Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070908
ISSN 2371-8838 Vol. 7 Num. 9

A review of machine learning-based prediction of lncRNA subcellular localization

Xi Deng[1,a], Lin Liu[1,b,*]

[1]Yunnan Normal University, Kunming, China
[a]18213835546@163.com, [b]liulinrachel@163.com
[*]Corresponding author

Keywords: Long non-coding RNA (lncRNA), subcellular localization, machine learning, prediction model, bioinformatics

Abstract: With the continuous development of the field of bioinformatics, the subcellular localization of long non-coding RNA (lncRNA) has become a highly prominent frontier. LncRNAs play crucial regulatory roles in cellular processes, and understanding their subcellular localization is essential for comprehending their functions and mechanisms. However, traditional experimental methods face challenges of high costs and time consumption when predicting the subcellular localization of lncRNAs on a large scale, which has led to the emergence of research methods based on machine learning. This review aims to recap the latest advancements and trends in machine learning-based prediction of lncRNA subcellular localization in recent years. It not only provides new opportunities for a better understanding of lncRNA functions and cellular processes but also propels advancements in the fields of bioinformatics and molecular biology.

1. Introduction

Long non-coding RNAs (LncRNAs) are non-coding RNAs with a length exceeding 200 nucleotides. Human LncRNAs participate in a wide range of biological processes, including epigenetics, nuclear transport, alternative splicing, RNA decay, and translation. LncRNAs can also serve as precursors for small RNAs, and therefore, the aberrant expression of LncRNAs can lead to various human diseases and disorders. With the increasing reports of dysregulated LncRNA expression in many types of cancer, it also suggests that LncRNAs may function as potential tumor-suppressive RNAs[1]. Exploring the cellular functions of LncRNAs has become a central task in the post-genomic era.

A significant body of research indicates that the subcellular localization of lncRNAs has profound biological significance and exerts a significant impact on cellular processes and molecular mechanisms. The subcellular localization of lncRNAs can influence their interactions with chromatin, thereby regulating the expression of specific genes. For instance, lncRNAs located in the cell nucleus can interact with target genes in the genome, modulating gene transcription and expression levels. Furthermore, the subcellular localization of lncRNAs plays a crucial role in cell differentiation. For example, some nuclear lncRNAs can regulate the expression of key genes during stem cell differentiation, thereby influencing the ultimate differentiation state of cells.

Additionally, certain cytoplasmic lncRNAs can regulate cell growth and division by interacting with cell cycle-related proteins.

On the other hand, the subcellular localization of lncRNAs is closely related to the mechanisms of various diseases. For instance, nuclear lncRNAs may interact with tumor-related genes, participating in the initiation and progression of cancer. In the context of drug development and therapy, understanding the subcellular localization of specific lncRNAs associated with diseases can provide important insights into targeted therapies and drug development. By intervening with cytoplasmic or nuclear lncRNAs, it is possible to modulate signaling pathways associated with diseases. Some lncRNAs in the cytoplasm can regulate cell signaling pathways, such as serving as "sponges" for miRNAs to modulate protein expression. The subcellular localization of lncRNAs can also influence the intracellular subcellular environment, such as cytoplasmic transport and protein localization, which are crucial for maintaining normal cell function.

In summary, the subcellular localization of lncRNAs plays a pivotal role at multiple biological levels, holding profound biological significance for our understanding of cellular processes and disease mechanisms, as well as the development of new therapeutic strategies.

In the past, the subcellular localization prediction of lncRNAs in biology often relied on experimental techniques and cell biology methods. These methods included isolating different subcellular components of cells, using immunocytochemistry techniques and antibody labeling to detect and locate lncRNA molecules, as well as fluorescent labeling, among others. While these experimental methods are direct and accurate, they require more experimental work and resources.

Compared to traditional biological experimental methods, machine learning techniques have found widespread application in the field of bioinformatics, particularly in handling large-scale biological data. Machine learning methods excel in various aspects such as feature learning, classification prediction, generalization capability, and model interpretability. They assist biologists in more efficiently processing and analyzing biological data, thus enhancing the efficiency of biological research. The subcellular localization of LncRNAs, based on computable models, fundamentally belongs to a classification problem in biological data. The process of solving this problem mainly involves four aspects: the construction of data (sample) sets, the extraction and representation of LncRNA-related features, the design of classification models, and performance evaluation.

While machine learning methods have achieved success in the prediction of lncRNA subcellular localization, there are still some challenges to be addressed. These challenges include dealing with data imbalance, enhancing model generalization, ensuring data quality, and improving model interpretability. Addressing these challenges will contribute to further enhancing the performance and applicability of lncRNA subcellular localization prediction models.

2. Database Introduction

Databases related to the subcellular localization of lncRNAs are essential resources in biological research. They are used to collect, organize, and provide information on the localization of lncRNAs within different subcellular structures. These databases contain a wealth of experimental data and analysis results, offering researchers valuable insights into the subcellular localization of lncRNAs. However, there is currently a relatively limited number of databases specifically dedicated to lncRNA subcellular location information.

2.1. lncLocator

lncLocator1.0 is an online database and the first tool dedicated to predicting the subcellular localization of lncRNAs. It offers browsing, querying, and analysis interfaces for subcellular

location-related information for 10 types of RNA, including CircRNA, CsRNA, LncRNA, MiRNA, PiRNA, rRNA, SnoRNA, SnRNA, tRNA, and mRNA. It is currently one of the most popular tools for lncRNA localization prediction and serves as a primary source of data for numerous experiments in lncRNA subcellular localization prediction. The database currently contains information on the localization of 9,587 lncRNAs, with 6,636 originating from human lncRNAs. It is evident that the number of lncRNAs in different subcellular locations exhibits a highly imbalanced distribution, particularly after applying data preprocessing methods that result in a significant reduction in the number of lncRNA samples for each location label.

lncLocator 2.0 is an updated version of lncLocator 1.0 and includes the option to select cell lines. As research has shown, the subcellular localization of lncRNAs is closely related to the type of cells and tissues[2]. For example, the gene LINC00476 exhibits varying subcellular localization results in 12 different cell lines. In six of these cell lines, the lncRNA is located in the cytoplasm, while in the other six cell lines, it is located in the nucleus. Therefore, the addition of a cell line selection option is essential.

2.2. lncATLAS

lncATLAS contains subcellular localization data of lncRNAs in various cell lines. It currently includes over 6,700 lncRNAs annotated by GENCODE and covers localization information in 15 different human cell lines. The database offers comprehensive lncRNA annotation information, including lncRNA names, gene structures, sequence details, and functional annotations related to lncRNAs. It also provides a graphical query interface, aiding researchers in understanding the biological functions and pathways in which specific lncRNAs are involved.

2.3. lncSLdb

lncSLdb(Long Non-Coding RNA Subcellular Localization Database) is a comprehensive website for lncRNA subcellular localization. It compiles subcellular localization information for over 11,000 lncRNA transcripts or genes from three species: human, mouse, and Drosophila. The database integrates a wealth of experimental data and literature information, including data from RNA localization experiments, RNA-seq data, cell fractionation techniques, and relevant reports in the literature. The website also provides convenient browsing and search options.

3. Feature extraction

Traditional machine learning methods for classification tasks focus more on feature extraction. For the subcellular localization of LncRNA, various classic RNA sequence feature extraction methods or sequence-related databases can be extensively utilized. These include nucleotide composition (k-mer), pseudo-nucleotide composition (PseKNC, PseDNC), three-dimensional reading frames, conserved motif information, triple methods for secondary structure, and topological secondary structure parameter methods, most of which are based on sequence-based feature extraction.

K-mer is a widely used feature extraction and sequence analysis technique in bioinformatics, primarily applied to DNA, RNA, or protein sequences. It is commonly employed in genomics, transcriptomics, proteomics, and related fields. K-mer refers to a continuous subsequence segment of length K, typically a part of DNA, RNA, or protein sequences. For instance, in a DNA sequence, a 2-mer represents a two-nucleotide segment, and a 3-mer represents a three-nucleotide segment, and so on.

PseKNC is a feature extraction method constructed using K-tuple Nucleotide Composition,

where K represents combinations of K consecutive nucleotides in the sequence. It introduces "pseudo KNC" features, enhancing the expressiveness of features by mathematically transforming and adjusting the weights of nucleotide combination frequencies.

PseDNC is a feature extraction method built using Dinucleotide Composition (frequency of dinucleotide combinations). Similar to PseKNC, PseDNC introduces "pseudo DNC" features. It constructs a feature vector by calculating the frequencies of different dinucleotide combinations and applying weight functions and mathematical transformations. This feature vector includes information about different dinucleotide combinations in the sequence and their weighted information.

To date, in addition to the above three common methods, DeepLncRNA experiments have utilized relevant transcription annotation features, genome loci, and RNA binding motifs to express sequence features. Locate-R experiments employ a variety of nucleotide composition features to help differentiate lncRNAs in different subcellular locations. lncLocPred experiments use a combination of K-mer, PseDNC, and Triplet features for model training. RNAlight[3] experiments, in addition to K-mer, PseKNC, and RNA secondary structure features, also incorporate RNA functional features, protein interactions, and other biological features. LightGBM-LncLoc[4] experiments have also adopted the RCKmer method, treating two complementary k-mer nucleotides as the same k-mers. These sequence-based multi-perspective feature extraction methods help the model understand the feature composition and structure of lncRNA sequences more accurately, thereby determining their subcellular localization more precisely.

4. lncRNA Subcellular Localization Prediction

4.1. Algorithm Design for lncRNA Subcellular Localization Prediction

In the initial stages, researchers primarily employed traditional machine learning methods based on feature engineering. These methods relied on feature selection strategies, such as K-mer frequencies and combinations of RNA structural features. They used classifiers such as Support Vector Machines and Random Forests for predicting lncRNA subcellular localization. These methods achieved some success but were reliant on feature engineering and couldn't fully capture the complex information in sequences.

With the development of deep learning technology, it has gained widespread applications not only in areas like image processing, speech recognition, natural language processing, and big data feature extraction but also in computational biology, such as protein structure prediction and genome editing. Deep learning methods can automatically capture high-level features in data. In the study by Fan et al., deep neural networks were employed for predicting subcellular localization in both the nucleus and cytoplasm. Wang et al. proposed the ensemble model IDDLncLoc based on convolutional neural networks and Support Vector Machines. Experiments conducted in GM-lncLoc and GraphLncLoc [5] transformed lncRNA sequences into graph structures, where nodes represented sequence features and edges represented relationships between them. Graph Convolutional Networks (GCN) or Graph Neural Networks (GNN) were used to learn more representative feature representations, enhancing prediction performance.

4.2. Challenges in lncRNA Subcellular Localization Prediction Algorithm

From the current state of research, the field of lncRNA subcellular localization faces the following challenges:

Data Quality and Label Issues: Data quality is crucial for training and testing lncRNA subcellular localization models. However, existing label information may suffer from inaccuracies

and noise, negatively impacting model performance.

Sample Imbalance: The number of lncRNAs in different subcellular locations can vary significantly, leading to class imbalance issues. This may cause models to be biased toward majority classes, neglecting minority classes, thereby reducing the accuracy of predictions for minority subcellular locations.

Integration of Multi-modal Information: lncRNA subcellular localization is influenced by various factors, including sequence information and lncRNA-protein interactions. Effectively integrating this multi-modal information and determining the contributions of different types of information is a complex challenge.

Model Interpretability: Complex models such as deep learning and graph neural networks often lack interpretability, making it difficult to understand the rationale behind specific predictions. For biologists, understanding the underlying principles of model predictions is crucial.

Cell-Specific Analysis: The same lncRNA can exhibit different localization patterns in various cell lines. The transfer of models from one cell line to another or cross-cell line prediction poses a challenging problem.

Scarcity of Datasets: The limited availability of effective training samples makes it challenging to establish high-quality training datasets, restricting model performance.

Multi-localization Prediction: As research has revealed, a single lncRNA may exist in multiple subcellular locations within a cell. Labels for different subcellular locations may overlap or have interdependencies, making accurate multi-label prediction more challenging.

Currently, researchers are making continuous efforts to address these challenges by adopting advanced machine learning and deep learning techniques, designing more complex model architectures, integrating information from various sources, and dealing with data issues. These efforts aim to improve the accuracy and applicability of lncRNA subcellular localization prediction. As the field continues to evolve, we can expect these challenges to be better resolved.

5. Conclusion

In this comprehensive review, we delved into the field of lncRNA subcellular localization, along with its associated methods and challenges. We learned that current research not only emphasizes the accuracy of prediction models but also focuses on improving data quality and label reliability, addressing sample imbalance, integrating multi-modal information, enhancing model interpretability, achieving cell-specific analysis, and tackling challenges like multi-localization prediction. These efforts are expected to provide more robust tools and methods for the study of lncRNA subcellular localization, potentially impacting fields such as cancer therapy, disease diagnosis, and fundamental biological research significantly.

With the continuous advancement of technology and deeper research, the field of lncRNA subcellular localization offers numerous opportunities yet to be explored. In the future, we can anticipate more interdisciplinary collaborations, high-quality data resources, and innovative methods to emerge, providing a more comprehensive understanding of the precise localization and function of lncRNA within cells. In summary, the field of lncRNA subcellular localization research holds immense potential, offering a solid foundation for deepening our understanding and application of lncRNA, making valuable contributions to the further development of the biomedical field.

Acknowledgements

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

References

[1] Atianand, M. K., Hu, W., Satpathy, A. T., Shen, Y., Ricci, E. P., Alvarez-Dominguez, J. R., ... & Fitzgerald, K. A. (2016). A long noncoding RNA lincRNA-EPS acts as a transcriptional brake to restrain inflammation. Cell, 165(7), 1672-1685.

[2] Lin, Y., Pan, X., & Shen, H. B. (2021). lncLocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding RNAs with interpretable deep learning. Bioinformatics, 37(16), 2308-2316.

[3] Yuan, G. H., Wang, Y., Wang, G. Z., & Yang, L. (2023). RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. Briefings in Bioinformatics, 24(1), bbac509.

[4] Lyu, J., Zheng, P., Qi, Y., & Huang, G. (2023). LightGBM-LncLoc: A LightGBM-Based Computational Predictor for Recognizing Long Non-Coding RNA Subcellular Localization. Mathematics, 11(3), 602.

[5] Li, M., Zhao, B., Yin, R., Lu, C., Guo, F., & Zeng, M. (2023). GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. Briefings in Bioinformatics, 24(1), bbac565.