Hans Journal of Computational Biology 计算生物学, 2023, 13(2), 11-21
Published Online June 2023 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2023.132002

Hans汉斯

LPI-MAM：以miRNAs为中介基于深度学习预测lncRNA-蛋白质相互作用

屈文燕，颜　静，李晓毅，谭建军*

北京工业大学环境与生命学院生物医学工程系，北京

收稿日期：2023年5月10日；录用日期：2023年6月12日；发布日期：2023年6月19日

摘　要

长链非编码RNA (Long non-coding RNAs, lncRNAs)是细胞增殖和死亡的重要调控因子，它的失调可能会导致多种疾病发生。LncRNAs主要是通过与蛋白质相互作用(lncRNA-protein interactions, lncRPIs)来发挥生物学功能。因此，研究lncRPIs对了解lncRNAs的功能及相关疾病至关重要。目前，多数计算方法依赖于已知的验证过的lncRPIs构建模型，但经过实验验证的样本是有限的。MiRNAs主要是与mRNAs结合导致基因沉默，而lncRNAs可作为竞争性内源性RNA，竞争性的结合miRNAs来间接地调节基因表达。本文提出LPI-MAM方法，使用miRNAs作为中间体来扩大lncRPIs的预测范围。该方法将序列、结构和组成转换分布特征融合，输入卷积神经网络和独立循环神经网络的集成深度学习框架中。结果表明，LPI-MAM在基准数据集上取得了良好的性能。并且通过构建可视化交互网络发现该模型具有预测未知lncRPIs的能力。

关键词

LncRNA-蛋白质相互作用，miRNA，中间体，组成转换分布，深度学习

LPI-MAM: Predicting lncRNA-Protein Interactions with miRNAs as Mediators Based on Deep Learning

Wenyan Qu, Jing Yan, Xiaoyi Li, Jianjun Tan*

Department of Biomedical Engineering, Faculty of Environment and Life, Beijing University of Technology, Beijing

*通讯作者。

文章引用: 屈文燕, 颜静, 李晓毅, 谭建军. LPI-MAM: 以 miRNAs 为中介基于深度学习预测 lncRNA-蛋白质相互作用[J]. 计算生物学, 2023, 13(2): 11-21. DOI: 10.12677/hjcb.2023.132002

Received: May 10th, 2023; accepted: Jun. 12th, 2023; published: Jun. 19th, 2023

Abstract

Long non-coding RNAs (lncRNAs) are crucial regulatory factors of cell proliferation and death, its dysregulation may lead to the occurrence of a variety of diseases. LncRNAs play biological functions mainly through lncRNA-protein interactions (lncRPIs). Therefore, it becomes essential to study the interactions between lncRNA and protein (lncRPIs) for exploring the function of lncRNAs. At present, almost computational methods depend on known lncRPIs to build a model. However, the samples that have been verified are limited. MiRNAs mainly bind to mRNAs to cause gene silencing. As competitive endogenous RNAs (ceRNAs), lncRNAs can indirectly regulate gene expression by competitively binding miRNAs. This study proposes the LPI-MAM method, which uses miRNAs as mediators to expand the prediction range of lncRPIs. The features of sequence, structure and composition transformation distribution (CTD) are fused and then input into the integrated deep learning framework of convolutional neural network (CNN) and independent recurrent neural network (IndRNN). The results indicate that LPI-MAM has achieved good performance on benchmark dataset. And by constructing a visual interaction network, it is found that the model has the ability to predict unknown lncRPIs.

Keywords

LncRNA-Protein Interactions, MicroRNA, Mediators, Composition Transformation Distribution, Deep Learning

1. 引言

非编码 RNAs (non-coding RNAs, ncRNAs)是指不编码蛋白质的 RNA。根据 ncRNAs 的长度可以将其分为小 ncRNAs 和长 ncRNAs (long non-coding RNAs, lncRNAs)。LncRNAs 控制着转录、翻译和剪接等许多细胞基本过程，它的失调可能会导致许多疾病发生。因此，探索 lncRNAs 的功能有助于理解相关疾病的机理。一般来说，lncRNAs 可以结合许多类型的生物分子，如 miRNAs、mRNA、DNA 和蛋白质[1]。其中，lncRNAs 与蛋白质的相互作用(lncRNA-protein interactions, lncRPIs)是 lncRNAs 发挥其功能的重要途径。因此，可以通过研究 lncRPIs 来了解 lncRNAs 的功能和相关疾病的发病机制[2] [3]。

LncRPIs 可以通过实验方法验证，也可以通过计算方法预测。计算方法通常是基于机器学习、深度学习或网络算法，它可以为生物实验提供指导意见。目前已经有许多计算方法可以预测 lncRPIs。Muppirala 等人[4]提出了 RPISeq，该方法使用 RNA 和蛋白质序列数据作为输入来预测 RNA-蛋白质相互作用。然后提出 RPISeq 的两种变体，分别是使用支持向量机(support vector machine, SVM)的 RPISeq-SVM 和使用随机森林(random forest, RF)的 RPISeq-RF。Peng 等人[5]提出了 RPITER。该方法通过改进联合三元特征(conjoint triad feature, CTF)编码，使 RNA 和蛋白质具有全面的序列和结构特征。Pan 等人[6]提出了 IPMiner，它使用堆叠的自编码器从蛋白质和 RNA 序列的综合特征中提取隐藏的序列相互作用。然后将提取的信息输入到 RF 中。Xiao 等人[7]提出了 PLPIHS，它由三个子网络组成，然后通过 HeteSim 指标计算 lncRPIs

在网络中的相互关系评分。最后，SVM 分类器基于该评分预测 lncRPIs。Hu 等人[8]提出了 HLPI-ensemble，集成了 SVM、RF 和极限梯度增强(extreme gradient boosting, XGBoost)三种算法来预测 lncRPIs。Ge 等人[9]提出了 LPBNI，该方法建立了一个二分网络，把连接起来的 lncRNA-蛋白认为是有相互作用的。然后在一种输入与输出均为二进制的人工神经网络中对每个 lncRNA 对应的蛋白进行评分和排序。本课题组 Wang 等人[10]基于深度学习，用 EDLMFC 融合序列、二级结构和三级结构来预测非编码 RNA 和蛋白质相互作用(ncRNA-protein interactions, ncRPIs)。Huang 等人[11]提出了 LPI-CSFFR，该方法将原始序列信息、二级结构信息和物理化学信息串联融合，最大限度地提高了各特征对预测结果的贡献。

以往研究中，几乎所有模型都依赖于已知的、经实验验证的 lncRPIs 来构建预测模型。然而，由于已被验证的 lncRPIs 数量有限，这就对先前多数方法的模型性能造成了一定的影响。为了扩大 lncRPIs 的可预测范围，Zhou 等人[12]提出 LPI-MMHN，选择将 miRNAs 作为中间体基于网络的方法构建相似矩阵，然后通过网络算法计算相关评分来预测 lncRPIs。从生物层面看，MiRNAs 通过抑制目标 mRNA 的翻译和破坏其稳定性来发挥调控作用。而 lncRNAs 可作为竞争内源性 RNA (competition endogenous RNA, ceRNA)，与 miRNAs 竞争性的结合从而调节基因表达[13]。从数据层面看，Zhou 等人[12]通过卡方检验统计分析也证明有共同 miRNAs 作为中间体的 lncRPIs 在所有的 lncRPIs 中所占的比例是很高的。因此，miRNAs 被考虑作为预测 lncRPIs 的中间体。但是基于网络的方法预测 lncRPIs 时，只有 lncRPIs 在网络中才能够有效，这就使得模型很有局限性。另外 lncRPIs 网络由多个子网络组成，每个节点的分布不平衡也会影响其预测结果[9]。

在本研究中，miRNAs 将作为中间体基于深度学习集成框架来预测 lncRPIs，称为 LPI-MAM。该方法以序列信息、二级结构信息和组成转换分布(composition transformation distribution, CTD)信息为输入，通过卷积神经网络(convolutional neural network, CNN)和独立循环神经网络(independent recurrent neural network, IndRNN)的集成深度学习模型提取特征，最后利用 softmax 函数对 lncRPIs 进行预测。结果表明，在 RAIDv2.0 和 RPI5392 数据集上，LPI-MAM 的准确率分别为 92.23%和 98.83%。与 LPI-MMHN [12]、RPITER [5]、IPMiner [6]和 PRPI-SC [14]方法相比，本文提出的 LPI-MAM 方法综合性能最好。此外，通过构建可视化预测网络表明 LPI-MAM 有预测新的交互关系的能力。

2. 材料和方法

2.1. 基准数据集

RAIDv2.0 [15]是一个大规模的生物分子相互作用数据库。本研究从 Zhou 等人[12]提出的 LPI-MMHN 研究中下载了 1356 个 lncRNA-miRNA 相互作用、1156 个蛋白质-miRNA 相互作用以及 1925 个 lncRNA-蛋白质相互作用。其中 lncRNA-miRNA 相互作用和蛋白质-miRNA 相互作用是通过共同 miRNAs 筛选出来的。然后，根据已验证的 1925 个 lncRNA-蛋白互作对，从 1356 个 lncRNA-miRNA 互作对和 1156 个蛋白-miRNA 互作对中筛选出 2329 个 lncRNA-miRNA-蛋白互作对作为阳性样本。由于 RNAfold [16]预测 lncRNA 的最优二级结构时，为保证对应的自由能(minimum corresponding free energy, MFE)最小，序列长度被限制在 1 万内。最终筛选出 2134 个 lncRNA-miRNA-蛋白相互作用对作为阳性样本。阴性样本为 2089 个 lncRNA-miRNA 有相互作用，miRNA-蛋白有相互作用，但 lncRNA-蛋白之间没有相互作用的样本。数据按 7:3 分成两部分。一部分是训练集，训练集有 1493 对正样本和 1462 对负样本。另一部分是测试集，有 641 个正样本和 627 个负样本。然后从 Zhou 等人[12]提出的 LPI-MMHN 研究中，再下载 20425 对 lncRNA-miRNA，1349 对蛋白-miRNA 和 2803 对 lncRNA-蛋白质。用类似的方法筛选训练集，命名为 RPI5392。该数据集是将 RAIDv2.0 数据库中的大多数被丢弃的预测交互作用集成起来的一个更大的

数据集。

2.2. 特征提取

2.2.1. 序列编码

LncRNA 序列来源于 NCBI 基因库[17]。MiRNA 序列从 miRBase 数据库中获得[18]。蛋白质序列来源于 UniProt 数据库[19]。对于无法找到的 lncRNA，从 Ensemble 数据库中搜索[20]。本研究采用 RPITER [5]提出的序列编码方法，然后使用联合 k-mer 对 lncRNA、miRNA 和蛋白质进行编码。对于 lncRNA 选择 1-4mer 编码得到 340 ($\sum_{k=1}^{4}4^{k}$)维的数值向量。由于 miRNA 的序列较短，若使用 1-4mer 编码，会使矩阵过于稀疏[21]。因此选择 1-3mer 对 miRNA 编码，得到 84 ($\sum_{k=1}^{3}4^{k}$)个元素的编码向量。对于蛋白质，将 20 种氨基酸按照侧链体积和偶极矩分为 G1~G7 七类，选择 1-3mer 得到 399 个($\sum_{k=1}^{3}7^{k}$)个元素的编码向量。

2.2.2. 结构编码

LncRNA 的结构信息通过 RNAfold [16]预测得到。LncRNA 的二级结构包括环区和茎区，用括号和点表示。蛋白质的二级结构由 SOPMA [22]预测得到，它包括α-helix (H)，β-sheet (E)和 coil (C)三种结构，用“hec”表示。与序列特征编码类似，结构特征也使用联合 k-mer 编码。LncRNA 选择 1-4mer 得到 30 维($\sum_{k=1}^{4}2^{k}$)编码向量，蛋白质选择 1-3mer 得到 39 维($\sum_{k=1}^{3}3^{k}$)编码向量。

2.2.3. 组成转换分布特征

CTD 来源于组成、转换和分布[23] [24]，共 30 维。本研究使用 CTD 特征来表示 lncRNA 和 miRNA 的结构信息[21]。组成特征是具有特定性质的核苷酸的数量与核苷酸总数的比。转换特征表示具有特定性质的核苷酸后紧跟的具有不同性质的核苷酸的频率百分比。分布特征是对具有特定性质的核苷酸位于链长的第一、1/4、1/2、3/4 和最后一个位置进行链长测量。假设一个 RNA 的序列为“CTGTAATCACAGCTGTCAGG”，下面说明该 RNA 的 CTD 特征计算过程。该 RNA 序列包括 5 个 A、5 个 G、5 个 T 和 5 个 C。那么组成特征分别为 0.25、0.25、0.25 和 0.25。转换特征分别为 0.105 (AT)、0.211 (AC)、0.105 (AG)、0.211 (TG)、0.211 (TC)、0.053 (GC)。分布特征的计算以核苷酸 A 为例，第一个节点位于序列的第 5 个位置。1/4、1/2、3/4 和最后一个节点分别位于第 6、9、11、18 位。因此，A 的分布特征为 0.25、0.3、0.45、0.55、0.9。CTD 特征考虑的是每个核苷酸前后的相关性，每个 lncRNA 和 miRNA 都可以被表示为一个 30 维的特征向量。

2.3. 模型设计

联合编码的 lncRNA、miRNA 和蛋白质的序列特征、二级结构特征结合 CTD 特征，分别形成 400 维、438 维和 114 维的特征向量。将这三个特征向量输入 CNN 网络，提取隐藏的高级生物特征。然后再将学习到的高级特征输入到 IndRNN 层，学习特征之间的关系。最后将 IndRNN 层的三个输出连接在一起，利用 softmax 激活函数完成预测。整个工作流程如图 1 所示。

经过一系列调整，选择三层 CNN 结合一层 IndRNN 构建训练模型。通过优化模型的主要参数，如学习率、滤波器大小、内核大小、IndRNN 隐藏大小等，为随机选择的验证集优化性能指标。蛋白质子网络的参数值为层数 3；滤波器尺寸 45、96、128；核大小 6、6、6；丢包率：0.2、0.2、0.2；IndRNN 隐藏大小 64；全连接层大小 64。对于 lncRNAs 子网络，除了内核大小分别为 6、5 和 5 外，其他参数值与蛋白质相同。MiRNA 子网络的参数与 lncRNA 一致，只是内核大小分别为 6、4 和 5。最后三层完全连接的神经元分别为 128、64 和 2。该模型由 Keras2.3.1 完成。

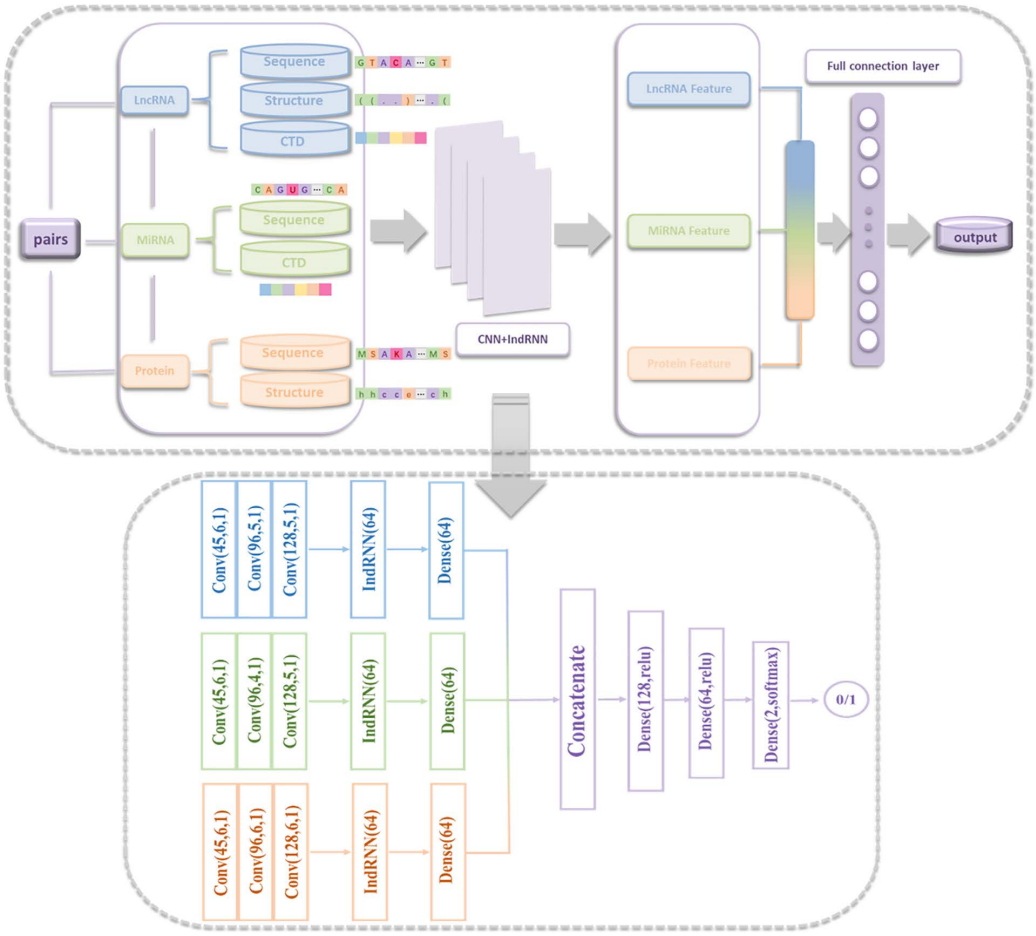


Figure 1. Flowchart of the proposed method LPI-MAM
图 1. LPI-MAM 流程图

2.4. 评估指标

本研究采用 5 倍交叉验证的方法，重复多次。评估指标的公式如下所示。此外，受试者工作特征曲线(receiver operating characteristic, ROC)下面积(AUC)指标也被使用。

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{SEN}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{SPE}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$

$$\mathrm{PRE}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}\times\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{5}$$

$$\mathrm{F1}=\frac{2\times\mathrm{SEN}\times\mathrm{PRE}}{\mathrm{SEN}+\mathrm{PRE}} \tag{6}$$

其中TP代表真阳性样本，TN为真阴性样本，FP表示假阳性样本，FN是假阴性样本。准确率(accuracy, ACC)是表示预测正确的样本占总数据的比例。灵敏度(sensitivity, SEN)表示所有阳性样本的配对比例。特异性(specificity, SPE)反映阴性样本的预测能力。精度是PRE (precision)。马修斯相关系数(Matthews correlation coefficient, MCC)用于正负样本量不平衡的情况。F1分数是将SEN和PRE都考虑在内的综合指标。

3. 结果

3.1. 模型在不同数据集上的性能表现

构建的LPI-MAM模型在数据集RAIDv2.0和RPI5392上进行训练。从表1可以看出，RAIDv2.0数据集上的ACC值为92.23%，AUC值为96.80%，RPI5392数据集上的ACC值为98.83%，AUC值为99.67%。在两种数据集上均表现出较好的预测性能。

Table 1. Training results on RAIDv2.0 and RPI5392 datasets
表1. 数据集RAIDv2.0和RPI5392上的训练结果

数据集	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	MCC (%)	AUC (%)
RAIDv2.0	92.23 ± 0.12	93.47 ± 0.28	90.97 ± 0.31	91.40 ± 0.26	92.40 ± 0.11	84.53 ± 0.23	96.80 ± 0.09
RPI5392	98.83 ± 0.02	96.01 ± 0.10	99.44 ± 0.02	97.35 ± 0.09	96.67 ± 0.06	95.97 ± 0.07	99.67 ± 0.02

3.2. 特征分析

为了分析模型输入不同特征和不同特征组合时的性能差异，模型分别采用序列，序列与结构，序列与CTD，结构与CTD，序列、结构与CTD 5种情况在RAIDv2.0和RPI5392数据集上进行训练。结果见表2，图2是模型在数据集RAIDv2.0上训练得到的ROC曲线和精确率－召回率(precision-recall, PR)曲线。

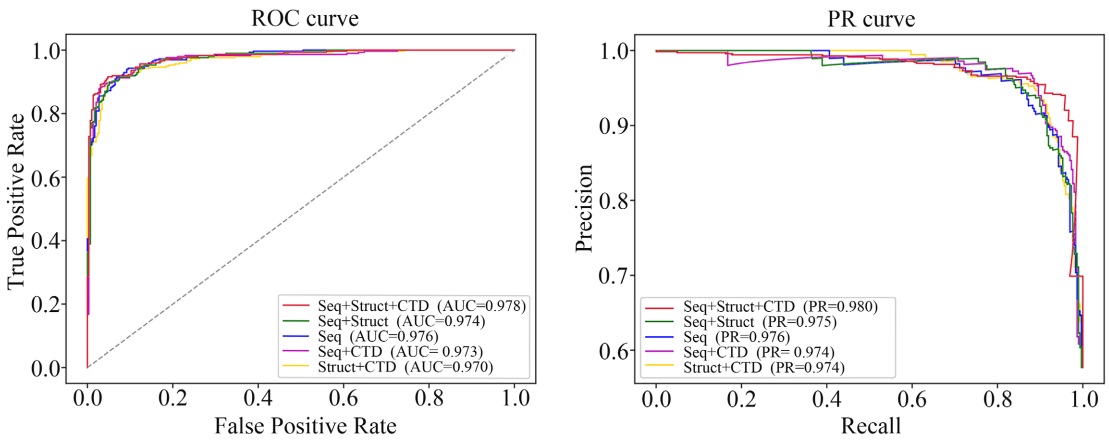


Figure 2. Model performance with different feature inputs on RAIDv2.0. The left figure is the ROC curve. The right figure shows the PR curve
图2. 在RAIDv2.0上不同特征输入时的模型性能。左图为ROC曲线。右图为PR曲线

从表2可以看出，当只输入结构特征和CTD特征时，模型的7个指标显著低于其他特征组合，说明序列信息的重要性。在RAIDv2.0数据集上，当所有特征都作为输入时，模型的七个指标都是最高的。其中ACC达到92.23%，AUC达到96.80%。同样，从表3可以看出，在RPI5392数据集中，当所有特征都作为输入时，模型的ACC、SPE、PRE、F1和MCC指标最高，只有SEN值略低于输入序列和结构组合，

AUC 值略低于仅输入序列时的情况。可见这些特征相互补充，覆盖了更全面的信息，使模型的预测性能达到最高。

Table 2. Training results of different features on RAIDv2.0 dataset
表 2. 数据集 RAIDv2.0 上不同特征的训练结果

数据集	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	MCC (%)	AUC (%)
序列	91.88 ± 0.12	93.02 ± 0.22	90.72 ± 0.31	91.16 ± 0.25	92.05 ± 0.11	83.84 ± 0.24	96.73 ± 0.15
序列 + 结构	91.98 ± 0.33	93.09 ± 0.49	90.85 ± 0.40	91.27 ± 0.36	92.14 ± 0.33	84.03 ± 0.66	96.71 ± 0.10
序列 + CTD	91.74 ± 0.22	92.86 ± 0.26	90.59 ± 0.33	91.02 ± 0.29	91.90 ± 0.21	83.54 ± 0.44	96.54 ± 0.14
结构 + CTD	90.91 ± 0.28	91.67 ± 0.39	90.12 ± 0.27	90.52 ± 0.25	91.06 ± 0.28	81.87 ± 0.55	96.37 ± 0.16
序列 + 结构 + CTD	92.23 ± 0.12	93.47 ± 0.28	90.97 ± 0.31	91.40 ± 0.26	92.40 ± 0.11	84.53 ± 0.23	96.80 ± 0.09

Table 3. Training results of different features on RPI5392 dataset
表 3. 数据集 RPI5392 上不同特征的训练结果

数据集	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	MCC (%)	AUC (%)
序列	98.80 ± 0.03	96.02 ± 0.13	99.40 ± 0.04	97.16 ± 0.16	96.58 ± 0.08	95.86 ±0.10	99.70 ± 0.02
序列 + 结构	98.82 ± 0.02	96.07 ± 0.10	99.41 ± 0.02	97.22 ± 0.08	96.64 ± 0.05	95.93 ± 0.06	99.69 ± 0.02
序列 + CTD	98.80 ± 0.02	96.02 ± 0.11	99.40 ± 0.02	97.18 ± 0.09	96.60 ± 0.05	95.87 ± 0.05	99.70 ± 0.01
结构 + CTD	98.61 ± 0.03	95.35 ± 0.13	99.31 ± 0.05	96.76 ± 0.22	96.04 ± 0.10	95.21 ± 0.12	99.63 ± 0.02
序列 + 结构 + CTD	98.83 ± 0.02	96.01 ± 0.10	99.44 ± 0.02	97.35 ± 0.09	96.67 ± 0.06	95.97 ± 0.07	99.67 ± 0.02

3.3. 与其他方法比较

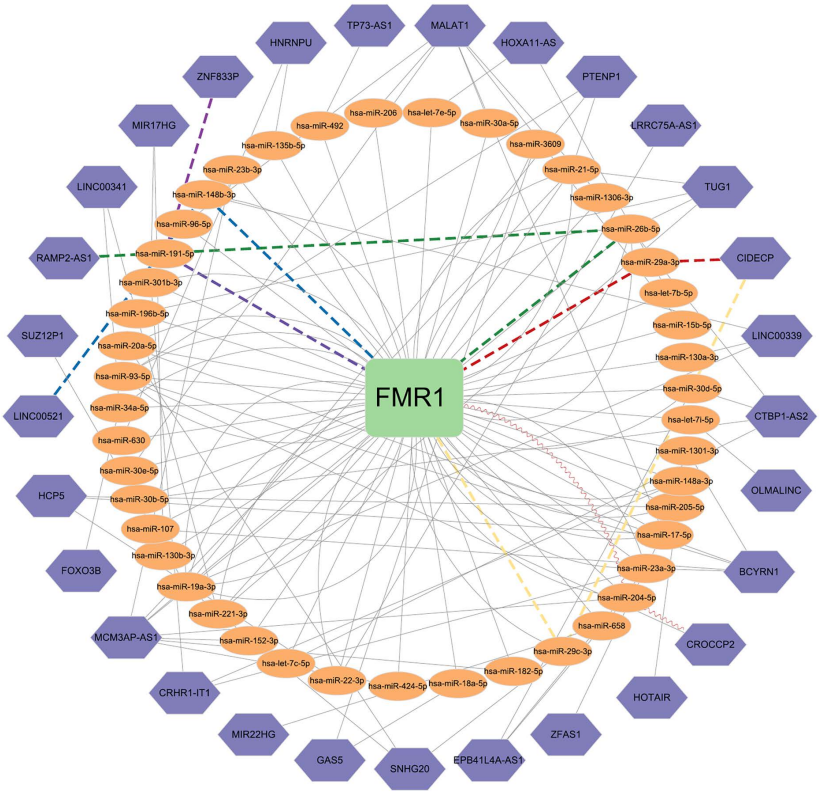
将该模型 LPI-MAM 与 LPI-MMHN [12]、RPITER [5]、IPMiner [6]和 PRPI-SC [14]四种方法比较，结果如表 4 所示。从表 4 数据可以看出，LPI-MAM 的 ACC、SEN、F1 和 MCC 显著高于其他模型。只有 SPE 值略低于 RPITER 和 PRPI-SC，PRE 和 AUC 值略低于 PRPI-SC。总的来说，本研究提出的 LPI-MAM 整体性能是最好的。

Table 4. Compared with other different methods on RAIDv2.0 dataset
表 4. 在数据集 RAIDv2.0 上与其他不同方法比较

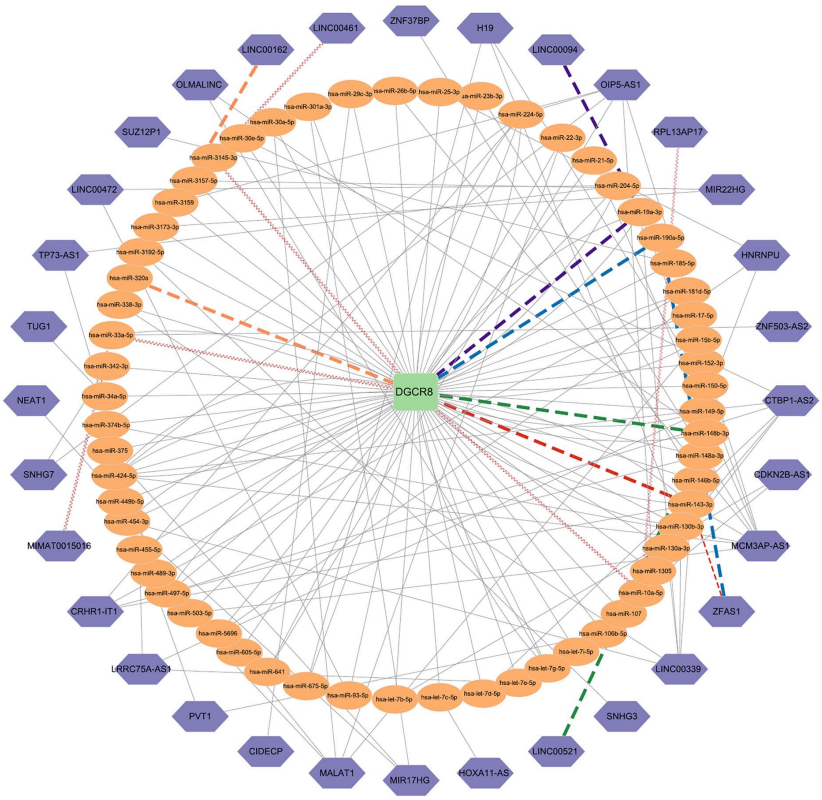
方法	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	MCC (%)	AUC (%)
LPI-MMHN	72.37 ± 0.00	62.12 ± 0.00	73.84 ± 0.00	32.80 ± 0.00	34.94 ± 0.00	29.04 ± 0.00	82.10 ± 0.00
RPITER	73.58 ± 0.95	47.38 ± 3.95	91.67 ± 1.35	81.50 ± 1.70	58.67 ± 3.29	45.38 ± 1.63	83.96 ± 0.20
IPMiner	77.37 ± 1.16	61.00 ± 2.45	88.67 ± 0.54	78.95 ± 1.19	68.55 ± 2.03	52.61 ± 2.50	74.83 ± 1.35
PRPI-SC	92.18 ± 0.21	92.73 ± 0.84	91.61 ± 1.05	92.04 ± 0.83	92.29 ± 0.19	84.52 ± 0.33	97.56 ± 0.15
LPI-MAM	92.23 ± 0.12	93.47 ± 0.28	90.97 ± 0.31	91.40 ± 0.26	92.40 ± 0.11	84.53 ± 0.23	96.80 ± 0.09

3.4. 构建 lncRNA-miRNA-蛋白质网络

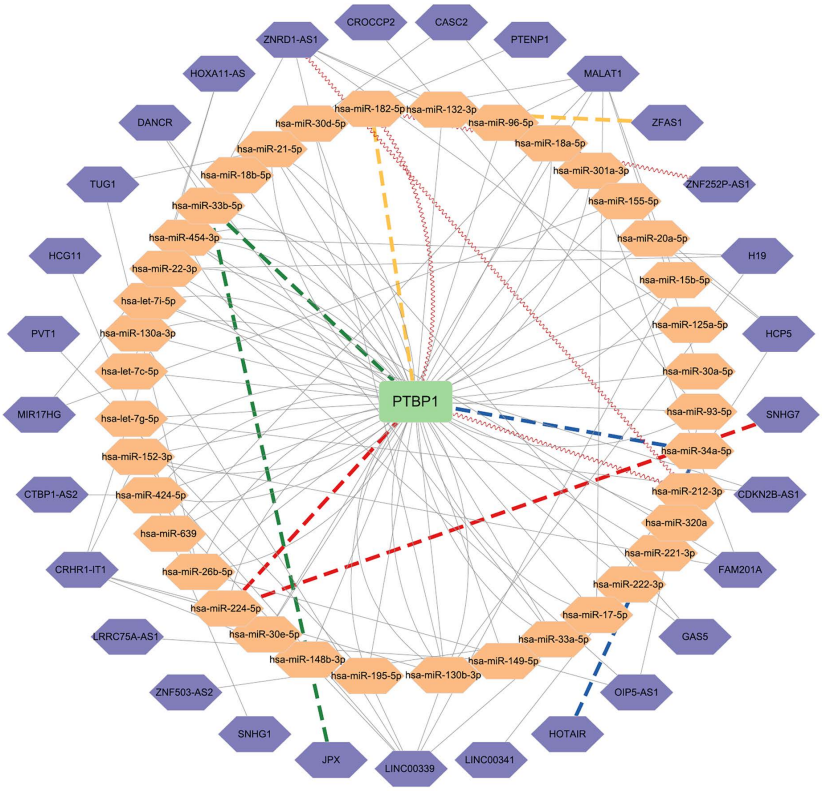
将从 RAIDv2.0 分出的三份数据作为测试集，输入到训练保存模型中，得到预测结果。将结果输入 cytoscape [25]构建 lncRNA-miRNA-蛋白网络。由于有大量的相互作用，最终选择构建 3 个热点蛋白。


FMR1
TP73-AS1
MALAT1
HOXA11-AS
PTENP1
LRRC75A-AS1
TUG1
CIDECP
LINC00339
CTBP1-AS2
OLMALINC
BCYRN1
CROCCP2
HOTAIR
ZFAS1
EPB41L4A-AS1
SNHG20
GAS5
MIR22HG
CRHR1-IT1
MCM3AP-AS1
FOXO3B
HCP5
LINC00521
SUZ12P1
RAMP2-AS1
LINC00341
MIR17HG
ZNF833P
HNRNPU
hsa-miR-206
hsa-let-7e-5p
hsa-miR-30a-5p
hsa-miR-492
hsa-miR-135b-5p
hsa-miR-3609
hsa-miR-23b-3p
hsa-miR-21-5p
hsa-miR-148b-3p
hsa-miR-1306-3p
hsa-miR-96-5p
hsa-miR-26b-5p
hsa-miR-191-5p
hsa-miR-29a-3p
hsa-miR-301b-3p
hsa-let-7b-5p
hsa-miR-196b-5p
hsa-miR-15b-5p
hsa-miR-20a-5p
hsa-miR-130a-3p
hsa-miR-93-5p
hsa-miR-30d-5p
hsa-miR-34a-5p
hsa-let-7i-5p
hsa-miR-630
hsa-miR-1301-3p
hsa-miR-30e-5p
hsa-miR-148a-3p
hsa-miR-30b-5p
hsa-miR-205-5p
hsa-miR-107
hsa-miR-17-5p
hsa-miR-130b-3p
hsa-miR-23a-3p
hsa-miR-19a-3p
hsa-miR-204-5p
hsa-miR-221-3p
hsa-miR-658
hsa-miR-152-3p
hsa-miR-29c-3p
hsa-let-7c-5p
hsa-miR-182-5p
hsa-miR-22-3p
hsa-miR-424-5p
hsa-miR-18a-5p

(a)


DGCR8
LINC00461
ZNF37BP
H19
LINC00162
LINC00094
OLMALINC
OIP5-AS1
SUZ12P1
RPL13AP17
LINC00472
MIR22HG
TP73-AS1
HNRNPU
TUG1
ZNF503-AS2
NEAT1
CTBP1-AS2
SNHG7
CDKN2B-AS1
MIMAT0015016
MCM3AP-AS1
CRHR1-IT1
ZFAS1
LRRC75A-AS1
LINC00339
PVT1
SNHG3
CIDECP
LINC00521
MALAT1
MIR17HG
HOXA11-AS
hsa-miR-29c-3p
hsa-miR-26b-5p
hsa-miR-25-3p
hsa-miR-23b-3p
hsa-miR-301a-3p
hsa-miR-224-5p
hsa-miR-30a-5p
hsa-miR-22-3p
hsa-miR-30e-5p
hsa-miR-21-5p
hsa-miR-3145-3p
hsa-miR-204-5p
hsa-miR-3157-5p
hsa-miR-3159
hsa-miR-19a-3p
hsa-miR-3173-3p
hsa-miR-190a-5p
hsa-miR-3192-5p
hsa-miR-185-5p
hsa-miR-320a
hsa-miR-181d-5p
hsa-miR-338-3p
hsa-miR-17-5p
hsa-miR-33a-5p
hsa-miR-15b-5p
hsa-miR-342-3p
hsa-miR-152-3p
hsa-miR-34a-5p
hsa-miR-150-5p
hsa-miR-374b-5p
hsa-miR-149-5p
hsa-miR-375
hsa-miR-148b-3p
hsa-miR-424-5p
hsa-miR-148a-3p
hsa-miR-449b-5p
hsa-miR-146b-5p
hsa-miR-454-3p
hsa-miR-143-3p
hsa-miR-455-5p
hsa-miR-130b-3p
hsa-miR-489-3p
hsa-miR-130a-3p
hsa-miR-497-5p
hsa-miR-1305
hsa-miR-503-5p
hsa-miR-10a-5p
hsa-miR-5696
hsa-miR-107
hsa-miR-605-5p
hsa-miR-106b-5p
hsa-miR-641
hsa-let-7i-5p
hsa-miR-675-5p
hsa-let-7g-5p
hsa-miR-93-5p
hsa-let-7e-5p
hsa-let-7b-5p
hsa-let-7c-5p
hsa-let-7d-5p

(b)



(c)

Figure 3. Prediction network of hot spot protein FMR1, DGCR8 and PTBP1. (a) Prediction network of hot spot protein FMR1. (b) Prediction network of hot spot protein DGCR8. (c) Prediction network of hot spot protein PTBP1
图 3. 热点蛋白 FMR1，DGCR8，PTBP1 的预测网络。(a) 热点蛋白 FMR1 的预测网络，(b) 热点蛋白 DGCR8 的预测网络，(c) 热点蛋白 PTBP1 的预测网络

图 3 显示了构建的三种热点蛋白的网络。分别是 FMR1 蛋白，DGCR8 蛋白，PTBP1 蛋白。最外层紫色六边形节点为 lncRNA，黄色圆形节点的中间层为 miRNA，最内层的中心节点为蛋白质。它们之间的关系用边表示。黑色实线表示正确预测的相互作用，红色波浪线表示预测错误的相互作用，彩色虚线表示预测的可能的新相互作用。从图 3 可以看出，LPI-MAM 模型具有预测新的相互作用的功能。例如，通过 has-miR-29a-3p 的介导，预测了 FMR1 蛋白与 CIDECP 之间的相互作用。DGCR8 蛋白与 LINC00094 的相互作用则是通过 has-miR-19a-3p 的介导预测出来的。其中，FMR1 的预突变与原发性卵巢功能衰竭和共济失调有关[26]。DGCR8 对 miRNA 前体的加工很重要[27]，它还可以促进肿瘤对 x 射线辐射的抵抗[28]。PTBP1 在剪接中起重要作用。PTBP1 还在多种疾病中发挥糖酵解、肿瘤发生、侵袭和浸润的调节作用[29]。因此，lncRNA-miRNA-蛋白质网络的构建有助于识别关键蛋白与 lncRNA 之间的关系，这对于进一步探索关键蛋白与 lncRNA 的功能及相关疾病非常有帮助。

4. 总结与讨论

本研究提出 LPI-MAM，它使用 miRNAs 作为中间体来预测 lncRPIs。通过联合 K-mer 方法对序列、结构特征信息编码，再融合 CTD 特征输入到三层 CNN 中来提取隐藏的抽象特征，然后输入 IndRNN 学习特征之间的关系。分别形成 lncRNA、miRNAs 和蛋白质三个特征子网络，最后整合输出。该方法可以在先前研究的基础上扩大 lncRPIs 的可预测范围。与 LPI-MMHN [12]、RPITER [5]、IPMiner [6]和 PRPI-SC

[14]相比，LPI-MAM 在 RAIDv2.0 和 RPI5392 数据集中综合性能是最好的。一个原因是使用 miRNAs 作为中间体扩大了 lncRPIs 的可预测范围。另一个原因是不仅考虑了序列和二级结构，还考虑了序列中每个核苷酸之间的联系，这使得特征更加丰富。最后一个原因是深度集成模型是 CNN 和 IndRNN 的结合，在学习了隐藏的高级特征之后，模型还学习了特征之间的关系。此外，通过构建可视化预测网络，表明该方法有预测新的 lncRPIs 的能力。虽然本模型已经取得了良好的性能，但其可解释性还有待进一步的探索，这也是今后下一步工作研究的方向。

参考文献

[1] Kazimierczyk, M., Kasprowicz, M.K., Kasprzyk, M.E. and Wrzesinski, J. (2020) Human Long Noncoding RNA Interactome: Detection, Characterization and Function. *International Journal of Molecular Sciences*, **21**, Article No. 1027. https://doi.org/10.3390/ijms21031027

[2] Zhao, D., Wang, C., Yan, S. and Chen, R. (2022) Advances in the Identification of Long Non-Coding RNA Binding Proteins. *Analytical Biochemistry*, **639**, Article ID: 114520. https://doi.org/10.1016/j.ab.2021.114520

[3] Marchese, D., de Groot, N.S., Lorenzo Gotor, N., Livi, C.M. and Tartaglia, G.G. (2016) Advances in the Characterization of RNA-Binding Proteins. *Wiley Interdisciplinary Reviews*. *RNA*, **7**, 793-810. https://doi.org/10.1002/wrna.1378

[4] Muppirala, U.K., Honavar, V.G. and Dobbs, D. (2011) Predicting RNA-Protein Interactions Using Only Sequence Information. *BMC Bioinformatics*, **12**, Article No. 489. https://doi.org/10.1186/1471-2105-12-489

[5] Peng, C., Han, S., Zhang, H. and Li, Y. (2019) RPITER: A Hierarchical Deep Learning Framework for ncRNA-Protein Interaction Prediction. *International Journal of Molecular Sciences*, **20**, Article No. 1070. https://doi.org/10.3390/ijms20051070

[6] Pan, X., Fan, Y.X., Yan, J. and Shen, H.B. (2016) IPMiner: Hidden ncRNA-Protein Interaction Sequential Pattern Mining with Stacked Autoencoder for Accurate Computational Prediction. *BMC Genomics*, **17**, Article No. 582. https://doi.org/10.1186/s12864-016-2931-8

[7] Xiao, Y., Zhang, J. and Deng, L. (2017) Prediction of lncRNA-Protein Interactions Using HeteSim Scores Based on Heterogeneous Networks. *Scientific Reports*, **7**, Article No. 3664. https://doi.org/10.1038/s41598-017-03986-1

[8] Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q. and Liu, H. (2018) HLPI-Ensemble: Prediction of Human lncRNA-Protein Interactions Based on Ensemble Strategy. *RNA Biology*, **15**, 797-806. https://doi.org/10.1080/15476286.2018.1457935

[9] Ge, M., Li, A. and Wang, M. (2016) A Bipartite Network-Based Method for Prediction of Long Non-Coding RNA-Protein Interactions. *Genomics Proteomics Bioinformatics*, **14**, 62-71. https://doi.org/10.1016/j.gpb.2016.01.004

[10] Wang, J., Zhao, Y., Gong, W., Liu, Y., Wang, M., Huang, X. and Tan, J. (2021) EDLMFC: An Ensemble Deep Learning Framework with Multi-Scale Features Combination for ncRNA-Protein Interaction Prediction. *BMC Bioinformatics*, **22**, Article No. 133. https://doi.org/10.1186/s12859-021-04069-9

[11] Huang, X., Shi, Y., Yan, J., Qu, W., Li, X. and Tan, J. (2022) LPI-CSFFR: Combining Serial Fusion with Feature Reuse for Predicting LncRNA-Protein Interactions. *Computational Biology and Chemistry*, **99**, Article ID: 107718. https://doi.org/10.1016/j.compbiolchem.2022.107718

[12] Zhou, Y.K., Shen, Z.A., Yu, H., Luo, T., Gao, Y. and Du, P.F. (2020) Predicting lncRNA-Protein Interactions with miRNAs as Mediators in a Heterogeneous Network Model. *Frontiers in Genetics*, **10**, Article No. 1341. https://doi.org/10.3389/fgene.2019.01341

[13] Chen, D., Wang, H., Zhang, M., Jiang, S., Zhou, C., Fang, B. and Chen, P. (2018) Abnormally Expressed Long Non-Coding RNAs in Prognosis of Osteosarcoma: A Systematic Review and Meta-Analysis. *Journal of Bone Oncology*, **13**, 76-90. https://doi.org/10.1016/j.jbo.2018.09.005

[14] Zhou, H., Wekesa, J.S., Luan, Y. and Meng, J. (2021) PRPI-SC: An Ensemble Deep Learning Model for Predicting Plant lncRNA-Protein Interactions. *BMC Bioinformatics*, **22**, Article No. 415. https://doi.org/10.1186/s12859-021-04328-9

[15] Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., Liu, L., Hou, P., Cui, T., Tan, P., Hu, Y., Zhang, T., Huang, Y., Li, X., Yu, J. and Wang, D. (2017) RAID v2.0: An Updated Resource of RNA-Associated Interactions across Organisms. *Nucleic Acids Research*, **45**, D115-D118. https://doi.org/10.1093/nar/gkw1052

[16] Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms for Molecular Biology*: *AMB*, **6**, Article No. 26. https://doi.org/10.1186/1748-7188-6-26

[17] Brown, G.R., Hem, V., Katz, K.S., *et al*. (2015) Gene: A Gene-Centered Information Resource at NCBI. *Nucleic Acids Research*, **43**, D36-D42. https://doi.org/10.1093/nar/gku1055

[18] Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: Integrating microRNA Annotation and Deep-Sequencing Data. *Nucleic Acids Research*, **39**, D152-D157. https://doi.org/10.1093/nar/gkq1027

[19] UniProt Consortium (2021) UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480-D489.

[20] Hunt, S.E., McLaren, W., Gil, L., *et al*. (2018) Ensembl Variation Resources. *Database* (*Oxford*), **2018**, bay119. https://doi.org/10.1093/database/bay119

[21] Yang, S., Wang, Y., Lin, Y., Shao, D., He, K. and Huang, L. (2020) LncMirNet: Predicting LncRNA-miRNA Interaction Based on Deep Learning of Ribonucleic Acid Sequences. *Molecules* (*Basel*, *Switzerland*), **25**, Article No. 4372. https://doi.org/10.3390/molecules25194372

[22] Geourjon, C. and Deléage, G. (1995) SOPMA: Significant Improvements in Protein Secondary Structure Prediction by Consensus Prediction from Multiple Alignments. *Computer Applications in the Biosciences*: *CABIOS*, **11**, 681-684. https://doi.org/10.1093/bioinformatics/11.6.681

[23] Yang, S., Wang, Y., Zhang, S., Hu, X., Ma, Q. and Tian, Y. (2020) NCResNet: Noncoding Ribonucleic Acid Prediction Based on a Deep Resident Network of Ribonucleic Acid Sequences. *Frontiers in Genetics*, **11**, Article No. 90. https://doi.org/10.3389/fgene.2020.00090

[24] Tong, X. and Liu, S. (2019) CPPred: Coding Potential Prediction Based on the Global Description of RNA Sequence. *Nucleic Acids Research*, **47**, e43. https://doi.org/10.1093/nar/gkz087

[25] Otasek, D., Morris, J.H., Bouças, J., Pico, A.R. and Demchak, B. (2019) Cytoscape Automation: Empowering Workflow-Based Network Analysis. *Genome Biology*, **20**, Article No. 185. https://doi.org/10.1186/s13059-019-1758-4

[26] Mila, M., Alvarez-Mora, M.I., Madrigal, I. and Rodriguez-Revenga, L. (2018) Fragile X Syndrome: An Overview and Update of the FMR1 Gene. *Clinical Genetics*, **93**, 197-205. https://doi.org/10.1111/cge.13075

[27] Pabit, S.A., Chen, Y.L., Usher, E.T., Cook, E.C., Pollack, L. and Showalter, S.A. (2020) Elucidating the Role of Microprocessor Protein DGCR8 in Bending RNA Structures. *Biophysical Journal*, **119**, 2524-2536. https://doi.org/10.1016/j.bpj.2020.10.038

[28] Hang, Q., Zeng, L., Wang, L., *et al*. (2021) Non-Canonical Function of DGCR8 in DNA Double-Strand Break Repair Signaling and Tumor Radioresistance. *Nature Communications*, **12**, Article No. 4033. https://doi.org/10.1038/s41467-021-24298-z

[29] Zhu, W., Zhou, B.L., Rong, L.J., *et al*. (2020) Roles of PTBP1 in Alternative Splicing, Glycolysis, and Oncogensis. *Journal of Zhejiang University Science B*, **21**, 122-136. https://doi.org/10.1631/jzus.B1900422