PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2023, 50(3): 657~667
www.pibb.ac.cn

基于注意力机制的RNA碱基关联图预测方法*

曹一航　黄　强**
（复旦大学生命科学学院，上海 200438）

摘要　目的　长链非编码RNA在遗传、代谢和基因表达调控等方面发挥着重要作用。然而，传统的实验方法解析RNA的三级结构耗时长、费用高且操作要求高。此外，通过计算方法来预测RNA的三级结构在近十年来无突破性进展。因此，需要提出新的预测算法来准确的预测RNA的三级结构。所以，本文发展可以用于提高RNA三级结构预测准确性的碱基关联图预测方法。**方法**　为了利用RNA理化特征信息，本文应用多层全卷积神经网络和循环神经网络的深度学习算法来预测RNA碱基间的接触概率，并通过注意力机制处理RNA序列中碱基间相互依赖的特征。**结果**　通过多层神经网络与注意力机制结合，本文方法能够有效得到RNA特征值中局部和全局的信息，提高了模型的鲁棒性和泛化能力。检验计算表明，所提出模型对序列长度L的4种标准（L/10、L/5、L/2、L）碱基关联图的预测准确率分别达到0.84、0.82、0.82和0.75。**结论**　基于注意力机制的深度学习预测算法能够提高RNA碱基关联图预测的准确率，从而帮助RNA三级结构的预测。

关键词　深度学习，RNA碱基关联图，结构预测，注意力机制
中图分类号　Q812　　**DOI**：10.16476/j.pibb.2022.0241

RNA作为生物大分子在遗传代谢、细胞表达调控等方面作为调控因子发挥生物学功能[1-2]。RNA分子通过三级结构形成的空间构象与其他分子发生相互作用或反应。因此，解析其结构能够更有效地研究它与其他分子的相互作用机制。RNA分子的三级结构包括所有碱基原子的空间坐标及其在三维空间中的相互作用关系[3]。三级结构的相互作用主要包括共价键的相互作用、氢键相互作用、范德华力及其他非键相互作用[4]。目前，冷冻电镜技术[5]是获得RNA分子三级结构的常用手段。尽管该方法能得到精确的三级结构，但是价格昂贵、实验周期长等缺点限制了该技术的大规模使用。这促使研究人员开发基于计算的预测算法来快速地预测RNA的三级结构。

RNA三级结构的预测算法主要可分为两类，基于先验知识的预测算法和基于核酸分子物理化学特征的预测算法。其中，基于先验知识的三级结构预测方法包括碎片组装算法[6]和同源序列比较算法[7-8]。前者将已有的RNA三级结构按照不同的标准切割成碎片再重新组合，例如按照二级结构、原子的三维空间坐标或内部碱基间的接触概率[9-10]；后者利用模版序列间的协同进化信息来预测目标序列的三级结构。基于核酸分子物理化学特征的预测算法通过计算最低的自由能构象来预测RNA分子的三级结构[11]，该方法通常采用蒙特卡罗模拟退火算法或分子动力学模拟算法[12]，利用动态规划方法迭代来得到目标序列的三级结构，例如SWA、FARFAR、FARNA等算法。研究表明，优化RNA内部碱基间的接触概率矩阵能显著提高碎片组装算法的预测准确率[3]。此外，内部碱基间的接触概率矩阵能提升直接耦合分析的准确率，这也有助于核酸的结构预测[13-14]。因此，RNA碱基间的接触问题受到了密切的关注。

以往研究发现，蛋白质残基接触的准确性能够影响蛋白质三级结构的预测结构。通过蛋白质残基的接触绘制接触概率矩阵，并将该矩阵称为关联图。近年来，机器学习算法已经广泛应用于蛋白质

* 国家重大科技专项“重大新药开发”课题（2018ZX09J18112）和国家自然科学基金（31971377）资助项目。
** 通讯联系人。
Tel：021-31246589，E-mail：huangqiang@fudan.edu.cn
收稿日期：2022-05-26，接受日期：2022-07-11

的关联图预测中，例如：Li等[15]提出残差神经网络来预测蛋白质的关联图；RaptorX[16]将进化偶联算法和序列保守信息整合至深度学习算法来预测蛋白质关联图；R2C算法[17]将支持向量机用于关联图的预测并得到了较好的准确率。此外，深度学习算法也已广泛应用于RNA变异位点的分类[18]、RNA结合蛋白[19]以及RNA三级结构[20]的预测中。然而，深度学习算法在RNA关联图的预测中进展缓慢。目前，仅有RNAcontact[21]将深度学习算法用于RNA关联图的预测。虽然众多研究已经表明基于深度学习的预测算法往往优于传统的计算方法，但是RNAcontact仅采用深度卷积神经网络来提取特征值，该方法无法提取序列间的特征值，因此预测准确率还有待提高。

针对现有的深度学习方法不能处理RNA序列中碱基间相互依赖的特征，本文提出了一种基于注意力机制的深度学习预测模型（命名为ATTcontact），用于预测RNA的关联图。其中，用文本注意力机制提取RNA序列的特征值，而图像注意力机制则用于提取RNA二级结构、位置特异性矩阵和协方差矩阵的特征值。首先，ATTcontact的运行仅需要RNA的序列信息和多序列比对文件。其次，为了处理特征值间对输出的影响，模型采用了长短期记忆网络（long short-term memory networks，LSTM）、卷积神经网络（convolutional neural networks，CNN）、注意力机制和全连接层进行学习。最后，通过特征值间维度的转换，用全连接层预测每个碱基间的接触概率。检验计算结果表明，ATTcontact模型优于基于进化偶联算法和以往的深度学习预测方法。

1 数据与方法

1.1 数据集

首先，从Protein Data Bank（PDB）库中下载具有3D结构的单链RNA序列，共得到了729个PDB结构文件。随后，去除同源度高于90%的RNA序列。为了优化运行速度，本文去除了序列长度大于300个碱基的PDB结构。预处理之后，用余下的649个PDB结构建立数据集，利用DSSR[22]软件提取每个PDB结构所对应的二级结构，并从649条序列随机选取500条序列作为模型的训练集，74条序列为验证集，75条序列为测试集。

根据Weinreb等[14]的工作以及蛋白质关联图的相关研究，当RNA中碱基上的任意一个原子与另一个碱基的任意原子距离小于8 Å时，设定这两个碱基在三级结构上接触（设碱基对的关联值为1），否则，就认为它们在三级结构中没有发生接触（设碱基对的关联值为0）。基于这个判定标准，本研究把RNA序列中每个碱基之间的接触关联特征转换为$L\times L$大小的矩阵（L是RNA序列的长度）。因此，碱基间的接触问题可以视为二分类问题，可以应用深度学习算法进行预测。

1.2 深度学习网络模型

本研究利用图1的深度学习模型来训练RNA数据集。模型结构包括CNN、LSTM、残差神经网络（residual net，ResNet）[23]和注意力机制[24]。注意力机制包括基于文本的注意力机制和基于图像的注意力机制。此外，注意力机制整合进残差神经网络中来提取RNA特征值中更深维度的信息。RNA序列经过独热编码（one-hot encoding）的处理后直接与文本注意力机制相连，随后用3层LSTM充分挖掘RNA的序列信息。第一部分的ResNet包含着3个一维的卷积层并以LSTM的输出作为输入。卷积层的卷积核分别为2、3、3，ReLU函数作为激活函数。该部分的网络引入0.02的丢弃率来避免过拟合。序列特征值经过一系列非线性变换后得到输出，并经过扩维运算后转换为三维张量与另外3个特征值合并，形成$L\times L\times 4$的张量并作为第二部分神经网络的输入。在第二部分，每个ResNet中都包含了二维的卷积层和图像注意力机制。同样，每个卷积层的卷积核都是2，并用ReLU函数作为激活函数。模型采用0.02的丢弃率和标准化来避免模型的过拟合。在模型的最后，加入了两层全连接层，第一层采用ReLU激活函数，并使用0.02的丢弃率和标准化来避免过拟合，最后一层全连接层只有一个节点，用Sigmoid函数作为激活函数，将模型最后的输出转化为［0，1］之间的概率。

模型构建采用Google公司的TensorFlow[25]框架（V2.16）。网络使用Adam优化算法，并将Cross-entropy作为损失函数来进行模型的优化。研究过程中使用了Nvidia GTX3080图形处理器来加速模型的优化速度。

1.3 输入值

完成RNA三级结构数据集的构建后，生成RNA序列的特征矩阵。特征矩阵越体现RNA的理化信息，预测结果就越准确。根据Sun等[21]的研究，有多种参数信息对RNA三级结构的预测有影响，包括碱基的排列顺序、RNA的二级结构、溶

剂可及表面积、碱基间的相对位置、协同进化信息等。在本研究中，由于溶剂可及表面积等参数需要利用其他软件进行预测，这一步骤往往会引入错误的预测结果，使关联图的预测过程把错误结果视为真实值进行学习，从而影响模型的准确性。基于上述原因，本研究把RNA序列、位置特异性矩阵、二级结构以及基于协同进化信息得到的协方差矩阵作为神经网络的特征值。

a. RNA序列。从PDB结构文件中提取RNA的碱基，并根据碱基的种型把每个碱基都转换为one-hot向量，即M_sequence∈R^(L×4)。为了处理不同长度的RNA序列，本文将RNA序列的长度设置为300，未满300个核苷酸的RNA序列用0向量填充至300，以便于模型运算。

b. 二级结构（secondary structure，SS）。对于已有PDB结构文件的RNA序列，利用DSSR软件分析其三级结构的特征来得到准确的二级结构，并把二级结构转化为L×L大小的矩阵，M_secondary∈R^(L×L)。当碱基互补配对时该位置处的元素为1，否则为0。对于没有PDB结构文件的RNA序列，则利用SPOT-RNA软件预测其二级结构并经过上述相同的方法得到二维的特征矩阵。同样地，本文将未满300个碱基的矩阵用0元素进行填充，并对矩阵进行扩维，得到L×L×1大小的三维矩阵。

c. 协方差矩阵（covariance matrix，Cov）。从NCBI数据库中下载FASTA参考数据集，并利用BLASTN软件将待测RNA序列与参考数据集进行比较，最后用Muscle软件[26]得到多序列比对文件。为了得到更具有特异性的序列，本文去除了具有80%相似性的序列和空位超过50%的序列。最后，使用pydca软件[27]中的MeanField算法计算序列中每个碱基与另一个碱基之间的协方差，得到协方差矩阵M_Cov∈R^(L×L)。对该矩阵进行填充和扩维操作后得到L×L×1大小的三维矩阵。

d. 位置特异性矩阵（position specific scoring matrix，PSSM）。采用动态规划算法计算RNA序列中每个碱基的相对位置。得分规则如下：

$$\text{if seq}[i]=\text{seq}[j],\ A_{i,j}=A_{i-1,j-1}+1 \tag{1a}$$

$$\text{if seq}[i]\neq\text{seq}[j],\ A_{i,j}=A_{i-1,j-1}+1 \tag{1b}$$

其中，i和j分别是碱基在序列中的位置，seq[i]（seq[j]）是在i(j)位置处的碱基，$A_{i,j}$是第i行第j列时矩阵的元素。基于上述计算方法构建了位置特异性矩阵，M_pssm∈R^(L×L)，并对该矩阵进行相同的填充和扩维操作。

最后，模型将经过独热编码后的RNA序列作为第一部分网络的输入，即经过文本注意力机制、LSTM和一维的CNN等非线性计算后进行扩维操作，得到L×L×1的张量。此外，二级结构、协方差矩阵和位置特异性矩阵等特征值合并为L×L×3的矩阵张量，并与第一部分网络的输出一起作为深度学习模型第二部分的输入。

1.4　注意力机制

模型包含文本注意力机制[28]和图像注意力机制[29]。文本注意力机制主要从二维的RNA序列信息中提取特征值，并与LSTM相连。图像注意力机制包含了通道注意力机制和空间注意力机制，并与卷积神经网络整合，来提取RNA二级结构、协方差矩阵、位置特异性矩阵和RNA序列的高维特征值。图像注意力机制分别在特征值的通道和大小中寻找对输出权重较大的部分特征。

在文本注意力机制（self-attention）中，模型将一维的输入经过非线性计算变化为2个矩阵，Query（查询值）和Key（键），并设置Value（值）与Key相同。即对于每一个序列中的碱基X_i，$Q_i=\sigma(X_iW)$，$K_i=\sigma(X_iW)$，其中，X_i∈R^(1×4)，W∈R^(4×m)，Q_i∈R^(1×m)，K_i∈R^(1×m)，σ为非线性激活函数。对Q_i与K_i进行矩阵的点积运算并用Softmax将计算结果标准化。最后，将V_i与Softmax函数对计算结果相乘，即

$$Z_i=\text{Softmax}(Q_i\otimes K_i)\times V_i \tag{2}$$

其中⊗是矩阵的点积运算。上述计算过程可总结为：

$$\text{Attention}(Q,K,V)=\text{Softmax}(QK^T)V \tag{3}$$

公式（3）中，K^T是K的转置矩阵。

本文将通道注意力机制和空间注意力机制整合形成卷积模块注意力机制（convolutional block attention module，CBAM）。该模块能从多维的空间矩阵中提取三维矩阵的特征值信息。因此，CBAM模块能够有效学习三维特征数据并得到每个特征值对输出结果的贡献权重。公式（4a）和（4b）表示CBAM模块的计算过程。

$$F'=M_C(F)\otimes F \tag{4a}$$

$$F''=M_S(F')\otimes F' \tag{4b}$$

公式（4a）中F是一个中间层的三维矩阵，M_C是通道注意力；公式（4b）中M_S是空间注意力，

⊗是矩阵的点积运算。

1.4.1 通道注意力机制

通道注意力机制使用平均池化和最大池化的降维方法，利用ReLU函数进行非线性变化，分别得到代表通道的张量。在经过Sigmoid函数进行非线性操作后，得到通道注意力矩阵。公式如下：

$$M_C(F)=\sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(F)\right)+\mathrm{MLP}\left(\mathrm{MaxPool}(F)\right)\right) \quad (5)$$

式中σ是Sigmoid函数，MLP是进行全连接层计算，AvgPool和MaxPool分别代表平均池化和最大池化的降维操作，F是中间层的三维矩阵，$F\in$ R^（$L\times L\times n$）。

1.4.2 空间注意力机制

空间注意力机制也使用最大池化计算和平均池化计算得到两个描述不同特征的值，并将这两个特征合并进行卷积运算从而产生二维的空间注意力矩阵。公式如下：

$$M_S(F)=\sigma\left(f^{3\times3}\left[\mathrm{AvgPool}(F),\mathrm{MaxPool}(F)\right]\right) \quad (6)$$

式中$f^{3\times3}$代表在进行卷积运算时利用了3×3的卷积核。

1.5 效果评估

类似于蛋白质内部残基间接触概率的预测，本文对RNA内部碱基间的接触概率预测同样由精确率（Precision）、召回率（Recall）、准确率（Accuracy）等指标进行计算。具体如下：

$$Precision=\frac{TP}{TP+FP} \quad (7a)$$

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \quad (7b)$$

$$Recall=\frac{TP}{TP+FN} \quad (7c)$$

$$F1(\text{调和平均})=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (7d)$$

其中：*TP*是真阳性，表示预测的碱基配对结果与实验结构的值都为阳性，即两个碱基预测的原子距离小于8 Å，在实验结构也小于8 Å；*FP*是假阳性，表示预测的值为阳性，而实验结构的值为阴性，即两个碱基预测的原子距离小于8 Å，而真实大于8 Å；*TN*是真阴性，即预测的碱基配对结果与实验结构的值都为阴性，即两个碱基预测的原子距离大于8 Å，而真实也大于8 Å；*FN*是假阴性，表示预测的值为阴性，而实验结构的值为阳性，即两个碱基预测的原子距离大于8 Å，而真实大于8 Å。一般来说，top *L*/*n*（*n*=1，2，5，10）表示在该范围内模型预测得到的最高准确率。

此外，为了进一步评价ATTcontact的预测效果，本文还构建了两个评价指标，受试者工作特征曲线（receiver operating characteristic curve，ROC）和精确率-召回率曲线（precision-recall curve，PRC）。其中，ROC上各点反映着每个待测软件相同的特性，即它们都是对同一输入数据的不同反应。在几种不同的判定标准下，不同的待测软件往往会得到不同的反应结果。ROC曲线是指在特定的阈值条件下，模型在不同判断标准下所得的各个信号的连线。通常情况下，ROC曲线覆盖的面积即AUC（area under curve）值越大，模型的预测效果越好。其横轴为假阳率（false positive rate，FPR），纵轴为真阳率（true positive rate，TPR），即：

$$FPR=\frac{TP}{\mathrm{TP}+\mathrm{FN}} \quad (8a)$$

$$TPR=\frac{FP}{\mathrm{FP}+\mathrm{TN}} \quad (8b)$$

另外，PRC的横轴为召回率，纵轴为精确率，因此PRC表示精确率与召回率的关系。模型在0~1的范围内设置阈值区间并设置频率得到一定数量的阈值，即每次模型的输出与该阈值进行比较，当模型的输出值大于阈值时定义为正样本，当模型的输出值小于阈值时定义为负样本。基于上述理论，本研究得到与模型输出相关的精确率和召回率。当曲线越靠近右上角，召回率和精确率越大。

2 结　果

2.1 算法概览

模型在LSTM和CNN的基础上添加了注意力机制，用于计算对模型输出结果更为重要的部分特征（图1）。模型的输入是RNA序列、二级结构、位置特异性矩阵以及协方差矩阵。首先，序列提交给SPOT-RNA和Muscle，分别获得序列的二级结构和多序列比对文件。随后，用pydca软件比较多序列比对文件得到协方差矩阵。最后，分别把特征值输入至一维的神经网络和二维的神经网络中，来预测所有碱基间的接触概率。

2.2 不同注意力机制的预测结果

为了探究注意力机制对RNA关联图预测的影响，本实验维持其他参数（模型的层数、深度、卷积核、优化函数以及学习率等）不变，分别比较ATTcontact、保留文本注意力机制而不采用图像注

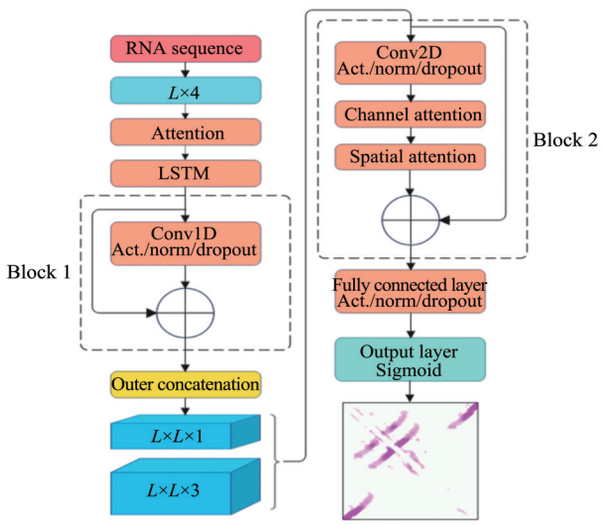


Fig. 1　Framework of neural network model

After extracting features, the model is trained by the features and predict the contact map.

意力机制的模型、保留图像注意力机制而不采用文本注意力机制的模型，以及不采用注意力机制的模型来预测测试集。由表1可知，当模型不采用注意力机制，而仅使用CNN和LSTM时，模型在*L*/10、*L*/5、*L*/2、*L*的范围内分别达到了0.82、0.82、0.80、0.75的准确率。当模型采用了注意力机制后，模型的准确率得到了提高，在*L*/10、*L*/5、*L*/2和*L*的范围内分别达到了0.84、0.82、0.82和0.75的准确率。此外，文本注意力机制未能有效提高模型的准确率，而图像注意力机制能够相对提高模型的准确率，在*L*/10、*L*/5、*L*/2和*L*的范围内分别达到了0.83、0.80、0.79和0.75的准确率。由此，认为图像注意力机制的插入对RNA关联图预测有更重要的贡献。由于模型的二级结构不是在自然状态下的真实值，因此在关联图预测时ATTcontact可能将假阳性的碱基配对视为真实的二级结构从而影响关联图的预测。因此，模型提高的准确率远比表格显示的大。

Table 1　Comparison of accuracies on different attentional models

Parameters	*L*/10	*L*/5	*L*/2	*L*
CBAM+Self-attention	0.84	0.82	0.82	0.75
CBAM	0.83	0.80	0.79	0.75
Self-attention	0.82	0.82	0.80	0.75
Non-attention	0.82	0.82	0.80	0.75

2.3　不同预测方法比较

为了比较ATTcontact的性能，分别用RNAcontact、PLMC[14]和RNAcmap[30]预测测试集。RNAcontact利用了传统的深度学习方法来预测RNA的关联图，仅将卷积神经网络整合入残差神经网络，来提高模型提取特征值的能力。PLMC和RNAcmap则利用了协同进化方法预测RNA的关联图。本部分实验把RNA序列作为RNAcontact的输入，并使用它的默认参数设置得到待测RNA的关联图。对于PLMC和RNAcmap，将测试集中的RNA序列利用BLAST软件在参考数据集中寻找同源序列，并将多序列比对文件作为上述软件的输入。表2说明了ATTcontact在关联图预测中有着最高的准确率，明显优于其他软件。在测试数据集中，ATTcontact在L/10、L/5、L/2和L中分别达到了0.84、0.82、0.82和0.75的准确率，比RNAcontact高了4%、2%、12%和14%。此外，与传统的协同进化方法ATTcontact与PLMC相比，ATTcontact有着巨大的提升。

Table 2　Comparison of accuracies on different software in test sets

Softwares	Dataset	*L*/10	*L*/5	*L*/2	*L*
ATTcontact	Rfam	0.84	0.82	0.82	0.75
RNAcontact	Rfam	0.80	0.80	0.70	0.61
PLMC	Rfam	0.67	0.61	0.44	0.31
RNAcmap	Rfam	0.08	0.13	0.08	0.15

为了进一步和RNAcontact比较软件性能，ROC曲线和PRC作为评价指标来展示在不同阈值条件下模型的预测结果（图2）。图2a表示在测试集中两种软件在不同情况下的ROC曲线和AUC值。由图可知，ATTcontact的AUC值为0.796，而RNAcontact的AUC值为0.776，这表明ATTcontact略微优于RNAcontact。用真实的二级结构作为ATTcontact二级结构特征值的输入，模型的AUC值达到了0.868。这一结果不仅证明了模型在预测RNA关联图时，准确的二级结构对模型的预测结果至关重要。这一结果也证明了模型的稳定性和鲁棒性。图2b表示在测试集中两种软件在不同情况下的PRC。同样地，真实的二级结构也作为特征值输入模型中进行预测。由图可知，相较于RNAcontact，ATTcontact取得了更好的结果（蓝色曲线比红色曲线更靠近右上角）。当ATTcontact利用了真实的二级结构进行预测时，结果也明显优于

预测的二级结构。这一结果不仅补充证明了ATTcontact的准确性，也反应了准确的二级结构对ATTcontact的预测结果的影响。

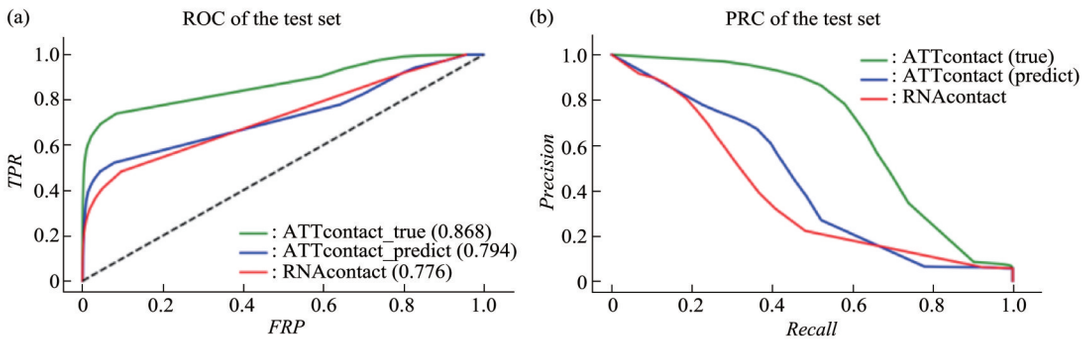


Fig. 2　Comparison of ATTcontact and RNAcontact

(a) The ROC of the test set. (b) The Precision-Recall curve of the test set.

2.4　特征值影响

为了探究不同特征值对模型的影响，本部分实验保持一维特征值（RNA序列）不变，ATTcontact在不同的二维特征值组合下的预测结果（表3）。结果表明，当只有1个二维特征输入时，二级结构对模型具有最重要的贡献，在*L*/10、*L*/5、*L*/2和*L*范围内预测的准确率分别达到0.82、0.82、0.80和0.80。而协方差矩阵和位置特异性矩阵作为单独的二维特征值输入时，模型的准确率远低于二级结构。将不同的特征值进行组合，结果证明特征值的组合对模型的性能具有协同作用。例如，当协方差矩阵和二级结构组合时，模型分别在*L*/10、*L*/5、*L*/2，*L*的范围中分别达到了0.82、0.81、0.78、0.74。当输入3个特征值时，模型达到最高的准确率。这表明模型能从复杂的输入特征中提取到更复杂的信息。此外，把真实的二级结构作为输入时，模型在*L*/10、*L*/5、*L*/2、*L*的范围中分别达到了1.00、0.99、0.98和0.96。因此，ATTcontact具有良好的鲁棒性及稳定性。

计算结果表明，RNA的二级结构对RNA碱基关联图具有最重要的影响，其次为协方差矩阵，最后为位置特异性矩阵。这是因为RNA碱基在三维空间中发生接触，最普遍的是发生在两个碱基互补配对的位置。二级结构附近几个碱基虽然未形成碱基对，但在三维空间中的原子产生接触，从而在整体上维持着RNA三级结构的稳定。协方差矩阵在*L*/10的范围内也得到了较好的结果。这表明模型在处理局部类似的RNA序列时，能够从多序列比对文件中分析出对结果有作用的碱基间配对信息。

Table 3　Comparison of accuracies on different feature combinations

Features	*L*/10	*L*/5	*L*/2	*L*
PSSM	0.35	0.34	0.25	0.23
SS	0.82	0.82	0.80	0.75
Cov	0.60	0.37	0.24	0.18
PSSM+SS	0.83	0.81	0.79	0.75
PSSM+Cov	0.72	0.65	0.53	0.44
Cov+SS	0.82	0.81	0.78	0.74
Cov+SS+PSSM	0.84	0.82	0.82	0.75
Cov+SS (True)+PSSM	1.00	0.99	0.98	0.96

2.5　预测实例

本部分实验分别比较3个不同的RNA序列（PDB ID分别为1P50、3F2Y、2A64）来展示ATTcontact和RNAcontact的差异。图3~5分别表示在*L*的范围下，不同的RNA序列的关联图及三维结构图。在每个图中，（a）为ATTcontact利用真实的二级结构预测到的关联图；（b）为ATTcontact利用预测的二级结构得到的关联图；（c）为该条RNA在三维空间中的展示，红色部分表示对三级结构有重要作用的非二级结构配对，蓝色部分表示对三级结构有重要影响的碱基配对；（d）表示RNAcontact利用预测的二级结构得到的关联图。

图3为3F2Y在不同条件下的关联图。在本示例中，该条RNA序列是一条有107个核苷酸的序列。为了比较ATTcontact和RNAcontact的准确率，

二级结构的预测值作为这两个软件的输入。由图3a中蓝圈和红圈部分可知，ATTcontact能够成功预测第40以及100位左右的碱基。而RNAcontact则不能成功预测，只能预测到序列两端的配对。当采用二级结构的真实值时，正如图3a中红圈和蓝圈部分所示，ATTcontact不仅能够得到与其整体结构相匹配的碱基间配对，还预测出二级结构中未形成配对的碱基间接触。

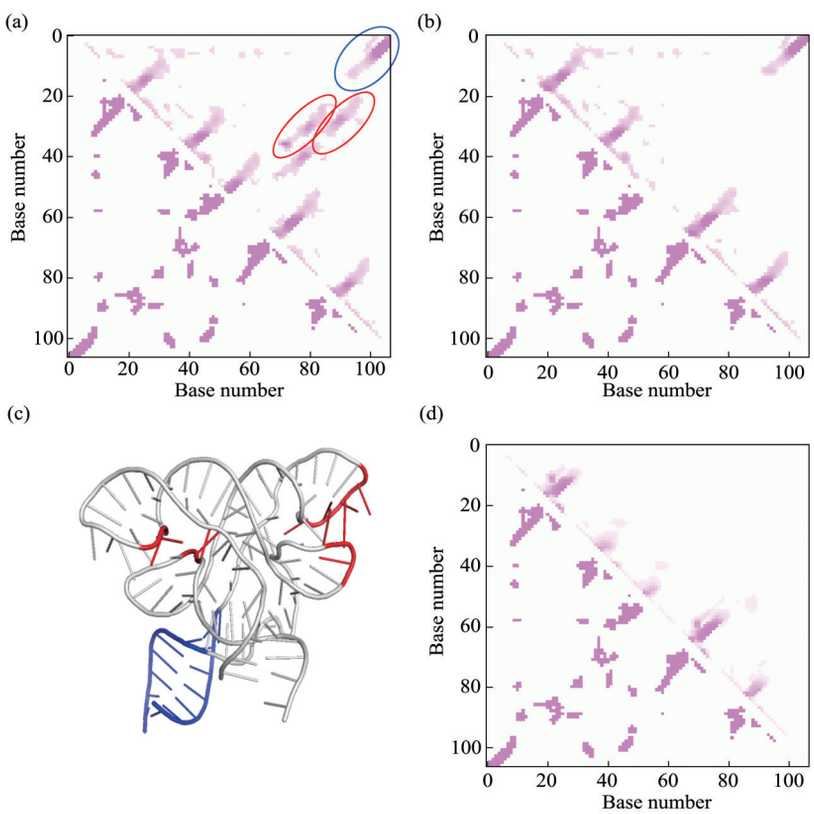


Fig. 3　Contact map and 3D structure of 3F2Y

(a) The contact map of ATTcontact by native secondary structure. (b) The contact map of ATTcontact by predicted secondary structure. (c) 3D structure of 3F3Y by PyMOL. (d) The contact map of RNAcontact by predicted secondary structure.

图4为2A64在不同条件下的关联图。本例的RNA序列是一条有297个核苷酸的RNA序列。为了比较ATTcontact和RNAcontact，同样选择了二级结构的预测值作为RNAcontact的输入。相较于RNAcontact的关联图，ATTcontact预测到了更多、更准确的碱基配对信息，更能反映该条序列在三维空间中碱基的接触信息，而RNAcontact仅预测到了两端碱基配对。当采用二级结构的真实值时，正如图4a中红圈和蓝圈部分所示，ATTcontact不仅能够得到与其整体结构相匹配的碱基间配对，还预测出二级结构中未形成配对的碱基间接触。

图5为1P50在不同条件下的关联图。本例的RNA序列是一条有74个核苷酸的RNA序列。为了比较ATTcontact和RNAcontact，同样用二级结构的预测值作为这两个软件的输入。由图可知，ATTcontact能够准确预测出序列两端及短距离内的碱基配对，而RNAcontact仅预测到了两端的碱基配对，且引入了大量的假阳性。把二级结构的真实值作为输入时，正如图5a中红圈和蓝圈部分所示，ATTcontact不仅能够得到与其整体结构相匹配的碱基间配对，还预测出二级结构中未形成配对的碱基间接触。

综上所述，RNAcontact往往遗漏很多对RNA三级结构有帮助的碱基配对能看出来，ATTcontact能够准确的预测在三维结构折叠中具有重要作用的碱基配对。

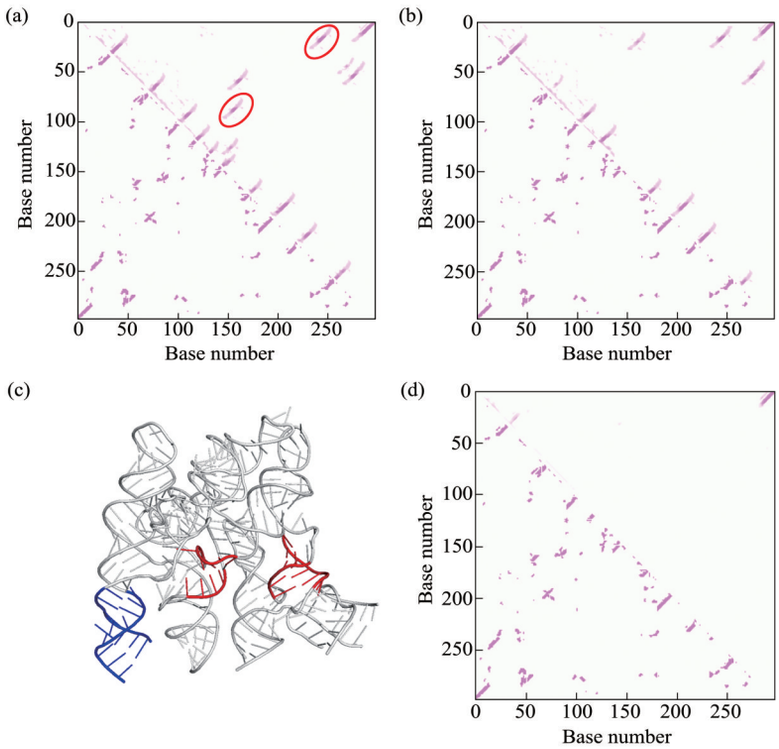


Fig. 4　Contact map and 3D structure of 2A64

(a) The contact map of ATTcontact by nature secondary structure. (b) The contact map of ATTcontact by predicted secondary structure. (c) 3D structure of 2A64 by PyMOL. (d) The contact map of RNAcontact by predicted secondary structure.

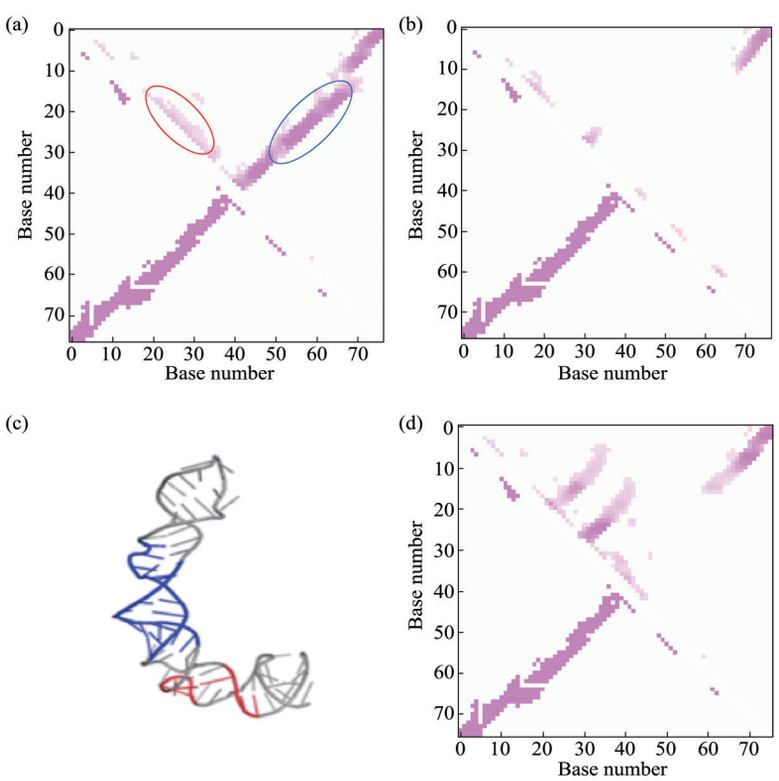


Fig. 5　Contact map and 3D structure of 1P50

(a) The contact map of ATTcontact by nature secondary structure. (b) The contact map of ATTcontact by predicted secondary structure. (c) 3D structure of 1P50 by PyMOL. (d) The contact map of RNAcontact by predicted secondary structure.

3 结　论

本文提出了一种基于注意力机制的预测方法来构建RNA的关联图。该方法利用CNN和LSTM进行特征值提取。此外，注意力机制用于计算对关联图预测有显著影响的部分特征值，并利用残差神经网络深度提取特征值的信息。CNN和LSTM用于提取特征值间复杂的信息。这些网络结构能够有效处理特征值之间的关系以及对碱基配对的影响。计算结果表明本文的模型优于其他软件，这证明了基于注意力机制的深度学习模型在未监督的特征值提取中达到了最优的效果，也证明了注意力机制模型的解码作用能广泛用于RNA序列的处理。

RNA关联图的准确性能够提高RNA三级结构，通过加入文本注意力机制以及图像注意力机制提高了RNA关联图预测的准确性。文本注意力机制的优势是将RNA序列看作由字符串组成的文字，并解码为中间矩阵，并得到与标签最为密切的部分。而空间注意力机制和通道注意力机制则分别在不同的维度对三维的RNA特征值进行处理，同样得到图片中与标签相关的信息。本文混合了深度学习中的语义分析和计算机视觉，利用最先进的计算机技术处理生物学难题，并得到了良好的结果。这不仅仅为RNA结构预测提供了新的思路，还提供了一种新的技术方法来预测RNA的关联图。

然而，本研究还有一些不足，如序列长度的限制。ATTcontact对RNA序列的输入要求小于300个碱基对。这一点可以通过增加显存或改善数据集的存储方式来实现。此外，神经网络中过多的节点有可能会产生过拟合现象。尽管模型采用了各种措施来避免过拟合，过拟合现象仍然是神经网络需要克服的问题。针对这一问题，可以通过改善网络结构或改变网络的训练目标来缓解模型的过拟合问题。例如，能将深度学习与能量参数结合从而降低网络中的参数从而避免过拟合。

致谢　感谢中国人民解放军海军军医大学基础医学院王梁华教授、药学院陆峰教授对本研究的指导和帮助。

参　考　文　献

[1] Morris K V, Mattick J S. The rise of regulatory RNA. Nat Rev Genet, 2014, **15**(6): 423-437

[2] Corra F, Agnoletto C, Minotti L, *et al*. The network of non-coding RNAs in cancer drug resistance. Front Oncol, 2018, **8**:327

[3] Leonardis D E, Lutz B, Ratz S, *et al*. RNA secondary and tertiary structure prediction by tracing nucleotide co-evolution with direct coupling analysis. Biophys J, 2016, **110**(3): 364a

[4] Wang J, Zhao Y J, Zhu C Y, *et al*. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. Nucleic Acids Res, 2015, **43**(10): e63

[5] Henderson R. The potential and limitations of neutrons, electrons and X-rays for atomic-resolution microscopy of unstained biological molecules. Q Rev Biophys, 1995, **28**(2): 171-193

[6] Massire C, Westhof E. MANIP: an interactive tool for modelling RNA. J Mol Garph Model, 1998, **16**(4-6): 197-205

[7] Rother M, Rother K, Puton T, *et al*. ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Res, 2011, **39**(10): 4007-4022

[8] Flores S C, Altman R B. Turning limited experimental information into 3D models of RNA. RNA, 2010, **16**(9): 1769-1778

[9] Antczak M, Popenda M, Zok T, *et al*. New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure. Acta Biochim Pol, 2016, **63**(4): 737-744

[10] Wang J, Mao K K, Zhao Y J, *et al*. Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. Nucleic Acids Res, 2017, **45**(11): 6299-6309

[11] Liu Z D, Yang Y R, Li D Y, *et al*. Prediction of the RNA tertiary structure based on a random sampling strategy and parallel mechanism. Front Genet, 2022, **12**: 813604

[12] Takahashi S, Sugimoto N. Stability prediction of canonical and non-canonical structures of nucleic acids in various molecular environments and cells. Chem Soc Rev, 2020, **49**(23): 8439-8468

[13] Leonardis D E, Lutz B, Ratz S, *et al*. Direct-Coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic Acids Res, 2015, **43**(21): 10444-10455

[14] Weinreb C, Riesselman A J, Ingraham J B, *et al*. 3D RNA and functional interactions from evolutionary couplings. Cell, 2016, **165**(4): 963-975

[15] Li Z, Lin Y L, Elofsson A, *et al*. Protein contact map prediction based on ResNet and DenseNet. Biomed Res Int, 2020, **2020**: 7584968

[16] Wang S, Sun S, Li Z, *et al*. Accurate *de novo* prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol, 2017, **13**(1): e1005324

[17] Yang J, Jin Q Y, Zhang B, *et al*. R2C: improving *ab initio* residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. Bioinformatics, 2016, **32**(16): 2435-2443

[18] Song Z, Huang D, Song B, *et al*. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat Commun, 2021, **12**(1): 4011

[19] Du B, Liu Z Y, Luo F L. Deep multi-scale attention network for RNA-binding proteins prediction. Inform Sci, 2022, **582**: 287-301

[20] Townshend R J L, Eismann S, Watkins A M, *et al*. Geometric deep learning of RNA structure. Science, 2021, **373**(6558): 1047-1051

[21] Sun S S, Wang W K, Peng Z L, *et al*. RNA inter-nucleotide 3D closeness prediction by deep residual neural networks. Biomed Res Int, 2021, **37**(8): 1093-1098

[22] Lu X J, Bussemaker H J, Olson W K, *et al*. DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic Acids Res, 2015, **43**(21): e142

[23] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition//IEEE. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770-778

[24] Niu Z Y, Zhong G Q, Yu H. A review on the attention mechanism of deep learning. Neurocomputing, 2021, **452**: 48-62

[25] Abadi M, Barham P, Chen J, *et al*. TensorFlow: a aystem for large-scale machine learning// USENIX Association. Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. Savannah, GA: USENIX Association, 2016: 265-283

[26] Pei J. Multiple protein sequence alignment. Curr Opin Struc Biol, 2008, **18**(3): 382-386

[27] Zerihun M B, Pucci F, Peter E K, *et al*. pydca v1.0: A comprehensizve software for direct coupling analysis of RNA and protein sequences. Bioinformatics, 2020, **36**(7): 2264-2265

[28] Ayana, Shen S Q, Lin Y K, *et al*. Recent advances on neural headline generation. J Comput Sci Tech, 2017, **32**: 768-784

[29] Wang F, Jiang M, Qian C, *et al*. Residual attention network for image classification//IEEE. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6450-6458

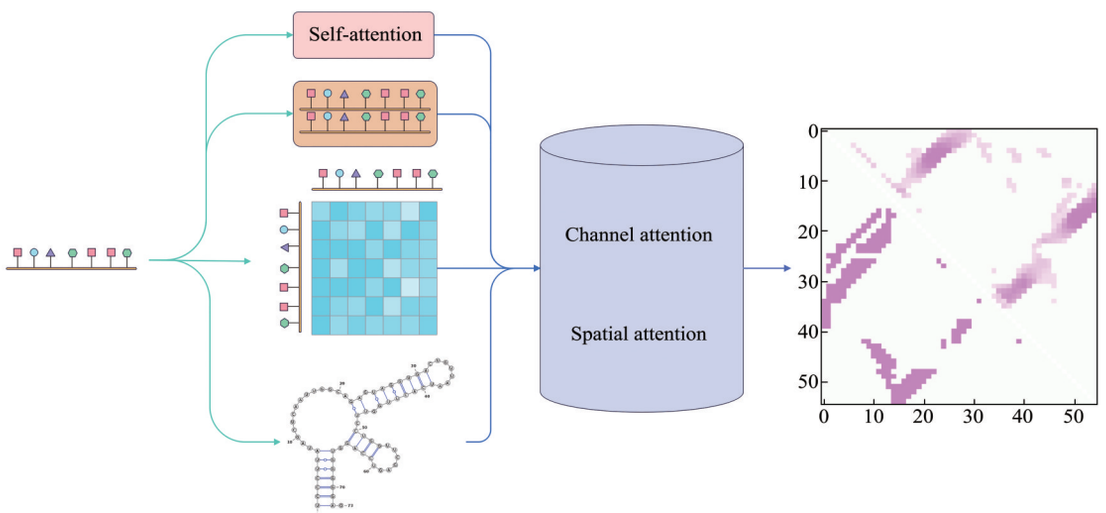
[30] Zhang T C, Singh J, Litfin T, *et al*. RNAcmap: a fully automatic pipeline for predicting contact maps of RNAs by evolutionary coupling analysis. Bioinformatics, 2021, **37**(20): 3494-3500

Prediction Method of RNA Contact Map Based on Attention Mechanism*

CAO Yi-Hang, HUANG Qiang**

(*School of Life Sciences, Fudan University, Shanghai* 200438, China)

Graphical abstract



Abstract　**Objective**　Long non-coding RNA play an important role in genetics, metabolism and gene expression regulation. But it is time-consuming and costly to analyze the RNA structure by experimental approaches. However, prediction software based on co-evolutionary algorithm has not made breakthrough progress in prediction accuracy in recent ten years. Therefore, it is necessary to propose a new prediction algorithm to accurately predict the tertiary structure of RNA. So, this paper develops prediction method of base contact map of RNA that can be used to improve the accuracy of tertiary structure prediction. **Methods**　To utilize the physical and chemical characteristics of RNA, we propose a deep learning algorithm based on multi-layer convolutional neural network and long short-term memory networks to predict the contact map between base pair. In addition, we employ attention mechanism to deal with complex global spatial independence features in RNA sequences. **Results**　By combining multilayer neural networks with the attention mechanism, our method can effectively obtain local and global information in RNA features, which improves the robustness and generalization ability of the model. The computations show that the proposed model achieves 0.84, 0.82, 0.82 and 0.75 prediction accuracies for the base contact map of 4 criteria (L/10, L/5, L/2, L) of sequence length L. **Conclusion**　Prediction method based on attention method is better than traditional computational methods and common deep learning algorithms, respectively.

Key words　deep learning, RNA contact map of base pair, structure prediction, attention mechanism

DOI：10.16476/j.pibb.2022.0241

* This work was supported by grants from the National Major Scientific and Technology Special Project for “Significant New Drugs Development”（2018ZX09J18112）and The National Natural Science Foundation of China（31971377）.

** Corresponding author.

Tel：86-21-31246589，E-mail：huangqiang@fudan.edu.cn

Receiced：May 26，2022　Accepted：July 11，2022