ISSN 2616-5775 Vol. 7, Issue 3: 68-72, DOI: 10.25236/AJCIS.2024.070310

Application of Graph Attention Networks in LncRNA Subcellular Localization Prediction

Xi Deng[a,*]

Yunnan Normal University, Kunming, China
[a]18213835546@163.com
**corresponding author*

Abstract: *LncRNAs are crucial in gene regulation and associated with diseases and biological processes. Predicting their subcellular localization accurately remains a challenge due to sequence complexity and data imbalance. We propose a graph neural network method based on LncRNA sequence features, emphasizing enhanced prediction accuracy through optimized graph structure and attention mechanisms. Our approach addresses data imbalance by introducing a weighted graph attention mechanism and a corrective network for improved generalization with fewer samples. This study introduces a potential method for LncRNA subcellular localization prediction, highlighting GNN applicability in bioinformatics tasks. These innovations contribute to biological data analysis and understanding LncRNA function, with practical applications in experimental validation.*

Keywords: *Long non-coding RNA (LncRNA), Graph Neural Network (GNN), Sequence feature similarity, Subcellular localization, Weighted graph attention mechanism*

1. Introduction

Long non-coding RNA (lncRNA) is a type of RNA molecule that does not encode proteins and has a length of over 200 nucleotides. It plays crucial roles in various key biological processes, including gene expression regulation and cell differentiation. An increasing number of studies have revealed the relevance of lncRNAs to the pathogenesis of numerous diseases, especially in cancer, cardiovascular, and neurodegenerative diseases. Unlike coding RNAs, the functions of lncRNAs are often closely linked to their subcellular localization within cells, making accurate subcellular localization prediction essential for understanding their functions and associated disease mechanisms.

When exploring the subcellular localization of lncRNAs, understanding their sequence features significantly contributes to localization prediction. Sequence features provide foundational clues to reveal the structure and function of lncRNAs, and they are core executors of various key biological processes within cells, such as transcriptional regulation, gene silencing, and chromatin remodeling. Therefore, accurate extraction and interpretation of these sequence features play a pivotal role in predicting the subcellular localization of lncRNAs.

However, predicting the subcellular localization of lncRNAs using computational models is a highly challenging bioinformatics task, primarily due to two major difficulties: first, compared to coding RNAs, there is often a lack of sufficient and balanced training data available for lncRNAs; second, although existing studies have attempted to alleviate these issues through techniques such as sampling algorithms, they still struggle to effectively avoid risks such as overfitting. These circumstances greatly limit the accuracy and reliability of lncRNA localization prediction models.

To address these challenges, we employ deep learning technologies, especially the powerful processing capabilities of Graph Neural Networks (GNNs), to delve into the hidden, deep-level features within lncRNA sequences. GNNs can efficiently encode complex sequence data and capture the intrinsic relationships between lncRNAs through their graph structure. Furthermore, our proposed model combines graph attention mechanisms and weighted loss strategies to mitigate the adverse effects of sample imbalance on prediction models, aiming to enhance the accuracy and generalization performance of the model to the fullest extent without expanding the overall sample size. Through this innovative approach, we aim to provide a more reliable and precise computational prediction tool for understanding the complex functions of lncRNAs and their disease mechanisms.

2. Experiment and Methods

2.1. Dataset

In the implementation of LncRNA subcellular localization research, a large amount of experimental data was obtained, primarily sourced from the well-known RNA subcellular localization database RNALocate (http://www.rna-society.org/rnalocate/)[1]. This database provides researchers with an interactive platform that records extensive RNA subcellular localization information and allows users to query and browse detailed data in a concise and intuitive manner. In conducting LncRNA localization research, we adhered to a core principle, the richer the data, the more meaningful the research results.

Following this principle, we constructed an initial five-class dataset containing 612 LncRNA sequences, referred to as Dataset1. To improve data quality, we proactively removed sequences that could lead to information redundancy and noise interference. Specifically, an abnormally long LncRNA sequence (length of 91,671 nucleotides) was removed, and 11 LncRNA sequences containing non-standard nucleotide symbols (such as "N", "R", "S", "Y") were filtered out. Through this cleaning step, we obtained the refined dataset Dataset2, which consists of 600 high-quality LncRNA sequences. These sequences were classified into five categories based on their subcellular localization, with the specific distribution as follows: Cytoplasm 292 sequences, Nucleus 149 sequences, Cytosol 91 sequences, Ribosome 43 sequences, Exosome 25 sequences, as shown in Table 1.

Table 1: Benchmark dataset

	Dataset1(original)	*Dataset2(after filter)*
Cytoplasm	*301*	*292*
Nucleus	*152*	*149*
Cytosol	*91*	*91*
Ribosome	*43*	*43*
Exosome	*25*	*25*
Total	*612*	*600*

2.2. Sequence Feature Extraction: k-mer Method

The k-mer method is a popular technique in feature extraction, where "k-mer" refers to all possible combinations of nucleotide sequences with a length of k. By counting the frequencies of various k-mers in a given sequence, it can be converted into quantitative features, which can then be used for training graph neural network models. For example, in the case of k equals 3 (i.e., 3-mer), all possible combinations of three nucleotides (such as AAA, AAC, AAG, …, TTT) will be counted for their occurrences in the LncRNA sequence.

The advantages of the k-mer method lie in its simplicity and directness. It can reveal inherent patterns in sequences without the need for complex biological annotations. k-mer features can capture local sequence information and can be used to train models to identify sequence feature patterns related to specific subcellular localization.

2.3. Building Graph Structure

Through k-mer feature extraction, we obtain high-dimensional vectors that reflect the properties of LncRNA sequences. The next step is to construct a graph structure to convert this high-dimensional data into a representation of LncRNA similarity or connection. In this graph, each node represents an LncRNA, and edges connect LncRNA nodes that are similar or functionally related.

The process of constructing the graph is as follows:

Node Definition: Each node in the graph represents an LncRNA sequence, and the node's features are represented by the k-mer vector of that LncRNA.

Edge Construction: The establishment of edges is based on the similarity of features between nodes, which can be calculated using cosine similarity. The edges between nodes directly reflect the degree of similarity between two nodes. In this experiment, we consider the top m nodes with the highest cosine similarity to each node as neighbor nodes.

Through this approach, we not only extract the base-level associations between LncRNAs but also capture the complex network among them. This provides an information-rich platform for predictive models based on graph neural networks to identify the subcellular localization of LncRNAs.

2.4. Weighted Graph Attention Network

In this study, we designed a Weighted Graph Attention Network (R-GAT) to more effectively address the issue of class imbalance in LncRNA subcellular localization problems. By introducing attention mechanisms, our network can assign different importance weights to each node during the training process, combined with weighted losses to guide the model to focus more on nodes belonging to the minority classes.

Attention Mechanism

Graph Attention Networks are an effective architecture of graph neural networks that can adaptively learn the weights between nodes[2]. The attention mechanism calculates the attention weights between nodes based on the features of each node and its neighbors. This dynamically highlights important nodes and suppresses less important nodes, thereby improving the model learning process.

Weighted Loss

In R-GAT, we have enhanced traditional attention mechanisms to address the issue of class imbalance among nodes. Specifically, we modulate the attention weights to reflect the distribution of classes.

Weight Calculation: To alleviate data imbalance issues, class-balanced weights are calculated inversely proportional to the effective number of samples, connecting each sample with a small neighborhood region. By combining the optimal balancing loss from the uneven margins of the imbalanced dataset, and minimizing the generalized boundary based on margins, a larger margin is provided for minority classes[3]. In this experiment, compared to existing inverse class frequency weighting methods, this approach demonstrates better performance.

$$\text{Weight} = \frac{1-\partial}{1-\partial^{\mathrm{n_c}}}$$

Representing the number of samples for class c, hyperparameters, in most experiments are typically set to [0.9, 1) for optimal results, as in this experiment.

Attention Layer: We designed multiple weighted attention layers that refine the representation of nodes. These layers learn complex node representations by adaptively reorganizing node features and information from neighbors.Adjusting Attention Mechanism，In R-GAT, the attention weight of each node towards its neighbors is defined by the following formula:

$$\theta_{ij} = \frac{exp\left(LeakyReLU\left(\delta\left(Wx_i, Wx_j\right)\right)\right)}{\sum_{k\in v_i} exp\left(LeakyReLU\left(\delta\left(Wx_i, Wx_j\right)\right)\right)}$$

In the above formula,$\mathrm{k} \in \mathrm{v_i}$represents the neighboring nodes of i node, the weight matrix $\mathrm{W} \in \mathrm{R}^{\mathrm{F*F'}}$ implements a linear transformation from input features to output features,δ represents a learnable weight vector, with LeakyReLU acting as a non-linear activation function.

Through the above design, R-GAT can maintain the flexibility and strong representational power of graph attention networks while mitigating information loss and overfitting issues during training caused by class imbalance. This design not only helps improve predictive performance for minority classes but also aims to enhance the overall model's generalization ability, thereby achieving more accurate results in predicting the subcellular localization of LncRNAs.

3. Experimental Results

3.1. Evaluation Metrics

To facilitate comparison with other methods, the models will be evaluated using a 5-fold cross-validation approach to assess accuracy, recall, and three other metrics. Accuracy (Acc) is computed by dividing the number of correctly classified samples by the total number of samples, and it serves as an intuitive performance metric. Recall measures the extent of coverage, quantifying how many positive instances are correctly identified as positive.

$$\mathrm{Recall}^{(\mathrm{i})} = \frac{\mathrm{TP}^{(\mathrm{i})}}{\mathrm{TP}^{(\mathrm{i})} + \mathrm{FN}^{(\mathrm{i})}} \qquad \mathrm{Recall} = \frac{1}{\mathrm{c}} \sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{Recall}^{(\mathrm{i})}$$

The F1 Score is an index used in statistics to measure the accuracy of binary classification models. It takes into account both the precision and recall of the classification model. The F1 Score can be seen as a weighted average, or harmonic mean, of the model's precision and recall, where its maximum value is 1, and the minimum is 0:

$$\mathrm{Precision}^{(i)} = \frac{\mathrm{TP}^{(\mathrm{i})}}{\mathrm{TP}^{(\mathrm{i})} + \mathrm{FP}^{(\mathrm{i})}}$$

$$\mathrm{F1} = \frac{1}{\mathrm{c}} \sum_{\mathrm{i}=1}^{\mathrm{c}} \frac{2 \times \mathrm{Precision}^{(i)} \times \mathrm{Recall}^{(\mathrm{i})}}{\mathrm{Precision}^{(i)} + \mathrm{Recall}^{(\mathrm{i})}}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives.

3.2. Proportion of Same-Class Samples in Neighbor Nodes in Graph Structure

In graph neural networks, experiments show that graphs with high homogeneity are more conducive to performance, especially when there are many connections between samples of the same class, indicating a high proportion of same-class samples among neighboring nodes. Experimental results suggest that the performance is better when (m = 20), meaning each node has 20 neighbor nodes. Statistical analysis reveals that when (m = 20), the proportion of nodes in the 20 neighboring nodes that belong to the same class as the source node is relatively high, as shown in Table 2.

Table 2: Proportion of Same-Class Samples

K-mer	*Optimal Feature Number (f)*	*Number of Edges (m)*	*Proportion (%)*
6-mer	*2000*	*20*	*64.51*
		30	*61.44*
		40	*58*
	2500	*20*	*63.86*
		30	*59.54*
		40	*56.54*
	3000	*20*	*64.68*
		30	*61.03*
		40	*56.04*

3.3. Results Comparison and Analysis

In the experiments targeting 5 cellular substructures, we evaluated the R-GAT model using 5-fold cross-validation method, and the results showed that the model performed outstandingly in predicting LncRNA subcellular localization. From the experimental data in Table 3, it can be seen that the model achieved satisfactory performance: the F1 score reached 0.851, indicating a good balance between precision and recall; the recall rate was as high as 0.859, demonstrating the model successfully identified 85.9% of positive samples; while the accuracy reached 91.3%, indicating the model's robust performance in overall classification tasks. Overall, the R-GAT model exhibited excellent classification and generalization capabilities in the LncRNA subcellular localization classification task, providing directions for further model optimization in the future.

Table 3: Dataset with 5 subcellular compartments

	F1	*Recall*	*Acc(%)*
R-GAT	*0.851*	*0.859*	*91.3*

4. Results and Discussion

Existing research models for the 5-classification of LncRNA subcellular localization include lncLocator[4] and DeepLncLoc[5]. As shown in Table 4, comparing based on the three metrics of F1, Recall, and Acc, the R-GAT model outperforms in all three indicators. The R-GAT model, based on graph neural networks, demonstrated significant superiority in the task of lncRNA sequence classification.

Compared to traditional methods based on sequence feature extraction and optimized selection, R-GAT extracts advanced features directly from the optimal features of lncRNA sequences to achieve more precise classification. The advantage of R-GAT is not only reflected in its capability; through learning and integrating rich sequence features using graph neural networks, it also leverages the advantage of neighboring node information, especially the enhancement of same-class sample information.

In the experiments, R-GAT enhances connections between same-class samples to focus more on learning differences and commonalities from similar lncRNAs. This mechanism is manifested in the tight connections of same-class nodes in the graph structure, facilitating the flow and integration of highly relevant feature information. Additionally, R-GAT adopts a weighted loss strategy to address the issue of data imbalance, a common challenge in bioinformatics. By assigning higher loss weights to minority classes, the model can treat all classes more fairly during the learning process, thereby improving overall performance and sensitivity to minority classes.

In summary, the advanced feature extraction capability of the R-GAT model, along with the enhanced utilization of same-class sample information and the weighted loss method for addressing data imbalance, collectively contribute to its superior performance in the precise classification of lncRNA sequences. Particularly in key metrics such as F1 score, recall, and precision, the R-GAT model significantly surpasses other existing classification methods. This performance improvement provides insights for its application in complex biological problems.

Table 4: Comparison with the existing predictor(5 subcellular compartments).

Method	*F1*	*Recall*	*Acc(%)*
lncLocator[4]	*0.367*	*0.363*	*59.1*
DeepLncLoc[5]	*0.563*	*0.524*	*53.7*
R-GAT	*0.851*	*0.859*	*91.3*

5. Conclusion

The R-GAT model exhibits significant advantages and outstanding performance in the task of lncRNA subcellular localization. The model uses graph neural networks to extract underlying features for accurate classification; it also capitalizes on the information from neighboring nodes within the graph structure to reinforce connections among samples within the same class, aiding the identification of commonalities and variances among similar lncRNAs. Moreover, a weighted loss strategy is applied to tackle issues of data imbalance, increasing the model's sensitivity to minority classes and thereby enhancing overall classification effectiveness. Experimental results show that the R-GAT model significantly outperforms traditional methods on critical metrics, offering an effective solution for predicting lncRNA subcellular localization, with broad prospects for application.

Acknowledgements

Yunnan Normal University Graduate Research Innovation Fund (YJSJJ23- B173)

References

[1] T. Zhang, P.W. Tan, L.Q. Wang, N.N. Jin, Y.N. Li, L. Zhang, H. Yang, Z.Y. Hu, L. N. Zhang, C.Y. Hu, C.H. Li, K. Qian, C.J. Zhang, Y. Huang, K.N. Li, H. Lin, D. Wang, RNALocate: a resource for RNA subcellular localizations, Nucleic Acids Res. 45 (2017) 135–138.
[2] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks[J]. stat, 2017, 1050: 20.
[3] Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9268-9277.
[4] Cao Z, Pan X, Yang Y, et al. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier[J]. Bioinformatics, 2018, 34(13): 2185-2194.
[5] Zeng M, Wu Y, Lu C, et al. DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding[J]. Briefings in Bioinformatics, 2022, 23(1): bbab360.