• 生物信息专栏 •

细胞穿膜肽识别问题的多特征融合卷积网络预测算法

周丰丰[1,2*]，牛甲昱[1,2]

(1. 吉林大学计算机科学与技术学院　长春　130012；2. 吉林大学符号计算与知识工程教育部重点实验室　长春　130012)

【摘要】细胞穿膜肽是一类特殊的多肽，具有独特的医学价值，因此如何通过计算方法高效地识别细胞穿膜肽是一个值得研究的重要问题。目前的主流方法是使用各种特征表示算法获取序列特征，然后使用机器学习分类器进行分类。提出了一种新的识别算法 ConvCPP，利用改进的卷积神经网络提取蛋白质序列特征。改进之处包括在卷积层之前添加注意力层，并且优化了池化层的池化方式。设计消融实验来验证改进的有效性，之后结合多种其他基于蛋白质序列特征的特征提取算法，并测试了两种特征选择算法，最终得到最优的向量表示。再根据得到的向量表示，结合多种机器学习分类器对蛋白质序列进行分类识别。在基准数据集上的实验表明，该算法比当前的细胞穿膜肽识别方法具有更好的预测性能。

关　键　词　注意力机制；　计算机应用技术；　卷积神经网络；　细胞穿膜肽；　机器学习

中图分类号　TP399　　**文献标志码**　A　　doi:10.12178/1001-0548.2021391

Integrating Multiple Feature Representations in the Convolution Neural Network Prediction Algorithm for the Cell-Penetrating Peptides

ZHOU Fengfeng[1,2*] and NIU Jiayu[1,2]

(1. College of Computer Science and Technology, Jilin University　Changchun　130012;

2. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University　Changchun　130012)

Abstract　Cell-penetrating peptides(CPPs) are a special class of peptides with unique medical value. Therefore, it's important to efficiently identify CPPs by computational methods. The current mainstream method is to use different feature extraction algorithms to extract sequence features and then use machine learning classifier for classification. This paper proposed a novel classification method: ConvCPP, which uses an improved convolution neural network to extract protein sequence features. The improvements include adding an attention layer before the convolution layer and optimizing the pooling layer. Ablation experiments were designed to verify the effectiveness of the improvement. After that, a variety of other feature extraction algorithms based on protein sequence features are combined, and two feature selection algorithms are tested, and finally the optimal vector representation is obtained. Then, according to the vector, the protein sequences are classified and recognized with a variety of machine learning classifiers. On the benchmark data set, the proposed method shows better prediction performance than the state-of-the-art CPPs predictors.

Key words　attention;　computer application technology;　convolution neural network;　cell-penetrating peptides;　machine learning

自从 30 多年前第一个细胞穿膜肽 (cell penetrating peptides, CPPs) 被发现以来，CPPs 已经在医学领域得到了广泛应用[1]。CPPs 具有能够携带各种物质 (包括小分子物质和蛋白质等) 直接进入细胞而不会显著损伤细胞膜的特性[2]。这种特性使它们可以成为将物质转运到细胞中的药物递送剂，目前已被制药公司应用于多种治疗分子的局部和全身给药的临床实验中[3]。因为 CPPs 的实用性，识别 CPPs 也成为了一个有价值的研究领域。

传统的实验方法费时费力，当前生物序列预测领域采用更加省时高效的计算方法对序列进行预测[4-5]。目前，已有不少识别 CPPs 的算法模型，其

收稿日期：2021－12－20；修回日期：2022－02－20

基金项目：国家自然科学基金 (62072212，U19A2061)；吉林省中青年科技创新创业卓越人才 (团队) 项目 (创新类)(20210509055RQ)；吉林省大数据智能计算实验室 (20180622002JC)

作者简介：周丰丰 (1977－)，男，博士，教授，主要从事健康大数据方面的研究.

*通信作者：周丰丰，E-mail：FengfengZhou@gmail.com

中大多基于机器学习算法。如文献 [6] 鉴定了 111 种已知的 CPPs 和 34 种已知的 non-CPPs，并使用基于不同生化特性的特征表示算法和支持向量机 (support vector machine, SVM) 对 CPP 进行分类。文献 [7] 提出了一种同样基于 SVM 的预测方法 CellPPD，它的改进之处在于使用了更多的特征表示算法和更大的数据集。文献 [8] 提出的 C2Pred 使用基于二肽的特征提取方法，降维后用 SVM 分类器对 CPP 进行了分类，也取得了很好的结果。文献 [9] 构建了一个基于随机森林分类器的两层预测模型，该模型不仅可以对 CPPs 和 non-CPPs 进行分类，还可以预测 CPPs 的吸收效率高低。文献 [10] 开发了 CPPred FL，该模型使用了 19 种不同的特征表示算法和多种不同的机器学习分类器。

最近也有一些研究者使用深度学习方法来进行 CPPs 的识别，如文献 [11] 基于 Transformer 模型提出了 CPPFormer，该算法根据 CPPs 序列较短的特征重构了 Transformer 网络模型，并结合了多种基于生化特性的特征表示算法进行分类。文献 [12] 提出了 DeepCPPred，是一种基于多级深度神经网络的两层预测模型。

目前基于不同的蛋白质序列特征编码算法已经有了多种识别方法，识别精度也在逐步提高。本文提出了一种基于卷积神经网络的特征表示算法 ConvCPP，获得了更好的特征表示能力。ConvCPP 的创新之处在于首次使用了基于卷积神经网络的算法进行细胞穿膜肽的特征表示工作，且在卷积神经网络加入了注意力模块以更好地提取序列特征。算法的主要流程为将蛋白质序列中的氨基酸编码为不同的向量，然后将编码后的序列输入到卷积神经网络中，提取网络的最后一层作为特征表示。最后，将卷积网络提取到的特征和一些传统特征表示方法得到的特征相结合，并集成了多种分类器得到最终的结果。实验结果表明，ConvCPP 在 SN、SP、ACC、MCC 这 4 项指标上分别达到了 0.950、0.935、0.943 和 0.885，在预测精确度 ACC 上相对当前主流分类算法有 2.2% 的提升，具有更好的分类性能。本文将模型做成了 CPPs 预测软件包，可在网址链接：https://pan.baidu.com/s/1Lx60bAQe_MfFa0QDKJ_rcw?pwd=hilb 下载使用。

1　细胞穿膜肽识别算法流程

1.1　模型框架

ConvCPP 的流程如图 1 所示。其中，特征表示分为两部分，一部分是经过改进的卷积神经网络，另一部分是基于蛋白质理化性质的传统特征表示算法，每部分得到的特征向量都使用了 T 检验进行特征选择。在分类算法部分，本文使用了朴素贝叶斯 (naive Bayes, NBayes)、随机森林 (random forest, RF)、SVM、K-最近邻算法 (K-nearest neighbor, KNN) 和极限梯度推进算法 (extreme gradient boosting, XGBoost) 进行分类，并用投票方式集成其结果作为最终分类结果。

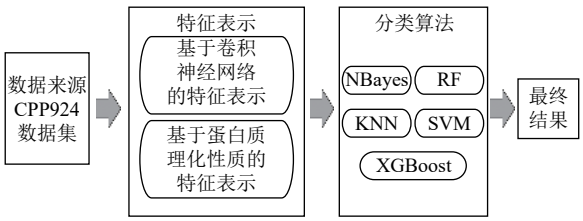

图 1　ConvCPP 模型框架

1.2　基于卷积神经网络的特征表示算法

卷积神经网络最初被发明应用于计算机视觉领域，后来逐渐在自然语言处理领域得到了广泛应用[13]，鉴于蛋白质序列和文本序列在形式上的相似性，考虑用卷积神经网络来进行特征提取。本文模型结构参考了 TextCNN 的标准结构，并在其基础上进行了改进。

首先，需要对蛋白质序列中每个氨基酸进行编码，使用了随机编码的方式将 20 种氨基酸编码为一维向量，设其长度为d，则长度为s的蛋白质序列可被表示为$s\times d$维的向量矩阵。

在将向量矩阵输入卷积层之前，加入了注意力机制来更好地获取序列的位置信息。当前注意力机制在序列数据处理领域已经得到了广泛应用[14-16]，本文使用的是多头注意力机制，表示如下：

$$\mathrm{multihead}(x)=\mathrm{concat}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_h)\boldsymbol{W}_0 \tag{1}$$

式中，$\boldsymbol{W}_0$是权重矩阵；$\boldsymbol{h}_i$表示第i个注意力头的计算结果，每个注意力头都是一个拥有不同权重矩阵的自注意力计算流程。其计算流程如下：首先，每行输入向量$\boldsymbol{x}_i$，分别和 3 个权重矩阵$\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v$相乘，得到向量$\boldsymbol{q}_i,\boldsymbol{k}_i,\boldsymbol{v}_i$。之后计算注意力权重$\boldsymbol{a}_i$，公式如下：

$$\boldsymbol{a}_i(\boldsymbol{q}_i,\boldsymbol{k}_i)=\frac{\boldsymbol{q}_i\boldsymbol{k}_i}{\sqrt{|\boldsymbol{k}_i|}} \tag{2}$$

之后对$\boldsymbol{a}_1$进行 Softmax 归一化，最后计算得到最终的输入向量z_i：

$$z_i=\sum_{i=1}^{k}\boldsymbol{a}_i\boldsymbol{v} \tag{3}$$

注意力层运算完毕之后，将结果输入卷积层。在卷积层，模型使用了多个卷积核，卷积核的个数可以作为超参数进行调整，卷积核的宽度固定为输入向量矩阵$\boldsymbol{A}$的宽度d，不同的卷积核可以设置为不同的高度。设卷积核w的大小为$h\times d$，则卷积操作的输出$\boldsymbol{o}_i$可以表示为：

$$\boldsymbol{o}_i = \boldsymbol{w}\boldsymbol{A}[i:i+h-1] \quad i=1,2,\cdots,s-h+1 \tag{4}$$

式中，$\boldsymbol{A}[i:i+h-1]$表示输入向量矩阵$\boldsymbol{A}$的第i行到第$i+h-1$行。之后叠加偏置$\boldsymbol{b}$，经过激活函数f得到特征图$\boldsymbol{c}$，其中：

$$\boldsymbol{c}_i = f(\boldsymbol{o}_i+\boldsymbol{b}) \quad i=1,2,\cdots,s-h+1 \tag{5}$$

在池化层，经典结构使用的是 1-max pooling 方法，即每个卷积核只取特征图中的最大值，最后将每个卷积核所得的值拼接，得到最终的特征向量。但是这种方法只取最大值，容易丢失信息，因此，本文将池化层改为动态 k-max pooling 方法。这种方法根据序列的长度不同，动态保留前 k 个最大值，尽可能的保留了序列的特征信息。

整个卷积神经网络模型结构如图 2 所示，模型共包含一个注意力层、一个卷积层和一个池化层，池化层的结果作为序列的特征输出。

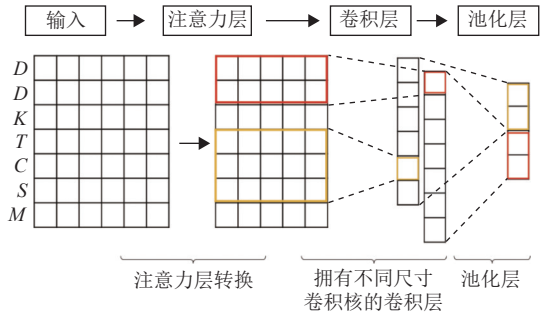

图 2 卷积神经网络模型结构

1.3 基于蛋白质理化性质的特征表示算法

除了利用卷积神经网络得到的特征，本文还结合了 3 种利用蛋白质理化性质的特征表示算法。

1) 氨基酸组成 (amino acid composition, AAC) 是最简单且常用的蛋白质序列特征表示算法之一[10]。此算法主要利用了每种类型的氨基酸在蛋白质序列中出现的频率，如下所示：

$$f(\mathrm{AAC}) = [n_1, n_2, \cdots, n_{20}] \tag{6}$$

式中，n_1表示第i种氨基酸在蛋白质序列中出现的频率，即出现次数除以序列的长度。对于给定的蛋白质序列，通过分别计算其中 20 种不同氨基酸的频率，可以得到一个 20 维的特征向量。

2) 重叠属性表示算法，文献 [17] 发现每种氨基酸通常具有不只一种物理化学性质，因此，该文根据 10 种理化性质将标准氨基酸分为 10 组，每组可能有重叠。用 10 维的 0/1 向量对每个氨基酸进行编码，其中向量的每个位置代表一个氨基酸组。如果氨基酸属于相应的氨基酸组，则向量的位置设置为 1，否则设置为 0。最后，对长度为L的蛋白质序列进行编码，得到$10\times L$大小的特征向量。

3) 20 位表示算法，文献 [7] 将 20 种氨基酸分别用 one-hot 编码方式编码，则每个氨基酸可以用 20 维的一维向量来表示，则一条长度为L的蛋白质序列可以表示为一个$20\times L$的二维向量。

2 实验结果与分析

2.1 实验环境和模型参数

在实验中，本文使用的 Python 版本为 3.6.13，Pytorch 框架的版本为 1.3.1，numpy 库的版本为 1.19.2，scipy 库的版本为 1.5.4，pandas 库版本为 1.1.5。实验使用的显卡型号为 NVIDIA Tesla P100，16 GB，驱动版本为 450.80.02，CUDA 版本为 11.0。

本文进行特征提取的卷积网络超参数的取值设置参见表 1。

表 1 模型超参数设置

超参数	值区间
批次大小/个	32
训练轮数/轮	150
随机失活率	0.5
激活函数类型	ReLu
卷积核大小	(3,4,5,6)
特征向量维度	64

2.2 数据来源及评价指标

为了便于与现有方法进行比较，本文使用了最常用的 CPPs 数据集 CPP924[18]，包含 924 个蛋白质序列，其中包括 462 个 CPPs 和 462 个 non-CPPs。CPP924 数据集中的数据都来自目前最大的 CPPs 数据集 CPPsite v2.0，且其中任意两个序列的序列相似性均小于 80%。这一点非常重要，因为过高的序列相似性可能会使模型的预测性能表现受到影响。

本文用到的评价指标包括敏感性值 (sensitivity, SN)、特异性值 (specificity, SP)、准确率 (accuracy, ACC) 和马修斯相关系数 (Mathew’s correlation

coefficient, MCC)。其计算公式分别如下：

$$SN = \frac{TP}{TP + FN} \tag{7}$$

$$SP = \frac{TN}{FP + TN} \tag{8}$$

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{10}$$

2.3　卷积神经网络模型的消融实验

为了验证本文对卷积神经网络模型改进的有效性，本文做了消融实验来验证各模块的必要性。图 3 展示了消融实验的结果，将改进分为注意力机制模块和池化层改进模块，其中模型 1 表示引入注意力机制模块和池化层改进模块的模型，模型 2 表示只引入注意力机制模块的模型，模型 3 表示只引入池化层改进模块的模型，模型 4 表示未经修改的原始模型。实验数据采用 10 倍交叉验证获得。4 组实验考察了注意力模块和池化模块的不同组合方式。

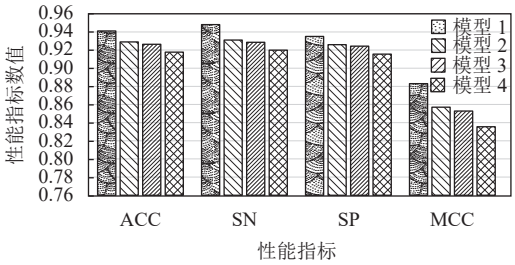

图 3　模型消融实验结果

图 3 的数据表明，注意力模块和池化模块对卷积神经网络原始模型具有较好的性能提升，且二者联合使用会对原始模型的改进程度更大。注意力模块将原始模型准确率 (ACC) 提升了 0.011，而池化模块将原始模型准确率 (ACC) 提升了 0.009。本文提出的 ConvCPP算法整合了注意力模块和池化模块的优势功能，将原始模型的准确率提升了 0.023。原始模型的马修斯相关系数 (MCC) 也提升了 0.047。

2.4　不同特征表示方法的性能对比

为了验证两种特征表示方法都是有效的，本文分别将只用卷积网络得到的特征和只用利用理化性质算法得到的特征进行分类，实验结果如图 4 所示。其中特征集 1 包括两种方法得到的特征，特征集 2 表示用卷积网络得到的卷积特征，特征集 3 表示利用理化性质算法得到的理化特征。图 4 显示特征集 1 的结果在 4 个指标上都是最好的，表明两种方法得到的特征都是有效的，而特征集 2 取得的准确率 (ACC) 优于特征集 3 至少 0.016。这表明本文提出的卷积特征提取算法相对于利用理化性质的特征表示算法，能更好地利用序列中各氨基酸之间的内在关联信息来发现 CPP 特异的序列特性。实验数据采用 10 倍交叉验证获得。本实验考察了从序列数据中提取出来的理化特征和卷积特征对 CPP 预测问题的性能贡献。

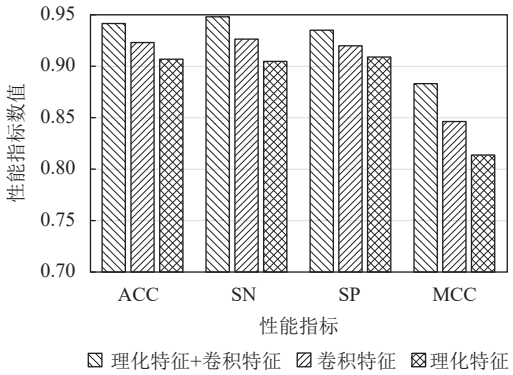

图 4　不同特征集合的实验结果

2.5　不同分类器的实验结果对比

本文使用 5 种分类器，通过投票方式集成为最终分类结果。这 5 种分类器包括 NBayes[19]、RF[20-21]、SVM[22]、KNN[23-24]、和 XGBoost[25]。每种分类器各自的性能结果以及集成后模型的分类结果如表 2 所示。可以看出，SVM 和 XGBoost 是效果相对较好的分类器，均取得了 0.930 以上的准确率 (ACC)，后续如需优化分类器数量可保留这两个分类器。其中 XGBoost 取得了最好的准确率 0.938，而集成后的分类精确度指标是最优的 (ACC=0.942)，比性能最好的独立分类器 XGBoost 的准确率还要高 0.006。实验结果说明，本文采用的投票集成法可以有效利用优势互补的各个分类器预测结果。

表 2　不同分类器的结果对比

分类器	SN	SP	ACC	MCC
NBayes	0.885	0.881	0.883	0.766
RF	0.898	0.909	0.904	0.807
SVM	0.916	0.95	0.933	0.866
KNN	0.913	0.903	0.908	0.816
XGBoost	0.952	0.924	0.938	0.877
本文方法	0.948	0.935	0.942	0.883

2.6 特征选择结果对比

为了消除两种不同特征提取方式的信息冗余，本文采用了最大相关−最小冗余 (mRMR) 和 T 检验两种不同的特征选择算法，分别对拼接后的特征向量进行特征选择，实验结果如图 5 所示，使用 mRMR 算法进行特征选择，特征数为 125 时得到了最优分类效果 (ACC=0.943)，达到了消除冗余和优化分类结果的目的。实验数据采用 10 倍交叉验证获得。本实验测试了两种不同特征选择算法对分类准确率的影响。

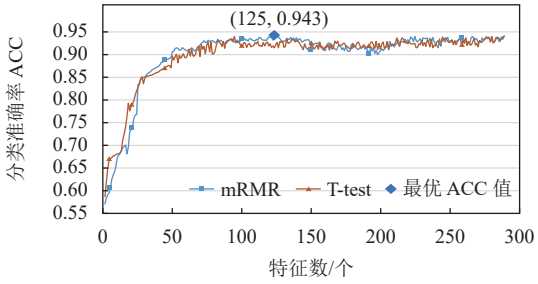

图 5 两种不同特征选择算法的实验结果

2.7 与其他模型的结果对比

为了证明本文提出的 ConvCPP 算法的有效性，本文和已发表的其他 CPPs 预测模型的结果进行了对比。这些模型都是基于 CPP924 数据集进行的预测，包括 CPPred-RF [9]、文献 [26] 提出的算法模型、SkipCPP-Pred[18]、CellPPD 和 CPPred-FL。其中，CellPPD 根据不同的特征提取算法被分为 3 个不同的模型，分别称为 CellPPD-1、CellPPD-2 和 CellPPD-3。实验的验证方法均为十折交叉验证，验证指标包括敏感性值 (SN)、特异性值 (SP)、准确率 (ACC) 和马修斯相关系数 (MCC)。表 3 给出了 4 个分类性能指标。可以看出，本文提出的 ConvCPP 模型在 4 个指标上，都优于其他模型。准确率 ACC 至少被改进了 0.022，综合性能指标 MCC 也至少被改进了 0.043。足以证明本文算法模型的有效性与精确性。

表 3 与其他模型的结果对比 /%

模型	ACC	SN	SP	MCC
CellPPD-1	90.7	90.9	90.5	81.6
CellPPD-2	87.0	83.3	90.7	74.5
CellPPD-3	83.7	78.1	89.2	68.0
SkipCPP-Pred	90.6	88.5	92.6	81.2
文献[26]	91.2	90.3	92.2	82.5
CPPred-RF	91.6	90.5	92.6	83.1
CPPred-FL	92.1	92.4	91.8	84.2
ConvCPP(本文)	94.1	94.8	93.5	88.3

3 结束语

本文提出了一种基于卷积神经网络的细胞穿膜肽识别算法 ConvCPP，结合了引入注意力机制的卷积神经网络和利用蛋白质理化性质的两种特征表示方法获取的蛋白质序列特征，并结合了多种分类器进行了最终分类。经实验验证，ConvCPP 在各项指标上都达到了最佳结果，预测准确率 ACC 相对当前其他分类算法提升了 2.2%。相关实验结果表明，深度卷积网络可以有效提取蛋白质序列的隐含模式信息，具有优秀的对细胞穿膜肽的检测能力。

参 考 文 献

[1] HEITZ F, MORRIS M C, DIVITA G. Twenty years of cell-penetrating peptides: From molecular mechanisms to therapeutics[J]. Brit J Pharmacol, 2009, 157(2): 195-206.

[2] MADANI F, LINDBERG S, LANGEL U, et al. Mechanisms of cellular uptake of cell-penetrating peptides[J]. J Biophys, 2011, 2011(6): 414729.

[3] GUIDOTTI G, BRAMBILLA L, ROSSI D. Cell-penetrating peptides: From basic research to clinics[J]. Trends Pharmacol Sci, 2017, 38(4): 406-424.

[4] LIU K W, CHEN W, LIN H. XG-PseU: An eXtreme Gradient Boosting based method for identifying pseudouridine sites[J]. Mol Genet Genomics, 2020, 295(1): 13-21.

[5] ZHANG Z Y, YANG Y H, DING H, et al. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens[J]. Brief Bioinform, 2021, 22(1): 526-535.

[6] SANDERS W S, JOHNSTON C I, BRIDGES S M, et al. Prediction of cell penetrating peptides by support vector machines[J]. Plos Comput Biol, 2011, 7(7): e1002101.

[7] GAUTAM A, CHAUDHARY K, KUMAR R, et al. In silico approaches for designing highly effective cell penetrating peptides[J]. J Trans Med, 2013, 11: 74.

[8] TANG H, SU Z D, WEI H H, et al. Prediction of cell-penetrating peptides with feature selection techniques[J]. Biochem Bioph Res Co, 2016, 477(1): 150-154.

[9] WEI L Y, XING P W, SU R, et al. CPPred-RF: A sequence-based predictor for identifying cell penetrating peptides and their uptake efficiency[J]. J Proteome Res, 2017, 16(5): 2044-2053.

[10] QIANG X L, ZHOU C, YE X C, et al. CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning[J]. Brief Bioinform, 2020, 21(1): 11-23.

[11] XUE Y Y, YE X C, WEI L S, et al. Better performance with transformer: CPPFormer in precise prediction of cell-penetrating peptides[J]. Current Medicinal Chemistry, 2022, 29(5): 881-893.

[12] ARIF M, KABIR M, AHMAD S, et al. DeepCPPred: A deep learning framework for the discrimination of cell-

penetrating peptides and their uptake efficiencies[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, DOI: 10.1109/TCBB.2021.3102133.

[13] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2014: 1725-1732.

[14] CHEN J Y, ZHANG H W, HE X N, et al. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention[C]//International ACM Sigir Conference. [S.l.]: ACM, 2017: 335-344.

[15] CHEN Q, HU Q M, HUANG X J, et al. Enhancing recurrent neural networks with positional attention for question answering[C]//Sigir'17: Proceedings of the 40th International ACM Sigir Conference on Research and Development in Information Retrieval. [S.l.]: ACM, 2017: 993-996.

[16] REN P J, CHEN Z M, REN Z C, et al. Leveraging contextual sentence relations for extractive summarization using a neural attention model[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. [S.l.]: ACM, 2017: 95-104.

[17] DOU Y C, YAO B, ZHANG C. PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine[J]. Amino Acids, 2014, 46(6): 1459-1469.

[18] WEI L Y, TANG J J, ZOU Q. SkipCPP-Pred: Promising prediction method for cell-penetrating peptides using adaptive k-Skip-n-Gram features on a high-quality dataset[J]. Lect N Bioinformat, 2016, 9683: 299-300.

[19] MAKOND B, WANG K J, WANG K M. Benchmarking prognosis methods for survivability-A case study for patients with contingent primary cancers[J]. Comput Biol Med, 2021, 138: 104888.

[20] GAO W P, WANG J S, ZHOU L, et al. Prediction of acute kidney injury in ICU with gradient boosting decision tree algorithms[J]. Comput Biol Med, 2021, 140: 105097.

[21] ERIKSSON P, MARZOUKA N A, SJODAHL G, et al. A comparison of rule-based and centroid single-sample multiclass predictors for transcriptomic classification[J]. Bioinformatics, 2021, 38(4): 1022-1029.

[22] WANG W, HAN R J, ZHANG M H, et al. A network-based method for brain disease gene prediction by integrating brain connectome and molecular network[J]. Brief Bioinform, 2022, 23(1): bbab459.

[23] ALI M M, AHMED K, BUI F M, et al. Machine learning-based statistical analysis for early stage detection of cervical cancer[J]. Comput Biol Med, 2021, 139(1): 104985.

[24] TAYLOR S, PONZINI M, WILSON M, et al. Comparison of imputation and imputation-free methods for statistical analysis of mass spectrometry data with missing data[J]. Brief Bioinform, 2022, 23(1): bbab353.

[25] LYU H, ZHANG Y, WANG J S, et al. iRice-MS: An integrated XGBoost model for detecting multitype post-translational modification sites in rice[J]. Brief Bioinform, 2022, 23(1): bbab486.

[26] DIENER C, MARTINEZ G G R, BLAS D M, et al. Effective design of multifunctional peptides by combining compatible functions[J]. Plos Comput Biol, 2016, 12(4): e1004786.

编　辑　刘飞阳