带有蛋白质输入的 RNA-蛋白质结合位点预测方法

梅 杰 何如吉 吕 强
(苏州大学计算机科学与技术学院 江苏 苏州 215006)
(江苏省计算机信息处理技术重点实验室 江苏 苏州 215006)

摘 要 RNA 及 RNA 结合蛋白之间的相互作用在基因调控中扮演着重要角色。许多预测 RNA-蛋白质结合位点的深度学习方法陆续提出。目前多数研究没有将 RNA 结合蛋白作为模型输入,限制了深度学习模型的规模。对此问题,提出一个带有 RNA 结合蛋白输入的深度学习模型,通过扩大训练集的规模挖掘 RNA-蛋白质结合位点的公共知识。模型将 RNA 序列先后经过卷积神经网络和门控循环单元来得到序列特征;将序列特征与 RNA 结合蛋白的独热编码拼接,作为全连接层的输入;通过一个 Sigmoid 单元输出该 RNA 结合蛋白对 RNA 序列的结合概率。在两个权威数据集上,该方法相比其他模型均具有一定优势。

关键词 RNA RNA 结合蛋白 RNA-蛋白质结合位点 深度学习

中图分类号 TP3 Q811.4 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.007

MODEL WITH PROTEIN AS INPUT FOR PREDICTING RNA-PROTEIN BINDING SITES

Mei Jie He Ruji Lü Qiang
(*School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China*)
(*Jiangsu Province Key Laboratory for Information Processing Technologies, Suzhou 215006, Jiangsu, China*)

Abstract The interactions between RNAs and RNA binding proteins are crucial for gene regulation. A wide range of deep learning methods have been proposed for predicting RNA-protein binding sites. Till now, most of them do not take RNA binding protein as input, which restricts the scale of deep learning model. We propose a deep learning model whose input includes RNA binding protein, which enlarges the scale of training set and mines meta information from RNA-protein binding sites. The model fed the RNA sequence into convolutional neural network and gated recurrent unit to extract sequence feature. It took the concatenation of sequence feature and RNA binding protein in the format of one hot encoding as the input of fully connected layer. A sigmoid unit was used to output the binding probability of RNA binding protein to RNA sequence. This method has advantages over other models on two authoritative data sets.

Keywords RNA RNA binding protein RNA-protein binding site Deep learning

0 引 言

RNA 和 RNA 结合蛋白(RNA Binding Proteins,RBP)的交互作用是理解转录后调控机制的关键,对蛋白质合成、基因融合和可变 mRNA 加工具有广泛的影响[1-3]。RNA-蛋白质结合位点预测是指仅以 RNA 作为模型输入,并为每一个 RBP 训练一个模型用于预测 RBP 是否结合于输入的 RNA。得益于高通量测序技术的高速发展如 CLIP-Seq[4],数以百计的 RBP 对应的大量 RNA-蛋白质结合位点得以发现[5-8]。因此,通过机器学习方法预测 RNA 上的 RNA-蛋白质结合位点成为了当前的研究热点。其中深度学习相比传统机器学习方法由于无需特征工程即可获得良好的性能,近年

收稿日期:2020-02-01。国家自然科学基金项目(31801108);苏州大学江苏省计算机信息处理重点实验室开放课题(KJS1843);江苏高校优势学科建设工程资助项目。**梅杰**,硕士生,主研领域:生物信息学。**何如吉**,硕士生。**吕强**,教授。

来被广泛应用到此问题上。

DeepBind 第一个将卷积神经网络(Convolutional Neural Network,CNN)用于提取 RNA 序列特征,在当时取得了突破性的进展[9]。随后,沈红斌课题组的系列模型(iDeep[10]、iDeepM[11]、iDeepA[12]、iDeepV[13]、iDeepE[14]、iDeepS[15] 和 CRIP[16])及 Deepnet-rbp[17]、mmCNN[18]、CircSLNN[19]等模型运用深度学习方法对 RNA-蛋白质结合位点预测问题进行了广泛而深入的研究,包括长短时记忆网络[20](Long Short-Term Memory,LSTM)、残差神经网络[21](Residual Network,ResNet)及注意力机制[22](Attention Mechanism)等方法都陆续被使用。尽管如此,这些方法都没有考虑将 RBP 本身作为模型的输入之一从而进一步扩大数据集并挖掘不同 RNA-蛋白质结合位点问题的联系。

从更高的视角来看,不同于 RNA-蛋白质结合位点问题,RNA-蛋白质交互作用对问题同时需要 RNA 和 RBP 以一定形式作为输入。由于同时获取 RNA 和 RBP 数据的成本高昂,有限的数据量使得端到端的深度学习方法仍不能有效应用于这一问题[23]。而尽管高通量测序技术可以获得单个 RBP 在特定细胞系和组织下的大量 RNA-蛋白质结合位点,但将不同体内环境下的 RNA-蛋白质结合位点进一步整合并构建更大的数据集可以进一步发挥深度学习模型的优势。另一方面,模型通过对其他 RNA-蛋白质结合位点数据的学习可能挖掘出与自身有关的知识。如在命名实体识别任务中,BioNER 通过整合不同类型实体的数据集取得了性能的提升[24]。因此,本文提出一个整合不同 CLIP 数据的模型,并将 RBP 实验编号以独热编码的形式作为模型的输入之一用来区别 RNA 序列将被哪个 RBP 结合。

在评估该模型效果时,将该模型在两个 RNA-蛋白质结合位点预测的权威数据集上与其他模型进行对比,结果表明该模型在这两个数据集上相比其他模型均具有一定优势。

1　算法介绍

本文提出的模型将不同 RBP 对应的实验数据合并作为本模型的数据集,将 RNA 序列和 RBP 的实验编号作为输入并最终输出对两者结合概率的预测,模型结构如图 1 所示。

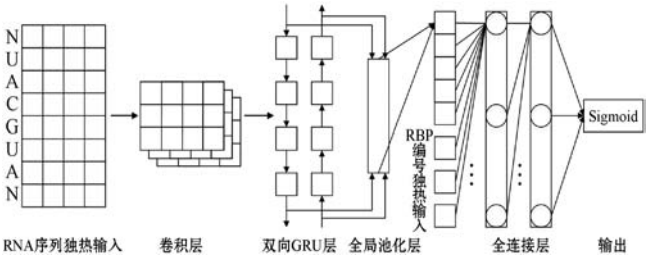

图 1　模型结构

RNA 序列以独热编码的形式表示,对于序列不等长的数据集,取训练集中序列的最大长度 n 作为序列的输入长度,并对长度不足的序列两端以 N 补齐,其中 $N=[0.25,0.25,0.25,0.25]$。RBP 实验编号的输入向量宽度与训练集中的实验总数 m 一致,如 RBP 实验编号为 0 的独热编码表示为第 0 位为 1 而其他 $m-1$ 位均为 0 的向量。

将 RNA 序列的独热编码作为卷积层的输入。第 k 个卷积核对齐到 RNA 序列位置 i 的输出如式(1)所示。

$$conv(s)_{i,k}=\sum_{j=1}^{l}\sum_{b=1}^{4}\boldsymbol{S}_{i+j,b}\boldsymbol{M}_{k,j,b} \tag{1}$$

式中:$\boldsymbol{S}$ 是 RNA 序列的独热编码表示,它是一个 $n\times4$ 的矩阵;$\boldsymbol{M}_k$ 表示第 k 个卷积核的权重矩阵;b 取值为 1 到 4,表示 A、U、C 和 G 四种碱基,l 表示卷积核长度;$1\leqslant i\leqslant n-l+1$ 且 $1\leqslant k\leqslant f$,其中 f 指卷积核的数量。按以上步骤依次计算 f 个卷积核对 RNA 序列的输出,则能得到一个大小为 $n\times f$ 的矩阵,即为 CNN 的输出。一个卷积核就相当于一个特征选择器,这里卷积运算用于学习 RNA 序列的局部特征,类比于图像处理任务中的卷积运算得到图像特征。

CNN 层卷积处理后,应用修正线性单元(Rectified Linear Unit,ReLU)激活函数和批量归一化层处理[25](Batch Normalization,BN)。ReLU 可以对 CNN 的输出进行非线性形变,批量归一化层则可以加速模型收敛,并在一定程度上避免过拟合。

然后使用双向门控神经单元[26](Bidirectional Gated Recurrent Unit,Bi-GRU)进一步提取 RNA 序列的全局特征。考虑到 RBP 对 RNA 序列的结合在生物学上并没有一定的方向,所以这里使用双向的设计。

一个 GRU 对于 t 时的输入 x_t 按式(2)-式(5)进行运算。而 Bi-GRU 包含正反两个方向的 GRU,它们在 t 时的输出按式(6)合并。

$$\boldsymbol{z}_t=\sigma(\boldsymbol{W}_z\times[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]) \tag{2}$$

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_r\times[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]) \tag{3}$$

$$\tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{W}\times[\boldsymbol{r}_t\odot\boldsymbol{h}_{t-1},\boldsymbol{x}_t]) \tag{4}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t \tag{5}$$

式中：⊙指逐位乘；σ 指 Sigmoid 层；$\boldsymbol{z}_t$ 和 $\boldsymbol{r}_t$ 分别是 GRU 的更新门和重置门；$\boldsymbol{W}$、$\boldsymbol{W}_z$ 和 $\boldsymbol{W}_r$ 分别是 $\boldsymbol{h}_t$、$\boldsymbol{z}_t$ 和 $\boldsymbol{r}_t$ 的权重矩阵；$\boldsymbol{h}_{t-1}$ 用于记忆前一时刻输入的隐状态；$\tilde{\boldsymbol{h}}_{tt}$ 则用来更新 $\boldsymbol{h}_t$。

$$\overleftrightarrow{\boldsymbol{h}}_t = \overrightarrow{\boldsymbol{h}}_t \oplus \overleftarrow{\boldsymbol{h}}_t \tag{6}$$

式中：$\overrightarrow{\boldsymbol{h}}_t$ 和 $\overleftarrow{\boldsymbol{h}}_t$ 分别是正向 GRU 和反向 GRU 在 t 时输出的 $\boldsymbol{h}_t$；$\overleftrightarrow{\boldsymbol{h}}_t$ 是 Bi-GRU 在 t 时刻的输出；⊕表示逐位加。

Bi-GRU 的输出经过全局最大池化层，得到了代表序列信息的特征向量。将特征向量与 RBP 实验的独热编码拼接起来，作为一个两层均带 Dropout[27] 的全连接层的输入，它的输出经过一个 Sigmoid 单元得到一个 0 到 1 之间的预测值，表示 RNA 序列在 RBP 实验中的结合概率。

2 实验与结果分析

为了评估本文模型的性能，本文选择在 RNA-蛋白质结合位点问题的两个权威数据集（RBP-24 和 RBP-31）上与其他模型进行对比。

2.1 实验数据与评价指标

RBP-24 可在 GraphProt[28] 处下载，它由 24 个 CLIP 实验组成并包含了 21 个不同的 RBP。24 个实验中每个实验的数据量不同，将 24 个实验的训练集合并并打乱后作为模型的训练集，总计包含约 120 万个样本，训练集序列的长度范围在 38 个碱基对到 375 个碱基对之间。由于 CLIP 数据仅包含正样本（即结合序列），GraphProt 通过打乱结合序列顺序的方式提供了数量相当的负序列。

RBP-31 可在 iONMF[29] 处下载，它由 31 个 CLIP 实验组成并包含了 19 个 RBP。iONMF 为每个 CLIP 实验提供了划分好的三组交叉验证数据，每组数据的训练集为 30 000 条，测试集为 10 000 条。iONMF 选择使用基因组上未被任何 RBP 结合的序列作为负序列，训练集和测试集的正负样本比例均为 1∶4，序列长度为固定的 101 个碱基对。

两组数据均以 AUC（Area Under the ROC Curve）作为评价指标，见式（7）。

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{7}$$

式中：M 是正样本的数量；N 是负样本的数量；*positiveClass* 是正样本的集合；通过对正样本预测值进行排序，正样本的最小预测值对应为 $rank_1$，以此类推 $rank_i$。

2.2 实验环境与模型参数

实验机器硬件配置为：CPU 为两块 Intel（R）Xeon（R）CPU E5-2620 v4 @ 2.10 GHz，GPU 为三块 GeForce GTX 1080 Ti（每次训练只使用一块），内存大小为 128 GB。模型是由 Keras 2.2.2 以 TensorFlow 1.9.0 为后端（backend）编程实现的。

本模型在 RBP-24 和 RBP-31 数据集上的主要参数设置分别如表 1 所示。权重和偏置均为 Keras 2.2.2 的默认设置。本模型的损失函数为交叉熵损失函数（CrossEntropy Loss），优化器选择了 Adam[30]，学习率初始值是 0.001。同时，本文还使用了早停技术及时中断训练，同时设置了检查点来保存验证损失最小的模型。

表 1 模型主要超参数

参数名	数据集	
	RBP-24	RBP-31
RNA 输入 shape	(375,4)	(101,4)
RBP 编码输入 shapec	(24)	(31)
卷积核核数	80	
卷积核步长	9	
GRU 单元数	200	
第一层全连接层宽度	400	
第一层 Dropout 率	0.25	
第二层全连接层宽度	220	
第二层 Dropout 率	0.25	
Batch size	1 000	

2.3 RBP-24 数据集上的实验结果

本模型在 RBP-24 测试集上计算的 AUC 分布与其他模型对比的结果如图 2 所示，对比模型的结果均来源于公开发表的论文。其中仅以序列作为输入而不使用其他外源数据（如 RNA 的二/三级结构、region type 和 clip-cobinding 等）的模型均以 * 标记，是对比的主要对象。特别要说明的是，Deepnet-rbp 在原文中同时提供了仅以 RNA 序列作为输入的结果，以及以 RNA 序列和 RNA 三级结构作为输入的结果，本文使用的是前者。图 2 中竖线表示模型在 RBP-24 测试集上 24 个 RNA-蛋白质结合位点预测的 AUC 的平均值；圆圈为异常值，表示模型在某个测试集上的表现显著低于其他测试集。

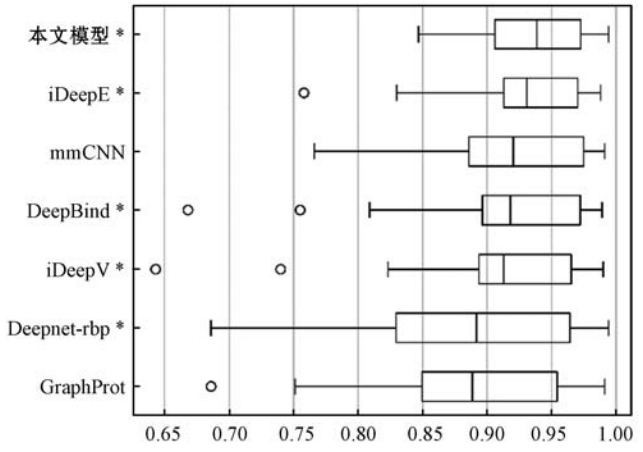

图 2　不同模型在 RBP-24 测试集上的 AUC 分布

可以看出,本模型的预测结果整体分布较好,下界和均值分别是 84.7% 和 93.9%,分别比 iDeepE 高出了 8.9 百分点和 0.8 百分点。

2.4　RBP-31 数据集上的实验结果

本模型在 RBP-31 测试集上计算的 AUC 分布和其他模型对比的结果如图 3 所示。可以看到,使用传统机器学习方法的 GraphProt 和 iONMF 尽管使用了 RNA 序列之外的数据源,相比深度学习方法仍然没有优势。而在所有仅以 RNA 序列作为输入的模型中,本模型的平均 AUC 为 87.3% 排名第一,比 DeeperBind[31] 高出 1.6 百分点,甚至比额外使用了 RNA 结构信息的 iDeepS 还要高 1.2 百分点,与 iDeepS 的成对 t-检验的单尾 p-value 远小于 1 百分点,具有显著差异性。而 iDeep 使用了 RNA 序列信息、结构信息、region type motif 及 clip-cobinding 作为模型输入,相比本文模型仍然有较大优势。

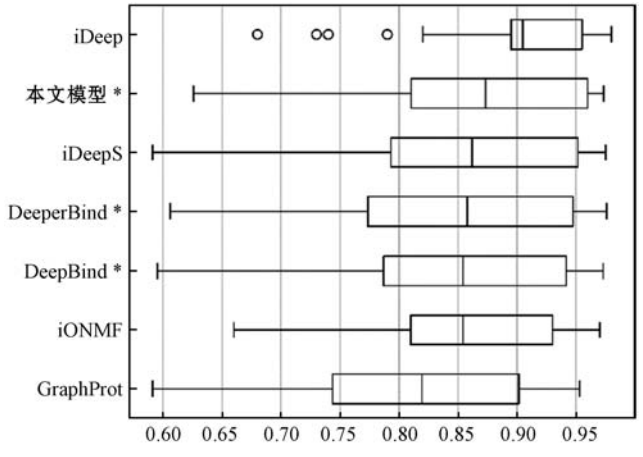

图 3　不同模型在 RBP-31 测试集上的 AUC 分布

2.5　结果分析

从 RBP-24 和 RBP-31 的对比结果来看,本文模型相比现有的仅以 RNA 序列作为输入的模型具有一定优势。本文在处理 RNA 序列信息时使用了独热编码 + CNN + Bi-GRU 的结构,这与 DeeperBind 的独热编码 + CNN + LSTM 及 iDeepS 的独热编码 + CNN + Bi-LSTM 均较为相似。但是,本文模型在 RBP-31 上却取得了更好的结果,这说明了由于不同的 RNA-蛋白质结合位点问题中存在着公共知识,具有生物学上的相关性,通过对其他 RNA-蛋白质结合位点任务的学习确实带来了目标任务的性能提升。同时,本文模型在参数规模上也有一定优势。以 iDeepS 为例,它的卷积核数为 16 个,那么在 RBP-31 上的总卷积核数为 496 个,而本文模型仅使用了 80 个卷积核,这表明本文模型更有效地利用了模型参数。另外,从表 1 可以发现,尽管 RBP-24 和 RBP-31 在 RNA 序列长度上存在巨大差别,数据规模和 RBP 实验数量也有不同,但是本文提出的模型却可以以一套相同的超参数在两个数据集上均取得出色的成绩,这说明本文模型具有较强的泛化性能,不易过拟合,这也与训练集的规模扩大有关。

3　结　语

本文提出的模型通过 CNN-GRU 结构提取 RNA 的序列特征,并通过将 RBP 的实验编号以独热编码的形式作为模型的另一输入,扩展了模型训练集的规模,深挖不同 RNA-蛋白质结合位点问题的公共知识,进一步发挥了深度学习模型的优势。在 RNA-蛋白质结合位点预测任务中,本文模型在 RBP-24 和 RBP-31 这两个数据集上均取得了比其他模型更好的结果。但是,如此大规模的训练集对于 RNA-蛋白质结合位点预测问题是一个相对陌生的领域,尽管使用 CNN-GRU 这一结构已经取得了一定进步,但是如何使用更复杂的技术、更深的网络模型去充分挖掘数据中的信息仍然有进一步研究的空间。

参　考　文　献

[1] Dolinar A, Koritnik B, Glavac D, et al. Circular RNAs as potential blood biomarkers in amyotrophic lateral sclerosis [J]. Molecular Neurobiology, 2019, 56(12): 8052 – 8062.

[2] Mackenzie I R, Rademakers R, Neumann M. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia[J]. The Lancet Neurology, 2010, 9(10): 995 – 1007.

[3] Bolognani F, Perrone-bizzozero N I. RNA-protein interactions and control of mRNA stability in neurons[J]. Journal of Neuroscience Research, 2008, 86(3): 481 – 489.

[4] Ascano M, Hafner M, Cekan P, et al. Identification of

RNA-protein interaction networks using PAR-CLIP[J]. Wiley Interdisciplinary Reviews: RNA,2012,3(2):159 - 177.

[5] Blin K, Dieterich C, Wurmus R, et al. DoRiNA 2.0-upgrading the doRiNA database of RNA interactions in post-transcriptional regulation[J]. Nucleic Acids Research, 2015, 43(D1):160 - 167.

[6] Li J H, Liu S, Zhou H, et al. StarBase v2. 0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein—RNA interaction networks from large-scale CLIP-Seq data[J]. Nucleic Acids Research,2014,42(D1):92 - 97.

[7] Hu B, Yang Y C T, Huang Y, et al. POSTAR: A platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins[J]. Nucleic Acids Research, 2017, 45(D1): 104 - 114.

[8] Nostrand E L V, Pratt G A, Shishkin A A, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP(eCLIP)[J]. Nature Methods, 2016, 13(6):508 - 514.

[9] Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning[J]. Nature Biotechnology,2015,33(8): 831 - 838.

[10] Pan X, Shen H B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach[J]. BMC Bioinformatics,2017,18(1): 136.

[11] Pan X, Fan Y X, Jia J, et al. Identifying RNA-binding proteins using multi-label deep learning[J]. Science China Information Sciences,2019,62(1):19103.

[12] Pan X, Yan J. Attention based convolutional neural network for predicting RNA-protein binding sites[EB]. arXiv:1712.02270,2017.

[13] Pan X, Shen H B. Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network[J]. Neurocomputing,2018,305:51 - 58.

[14] Pan X, Shen H B. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks[J]. Bioinformatics,2018,34(20):3427 - 3436.

[15] Pan X, Rijnbeek P, Yan J, et al. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks[J]. BMC Genomics,2018,19(1):511.

[16] Zhang K, Pan X, Yang Y, et al. CRIP: Predicting circRNA-RBP-binding sites using a codon-based encoding and hybrid deep neural networks[J]. RNA,2019,25(12):1604 - 1615.

[17] Zhang S, Zhou J, Hu H, et al. A deep learning framework for modeling structural features of RNA-binding protein targets[J]. Nucleic Acids Research,2015,44(4):e32.

[18] Chung T, Kim D. Prediction of binding property of RNA-binding proteins using multi-sized filters and multi-modal deep convolutional neural network[J]. PLoS One, 2019, 14(4): e0216257.

[19] Ju Y, Yuan L, Yang Y, et al. CircSLNN: Identifying RBP-binding sites on circRNAs via sequence labeling neural networks[J]. Frontiers in Genetics,2019,10:1184.

[20] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation,1997,9(8):1735 - 1780.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:770 - 778.

[22] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[EB]. arXiv:1409.0473,2014.

[23] Pan X, Yang Y, Xia C Q, et al. Recent methodology progress of deep learning for RNA-protein interaction prediction [J]. Wiley Interdisciplinary Reviews: RNA, 2019, 10(6):e1544.

[24] Wang X, Zhang Y, Ren X, et al. Cross-type biomedical named entity recognition with deep multi-task learning[J]. Bioinformatics,2019,35(10):1745 - 1752.

[25] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB]. arXiv:1502.03167,2015.

[26] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB]. arXiv:1412.3555,2014.

[27] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929 - 1958.

[28] Maticzka D, Lange S J, Costa F, et al. GraphProt: Modeling binding preferences of RNA-binding proteins[J]. Genome Biology,2014,15(1):R17.

[29] Stražar M, Žitnik M, Zupan B, et al. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins[J]. Bioinformatics,2016,32(10):1527 - 1535.

[30] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB]. arXiv:1412.6980,2014.

[31] Hassanzadeh H R, Wang M D. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins [C]//2016 IEEE International Conference on Bioinformatics and Biomedicine(BIBM),2016:178 - 183.