

基于三维卷积神经网络的 RNA 结构上镁离子结合位点预测

赵彦彭 巩卫康 刘洋 王京京 李春华

摘 要 目的 核糖核酸(ribonucleic acid, RNA)在许多生命过程中起着关键作用,它几乎参与了遗传信息转录和翻译的各个方面。镁离子(Mg^{2+})是 RNA 折叠成稳定三级结构所必需的,而且常常与核酶的催化活性有关。尽管实验解析的 RNA 结构越来越多,但实验确定 RNA 结构中的离子存在诸多困难。三维卷积神经网络(three-dimensional convolutional neural network, 3D-CNN)能够直接从原始数据中学习有用的特征,已被应用于一些基于结构的预测工作。为此本文拟提出一种基于 3D-CNN 的 RNA 结构上 Mg^{2+} 结合位点的预测方法 RNAMg。方法 首先收集蛋白质数据库(protein data bank, PDB)中的 RNA 结构构建了高分辨率的非冗余 RNA- Mg^{2+} 结合位点数据库(113 个结构);随后,对每个结合位点建立局部坐标系,以周围微环境(碳、氧、氮、磷原子和电荷)构建 5 个特征通道,将 RNA 三级结构作为 3D 图像送入 3D-CNN 模型预测 RNA- Mg^{2+} 结合位点;最后,使用五折交叉验证评估模型的性能。结果 独立测试集上的结果表明,在识别 RNA 结构中的 Mg^{2+} 位点上, RNAMg 优于目前最先进的方法。结论 RNAMg 可以识别出 RNA 结构中的 Mg^{2+} 结合位点,对理解 RNA 折叠、核酶催化反应等各种关键生物学过程及相关疾病的产生机制有重要意义。

关键词 RNA- Mg^{2+} 结合位点; 三维卷积神经网络; RNA 三维结构; 特征贡献; 局部坐标系

DOI: 10.3969/j.issn.1002-3208.2021.06.004.

中图分类号 R318 **文献标志码** A **文章编号** 1002-3208(2021)06-0570-07

本文著录格式 赵彦彭, 巩卫康, 刘洋, 等. 基于三维卷积神经网络的 RNA 结构上镁离子结合位点预测[J]. 北京生物医学工程, 2021, 40(6): 570-576. ZHAO Yanpeng, GONG Weikang, LIU Yang, et al. Prediction of magnesium binding sites in RNA structures based on three-dimensional convolutional neural network [J]. Beijing Biomedical Engineering 2021, 40(6): 570-576.

Prediction of magnesium binding sites in RNA structures based on three-dimensional convolutional neural network

ZHAO Yanpeng, GONG Weikang, LIU Yang, WANG Jingjing, LI Chunhua
Faculty of Environment and Life, Beijing University of Technology, Beijing 100124
Corresponding author: LI Chunhua (E-mail: chunhuali@bjut.edu.cn)

【Abstract】 Objective Ribonucleic acids (RNA) play an important role in many life processes. They take part in almost all the aspects of gene transcription and translation. Magnesium ions (Mg^{2+}) are essential for RNA folding into stable tertiary structures and are often involved in the catalytic activity of ribozymes. Despite many RNA structures have been resolved experimentally, accurately detecting them still faces many challenges experimentally. Three-dimensional convolutional neural network (3D-CNN) are capable of learning useful

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

基金项目: 国家自然科学基金(31971180、11474013)资助
作者单位: 北京工业大学环境与生命学部(北京 100124)
通信作者: 李春华。E-mail: chunhuali@bjut.edu.cn

features directly from raw data, and have been used in many structure-based prediction works. Therefore, this paper proposes a 3D-CNN based method for predicting Mg^{2+} binding sites in RNA structure, RNAMg. **Methods**

Firstly, a database of RNA-Mg²⁺ binding sites (113 RNAs) is constructed by collecting RNA structures from PDB (protein data bank) database. Then, a local coordinate system is established for each binding site, the ambient microenvironment is represented as five 'channels' (corresponding to atom types of carbon, oxygen, nitrogen and phosphorus and charge), and the tertiary structure of RNA is sent into the 3D-CNN model as a 3D image to predict the RNA-Mg²⁺ binding sites. In the end, the performance of the model is evaluated by 5-fold cross validation. **Results** Results on an independent test set show that RNAMg is superior to state-of-the-art methods in identifying Mg²⁺ binding sites in RNA structures. **Conclusions** RNAMg can identify the Mg²⁺ binding sites in RNA structures, which is of great significance in understanding various key biological processes such as RNA folding, ribozyme-catalyzed reactions and the generation mechanism of related diseases.

【Keywords】 RNA-Mg²⁺ binding site; three-dimensional convolutional neural network; RNA 3D structure; feature contribution; local coordinate system

0 引言

自从发现 RNA 可以催化化学反应以来,它在分子生物学中的重要性不断增加^[1]。现在 RNA 被认为在许多细胞调控网络中扮演关键角色^[2]。基因表达和调控的实现常常依赖于 RNA 分子特有的结构和动力学。由于 RNA 的聚阴离子性质^[3],RNA 折叠会导致大量负电荷的堆积和链内的强库仑排斥,因此 RNA 折叠会吸引溶液中的金属离子与其结合,从而降低静电能垒,使折叠后的 RNA 结构更加稳定。Mg²⁺ 在 RNA 中普遍存在,长期以来被认为是最重要的稳定 RNA 结构的离子^[3],也被认为是 RNA 能否正确折叠的关键因素。另外,Mg²⁺ 的分布也能够极大地影响 RNA 的局部结构,并导致不同的局部动力学,这一现象和 Mg²⁺ 的浓度高度相关^[4]。例如,TPP-依赖核糖开关功能的发挥不仅受该小分子的调控,还强烈依赖于溶液中 Mg²⁺ 的浓度^[5]。除了稳定结构外,Mg²⁺ 还可以帮助核酶执行催化功能^[6]。Mg²⁺ 对核酶的催化有三重作用:促使核酶折叠成活性三维结构,配合底物和核酶的官能团来定位反应物,水合 Mg²⁺ 提供或接受质子以促进催化。通过上述作用,Mg²⁺ 可提升核酶的催化效率达 10¹¹ 倍^[7]。

在过去的十年中,许多高分辨率的 RNA 结构被解析出来,特别是磁共振(nuclear magnetic resonance, NMR)和单分子荧光共振能量转移光谱技术,提供了 RNA 折叠和构象变化动力学数据^[8]。然而实验确定 RNA 结构中的离子却存在诸多困难:大多数阳离子在光谱上无法被观察到或者很难被探测到^[9];其次由于 Mg²⁺ 和 H₂O 都有 10 个电子,所以结合 Mg²⁺ 很容易被误认为是水分子;再有 Mg²⁺ 的位

置也很难用 NMR 来确定。因此开发理论预测方法,从 RNA 序列和结构信息来识别可能的 Mg²⁺ 结合位点对于理解 RNA 的结构及功能机制非常重要。

目前预测离子结合位点的工作主要集中在蛋白质上,在 RNA 上的工作很少。2016 年, Hu 等^[10] 提出了基于支持向量机的 IonCom 方法,用于预测蛋白质上常见的 13 种金属离子和酸性自由基结合位点。2021 年, Sánchez-Aparicio 等^[11] 基于金属结合位点在蛋白质骨架结构上出现的偏好性,发展了 BioMetAll 方法来预测具有特定氨基酸基序的金属结合位点。在 RNA 方面,2012 年, Philips 等^[12] 发展了 MetalionRNA 方法来预测金属离子(镁、钠和钾)所在位置,该方法基于已知结构中获得的金属离子结合位点的统计势。2018 年, Wang 等^[13] 提出 RBind 方法,使用基于结构的核苷酸相互作用网络方法预测 RNA 结合位点。RNA 的三级结构被转化为一个网络,核苷酸作为节点,它们之间的非共价相互作用作为边。然后,计算了核苷酸的度中心性(degree)和接近中心性(closeness)以确定结合位点。2021 年, Su 等^[14] 提出了基于随机森林机器学习方法的 RNAsite 模型来预测 RNA-小分子/离子结合位点,方法是将 RNA 结构的 Laplacian 范数、拓扑特性以及溶剂可及性作为特征。然而,在 RBind 与 RNAsite 方法中,金属离子结合位点预测的精度低于其他小分子结合位点。这可能是因为许多金属离子往往位于 RNA 结构的表面,并没有形成形状良好的结合口袋,使其更难预测。另一个可能的原因是训练结构以非金属离子小分子为主,这说明需要发展离子特异性方法来预测金属离子结合位点。

目前深度学习方法由于其具有能够直接从原始

数据中提取特征的优势已开始生物分子结构功能预测领域得到应用。其中,三维卷积神经网络(three-dimensional convolutional neural network, 3D-CNN)模型可以将结构转换为多个三维网格数据通道作为输入信息,能够很好处理特征的高维性、不均匀性和方向信息丢失等问题。Li 等^[15]开发了 RNA3DCNN 方法,通过将结构三维网格化,提取网格中原子的个数、质量以及电荷信息,来评价 RNA 三维结构预测的质量。Kozlovskii 等^[16]开发了 BiteNet 方法,提取结构三维网格化后的原子密度信息来识别药物结合位点。随着 RNA 结构数据的增加,基于深度学习预测金属离子结合位点成为可能。

在本研究中,课题组提出了预测 RNA 三维结构中 Mg^{2+} 结合位点的方法 RNAMg。首先建立了 RNA- Mg^{2+} 结合位点数据库,随后建立局部坐标系提取局部 RNA 结构,并用 3D 图像表示。提取原子类型和电荷量通道信息,通过 3D-CNN 模型进行学习并预测。通过独立测试集评价了模型的性能,并与最先进的方法进行比较。

1 研究方法

RNAMg 以 RNA 结构中的每个核苷酸为样本预测其是否结合 Mg^{2+} 。核苷酸及其周围一定范围的微环境被视为 3D 图像,建立 3D-CNN 模型进行训练以及预测,模型输出核苷酸是否是 Mg^{2+} 结合位点。主要工作包括数据集的建立、特征提取、3D-CNN 的架构和配置以及评价指标。

1.1 数据处理与分析

首先从蛋白质数据库(protein data bank, PDB)^[17]下载所有的共计 1 528 个只含有 RNA 分子的结构(2021 年 1 月),之后筛选符合以下条件的结构:(1) 结构分辨率优于 3.0 Å (10 Å = 1 nm); (2) 结构中至少含有一个镁离子; (3) 对 NMR 结构,保留 PDB 文件中的第一个构象; (4) 序列一致性小于 90%,对于序列一致性大于 90%的 RNA 组,只保留其中分辨率最高的结构^[12]。最后,数据库由 113 个 RNA 结构组成。

为了能够与目前最先进的方法进行比较,本工作采用了 RNAsite 中正负样本的定义:如果核苷酸任一原子与 Mg^{2+} 距离不大于截断距离(4.0 Å),则定义这个核苷酸是正样本,标签设置为 1,反之被定义为负样本,标签设置为 0。根据以上定义,从数据

库所有结构中收集到正样本 1 888 个,负样本 11 504 个。由于负样本数量远远多于正样本,课题组使用欠抽样(under-sampling)方法随机减少负样本数量从而实现样本均衡。

使用 RNAsite 独立测试集中含有 Mg^{2+} 的结构作为本工作的独立测试集。RNAsite 使用以下筛选标准:两个 RNA 结构之间的 TM-scoreRNA^[18]大于 0.3,则保留阳性样本比例较高的那一个(定义为一个 RNA 中阳性样本的总数除以序列长度),结果从 712 条 RNA 链中筛选得到了 78 条 RNA 链。这 78 个结构以 30%的序列相似性作为条件,聚类为 57 个簇。根据聚类信息对这些 RNA 进行分割:选取 42 个包含 60 个 RNA 的簇作为训练集,选取 15 个包含 18 个 RNA 的簇作为独立测试集。在独立测试集的 18 个结构中,有 4 个结构含有 Mg^{2+} 结合位点,所以课题组使用这 4 个结构作为独立测试集,结构信息见表 1,它们与本研究的 113 个 RNA 的序列一致性均小于 90%。它们被用于对模型的测试。

表 1 独立测试集中的结构

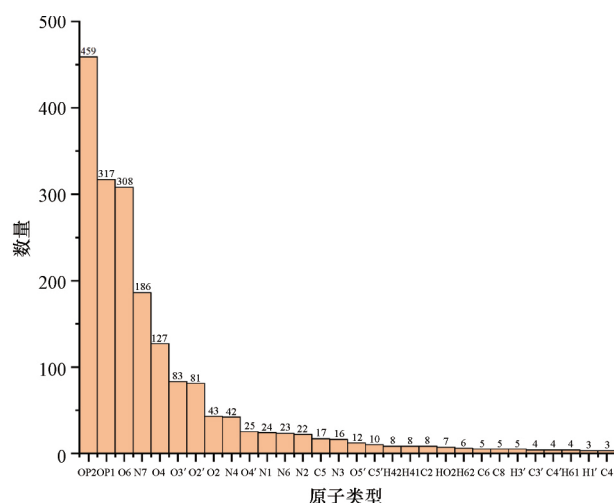
Table 1 Structures in the independent test set

PDB ID	名称	核苷酸数量	Mg^{2+} 数量
4pqv_A ^[19]	XRNI-resistant flaviviral RNA	68	7
430d_A ^[20]	SARCIN/RICIN LOOP FROM RAT 28S R-RNA	29	9
364d_C ^[21]	Fragment I from <i>E. coli</i> 5S rRNA	31	8
2mis_A ^[22]	VS Ribozyme	26	4

对训练集中的 1 888 个 Mg^{2+} 结合位点进行微环境分析,统计了最接近离子的原子类型的分布,图 1 列出了频数大于 2 的原子类型分布情况。从图 1 可以看出,频数最高的原子类型是磷酸盐上的氧原子 OP1 和 OP2,它们是带负电荷的 RNA 原子,通常需要正电荷来补偿,因此出现频数最高。相比于 OP1,OP2 更倾向与 Mg^{2+} 结合。此外,在碱基中更倾向结合离子的原子是 O6、N7 和 O4,特别是 O6 与 OP1 的结合频数十分接近。五碳糖中的 O3' 和 O2' 也有较强的结合离子的能力。

1.2 特征提取

核苷酸周围的环境指的是它的周围的原子分布。课题组建立局部笛卡尔坐标系来确定核苷酸的

图 1 结合核苷酸原子中距离 Mg^{2+} 最近的原子类型分布Figure 1 Distribution of atom types closest to Mg^{2+} in binding nucleotides

原子位置。为了获得更好的预测方法,本文构建了两种局部坐标系,分别是以核苷酸的 $C1'$ 原子和 $OP2$ 原子作为原点。使用 $C1'$ 作为原点的原因是该原子的特殊位置:连接了核苷酸的碱基与五碳糖。 $OP2$ 作为原点的原因是其最偏向与 Mg^{2+} 结合。另外,尝试了两种不同的特征组合,分别是碳、氧、磷、氮 4 种原子类型通道模型,和 4 种原子类型加电荷量的 5 通道模型。为此共构建 4 种模型:局部坐标系以 $OP2$ 为原点并使用 4 种原子类型通道($OP2+4$ 通道), $OP2$ 为原点并使用 4 种原子类型和电荷量通道($OP2+5$ 通道),以 $C1'$ 为原点并使用 4 种原子类型通道($C1'+4$ 通道), $C1'$ 为原点并使用 4 种原子类型和电荷量通道($C1'+5$ 通道)。

坐标系由 $C1'$ 、 $O5'$ 、 $C5'$ 、 $OP2$ 和嘧啶的 $N1$ 原子或嘌呤的 $N9$ 原子的位置确定。具体来说,根据局部坐标系的不同,原点位于 $C1'$ 或者 $OP2$ 所在坐标, x 轴、 y 轴和 z 轴分别用 x 、 y 、 z 表示,满足式(1)~式(3),其中 $r_{C1'}$ 、 $r_{O5'}$ 、 $r_{C5'}$ 和 r_N 表示全局坐标系中从原点指向原子 $C1'$ 、 $O5'$ 、 $C5'$ 以及 $N1$ (嘧啶)或者 $N9$ (嘌呤)的向量。

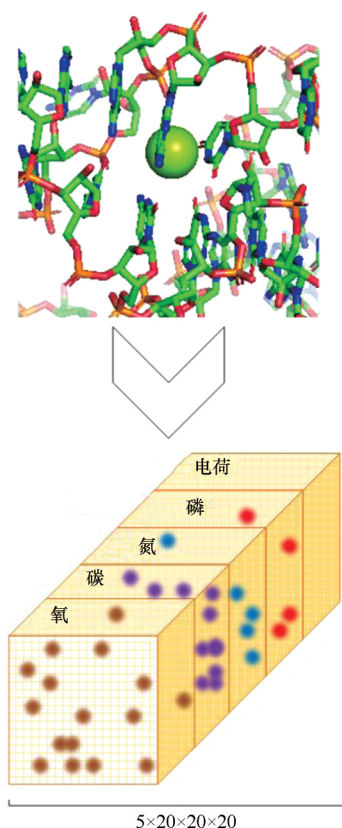
$$x = r_N - r_{C1'} \quad (1)$$

$$y = \frac{r_{O5'} + r_{C5'}}{2} - r_{C1'} \quad (2)$$

$$z = x \times y, \quad z = \frac{z}{\|z\|}, \quad x = \frac{x}{\|x\|}, \quad y = z \times x \quad (3)$$

$OP2$ 原子只作为局部坐标系原点,其 x 轴、 y 轴和 z 轴的单位向量的计算与 $OP2$ 无关,只由 $C1'$ 、 $O5'$ 、 $C5'$ 以及 $N1$ 或者 $N9$ 原子的位置决定。核苷酸周围的环境包含长度为 $20 \text{ \AA}$ 的立方体中的原子,该立方体以这个核苷酸的 $C1'$ 或者 $OP2$ 原子为中心。这样构建坐标系的目的是保持环境所在的立方体与核苷酸的方向具有一定的一致性,便于特征学习。

随后,每个边长 $20 \text{ \AA}$ 的盒子被分割成边长 $1 \text{ \AA}$ 的正方形 3D 格点,也称作体积像素(简称体素)。对每个体素,用原子类型通道记录碳、氧、磷和氮原子以及电荷量的存在。电荷量从 Gromacs 软件的 amber03 力场中提取。对于体素中的一个原子,如 N ,表示 N 的通道在其相应位置的值为 1,其他原子通道相应位置的值为 0,同时电荷量通道对应位置为该原子的电荷量。由此产生的 5 个 3D 矩阵作为输入通道堆叠在一起,如图 2 所示。需要说明的是,对于氢原子,课题组并没有建立相应的通道,因为其

图 2 Mg^{2+} 结合位点周围微环境的特征化Figure 2 Featurization of the microenvironment around a Mg^{2+} binding site

位置几乎由其他重原子的位置确定,因此它们在所建通道中已隐式表示。体素选择 $1 \text{ \AA}$ 可以确保每个体素只能容纳一个原子,从而使网格有更好的空间分辨率。这些通道数据在输入到深度学习网络之前,样本都需要进行零均值归一化。对于每种类型的通道,计算训练数据集中每种类型的通道中所有位置的平均值,并从训练、验证和测试样本中减去。

1.3 CNN

图 3 展示了 RNAMg 的架构。该神经网络有 4 个三维卷积层,每个卷积层中的滤波器数量分别为 32、64、128 以及 128,所有卷积层中滤波器的大小为 $3 \times 3 \times 3$ 体素,卷积步长设置为一个体素。在两个连续的卷积层之后放置一个步长为 2 个体素的极大池化层。随后,在卷积层之后是一个全连通层,包含 512 个隐藏单元。最后的输出层是被训练或者预测的样本即核苷酸是金属离子结合位点的概率。隐含层的所有单元均由 ReLU 非线性函数激活,输出层使用 Sigmoid 非线性函数激活。

训练神经网络的目的是减少真实标签和预测值之间的交叉熵。网络采用 Adam 算法,通过反向传播对网络参数进行优化。对所有的卷积层和全连通层添加 dropout 层和正则化层,训练的 dropout 率为 0.5。网络采用截断正态分布初始化权值,使用 128 的批次大小,运行了 200 次迭代。

1.4 性能评价

RNAMg 模型通过 5 倍交叉验证(5-fold cross-validation, 5CV)训练,并采用独立数据集测试模型的有效性。通过精确率(Precision)、召回率(Recall)和 Matthews 相关系数(Matthews correlation coefficient, MCC)以及接受者操作特性曲线(receiver

operating characteristic curve, ROC) 下面积(area under the ROC curve, AUC) 来评估分类器的预测性能,其定义如下:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

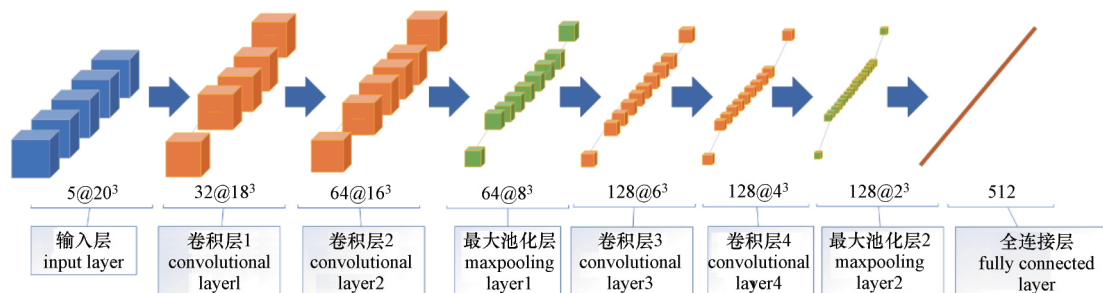
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}} \quad (6)$$

通过将每个核苷酸的预测标签与实际标签进行比较,得到真阳性(TP)、假阳性(FP)、真阴性(TN)和假阴性(FN)。其中, Precision 是指在所有模型分类为“阳性”的样本中,确实是阳性的占比; Recall 是指在所有确实为阳性的样本中,被模型分类为“阳性”的占比; MCC 描述的是实际分类与预测分类之间的相关系数,它的取值范围为 -1 到 1,取值为 1 时表示对测试对象的完美预测,取值为 0 时表示预测的结果几乎等于随机预测的结果, -1 是指预测分类和实际分类完全不一致。课题组还使用接受者操作特性曲线说明模型的性能,计算 AUC 来评估模型的准确性。AUC 分数范围为 0 到 1, AUC = 1 代表完美分类器, AUC = 0.5 时是随机分类器。

2 结果与讨论

2.1 特征贡献分析

4 种模型在五折交叉验证中的性能如表 2 所示,其中 OP2+5 通道模型的 Precision、Recall、MCC 和 AUC 结果为 0.661、0.668、0.325 以及 0.724,比 C1'+5 通道的方法分别提高了 4.0%、24.6%、



(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. <http://www.cnki.net>
图 3 3D-CNN 网络结构

Figure 3 3D-CNN architecture

表 2 考虑不同特征组合与不同局部坐标系原点的 3D-CNN 模型进行 5CV 的结果

Table 2 Average results of 5CV of 3D-CNN models with different feature combinations and different local coordinate system origins

方法	Precision	Recall	MCC	AUC
OP2+5 通道	0.661 ± 0.041	0.668 ± 0.059	0.325 ± 0.046	0.724 ± 0.025
OP2+4 通道	0.658 ± 0.021	0.671 ± 0.027	0.321 ± 0.029	0.723 ± 0.019
C1'+5 通道	0.635 ± 0.045	0.536 ± 0.095	0.224 ± 0.039	0.654 ± 0.018
C1'+4 通道	0.640 ± 0.027	0.505 ± 0.036	0.226 ± 0.033	0.659 ± 0.019

注: OP2 指局部坐标系原点为 OP2 原子, C1' 指局部坐标系原点为 C1' 原子; 4 通道指使用 C、O、N 以及 P 4 种原子类型作为通道提取特征, 5 通道指使用 C、O、N 以及 P 4 种原子类型加电荷量作为通道提取特征。

45.1%以及 10.7%。这说明选择 OP2 作为局部坐标系的原点更能突出离子结合位点与非结合位点周围微环境的差别。另一方面, OP2+5 通道模型在 Precision、MCC 和 AUC 3 个评价指标上都略优于 OP2+4 通道模型, 只有 Recall 稍稍低于 OP2+4 通道模型的结果。说明电荷量与原子类型特征在 Mg^{2+} 结合位点预测方面是互补的, 增加电荷量通道有助于 RNAMEg 更准确地预测金属离子结合位点。总体来说, 以 OP2 为局部坐标系原点, 使用碳、氧、磷、氮原子类型+电荷量的 5 通道模型(OP2+5 通道) 取得了最佳的结果。

2.2 模型在独立测试集上的表现

选择 OP2+5 通道模型作为最终的预测模型, 并在独立测试集上测试该模型, 结果如表 3 所示。此外, 表 3 还对比展示了 Wang 提出的预测工具 RBind 以及 Su 提出的预测服务器 RNAsite 在独立测试集的测试结果。从表 3 看出本方法在所有评价指标上都略优于另两个方法。RNAMEg 方法的 Precision、Recall、MCC 和 AUC 均高于 RBind 以及 RNAsite 方法的相应结果。尤其在 MCC 上, 指标提升明显。综上所述, 在 Mg^{2+} 结合位点预测上, RNAMEg 模型好于 RBind 与 RNAsite。课题组认为这主要归功于 3D-CNN 能够直接从初始原子分布中学习更多有用的特征。

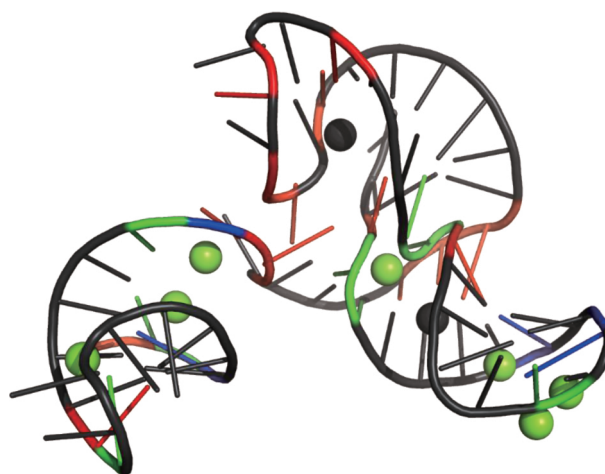
表 3 RNAMEg 与其他方法在独立测试集上的比较

Table 3 Comparison between RNAMEg and other methods on the independent test set

方法	Precision	Recall	MCC	AUC
RNAMEg	0.521	0.568	0.345	0.654
RNAsite	0.420	0.213	0.058	0.577
Rbind	0.338	0.488	0.109	0.560

2.3 样例分析

图 4 显示了在长度为 68 nt 的 4PQV_A 上的 Mg^{2+} 预测结果。该结构有 9 个 Mg^{2+} , 7 个离子距离 RNA 结构在 4 Å 以内, 其中的 5 个 Mg^{2+} 分别与 2 个核苷酸结合, 2 个 Mg^{2+} 分别与 1 个核苷酸结合, 所以共有 12 个结合位点。RNAMEg 成功预测出了 12 个结合位点中的 8 个。RNAMEg 在这个结构上得到的 Precision、Recall、MCC 和 AUC 分别为 0.363、0.667、0.340 和 0.749。



黑色表示 RNA 结构, 绿色、红色和蓝色分别表示 TP、FP 和 FN。

图 4 4PQV_A 结构上预测的 Mg^{2+} 结合位点Figure 4 Mg^{2+} ion binding site prediction on 4PQV_A structure

3 结论

课题组发展了一种有效预测 RNA 结构中 Mg^{2+} 结合位点的理论新方法 RNAMEg。该方法通过构建基于待预测核苷酸的局部坐标系, 从已知 RNA 结构

中提取离子结合位点周围原子分布以及电荷量作为特征,将 RNA 结构作为 3D 图像,使用三维卷积神经网络模型进行预测。通过 5 倍交叉验证测试找到了表现最好的局部坐标系以及特征组合,并证明 RNAMg 可以准确预测 Mg^{2+} 结合位点。

目前 PDB 数据库中具有离子结合的 RNA 结构数据还非常有限,这阻碍了深度学习预测金属离子结合位点的性能。随着实验解析 RNA 结构的增多,离子结合位点的数据会不断增长,可以预计深度学习预测模型会不断得到优化,模型的预测能力将不断提升。在独立测试集上的与其他最先进方法的对比证明, RNAMg 可以相对准确地预测 Mg^{2+} 在 RNA 上的位置,其预测能力优于最先进的预测服务器 RBind 以及 RNAsite,是一种实用的工具。这项工作对理解 RNA 折叠、核酶催化反应等各种关键生物学过程及相关疾病的产生机制有重要意义。

参考文献

- [1] Doudna JA, Cech TR. The chemical repertoire of natural ribozymes[J]. *Nature* 2002 418(6894): 222–228.
- [2] Morris KV, Mattick JS. The rise of regulatory RNA[J]. *Nature Reviews. Genetics* 2014 15(6): 423–437.
- [3] Bowman JC, Lenz TK, Hud NV, et al. Cations in charge: magnesium ions in RNA folding and catalysis[J]. *Current Opinion in Structural Biology* 2012 22(3): 262–272.
- [4] Wang J, Xiao Y. Types and concentrations of metal ions affect local structure and dynamics of RNA[J]. *Physical Review E*, 2016 94(4-1): 040401.
- [5] Yamauchi T, Miyoshi D, Kubodera T, et al. Roles of Mg^{2+} in TPP-dependent riboswitch[J]. *FEBS Letters* 2005 579(12): 2583–2588.
- [6] Schnabl J, Sigel RKO. Controlling ribozyme activity by metal ions[J]. *Current Opinion in Chemical Biology* 2010 14(2): 269–275.
- [7] Herschlag D, Cech TR. Catalysis of RNA cleavage by the tetrahymena-thermophila ribozyme .1. Kinetic description of the reaction of an RNA substrate complementary to the active-site[J]. *Biochemistry* 1990 29(44): 10159–10171.
- [8] Stellos K, Gatsiou A, Stamatiopoulos K, et al. Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation[J]. *Nature Medicine* 2016 22(10): 1140–1150.
- [9] Cruz-León S, Schwierz N. Hofmeister series for metal-cation-RNA interactions: the interplay of binding affinity and exchange kinetics[J]. *Langmuir* 2020 36(21): 5979–5989.
- [10] Hu X, Dong Q, Yang J, et al. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals[J]. *Bioinformatics* 2016 32(21): 3260–3269.
- [11] Sánchez-Aparicio JE, Tiessler-Sala L, Velasco-Careros L, et al. BioMetAll: identifying metal-binding sites in proteins from backbone preorganization[J]. *Journal of Chemical Information and Modeling* 2021 61(1): 311–323.
- [12] Philips A, Milanowska K, Lach G, et al. MetalionRNA: computational predictor of metal-binding sites in RNA structures[J]. *Bioinformatics* 2012 28(2): 198–205.
- [13] Wang K, Jian Y, Wang H, et al. RBind: computational network method to predict RNA binding sites[J]. *Bioinformatics* 2018, 34(18): 3131–3136.
- [14] Su H, Peng Z, Yang J. Recognition of small molecule-RNA binding sites using RNA sequence and structure[J]. *Bioinformatics* 2021 37(1): 36–42.
- [15] Li J, Zhu W, Wang J, et al. RNA3DCNN: local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks[J]. *PLoS Computational Biology* 2018, 14(11): 1–18.
- [16] Kozlovskii I, Popov P. Spatiotemporal identification of druggable binding sites using deep learning[J]. *Communications Biology*, 2020 3(1): 618.
- [17] Berman HM, Westbrook J, Feng Z, et al. The protein data bank[J]. *Nucleic Acids Research* 2000 28(1): 235–242.
- [18] Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA[J]. *Bioinformatics* 2019 35(21): 4459–4461.
- [19] Chapman EG, Costantino DA, Rabe JL, et al. The structural basis of pathogenic subgenomic flavivirus RNA (sRNA) production[J]. *Science* 2014 344(6181): 307–310.
- [20] Correll CC, Munishkin A, Chan YL, et al. Crystal structure of the ribosomal RNA domain essential for binding elongation factors[J]. *Proceedings of the National Academy of Sciences of the United States of America* 1998 95(23): 13436–13441.
- [21] Correll CC, Freeborn B, Moore PB, et al. Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain[J]. *Cell* 1997 91(5): 705–712.
- [22] Bonneau E, Legault P. NMR localization of divalent cations at the active site of the *neurospora* VS ribozyme provides insights into RNA – metal-ion interactions[J]. *Biochemistry* 2014 53(3): 579–590.

(2021-03-02 收稿 2021-04-16 修回)