Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080105
ISSN 2371-8838 Vol. 8 Num. 1

A review of computational model-based prediction of lncRNA subcellular localization

Rongneng Sun[1,a], Qingsong Guo[1,b,*], Xiaofen Yuan[1,c]

[1]*Yunnan Normal University, Kunming, China*
[a]*2546755854@qq.com*, [b]*3080790461@qq.com*, [c]*1528264906@qq.com*
[*]*Corresponding author*

Keywords: Long non-coding RNAs (lncRNA), subcellular localization, database, computational model

Abstract: Long non-coding RNAs (lncRNA) play pivotal roles in diverse cellular processes, and the determination of lncRNA subcellular localization serves as crucial information for elucidating their functional roles. However, conventional biochemical experimental methodologies employed for identifying lncRNA subcellular localization exhibit inherent complexities, challenges in reproducibility, and substantial costs. In the contemporary era of burgeoning bioinformatics, computational models for predicting the subcellular localization of biomolecules offer a viable alternative. Notably, these computational approaches boast high efficiency and relatively lower costs, presenting a substantial reduction in time and human resource expenditure compared to traditional experimental protocols. This comprehensive review encapsulates the latest strides in leveraging computational models for the prediction of lncRNA subcellular localization, offering novel avenues for a profound comprehension of lncRNA functionality and their intricate involvement in cellular processes.

1. Introduction

Long-stranded non-coding RNAs (lncRNAs) are a class of RNA molecules that are more than 200 nucleotides in length and have no or very low protein-coding ability[1]. Originally thought to be a transcriptional by-product, lncRNA has now been revealed to play a key role in cellular processes such as RNA regulation as research progresses. It is involved in a variety of RNA regulation and other processes such as cell cycle regulation, epigenetic regulation, gene expression regulation, transcription, splicing, mRNA decay and translation[2-4]. Regarding the location of lncRNA, they are mainly concentrated in the nucleus and cytoplasm, and even some of them exist in both nucleus and cytoplasm at the same time[5]. lncRNA in the nucleus can serve as molecular scaffolds and have the functions of assisting variable splicing and regulating chromosome structure [6-8]. In the cytoplasm, they may be involved in translation or signaling, and promote or inhibit mRNA degradation[9-11], suggesting that the biological functions of lncRNAs are closely related to their specific intracellular localizations, and thus understanding their subcellular locations has become particularly critical. This paper reviews lncRNA subcellular localization prediction research, focusing on the commonly used databases for lncRNA and the current advances in computational

modeling for lncRNA research.

2. Introduction to the database

The lncRNA databases provide comprehensive and reliable data resources, effectively simplifying research work in related fields and significantly enhancing research efficiency, the current commonly used lncRNA subcellular localization databases are as follows:

2.1. RNALocate

RNALocate is an online database focusing on the annotation of RNA subcellular locations[12],

It provides detailed labelling of multiple subcellular locations, which enables users to obtain targeted annotations of RNA subcellular locations of interest. Users can easily obtain the desired subcellular location information through an intuitive query interface by entering the name, sequence, or other relevant information of the RNA. RNALocate is also interlinked with other bioinformatics databases to provide users with more relevant information about RNA. In addition, RNALocate is regularly updated to maintain the timeliness of the database, ensuring that users are provided with the latest RNA subcellular location annotation information, which helps to promote in-depth research on RNA function and cell biology.

2.2. LncATLAS

The LncATLAS database[13] covers more than 6,700 lncRNAs annotated by GENCODE, including subcellular localization data in 15 different human cell lines. The database provides exhaustive lncRNA annotation information, including their names, gene structures, sequence details, and functional annotations related to lncRNA. To facilitate researchers' in-depth understanding of specific lncRNA, LncATLAS also provides an intuitive graphical query interface to help reveal the biological functions and pathways in which these lncRNA are involved. This enables researchers to make more comprehensive use of this database to dig deeper into the subcellular localization of lncRNA and their biological roles in cells.

2.3. LncSLdb

The LncSLdb database[14] was released in 2018 and describes in detail subcellular localization data from three species, including over 11,000 non-coding transcripts. These transcripts were categorised into three basic types of localization, i.e. nuclear, cytoplasmic and nuclear/cytoplasmic. The subcellular localization data were collected using two complementary approaches. Firstly, PubMed was searched to obtain 3000 papers containing the keywords "lncRNA" and "subcellular localization", and the number of papers focusing on the subcellular localization of lncRNA was reduced to 100 after manual screening. Then, we collected gene information and localization data of lncRNA from several databases, such as UCSC, Ensembl, MGD. Through these integrated approaches, LncSLdb offers comprehensive information on the subcellular localization of non-coding transcripts, providing researchers with a valuable tool to gain insight into the localization of lncRNA in cells and their potential biological functions.

2.4. lncRNAdb

lncRNAdb is a comprehensive database of lncRNAs[15], which provides researchers with a large amount of information on lncRNA from humans and other species. The database not only

contains detailed annotations of lncRNA genes, but also covers their genomic environments, expression profiles, and potential functions. lncRNAdb is unique in integrating experimental evidence to aid in the functional annotation of lncRNAs. Its user-friendly interface facilitates access to a wealth of data including links to the relevant literature, and the database's regular updating ensures that it remains a relevant and up-to-date repository for lncRNA research.

2.5. LNCipedia

LNCipedia is a wide-ranging public database[16] that assembles a large collection of rigorously selected lncRNA sequences and annotation information. It details gene structure, transcript length, genomic location, and the relationship to known protein-coding genes. It was last updated in August 2018 to version 5.2, containing approximately 127,802 records and 56,946 genes. The intuitive user interface and efficient search function make LNCipedia a great resource for researching and learning about lncRNA.

2.6. NCBI

The NCBI database[17] is an internationally recognised repository of biomedical information, providing researchers with a large amount of life science data. Its distinguishing feature is the integration of a wide range of biological datasets, including genomic information, protein structures and medical research literature. NCBI not only supports the retrieval of data, but also provides a series of bioinformatics analysis tools for researchers to use, supporting researchers to perform complex calculations such as gene sequence comparison and protein structure analysis. As an important resource in the field of global bioinformatics, NCBI continuously updates its database content, ensuring researchers have access to the latest and most comprehensive biological data and research tools.

3. Current research status

The currently proposed computational models for predicting subcellular localization of lncRNA are mainly divided into two major directions: machine learning and deep learning. Machine learning models utilize algorithmic approaches to analyze patterns in data, while deep learning models employ neural networks to automatically and adaptively learn data representations.

3.1. Machine learning-based subcellular localization of lncRNA

ZD Su et al. proposed a predictor of lncRNA subcellular localization called iLoc-lncRNA[18]. In this study, using the data provided by the RNALocate website, among 923 sequences of lncRNAs, the sequences with greater than 80% similarity were removed by CD-Hit filtering, and finally 655 sequences were obtained, which included 156 nuclear, 426 cytoplasmic, 43 ribosomal, and 30 exosome localization. The method of combining 8-mer components with PseKNC components as features was adopted in the study, and the optimal subset of features was established using the IFS strategy, and finally the support vector machine (SVM) method was applied for prediction. This predictor achieved an overall accuracy of 86.72% on the benchmark dataset through 5-fold cross validation, but its prediction performance for the minority samples exosome and ribosome was very poor.

Ahsan Ahmad et al. proposed a method called Locate-R[19], which is based on fusion features consisting of n-gapped l-mer and l-mer. This approach involves feeding the features into a local depth SVM for training, ultimately improving the overall accuracy to 90.69%. The prediction of

Locate-R on exosome and ribosome has a considerable improvement, but the accuracy on nucleus is only 65.92%, which needs to be further improved.

Yang et al. proposed a method for subcellular localization of lncRNA based on non-equilibrium pseudo k-nucleotide components[20]. In order to utilize the sequence information more comprehensively, they considered both the k-mer nucleotide composition of lncRNA and the sequence order correlation factors, and finally utilized the support vector machine to perform the prediction analysis. The method achieved 90.37% accuracy by leave-one-out cross-validation.

Fan Y et al[21] proposed a logistic regression based classifier, lncLocPred. In terms of features, they chose k-mer, triplet, and PseDNC sequences, and then utilized variance thresholding, binomial distribution, and F-score for the feature selection of the system. Finally, lncLocPred achieved an overall accuracy of 92.37% in jackknife test.

In order to identify the subcellular multi-localization of lncRNA, Yan introduced the k-mer nucleotide composition and sequence order correlation factors as the feature vectors of lncRNA [22]. To remove redundant information and improve model prediction, the authors used analysis of variance (ANOVA) to screen out the optimal subset of features, and based on the support vector machine algorithm to predict the subcellular multi-localization of lncRNA. The model was evaluated by a 5-fold cross-test. The results show that the predicted location coverage of the benchmark dataset and the independent dataset reach 87.22% and 71.56%, respectively.

Sun proposed a lncRNA subcellular localization prediction algorithm based on a multi-feature fusion algorithm[23], which uses nucleotide sequence features extracted from multiple perspectives , these include the traditional sequence assembly algorithm K-mer, subsequence-based takes into account the coding features of contextual information , biological structural properties of nucleotide sequences, and pseudo-dinucleotide composition of nucleotide sequences Pse-DNC. It employs a feature fusion algorithm, which assigns weights to each feature and fuses them accordingly, and the fused features are downscaled to obtain the best subset of features as input to the machine learning classifier. Finally, the best classifier is selected among multiple multi-label classifiers, with a 4.1 % increase in AP compared to the existing tools.

3.2. Deep learning-based subcellular localization of lncRNA

Since 2016, various deep learning algorithms based on neural networks have been gradually promoted and applied in various fields, and all of them can achieve better performance, which makes bioinformaticians become more and more serious in their attention to deep models.

Cao et al. conducted a study of lncRNA subcellular localization, firstly developing a predictor called lncLocator[24]. They constructed a standard dataset containing five subcellular locations based on the RNALocate database, and from a total of 1,361 sequences of lncRNAs, 612 sequences of lncRNAs were screened after screening conditions such as the number of datasets and sequence similarity, including 152 nuclear localizations, 301 cytoplasmic localizations,91 cytoplasmic matrix localizations, 43 ribosome-localised and 25 exosome-localised sequences. By extracting the raw k-mer frequency features and using supervised oversampling techniques to balance different categories of samples, the research team established the RF^R, SVM^R, RF^A and SVM^A four prediction models, while the lncLocator model was constructed based on the integration technique of neural network. Finally, the overall prediction success rate of 59.1% was obtained through 5-fold cross-validation.

Gudenas BL et al. developed DeepLncRNA[25], a deep learning algorithm that predicts the subcellular localization of lncRNA directly from lncRNA transcript sequences. The model was constructed using 93 samples of strand-specific RNA sequences from both nuclear and cytoplasmic fractions of multiple cell types. DeepLncRNA, a feed-forward, multi-layer deep neural network,

achieved an accuracy of 72.4%, a specificity of 62.4%, and a sensitivity of 83%. While the prediction results showed improvement, they were still not entirely satisfactory.

Yang Lin et al. proposed an updated cell lineage-specific prediction method, lncLocator 2.0[26], which can train an end-to-end depth model for each cell line to predict lncRNA subcellular localization from sequences. They first constructed a benchmark dataset of lncRNA subcellular localization for 15 cell lines. Then classification of lncRNA subcellular localization was performed by learning word embeddings and applying them to convolutional neural networks, long and short-term memory networks and multilayer perceptrons.

Zhu, X proposed a new model called IDDLncLoc[27]. In IDDLncLoc, an integrated framework of sequence features is introduced, after which three features are used to describe lncRNA sequences, and feature selection and recursive feature elimination are systematically processed by binomial distribution to identify the optimal features, and a novel CNN network, AFCNN, is proposed to predict lncRNA subcellular locations. The model achieves a high accuracy of 94.96% on the benchmark dataset.

Li M et al. proposed a predictor called GraphLncLoc[28], which converts lncRNA sequences into graphs and utilizes a graph convolutional network to capture high-level features, and the high-level feature vectors are transported into the fully-connected layer to perform the prediction task, which achieves a better performance than traditional machine learning models and existing predictors.

GM-lncLoc[29], proposed by JZ Cai extracts initial information from lncRNA sequences based on K-mer and combines graph structure information to extract high-level features of lncRNAs. To solve the problem of limited number of samples in lncRNA subcellular localization, GM-lncLoc introduces a meta-learning training model. Accuracy rates of 93.4% and 94.2% were achieved on the benchmark datasets of 5 subcellular compartments and 4 subcellular compartments, respectively.

Zeng M et al. proposed LncLocFormer[30], a multi-label lncRNA subcellular localization predictor that uses eight Transformer blocks to model remote dependencies in lncRNA sequences, shares information among lncRNA sequences, and employs a localization-specific attention mechanism to discover different localization patterns for different subcellular localizations .

4. Conclusion

This paper presents an extensive analysis of computational model-based prediction studies of lncRNA subcellular localization. It first discusses the biological importance of lncRNAs and the criticality of subcellular localization for functional understanding, followed by a discussion of currently popular lncRNA databases, and then introduces the latest research progress in machine learning and deep learning based lncRNA subcellular localization prediction methods, and explores the performance of the various algorithms in practical applications. Through these comprehensive reviews, this paper provides a comprehensive perspective and future research directions for lncRNA research and subcellular localization prediction.

References

[1] Taft, R.J.; Pang, K.C.; Mercer, T.R.; Dinger, M.E.; Mattick, J.S. Non-coding RNAs: Regulators of disease.J. Pathol. 2010, 220, 126- 139.[PubMed]

[2] Martin KC, Ephrussi A. mRNA localization: gene expression in the spatial dimension[J]. Cell. 2009 Feb 20;136(4):719-30.

[3] Hu L, Hou D, Huang D, et al. Long-stranded non-coding RNAs and epigenetic regulation[J]. Genomics and Applied Biology, 2016,35(012):3319-3324.

[4] Gupta, R. A., Shah, N., Wang, K.C., Kim, J., Horlings, H. M., & Wong, D.J.: Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010, 464(7291):1071.

[5] Rashid F,Shah A,Shan G. Long Non-coding RNAs in the Cytoplasm[J].Genomics Proteomics & Bioinformatics, 2016,14(2):73-80.
[6] Chen Y G, Satpathy A T, Chang H Y. Generegulation in the immune system by longnoncoding RNAs[J]. Nature immunology, 2017, 18(9): 962-972.
[7] Chen M T,Lin H S,Shen C,et al. PU.1-Regulated LongNoncoding RNA lnc-MC Controls Human Monocyte/ Macrophage Differentiation through Interaction with MicroRNA 199a-5p[J].Molecular & Cellular Biology,2015,35(18):3212-3224.
[8] Tay Y, Kats L, Salmena L, et al. Coding-independent regulation of the tumour suppressor PTEN by competing endogenous mRNAs[J]. Cell, 2011, 147(2): 344-357.
[9] Yao R W, Wang Y, Chen L L. Cellular functions of long noncoding RNAs[J]. Nature cell biology, 2019, 21(5):542-551.
[10] Chen L L. Linking long noncoding RNA localization and function[J]. Trends in biochemical sciences, 2016, 41(9): 761-772.
[11] Wen X, Gao L, Guo X, et al. lncSLdb: aresource for long non-coding RNA subcellular localization[J]. Database, 2018.
[12] Zhang T,Tan P,Wang L,et al. RNALocate:a resource for RNA subcellular localizations[J].Nucleic acids research,2017,45(D1):D135-D138.
[13] Mas-Ponte D, Carlevaro-Fita J,Palumbo E,et al.LncATLAS database for subcellular localization of long noncoding RNAs [J]. Rna, 2017, 23(7):1080- 1087.
[14] Wen X, Gao L, Guo X, Li X, Huang X, Wang Y, Xu H, He R, Jia C, Liang F. lncSLdb: a resource for long non-coding RNA subcellular localization[J]. Database (Oxford). 2018 Jan 1;2018:1-6.
[15] Amaral P P,Clark M B,Gascoigne D K,et al.lncRNAdb:a reference database for long noncoding RNAs[J].Nucleic acids research,2011,39(suppl_1). D146-D151.
[16] Volders P J,Anckaert J,Verheggen K,et al.LNCipedia 5:towards a reference set of human long non-coding RNAs[J].Nucleic acids research,2019,47(D1):D135-D139.
[17] Geer L Y, Marchler-Bauer A, Geer R C, et al. The NCBI biosystems database[J]. Nucleic acids research, 2010, 38(suppl_1): D492-D496.
[18] Su Z D, Huang Y, Zhang Z Y,et al. iLoc-LncRNA:predict the subcellular location of LncRNAs by incorporeating octamer composition into general PseKNC [J]. Bioinformatics, 2018, 34(24):4196-4204.
[19] Ahmad A, Lin H, Shatabda S. Locate-R: Subcellular localization of long non-coding RNAs using nucleotide compositions [J]. Genomics, 2020, 112(3): 2583-2589.
[20] Yang X-F, Zhou Y-K, Zhang L, et al. Predicting LncRNA Subcellular Localization Using Unbalanced Pseudo-k Nucleotide Composition[J]. Current Bioinformatics. 2020; 15(6).
[21] Y. Fan, M. Chen and Q. Zhu. lncLocPred: Predicting LncRNA Subcellular Localization Using Multiple Sequence Feature Information[J]. IEEE Access, vol. 2020; 8(124702-124711).
[22] Yan D X,CHEN Yingli. Subcellular multi-localization prediction of long-chain non-coding RNAs based on sequence information[J]. Journal of Inner Mongolia University (Natural Science Edition),2022,53(01):38-47.
[23] Sun X H. Research on multi-feature fusion algorithm for long-stranded non-coding RNA subcellular localization prediction problem [D]. Jilin University, 2022.
[24] 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.
[25] GUDENAS B L,WANG L. Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features [J] Scientific Reports, 2018, 8(1):1- 10.
[26] Lin Y, Pan X. Hong-Bin Shen. lncLocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding RNAs with interpretable deep learning[J]. Bioinformatics. 2021; 37(16):2308-2316.
[27] Zhu X P. A subcellular localization method of lncRNA based on unbalanced data distribution framework [D]. Jilin University, 2022.
[28] Li M, Zhao B Y, Yin R, et al. GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation, Briefings in Bioinformatics, Volume 24, Issue 1, January 2023, bbac565.
[29] Cai J Z, Wang T.,Deng, X. et al. GM-lncLoc: LncRNAs subcellular localization prediction based on graph neural network with meta-learning. MC Genomics 24, 52 (2023).
[30] Zeng M, Wu Y F, Li Y M, et al. LncLocFormer: a Transformer-based deep learning model for multi-label lncRNA subcellular localization prediction by using localization-specific attention mechanism, Bioinformatics, Volume 39, Issue 12, December 2023, btad752.