Advances in Applied Mathematics 应用数学进展, 2022, 11(9), 6670-6677
Published Online September 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.119706

Hans汉斯

基于PCA和神经网络预测长链非编码RNA

曹冰倩

青岛大学，数学与统计学院，山东 青岛

收稿日期：2022年8月21日；录用日期：2022年9月16日；发布日期：2022年9月23日

摘　要

长链非编码RNA是一类不编码蛋白质且长度不小于200 nt的转录本。近些年来，长链非编码RNA被发现在生命活动中发挥着重要的作用，而研究其功能的第一步就是准确识别出长链非编码RNA。在本文中，我们基于主成分分析和多层感知机提出了一种识别长链非编码RNA的新方法。我们选择转录本的k-mer作为原始特征向量，使用主成分分析进行降维得到新的特征向量，将其输入到一个含有五个隐藏层的多层感知机中来预测其是否为长链非编码RNA。我们使用人类、小鼠和斑马鱼物种的转录物序列来评估我们所提出的方法，最终在上述物种的normal类型测试集上准确率分别为94.74%，93.25%和93.04%。

关键词

长链非编码RNA，主成分分析，多层感知机，深度学习

Prediction of Long Non-Coding RNAs Based on PCA and Neural Network

Bingqian Cao

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Aug. 21st, 2022; accepted: Sep. 16th, 2022; published: Sep. 23rd, 2022

Abstract

Long non-coding RNAs are transcripts composed of more than 200 nucleotides that do not encode proteins. In recent years, long non-coding RNAs have been found to play important roles in many biological mechanisms, and the first step to study their functions is to identify long non-coding RNAs accurately. In this paper, we propose a novel method to identify long non-coding RNAs based on principal component analysis and multilayer perceptron. We select the k-mer of the transcript as the original feature vectors and use principal component analysis to reduce the dimension to obtain new feature vectors. The new feature vector of transcript was fed into a multilayer percep-

文章引用: 曹冰倩. 基于 PCA 和神经网络预测长链非编码 RNA [J]. 应用数学进展, 2022, 11(9): 6670-6677.
DOI: 10.12677/aam.2022.119706

tron with five hidden layers to predict the coding ability of the transcript. We used the transcript sequences of human, mouse and zebrafish to evaluate our proposed method and achieved 94.74%, 93.25% and 93.04% accuracies on the normal type test set of the above species, respectively.

Keywords

Long Non-Coding RNA, Principal Component Analysis, Multilayer Perceptron, Deep Learning

1. 引言

长链非编码 RNA (Long non-coding RNA, lncRNA)是一类不编码蛋白质且长度不小于 200 nt 的转录物序列[1] [2] [3]。长期以来，lncRNA 被视作没有生物学功能的“噪声”，但是最新的研究表明，lncRNA 与基因表达，蛋白质翻译和稳定性等细胞机制有关[4]。除此之外，lncRNA 还和许多疾病密切相关，例如癌症、糖尿病等[4] [5] [6]。因此，研究 lncRNA 具有十分重大的意义。近些年来，随着高通量测序技术的发展，产生了大量的新的序列数据，这其中就有许多长链非编码 RNA，而研究其功能的第一步就是准确识别出长链非编码 RNA，所以研究出能够准确快速识别出 lncRNA 的工具是迫切的需求。

目前已经发展出了一些识别 lncRNA 的生物信息学方法。第一大类是基于比对的方法。Kong 等在 2007 年提出了 CPC，其使用六个具有生物学意义的特征，并基于支持向量机模型预测转录本序列的编码能力[7]。这六个特征包括三个开放阅读框(open reading frame, ORF)特征以及和已知蛋白质的序列相似性特征。CPC 具有较高的蛋白质编码转录本识别能力，但也存在速度缓慢等不足之处。Lin 等在 2011 年提出了 PhyloCSF [8]。PhyloCSF 充分利用多重比对并产生非编码 RNA 和编码 RNA 的似然比。上述基于比对的方法都依赖于数据库，这限制了它们的预测能力。第二大类是不基于比对的方法，这类方法依赖于序列本身固有的信息，这提高了其预测转录本序列编码能力的效率。Wang 等在 2013 年提出了 CPAT，其使用开放阅读框的最大长度、开放阅读框覆盖率、Fickett 分数以及 hexamer 分数作为特征构建了逻辑回归模型去识别 lncRNA [9]。Hexamer 得分是 CPAT 所使用的特征中最有识别力的特征[10]。Li 等提出了 PLEK 去识别 lncRNA [11]。PLEK 使用校准的 k-mer 作为特征并使用支持向量机算法来区分长链非编码 RNA 和信使 RNA。Tong 等在 2019 年提出了 CPPred [12]。CPPred 基于支持向量分类器预测转录本序列的编码能力。近些年来，深度学习在诸多领域得到了广泛应用，随之也出现了一些基于深度学习技术去识别 lncRNA 的工具，例如 DeepCPP [13]、LncRNAnet [14]、LncADeep [15]等。

在本文中，我们提出了一种新方法去识别 lncRNA。我们使用 k-mer 作为原始特征，利用主成分分析法对 k-mer 特征进行降维。我们选择累积方差贡献率比较大的一些主成分作为新的特征向量，并将其作为我们构建的多层感知机神经网络的输入来预测转录本序列的编码能力。最后，我们使用灵敏度、特异度和准确率来评估我们提出方法的性能。

2. 材料和方法

2.1. 数据集

我们使用人类的编码 RNA 序列和长链非编码 RNA 序列去训练模型。得到模型后，使用人类、小鼠、

斑马鱼物种的转录物序列来评估我们所提出方法的性能。我们从 Ensembl 数据库中收集上述物种的转录物序列来建立各个物种的数据集。

我们从人类数据集中的长链非编码 RNA 序列和编码转录物序列中分别随机抽取三分之二的序列作为训练集，剩余的序列作为人类物种测试集，共得到了 16068 条编码转录物序列以及 19606 条长链非编码 RNA 序列作为训练集，8034 条编码 RNA 和 9804 条长链非编码 RNA 序列作为测试集。

当在其他物种数据集上测试模型性能时，将上述所有人类物种转录物序列作为测试集来训练模型。具有小开放阅读框的长链非编码 RNA (sORF)是长度不超过 300 nt 的长链非编码 RNA，对其进行正确预测比较困难，还有提升的空间。我们仍然构建了各个物种的 sORF 类型的测试集来评估我们所提出的方法，表 1 展示了各个数据集的详细信息。

Table 1. The data amount used for evaluating the performance of the method proposed in this work
表 1. 用于评估本文提出方法性能的数据集所含样本数目

数据集名称	编码转录本数量	lncRNA 数量
Human training dataset	16068	19606
Human normal test dataset	8034	9804
Human sORF test dataset	1575	966
Mouse normal test dataset	21798	11804
Mouse sORF test dataset	980	579
Zebrafish normal test dataset	23856	3067
Zebrafish sORF test dataset	624	188

2.2. 特征提取

K-mer 是一种从 DNA 序列和 RNA 序列中提取信息的常见且有效的方法，在生物信息学的许多领域中有着广泛的应用。对于一条给定的转录物序列，我们以 k 为滑动窗口的大小，每次滑动 1 个步长去分割该转录物序列，并计算 4^k 种子序列出现的频率，滑动窗口的长度为 k，且每个位置的碱基有四种(A：腺嘌呤，T：胸腺嘧啶，C：胞嘧啶，G：鸟嘌呤)，因此共有 4^k 种子序列。因此，我们得到了一个表示转录物序列的维度为 4^k 的向量。可以注意到，随着 k 的不断增大，k-mer 向量的维度呈指数级增长，这可能会引起“维度灾难”。在本研究中，我们将 k 设置为 3、4、5，因此我们可以将一条转录物序列表示为一个 1344 维的向量。

2.3. 主成分分析

特征的维度并不是越高越好，高维特征向量中可能会存在信息冗余及噪声，进而影响模型的精度和运算速度，导致过拟合问题，因此在将其输入到模型之前进行降维是非常有必要的。常见的降维方法有主成分分析，方差分析，因子分析等。本文选择主成分分析来进行特征降维。主成分分析使用方差来作为衡量信息量大小的标准，在信息损失尽可能小的情况下，将高维空间上的数据映射到低维空间，将含有冗余信息的多个变量映射成数量较少的不相关的新变量，也就是主成分。使用主成分分析法进行降维的步骤如下[16] [17]：

1) 对数据集进行中心化处理

2) 计算样本的协方差矩阵

3) 计算协方差矩阵的特征值与特征向量

4) 取最大的 i 个特征值所对应的特征向量，记作 $w_1, w_2, \cdots, w_i$，因此得到投影矩阵：

$$W = (w_1, w_2, \cdots, w_i)$$

其中 i 为选取的主成分的个数。

2.4. 多层感知机模型

近些年来，深度学习技术迅速发展，在许多领域有着广泛的应用，在生物信息学中也得到了越来越多的应用。多层感知机是一种经典的深度学习模型，它包括输入层、输出层和若干隐藏层，这显著提高了神经网络拟合非线性数据的能力。在本文中，我们构建了多层感知机模型去从转录本序列中识别 lncRNA。我们构建的多层感知机模型包括输入层，输出层以及若干个隐藏层。本文构建的多层感知机模型将累积方差贡献率比较高的主成分作为输入，输出该转录本序列是 lncRNA 和编码转录本的概率。激活函数有着非常重要的作用，它是非线性变换，若没有激活函数，多层神经网络实际上和单层神经网络没有区别。使用激活函数为神经网络加入了非线性变换，使得其可以拟合更加复杂的数据。常见的激活函数有 sigmoid、tanh 和 relu 函数。我们在本文中构建的多层感知机模型的每个隐藏层的神经元节点数都被设置为相同的，并使用 relu 函数作为激活函数，输出层使用 softmax 函数作为激活函数来输出转录本是 lncRNA 和编码转录本的概率。随机丢弃在深度学习中是一种常见且有效的技术。随机丢弃是指在神经网络训练的过程中，随机忽略一部分神经元，这有利于简化神经网络、缓解过拟合问题。本文在构建多层感知机时，在每一个隐藏层都使用随机丢弃方法随机忽略一半的神经元来减轻神经网络过拟合问题。本文提出的方法流程如图 1 所示：

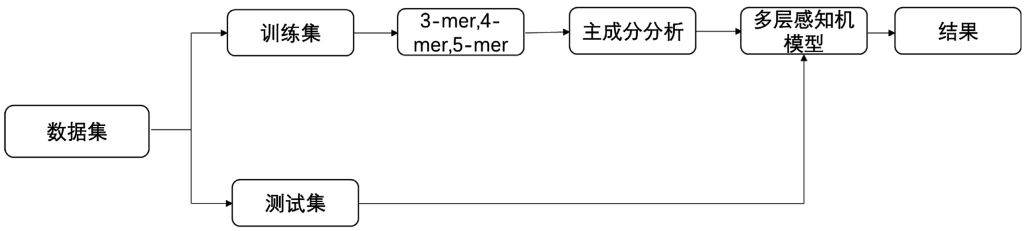

Figure 1. The flow chart of methods proposed in the paper
图 1. 本文方法流程图

2.5. 评估准则

我们使用准确率(ACC)，灵敏度(SN)和特异度(SP)等常见的机器学习评估指标来评估我们所提出方法的性能：

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

$$SN = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

其中 TP 代表被正确识别的阳性样本的数量，TN，FP，FN 分别代表真阴性，假阳性，假阴性样本的数量，

在本文中，lncRNA 为阳性样本。

3. 结果

3.1. 参数对模型性能的影响

特征数量过多时会造成计算效率低下，过拟合等问题，因此我们利用主成分分析来降低特征维度。我们选择前 333 个主成分，其累积方差贡献率在 95%左右。

在本文中我们使用 k-mer 作为原始特征向量，利用主成分分析进行降维，得到维度为 333 的特征向量，将其作为多层感知机的输入，进而得到转录本为 lncRNA 和编码转录本的概率，选取概率大的类别作为最后的预测结果。

K-mer 是生物信息学中常见且有效的特征提取方式，我们在测试集上验证了使用不同 k-mer 作为特征搭建的模型的性能。图 2 展示了不同 k-mer 作为特征时，不同测试集上模型的性能。我们可以发现在选择 3-mer,4-mer 和 5-mer 组合特征时，在相应的测试集上 ACC 达到最大，尤其是在 HumansORF testdataset 上的 ACC 的提高非常明显，因此我们选择 3-mer,4-mer 和 5mer 为原始特征。

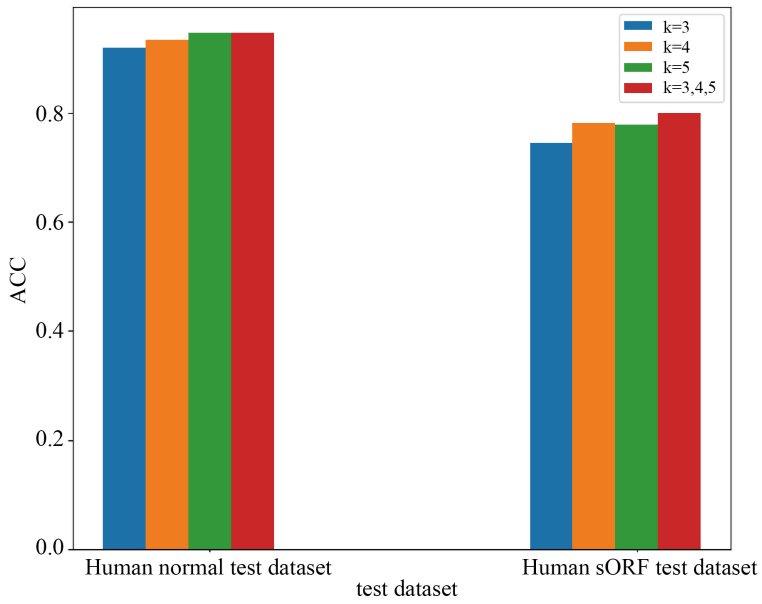

Figure 2. The effect of different values of k on the performance of the model
图 2. k 不同的取值对模型性能的影响

多层感知机中隐藏层的层数和每一层的节点数是神经网络中非常重要的参数，我们在本文中所搭建的模型每一个隐藏层中含有相同的节点个数。为了选取最好的参数以及展示不同参数对性能的影响，我们分别选择不同的隐藏层的层数和不同的节点个数来构建模型并使用人类物种数据来测试其性能。

当固定节点个数时，我们将隐藏层的个数分别设置为 1，2，3，4，5 搭建模型来测试其性能。可以发现在 Human normal test dataset 上时，不同的隐藏层的个数对应的 ACC 大体相同，但是在 Humans ORF test dataset 上不同的隐藏层层数对应的 ACC 变化较大，当隐藏层的层数为 5 时对应的 ACC 明显高于其他隐藏层层数对应的 ACC，因此我们将隐藏层的个数设置为 5。图 3 展示了隐藏层的层数对模型性能的影响。

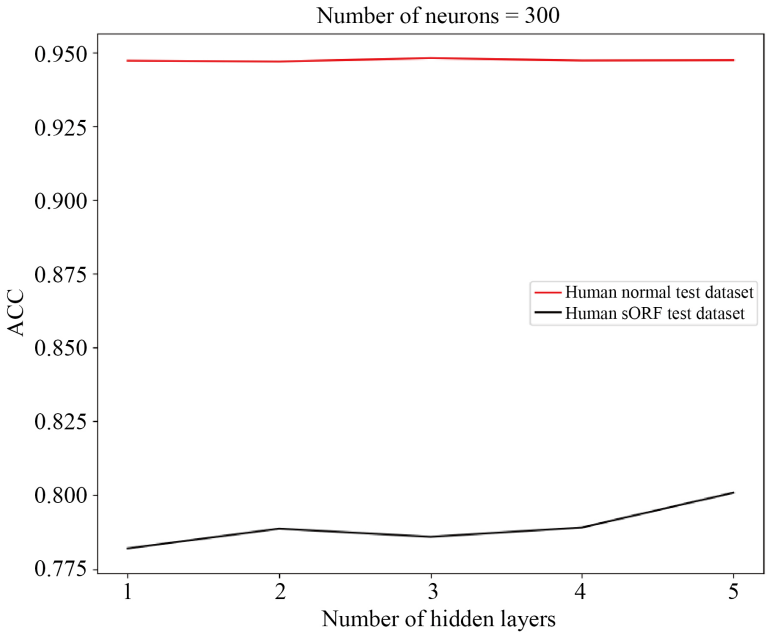

Figure 3. The effect of the number of hidden layers on model performance
图 3. 隐藏层的层数对模型性能的影响

当固定隐藏层的层数时，我们将节点数设置为 50、100、150、200、250 和 300 来构建不同的模型。在 Human normal test dataset 上，不同的节点数对应的 ACC 差别不大。在 Humans ORF test dataset 上不同的节点数对应的 ACC 波动较为明显，当节点数设置为 300 时，所搭建的模型在 Humans ORF test dataset 对应的 ACC 达到最大，因此我们将每一个隐藏层所含有的神经元个数设置为 300。图 4 展示了不同的节点数对模型性能的影响。

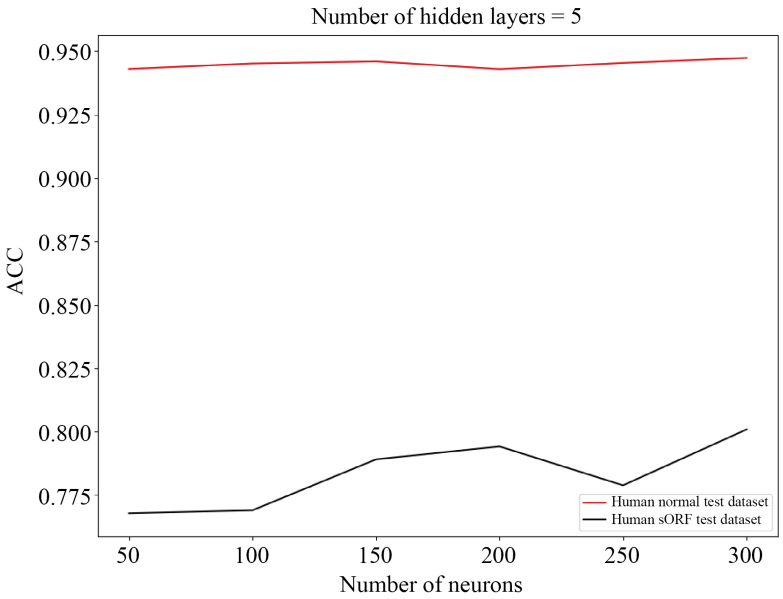

Figure 4. The effect of the number of neurons on model performance
图 4. 节点数对模型性能的影响

3.2. 模型性能

我们在人类、小鼠和斑马鱼的不同类型测试集上测试我们提出的方法的性能，就准确率而言，我们提出的方法在 Human normal test dataset 上最高，为 94.74%，另外两个物种的 normal 类型数据集即 Mouse normal test dataset，Zebrafish normal test dataset 次之，分别为 93.25%和 93.04%，并且 normal 类型数据集上的准确率高于 sORF 类型的数据集上的准确率。

对于灵敏度指标，Human normal test dataset 最高，为 95.24%，Mouse normal test dataset 次之，为 91.12%，其余数据集的灵敏度均在 90%之下，Human sORF test dataset，Mouse sORF test dataset 上的灵敏度均在 80%之上，分别为 82.30%，84.97%，而斑马鱼物种的测试集上的灵敏度均较低，Zebrafish normal test dataset 上的灵敏度 75.64%，Zebrafish sORF test dataset 上的灵敏度为 64.36%。

就特异度而言，Human normal test dataset、Mouse normal test dataset 和 Zebrafish normal test dataset 的特异度均超过了 90%，其中最高的为 Zebrafish normal test dataset 的 95.28%，Zebrafish sORF test dataset 的特异度最低，为 74.04%。表 2 展示了我们所搭建模型在各个测试集上的性能。

Table 2. The performance of the proposed method on each test set
表 2. 本文提出方法在各个测试集上的性能

数据集	灵敏度(%)	特异度(%)	准确率(%)
Human normal test dataset	95.24	94.14	94.74
Human sORF test dataset	82.30	78.73	80.09
Mouse normal test dataset	91.12	94.41	93.25
Mouse sORF test dataset	84.97	78.78	81.08
Zebrafish normal test dataset	75.64	95.28	93.04
Zebrafish sORF test dataset	64.36	74.04	71.80

4. 讨论

我们在本文中提出了一个从转录本序列中识别 lncRNA 的新方法，我们使用 3-mer，4-mer，5-mer 作为原始特征，并使用主成分分析对特征进行降维得到了 333 维的新特征向量，然后基于多层感知机模型来预测其是否为 lncRNA。我们提出的方法是不基于比对的方法，这利用了序列本身的信息。我们使用主成分分析对原始特征向量进行降维，这大大缓解了维度过高可能会导致的过拟合以及计算效率低下等问题。我们使用训练集训练模型后，使用不同物种的不同类型测试集评估了本文所提出的方法。

我们在本文中提出的方法不同于以往直接将特征送入神经网络的方法。我们首先使用主成分分析对原始特征进行了一次降维，这相当于对原始特征进行初步的提取有效信息，除此之外，这还有利于提高后续训练效率以及缓解过拟合问题。

5. 总结

本文提出了一个用于区分 lncRNA 和编码转录本的新方法。我们将 3-mer，4-mer 和 5-mer 作为原始特征，使用主成分分析法来降低原始特征的维度，进而得到最终的 333 维特征向量。我们搭建了包含 5 个隐藏层的多层感知机，并将降维后的 333 维特征向量作为其输入来识别 lncRNA。我们构建的模型在人类、小鼠和斑马鱼的 normal 类型测试集上的准确率较高，分别达到了 94.74%，93.25%和 93.04%。

参考文献

[1] Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C.P., Sorensen, P.H., Reaman, G., Milos, P., Arceci, R.J. and Thompson, J.F. (2010) The Majority of Total Nuclear-Encoded Non-Ribosomal RNA in a Human Cell Is 'Dark Matter' Un-Annotated RNA. *BMC Biology*, **8**, Article No. 149. https://doi.org/10.1186/1741-7007-8-149

[2] Laurent, G.S., Wahlestedt, C. and Kapranov, P. (2015) The Landscape of Long Noncoding RNA Classification. *Trends in Genetics*, **31**, 239-251. https://doi.org/10.1016/j.tig.2015.03.007

[3] Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) Long Non-Coding RNAs: Insights into Functions. *Nature Reviews Genetics*, **10**, 155-159. https://doi.org/10.1038/nrg2521

[4] Schmitz, S.U., Grote, P. and Herrmann, B.G. (2016) Mechanisms of Long Noncoding RNA Function in Development and Disease. *Cellular and Molecular Life Sciences*, **73**, 2491-2509. https://doi.org/10.1007/s00018-016-2174-5

[5] Morán, I., Akerman, I., Van De Bunt, M., Xie, R., Benazra, M., Nammo, T., Arnes, L., Nakić, N., García-Hurtado, J. and Rodríguez-Seguí, S. (2012) Human β Cell Transcriptome Analysis Uncovers LncRNAs That Are Tissue-Specific, Dynamically Regulated, and Abnormally Expressed in Type 2 Diabetes. *Cell Metabolism*, **16**, 435-448. https://doi.org/10.1016/j.cmet.2012.08.010

[6] Tsai, M.C., Spitale, R.C. and Chang, H.Y. (2011) Long Intergenic Noncoding RNAs: New Links in Cancer Progression. *Cancer Research*, **71**, 3-7. https://doi.org/10.1158/0008-5472.CAN-10-2483

[7] Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L. and Gao, G. (2007) CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine. *Nucleic Acids Research*, **35**, W345-W349. https://doi.org/10.1093/nar/gkm391

[8] Lin, M.F., Jungreis, I. and Kellis, M. (2011) PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-Coding Regions. *Bioinformatics*, **27**, i275-i282. https://doi.org/10.1093/bioinformatics/btr209

[9] Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P. and Li, W. (2013) CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model. *Nucleic Acids Research*, **41**, e74. https://doi.org/10.1093/nar/gkt006

[10] 刘珊珊. 基于深度神经网络的长非编码 RNA 预测方法研究[D]: [硕士学位论文]. 扬州: 扬州大学, 2019. https://doi.org/10.27441/d.cnki.gyzdu.2019.001657

[11] Li, A., Zhang, J. and Zhou, Z. (2014) PLEK: A Tool for Predicting Long Non-Coding RNAs and Messenger RNAs Based on an Improved *k*-Mer Scheme. *BMC Bioinformatics*, **15**, Article No. 311. https://doi.org/10.1186/1471-2105-15-311

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

[13] Zhang, Y., Jia, C., Fullwood, M.J. and Kwoh, C.K. (2021) DeepCPP: A Deep Neural Network Based on Nucleotide Bias Information and Minimum Distribution Similarity Feature Selection for RNA Coding Potential Prediction. *Briefings in Bioinformatics*, **22**, 2073-2084. https://doi.org/10.1093/bib/bbaa039

[14] Baek, J., Lee, B., Kwon, S. and Yoon, S. (2018) LncRNAnet: Long Non-Coding RNA Identification Using Deep Learning. *Bioinformatics*, **34**, 3889-3897. https://doi.org/10.1093/bioinformatics/bty418

[15] Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M.D. and Zhu, H. (2018) LncADeep: An *ab Initio* LncRNA Identification and Functional Annotation Tool Based on Deep Learning. *Bioinformatics*, **34**, 3825-3834. https://doi.org/10.1093/bioinformatics/bty428

[16] 刘敬浩, 孙晓伟, 金杰. 基于主成分分析和循环神经网络的入侵检测模型[J]. 中文信息学报, 2020, 34(10): 105-112.

[17] 林寒冰, 金秀玲, 王婷, 林云霞. 基于 PCA-CNN 的动态短文本分析研究[J]. 科技创新与应用, 2022, 12(11): 44-48+52. https://doi.org/10.19981/j.CN23-1581/G3.2022.11.007