

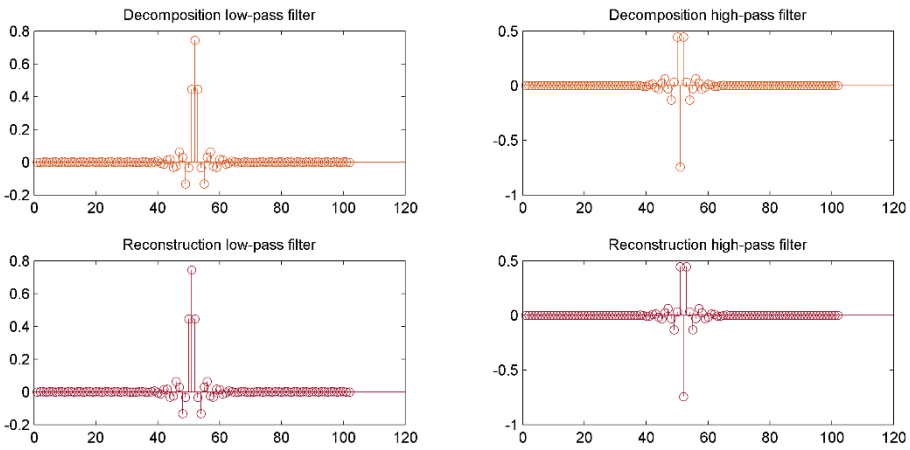

Figure 2. The four filters for dmey WF.

DWT is visually represented in Figure 3 across four distinct levels. At each stage, the data are divided between noisy information in the higher frequency bands and valuable signals in the lower frequency bands. These bands must undergo additional transformations in subsequent steps.

The high-frequency and low-frequency signals are separated at each level of the DWT. We calculated the standard deviation, minimum, maximum, and average values of the bands at each level to derive four features for the high-frequency and low-frequency bands, generating a total of eight features. Furthermore, since these first five elements encapsulate vital information regarding the sequence in the compressed low band, we obtained the initial five discrete cosine coefficients from the approximation coefficients.

Consequently, 52 features are obtained from each level of the DWT, comprising four features each in the high-frequency and low-frequency bands, and five features from the discrete cosine coefficients. By utilizing the 5-level DWT method, it is possible to extract 52 features for each attribute in the LP matrix. As a result, all six physical and chemical properties collectively provide a total of 390 features.

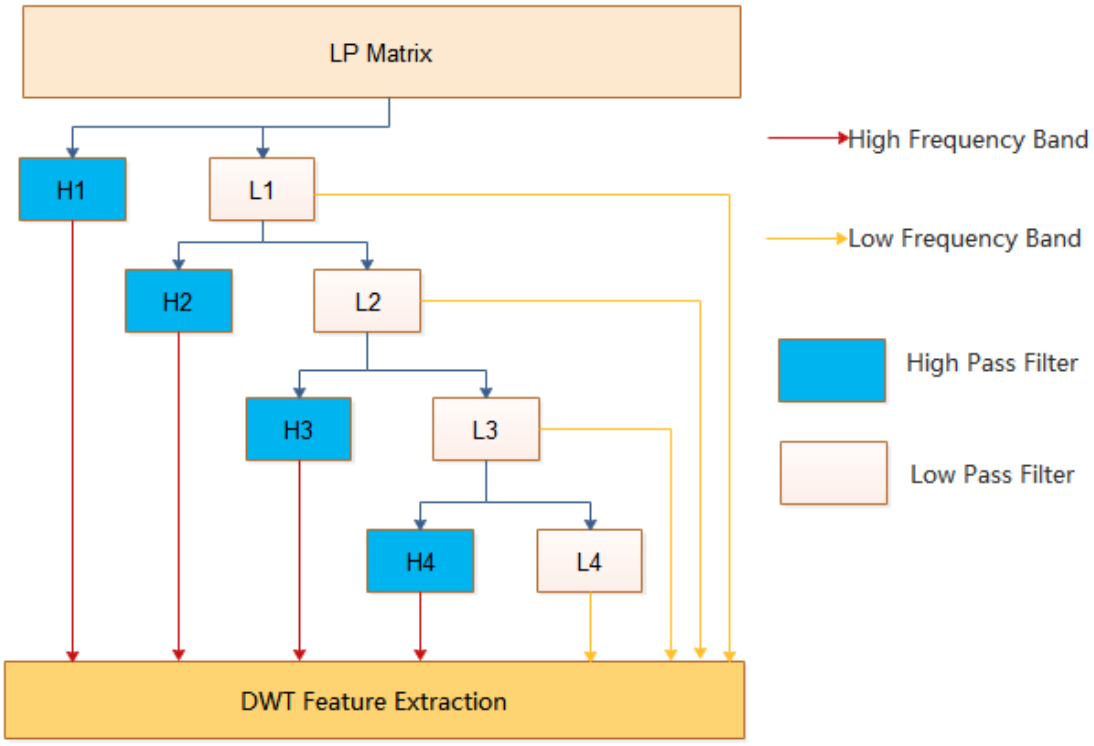

Figure 3. An example of a DWT process.

2.5. SMOTE

Table 1 shows noticeable disparities in the data distribution of lncRNAs across different subcellular locations. Since machine learning algorithms often tend to categorize new samples into the majority class, the presence of an unbalanced data distribution can adversely affect the classification performance of the minority class. This problem becomes more pronounced when dealing with multiple classification tasks. The SMOTE method mitigates the effects of data imbalance by generating synthetic samples and oversampling a few classes of samples. SMOTE technology achieves data balance by synthesizing new samples for the minority class.

The imbalance levels and sampling rates can be derived based on different samples in each category. The number of SLs is denoted by the symbol C, and num_i $(1 \leq i \leq \mathrm{C})$ is the number of samples in the ith subcellular of the benchmark dataset. The imbalance level in the ith category can be calculated using the following formula:

$$IL_i = \frac{num\_max}{num_i}, 1 \leq i \leq C \tag{6}$$

Here, the number of samples in the largest category in the dataset is denoted by num_ma. The sampling rate n_i is calculated by rounding the imbalance level IL_i, as defined below in the following equation:

$$n_i = round(IL_i), 1 \leq i \leq C \tag{7}$$

For each subcellular class with a non-maximum number of lncRNAs, find the nearest *K* samples for each sample in the class and randomly extract n_i samples from the *K* samples (n_i is the sampling magnification, $K > n_i$), denoted as y_1, y_2, ..., y_{n_i}. Therefore, a random interpolation operation between *X* and y_j is performed to obtain an interpolated sample syn_j. The interpolation equation is defined as follows:

$$syn_j = X + rand(0,1) * (y_j - X), j = 1,2,...,n_j \tag{8}$$

Here, *rand*(0,1) indicates a number that is randomly generated in the interval (0,1), and y_j represents a neighboring sample of a few subcellular samples.

2.6. LFDA

Here, we describe our use of a supervised dimensionality reduction technique called LFDA [47]. LFDA offers an embedded transformation with an analytical form and it can be efficiently solved by addressing generalized eigenvalue problems. The LP matrix of the lncRNA base is defined as LP = $[x_1, x_2, \dots, x_i, \dots, x_n]$, $x_i \in R^d$, where n is the number of lncRNA sequence samples, the symbol d denotes the dimension of the eigenvector, $y_i \in \{1,2, \dots, c\}$ indicates the category label and n_ℓ is the total number of lncRNAs of a subcellular ℓ; then, we have

$$\sum_{\ell=1}^{c} n_\ell = n \tag{9}$$

The local intraclass scattering matrix $S^{(w)}$ and the local interclass scattering matrix $S^{(b)}$ are respectively defined as follows:

$$S^{(\omega)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(\omega)} (x_i - x_j)(x_i - x_j)^T \tag{10}$$

$$S^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(\mathrm{b})} (x_i - x_j)(x_i - x_j)^T \tag{11}$$

where,

$$W_{i,j}^{(\omega)} = \begin{cases} A_{i,j} / n_\ell \ \ if \ y_i = y_j = \ell \\ 0 \ if \ y_i \neq y_j \end{cases} \tag{12}$$

$$W_{i,j}^{(b)} = \begin{cases} A_{i,j}((1/\mathrm{n}) - (1/n_\ell)) \ if \ y_i = y_j = \ell \\ 1/n \ if \ y_i \neq y_j \end{cases} \tag{13}$$

Here, A is an affinity matrix, and $A_{i,j} \in A$ is the affinity of x_i and x_j. In this paper, we use the affinity matrix $A_{i,j} = \exp(-\|x_i - x_j\| / \sigma_i \sigma_j)$ defined in [55], $\sigma_i = \|x_i - x_i^K\|$ denotes a local scaling of the samples surrounding x_i and x_i^K is the *k*th nearest neighbor of x_i. Solve the LFDA transformation matrix T_{LFDA} as follows:

$$T_{LFDA} = \arg\max_{T \in R^{t \times r}} \left[tr\left((T^T S^{(\omega)} T)^{-1} T^T S^{(b)} T \right) \right] \tag{14}$$

The resulting dimensionally reduced matrix is described as follows:

$$Z = T_{LFDA} LP \tag{15}$$

2.7. SVM

We implement the SVM module by using the LIBSVM package proposed by Chang and Lin [56], which adopts a one-to-one (OVO) strategy for multi-classification problems. We apply a dataset $\mathrm{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, y_i \in \{C_1, C_2, \ldots, C_k\}$, where y_i is the category label corresponding to x_i. The OVO method involves designing a classifier between the categories of each pair, which requires an aggregate of *k(k-1)*/2 classifiers for *k* categories. During the testing phase, the test sample is evaluated by all classifiers, producing *k(k-1)*/2 classification results. The final prediction is determined via a voting process, where the category receiving the most votes is selected. In cases where two categories have the same number of votes, the category that appeared first during voting is assigned as the final prediction category.

2.8. Evaluation criteria

For a well-established classification prediction model, there must be a test method and some indicators to evaluate the model. We choose to use 5-fold cross-validation and independent validation methods to evaluate the performance of LSL prediction models and apply the Matthew's correlation coefficient (MCC), OA, sensitivity (SE), specificity (SP), recall, and F_1-score (F_1) as the evaluation metrics. The equations to calculate these metrics are as follows:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

$$SE^{(i)} = \frac{TP^{(i)}}{TP^{(i)} + FN^{(i)}} \tag{17}$$

$$SP^{(i)} = \frac{TN^{(i)}}{TN^{(i)} + FP^{(i)}} \tag{18}$$

$$MCC^{(i)} = \frac{TP^{(i)} \times TN^{(i)} - FP^{(i)} \times FN^{(i)}}{\sqrt{(\mathrm{TP}^{(i)} + \mathrm{FN}^{(i)})(\mathrm{TN}^{(i)} + \mathrm{FP}^{(i)})(\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)})(\mathrm{TN}^{(i)} + \mathrm{FN}^{(i)})}} \tag{19}$$

$$F_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \times Precision^{(i)} \times Recall^{(i)}}{Precision^{(i)} + Recall^{(i)}} \tag{20}$$

$$\mathrm{Recall} = \frac{1}{n} \sum_{i=1}^{n} Recall^{(i)} \tag{21}$$

where $Precision^{(i)}$ indicates the precision of the *i*th class. TP, FP, TN and FN denote the numbers of true positive, true negative, false positive and false negative instances, respectively.

3. Results

3.1. Selection of WF

The concept of WT involves transforming a signal into a representation composed of wavelet basis functions. Different families of WFs generate distinct wavelet basis functions, each offering unique signal processing capabilities and outcomes [57,58]. The effectiveness of feature extraction from a sequence is enhanced when the characteristics of the WF align well with the structure of the analyzed signal. It is widely acknowledged that WFs possess several desirable features, including compact support, symmetry, orthogonality, smoothness, and high-order vanishing moments. However, the choice of WF involves certain conflicting constraints as no single one possesses all of these characteristics simultaneously. In order to select the characteristic information that can effectively extract the LSL, in this study, six WFs were mainly investigated: discrete meyer (dmey), daubechies of number 8 (Db8), biorthogonals of number 3.3 (Bior3.3), coiflet of number 2 (Coif2), symlets of number 3 (Sym3) and rbio2.4. The results of applying the LSL prediction models using the above six WFs on the benchmark datasets, Datasets 1 and 2 are shown in Tables 3 and 4, respectively. We used jackknife cross-validation as the evaluation method and selected sensitivity (SE) and overall accuracy as the evaluation indicators.

Based on Table 3, the dmey WF yields the highest OA, i.e., 87.07%, and the OA of the coif2 function is the smallest among the five WFs, i.e., rbio2.4, coif2, db8, sym3, and bior3.3; the OA of the rbio2.4 function is 86.53%, which is very close to that of the dmey function.According to the dmey WF, except for a lower sensitivity in predicting lncRNA subcellular localization in the nucleus (MCC = 0.655) compared to the rbio2.4 function, the prediction accuracies of lncRNA for the other three subcellular compartments (cytoplasm, ribosome, and exosome) are 0.714, 0.960, and 0.966, respectively, which are the highest values.

Table 3. Predictive performance of different WFs in terms of LSL on Dataset 1.

Algorithm	*MCC*				O*A*
	Nucleus	Cytoplasm	Ribosome	Exosome	
rbio2.4	0.695	0.709	0.921	0.950	86.53%
coif2	0.593	0.617	0.905	0.918	81.95%
db8	0.606	0.638	0.908	0.959	83.43%
sym3	0.622	0.645	0.901	0.958	83.84%
bior3.3	0.656	0.647	0.885	0.972	84.24%
dmey	0.655	0.714	0.960	0.966	87.07%

As shown in Table 4, the dmey WF achieves the maximum OA of 91.28%. Among the five WFs, rbio 2.4, coif2, db8, sym3 and bior 3.3, the OA of the sym3 function is the smallest (OA = 88.24%). The OA of the bior 3.3 function is 90.06%, which is very close to that of the dmey function.

In the case of the dmey WF, the MCC values of lncRNA predicted for the cytosol and nucleus are 0.962 and 0.757, respectively, which are the best among several methods. Although the MCC values for the exosome (MCC = 0.994) and nucleus (MCC = 0.996) are not the best, they are very close to the best. Note that the OA of the db8 WF is found to be 89.93%. Although this is not very

prominent, the MCC values for the cytoplasm and ribosome are 0.774 and 0.998, respectively. These values are better than those for the other WFs.

Table 4. Predictive performance of different WFs in terms of LSL on Dataset 2.

Algorithm	*MCC*					OA
	Cytoplasm	Cytosol	Exosome	Nucleus	Ribosome	
rbio2.4	0.764	0.890	0.998	0.697	0.994	89.45%
coif2	0.705	0.926	0.992	0.691	0.981	88.57%
db8	0.774	0.897	0.990	0.717	0.998	89.93%
sym3	0.693	0.907	0.994	0.695	0.977	88.24%
bior3.3	0.764	0.912	0.998	0.715	0.998	90.06%
dmey	0.753	0.962	0.994	0.757	0.996	91.28%

Figures 4 and 5 show the sensitivity and specificity of the six WFs on Datasets 1 and 2, respectively, where each sub-cell is mapped with a different color across a radar chart.

It can be seen in Figure 4 that the sensitivity of the six WFs is ranked as follows: nucleus, cytoplasm, ribosome, and exosome for the four sub-cells; also, the cytoplasm, nucleus, exosome, and ribosome have been sorted from small to large. Similarly, as shown in Figure 5(A), the exosome and ribosome curves almost coincide, indicating that the six WFs are almost identical in sensitivity to the exosome and ribosome; the sensitivity of the other three sub-cells has been ordered from small to large: nucleus, cytoplasm and cytosol. In Figure 5(B), the subcellular curve is divided into two groups: cytoplasm and nucleus curves, and exosome, cytosol and exosome curves.

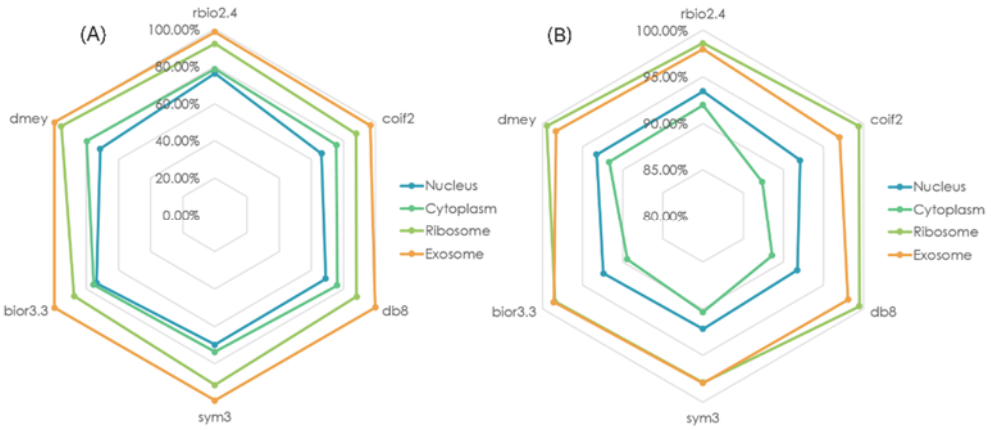

Figure 4. Values of sensitivity and specificity for the six WFs with four sub-cells on Dataset 1. (A) Sensitivity, and (B) specificity; each point-to-center distance measurement represents the value of sensitivity and specificity obtained for each WF, respectively.

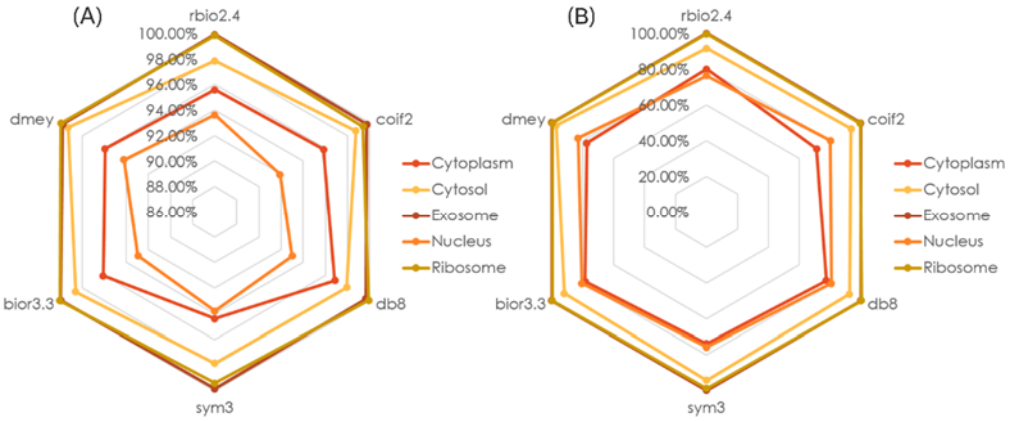

Figure 5. Values of sensitivity and specificity for the six WFs with five sub-cells on Dataset 2. (A) Sensitivity, and (B) specificity; each point-to-center distance measurement represents the value of sensitivity and specificity obtained for each WF, respectively.

Based on the above analysis, different WFs are used to predict the performance on different sub-cells, indicating that different WFs can capture different types of characteristic information of lncRNA sequences. The predicted overall performance of the dmey WF is found to be optimal. Therefore, we applied the dmey WF to construct the prediction model.

3.2. Comparison with state-of-the-art methods

In order to objectively assess the performance of the DlncRNALoc model in terms of predicting LSLs, we introduced three benchmark datasets. We compared their use with different existing methods on each benchmark dataset.

Table 5. Comparison of the DlncRNALoc model with existing methods on Dataset 1.

Method	Location	SE	SP	MCC	OA
iLoc-lncRNA	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%	100%	0.400	
lncLocator	Nucleus	38.15%	92.17%	0.357	66.50%
	Cytoplasm	88.01%	36.36%	0.288	
	Ribosome	7.00%	97.53%	0.070	
	Exosome	4.00%	97.27%	0.015	
DlncRNALoc	Nucleus	72.44%	93.44%	0.666	87.41%
	Cytoplasm	80.28%	91.97%	0.722	
	Ribosome	95.61%	99.45%	0.960	
	Exosome	100.00%	98.31%	0.966	

Table 5 presents the results of iLoc-lncRNA, lncLocator, and DlncRNALoc on Dataset 1,

respectively, using SE, SP, MCC and OA as the evaluation criteria. Table 5 shows that the OAs for the three LSL prediction methods are 86.72%, 66.50%, and 87.41%, respectively. Obviously, the OA of the DlncRNALoc model is the largest, and the OA of the lncLocator model is the lowest, at only 66.50%. The DcncRNALoc model predicted MCC values for lncRNA for the four sub-cells of nucleus, cytoplasm, ribosome and exosome: 0.666, 0.722, 0.960 and 0.966, respectively. Among them, the MCC values for the ribosome and exosome are the best, and at least 30% higher than the second-ranked iLoc-lncRNA prediction model; additionally, the MCCs for the nucleus and cytoplasm sub-cells are ranked second, but the value of the MCC for the first-ranked iLoc-lncRNA is very close.

Table 6. Comparison of the DlncRNALoc model with existing methods on Dataset 2.

Method	OA	F1	Recall
LoR ensemble	58.10%	23.50%	24.60%
Average ensemble	60.00%	25.50%	26.40%
lncLocator	59.10%	29.10%	30.20%
DlncRNALoc	91.08%	91.19%	91.22%

Table 6 indicates the experimental results of the DlncRNALoc prediction model, LoR ensemble, average ensemble, and lncLocator prediction model on Dataset 2, and in terms of the OA, recall, and F_1 as evaluation indicators. Among them, the LoR ensemble, average ensemble, and lncLocator methods' experimental results were taken from the literature [8]. The values of OA, F_1 and recall for the DlncRNALoc model are 91.08%, 91.19% and 91.22%, respectively, which are larger than those for the other prediction methods.

In order to compare advanced deep learning methods, we have utilized a dataset from the literature [49] called Dataset 3. We selected three baseline methods, namely, the convolutional neural network (CNN), long short-term memory (LSTM), and GraphLncLoc [49]. The results were verified via a five-fold cross-validation experiment, and they are shown in Table 7. The experimental results of the CNN and LSTM methods were obtained from the literature [49]. As depicted in Table 7, the DlncRNALoc method yields performance metrics of 69.44% for OA, 57.56% for the F_1 score and 55.00% for recall. These results are superior to those obtained via the other evaluated methods.

Table 7. Comparison of the DlncRNALoc with existing prediction models on Dataset 3.

Method	OA	F_1	Recall
CNN	58.00%	40.20%	39.400%
LSTM	56.60%	42.50%	42.50%
GraphLncLoc	61.20%%	50.60%	47.50%
DlncRNALoc	**69.44%**	**57.56%**	**55.00%**

To comprehensively assess the efficacy of DlncRNALoc in predicting the LSL, we evaluate it alongside existing methods for comparison. We employed an independent test set—furnished by GraphLncLoc—comprising sequences from the cytoplasm (20), ribosome (10), nucleus (20), and exosome (7). Our chosen baseline methods included lncLocator [8], iLoc-lncRNA [42], Locate-R [59], DeepLncLoc [60], and GraphLncLoc [49]. Table 8 shows the results of this independent test. The DlncRNALoc method outperforms other methods, as evidenced by an OA of 59.65%, an F_1 score of

63.16%, and a recall of 61.43%.

Table 8. Comparison of DlncRNALoc with existing methods on an independent test set.

Method	OA	F_1	Recall
lncLocator	42.10%	28.9%	32.50%
iLoc-lncRNA	50.90%	47.4%	47.00%
Locate-R	36.8%	32.1%	32.10%
DeepLncLoc	56.1%	58.2%	54.30%
GraphLncLoc	57.9%	58.4%	55.70%
DlncRNALoc	**59.65%**	**63.16%**	**61.43%**

The main reasons for the superior performance of DlncRNALoc compared to deep learning methods can be attributed to the following factors: 1) Deep learning models achieve high prediction accuracy by leveraging large-scale datasets and increasingly complex models. Consequently, training these models becomes progressively challenging, often leading to suboptimal performance. 2) lncRNA sequences vary significantly in length, such as in Dataset 3, where the sequence length ranged from 126 to 551,120. Deep learning models often adopt a fixed sequence length and employ rules to handle shorter or longer sequences. However, this approach leads to inevitable information loss or wasted space. It is to be noted that the experimental results featured in Tables 7 and 8 were obtained without the utilization of SMOTE for data processing. To summarize, The DlncRNALoc model has good predictive performance and effectiveness in LSL prediction.

4. Conclusions

The role of SL in understanding the intricate biological functions of lncRNAs underscores its importance in research, particularly in predicting LSL. We developed a DlncRNALoc approach for predicting LSL. This approach incorporates various features, encoding techniques, and machine learning methodologies to improve prediction performance and provide insight into the mechanisms that govern LSL.

The initial step involved the construction of the LP matrix, utilizing a 2-tuple of physicochemical properties of the bases in the lncRNA sequence. This was followed by the implementation of a DWT-based feature extraction, which employed the LP matrix. The SMOTE was then incorporated for sample generation, and the LFDA algorithm was implemented for optimal feature extraction through information dimensionality reduction. Finally, the LSL was predicted by using SVM algorithm. Evaluation of the DlncRNALoc predictions on two benchmark datasets yielded overall accuracies of 87.41% and 91.08%, respectively. Comparison with existing approaches revealed significant improvement of the performance for LSL, signifying the potential of the proposed method to predict other lncRNA properties and functions.

This study focuses on exploring the representation of lncRNA sequences. In the future, it is recommended to fuse structural information and physicochemical property information to further enhance the accuracy of LSL prediction. Additionally, conducting extensive benchmarking and comparative evaluations of DlncRNALoc with other computational methods on diverse datasets is necessary to evaluate their performance and identify their strengths and weaknesses. Moreover, it is essential to investigate methods for interpretable DlncRNALoc predictions that provide insights into

the specific features or characteristics contributing to the LSL. Such an investigation can improve model transparency and facilitate biological interpretation. By addressing these future research directions, this study can significantly advance our understanding and prediction of LSL, ultimately contributing to our knowledge of lncRNA function and cellular regulation.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China (Nos. 62002111, 62372158), the Natural Science Foundation of Hunan Province (No. 2022JJ40090). And it was also supported by a grant from the 'Macao Young Scholars Program' (Project code: AM2021025).

The authors would also like to express their gratitude to Prof. Yao Xiaojun from Macao Polytechnic University for his invaluable suggestions and enlightening discussions, which significantly improved the clarity and overall presentation of this manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Y. Wu, S. Ma, Impact of COVID-19 on energy prices and main macroeconomic indicators—evidence from China's energy market, *Green Finance*, **3** (2021), 383–402. https://doi.org/10.3934/GF.2021019
2. E. Assifuah-Nunoo, P. O. Junior, A. M. Adam, B. Ahmed, Assessing the safe haven properties of oil in African stock markets amid the COVID-19 pandemic: a quantile regression analysis, *Quant. Finance Econ.*, **6** (2022), 244–269. https://doi.org/10.3934/QFE.2022011
3. L. Katusiime, Time-Frequency connectedness between developing countries in the COVID-19 pandemic: The case of East Africa, *Quant. Finance Econ.*, **6** (2022), 722–748. https://doi.org/10.3934/QFE.2022032
4. Z. Li, B. Mo, H. Nie, Time and frequency dynamic connectedness between cryptocurrencies and financial assets in China, *Int. Rev. Econ. Finance*, **86** (2023), 46–57. https://doi.org/10.1016/j.iref.2023.01.015
5. A. Narvekar, D. Guha, Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession, *Data Sci. Finance Econ.*, **1** (2021), 180–195. https://doi.org/10.3934/DSFE.2021010
6. Q. Yang, F. Lin, Y. Wang, M. Zeng, M. Luo, Long noncoding RNAs as emerging regulators of COVID-19, *Front. Immunol.*, **12** (2021), 700184. https://doi.org/10.3389/fimmu.2021.700184
7. R. Wu, Y. Su, H. Wu, Y. Dai, M. Zhao, Q. Lu, Characters, functions and clinical perspectives of long non-coding RNAs, *Mol. Genet. Genomics*, **291** (2016), 1013–1033. https://doi.org/10.1007/s00438-016-1179-y

8. Z. Cao, X. Pan, Y. Yang, Y. Huang, H. Shen, The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier, *Bioinformatics*, **34** (2018), 2185–2194. https://doi.org/10.1093/bioinformatics/bty085
9. K. Chou, H. Shen, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, *Nat. Protoc.*, **3** (2008), 153–162. https://doi.org/10.1038/nprot.2007.494
10. J. Brennecke, A. Stark, R. B. Russell, S. M Cohen, Principles of MicroRNA–target recognition, *PLoS Biol.*, **3** (2005). https://doi.org/10.1371/journal.pbio.0030085
11. J. Wei, L. Zhuo, S. Pan, X. Lian, X. Yao, X. Fu, HeadTailTransfer: An efficient sampling method to improve the performance of graph neural network method in predicting sparse ncRNA-protein interactions, *Comput. Biol. Med.*, **157** (2023), 106783. https://doi.org/10.1016/j.compbiomed.2023.106783
12. L. Peng, J. Tan, W. Xiong, L. Zhang, Z. Wang, R. Yuan, et al., Deciphering ligand–receptor-mediated intercellular communication based on ensemble deep learning and the joint scoring strategy from single-cell transcriptomic data, *Comput. Biol. Med.*, **163** (2023), 107137. https://doi.org/10.1016/j.compbiomed.2023.107137
13. L. Peng, R. Yuan, C. Han, G. Han; J. Tan, Z. Wang, et al., CellEnBoost: A boosting-based ligand-receptor interaction identification model for cell-to-cell communication inference, *IEEE Trans. Nanobiosci.*, **22** (2023), 705–715. https://doi.org/10.1109/TNB.2023.3278685
14. L. Cai, X. Ren, X. Fu, L. Peng, M. Gao, X. Zeng, iEnhancer-XG: interpretable sequence-based enhancers and their strength predictor, *Bioinformatics*, **37** (2020), 1060–1067. https://doi.org/10.1093/bioinformatics/btaa914
15. L. Cai, X. Ren, X. Fu, M. Gao, P. Wang, J. Xu, et al., iEnhancer-CLA: Self-attention-based interpretable model for enhancers and their strength prediction, bioRxiv, 2021. https://doi.org/10.1101/2021.11.23.469658
16. L. Zhuo, B. Song, Y. Liu, Z. Li, X. Fu, Predicting ncRNA-protein interactions based on dual graph convolutional network and pairwise learning, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbac339
17. Z. Zhou, Z. Du, J. Wei, L. Zhuo, S. Pan, X. Fu, et al., MHAM-NPI: Predicting ncRNA-protein interactions based on multi-head attention mechanism, *Comput. Biol. Med.*, **163** (2023), 107143. https://doi.org/10.1016/j.compbiomed.2023.107143
18. W. Liu, T. Tang, X. Lu, X. Fu, Y. Yang, L. Peng, MPCLCDA: predicting circRNA-disease associations by using automatically selected meta-path and contrastive learning, *Briefings Bioinf.*, **24** (2023). https://doi.org/10.1093/bib/bbad227
19. L. Peng, C. Yang, Y. Chen, W. Liu, Predicting CircRNA-disease associations via feature convolution learning with heterogeneous graph attention network, *IEEE J. Biomed. Health. Inf.*, **27** (2023), 3072–3082. https://doi.org/10.1109/JBHI.2023.3260863
20. Z. Li, Y. Zhang, Y. Bai, X. Xie, L. Zeng, IMC-MDA: Prediction of miRNA-disease association based on induction matrix completion, *Math. Biosci. Eng.*, **20** (2023), 10659–10674. https://doi.org/10.3934/mbe.2023471
21. J. Wei, L. Zhuo, Z. Zhou, X. Lian, X. Fu, X. Yao, GCFMCL: predicting miRNA-drug sensitivity using graph collaborative filtering and multi-view contrastive learning, *Briefings Bioinf.*, **24** (2023). https://doi.org/10.1093/bib/bbad247

22. X. Fu, W. Zhu, L. Cai, B. Liao, L. Peng, Y. Chen, et al., Improved Pre-miRNAs Identification Through Mutual Information of Pre-miRNA Sequences and Structures, *Front. Genet.*, **10** (2019). https://doi.org/10.3389/fgene.2019.00119
23. Q. Qu, X. Che, B. Ning, X. Zhang, H. Ni, L. Zeng, et al., Prediction of miRNA-disease associations by neural network-based deep matrix factorization, *Methods*, **212** (2023), 1–9. https://doi.org/10.1016/j.ymeth.2023.02.003
24. W. F. Lawless, Autonomous human-machine teams: Reality constrains logic, but hides the complexity of data dependency, *Data Sci. Finance Econ.*, **2** (2022), 464–499. https://doi.org/10.3934/DSFE.2022023
25. L. Peng, F. Wang, Z. Wang, J. Tan, L. Huang, X. Tian, et al., Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies, *Briefings Bioinf.*, **23** (2022), bbac234. https://doi.org/10.1093/bib/bbac234
26. Z. Li, L. Ang, W. Shi, N. Xin, M. Chen, H. Tang, Informative SNP selection based on a fuzzy clustering and improved binary particle swarm optimization algorithm, *Comput. Math. Methods Med.*, **2022** (2022). https://doi.org/10.1155/2022/3837579
27. P. Zweifel, Expanding insurability through exploiting linear partial information, *Data Sci. Finance Econ.*, **2** (2022), 1–16. https://doi.org/10.3934/DSFE.2022001
28. A. Pierleoni, P. L. Martelli, R. Casadio, MemLoci: predicting subcellular localization of membrane proteins in eukaryotes, *Bioinformatics*, **27** (2011), 1224–1230. https://doi.org/10.1093/bioinformatics/btr108
29. H. Shen, K. Chou, Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, *Biochem. Biophys. Res. Commun.*, **355** (2007), 1006–1011. https://doi.org/10.1016/j.bbrc.2007.02.071
30. L. Cai, L. Wang, X. Fu, C. Xia, X. Zeng, Q. Zou, ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation, *Briefings Bioinf.*, **22** (2020). https://doi.org/10.1093/bib/bbaa367
31. X. Fu, W. Zhu, B. Liao, L. Cai, L. Peng, J. Yang, Improved DNA-Binding protein identification by incorporating evolutionary information into the Chou's PseAAC, *IEEE Access*, **6** (2018), 66545–66556. https://doi.org/10.1109/ACCESS.2018.2876656
32. X. Fu, L. Cai, X. Zeng, Q. Zou, StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency, *Bioinformatics*, **36** (2020). https://doi.org/10.1093/bioinformatics/btaa131
33. L. Cai, L. Wang, X. Fu, X. Zeng, Active Semisupervised model for improving the identification of anticancer peptides, *ACS Omega*, **6** (2021), 23998–24008. https://doi.org/10.1021/acsomega.1c03132
34. Y. Wang, Y. Zhai, Y. Ding, Q. Zou, SBSM-Pro: Support bio-sequence machine for proteins, 2023, preprint, arXiv:230810275. https://doi.org/10.48550/arXiv.2308.10275
35. R. Wang, Z. Zhou, X. Wu, X. Jiang, L. Zhuo, M. Liu, et al., An effective plant small secretory peptide recognition model based on feature correction strategy, *J. Chem. Inf. Model.*, **2023** (2023). https://doi.org/10.1021/acs.jcim.3c00868
36. H. Shen, K. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, *Anal. Biochem.*, **373** (2008), 386–388. https://doi.org/10.1016/j.ab.2007.10.012

37. L. Chen, G. G. Carmichael, Decoding the function of nuclear long non-coding RNAs, *Curr. Opin. Cell Biol.*, **22** (2010), 357–364. https://doi.org/10.1016/j.ceb.2010.03.003
38. M. N. Cabili, M. C. Dunagin, P. D. McClanahan, A. Biaesch, O. Padovan-Merhar, A. Regev, et al., Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution, *Genome Biol.*, **16** (2015), 1–16. https://doi.org/10.1186/s13059-015-0586-4
39. T. Zhang, P. Tan, L. Wang, N. Jin, Y. Li, L. Zhang, et al., RNALocate: a resource for RNA subcellular localizations, *Nucleic Acids Res.*, **45** (2016), D135–D138. https://doi.org/10.1093/nar/gkw728
40. D. Masponte, J. Carlevarofita, E. Palumbo, T. H. Pulido, R. Guigo, R. Johnson, LncATLAS database for subcellular localization of long noncoding RNAs, *RNA*, **23** (2017), 1080–1087. https://doi.org/10.1261/rna.060814.117
41. P. Feng, J. Zhang, H. Tang, W. Chen, H. Lin, Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions, *Interdiscip. Sci.: Comput. Life Sci.*, **9** (2017), 540–544. https://doi.org/10.1007/s12539-016-0193-4
42. Z. Su, Y. Huang, Z. Zhang, Y. Zhao, D. Wang, W. Chen, et al., iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, *Bioinformatics*, **34** (2018), 4196–4204. https://doi.org/10.1093/bioinformatics/bty508
43. B. L. Gudenas, L. Wang, Prediction of lncRNA subcellular localization with deep learning from sequence features, *Sci. Rep.*, **8** (2018), 16385. https://doi.org/10.1038/s41598-018-34708-w
44. B. Yu, Y. Zhang, The analysis of colon cancer gene expression profiles and the extraction of informative genes, *J. Comput. Theor. Nanosci.*, **10** (2013), 1097–1103. https://doi.org/10.1166/jctn.2013.2812
45. B. Yu, Y. Zhang, A simple method for predicting transmembrane proteins based on wavelet transform, *Int. J. Biol. Sci.*, **9** (2013), 22–33. https://doi.org/10.7150/ijbs.5371
46. P. M. Bentley, J. T. E. Mcdonnell, Wavelet transforms: an introduction, *Electron. Commun. Eng. J.*, **6** (1994), 175–186. https://doi.org/10.1049/ecej:19940401
47. M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, *J. Mach. Learn. Res.*, **8** (2007), 1027–1061.
48. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.1613/jair.953
49. M. Li, B. Zhao, R. Yin, C. Lu, F. Guo, M. Zeng, GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation, *Briefings Bioinf.*, **24** (2023), bbac565. https://doi.org/10.1093/bib/bbac565
50. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22** (2006), 1658. https://doi.org/10.1093/bioinformatics/btl158
51. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT, *Bioinformatics*, **28** (2012), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565
52. J. R. Goñi, A. Pérez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, *Genome Biol.*, **8** (2007), R263. https://doi.org/10.1186/gb-2007-8-12-r263
53. L. Nanni, S. Brahnam, A. Lumini, Wavelet images and Chou's pseudo amino acid composition for protein classification, *Amino Acids*, **43** (2012), 657–665. https://doi.org/10.1007/s00726-011-1114-9

54. L. Nanni, A. Lumini, S. Brahnam, An empirical study of different approaches for protein classification, *Sci. World J.*, **2014** (2014). https://doi.org/10.1155/2014/236717
55. L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, *Adv. Neural Inf. Process. Syst.*, **17** (2004).
56. C. Chang, C. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, **2** (2011), 27. https://doi.org/10.1145/1961189.1961199
57. J. Qiu, S. Luo, J. Huang, R. Liang, Using support vector machines for prediction of protein structural classes based on discrete wavelet transform, *J. Comput. Chem.*, **30** (2009), 1344–1350. https://doi.org/10.1002/jcc.21115
58. Y. Wang, Y. Ding, F. Guo, L. Wei, J. Tang, Improved detection of DNA-binding proteins via compression technology on PSSM information, *PLoS One*, **12** (2017). https://doi.org/10.1371/journal.pone.0185587
59. A. Ahmad, H. Lin, S. J. G. Shatabda, Locate-R: subcellular localization of long non-coding RNAs using nucleotide compositions, *Genomics*, **112** (2020), 2583–2589. https://doi.org/10.1016/j.ygeno.2020.02.011
60. M. Zeng, Y. Wu, C. Lu, F. Zhang, F. Wu, M. Li, DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding, *Briefings Bioinf.*, **23** (2022), bbab360. https://doi.org/10.1093/bib/bbab360

AIMS Press