

基于深度学习的长链非编码RNA与蛋白质相互作用关系预测的研究进展

赵靖轩

辽宁科技大学, 计算机与软件工程学院, 辽宁 鞍山

收稿日期: 2022年3月6日; 录用日期: 2022年4月5日; 发布日期: 2022年4月12日

摘 要

基因表现型受调控基因的表达来驱动,而基因的表达同时也受不同生物分子之间复杂的相互作用所影响。21世纪以来,蛋白质-DNA相互作用以及RNA-RNA相互作用都收获了很好的研究成果。相比起来,在受制于各种困难的情况下,长链非编码RNA-蛋白质相互作用(LPI)机制相对未知。然而LPI正在成为遗传机制中的关键相互作用,在人类发育过程中扮演着越来越重要的角色,因此人们对于长链非编码RNA与蛋白质相互作用关系研究的热情不断高涨。由于使用大规模实验方法来验证两者的相互作用需要耗费很多物质成本和时间成本,与此同时伴随着当代社会科学技术不断发展,计算机硬件持续迭代更新,研究人员开始将重心放在如何使用机器学习及深度学习方法来进行基础预测,可喜的是这些预测实验都取得了令人满意的结果。为了丰富该领域的科研论文资料,方便科研人员了解深度学习技术在长链非编码RNA与蛋白质相互作用关系预测的应用,在此我们回顾了长链非编码RNA与蛋白质相互作用预测的研究现状,调查了近期研究人员在实验中利用到的最新方法和数据库,并总结了相关数据,提出了一些可行性建议。

关键词

长链非编码RNA, 蛋白质, 相互作用, 数据库, 机器学习, 深度学习

Research Progress on Predicting Interaction between Long Noncoding RNA and Protein Based on Deep Learning

Jingxuan Zhao

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan Liaoning

Received: Mar. 6th, 2022; accepted: Apr. 5th, 2022; published: Apr. 12th, 2022

文章引用: 赵靖轩. 基于深度学习的长链非编码 RNA 与蛋白质相互作用关系预测的研究进展[J]. 计算机科学与应用, 2022, 12(4): 858-865. DOI: 10.12677/csa.2022.124087

Abstract

Gene phenotypes are driven by the expression of regulatory genes, which are also influenced by complex interactions between different biomolecules. Since the 21st century, protein-DNA interaction and RNA-RNA interaction have been well studied. In contrast, under many difficulties, the mechanism of long non-coding RNA-protein interaction (LPI) is relatively unknown. However, LPI is becoming a key interaction in the genetic mechanism and playing an increasingly important role in human development. Therefore, the research on the interaction between long non-coding RNAs and proteins is increasingly enthusiastic. Since the use of large-scale experimental methods to verify the interaction between the two requires a lot of material and time costs, at the same time, with the continuous development of contemporary social science and technology and the continuous iterative updating of computer hardware, researchers began to focus on how to use the machine learning and deep learning approach to predict. It is gratifying that these prediction experiments have achieved satisfactory results. In order to enrich scientific research papers in this field and facilitate researchers to understand the application of deep learning technology in predicting the interaction relationship between long non-coding RNA and protein, we reviewed the research status of long non-coding RNA and protein interaction prediction, investigated the latest methods and databases used by researchers in recent experiments, and summarized relevant data. Finally, some feasible suggestions are put forward.

Keywords

Long-Chain Noncoding RNA, Protein, Interaction, Database, Machine Learning, Deep Learning

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 介绍

转录组学是研究一个细胞内完整基因转录本的学科,可以测量不同条件下基因组的表达水平。现代转录组学通过用高通量测序技术来研究基因和其履行的生物职能,通常采用生物信息学方法,例如剪接位点识别、转录本变体识别或确定蛋白质编码转录子的替代启动子[1]。然而这些蛋白质编码转录本仅代表转录组的一小部分,一个大基因组里有相当多的RNA转录本是不会直接编码蛋白质的,研究证实在全部mRNA序列中外显子序列仅占2.94%,其余的被称为非编码RNA(ncRNA)[2],这些RNA的功能更难以表征。非编码RNA可以根据功能和长度分为多个类别,在此,我们着重介绍其中的长链非编码RNA(lncRNA),即长度大于200个核苷酸的非编码RNA。长非编码RNA发挥其生物学功能的一个重要作用机制即是与蛋白质进行相互作用,文章中,我们探讨了近期关于长非编码RNA与蛋白质相互作用关系的预测研究现状,对现有的利用深度学习预测长链非编码RNA与蛋白质相互作用关系模型进行了技术分析,另给出了模型的通用评价指标,便于读者对模型预测性能有一个更直观的了解。

2. 通用数据集

在2011年,Amaral等人构建了第一个长非编码RNA数据库,名为lncRNADB[3]。lncRNADB数据库提供长非编码RNA生物学功能的全面注释。随后,中国科学院生物物理研究所陈润生院士课题组通过

Pubmed 数据库收集整理非编码 RNA 相关的文章、补充材料以及其它资料,构建了 NONCODE 数据库[4]。该数据库是专门用于非编码 RNA (不包括 tRNA 和 rRNA)的综合知识数据库,它不仅记录了非编码 RNA 的序列信息、相关基因,还收集整理了非编码 RNA 在各组织器官的表达量等功能信息。由印度 CSIR 基因组和整合生物学研究所的研究人员开发的 lncRNome 是一个关于人类长非编码 RNA 的功能注释数据库[5]。该数据库拥有超过 17,000 个人类长非编码 RNA 的信息并提供有关长非编码 RNA 类型、染色体位置、生物功能和疾病关联的信息。

目前已知的主流长非编码 RNA 与蛋白质相互作用数据库包括 NPInter 数据库[6]和 RPI7317 数据库[7]等,其中 NPInter2.0 数据库中记录了经过实验验证的非编码 RNA (不包括 tRNA 和 rRNA)和生物分子(蛋白质、RNA 和 DNA)的相互作用关系信息,该数据库的信息主要来源于 NONCODE、miRBase 和 UniProt 等数据库,其中共包含 10,412 个经实验验证的 NPI,涉及来自六种不同生物的 4636 个非编码 RNA 和 449 个蛋白质。RPI7317 数据库是从 NPInter3.0 数据库中精选的,只从人类样本中选择交互。

3. 模型性能评估标准

在评估长链非编码 RNA 与蛋白质相互作用预测模型性能时,我们最多用到的评价标准有 ROC 曲线、AUC 值(ROC 曲线下面积)、ACC (准确率)、REC (召回率)、特异性(SPEC)、PRE (精确度)和 F1 值等。其中 ROC 曲线为受试者工作特性曲线,横坐标为 FPR (负正类率),纵坐标为 TPR (真正类率),它们分别由 TP、FP、TN、FN 计算而来,TP 代表实际为正类且预测为正类的样本,FP 代表实际为负类而预测为正类的样本,同理 TN 代表实际为负类且预测也为负类的样本,FN 代表实际为正类而预测为负类的样本。由公式不难得出,在实际情况中,AUC 值取值在(0, 1)之间,且 AUC 值越接近 1,则模型性能表现越好,当 AUC 取值为 0.5 时,模型性能等同于随机预测。全局留一交叉验证和 k 折交叉验证常用来测试模型在测试集上的表现。ACC 为准确率,表示预测正确的样本数在全部样本中的占比,且计算结果越高越好。REC 为召回率,又被称为灵敏性,为在判断为正类的样本中真实为正类的样本比率,类似的, SPEC 被称为特异性,为在判断为负类的样本中真实为负类的样本比率。PRE (PPV)为精确度,为在所有被预测为正类的样本中真实即为正类的比率。F1 值为 PRE 和 REC 的调和平均值。一般情况下,由上述指标值,我们可以对预测模型的性能有一个综合的评估。上述几项指标值的计算公式如下:

$$\begin{aligned} \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \\ \text{REC} &= \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{SPEC} &= \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{PRE} &= \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{F1} &= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} = 2 \cdot \frac{\text{PRE} \cdot \text{REC}}{\text{PRE} + \text{REC}} \end{aligned}$$

4. 预测模型

在 2020 年, Zhang 等人提出了一种基于卷积神经网络的并采用了复制填充技巧的深度学习模型,称为 LPI-CNNCP [8]。其中采用的复制填充技巧能够将非定长的蛋白质或 RNA 序列转换为定长序列,从而构建出 CNN 模型,并且采用了高阶 one-hot 编码来从序列信息中捕获到核苷酸或氨基酸之间的依赖关系,

转换为类似于图像形式的输入，最后将上面的处理过后的信息输入 CNN 中来预测长非编码 RNA 和蛋白质的相互作用数据。

卷积神经网络是一种有效的深度学习模型，能够从样本中学习隐藏的高级抽象特征，故广泛应用于各种分类问题。除了利用卷积神经网络模型之外，模型额外添加了复制填充技巧，这主要是因为卷积神经网络的输入要求为定长的序列，但在实际中 RNA 和蛋白质都是由非定长的核苷酸或氨基酸小分子序列组合而成。复制填充技巧可以简单描述为：若规定的长度恰好为当前序列长度的 n 倍，则直接将自身序列延长 n 倍作为输入即可；若规定的长度为当前序列长度的 n 倍且不能整除还剩余长度为 i 的短序列，则在将当前序列延长 n 倍后再在后面补上从第一个小分子开始长度为 i 的序列。具体流程如图 1 所示。

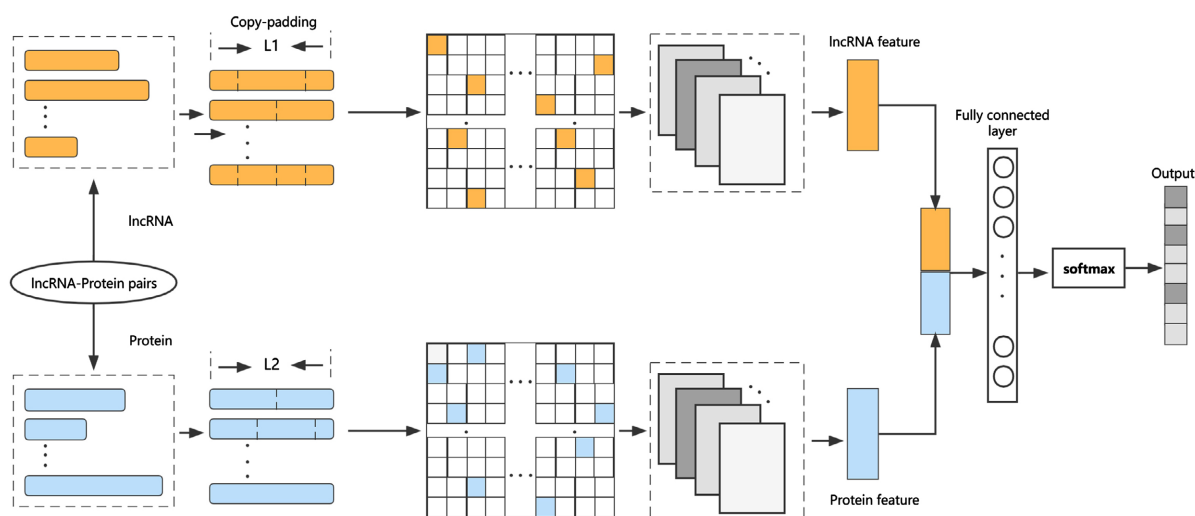


Figure 1. The flow chart of LPI-CNNCP

图 1. LPI-CNNCP 流程图

在 2021 年，Shen 等人提出了一种基于 GNN 的非编码 RNA-蛋白质相互作用预测方法，称为 NPI-GNN[9]。它能够根据网络信息和序列信息预测新的交互。它将 GraphSAGE 和 top-k pooling 结合到了 SEAL 框架中，构建了一个 ncRNA 和蛋白质相互作用的二部图网络，每个节点的特征表示由 ncRNA 和蛋白质的结构标签、node2vec 特征和 k-mer 特征组成。上述提到的 SEAL 框架是由 Zhang 等人在 2018 年提出的使用 GNN 进行链路预测的深度学习框架[10]，该框架使用 GNN 从局部子图学习启发式，以两个矩阵作为输入，一个是封闭子图的邻接矩阵 A ，另一个是节点信息矩阵 X ， X 代表了图中每个矩阵的特征，构建完两个矩阵之后再送入 GNN 中进行学习。

模型总体架构包括三个连续的特征处理模块，分别由 GraphSAGE、ReLU 和 top-k pooling 组成。它还包括三个全局池化模块，并用一个额外的加法模块来加和三个模块的结果，在这之后是三个连续的全连接层，分别包含 128 个、64 个、2 个神经元。当模型获取到输入的封闭图后，先经过第一个特征处理模块的处理，其次再依次进入到第二和第三个特征处理模块，由三个全局池化模块来读取三个特征处理模块的输出结果，并经过加法模块加和，结果送入全连接层，SoftMax 函数和 log 函数用来处理最后一个全连接层的输出来得到最终预测结果。在五倍交叉验证下证明，该模型可以有效进行非编码 RNA 和蛋白质的相互作用关系预测。具体流程如图 2 所示。

在 2021 年，Li 等人提出了一种新的多通道胶囊网络框架，集成用于 LPI 预测的多模特征，叫做 Capsule-LPI [11]。它集成了四种特征，包括序列特征、主题信息、理化性质和二级结构特征，该模型由

四个特征学习子网络和一个胶囊子网络构成。当模型接收到长非编码 RNA 和蛋白质对时,提取上述四种特征,并形成对应的四条特征向量,送进由四层全连接层构成的特征学习子网络中进行学习,每个特征学习子网络由四层全连接层组成,用该子网络不仅可以提取高层特征,而且可以统一特征向量的维数。最终形成一个三维输出向量,然后将从特征学习子网络中得到的三维输出向量作为“胶囊”处理获得 $U_1 - U_4$, $U_1 - U_4$ 包含根据每个特征预测得来的多个信息, $W_1 - W_4$ 是变换矩阵,它们能将 $U_1 - U_4$ 转换到相同的预测空间,得到的 $U'_1 - U'_4$ 即为利用四种特征产生的在相同预测空间下的预测,接下来,添加 $U'_1 - U'_4$ 作为新胶囊,使用激活函数将新胶囊的长度压缩到(0, 1)的范围,以获得最终胶囊 V ,取 V 的长度作为最终预测结果,长度大于 0.5 为阳性预测代表两者有关联,反之则为无关联。具体流程如图 3 所示。

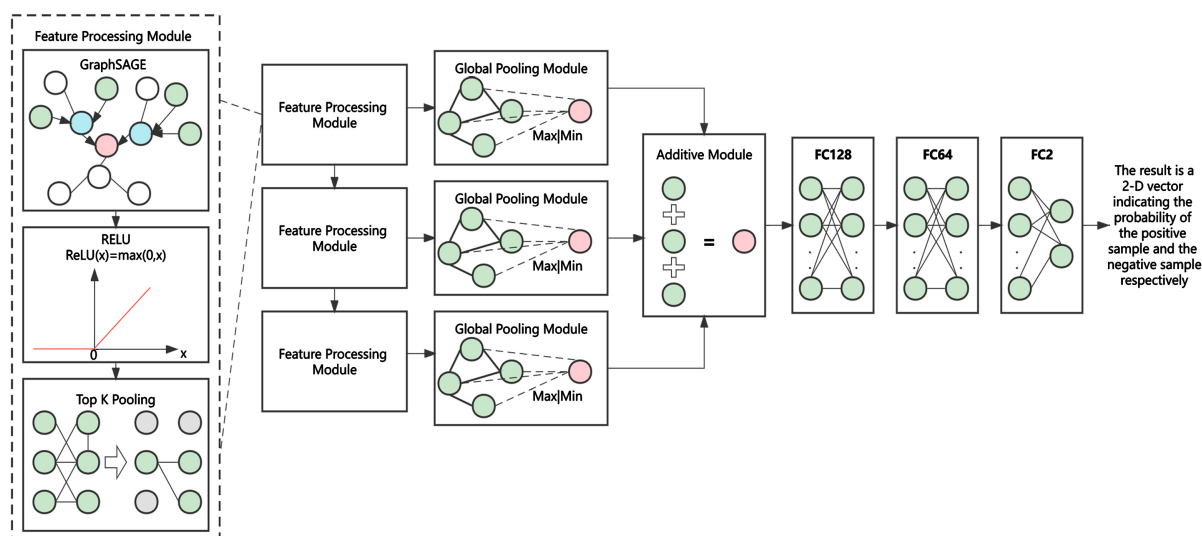


Figure 2. The flow chart of NPI-GNN

图 2. NPI-GNN 流程图

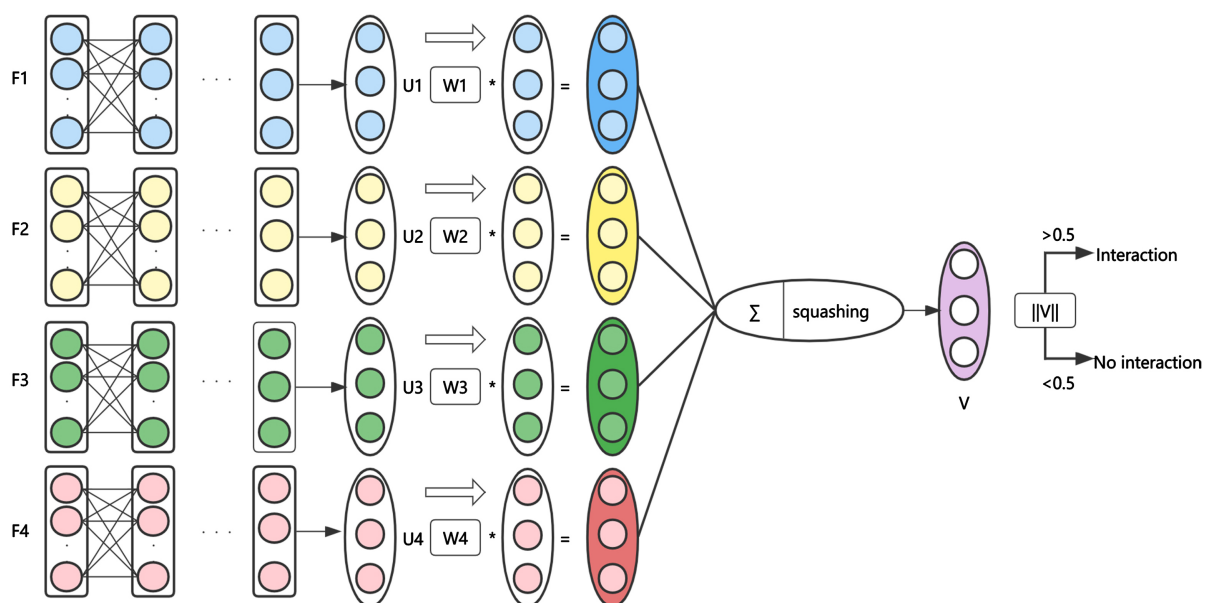


Figure 3. The flow chart of Capsule-LPI

图 3. Capsule-LPI 流程图

在 2021 年, Jin 等人提出一种基于图自动编码器和协同训练的端到端深度学习模型来预测长非编码 RNA 和蛋白质相互作用, 称为 LPIGAC [12]。与以往的研究不同, 该模型分别在长非编码 RNA 图和蛋白质图上实现了两个图自动编码器, 并对这两个编码器进行了协同训练。具体实现过程为, 首先整理出 lncRNAs 和蛋白质的相互作用矩阵 $Y = R^{m \times n}$, 其中 m 和 n 分别为 lncRNA 和蛋白质的数量, 同时构建出 lncRNAs 和蛋白质的相似性矩阵分别为 $S_l \in R^{m \times m}$ 和 $S_p \in R^{n \times n}$ 。将 Y , S_l , S_p 作为输入送入图自动编码器中进行学习训练, 最终得到打分矩阵 $F \in R^{m \times n}$, 其中 $F(i, j) \in [0, 1]$ 代表长非编码 RNA i 和蛋白质 j 的关联预测得分, 分数越高表明两者关联的可能性越大。

上面提到的图自动编码器(GAE)是一个拥有两层图卷积层的自动编码器, Shi 等人[13]证明了 GAE 能够模拟标签传播过程, 若设 Z_l 和 Z_p 分别为经过两层图卷积层后的输出, 则 $Z_l = \tanh\left(A_l \text{ReLU}\left(A_l Y \Theta^{(0)}\right) \Theta^{(1)}\right)$, $Z_p = \tanh\left(A_p \text{ReLU}\left(A_p Y^T \Phi^{(0)}\right) \Phi^{(1)}\right)$ 。其中 Θ 和 Φ 为神经网络的权重矩阵, A_l 和 A_p 代表 lncRNA 图和蛋白质图的归一化邻接矩阵, 由 S_l 和 S_p 计算而来, $A_l = D_l^{-1/2} S_l D_l^{-1/2}$, $A_p = D_p^{-1/2} S_p D_p^{-1/2}$ 。 D_l 和 D_p 为邻接矩阵的度矩阵, 对应的打分矩阵 $F_l \in R^{m \times n}$ 和 $F_p \in R^{n \times m}$ 可以分别经由 Z_l 和 Z_p 解码得来, 表示如下 $F_l = \text{sigmoid}\left(A_l \text{ReLU}\left(A_l Z_l \Theta^{(2)}\right) \Theta^{(3)}\right)$, $F_p = \text{sigmoid}\left(A_p \text{ReLU}\left(A_p Z_p \Phi^{(2)}\right) \Phi^{(3)}\right)$, 最终的关联打分矩阵 F 通过 F_l 和 F_p 计算得来, $F = \alpha F_l + (1 - \alpha) F_p^T$, 公式中的 $\alpha \in (0, 1)$ 是为平衡从 lncRNA 空间和从蛋白质空间得到的信息而设置的权重参数。具体流程如图 4 所示。

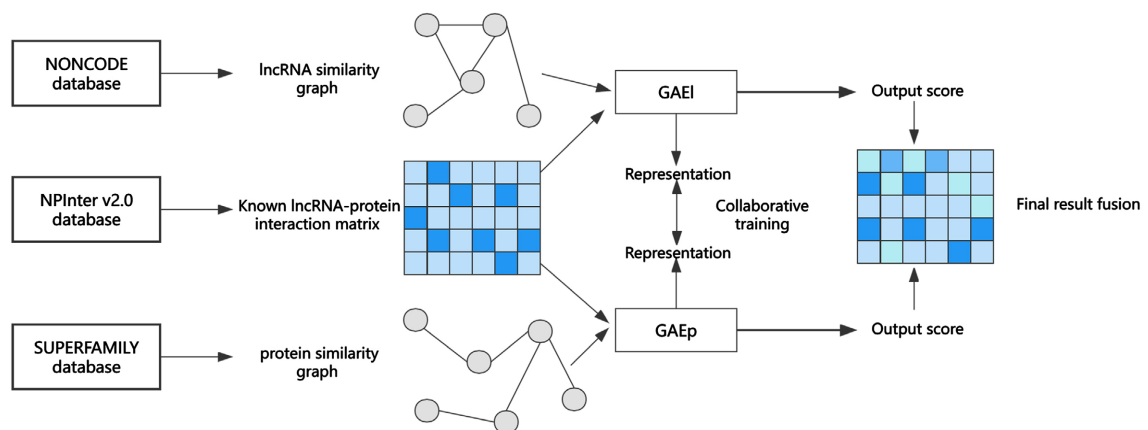


Figure 4. The flow chart of LPIGAC
图 4. LPIGAC 流程图

5. 结果比较

表 1 为四种深度学习模型在长非编码 RNA 与蛋白质关系预测表现上的比较。由指标值可见, 四种模型都取得了 0.9 以上的 AUC 或 ACC, 说明四种模型能够有效的预测出潜在的 LPI, 同时 NPI-GNN 和 LPIGAC 等利用了图学习理论来构建预测模型的方法性能要好于单纯使用了基础深度学习模型卷积神经网络、胶囊神经网络的 LPI-CNNCP 和 Capsule-LPI, 这表明图学习能够充分提取图结构数据中的潜在信息, 很好的适用于了长链非编码 RNA 与蛋白质的相互作用关系预测问题。故在以后的研究中, 图学习仍是一大热点。

Table 1. Comparison of experimental results of four models
表 1. 四种模型实验表现结果对比

Method	Classifier	Database	Performance (ACC, AUC)
LPI-CNNCP	CNN	RPI1446	0.866 ± 0.2、/

Continued

NPI-GNN	Graph SAGE	NPInter2.0	0.933、/
Capsule-LPI	Capsule network	NPInter2.0	0.917 ± 0.86、0.953 ± 0.41
LPIGAC	GAE	NPInter2.0	/、0.936

6. 结论与展望

研究证明长链非编码 RNA 与蛋白质之间的相互作用机制能影响各种各样的生命活动,因此使用深度学习工具来进行长非编码 RNA 与蛋白质相互作用关系预测是很有必要的,这样能节省大量的时间和物质成本。但目前的模型仍存在局限性,例如现有的长非编码 RNA 与蛋白质关联数据十分有限,这导致了我們只能使用相对小规模的数据集来训练计算模型,从而导致计算精度偏低。另外负样本的获取同样是一个不可避免的问题,由于已知的都是有相互关联的作用对,故大多数研究人员采用从排除掉正样本的所有两两组合数据中抽样作为负样本,由于我们不知道某个未知长非编码 RNA 和蛋白质究竟是否有关联就盲目把它当作负样本使用,故模型也会因此出现误差。为了克服这些缺点,故特为以后的实验模型提出几点建议: 1、构建更大规模、更全面的长非编码 RNA-蛋白质相互作用数据库,大量的相互作用数据有助于获得更可靠的预测模型; 2、尽量寻找一些经过实验验证的可靠的负本来训练模型,或者使用 dropout 策略,减小误差。3、找到更好的深度学习模型来从原始数据中提取更高级的特征。这些都有利于改良我们的模型,以获得更准确的预测。

在这篇综述中,我们全面回顾了近些年来使用深度学习工具来进行长非编码 RNA 与蛋白质相互作用关系预测的一些方法,这些方法已经成功预测了长非编码 RNA-蛋白质关联,并取得了优良结果。此外,在文章最后我们也总结了上面模型中存在的缺点,并为未来从事此研究方向的学者提出了一些建议,希望这些能帮助改进现有的方法来取得更优的结果。

参考文献

- [1] Lowe, R., Shirley, N., Bleackley, M., Dolan, S. and Shafee, T. (2017) Transcriptomics Technologies. *PLoS Computational Biology*, **13**, e1005457. <https://doi.org/10.1371/journal.pcbi.1005457>
- [2] Djebali, S., Davis, C.A., Merkel, A., et al. (2012) Landscape of Transcription in Human Cells. *Nature*, **489**, 101-108. <https://doi.org/10.1038/nature11233>
- [3] Amaral, P.P., Clark, M.B., Gascoigne, D.K., et al. (2011) lncRNADB: A Reference Database for Long Noncoding RNAs. *Nucleic Acids Research*, **39**, 146-151. <https://doi.org/10.1093/nar/gkq1138>
- [4] Zhao, Y., Li, H., Fang, S., et al. (2016) NONCODE 2016: An Informative and Valuable Data Source of Long Non-Coding RNAs. *Nucleic Acids Research*, **44**, D203-D208. <https://doi.org/10.1093/nar/gkv1252>
- [5] Bhartiya, D., Pal, K., Ghosh, S., et al. (2013) lncRNome: A Comprehensive Knowledge Base of Human Long Non-coding RNAs. *Database*, **2013**, bat034. <https://doi.org/10.1093/database/bat034>
- [6] Yuan, J., Wu, W., Xie, C., et al. (2014) NPInter v2.0: An Updated Database of ncRNA Interactions. *Nucleic Acids Research*, **42**, D104. <https://doi.org/10.1093/nar/gkt1057>
- [7] Fan, X.-N. and Zhang, S.-W. (2019) LPI-BLS: Predicting lncRNA-Protein Interactions with a Broad Learning System-Based Stacked Ensemble Classifier. *Neurocomputing*, **370**, 88-93. <https://doi.org/10.1016/j.neucom.2019.08.084>
- [8] Zhang, S.W., Zhang, X.X., Fan, X.N. and Li, W.N. (2020) Lpi-cnncp: Prediction of lncRNA-Protein Interactions by Using Convolutional Neural Network with the Copy-Padding Trick. *Analytical Biochemistry*, **601**, Article ID: 113767. <https://doi.org/10.1016/j.ab.2020.113767>
- [9] Shen, Z.A., Luo, T., Zhou, Y.K., Yu, H. and Du, P.F. (2021) NPI-GNN: Predicting ncRNA-Protein Interactions with Deep Graph Neural Networks. *Briefings in Bioinformatics*, **22**, No. 5. <https://doi.org/10.1093/bib/bbab051>
- [10] Zhang, M. and Chen, Y. (2018) Link Prediction Based on Graph Neural Networks. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 3-8 December

-
- 2018, Montréal, 5171-5181.
- [11] Li, Y., Sun, H., Feng, S., *et al.* (2021) Capsule-LPI: A LncRNA-Protein Interaction Predicting Tool Based on a Capsule Network. *BMC Bioinformatics*, **22**, Article Number: 246. <https://doi.org/10.1186/s12859-021-04171-y>
 - [12] Jin, C., Shi, Z., Zhang, H. and Yin, Y. (2021) Predicting lncRNA-Protein Interactions Based on Graph Autoencoders and Collaborative Training. 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, 9-12 December 2021, 38-43. <https://doi.org/10.1109/BIBM52615.2021.9669316>
 - [13] Shi, Z., Zhang, H., Jin, C., Quan, X. and Yin, Y. (2021) A Representation Learning Model Based on Variational Inference and Graph Autoencoder for Predicting LncRNA-Disease Associations. *BMC Bioinformatics*, **22**, Article Number: 136. <https://doi.org/10.1186/s12859-021-04073-z>