PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(3):328~335
www.pibb.ac.cn

基于卷积神经网络和循环神经网络的环形RNA剪接位点识别研究*

孙 凯 魏庆功 臧超禹 孙如轩 姜 丹 孙晓勇**
（山东农业大学信息科学与工程学院农业大数据研究中心，泰安 271000）

摘要 本文提出了一种基于卷积神经网络和循环神经网络的深度学习模型，通过分析基因组序列数据，识别人基因组中环形RNA剪接位点.首先，根据预处理后的核苷酸序列，设计了2种网络深度、8种卷积核大小和3种长短期记忆（long short term memory，LSTM）参数，共8组16个模型；其次，进一步针对池化层进行均值池化和最大池化的测试，并加入GC含量提高模型的预测能力；最后，对已经实验验证过的人类精浆中环形RNA进行了预测.结果表明，卷积核尺寸为32×4、深度为1、LSTM参数为32的模型识别率最高，在训练集上为0.9824，在测试数据集上准确率为0.95，并且在实验验证数据上的正确识别率为83%.该模型在人的环形RNA剪接位点识别方面具有较好的性能.

关键词 深度学习，卷积神经网络，循环神经网络，环形RNA，剪接位点
中图分类号 TP391，Q52 **DOI**：10.16476/j.pibb.2020.0298

科学界报道环形RNA（circRNA）已经40余年[1]，但是由于当时知识水平和技术水平的限制，环形RNA一直被认为是异常可变剪接造成的副产物，在很长一段时间都没有引起重视[2-5].随着高通量测序等生物信息技术的发展，人们发现环形RNA并不像之前认为的是一种异常意外的现象[6-8].近几年，环形RNA逐渐成为生物领域的研究热点.目前研究发现环形RNA具有作为mRNA和RNA结合蛋白（RBP）的海绵作用，这是环形RNA功能上的重大发现[9]，除此之外环形RNA还具有转录调控因子作用以及高保守性、稳定性、时空特异性等特点，这些与人类的某些重大疾病尤其是癌症有密切关系[10].但是目前环形RNA的剪接和成环机制尚未明确.有学者针对拟南芥的转录组数据进行了分析，揭示了环形RNA的直观剪接机制[11]；有学者提出是由于RNA结合蛋白结合外显子侧翼的内含子从而促进环形RNA的形成[12]；也有学者发现许多环形RNA首尾两端外显子的侧翼内含子上有一段互补的核苷酸序列，从而促进了环形RNA的形成[13].因此，能够准确地识别出环形RNA，对于环形RNA的生物机制和人类疾病的研究具有重大意义.

目前识别环形RNA的方法可以分为两大类，一种是基于RNA-seq数据的比对方法：2010年，Wang等[14]提出一种基于RNA-seq的方法Mapsplice用于检测非规范拼接和新规范拼接，此方法不依赖于剪接位点特征和内含子长度；2014年，Zhang等[13]设计了一种基于RNA-seq比对的方法circexplorer来发现环形RNA的剪接位点；2015年，Sazabo等[15]设计了一种基于RNA-seq的统计学方法KNIFE来检测环形RNA，此方法不依靠read数和外显子同源性；同年，Gao等[16]提出一种基于切屑剪裁信号的方法CIRI针对RNA-seq数据来对环形RNA进行不偏倚、准确的检测，首次识别并验证了内含子/基因间环形RNA特有片段在人类转录组中的流行程度；2016年，You等[17]提出一种可以对单端和双端RNA数据进行环形RNA丰度检测的方法Acfs，具有识别融合环形

* 国家自然科学基金（32070684，31571306）资助项目.
** 通讯联系人.
Tel：0538-8249879，E-mail：johnsunx1@126.com
收稿日期：2020-08-18，接受日期：2020-09-17

RNA的能力.

另一种是利用机器学习对RNA-seq数据的序列特征进行学习预测：2015年Pan等[18]提出PredcircRNA，一种基于机器学习的多核学习去区分环形RNA和其他lncRNA，准确率为0.778；2018年Chen等[19]提出一种基于分层极限学习机（hierarchical extreme learning machine，H-ELM）的方法来区分环形RNA和其他lncRNA，准确率为0.789；2019年Chaabane等[20]提出基于端到端的深度学习方法，叫做circDeep，也是区分环形RNA和其他lncRNA，准确率达到0.893；2020年，Niu等[21]提出一种将极限学习机与粒子群优化算法相结合的模型，名为CirRNAPL，来区分环形RNA和其他lncRNA，结果显示，环形RNA和蛋白质结合RNA、茎中表达的环形RNA、非茎中表达的环形RNA，区分准确率分别为0.815、0.802、0.782.

以上两类方法均是基于RNA-seq数据，该类数据需要相关实验室进行生物实验并需要测序公司进行测序，而基因组数据的获取非常方便，可以在网站直接下载，相对于RNA-seq数据可以节约大量的实验和经济资源. 目前，有关环形RNA的数据库并不多，其中人和动物的环形RNA数据库CircBase、拟南芥的环形RNA数据库AtCircDB[22]、玉米和水稻的环形RNA数据库CropCircDB[23]应用广泛. 本文提出了一种基于卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural network，RNN）的网络结构，对基因组数据中的剪接位点这一区域进行训练，受Wang和Thölken两个团队工作的启发[24-25]，尝试将GC含量加入训练，来区分环形RNA与其他线性RNA. 在人类环形RNA的训练数据集上准确率达到了0.9824，测试数据集上达到了0.9570.

1　材料与方法

1.1　材料

人的环形RNA数据下载自CircBase数据库（http：//www.circbase.org/），该数据库收录了人类、小鼠等多个物种的环状RNA信息，采用find_circ软件预测去核糖体文库中的环形RNA；人的基因组数据和基因组注释文件全部下载自NCBI网站（https：//www.ncbi.nlm.nih.gov/），该网站是全球最有影响的生物学网站之一，提供多种生物数据分析工具和多物种基因组数据.

负数据集的获取：a. 将注释文件的“exon”项保留下来，并且删掉除位置信息以外的信息；b. 根据环形RNA .bed文件中的“start”和“end”项，与保留“exon”文件中的位置信息进行比较，相同的去掉.

为了得到剪接位点区域的基因序列并且处理成模型可接受的输入，需要对正、负数据的.bed文件进行如下处理. a. 将正、负数据按照“start”、“end”分成两个文件，并且均保留“chromosome”、“strand”列；b. 根据人类基因组exon、intron的一般长度，采取如下方案：从“start”出发，往相邻intron方向取50 bp，往exon方向取30 bp，共80 bp（图1），得到2个.bed文件；c. 利用bedtools工具和全基因组文件对这2个.bed文件进行序列提取，得到2个.fasta文件，对提取的序列进行独热编码处理（One-Hot Encoding），将碱基A处理为0001、碱基T处理为0010、碱基G处理为0100、碱基C处理为1000、杂质如N处理为0000；d. 同时对于每一条序列进行GC含量的计算. 预处理过程如图2所示.

1.2　方法

1.2.1　不同网络深度、卷积核大小和LSTM参数的模型设计

本文的模型主要是由两个核心结构组成，CNN和RNN，其中RNN选择使用长短期记忆（long short term memory，LSTM）结构. 可将预处理后的序列看成是80×4的图片，利用CNN去提取特征，然后通过LSTM的“记忆”功能去理解整体序列信息. 本文从卷积核尺寸、卷积网络深度、LSTM参数入手，设计不同的模型结构组合，探索出最优的环形RNA剪接位点识别模型.

模型部分参数为：学习率0.0001，批处理量64，训练轮数20，训练validation比例0.1，卷积层后的dropout比例0.3，LSTM层后的dropout比例0.1，卷积层使用“relu”作为激活函数，“sigmoid”作为分类函数. 采用正确识别率P作为模型性能的评价标准，计算公式如下：

$$P = C_i^T/C_i^T+C_i^F \tag{1}$$

式（1）中C_i^T表示环形RNA剪接位点标签为i识别正确的数量，C_i^F表示环形RNA剪接位点标签为i识别错误的数量.

环形RNA剪接位点的序列经过预处理可以看成是80×4的图片，但是由于其不同于一般图像识别的形态和内容，所以我们选用了当下比较流行的

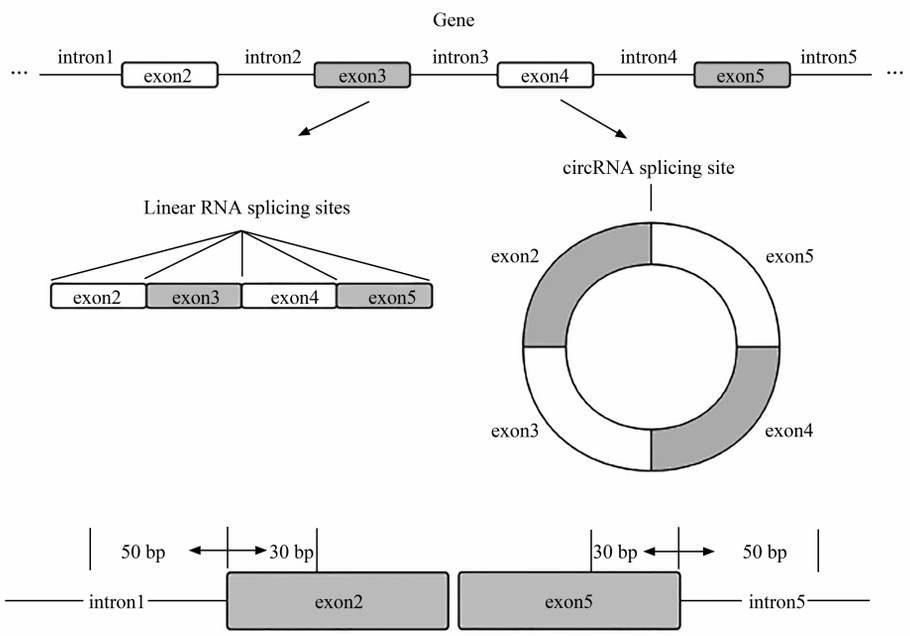

Fig. 1　Schematic diagram of linear RNA，circular RNA splicing site and 80 bp sequence acquisition

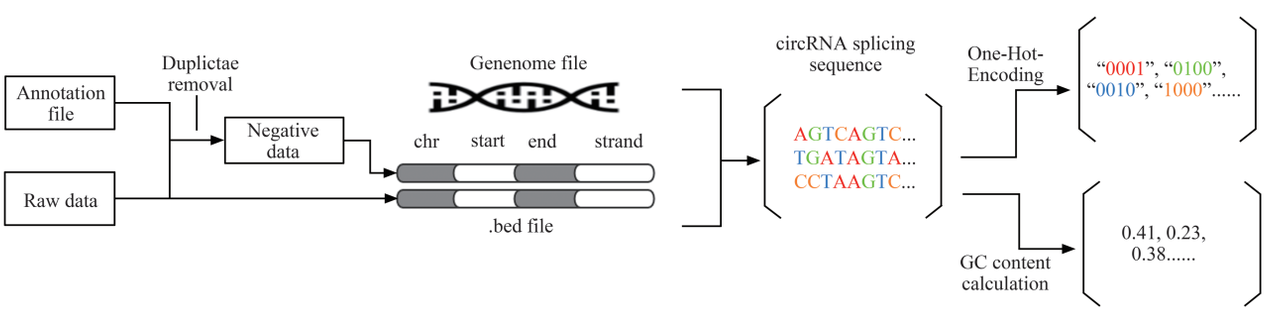

Fig. 2　Schematic diagram of data preprocessing

5×5、3×3、3×1卷积核和根据当前实际问题设计的其他大小的卷积核和LSTM参数（表1）.

1.2.2　基于CNN不同池化层的模型构建选择

池化层可以进行采样，在降低特征图维度的同时，还具有去除冗余信息，避免过拟合和提高泛化能力等优点. 当下比较流行的池化层选择为均值池化（average pooling）和最大池化（max pooling）. 本文从同组模型中选出性能更好的模型分别进行两种池化层的测试，从而选择一种最优组合.

2　结果与分析

2.1　不同卷积核尺寸、深度和LSTM参数对模型识别率的影响

试验的16个模型卷积层均采用padding补0的方式，采样层则选用均值池化层. 环形RNA剪接位点和线性RNA剪接位点数据各32 933个.

16种模型的准确率变化曲线（图3、4）表明，模型*2*、*8*、*11*、*12*、*13*、*14*、*16*初始识别率较低，均低于0.7，在前三轮迭代中，准确率增长较快，之后增长趋势相对较慢，逐渐趋于稳定；模型*1*、*3*、*4*、*5*、*6*、*7*、*9*、*10*初始准确率较高，均大于0.7，特别是模型*5*、*10*，初始准确率达到0.75以上，在前两轮迭代中，准确率增长幅度较大，在之后的18轮中增长幅度较小，直到最后两轮逐渐趋于平稳；将相同参数但深度不同的两个模型看成一组去比较（共8组），发现其中6组深度为1的模型最终准确率高于深度为2的模型，且初始识别率高的，最终识别准确率也高于初始识别率低的；将深度为1的8个模型纵向去比较，发现模型*1*、*3*、*7*、*9*、*11*的性能相对较好，准确率在0.9以上，这些

Table 1　Based on different convolution depth，convolution kernel size and LSTM parameters，16 models are combined

Model number	CNN kernal	CNN Depth	LSTM	RNN Depth
1	16×4	1	16	1
2	16×4	2	16	1
3	32×4	1	32	1
4	32×4	2	32	1
5	64×4	1	64	1
6	64×4	2	64	1
7	16×3	1	16	1
8	16×3	2	16	1
9	32×3	1	32	1
10	32×3	2	32	1
11	5×5	1	16	1
12	5×5	2	16	1
13	3×3	1	16	1
14	3×3	2	16	1
15	3×1	1	16	1
16	3×1	2	16	1

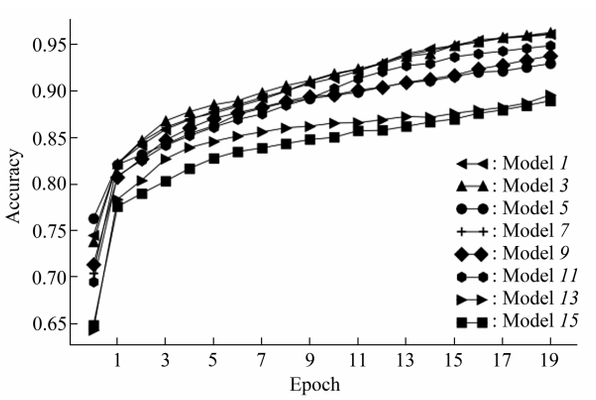

Fig. 3　Training process of eight models with depth 1

模型的卷积核相对较大，最小为5×5；将深度为2的8个模型纵向去比较，发现模型*2*、*4*、*8*、*10*、*12*的性能相对较好，与前面深度为1的正好对应，具有相同的特点.

据此，可以得到如下结论：a. 随着迭代次数的增加，模型的准确率也越来越高；b. 对于同一深度的网络，卷积核的大小不宜选择过小，16×4、32×4、16×3、32×3、5×5的相对较好；c. 随着网络深度的增加，大多数模型的性能并没有提升，8组中有6组都是深度为1的准确率更高.

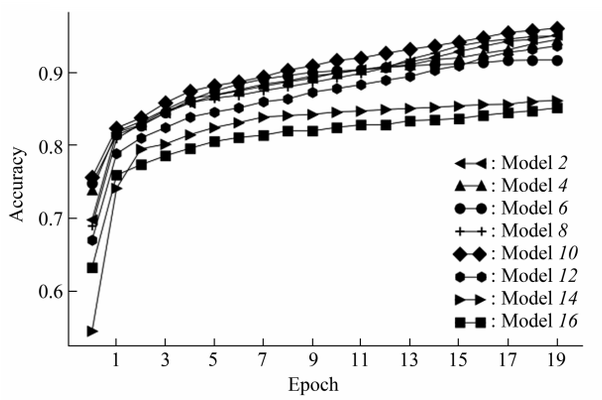

Fig. 4　Training process of eight models with depth 2

2.2　采样层对模型准确率的影响

由表2可知，均值池化与最大池化对于模型准确率的影响非常接近，模型*1*、*3*、*7*、*11*当使用均值池化时，效果更好，模型*5*、*10*、*13*、*15*则在使用最大池化时效果较好. 基于上述结果，采样层应根据实际效果择优选择.

Table 2　Comparison of training accuracy of eight models in average pooling layer and maximum pooling layer

Model	Recognition rate	
	Average pooling	Max pooling
1	0.9803	0.9675
3	0.9824	0.9625
5	0.9458	0.9555
7	0.9810	0.9608
10	0.9703	0.9715
11	0.9753	0.9740
13	0.9346	0.9385
15	0.9381	0.9449

2.3　GC含量对于模型准确率的影响

在序列本身这一自变量的基础上，加入了GC含量这一变量，未加入GC含量16个模型的训练过程如图5、6所示. 可以看出深度相同的模型在未加入GC含量前，无论是初始识别率还是最终识别率均低于加入GC含量后的对应模型，具体准确率对比见表3.

由表3可知，在模型深度和参数相同的情况下，带GC含量这一自变量的模型准确率均高于不带GC含量的模型，准确率最大差值为13.64%，最小差值为6.50%，提升效果十分显著.

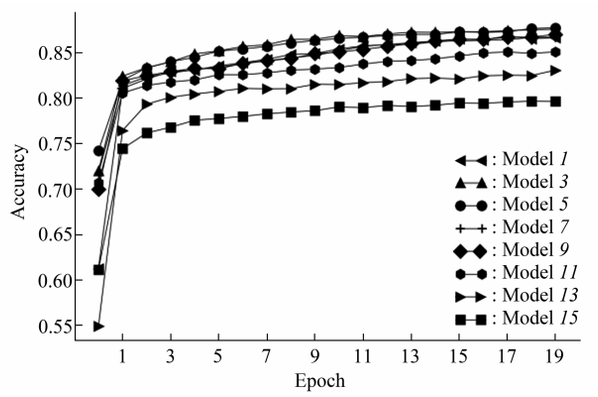

Fig. 5　Training process of eight models with depth 1 and without GC content

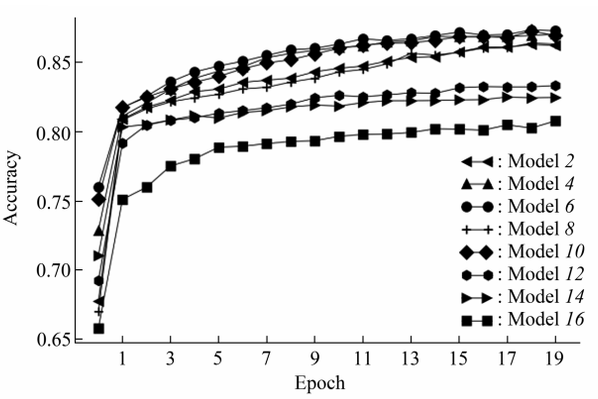

Fig. 6　Training process of eight models with depth 2 and without GC content

2.4　模型在测试集和实验验证的环形RNA数据上的表现

为了进一步测试模型的性能，分别进行了两次测试：a. 将数据集按照6∶2∶2的比例进行划分，分别用于training、evaluate和predict（图7，表4）. 其中evaluate准确率为0.9602，predict准确率为0.9570，从总体曲线图上也可以看到预测值两端所占比例较大，说明模型的识别能力较强；b. 同时对一组人类精浆中的环形RNA数据[26]进行测试（表5），预测值大于等于0.85的认为是环形RNA，预测值小于0.85的则认为不是环形RNA. 对比实验验证结果，24个数据预测正确，正确率为83%. 由此可以得出结论，本文所构造出的识别环形RNA剪接位点模型对生物实验具有重要的指导意义.

2.5　在线预测工具

为了方便人们使用本文的模型进行环形RNA的预测，我们提供了一个在线工具，名为DeepCircRNA，其网址为http://www.deepbiology.cn/DeepCircRNA/. 根据网站提供的输入示例，不仅可以对单组DNA序列进行预测，而且可以实现上传文件的功能，对批量的序列进行预测，简单快捷.

Table 3　Comparison of accuracy of 16 models before and after adding GC content

Model	Accuracy	
	with GC	without GC
1	0.9654	0.8772
2	0.9210	0.8517
3	0.9724	0.8734
4	0.9670	0.8661
5	0.9576	0.8756
6	0.9435	0.8792
7	0.9723	0.8793
8	0.9279	0.8454
9	0.9584	0.8721
10	0.9546	0.8606
11	0.9557	0.8589
12	0.8964	0.8277
13	0.9278	0.8277
14	0.9164	0.8514
15	0.9600	0.8236
16	0.9012	0.7848

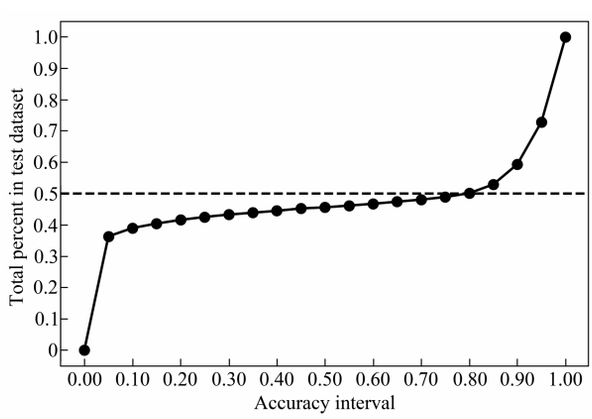

Fig. 7　Cumulative plot for 6 586 data

The x label means the model prediction accuracy interval.

Table 4　Results of 6 586 data for predicting

n=6 586	Predict：No	Predict：Yes
Actual：No	3070	223
Actual：Yes	60	3233

Table 5　The prediction results and experimental verification results of the model

Chr	Start	End	Prediction	Experiment
chr *3*	176782708	176816329	0.9969	+
chr *7*	102106264	102106717	0.9969	+
chr *12*	69644909	69656342	0.9969	+
chr *13*	23775175	23776697	0.9969	+
chr *19*	47653459	47658470	0.9969	+
chr *22*	4205907	42054356	0.9969	+
chr *1*	224605962	224612356	0.9968	+
chr *1*	92428275	92430321	0.9968	+
chr *1*	92446446	92446735	0.9967	+
chr *19*	47653459	47656297	0.9967	+
chr *13*	23751718	23776697	0.9966	+
chr *13*	25352433	25356082	0.9966	+
chr *19*	47646751	47658470	0.9966	+
chr *13*	25341410	25356082	0.9950	+
chr *13*	23792336	23794093	0.9949	+
chr *19*	47646751	47673180	0.9904	+
chr *4*	144464662	144465125	0.9900	+
chr *6*	25727079	25727268	0.9826	+
chr *22*	45749858	45750995	0.9729	+
chr *1*	92428275	92433817	0.9721	+
chr *8*	74585342	74601048	0.9713	+
chr *13*	25348951	25356082	0.9913	+
chr Y	2821950	2829687	0.8771	+
chr *10*	111878345	111886261	0.8594	+
chr *22*	42046727	42046895	0.8585	+
chr *3*	172694758	172766884	0.6996	+
chr *9*	136302869	136303486	0.6475	+
chr *10*	111883775	111890244	0.6360	+
chr *16*	3900298	3901010	0.0183	+
chr *17*	56690782	56693666	0.0116	+

+: Circular RNA was validated by experiments.

3　结论与展望

本研究通过对环形RNA剪接位点的基因序列进行预处理，从而可将序列数据看成80×4的二维“图像”，在此基础上，利用CNN和RNN的组合进行深度学习建模，对不同卷积核参数、不同深度和不同LSTM参数的16种组合进行测试，又进行了均值、最大两种池化层的比较，并加入第二个自变量GC含量，前后对比了32个模型的准确率后，最终筛选出准确率为0.9824的模型. 最后在实验验证数据上的表现进一步证明模型的可靠性和优秀的识别性能. 本文研究可以帮助从事环形RNA相关研究的实验室更快、更精确地识别发现环形RNA，揭示环形RNA的生物机制与功能.

下一步计划分析拟南芥、玉米、水稻三种植物的环形RNA数据，建立适用于植物环形RNA剪接位点预测的模型并且尝试找出植物与人和动物环形RNA的不同.

参　考　文　献

[1] Sanger H L, Klotz G, Riesner D, *et al*. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. Proc Natl Acad Sci USA, 1976, **73**(11): 3852-3856

[2] Hsu M T , Coca-Prados M . Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature, 1979, **280**(5720): 339-340

[3] Nigro J M, Cho K R, Fearon E R, *et al*. Scrambled exons. Cell, 1991, **64**(3): 607-613

[4] Cocquerelle C, Daubersies P, Majérus M A, *et al*. Splicing with inverted order of exons occurs proximal to large introns. The EMBO Journal, 1992, **11**(3):1095-1098

[5] Capel B , Swain A , Nicolis S , *et al*. Circular transcripts of the testis-determining gene Sry in adult mouse testis. Cell, 1993, **73**(5):1019-1030

[6] Danan M, Schwartz S, Edelheit S, *et al*. Traascriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Res, 2012, **40**(7): 3131-3142

[7] Salzman J, Gawad C, Wang P L, *et al*. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types.Plos One, 2012, **7**(2): e30733

[8] Jeck W R, Sorrentino J A, Wang K, *et al*. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA, 2013, **19**(2): 141-157

[9] Hansen T B, Jensen T I, Clausen B H, *et al*. Natural RNA circles function as efficient microRNA sponges. Nature, 2013, **495**(7441): 384-388

[10] Memczak S, Jens M, Elefsinioti A, *et al*. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 2013, **495**(7441): 333-338

[11] Sun X X, Wang L, Ding J, *et al* . Integrative analysis of arabidopsis thaliana transcriptomics reveals intuitive splicing mechanism for circular RNA. FEBS Lett. 2016, **590**(20): 3510-3516

[12] Chen L L. The biogenesis and emerging roles of circular RNAs. Nature Reviews Molecular Cell Biology, 2016, **17**(4): 205-211

[13] Zhang X O, Wang H B, Zhang Y, *et al*. Complementary sequence-mediated exon circularization. Cell, 2014, **159**(1):134-147

[14] Wang K, Singh D, Zeng Z, *et al*. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nuclc Acids Research, 2010, **38**(18):e178

[15] Szabo L, Morey R, Palpant N J, *et al*. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome

Biology, 2015,**16**(1): 126

[16] Gao Y, Wang J F, Zhao F Q. CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. Genome Biology, 2015, **16**(1): 4

[17] You X T, Conrad T O. Acfs: accurate circRNA identification and quantification from RNA-Seq data. Scientific Reports, 2016, **6**: 38820

[18] Pan X Y, Xiong K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. Molecular Biosystems, 2015, **11**(8): 2219-2226

[19] Chen L, Zhang Y H, Huang G H*et al*. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. Mol Genet Genomics, 2018, **293**(1): 137-149

[20] Chaabane M, Williams R M, Stephens A T, *et al*. CircDeep: deep learning approach for circular RNA classification from other long non-coding RNA. Bioinformatics, 2019, **36**(1): 73-80

[21] Niu M T, Zhang J, Li Y J, *et al*. CirRNAPL: a web server for the identification of circRNA based on extreme learning machine. Comput Struct Biotechnol J, 2020, **18**: 834-842

[22] Jiazhen Y, Lin W, Shuzhang L, *et al*. AtCircDB: a tissue-specific database for arabidopsis circular RNAs. Brief Bioinform, 2019, **20**(1): 58-65

[23] Kai W, Chong W, Baohuan G, *et al*. CropCircDB:a comprehensive circular RNA resource for crops in response to abiotic stress. Database(Oxford), 2019, **2019**: baz053

[24] Wang M, Marín A. Characterization and prediction of alternative splice sites. Gene, 2006, **366**(2):219-227

[25] Thölken C, Thamm M, Erbacher C, *et al*. Sequence and structural properties of circular RNAs in the brain of nurse and forager honeybees (Apis mellifera). BMC Genomics, 2019, **20**(1): 88

[26] Dong W W, Li H M, Qing X R, *et al*. Identification and characterization of human testis derived circular RNAs and their existence in seminal plasma. Scientific Reports, 2016, **6**: 39080

Identifying Circular RNA Splicing Sites Based on Convolutional Neural Networks and Recurrent Neural Networks*

SUN Kai, WEI Qing-Gong, ZANG Chao-Yu, SUN Ru-Xuan, JIANG Dan, SUN Xiao-Yong**

(*College of Information Science and Engineering , Shandong Agricultural University,Tai'an* 271000, China)

Abstract In this paper, we propose a deep learning model based on convolutional neural network and recurrent neural network, which uses genome sequence data to identify human circular RNA splicing sites. Firstly, we preprocessed the original genome sequences and designed 16 models with two network depths, eight convolution kernel sizes and three LSTM parameters; secondly, the pooling layer was further tested for average pooling and maximum pooling; and GC content was added to improve the prediction ability of the model; finally, we predicted the circRNA in human seminal plasma. The results show that the model with convolution kernel of 32 × 4, depth of 1 and LSTM parameter of 32 has the highest recognition rate of 0.9824 on training data set, and 0.95 on test data set. Also, we tested our model with a published study and the accuracy reaches 0.83. The model has good performance in the recognition of human circular RNA splicing sites.

Key words deep learning, convolutional neural networks, recurrent neural networks, circular RNA, splicing sites

DOI：10.16476/j.pibb.2020.0298

* This work was supported by grants from The National Natural Science Foundation of China（32070684，31571306）.

** Corresponding author.

Tel：86-538-8249879，E-mail：johnsunx1@126.com

Received：August 18，2020 Accepted：September 17，2020